Open for Innovation
KNIME

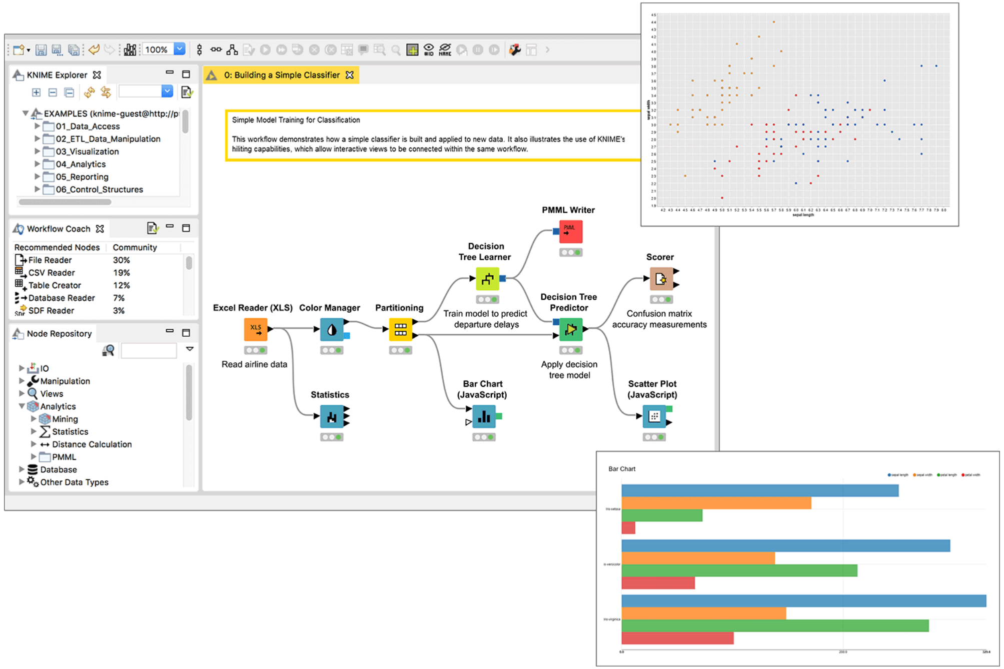
[L1+L2-DW] KNIME Analytics Platform for Data Wranglers: from Basics to Advanced

KNIME AG

Overview
KNIME Analytics Platform

What is KNIME Analytics Platform?

- A tool for data analysis, manipulation, visualization, and reporting
- Based on a graphical interface
- Provides a diverse array of extensions:
  - Text mining
  - Network mining
  - Cheminformatics
  - Many integrations, such as Java, R, Python, Weka, Keras, Plotly, H2O, etc.



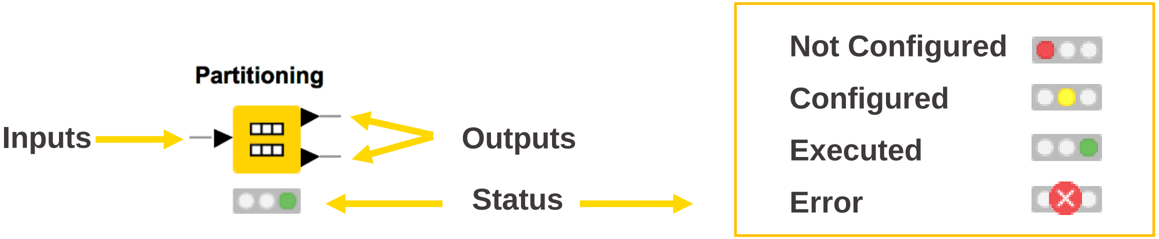
Simple Model Training for Classification

This workflow demonstrates how a simple classifier is built and applied to new data. It also illustrates the use of KNIME's hiliting capabilities, which allow interactive views to be connected within the same workflow.

Recommended Nodes	Community
File Reader	30%
CSV Reader	19%
Table Creator	12%
Database Reader	7%
SDF Reader	3%

Visual KNIME Workflows

NODES perform tasks on data



Partitioning

Inputs

Outputs

Status

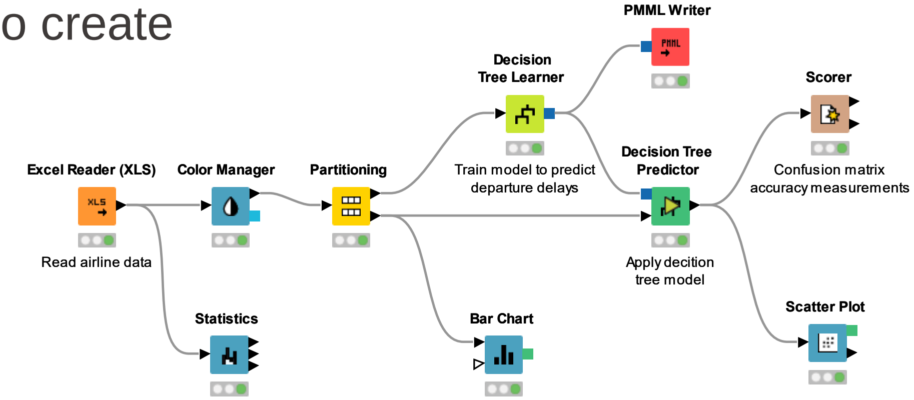
Not Configured

Configured

Executed

Error

Nodes are combined to create **WORKFLOWS**



PMML Writer

Decision Tree Learner

Train model to predict departure delays

Scorer

Confusion matrix accuracy measurements

Excel Reader (XLS)

Read airline data

Color Manager

Partitioning

Decision Tree Predictor

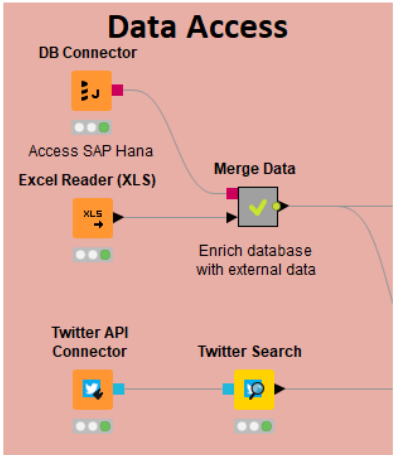
Apply decition tree model

Statistics

Bar Chart

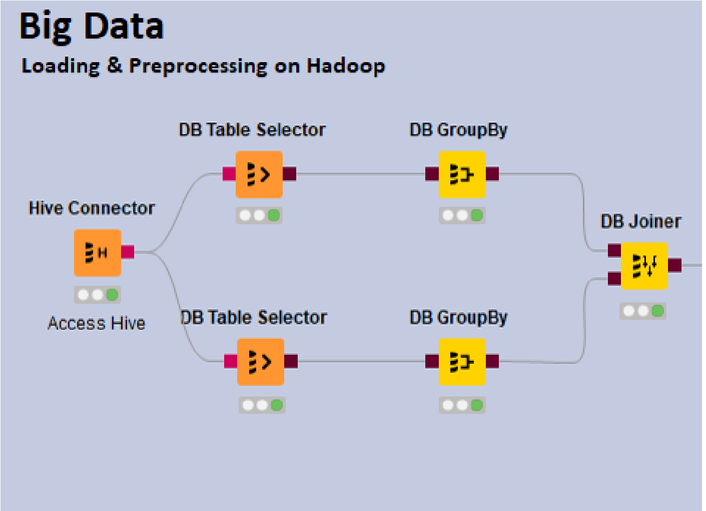
Scatter Plot

Data Access



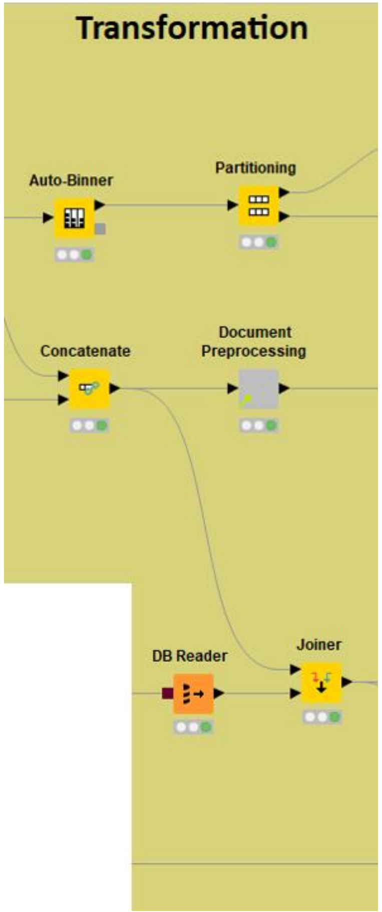
- Databases
  - MySQL, PostgreSQL, Oracle
  - Theobald
  - any JDBC (DB2, MS SQL Server)
  - Amazon DynamoDB
- Files
  - CSV, txt, Excel, Word, PDF
  - SAS, SPSS
  - XML, JSON, PMML
  - Images, texts, networks
- Other
  - Twitter, Google
  - Amazon S3, Azure Blob Store
  - Sharepoint, Salesforce
  - Kafka
  - REST, Web services

Big Data



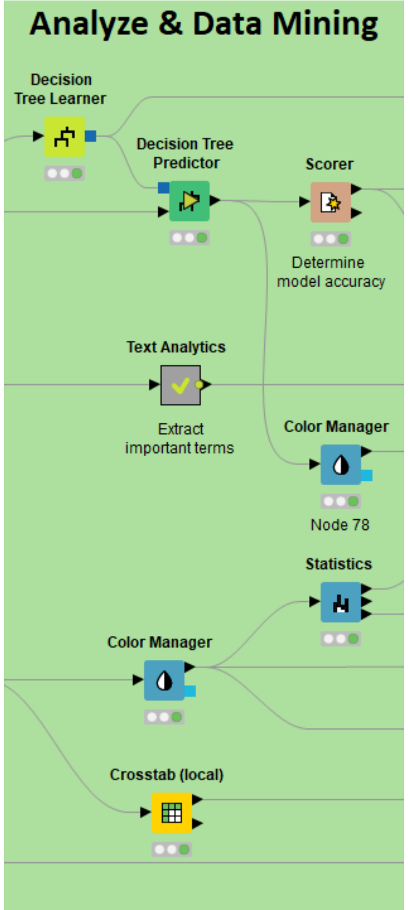
- Spark & Databricks
- HDFS support
- Hive
- Impala
- In-database processing

Transformation



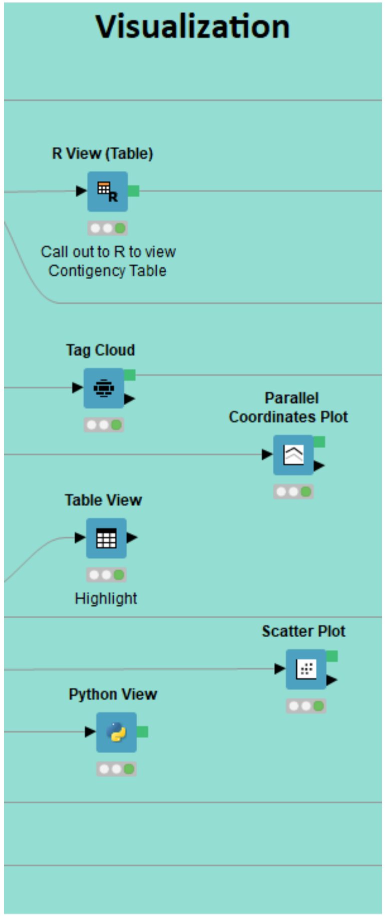
- Preprocessing
  - Row, column, matrix based
- Data blending
  - Join, concatenate, append
- Aggregation
  - Grouping, pivoting, binning
- Feature creation and selection

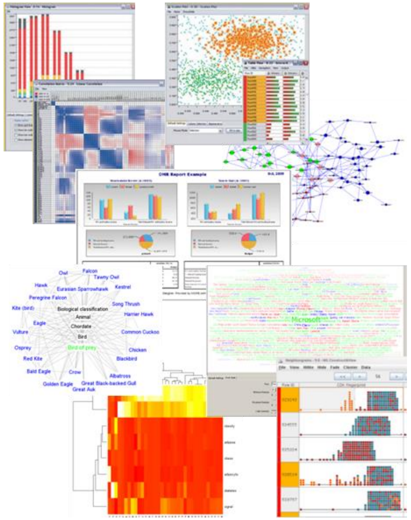
Analysis & Data Mining



- Regression
  - Linear, regression tree
- Classification
  - Decision tree, ensembles, SVM, MLP, Naïve Bayes, logistic regression
- Clustering
  - k-means, DBSCAN, hierarchical
- Validation
  - Cross-validation, scoring, ROC
- Deep Learning
  - Keras, DL4J
- External
  - R, Python, Weka, H2O, Keras

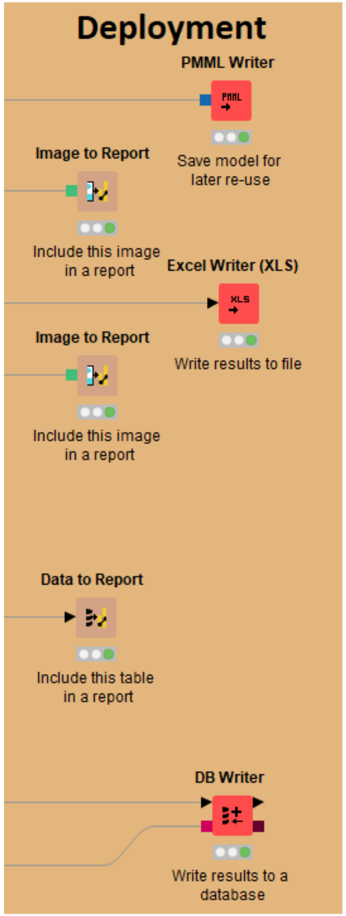
Visualization





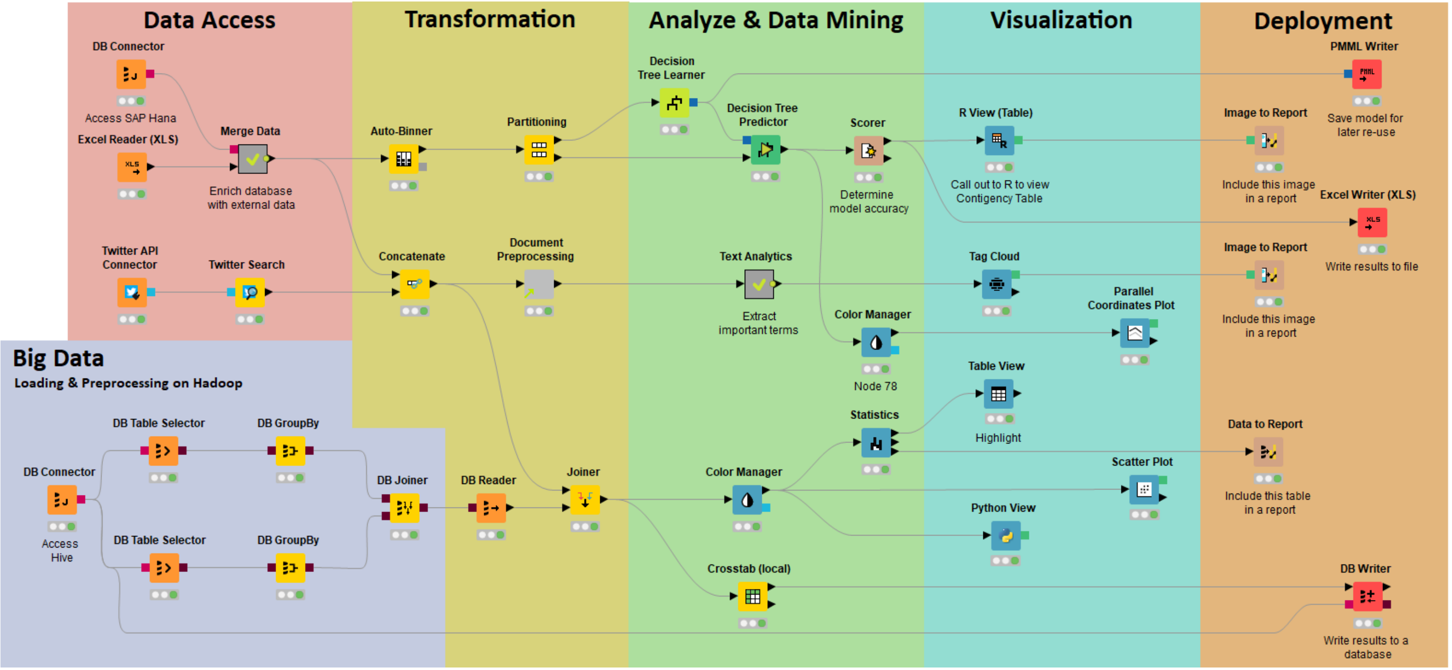
- Interactive visualizations
- JavaScript-based nodes
  - Scatter Plot, Box Plot, Line Plot
  - Networks, ROC Curve, Decision Tree
  - Plotly Integration
  - Adding more with each release!
- Misc
  - Tag cloud, open street map, molecules
- Script-based visualizations
  - R, Python

Deployment


Deployment
PMML Writer
Save model for later re-use
Image to Report
Include this image in a report
Excel Writer (XLS)
Write results to file
Image to Report
Include this image in a report
Data to Report
Include this table in a report
DB Writer
Write results to a database

- Database
- Files
  - Excel, CSV, txt
  - XML
  - PMML
  - to: local, KNIME Server, Amazon S3, Azure Blob Store
- BIRT reporting

Over 4000 Native and Embedded Nodes Included:



Data Access
MySQL, Oracle, ...
SAS, SPSS, ...
Excel, Flat, ...
Hive, Impala, ...
XML, JSON, PMML
Text, Doc, Image, ...
Web Crawlers
Industry Specific
Community / 3rd

Transformation
Row
Column
Matrix
Text, Image
Time Series
Java
Python
Community / 3rd

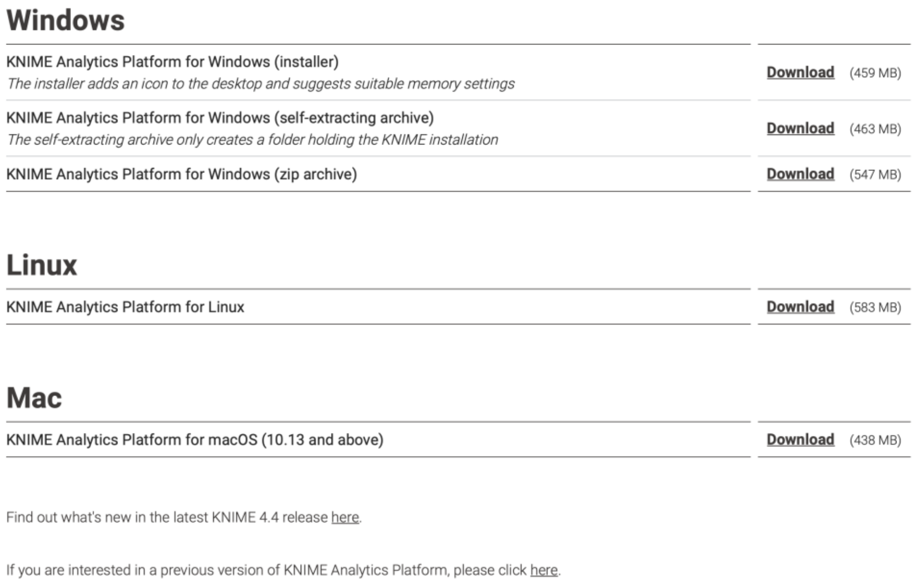
Analysis & Mining
Statistics
Data Mining
Machine Learning
Web Analytics
Text Mining
Network Analysis
Social Media Analysis
R, Weka, Python
Community / 3rd

Visualization
R
JFreeChart
JavaScript
Plotly
Community / 3rd

Deployment
via BIRT
PMML
XML, JSON
Databases
Excel, Flat, etc.
Text, Doc, Image
Industry Specific
Community / 3rd

Install KNIME Analytics Platform

- Select the KNIME version for your computer:
  - Mac
  - Windows – 32 or 64 bit
  - Linux
- Download the archive and extract the file, or download installer package and run it



Windows

KNIME Analytics Platform for Windows (installer) *The installer adds an icon to the desktop and suggests suitable memory settings*	Download	(459 MB)
KNIME Analytics Platform for Windows (self-extracting archive) *The self-extracting archive only creates a folder holding the KNIME installation*	Download	(463 MB)
KNIME Analytics Platform for Windows (zip archive)	Download	(547 MB)

Linux

KNIME Analytics Platform for Linux	Download	(583 MB)

Mac

KNIME Analytics Platform for macOS (10.13 and above)	Download	(438 MB)

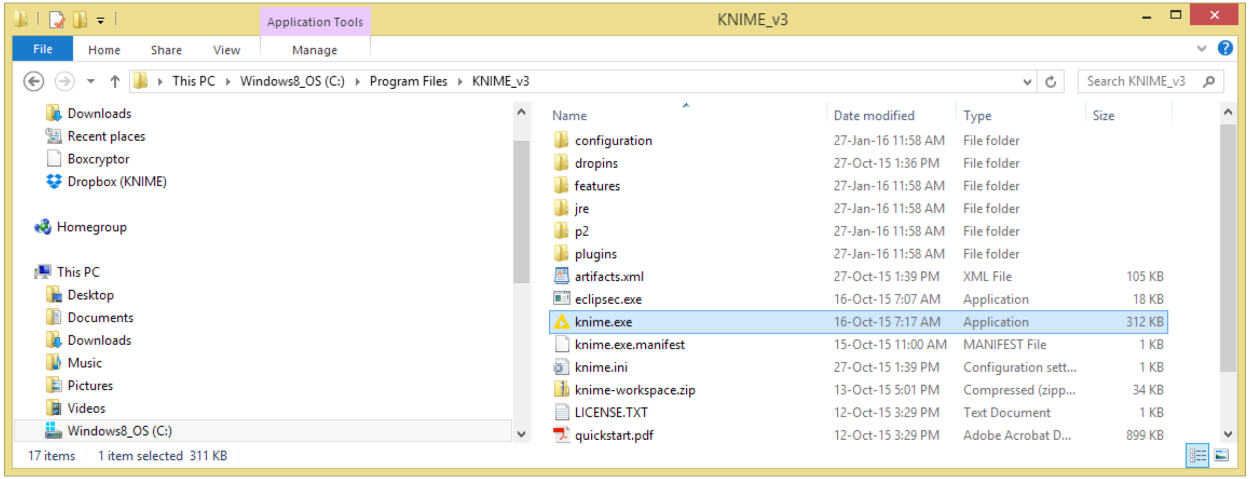
Find out what's new in the latest KNIME 4.4 release here.

If you are interested in a previous version of KNIME Analytics Platform, please click here.

Start KNIME Analytics Platform

- Use the shortcut created by the installer

- Or go to the installation directory and launch KNIME via the knime.exe

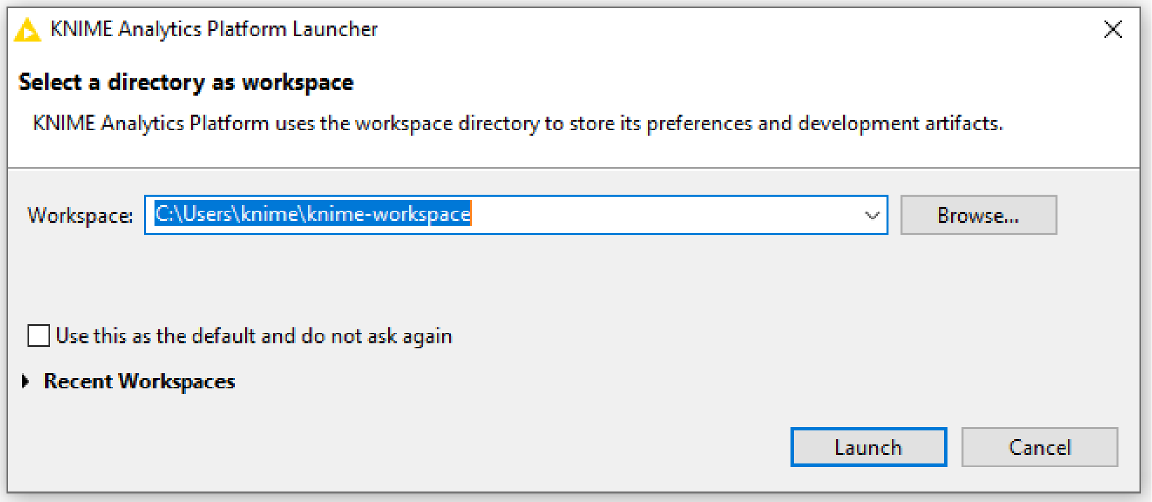


Name	Date modified	Type	Size
configuration	27-Jan-16 11:58 AM	File folder	
dropins	27-Oct-15 1:36 PM	File folder	
features	27-Jan-16 11:58 AM	File folder	
jre	27-Jan-16 11:58 AM	File folder	
p2	27-Jan-16 11:58 AM	File folder	
plugins	27-Jan-16 11:58 AM	File folder	
artifacts.xml	27-Oct-15 1:39 PM	XML File	105 KB
eclipsec.exe	16-Oct-15 7:07 AM	Application	18 KB
knime.exe	16-Oct-15 7:17 AM	Application	312 KB
knime.exe.manifest	15-Oct-15 11:00 AM	MANIFEST File	1 KB
knime.ini	27-Oct-15 1:39 PM	Configuration sett...	1 KB
knime-workspace.zip	13-Oct-15 5:01 PM	Compressed (zipp...	34 KB
LICENSE.TXT	12-Oct-15 3:29 PM	Text Document	1 KB
quickstart.pdf	12-Oct-15 3:29 PM	Adobe Acrobat D...	899 KB

17 items 1 item selected 311 KB

The KNIME Workspace

- The workspace is the **folder/directory** in which workflows (and potentially data files) are stored for the current KNIME session
- Workspaces are portable (just like KNIME)



KNIME Analytics Platform Launcher

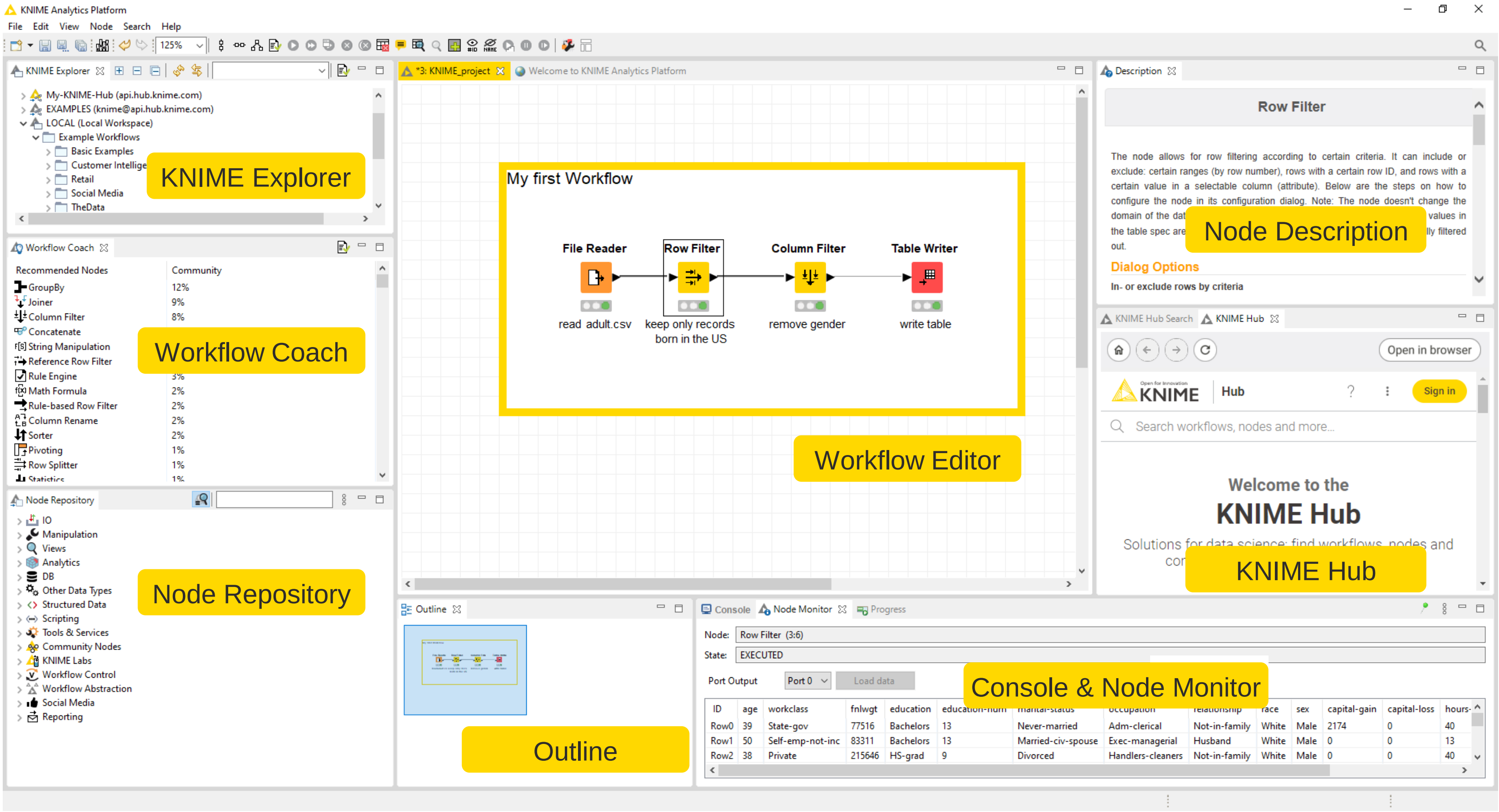
Select a directory as workspace

KNIME Analytics Platform uses the workspace directory to store its preferences and development artifacts.

Workspace: C:\Users\knime\knime-workspace

□ Use this as the default and do not ask again

The KNIME Analytics Platform Workbench



Recommended Nodes	Community
GroupBy	12%
Joiner	9%
Column Filter	8%
Concatenate	
String Manipulation	
Reference Row Filter	
Rule Engine	3%
Math Formula	2%
Rule-based Row Filter	2%
Column Rename	2%
Sorter	2%
Pivoting	1%
Row Splitter	1%
Statistics	1%

Row Filter

The node allows for row filtering according to certain criteria. It can include or exclude: certain ranges (by row number), rows with a certain row ID, and rows with a certain value in a selectable column (attribute). Below are the steps on how to configure the node in its configuration dialog. Note: The node doesn't change the domain of the dat values in the table spec are lly filtered out.

Dialog Options

In- or exclude rows by criteria

Node: Row Filter (3:6)

State: EXECUTED

Port Output: Port 0 — Load data

ID	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-
Row0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40
Row1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13
Row2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40

KNIME Explorer

Workflow Coach

Node Repository

Outline

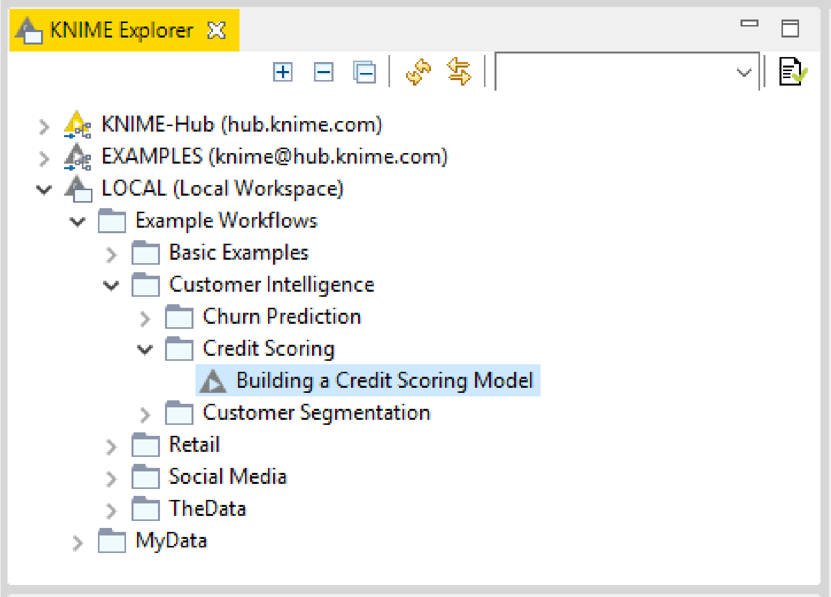
Workflow Editor

Node Description

KNIME Hub

Console & Node Monitor

KNIME Explorer

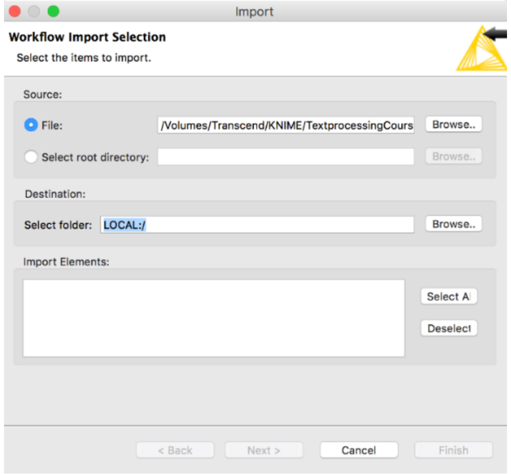


- KNIME-Hub (hub.knime.com)
- EXAMPLES (knime@hub.knime.com)
- LOCAL (Local Workspace)
  - Example Workflows
    - Basic Examples
    - Customer Intelligence
      - Churn Prediction
      - Credit Scoring
        - Building a Credit Scoring Model
      - Customer Segmentation
    - Retail
    - Social Media
    - TheData
  - MyData

- In LOCAL you can access your own workflow projects.
- Other mountpoints allow you to connect to
  - EXAMPLE Server
  - KNIME Hub
  - KNIME Server
- The Explorer toolbar on the top has a search box and buttons to
  - ⇆ select the workflow displayed in the active editor
  - ⟳ refresh the view
- The KNIME Explorer can contain 4 types of content:
  - Workflows
  - Workflow groups
  - Data files
  - Shared Components

Creating New Workflows, Importing, and Exporting

- Right-click inside the KNIME Explorer to create a new workflow or a workflow group, or to import a workflow
- Right-click the workflow or workflow group to export


Import

Workflow Import Selection

Select the items to import.

Source:

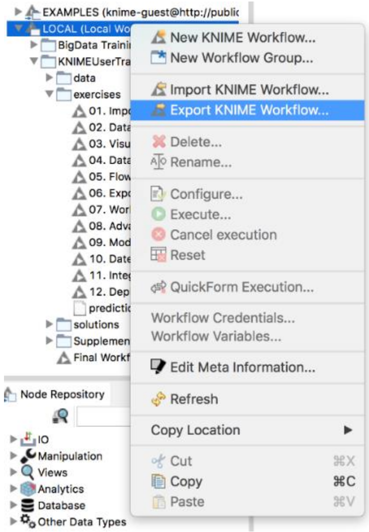
● File: /Volumes/Transcend/KNIME/TextprocessingCours

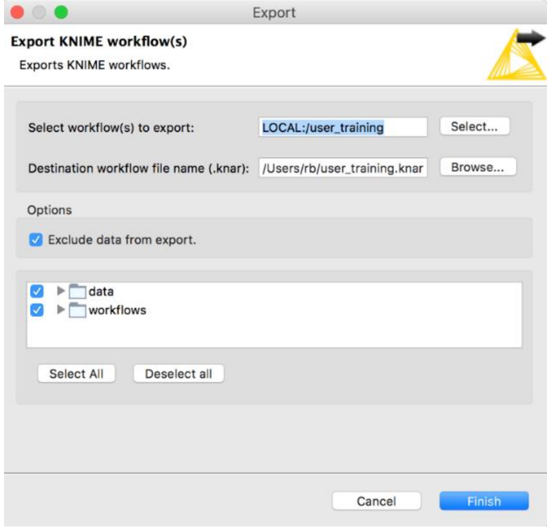
○ Select root directory:

Destination:

Select folder: LOCAL:/

Import Elements:




Export

Export KNIME workflow(s)

Exports KNIME workflows.

Select workflow(s) to export: LOCAL:/user_training

Destination workflow file name (.knar): /Users/rb/user_training.knar

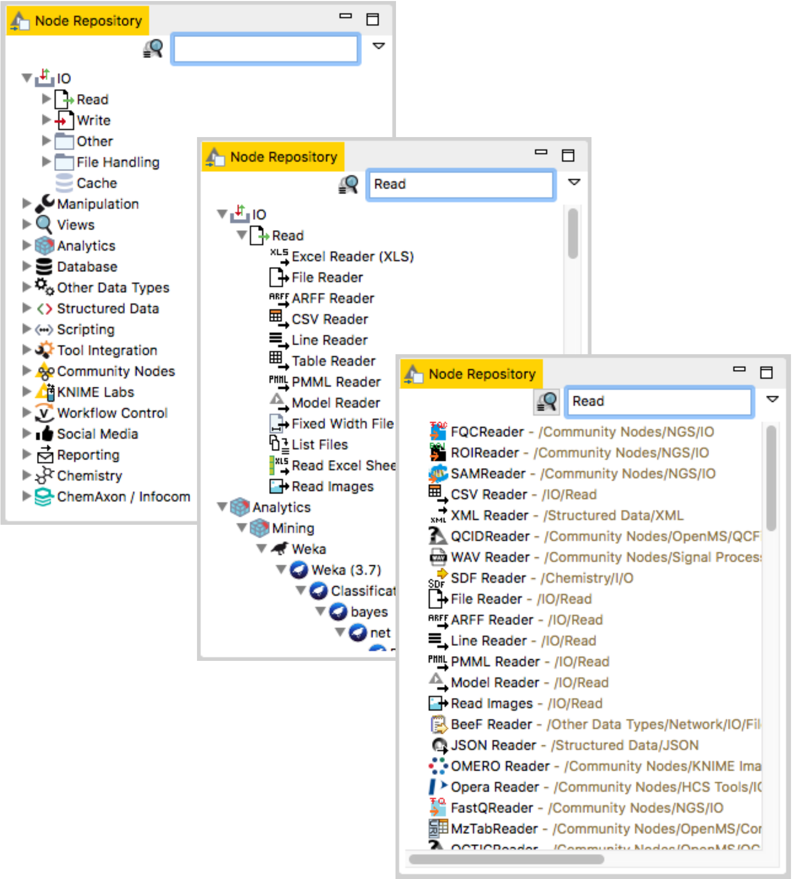
Options

☑ Exclude data from export.

☑ data

☑ workflows

Node Repository



Node Repository

- IO
  - Read
  - Write
  - Other
  - File Handling
  - Cache
- Manipulation
- Views
- Analytics
- Database
- Other Data Types
- Structured Data
- Scripting
- Tool Integration
- Community Nodes
- KNIME Labs
- Workflow Control
- Social Media
- Reporting
- Chemistry
- ChemAxon / Infocom

Node Repository

Read

- IO
  - Read
    - Excel Reader (XLS)
    - File Reader
    - ARFF Reader
    - CSV Reader
    - Line Reader
    - Table Reader
    - PMML Reader
    - Model Reader
    - Fixed Width File
    - List Files
    - Read Excel Shee
    - Read Images
- Analytics
  - Mining
    - Weka
      - Weka (3.7)
        - Classificat
          - bayes
            - net

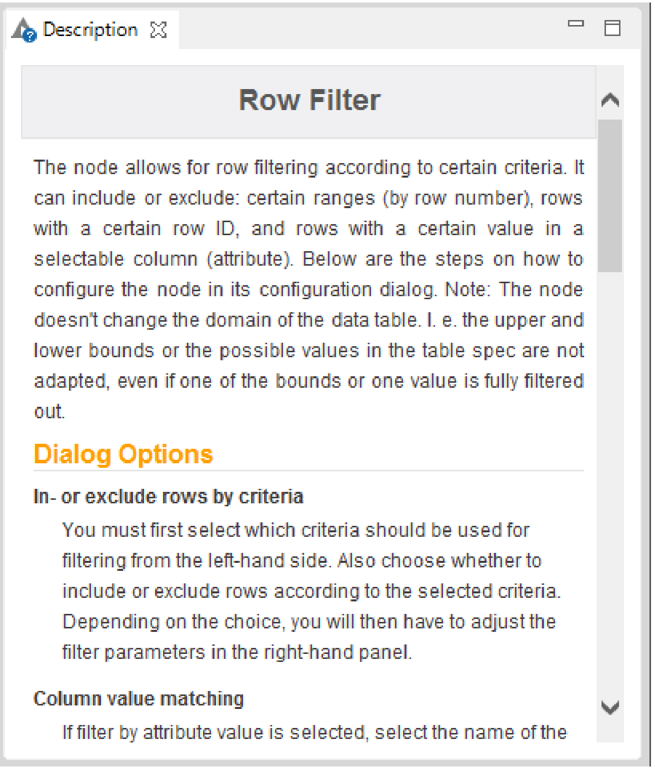
Node Repository

Read

FQCReader - /Community Nodes/NGS/IO
ROIReader - /Community Nodes/NGS/IO
SAMReader - /Community Nodes/NGS/IO
CSV Reader - /IO/Read
XML Reader - /Structured Data/XML
QCIDReader - /Community Nodes/OpenMS/QCF
WAV Reader - /Community Nodes/Signal Proces
SDF Reader - /Chemistry/I/O
File Reader - /IO/Read
ARFF Reader - /IO/Read
Line Reader - /IO/Read
PMML Reader - /IO/Read
Model Reader - /IO/Read
Read Images - /IO/Read
BeeF Reader - /Other Data Types/Network/IO/Fil
JSON Reader - /Structured Data/JSON
OMERO Reader - /Community Nodes/KNIME Ima
Opera Reader - /Community Nodes/HCS Tools/IC
FastQReader - /Community Nodes/NGS/IO
MzTabReader - /Community Nodes/OpenMS/Cor

- The Node Repository lists all KNIME nodes
- The search box has 2 modes
  - Standard Search – exact match of node name
  - Fuzzy Search – finds the most similar node name

Description



Row Filter

The node allows for row filtering according to certain criteria. It can include or exclude: certain ranges (by row number), rows with a certain row ID, and rows with a certain value in a selectable column (attribute). Below are the steps on how to configure the node in its configuration dialog. Note: The node doesn't change the domain of the data table. I. e. the upper and lower bounds or the possible values in the table spec are not adapted, even if one of the bounds or one value is fully filtered out.

Dialog Options

In- or exclude rows by criteria

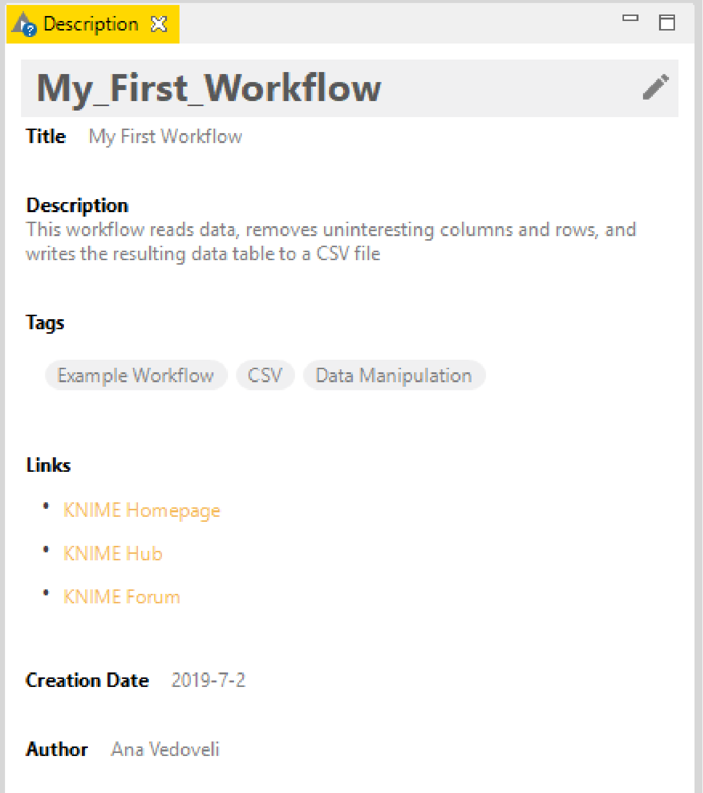
You must first select which criteria should be used for filtering from the left-hand side. Also choose whether to include or exclude rows according to the selected criteria. Depending on the choice, you will then have to adjust the filter parameters in the right-hand panel.

Column value matching

If filter by attribute value is selected, select the name of the

- The Description view provides information about:
  - Node functionality
  - Input & output
  - Node settings
  - Ports
  - References to literature

Workflow Description



My_First_Workflow

Title My First Workflow

Description
This workflow reads data, removes uninteresting columns and rows, and writes the resulting data table to a CSV file

Tags

Example Workflow CSV Data Manipulation

Links

- KNIME Homepage
- KNIME Hub
- KNIME Forum

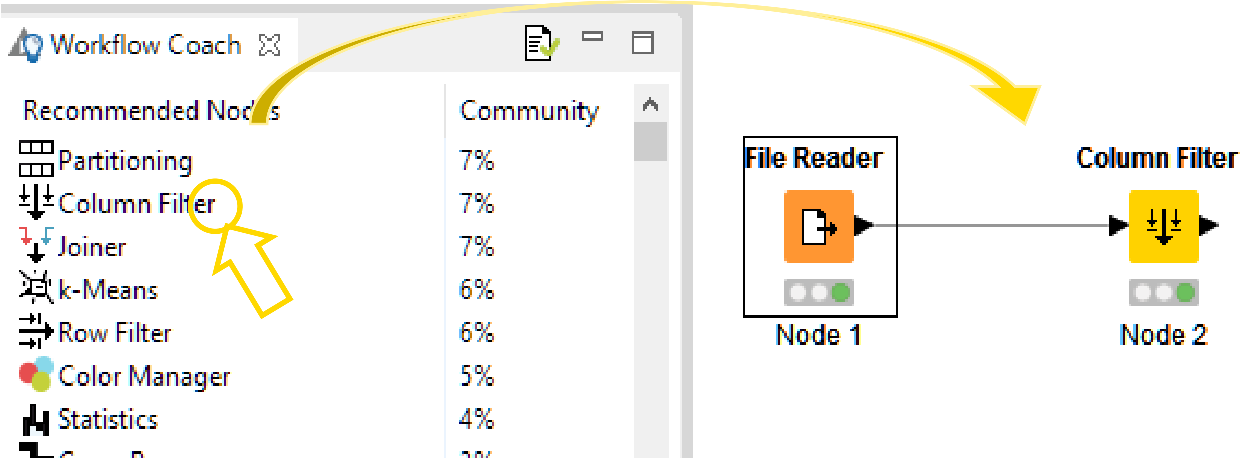
Creation Date 2019-7-2

Author Ana Vedoveli

- When selecting the workflow, the Description view gives you information about the workflow:
  - Title
  - Description
  - Associated tags and links
  - Creation date
  - Author

Workflow Coach

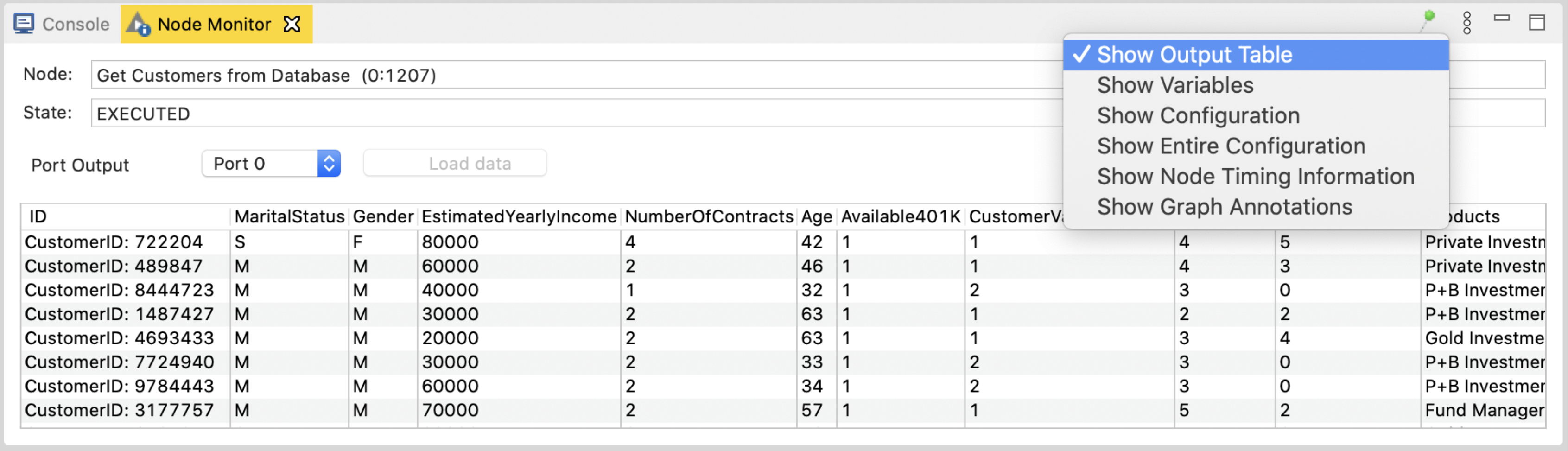
- Node recommendation engine
  - Gives hints about which node to use next in the workflow
  - Based on KNIME communities' usage statistics
  - Based on own KNIME workflows



Recommended Nodes	Community
Partitioning	7%
Column Filter	7%
Joiner	7%
k-Means	6%
Row Filter	6%
Color Manager	5%
Statistics	4%

Node Monitor

- By default the Node Monitor shows you the output table of the node selected in the workflow editor
- Click on the three dots on the upper right to show the flow variables, configuration, etc.



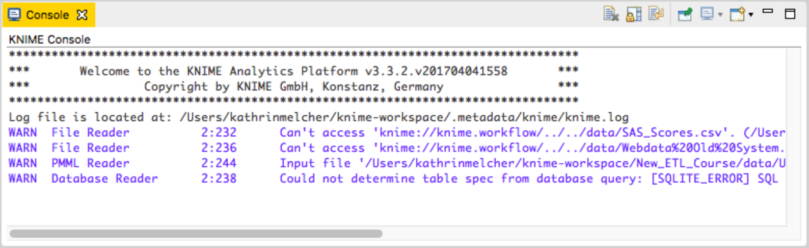
Node: Get Customers from Database (0:1207)

State: EXECUTED

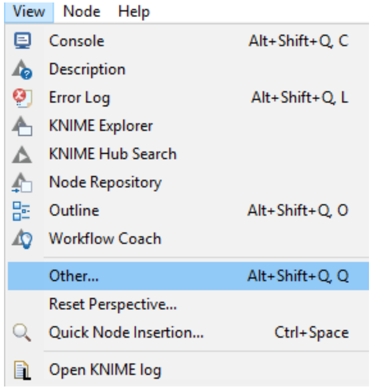
Port Output: Port 0

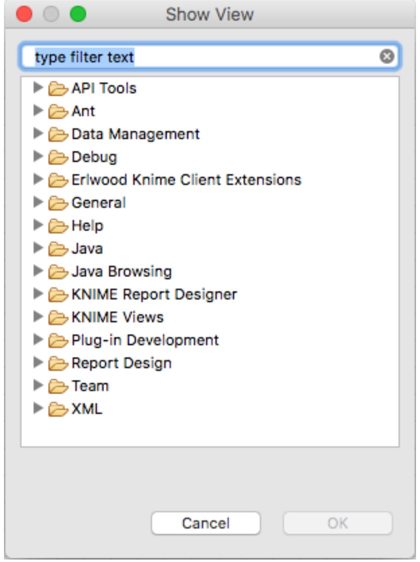
ID	MaritalStatus	Gender	EstimatedYearlyIncome	NumberOfContracts	Age	Available401K	CustomerV			ducts
CustomerID: 722204	S	F	80000	4	42	1	1	4	5	Private Investn
CustomerID: 489847	M	M	60000	2	46	1	1	4	3	Private Investn
CustomerID: 8444723	M	M	40000	1	32	1	2	3	0	P+B Investmer
CustomerID: 1487427	M	M	30000	2	63	1	1	2	2	P+B Investmer
CustomerID: 4693433	M	M	20000	2	63	1	1	3	4	Gold Investme
CustomerID: 7724940	M	M	30000	2	33	1	2	3	0	P+B Investmer
CustomerID: 9784443	M	M	60000	2	34	1	2	3	0	P+B Investmer
CustomerID: 3177757	M	M	70000	2	57	1	1	5	2	Fund Manager

Console and Other Views



```
********************************************************************************
***       Welcome to the KNIME Analytics Platform v3.3.2.v201704041558      ***
***                Copyright by KNIME GmbH, Konstanz, Germany               ***
********************************************************************************
Log file is located at: /Users/kathrinmelcher/knime-workspace/.metadata/knime/knime.log
WARN  File Reader          2:232      Can't access 'knime://knime.workflow/../../data/SAS_Scores.csv'. (/User
WARN  File Reader          2:236      Can't access 'knime://knime.workflow/../../data/Webdata%20Old%20System.
WARN  PMML Reader          2:244      Input file '/Users/kathrinmelcher/knime-workspace/New_ETL_Course/data/U
WARN  Database Reader      2:238      Could not determine table spec from database query: [SQLITE_ERROR] SQL
```



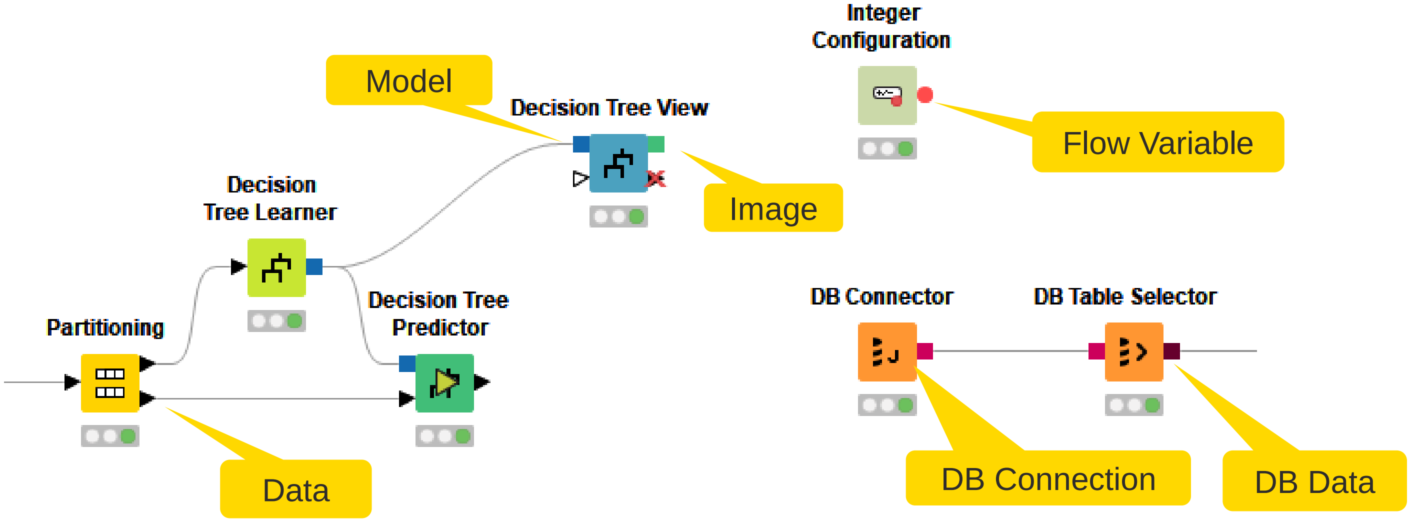


- Console view prints out error and warning messages about what is going on under the hood

- Click View and select Other… to add different views
  - Node Monitor, Licenses, etc.

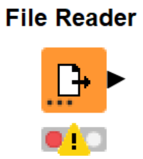
Inserting and Connecting Nodes

- Insert nodes into workspace by dragging them from the Node Repository or by double-clicking in the Node Repository
- Connect nodes by left-clicking the output port of Node A and dragging the cursor to the (matching) input port of Node B
- Common port types:

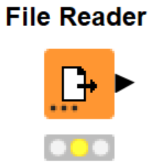


More on Nodes…

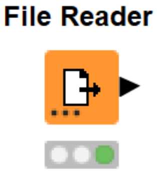
- A node can have 4 states:



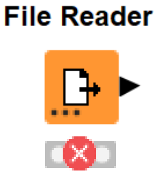
Not Configured:
The node is waiting for configuration or incoming data.



Configured:
The node has been configured correctly and can be executed.



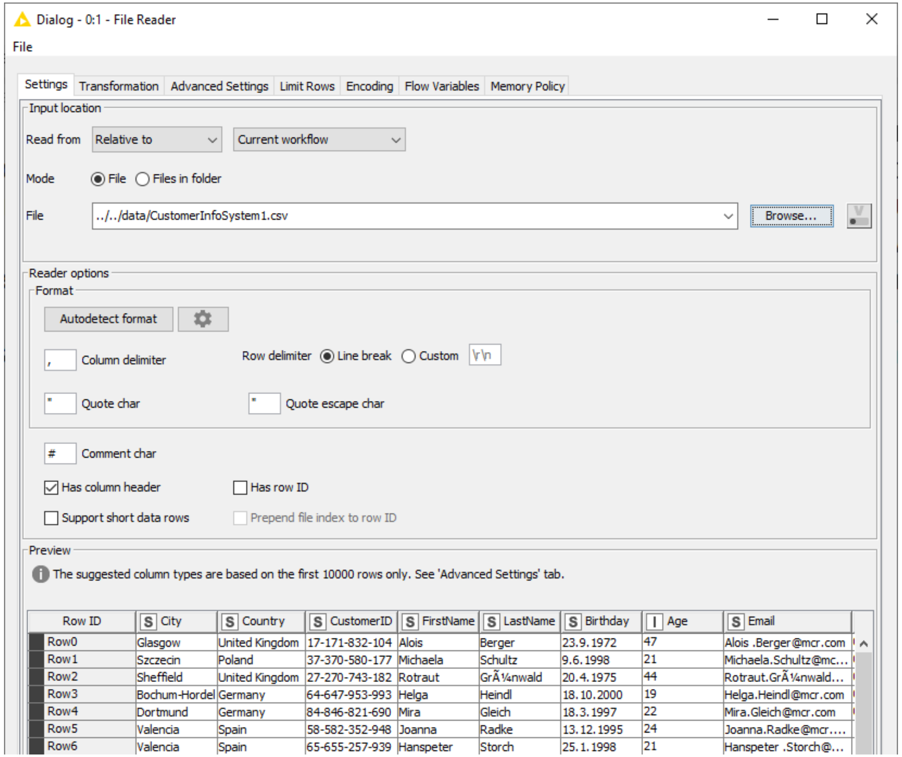
Executed:
The node has been successfully executed. Results may be viewed and used in downstream nodes.



Error:
The node has encountered an error during execution.

Node Configuration

- Most nodes need to be configured
- To access a node configuration dialog:
  - Double-click the node
  - Right-click -> Configure



Input location

Read from: Relative to | Current workflow

Mode: ● File ○ Files in folder

File: ../../data/CustomerInfoSystem1.csv

Reader options

Format

, Column delimiter

Row delimiter ● Line break ○ Custom \r\n

" Quote char

" Quote escape char

# Comment char

☑ Has column header

☐ Has row ID

☐ Support short data rows

☐ Prepend file index to row ID

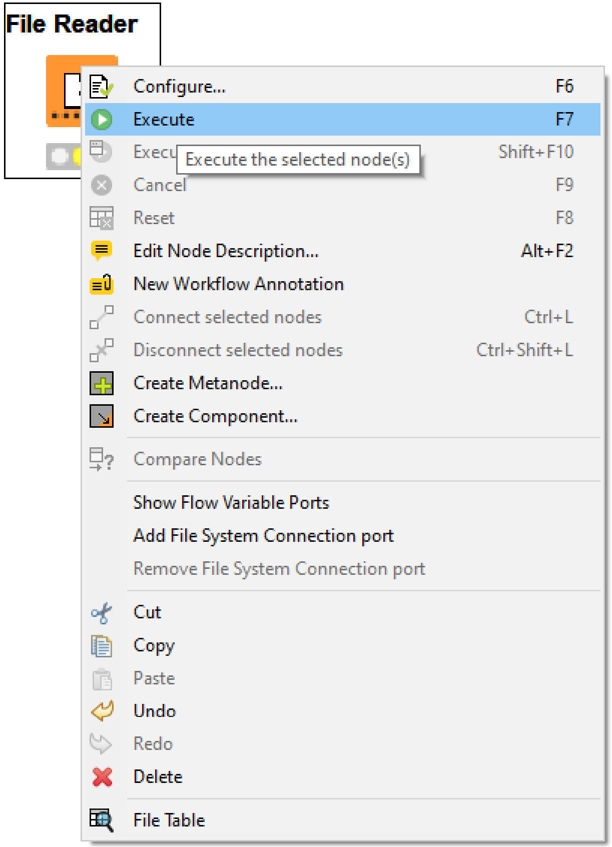
Preview

The suggested column types are based on the first 10000 rows only. See 'Advanced Settings' tab.

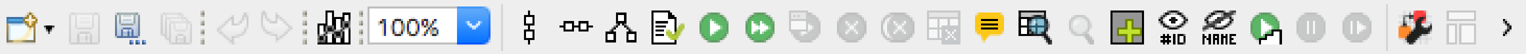
Row ID	S City	S Country	S CustomerID	S FirstName	S LastName	S Birthday	I Age	S Email
Row0	Glasgow	United Kingdom	17-171-832-104	Alois	Berger	23.9.1972	47	Alois .Berger@mcr.com
Row1	Szczecin	Poland	37-370-580-177	Michaela	Schultz	9.6.1998	21	Michaela.Schultz@mc...
Row2	Sheffield	United Kingdom	27-270-743-182	Rotraut	GrÃ¼nwald	20.4.1975	44	Rotraut.GrÃ¼nwald...
Row3	Bochum-Hordel	Germany	64-647-953-993	Helga	Heindl	18.10.2000	19	Helga.Heindl@mcr.com
Row4	Dortmund	Germany	84-846-821-690	Mira	Gleich	18.3.1997	22	Mira.Gleich@mcr.com
Row5	Valencia	Spain	58-582-352-948	Joanna	Radke	13.12.1995	24	Joanna.Radke@mcr....
Row6	Valencia	Spain	65-655-257-939	Hanspeter	Storch	25.1.1998	21	Hanspeter .Storch@...

Node Execution

- Right-click node
- Select Execute in the context menu
- If execution is successful, status shows green light
- If execution encounters errors, status shows red light



Tool Bar



- The buttons in the toolbar can be used for the active workflow. The most important buttons are:

Execute selected and executable nodes (F7)

Execute all executable nodes

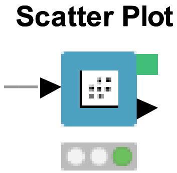
Execute selected nodes and open first view

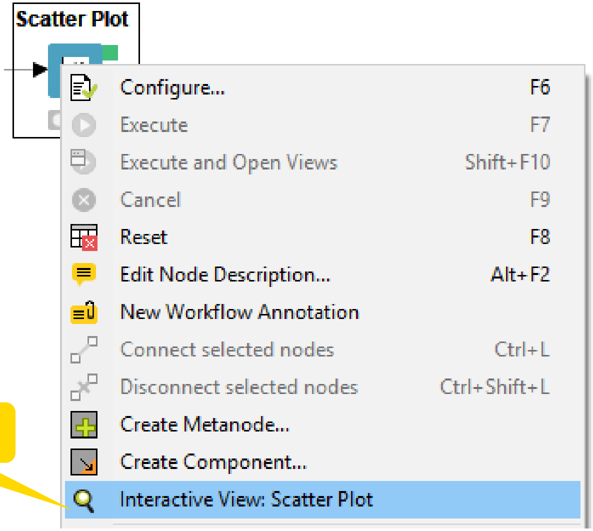
Cancel all selected, running nodes (F9)

Cancel all running nodes

Node Views

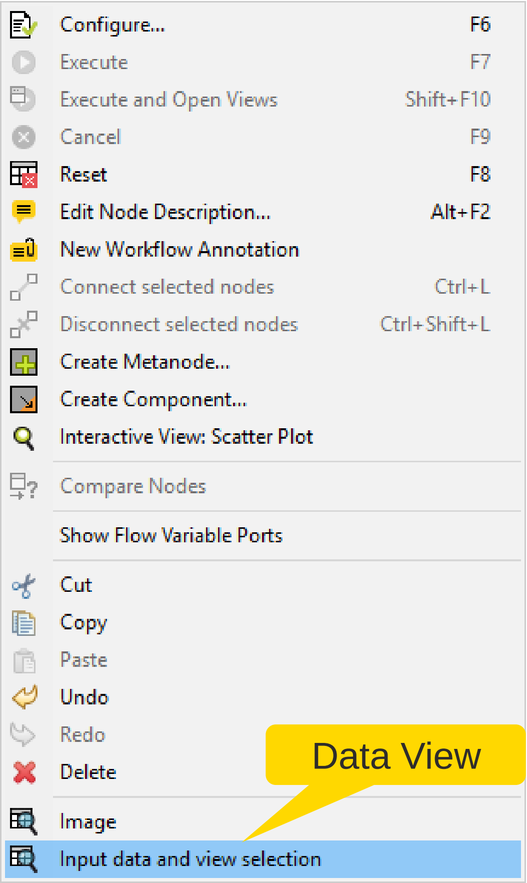
- Right-click node to inspect the execution results by
  - selecting output ports (last option in the context menu) to inspect tables, images, etc.
  - selecting Interactive View to open visualization results in a browser





Scatter Plot
Configure... F6
Execute F7
Execute and Open Views Shift+F10
Cancel F9
Reset F8
Edit Node Description... Alt+F2
New Workflow Annotation
Connect selected nodes Ctrl+L
Disconnect selected nodes Ctrl+Shift+L
Create Metanode...
Create Component...
Interactive View: Scatter Plot

Plot View



Configure... F6
Execute F7
Execute and Open Views Shift+F10
Cancel F9
Reset F8
Edit Node Description... Alt+F2
New Workflow Annotation
Connect selected nodes Ctrl+L
Disconnect selected nodes Ctrl+Shift+L
Create Metanode...
Create Component...
Interactive View: Scatter Plot
Compare Nodes
Show Flow Variable Ports
Cut
Copy
Paste
Undo
Redo
Delete
Image
Input data and view selection

Data View

KNIME File Extensions

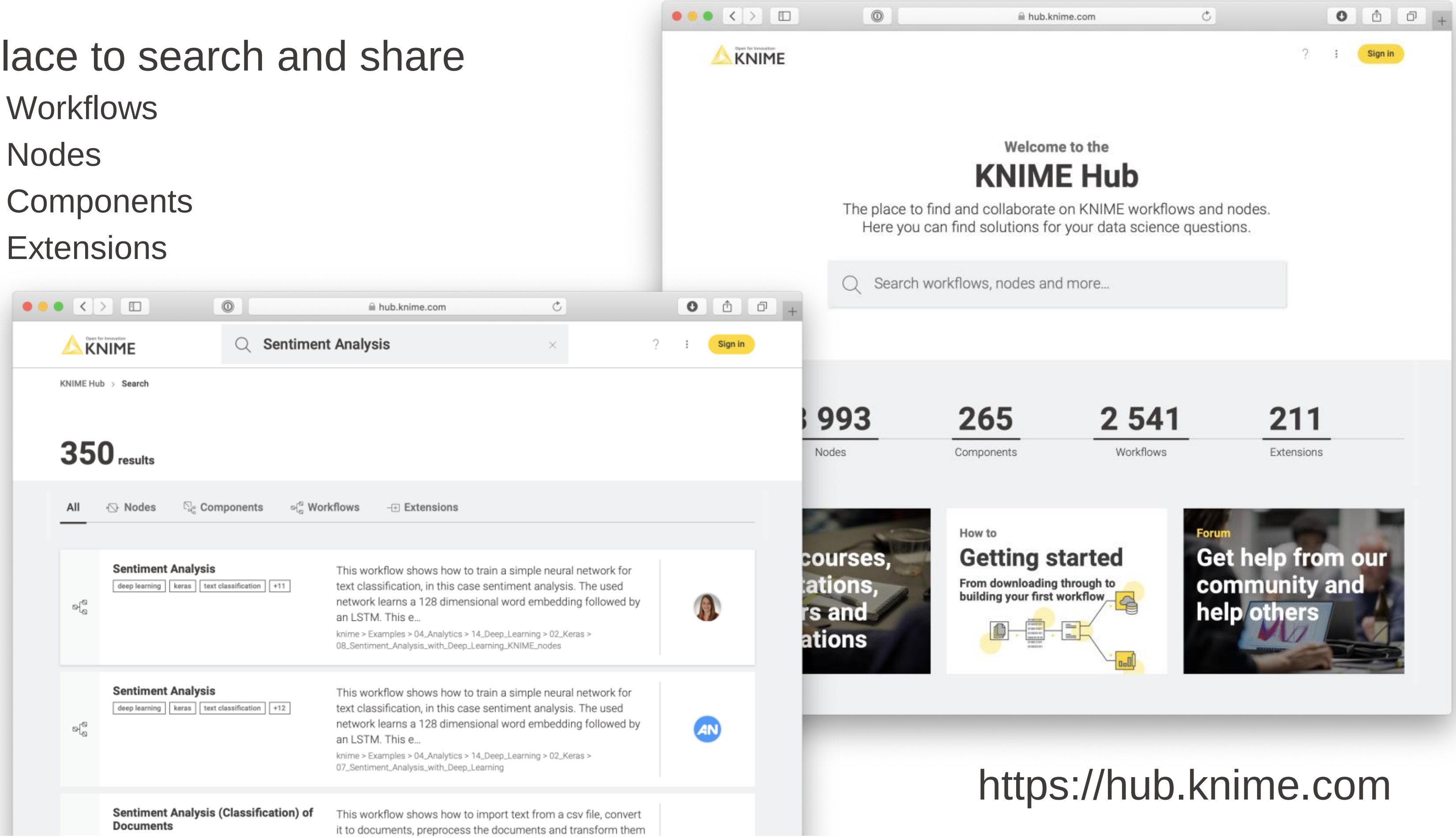
Dedicated file extensions for workflows and workflow groups associated with KNIME Analytics Platform

- ***.knwf** for KNIME Workflow Files

- ***.knar** for KNIME Archive Files

Getting Started: KNIME Hub

- Place to search and share
  - Workflows
  - Nodes
  - Components
  - Extensions



hub.knime.com

Open for Innovation KNIME

Sign in

Welcome to the

KNIME Hub

The place to find and collaborate on KNIME workflows and nodes.
Here you can find solutions for your data science questions.

Search workflows, nodes and more...

993	265	2 541	211
Nodes	Components	Workflows	Extensions

courses, ations, s and ations

How to
Getting started
From downloading through to building your first workflow

Forum
Get help from our community and help others

hub.knime.com

Sentiment Analysis

Sign in

KNIME Hub › Search

350 results

All | Nodes | Components | Workflows | Extensions

Sentiment Analysis
deep learning | keras | text classification | +11
This workflow shows how to train a simple neural network for text classification, in this case sentiment analysis. The used network learns a 128 dimensional word embedding followed by an LSTM. This e...
knime > Examples > 04_Analytics > 14_Deep_Learning > 02_Keras > 08_Sentiment_Analysis_with_Deep_Learning_KNIME_nodes

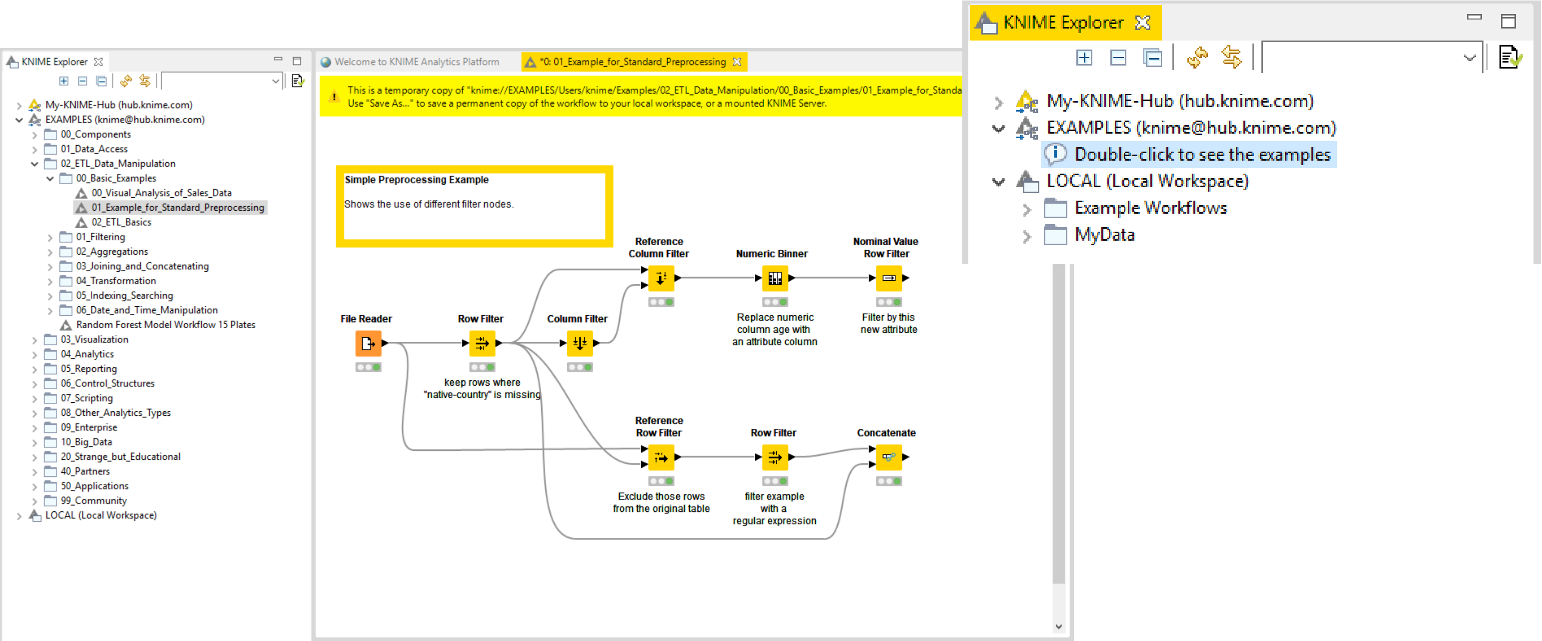
Sentiment Analysis
deep learning | keras | text classification | +12
This workflow shows how to train a simple neural network for text classification, in this case sentiment analysis. The used network learns a 128 dimensional word embedding followed by an LSTM. This e...
knime > Examples > 04_Analytics > 14_Deep_Learning > 02_Keras > 07_Sentiment_Analysis_with_Deep_Learning

Sentiment Analysis (Classification) of Documents
This workflow shows how to import text from a csv file, convert it to documents, preprocess the documents and transform them

https://hub.knime.com

Getting Started: KNIME Example Server

- Connect via KNIME Explorer to a public repository with large selection of example workflows for many, many applications



KNIME Explorer

- My-KNIME-Hub (hub.knime.com)
- EXAMPLES (knime@hub.knime.com)
  - 00_Components
  - 01_Data_Access
  - 02_ETL_Data_Manipulation
    - 00_Basic_Examples
      - 00_Visual_Analysis_of_Sales_Data
      - 01_Example_for_Standard_Preprocessing
      - 02_ETL_Basics
    - 01_Filtering
    - 02_Aggregations
    - 03_Joining_and_Concatenating
    - 04_Transformation
    - 05_Indexing_Searching
    - 06_Date_and_Time_Manipulation
    - Random Forest Model Workflow 15 Plates
  - 03_Visualization
  - 04_Analytics
  - 05_Reporting
  - 06_Control_Structures
  - 07_Scripting
  - 08_Other_Analytics_Types
  - 09_Enterprise
  - 10_Big_Data
  - 20_Strange_but_Educational
  - 40_Partners
  - 50_Applications
  - 99_Community
- LOCAL (Local Workspace)

Welcome to KNIME Analytics Platform | *0: 01_Example_for_Standard_Preprocessing

This is a temporary copy of "knime://EXAMPLES/Users/knime/Examples/02_ETL_Data_Manipulation/00_Basic_Examples/01_Example_for_Standa
Use "Save As..." to save a permanent copy of the workflow to your local workspace, or a mounted KNIME Server.

Simple Preprocessing Example

Shows the use of different filter nodes.

File Reader
Row Filter
keep rows where "native-country" is missing
Column Filter
Reference Column Filter
Numeric Binner
Replace numeric column age with an attribute column
Nominal Value Row Filter
Filter by this new attribute
Reference Row Filter
Exclude those rows from the original table
Row Filter
filter example with a regular expression
Concatenate

KNIME Explorer

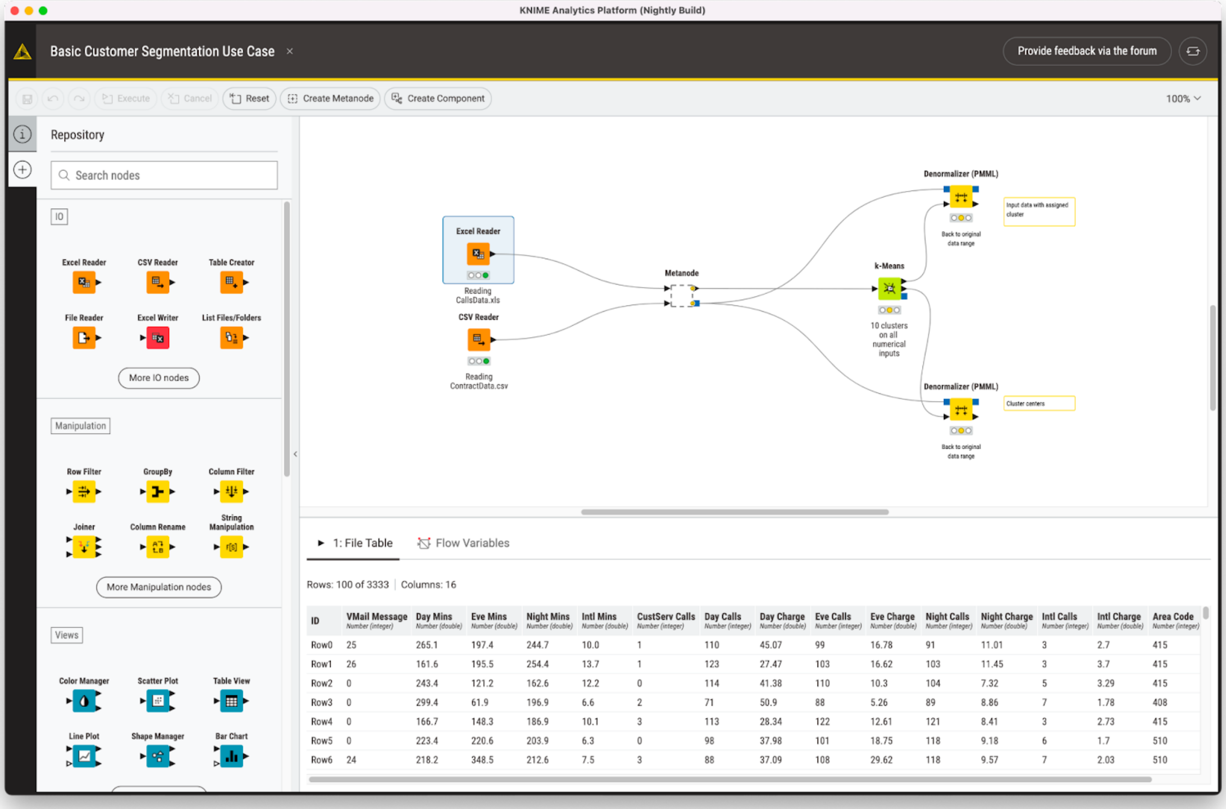
- My-KNIME-Hub (hub.knime.com)
- EXAMPLES (knime@hub.knime.com)
  - Double-click to see the examples
- LOCAL (Local Workspace)
  - Example Workflows
  - MyData

Hot Keys (for Future Reference)

Task	Hot key	Description
Node Configuration	F6	opens the configuration window of the selected node
Node Execution	F7	executes selected configured nodes
	Shift + F7	executes all configured nodes
	Shift + F10	executes all configured nodes and opens all views
	F9	cancels selected running nodes
	Shift + F9	cancels all running nodes
Node Connections	Ctrl + L	connects selected nodes
	Ctrl + Shift + L	disconnects selected nodes
Move Nodes and Annotations	Ctrl + Shift + Arrow	moves the selected node in the arrow direction
	Ctrl + Shift + PgUp/PgDown	moves the selected annotation in the front or in the back of all overlapping annotations
Workflow Operations	F8	resets selected nodes
	Ctrl + S	saves the workflow
	Ctrl + Shift + S	saves all open workflows
	Ctrl + Shift + W	closes all open workflows
Metanode	Shift + F12	opens metanode wizard

KNIME Modern UI Preview (Labs)

- Preview KNIME Analytics Platform’s makeover
  - Install KNIME Modern UI Preview extension and click the “Open KNIME Modern UI Preview”



Rows: 100 of 3333 | Columns: 16

ID	VMail Message Number (integer)	Day Mins Number (double)	Eve Mins Number (double)	Night Mins Number (double)	Intl Mins Number (double)	CustServ Calls Number (integer)	Day Calls Number (integer)	Day Charge Number (double)	Eve Calls Number (integer)	Eve Charge Number (double)	Night Calls Number (integer)	Night Charge Number (double)	Intl Calls Number (integer)	Intl Charge Number (double)	Area Code Number (integer)
Row0	25	265.1	197.4	244.7	10.0	1	110	45.07	99	16.78	91	11.01	3	2.7	415
Row1	26	161.6	195.5	254.4	13.7	1	123	27.47	103	16.62	103	11.45	3	3.7	415
Row2	0	243.4	121.2	162.6	12.2	0	114	41.38	110	10.3	104	7.32	5	3.29	415
Row3	0	299.4	61.9	196.9	6.6	2	71	50.9	88	5.26	89	8.86	7	1.78	408
Row4	0	166.7	148.3	186.9	10.1	3	113	28.34	122	12.61	121	8.41	3	2.73	415
Row5	0	223.4	220.6	203.9	6.3	0	98	37.98	101	18.75	118	9.18	6	1.7	510
Row6	24	218.2	348.5	212.6	7.5	3	88	37.09	108	29.62	118	9.57	7	2.03	510

Today’s Use Case

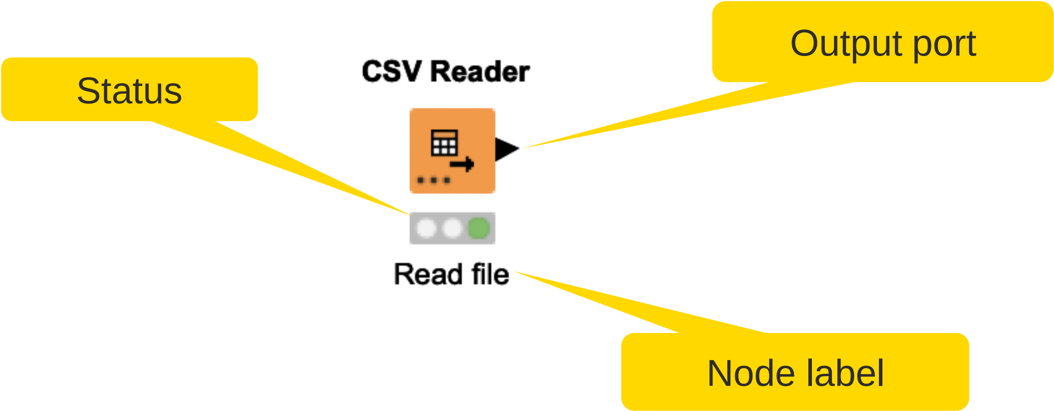
- Analyze data from a retail company, which has an online shop and stores
- Data:
  - Customer information from two different systems (.csv, .table)
  - Purchases from the online store (sqlite database)
  - List of product numbers and prices (sqlite database)
  - Purchases from the stores (.table)
  - Store information (.xls)
- Goal:
  - Single, clean table of our customers
  - Standardized list of all transactions

Importing Data

Data Source Nodes

Typically characterized by:

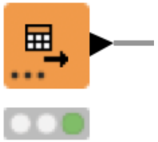
- Orange color
- By default no input ports, 1-2 output ports
- New file handling with KNIME 4.3.
  - Consistent user experience across all nodes and file systems
  - Managing of various file systems within the same workflow
  - Performance improvements

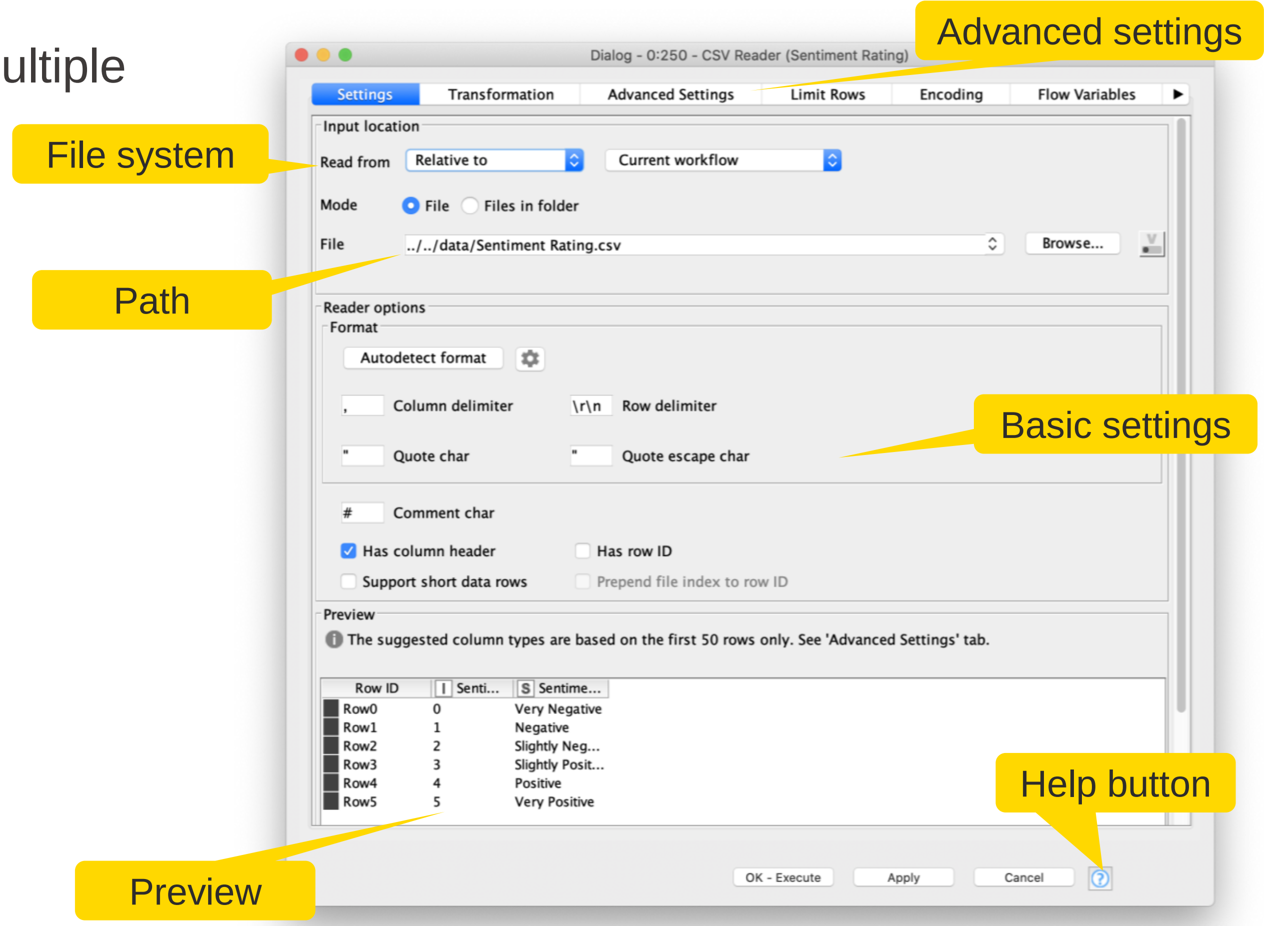


CSV Reader

- Reads either one or multiple .csv and .txt files
- Further tabs to
  - limit the rows
  - select encoding

CSV Reader





Advanced settings

File system

Path

Basic settings

Help button

Preview

Dialog - 0:250 - CSV Reader (Sentiment Rating)

Input location

Read from: Relative to — Current workflow

Mode: ● File ○ Files in folder

File: ../../data/Sentiment Rating.csv

Reader options

Format

, Column delimiter — \r\n Row delimiter

" Quote char — " Quote escape char

Comment char

☑ Has column header — □ Has row ID

□ Support short data rows — □ Prepend file index to row ID

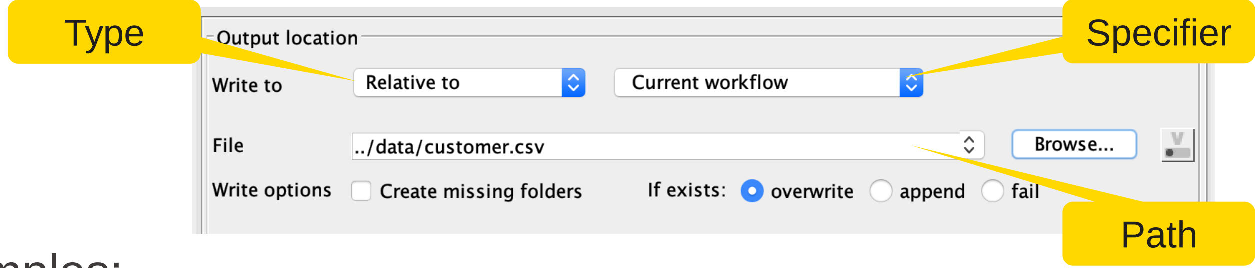
Preview

The suggested column types are based on the first 50 rows only. See 'Advanced Settings' tab.

Row ID	I Senti...	S Sentime...
Row0	0	Very Negative
Row1	1	Negative
Row2	2	Slightly Neg...
Row3	3	Slightly Posit...
Row4	4	Positive
Row5	5	Very Positive

Common Settings: File Path

- A path consists of three parts:
  - **Type**: Specifies the file system type e.g. local, relative, mountpoint, custome_url or connected.
  - **Specifier**: Optional string with additional file system specific information e.g. relative to which location (knime.workflow)
  - **Path**: Specifies the location within the file system



Output location

Write to: Relative to | Current workflow

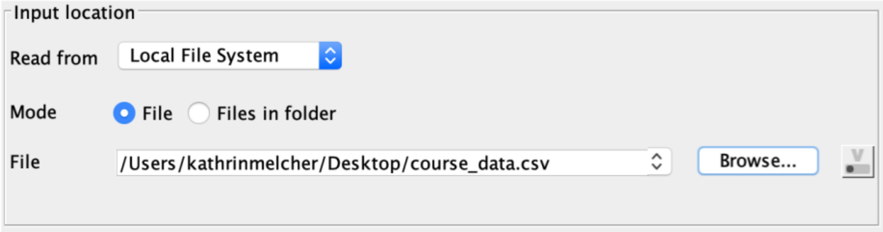
File: ../data/customer.csv Browse...

Write options: □ Create missing folders If exists: ● overwrite ○ append ○ fail

- Examples:
  - (LOCAL, , C:\Users\username\Desktop)
  - (RELATIVE, knime.workflow, file1.csv)
  - (MOUNTPOINT, MOUNTPOINT_NAME, /path/to/file1.csv)
  - (CONNECTED, amazon-s3:eu-west-1, /mybucket/file1.csv)

Common Settings: Four Default File Systems

- ## Local File System



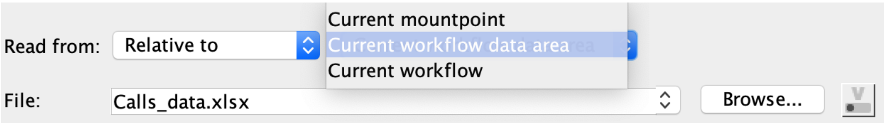
Input location

Read from: Local File System

Mode: ● File ○ Files in folder

File: /Users/kathrinmelcher/Desktop/course_data.csv

- ## Relative to …

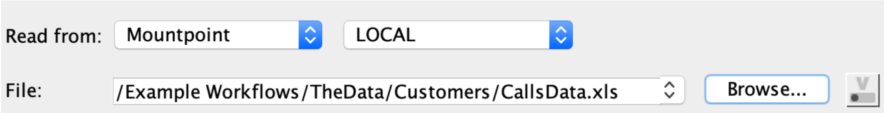


Read from: Relative to

Current mountpoint
Current workflow data area
Current workflow

File: Calls_data.xlsx

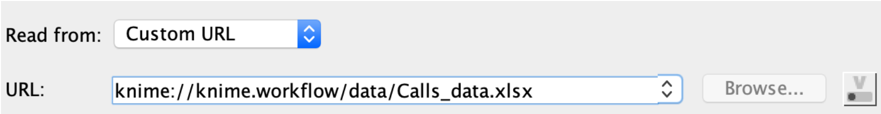
- ## Mountpoint



Read from: Mountpoint LOCAL

File: /Example Workflows/TheData/Customers/CallsData.xls

- ## Custom URL

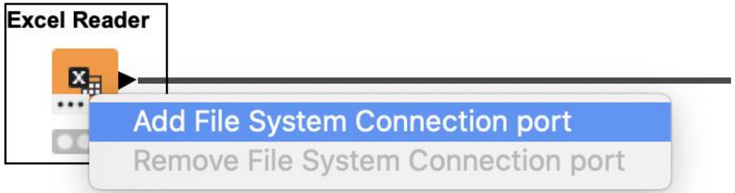


Read from: Custom URL

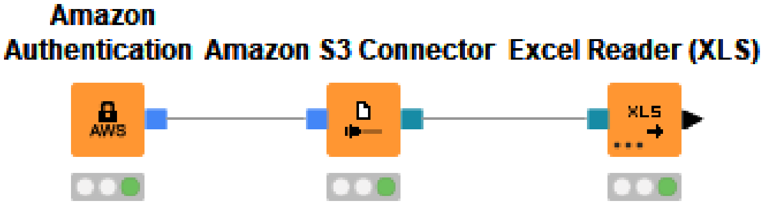
URL: knime://knime.workflow/data/Calls_data.xlsx

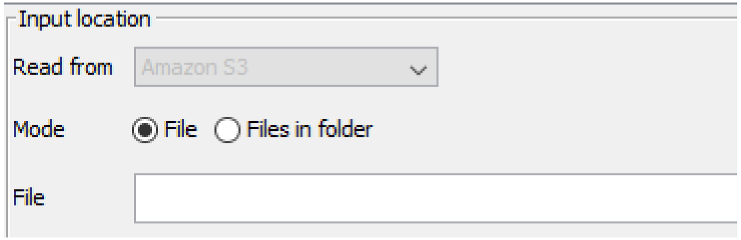
Common Settings: Connecting to other File Systems

- Add file system connection port to connect to another file system
  - Click on the three dots on the lower left to add or remove a dynamic port.


Excel Reader
Add File System Connection port
Remove File System Connection port

- Supported file systems
  - Microsoft Azure
  - Google
  - Amazon
  - Databricks
  - BigData file systems (hdfs, httpFS, …)
  - On-premise (e.g. ssh, ftp, …)




Input location

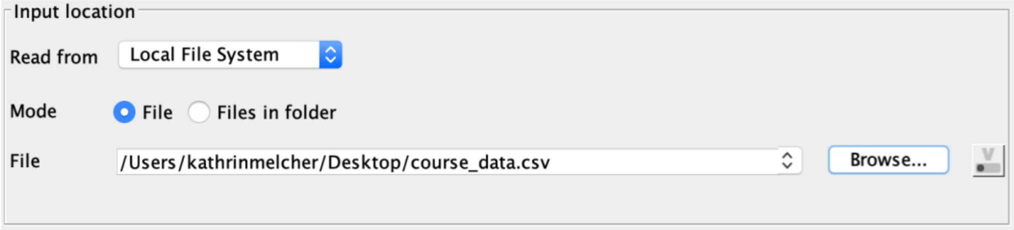
Read from: Amazon S3

Mode: ● File ○ Files in folder

File:

Common Settings: Read Single or Multiple Files

- Single file



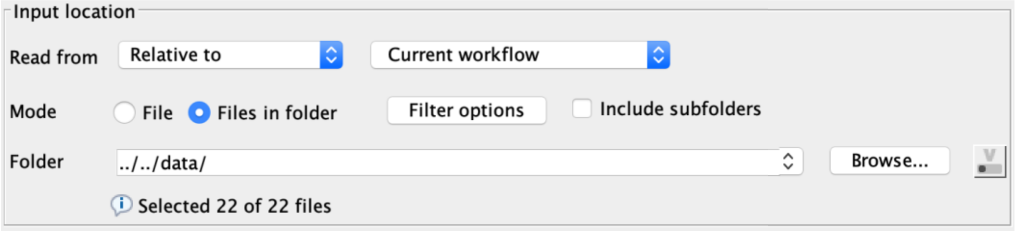
Input location

Read from: Local File System

Mode: ● File ○ Files in folder

File: /Users/kathrinmelcher/Desktop/course_data.csv

- Files in a folder



Input location

Read from: Relative to — Current workflow

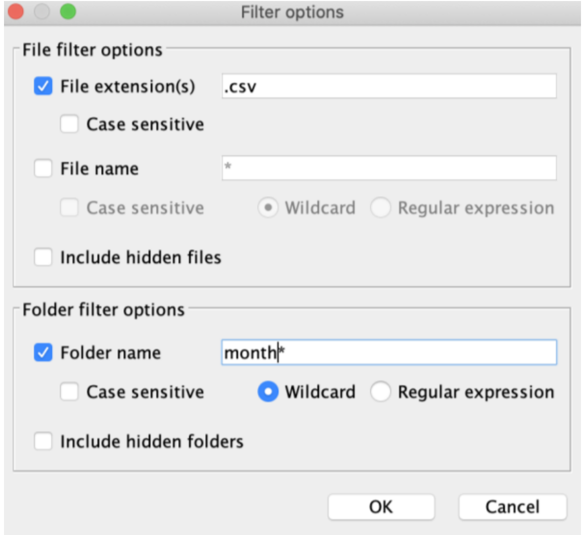
Mode: ○ File ● Files in folder

□ Include subfolders

Folder: ../../data/

Selected 22 of 22 files

  - Option to include subfolder
  - Option to define filter criterions



Filter options

File filter options

☑ File extension(s): .csv

□ Case sensitive

□ File name: *

□ Case sensitive ● Wildcard ○ Regular expression

□ Include hidden files

Folder filter options

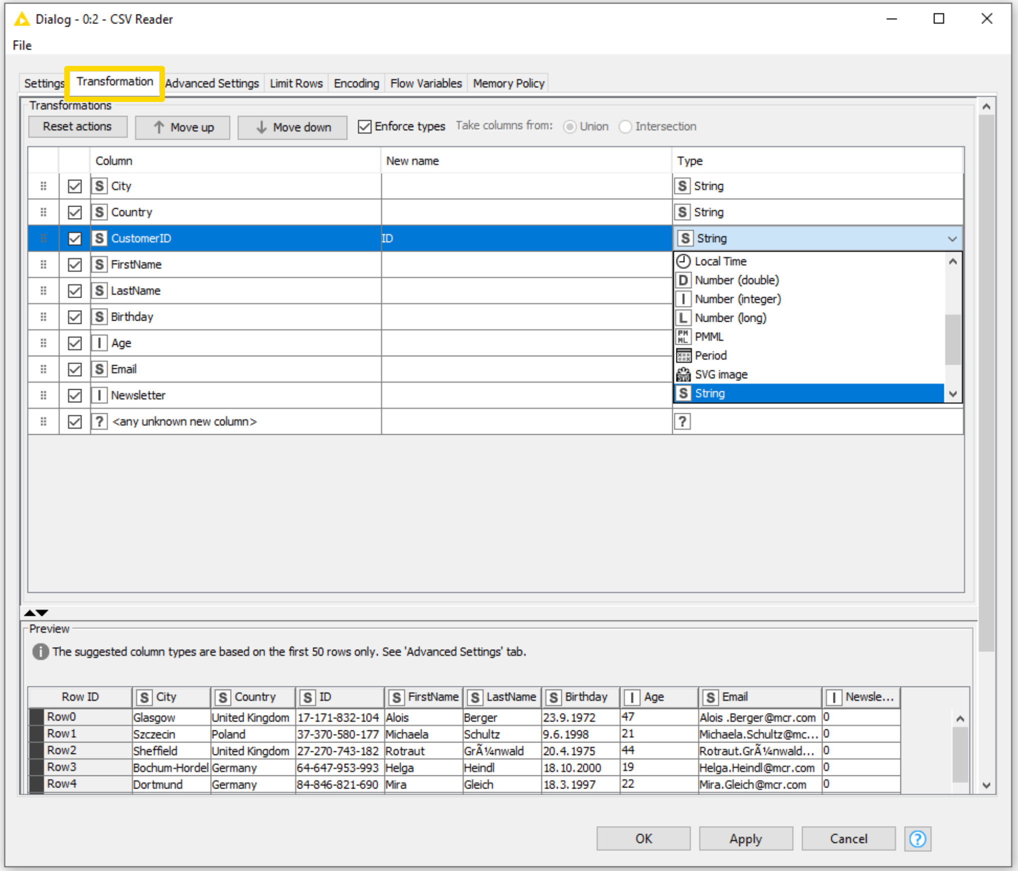
☑ Folder name: month*

□ Case sensitive ● Wildcard ○ Regular expression

□ Include hidden folders

Common Settings: Transformation Tab

- Supported operations
  - Column filtering
  - Column sorting
  - Column renaming
  - Column type mapping
  - Select between union or intersection of columns (in case of reading many files)

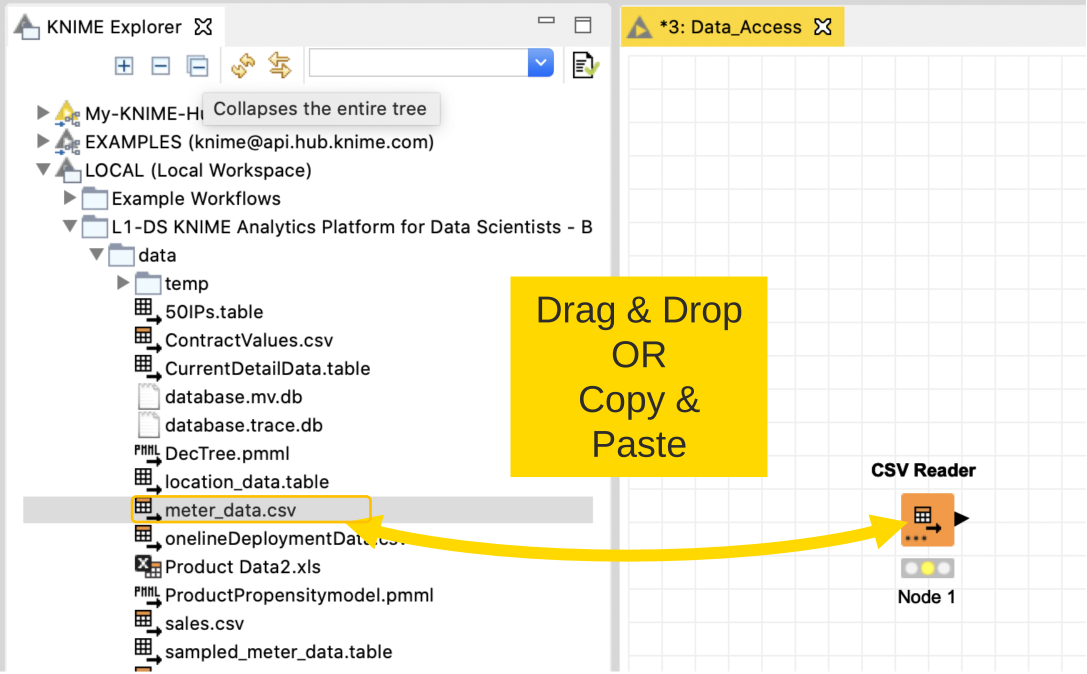


	Column	New name	Type
☑	S City		S String
☑	S Country		S String
☑	S CustomerID	ID	S String
☑	S FirstName		
☑	S LastName		
☑	S Birthday		
☑	I Age		
☑	S Email		
☑	I Newsletter		
☑	? <any unknown new column>		?

The suggested column types are based on the first 50 rows only. See 'Advanced Settings' tab.

Row ID	S City	S Country	S ID	S FirstName	S LastName	S Birthday	I Age	S Email	I Newsle...
Row0	Glasgow	United Kingdom	17-171-832-104	Alois	Berger	23.9.1972	47	Alois .Berger@mcr.com	0
Row1	Szczecin	Poland	37-370-580-177	Michaela	Schultz	9.6.1998	21	Michaela.Schultz@mc...	0
Row2	Sheffield	United Kingdom	27-270-743-182	Rotraut	GrÃ¼nwald	20.4.1975	44	Rotraut.GrÃ¼nwald...	0
Row3	Bochum-Hordel	Germany	64-647-953-993	Helga	Heindl	18.10.2000	19	Helga.Heindl@mcr.com	0
Row4	Dortmund	Germany	84-846-821-690	Mira	Gleich	18.3.1997	22	Mira.Gleich@mcr.com	0

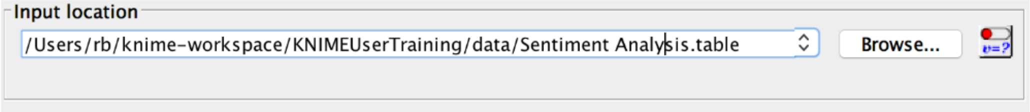
Alternative Faster Way …



Drag & Drop
OR
Copy &
Paste

File Path Options Old File Handling

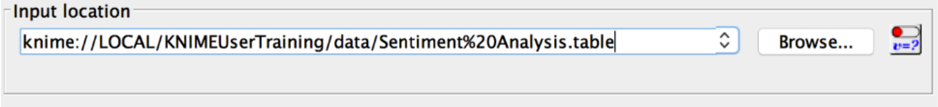
- Local path



Input location

/Users/rb/knime-workspace/KNIMEUserTraining/data/Sentiment Analysis.table

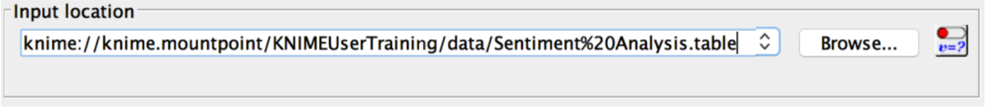
- Absolute URL



Input location

knime://LOCAL/KNIMEUserTraining/data/Sentiment%20Analysis.table

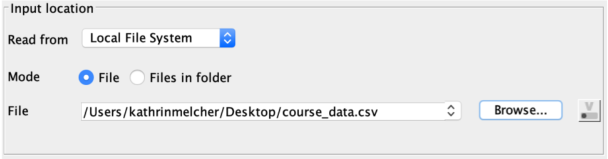
- Mountpoint-relative URL



Input location

knime://knime.mountpoint/KNIMEUserTraining/data/Sentiment%20Analysis.table

New file handling

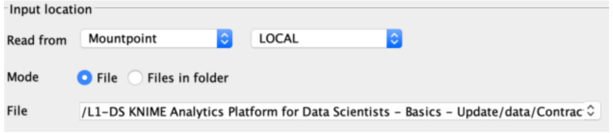


Input location

Read from: Local File System

Mode: ● File ○ Files in folder

File: /Users/kathrinmelcher/Desktop/course_data.csv

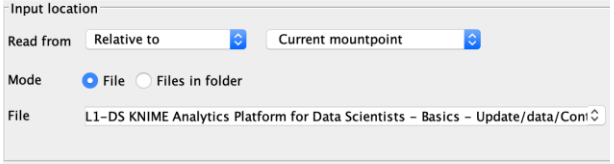


Input location

Read from: Mountpoint, LOCAL

Mode: ● File ○ Files in folder

File: /L1-DS KNIME Analytics Platform for Data Scientists – Basics – Update/data/Contrac



Input location

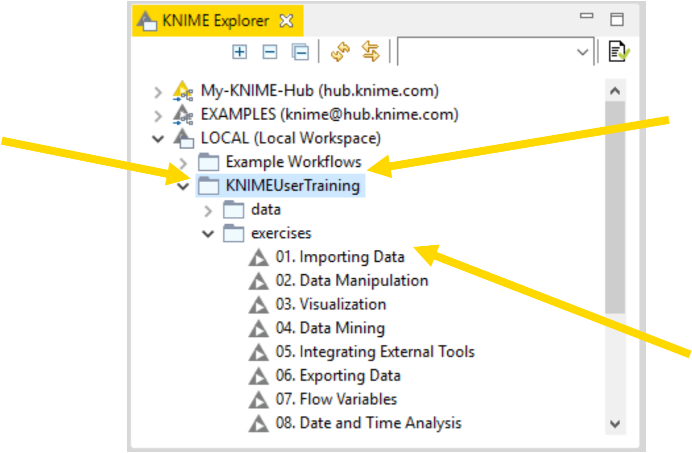
Read from: Relative to, Current mountpoint

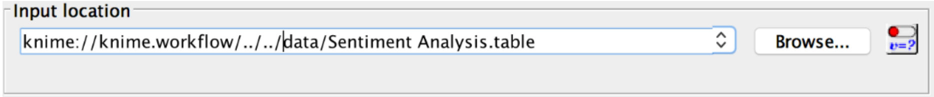
Mode: ● File ○ Files in folder

File: L1-DS KNIME Analytics Platform for Data Scientists – Basics – Update/data/Cont

Workflow-Relative File Paths (Old File Handling)

- Best choice if workflows are to be shared
- Requires matching folder structure within workflow group
  - Independent of environment outside of workflow group
- Example: Path to „Sentiment Analysis.table“
  - Local path:
  - C:\Users\rb\knime-workspace\KNIMEUserTraining\data\Sentiment Analysis.table
    - Workflow relative:





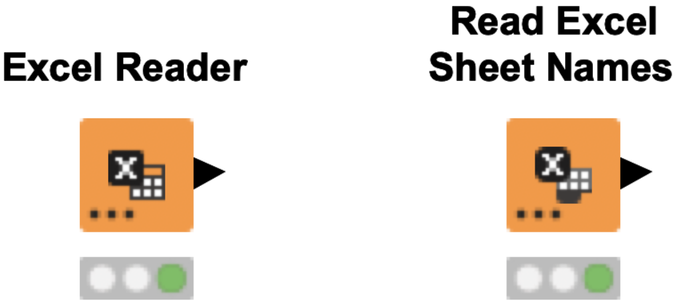
Input location

knime://knime.workflow/../../data/Sentiment Analysis.table

YouTube KNIME TV Channel:
https://youtu.be/U9sP4g4yGwY

Excel Reader (XLS)

- Reads .xls and .xlsx file from Microsoft Excel
- Supports reading from multiple sheets

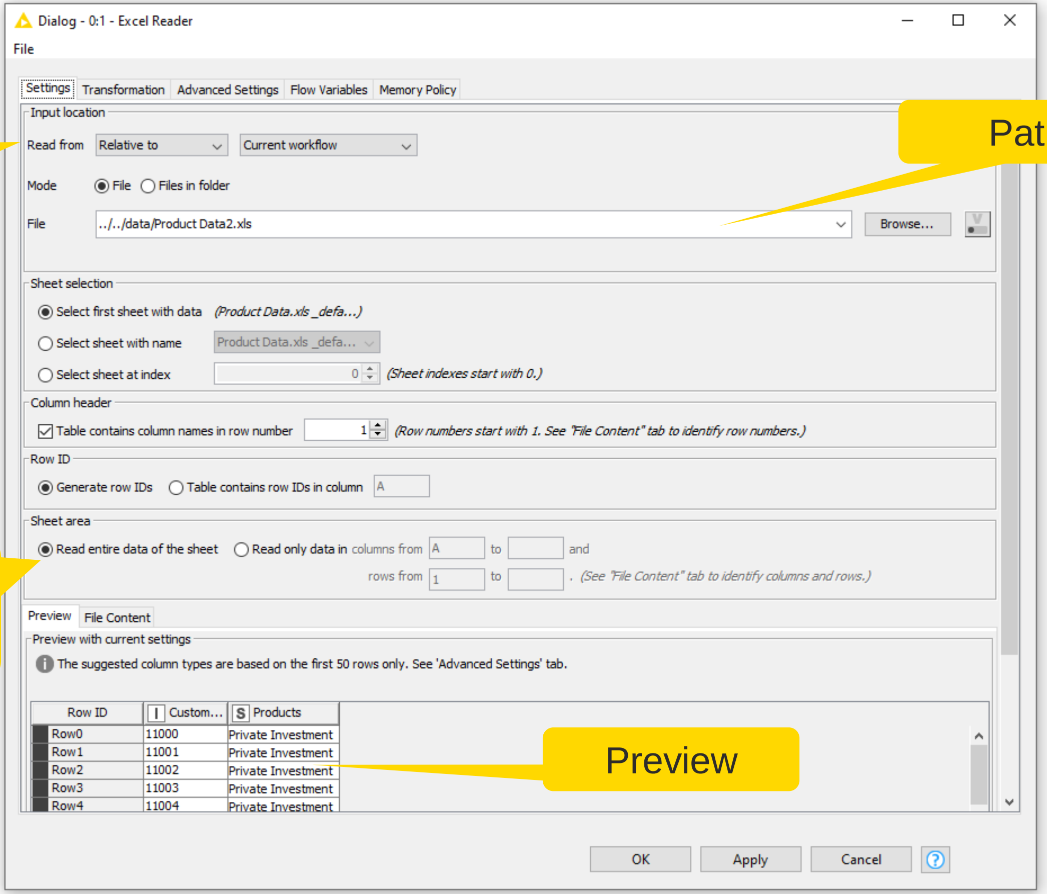


Excel Reader

Excel Reader

File system

Sheet specific settings



Dialog - 0:1 - Excel Reader

File

Settings | Transformation | Advanced Settings | Flow Variables | Memory Policy

Input location

Read from: Relative to — Current workflow

Mode: ● File ○ Files in folder

File: ../../data/Product Data2.xls — Browse...

Sheet selection

● Select first sheet with data *(Product Data.xls _defa...)*

○ Select sheet with name: Product Data.xls _defa...

○ Select sheet at index: 0 *(Sheet indexes start with 0.)*

Column header

☑ Table contains column names in row number 1 *(Row numbers start with 1. See "File Content" tab to identify row numbers.)*

Row ID

● Generate row IDs ○ Table contains row IDs in column A

Sheet area

● Read entire data of the sheet ○ Read only data in columns from A to and rows from 1 to . *(See "File Content" tab to identify columns and rows.)*

Preview | File Content

Preview with current settings

The suggested column types are based on the first 50 rows only. See 'Advanced Settings' tab.

Row ID	I Custom...	S Products
Row0	11000	Private Investment
Row1	11001	Private Investment
Row2	11002	Private Investment
Row3	11003	Private Investment
Row4	11004	Private Investment

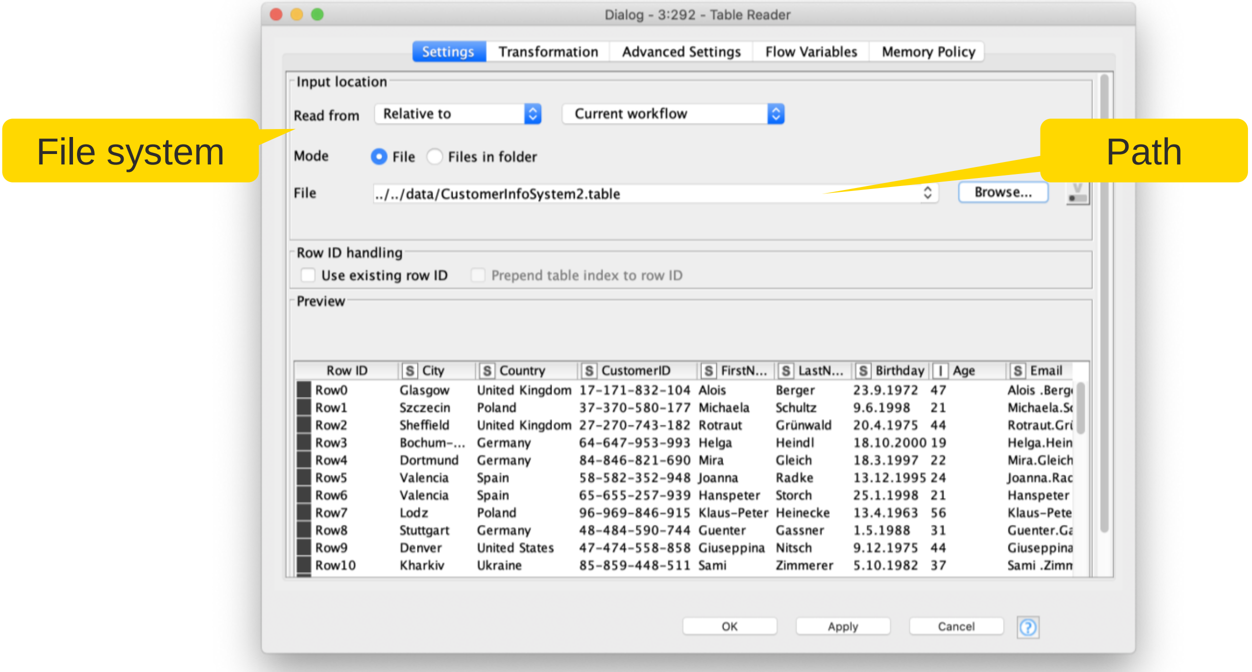
Path

Preview

Table Reader

- Reads tables from the native KNIME Format
- Maximum performance, minimum configuration

Table Reader



Dialog - 3:292 - Table Reader

Settings | Transformation | Advanced Settings | Flow Variables | Memory Policy

Input location

Read from: Relative to — Current workflow

Mode: ● File ○ Files in folder

File: ../../data/CustomerInfoSystem2.table — Browse...

Row ID handling

□ Use existing row ID □ Prepend table index to row ID

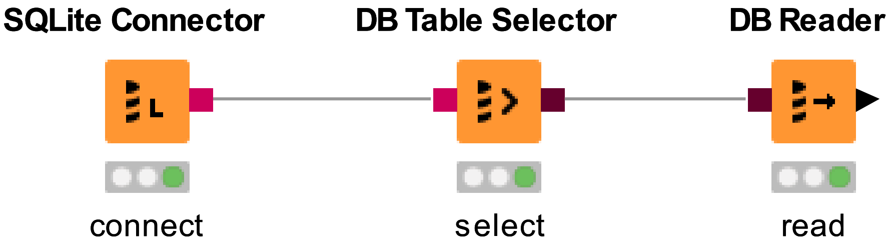
Preview

Row ID	S City	S Country	S CustomerID	S FirstN...	S LastN...	S Birthday	I Age	S Email
Row0	Glasgow	United Kingdom	17-171-832-104	Alois	Berger	23.9.1972	47	Alois .Berg
Row1	Szczecin	Poland	37-370-580-177	Michaela	Schultz	9.6.1998	21	Michaela.S
Row2	Sheffield	United Kingdom	27-270-743-182	Rotraut	Grünwald	20.4.1975	44	Rotraut.Gr
Row3	Bochum-...	Germany	64-647-953-993	Helga	Heindl	18.10.2000	19	Helga.Hein
Row4	Dortmund	Germany	84-846-821-690	Mira	Gleich	18.3.1997	22	Mira.Gleich
Row5	Valencia	Spain	58-582-352-948	Joanna	Radke	13.12.1995	24	Joanna.Rac
Row6	Valencia	Spain	65-655-257-939	Hanspeter	Storch	25.1.1998	21	Hanspeter
Row7	Lodz	Poland	96-969-846-915	Klaus-Peter	Heinecke	13.4.1963	56	Klaus-Pete
Row8	Stuttgart	Germany	48-484-590-744	Guenter	Gassner	1.5.1988	31	Guenter.Ga
Row9	Denver	United States	47-474-558-858	Giuseppina	Nitsch	9.12.1975	44	Giuseppina
Row10	Kharkiv	Ukraine	85-859-448-511	Sami	Zimmerer	5.10.1982	37	Sami .Zimn

OK | Apply | Cancel

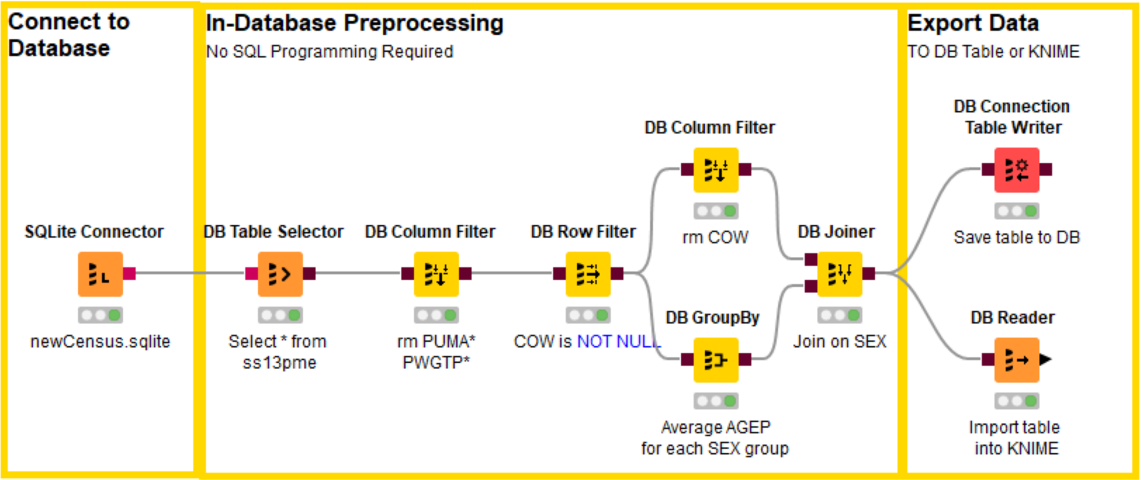
Database Connectivity

- Read data from any JDBC enabled database
- Write your own SQL or model it using dedicated nodes

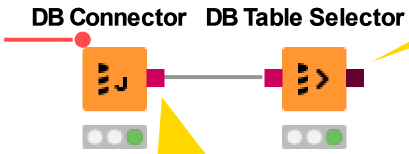


Database Extension

- Visually assemble complex SQL statements (no SQL coding needed)
- Connect to all JDBC-compliant databases
- Harness the power of your database within KNIME
- Complete rewrite in KNIME Analytics Platform 4.0



Database Port Types



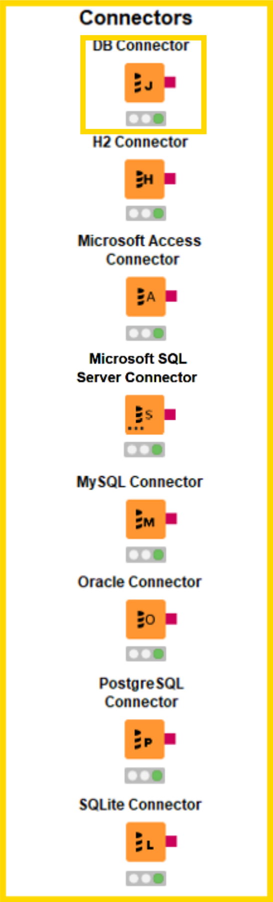
DB Data Port

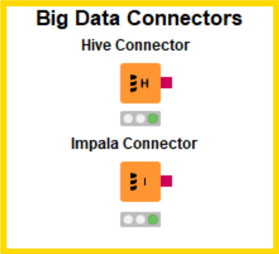
- Connection information
- SQL statement

DB Connection Port
Connection information

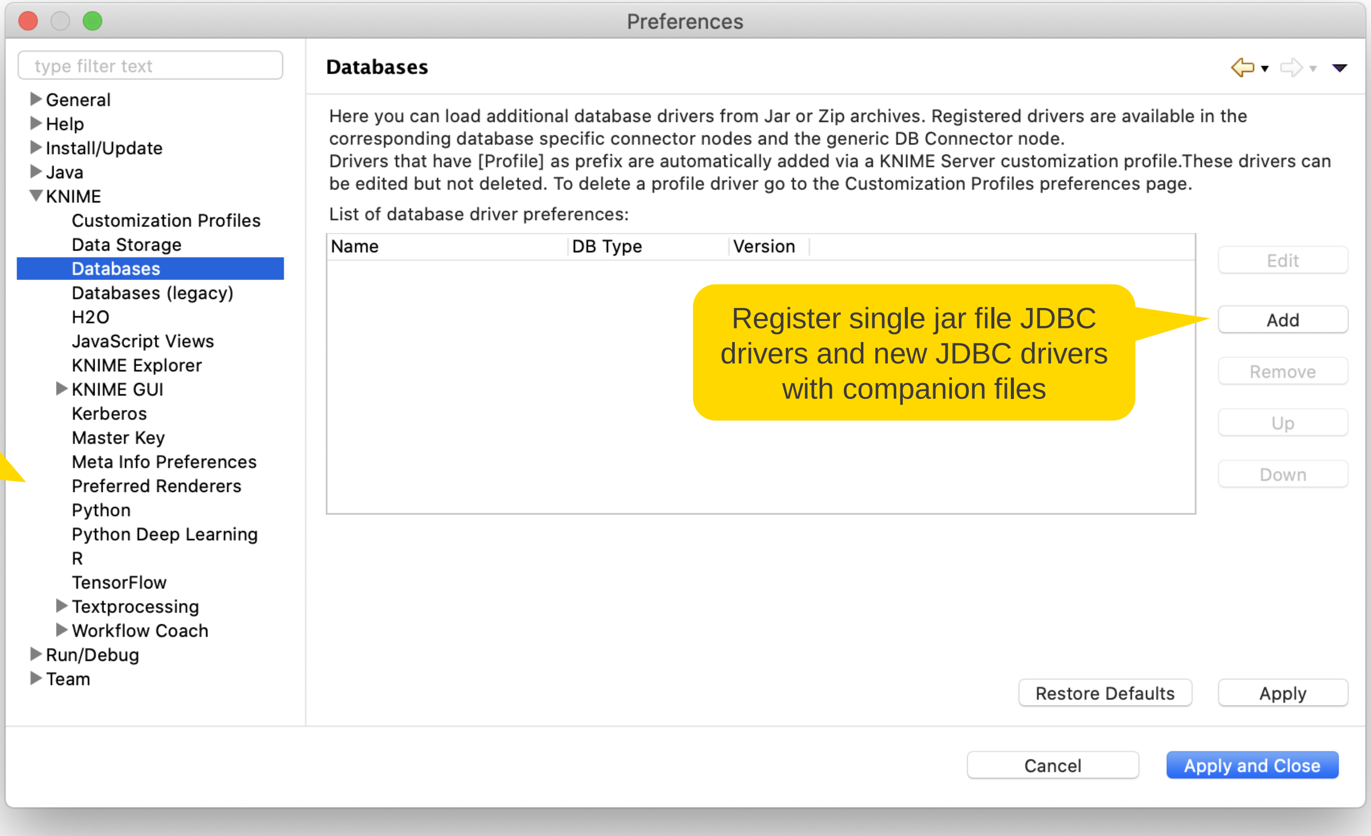
Database Connectors

- Dedicated nodes to connect to specific databases
  - Necessary JDBC driver included
  - Easy to use
  - Import DB specific behavior/capability
- Hive and Impala connectors are part of the KNIME Big Data Connectors extension
- General DB Connector
  - Can connect to any JDBC source
  - Register new JDBC driver via File -> Preferences -> KNIME -> Databases





Register JDBC Driver



Databases

Here you can load additional database drivers from Jar or Zip archives. Registered drivers are available in the corresponding database specific connector nodes and the generic DB Connector node.
Drivers that have [Profile] as prefix are automatically added via a KNIME Server customization profile.These drivers can be edited but not deleted. To delete a profile driver go to the Customization Profiles preferences page.

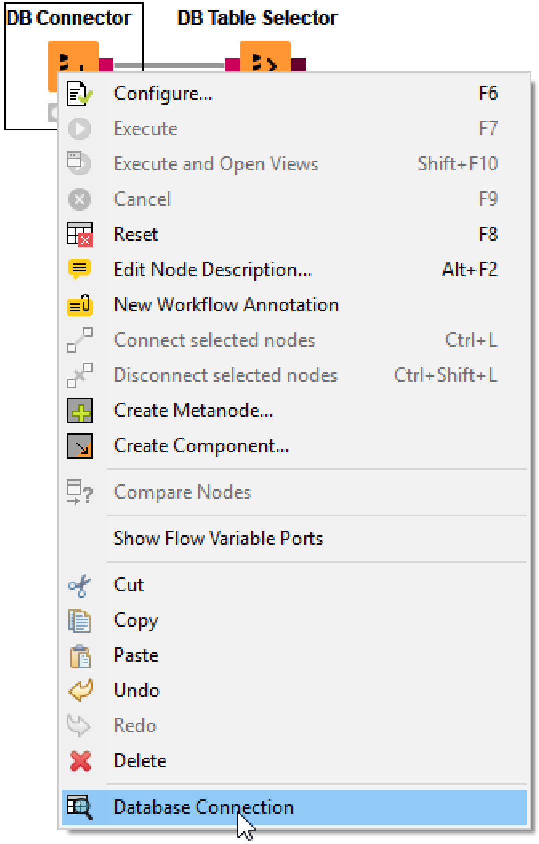
List of database driver preferences:

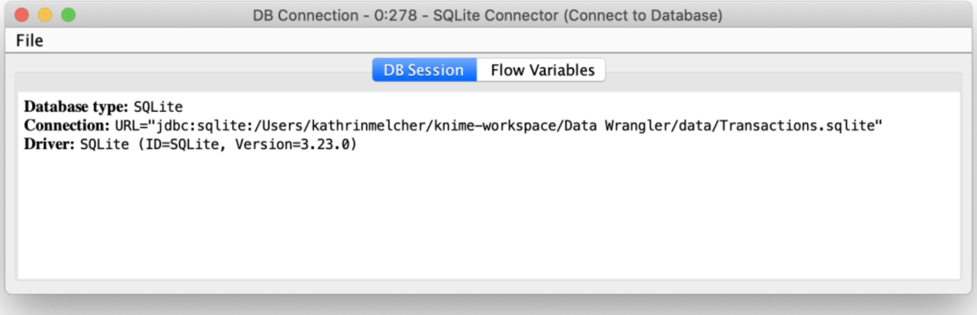
Name	DB Type	Version

Register single jar file JDBC drivers and new JDBC drivers with companion files

Open KNIME and go to File -> Preferences

Database JDBC Connection Port View

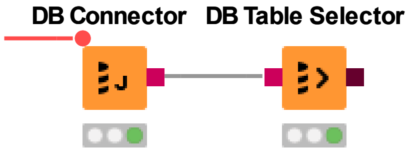


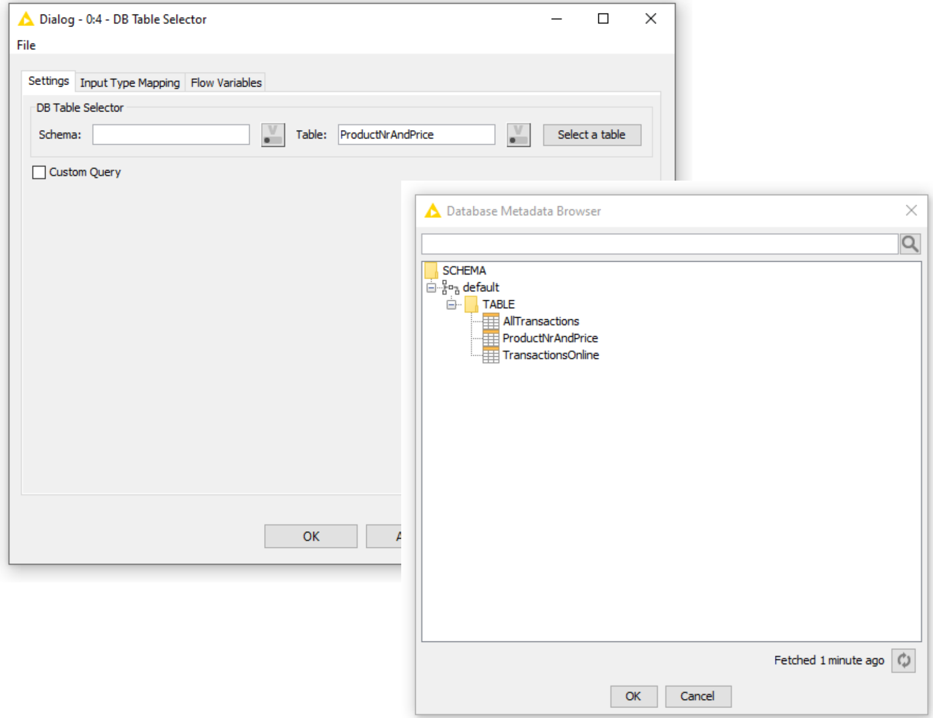

DB Connection - 0:278 - SQLite Connector (Connect to Database)
File
DB Session
Flow Variables

```
Database type: SQLite
Connection: URL="jdbc:sqlite:/Users/kathrinmelcher/knime-workspace/Data Wrangler/data/Transactions.sqlite"
Driver: SQLite (ID=SQLite, Version=3.23.0)
```

DB Table Selector

- Takes connection information and constructs a query
- Explores DB metadata
- Outputs a SQL query




Dialog - 0:4 - DB Table Selector

File

Settings | Input Type Mapping | Flow Variables

DB Table Selector

Schema:

Table: ProductNrAndPrice

Select a table

□ Custom Query

OK

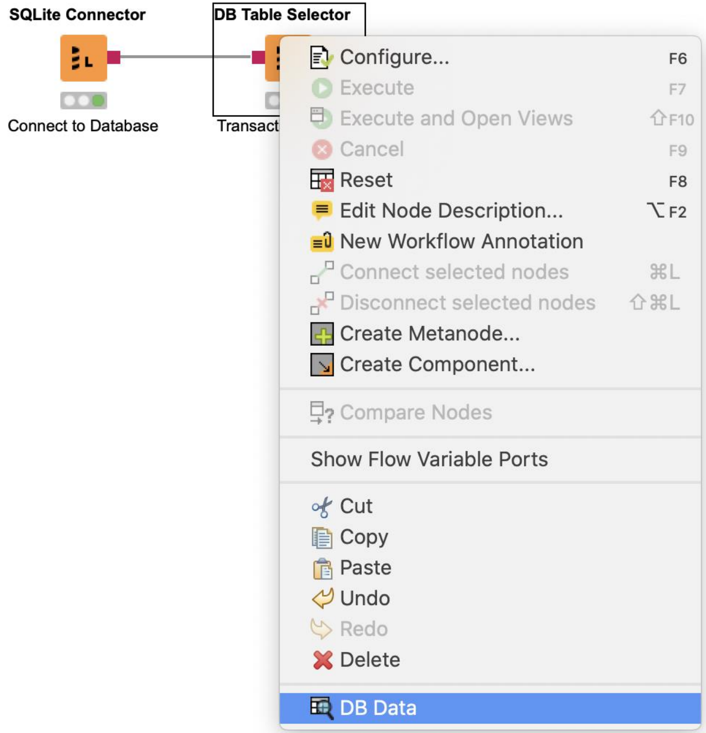
Database Metadata Browser

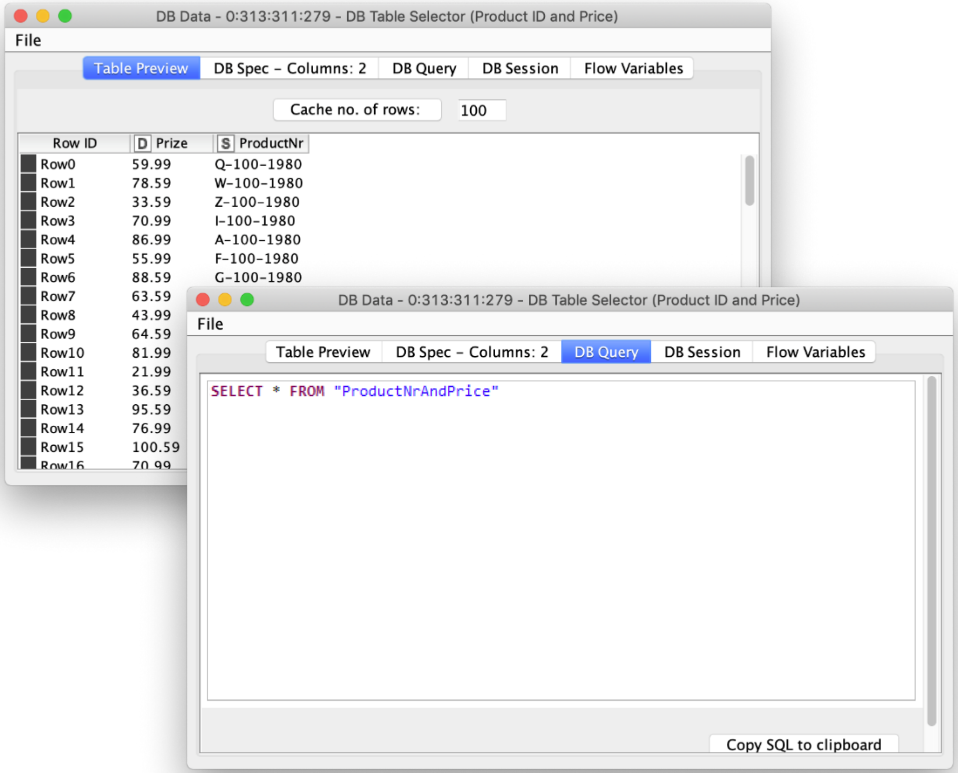
SCHEMA
- default
  - TABLE
    - AllTransactions
    - ProductNrAndPrice
    - TransactionsOnline

Fetched 1 minute ago

OK | Cancel

Database Connection Port View





Cache no. of rows: 100

Row ID	D Prize	S ProductNr
Row0	59.99	Q-100-1980
Row1	78.59	W-100-1980
Row2	33.59	Z-100-1980
Row3	70.99	I-100-1980
Row4	86.99	A-100-1980
Row5	55.99	F-100-1980
Row6	88.59	G-100-1980
Row7	63.59	
Row8	43.99	
Row9	64.59	
Row10	81.99	
Row11	21.99	
Row12	36.59	
Row13	95.59	
Row14	76.99	
Row15	100.59	
Row16	70.99	

```
SELECT * FROM "ProductNrAndPrice"
```

DB Reader

- Executes incoming SQL Query on database
- Reads results into a KNIME data table

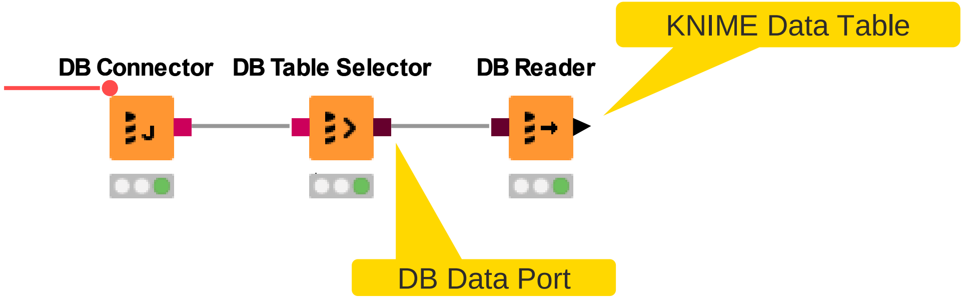
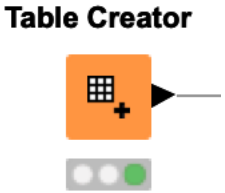
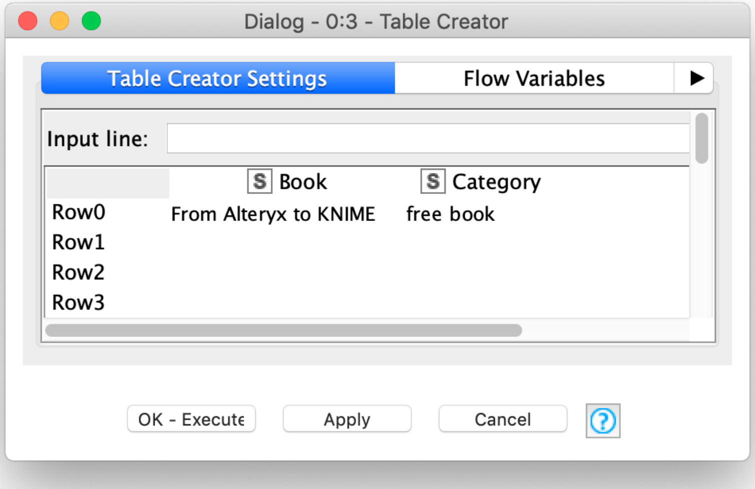


Table Creator

- Allows you to create data tables manually
- Data can be entered in a spreadsheet – like the table in the configuration dialog



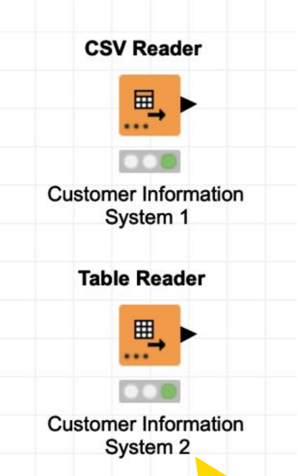

Dialog - 0:3 - Table Creator

Input line:

	S Book	S Category
Row0	From Alteryx to KNIME	free book
Row1		
Row2		
Row3		

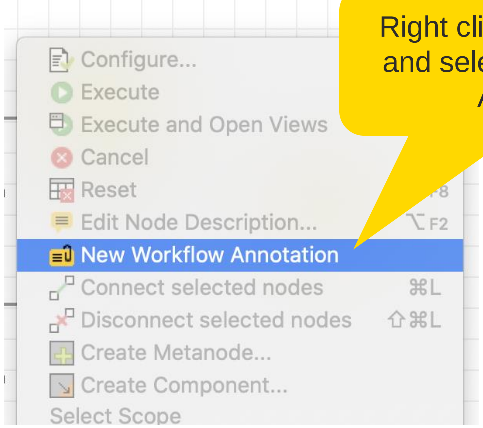
Comments & Annotations

Comments

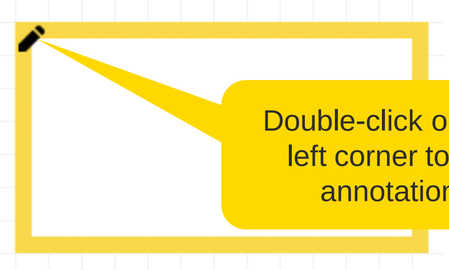


Double-click to change the node label

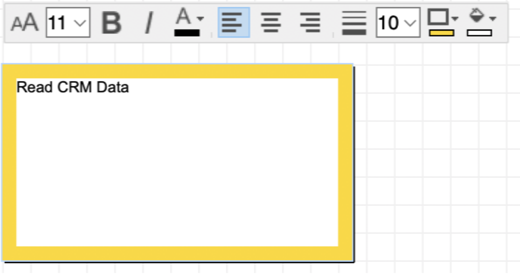
Annotations



Right click in the workflow and select New Workflow Annotation

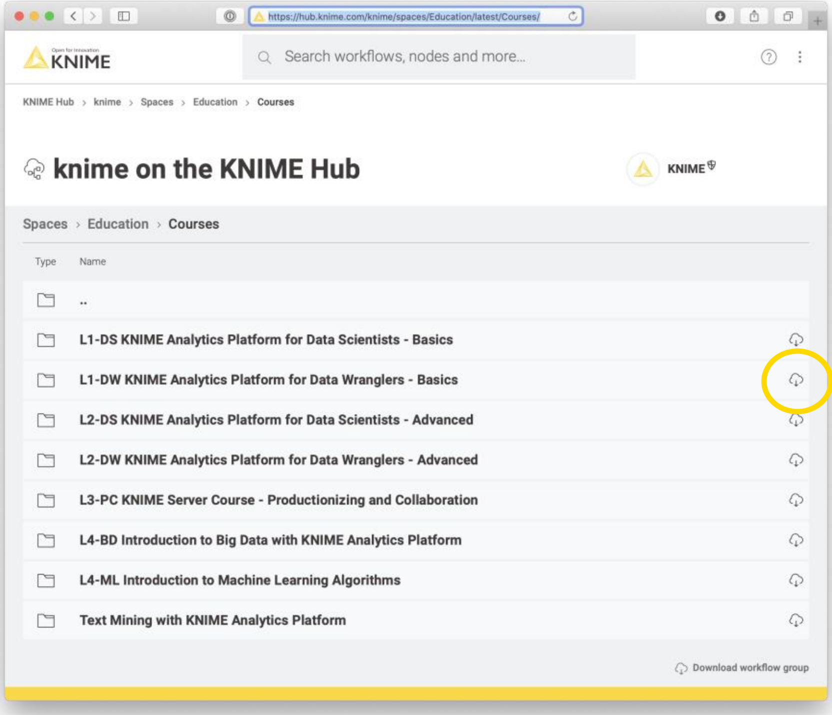


Double-click on the upper left corner to open the annotation editor



Downloading Exercises

- Download the course material from the KNIME Hub
  https://hub.knime.com/knime/spaces/Education/latest/Courses/



https://hub.knime.com/knime/spaces/Education/latest/Courses/

Search workflows, nodes and more...

KNIME Hub › knime › Spaces › Education › Courses

knime on the KNIME Hub

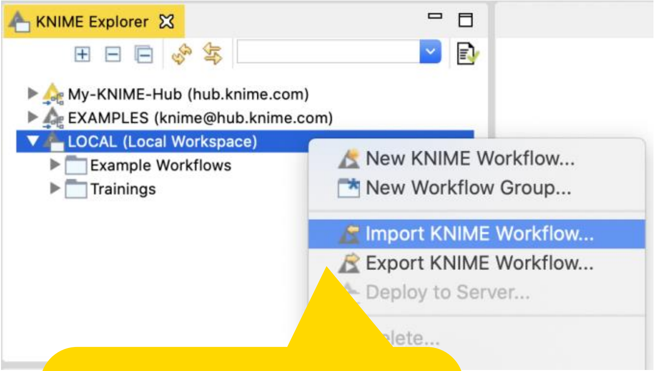
Spaces › Education › Courses

Type	Name
	..
	L1-DS KNIME Analytics Platform for Data Scientists - Basics
	L1-DW KNIME Analytics Platform for Data Wranglers - Basics
	L2-DS KNIME Analytics Platform for Data Scientists - Advanced
	L2-DW KNIME Analytics Platform for Data Wranglers - Advanced
	L3-PC KNIME Server Course - Productionizing and Collaboration
	L4-BD Introduction to Big Data with KNIME Analytics Platform
	L4-ML Introduction to Machine Learning Algorithms
	Text Mining with KNIME Analytics Platform

Download workflow group

Importing Exercises

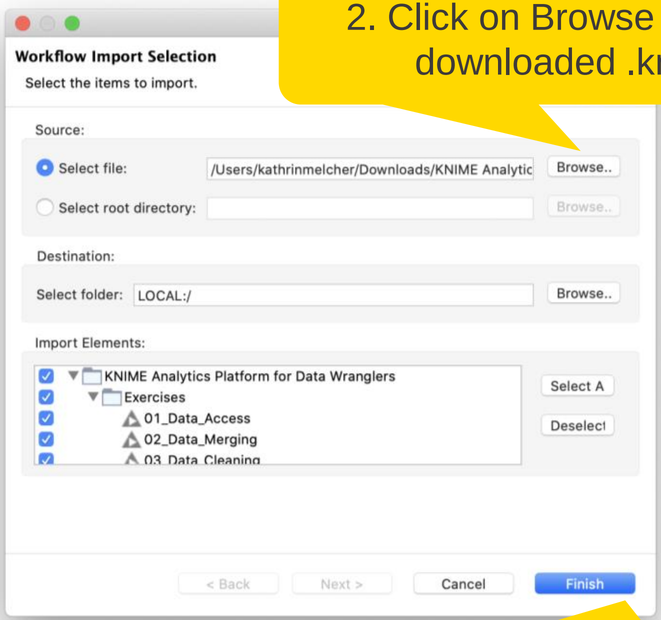
- Import the course material to KNIME Analytics Platform



KNIME Explorer

- My-KNIME-Hub (hub.knime.com)
- EXAMPLES (knime@hub.knime.com)
- LOCAL (Local Workspace)
  - Example Workflows
  - Trainings

1. Right click on LOCAL and select Import KNIME Workflow….



Workflow Import Selection

Select the items to import.

Source:

● Select file: /Users/kathrinmelcher/Downloads/KNIME Analytic

○ Select root directory:

Destination:

Select folder: LOCAL:/

Import Elements:

- ☑ KNIME Analytics Platform for Data Wranglers
  - ☑ Exercises
    - ☑ 01_Data_Access
    - ☑ 02_Data_Merging
    - ☑ 03_Data_Cleaning

2. Click on Browse and select downloaded .knar file

3. Click on Finish

Exercise: 01_Data_Access

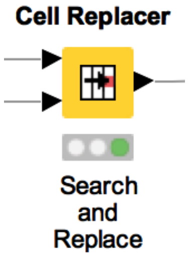
Open the workflow 01_Data_Access and read the following data files:

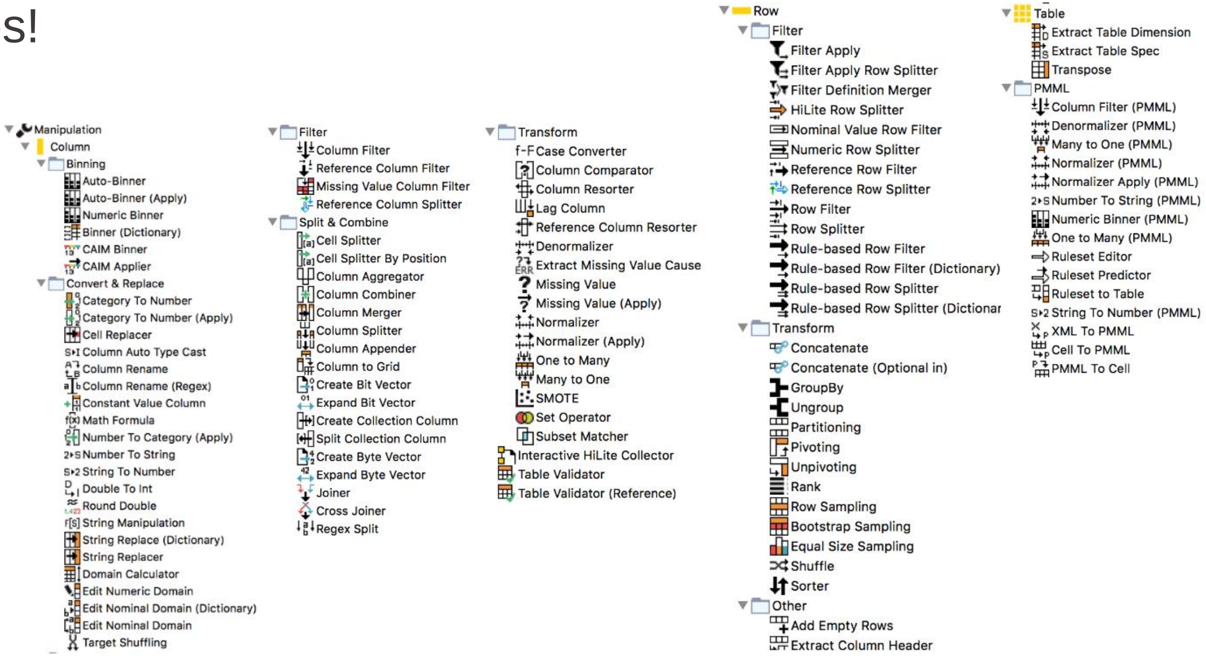
- Customer information
  - CustomerInfoSystem1.csv
  - CustomerInfoSystem2.table
- Online shop transactions, and product number & price information
  - TransactionOnline from Transations.sqlite
  - ProductNrAndPrice from Transations.sqlite
- Store transactions and information
  - Store.xlsx
  - TransactionsStore.table
- Try to use workflow relative-paths

Data Merging

Data Manipulation Nodes

- Yellow color with a variety of input and output ports
- Apply a transformation to input data
- Many, many nodes!



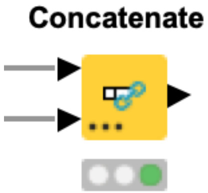


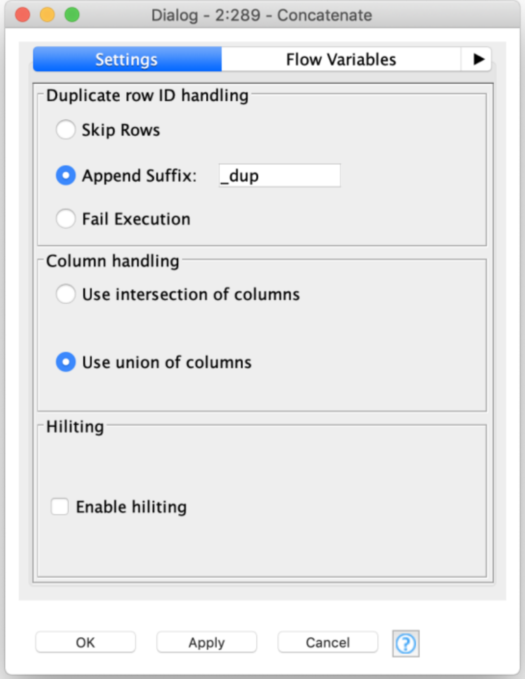
- Manipulation
  - Column
    - Binning
      - Auto-Binner
      - Auto-Binner (Apply)
      - Numeric Binner
      - Binner (Dictionary)
      - CAIM Binner
      - CAIM Applier
    - Convert & Replace
      - Category To Number
      - Category To Number (Apply)
      - Cell Replacer
      - Column Auto Type Cast
      - Column Rename
      - Column Rename (Regex)
      - Constant Value Column
      - Math Formula
      - Number To Category (Apply)
      - Number To String
      - String To Number
      - Double To Int
      - Round Double
      - String Manipulation
      - String Replace (Dictionary)
      - String Replacer
      - Domain Calculator
      - Edit Numeric Domain
      - Edit Nominal Domain (Dictionary)
      - Edit Nominal Domain
      - Target Shuffling
    - Filter
      - Column Filter
      - Reference Column Filter
      - Missing Value Column Filter
      - Reference Column Splitter
    - Split & Combine
      - Cell Splitter
      - Cell Splitter By Position
      - Column Aggregator
      - Column Combiner
      - Column Merger
      - Column Splitter
      - Column Appender
      - Column to Grid
      - Create Bit Vector
      - Expand Bit Vector
      - Create Collection Column
      - Split Collection Column
      - Create Byte Vector
      - Expand Byte Vector
      - Joiner
      - Cross Joiner
      - Regex Split
    - Transform
      - Case Converter
      - Column Comparator
      - Column Resorter
      - Lag Column
      - Reference Column Resorter
      - Denormalizer
      - Extract Missing Value Cause
      - Missing Value
      - Missing Value (Apply)
      - Normalizer
      - Normalizer (Apply)
      - One to Many
      - Many to One
      - SMOTE
      - Set Operator
      - Subset Matcher
    - Interactive HiLite Collector
    - Table Validator
    - Table Validator (Reference)
  - Row
    - Filter
      - Filter Apply
      - Filter Apply Row Splitter
      - Filter Definition Merger
      - HiLite Row Splitter
      - Nominal Value Row Filter
      - Numeric Row Splitter
      - Reference Row Filter
      - Reference Row Splitter
      - Row Filter
      - Row Splitter
      - Rule-based Row Filter
      - Rule-based Row Filter (Dictionary)
      - Rule-based Row Splitter
      - Rule-based Row Splitter (Dictionar
    - Transform
      - Concatenate
      - Concatenate (Optional in)
      - GroupBy
      - Ungroup
      - Partitioning
      - Pivoting
      - Unpivoting
      - Rank
      - Row Sampling
      - Bootstrap Sampling
      - Equal Size Sampling
      - Shuffle
      - Sorter
    - Other
      - Add Empty Rows
      - Extract Column Header
  - Table
    - Extract Table Dimension
    - Extract Table Spec
    - Transpose
  - PMML
    - Column Filter (PMML)
    - Denormalizer (PMML)
    - Many to One (PMML)
    - Normalizer (PMML)
    - Normalizer Apply (PMML)
    - Number To String (PMML)
    - Numeric Binner (PMML)
    - One to Many (PMML)
    - Ruleset Editor
    - Ruleset Predictor
    - Ruleset to Table
    - String To Number (PMML)
    - XML To PMML
    - Cell To PMML
    - PMML To Cell

Concatenate

Combine rows from two tables with shared columns

- Handles duplicate row keys gracefully
- Take the union or intersection of columns





Duplicate row ID handling

○ Skip Rows

● Append Suffix: _dup

○ Fail Execution

Column handling

○ Use intersection of columns

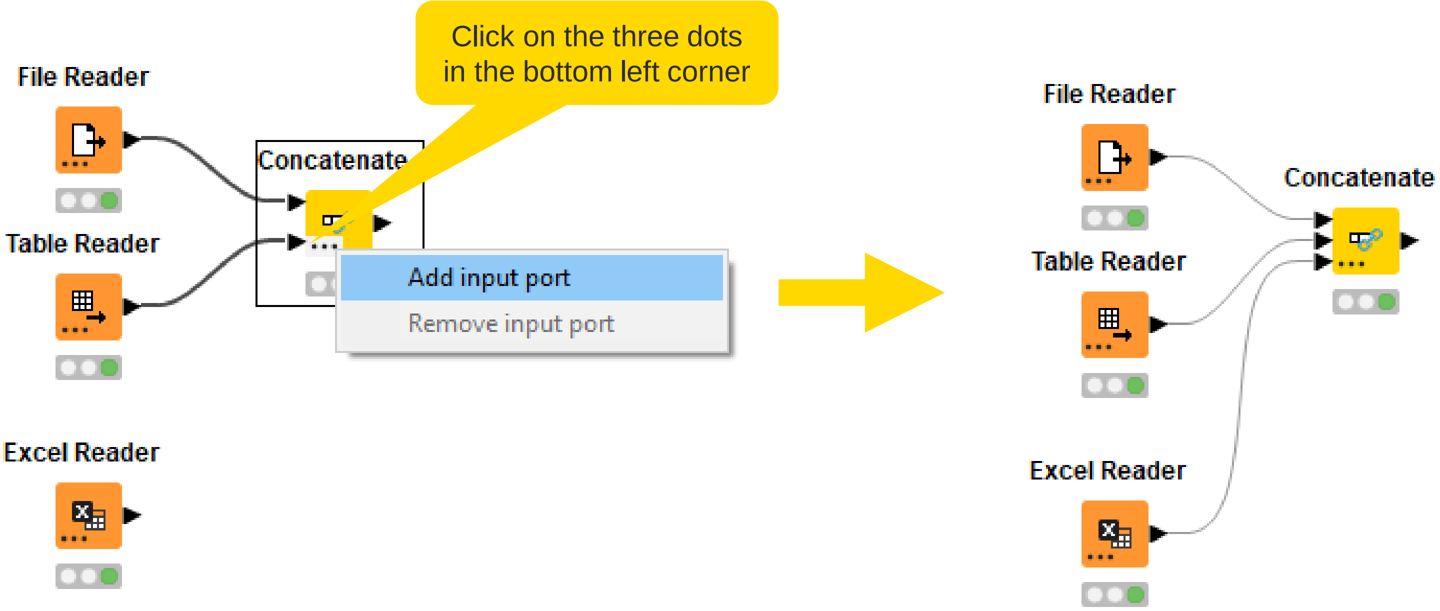
● Use union of columns

Hiliting

□ Enable hiliting

Dynamic Ports

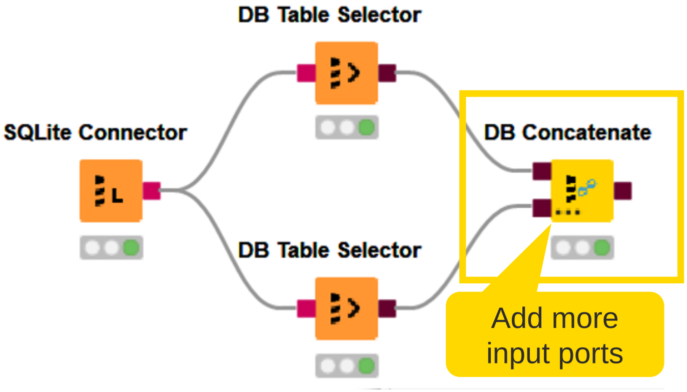
Add and remove node ports based on your needs, e.g. in order to concatenate three or more tables

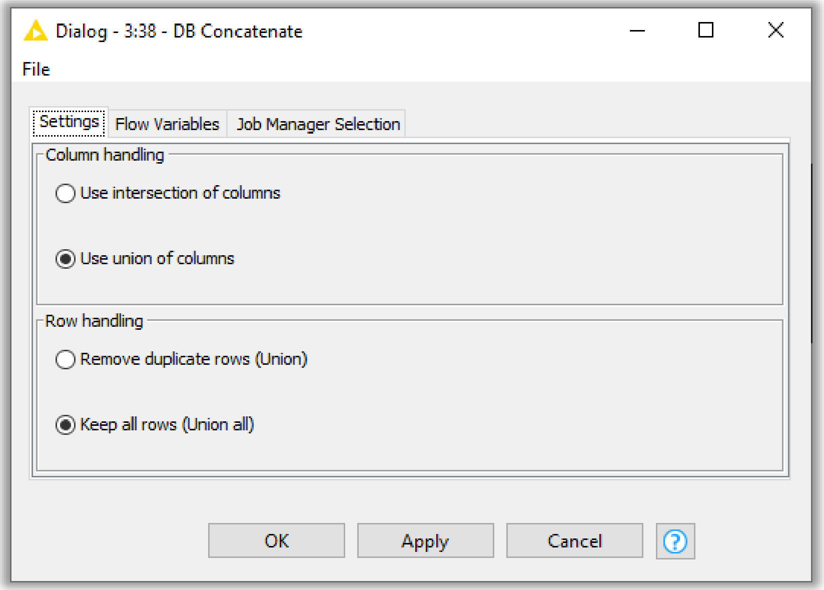


Click on the three dots in the bottom left corner

DB Concatenate

- Combine rows from 2 or more tables with shared columns
- Handles duplicate row keys gracefully
- Take the union or intersection of columns





Column handling

○ Use intersection of columns

● Use union of columns

Row handling

○ Remove duplicate rows (Union)

● Keep all rows (Union all)

Joining Columns of Data

Left Table

CustomerKey	OrderDate	OrderID
22	2019-09-23	#23444
24	2019-09-30	#23457
15	2019-10-07	#28985
10	2091-10-13	#29999

Join by CustomerKey

Right Table

CustomerKey	DoB	City	Gender
17	1974-02-23	Berlin	F
65	2001-05-25	Stuttgart	F
35	1988-08-05	Cologne	M
15	1983-07-20	Hamburg	M
10	1993-01-13	Berlin	M

Inner Join

CustomerKey	OrderDate	OrderID	DoB	City	Gender
15	2019-10-07	#28985	1983-07-20	Hamburg	M
10	2091-10-13	#29999	1993-01-13	Berlin	M

Left Outer Join

CustomerKey	OrderDate	OrderID	DoB	City	Gender
22	2019-09-23	#23444	?	?	?
24	2019-09-30	#23457	?	?	?
15	2019-10-07	#28985	1983-07-20	Hamburg	M
10	2091-10-13	#29999	1993-01-13	Berlin	M

Right Outer Join

CustomerKey	OrderDate	OrderID	DoB	City	Gender
17	?	?	1974-02-23	Berlin	F
65	?	?	2001-05-25	Stuttgart	F
35	?	?	1988-08-05	Cologne	M
15	2019-10-07	#28985	1983-07-20	Hamburg	M
10	2091-10-13	#29999	1993-01-13	Berlin	M

Joining Columns of Data

Left Table

CustomerKey	OrderDate	OrderID
22	2019-09-23	#23444
24	2019-09-30	#23457
15	2019-10-07	#28985
10	2091-10-13	#29999

Join by CustomerKey

Right Table

CustomerKey	DoB	City	Gender
17	1974-02-23	Berlin	F
65	2001-05-25	Stuttgart	F
35	1988-08-05	Cologne	M
15	1983-07-20	Hamburg	M
10	1993-01-13	Berlin	M

Full Outer Join

CustomerKey	OrderDate	OrderID	DoB	City	Gender
17	?	?	1974-02-23	Berlin	F
65	?	?	2001-05-25	Stuttgart	F
35	?	?	1988-08-05	Cologne	M
15	2019-10-07	#28985	1983-07-20	Hamburg	M
10	2091-10-13	#29999	1993-01-13	Berlin	M
22	2019-09-23	#23444	?	?	?
24	2019-09-30	#23457	?	?	?

Missing values in the left table

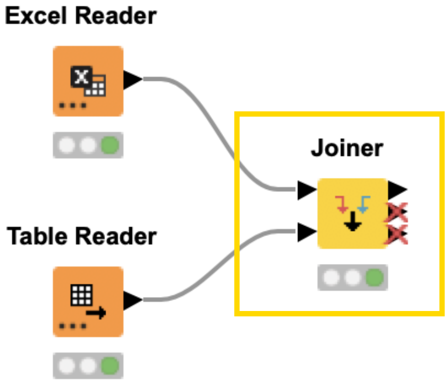
Missing values in the right table

Joiner

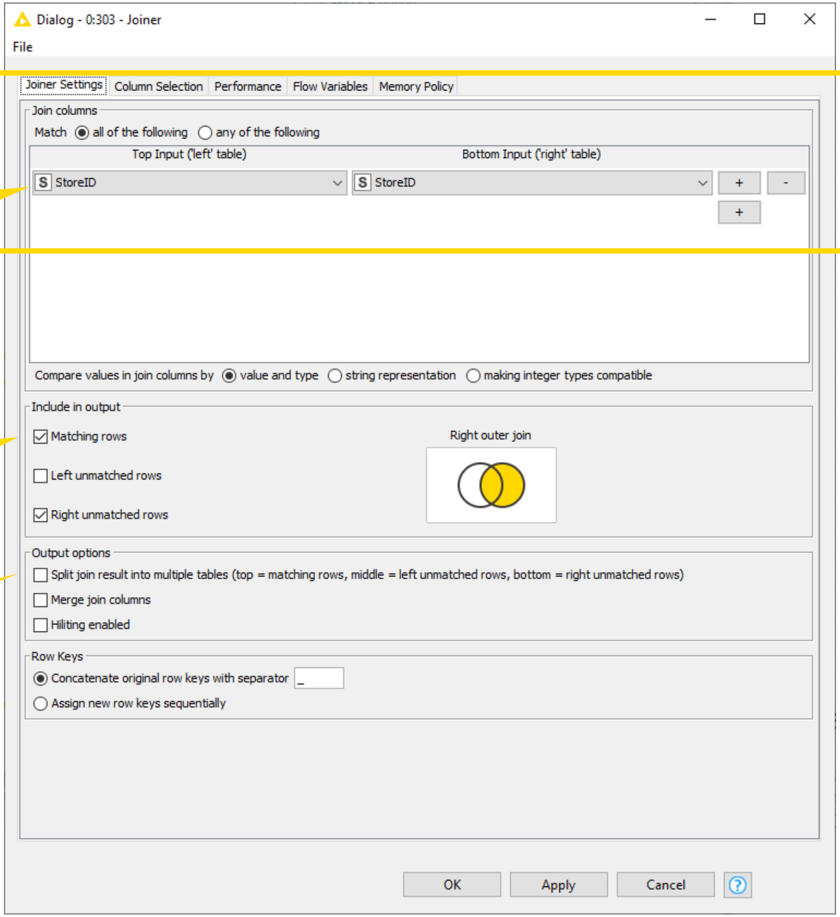
- Combines columns from two different tables
  - Top input port: “Left” data table
  - Bottom input port: “Right” data table

- Outputs:
  - Top port: Resulting joined table
  - Middle port: Unmatched rows from the left input table (top input port)
  - Bottom port: Unmatched rows from the right input table (bottom input port)

- By default the two bottom output ports are deactivated



Joiner Configuration – Linking Rows



Dialog - 0:303 - Joiner

File

Joiner Settings | Column Selection | Performance | Flow Variables | Memory Policy

Join columns

Match ● all of the following ○ any of the following

Top Input ('left' table)	Bottom Input ('right' table)		
S StoreID	S StoreID	+	-
		+	

Compare values in join columns by ● value and type ○ string representation ○ making integer types compatible

Include in output

☑ Matching rows

☐ Left unmatched rows

☑ Right unmatched rows

Right outer join

Output options

☐ Split join result into multiple tables (top = matching rows, middle = left unmatched rows, bottom = right unmatched rows)

☐ Merge join columns

☐ Hiliting enabled

Row Keys

● Concatenate original row keys with separator _

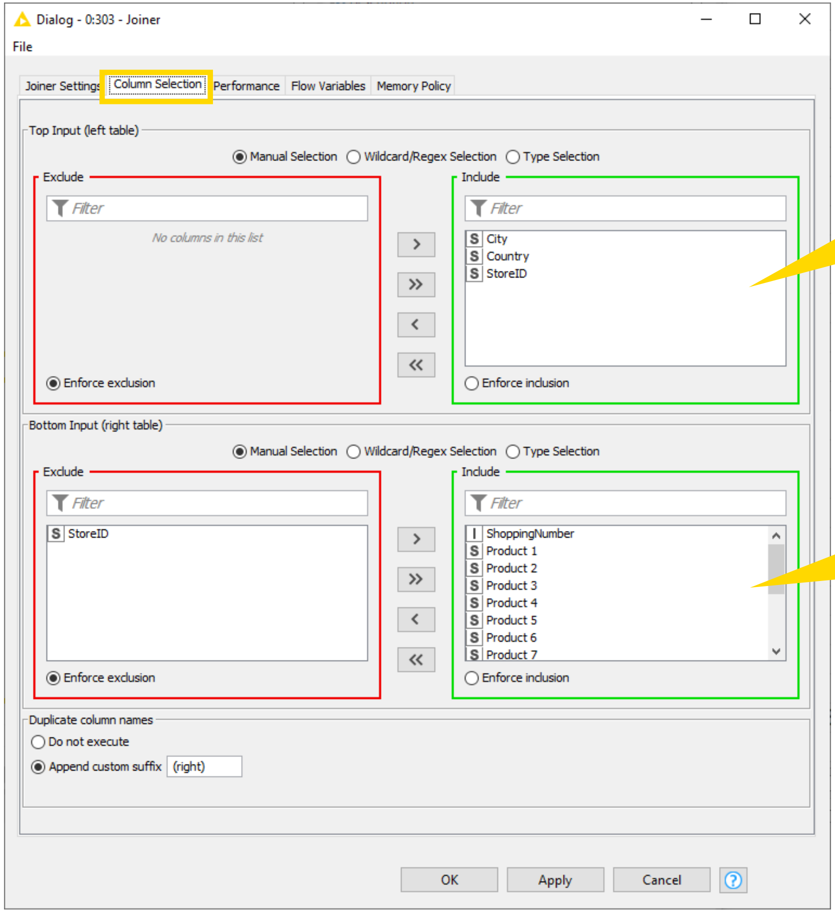
○ Assign new row keys sequentially

Values to join on. Multiple joining columns are allowed

Select the rows which should be included in the joined table

Activate this checkbox to activate the bottom output ports

Joiner Configuration – Column Selection



Dialog - 0:303 - Joiner

Top Input (left table)

● Manual Selection ○ Wildcard/Regex Selection ○ Type Selection

Exclude: Filter — No columns in this list

● Enforce exclusion

Include: Filter
- S City
- S Country
- S StoreID

○ Enforce inclusion

Bottom Input (right table)

● Manual Selection ○ Wildcard/Regex Selection ○ Type Selection

Exclude: Filter
- S StoreID

● Enforce exclusion

Include: Filter
- I ShoppingNumber
- S Product 1
- S Product 2
- S Product 3
- S Product 4
- S Product 5
- S Product 6
- S Product 7

○ Enforce inclusion

Duplicate column names

○ Do not execute

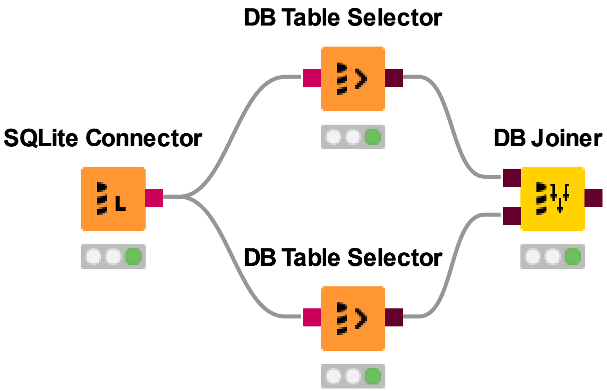
● Append custom suffix (right)

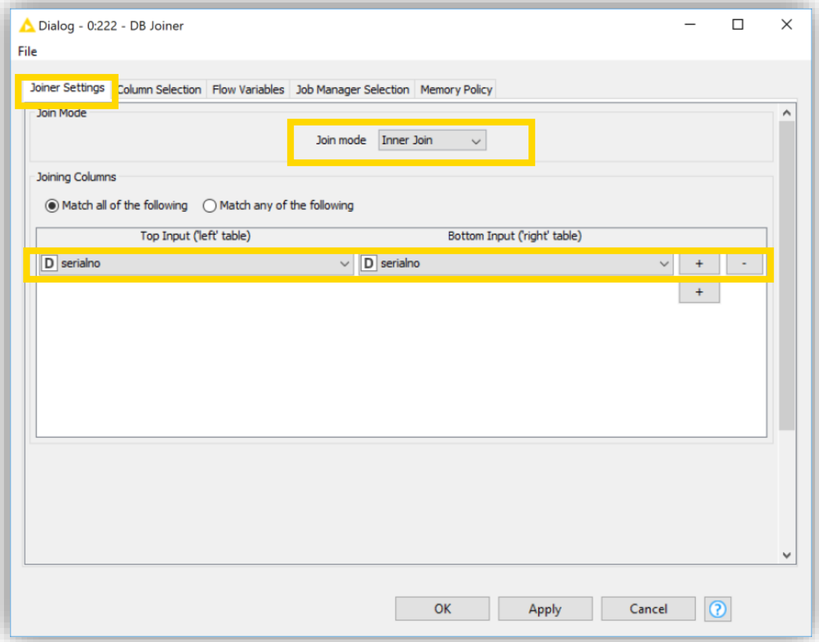
Columns from top table for joined table

Columns from lower table for joined table

DB Joiner

- In-database joiner
- Creates the SQL statement to join two tables stored in the same database
- No coding required




Dialog - 0:222 - DB Joiner

File

Joiner Settings | Column Selection | Flow Variables | Job Manager Selection | Memory Policy

Join Mode

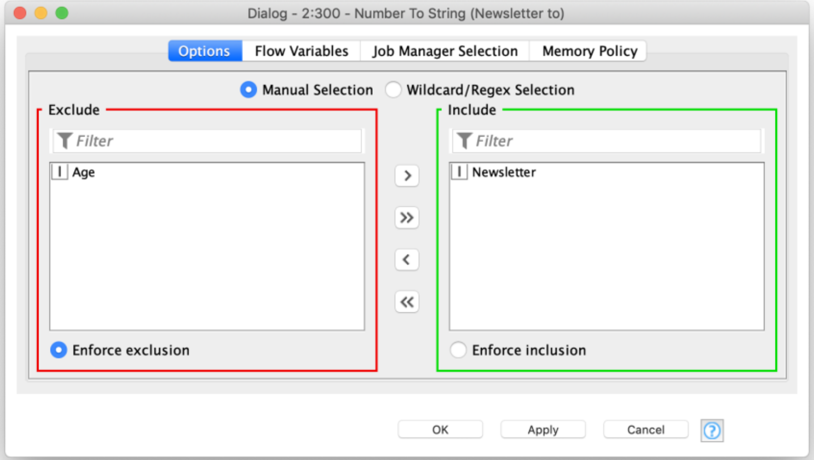
Join mode: Inner Join

Joining Columns

● Match all of the following ○ Match any of the following

Top Input ('left' table)	Bottom Input ('right' table)
D serialno	D serialno

Type Conversion



Dialog - 2:300 - Number To String (Newsletter to)

● Manual Selection ○ Wildcard/Regex Selection

Exclude

Filter

I Age

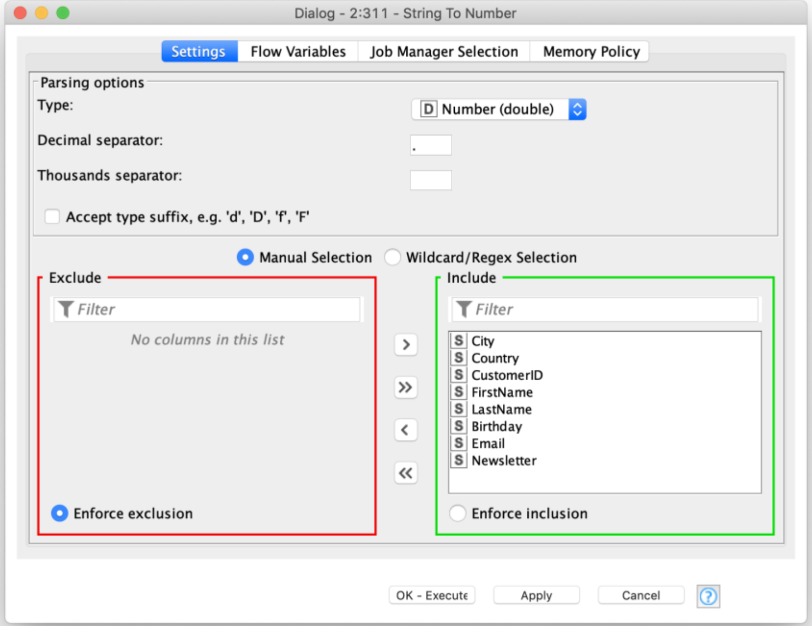
● Enforce exclusion

Include

Filter

I Newsletter

○ Enforce inclusion



Dialog - 2:311 - String To Number

Parsing options

Type: D Number (double)

Decimal separator: .

Thousands separator:

□ Accept type suffix, e.g. 'd', 'D', 'f', 'F'

● Manual Selection ○ Wildcard/Regex Selection

Exclude

Filter

No columns in this list

● Enforce exclusion

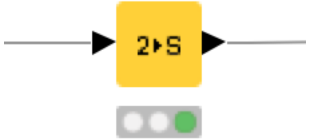
Include

Filter

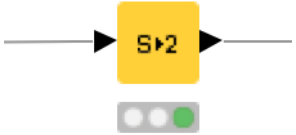
S City
S Country
S CustomerID
S FirstName
S LastName
S Birthday
S Email
S Newsletter

○ Enforce inclusion

Number To String

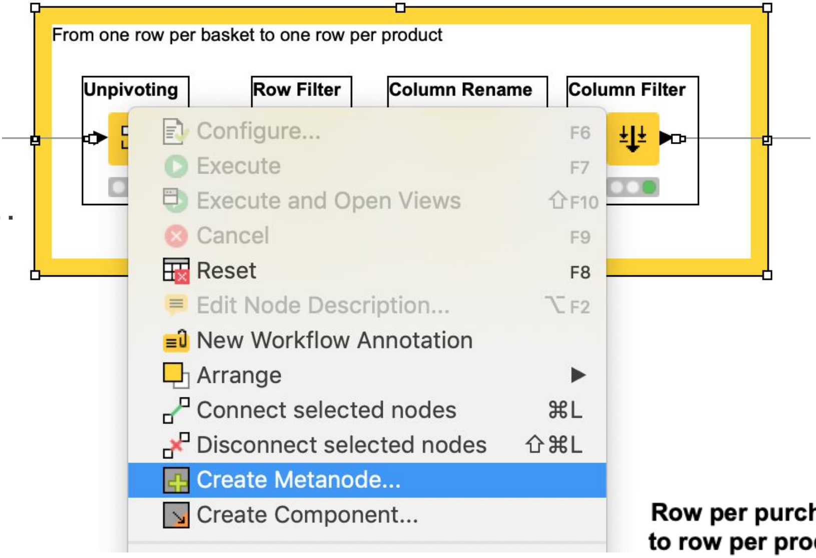


String To Number

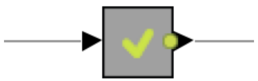


Workflow Organization – Good Practices

- Workflow annotations
- Node labels
- Metanodes
  - Organize workflow by task
  - Hide complexity & improve readability
  - Select nodes -> Right click -> Create Metanode…



Row per purchase to row per product



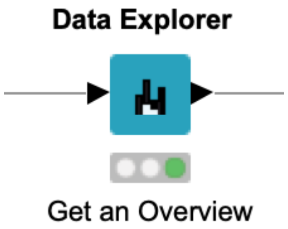
Exercise: 02_Data Merging

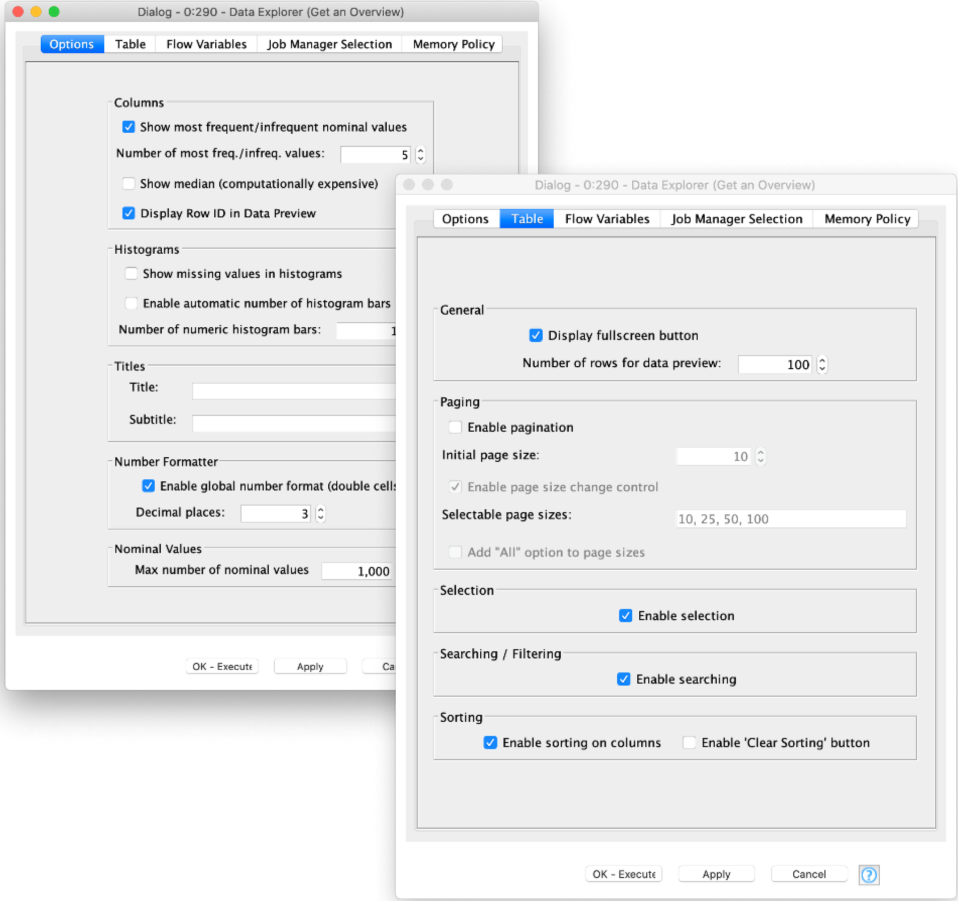
- **Concatenate** the customer information from the two systems
- Add the price information to each online product purchase (DB Joiner) and read the table into KNIME (DB Reader)
- Add the location information to each purchase in a store based on the StoreID (Joiner node)
- Create three metanodes to clean up your workflow
  - Customer data
  - Online transactions & product+price (two output ports)
  - Onsite purchases in stores

Numerical & Nominal Outlier Handling

Data Explorer

The Data Explorer node offers a range of options for displaying properties of the input data in an interactive view





Dialog - 0:290 - Data Explorer (Get an Overview)

Options | Table | Flow Variables | Job Manager Selection | Memory Policy

Columns

☑ Show most frequent/infrequent nominal values

Number of most freq./infreq. values: 5

☐ Show median (computationally expensive)

☑ Display Row ID in Data Preview

Histograms

☐ Show missing values in histograms

☐ Enable automatic number of histogram bars

Number of numeric histogram bars: 1

Titles

Title:

Subtitle:

Number Formatter

☑ Enable global number format (double cells

Decimal places: 3

Nominal Values

Max number of nominal values 1,000

Dialog - 0:290 - Data Explorer (Get an Overview)

Options | Table | Flow Variables | Job Manager Selection | Memory Policy

General

☑ Display fullscreen button

Number of rows for data preview: 100

Paging

☐ Enable pagination

Initial page size: 10

☑ Enable page size change control

Selectable page sizes: 10, 25, 50, 100

☐ Add "All" option to page sizes

Selection

☑ Enable selection

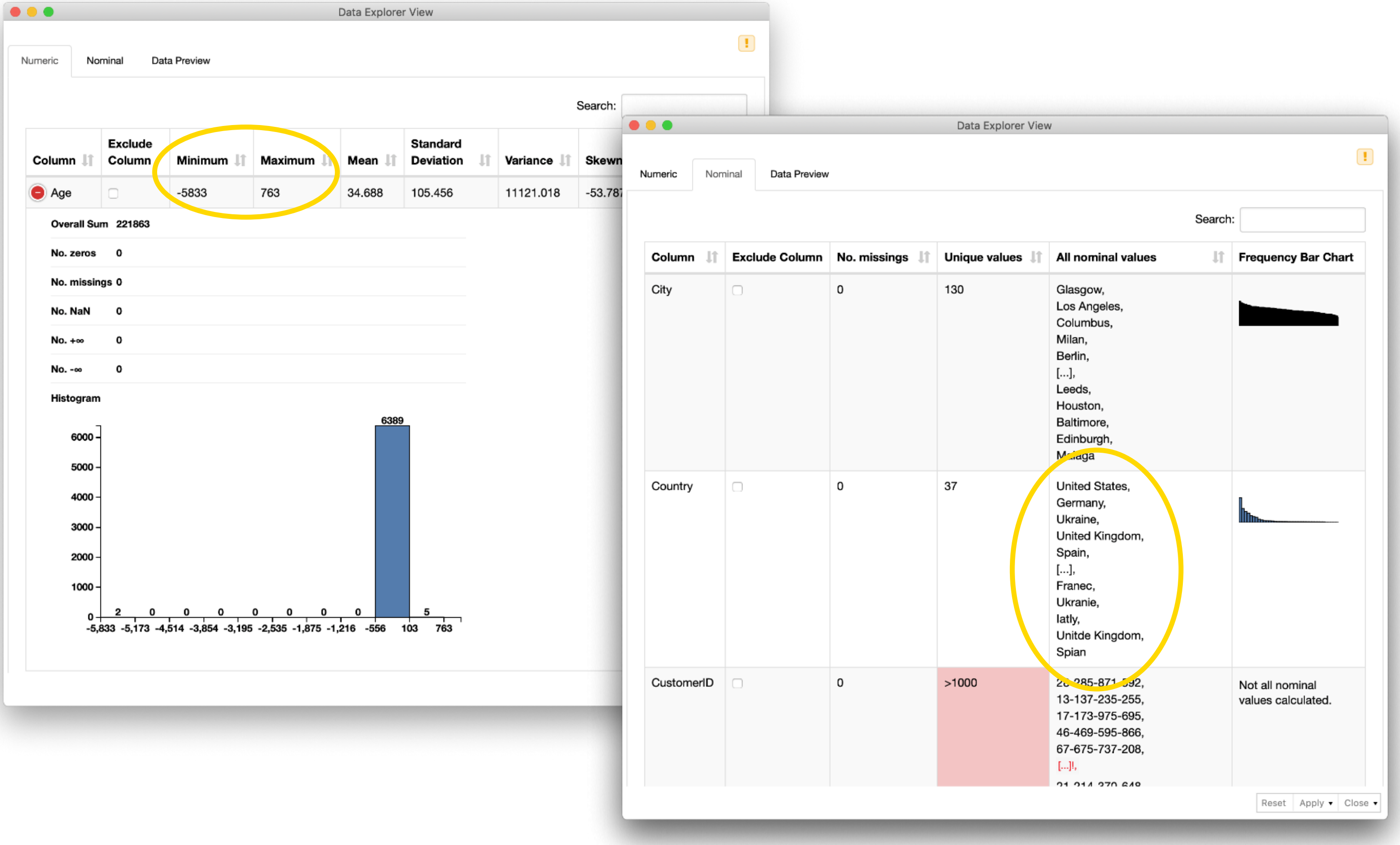
Searching / Filtering

☑ Enable searching

Sorting

☑ Enable sorting on columns ☐ Enable 'Clear Sorting' button

Customer Data Output



Column	Exclude Column	Minimum	Maximum	Mean	Standard Deviation	Variance	Skewn
Age	□	-5833	763	34.688	105.456	11121.018	-53.78

Overall Sum	221863
No. zeros	0
No. missings	0
No. NaN	0
No. +∞	0
No. -∞	0

Column	Exclude Column	No. missings	Unique values	All nominal values	Frequency Bar Chart
City	□	0	130	Glasgow, Los Angeles, Columbus, Milan, Berlin, [...], Leeds, Houston, Baltimore, Edinburgh, Malaga	
Country	□	0	37	United States, Germany, Ukraine, United Kingdom, Spain, [...], Franec, Ukranie, Iatly, Unitde Kingdom, Spian	
CustomerID	□	0	>1000	2[illegible]-285-871-[illegible]92, 13-137-235-255, 17-173-975-695, 46-469-595-866, 67-675-737-208, [...]!, 21-214-370-648	Not all nominal values calculated.

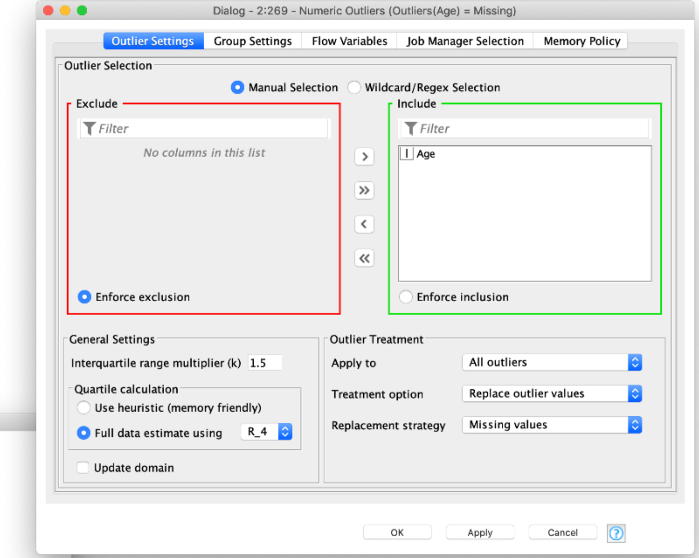
Numeric Outlier

- Detects and treats outliers
- x is a numeric outlier if

  $x < Q_1 - k * IQR$

  $x > Q_3 + k * IQR$

  with $IQR = Q_3 - Q_1$
- For $k = 1,5$ the boarders correspond to the whiskers of a box plot



Dialog - 2:269 - Numeric Outliers (Outliers(Age) = Missing)

Outlier Selection

● Manual Selection ○ Wildcard/Regex Selection

Exclude

Filter

No columns in this list

● Enforce exclusion

Include

Filter

Age

○ Enforce inclusion

General Settings

Interquartile range multiplier (k) 1.5

Quartile calculation

○ Use heuristic (memory friendly)

● Full data estimate using R_4

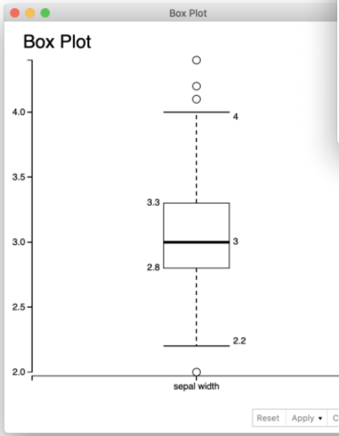
□ Update domain

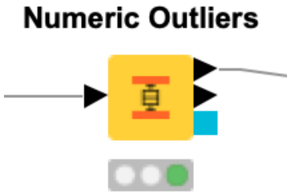
Outlier Treatment

Apply to: All outliers

Treatment option: Replace outlier values

Replacement strategy: Missing values



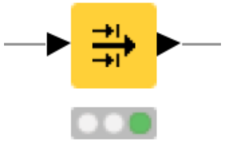


Numeric Outliers

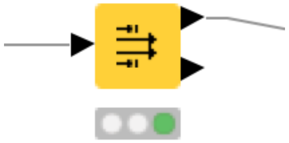
Row Filter and Row Splitter

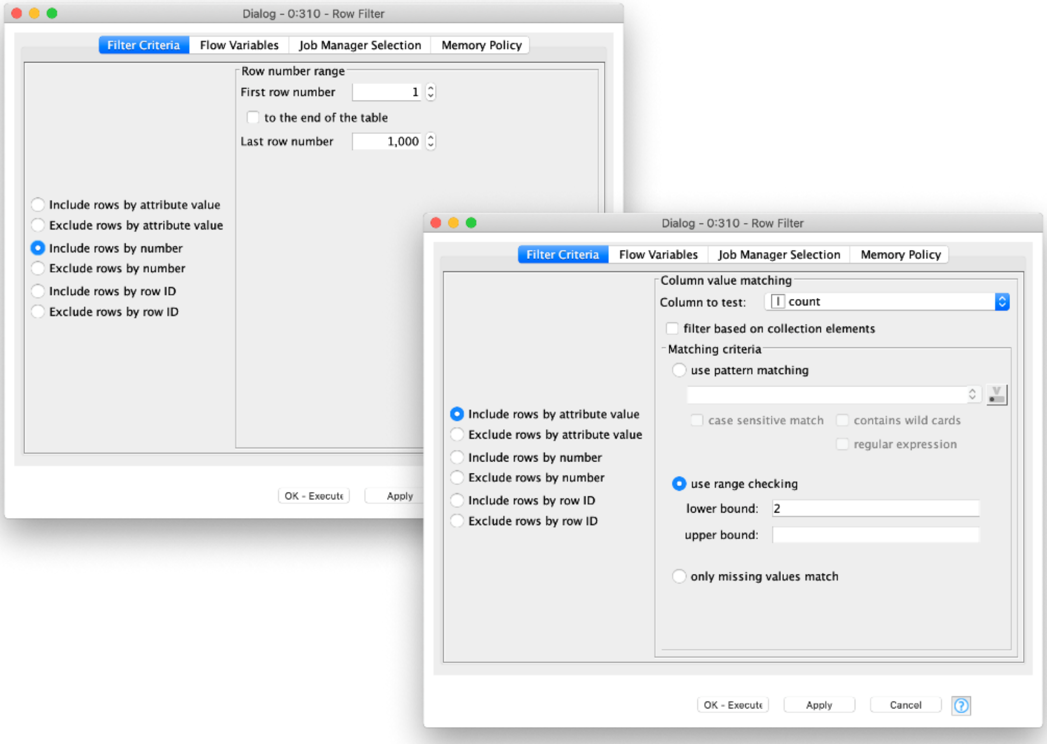
- Row filtering with include and exclude options according to certain criteria
  - Certain value or pattern in a selectable column
  - Row number
  - Row ID

Row Filter



Row Splitter





Dialog - 0:310 - Row Filter

Filter Criteria | Flow Variables | Job Manager Selection | Memory Policy

Row number range

First row number: 1

□ to the end of the table

Last row number: 1,000

○ Include rows by attribute value
○ Exclude rows by attribute value
● Include rows by number
○ Exclude rows by number
○ Include rows by row ID
○ Exclude rows by row ID

Dialog - 0:310 - Row Filter

Filter Criteria | Flow Variables | Job Manager Selection | Memory Policy

● Include rows by attribute value
○ Exclude rows by attribute value
○ Include rows by number
○ Exclude rows by number
○ Include rows by row ID
○ Exclude rows by row ID

Column value matching

Column to test: I count

□ filter based on collection elements

Matching criteria

○ use pattern matching

□ case sensitive match □ contains wild cards
□ regular expression

● use range checking

lower bound: 2

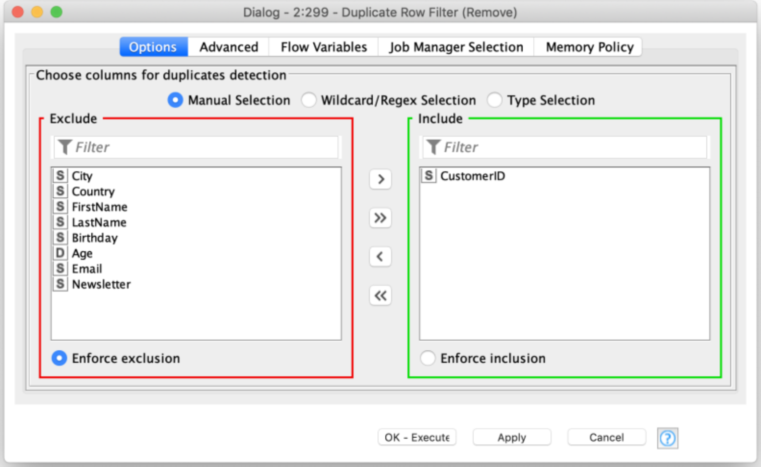
upper bound:

○ only missing values match

Duplicate Row Filter

Detects duplicate rows and apply a selected treatment

- First tab provides the option to select columns for duplicate detection
- Second tab provides options for treating duplicated values



Dialog - 2:299 - Duplicate Row Filter (Remove)

Options | Advanced | Flow Variables | Job Manager Selection | Memory Policy

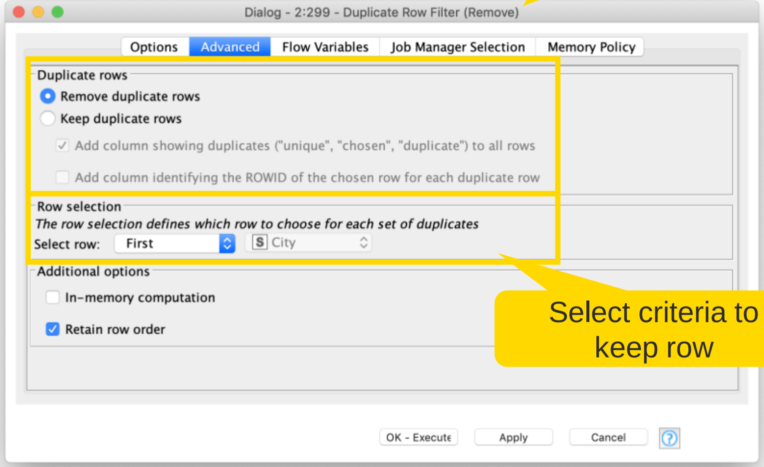
Choose columns for duplicates detection

● Manual Selection ○ Wildcard/Regex Selection ○ Type Selection

Exclude	Include
S City S Country S FirstName S LastName S Birthday D Age S Email S Newsletter	S CustomerID
● Enforce exclusion	○ Enforce inclusion

Filter
Filter
>
>>
<
<<
OK - Execute
Apply
Cancel

Duplicate Row Filter



Dialog - 2:299 - Duplicate Row Filter (Remove)

Options | Advanced | Flow Variables | Job Manager Selection | Memory Policy

Duplicate rows

● Remove duplicate rows

○ Keep duplicate rows

☑ Add column showing duplicates ("unique", "chosen", "duplicate") to all rows

□ Add column identifying the ROWID of the chosen row for each duplicate row

Row selection

The row selection defines which row to choose for each set of duplicates

Select row: First | S City

Additional options

□ In-memory computation

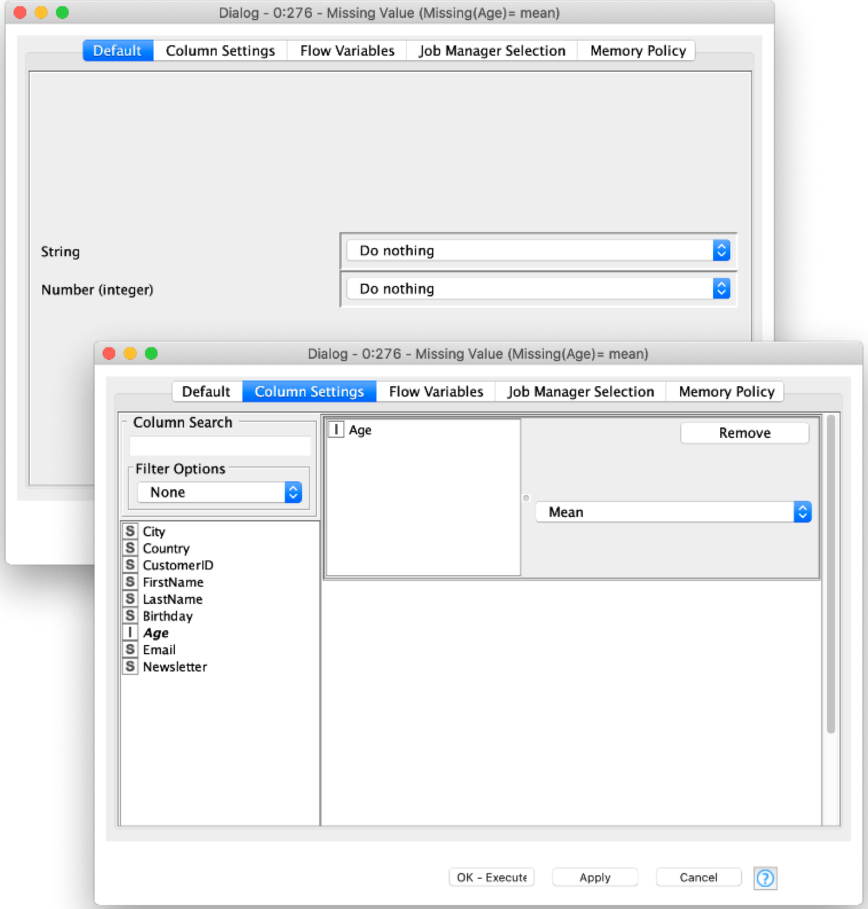
☑ Retain row order

Flag or Remove Duplicates

Select criteria to keep row

Missing Value

- Defines how to handle missing values for all columns of a given type
  - Affects all columns that are not explicitly mentioned in the second tab
- Defines how to handle missing values for each available column



Dialog - 0:276 - Missing Value (Missing(Age)= mean)

String: Do nothing

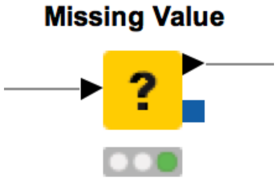
Number (integer): Do nothing

Dialog - 0:276 - Missing Value (Missing(Age)= mean)

Filter Options: None

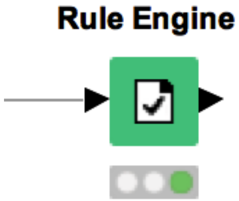
S City
S Country
S CustomerID
S FirstName
S LastName
S Birthday
I *Age*
S Email
S Newsletter

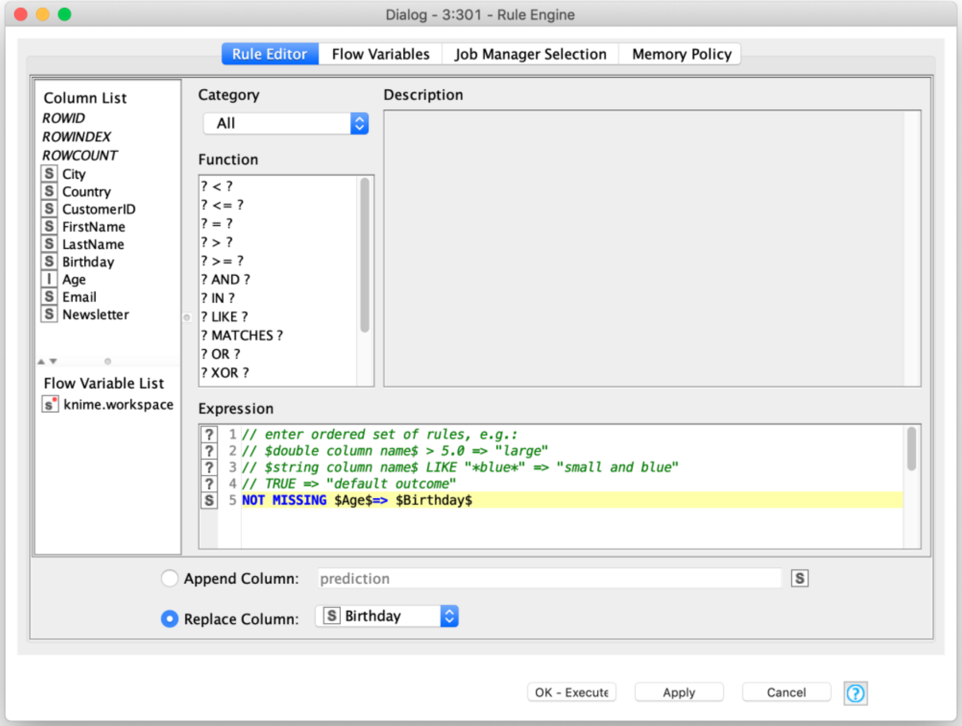
I Age: Mean



Rule Engine

- Defines custom logic to use simple rules
- Rules like: <Antecedent/Condition> => <Consequence>
  - (1=1 => “true”)
- Tries to match rules to each row of the input table




Dialog - 3:301 - Rule Engine

Column List
ROWID
ROWINDEX
ROWCOUNT
S City
S Country
S CustomerID
S FirstName
S LastName
S Birthday
I Age
S Email
S Newsletter

Flow Variable List
s knime.workspace

Category: All

Function
? < ?
? <= ?
? = ?
? > ?
? >= ?
? AND ?
? IN ?
? LIKE ?
? MATCHES ?
? OR ?
? XOR ?

Description

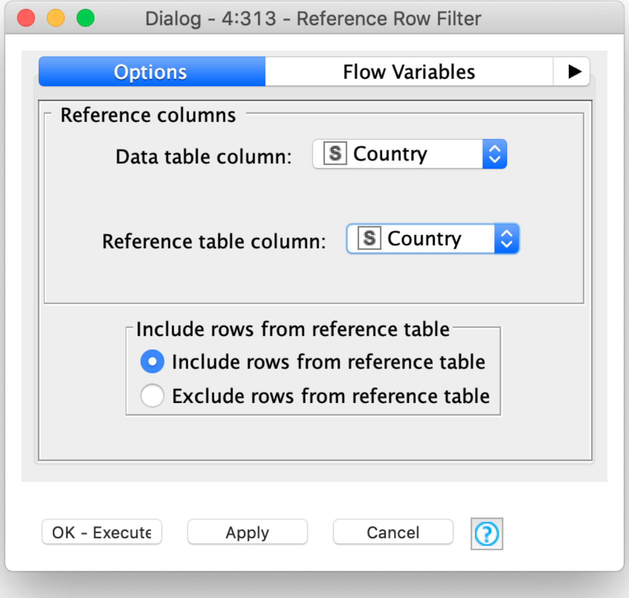
Expression

```
// enter ordered set of rules, e.g.:
// $double column name$ > 5.0 => "large"
// $string column name$ LIKE "*blue*" => "small and blue"
// TRUE => "default outcome"
NOT MISSING $Age$=> $Birthday$
```

○ Append Column: prediction S

● Replace Column: S Birthday

Other Options to Filter Rows



Reference columns

Data table column: S Country

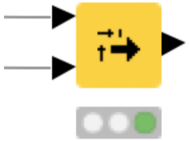
Reference table column: S Country

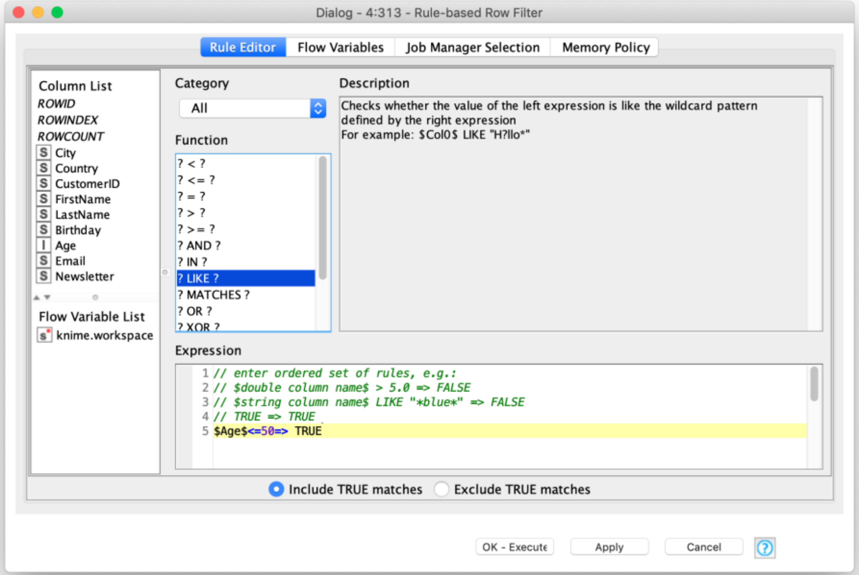
Include rows from reference table

● Include rows from reference table

○ Exclude rows from reference table

Reference Row Filter





Column List
- *ROWID*
- *ROWINDEX*
- *ROWCOUNT*
- S City
- S Country
- S CustomerID
- S FirstName
- S LastName
- S Birthday
- I Age
- S Email
- S Newsletter

Flow Variable List
- s knime.workspace

Category: All

Function
- ? < ?
- ? <= ?
- ? = ?
- ? > ?
- ? >= ?
- ? AND ?
- ? IN ?
- ? LIKE ?
- ? MATCHES ?
- ? OR ?
- ? XOR ?

Description

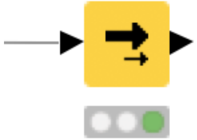
Checks whether the value of the left expression is like the wildcard pattern defined by the right expression
For example: $Col0$ LIKE "H?llo*"

Expression

```
// enter ordered set of rules, e.g.:
// $double column name$ > 5.0 => FALSE
// $string column name$ LIKE "*blue*" => FALSE
// TRUE => TRUE
$Age$<=50=> TRUE
```

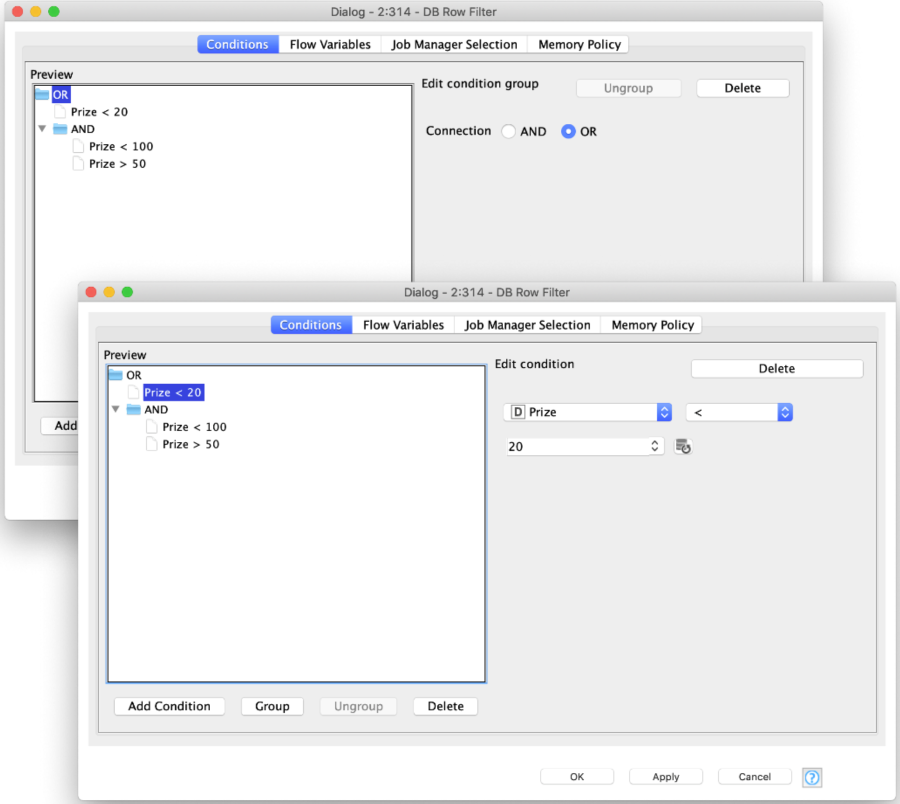
● Include TRUE matches ○ Exclude TRUE matches

Rule-based Row Filter



In-Database Row Filtering – DB Row Filter

- Creates a SQL statement to filter the rows that don’t match the conditions
- More than one condition is possible
- Allows you to create logical groups for AND and OR



Dialog - 2:314 - DB Row Filter

Conditions | Flow Variables | Job Manager Selection | Memory Policy

Preview

```
OR
  Prize < 20
▼ AND
    Prize < 100
    Prize > 50
```

Edit condition group

Connection ○ AND ● OR

Dialog - 2:314 - DB Row Filter

Conditions | Flow Variables | Job Manager Selection | Memory Policy

Preview

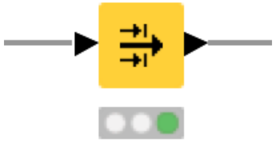
```
OR
  Prize < 20
▼ AND
    Prize < 100
    Prize > 50
```

Edit condition

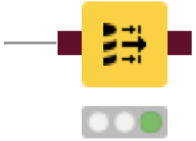
D Prize | <

20

Row Filter (Labs)

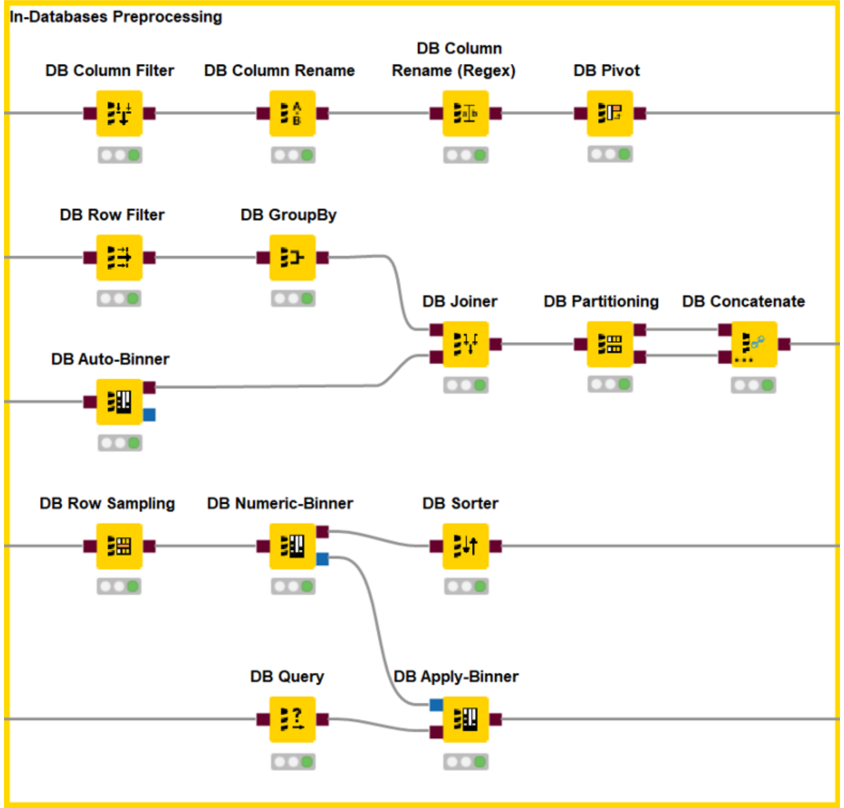


DB Row Filter



Query Nodes

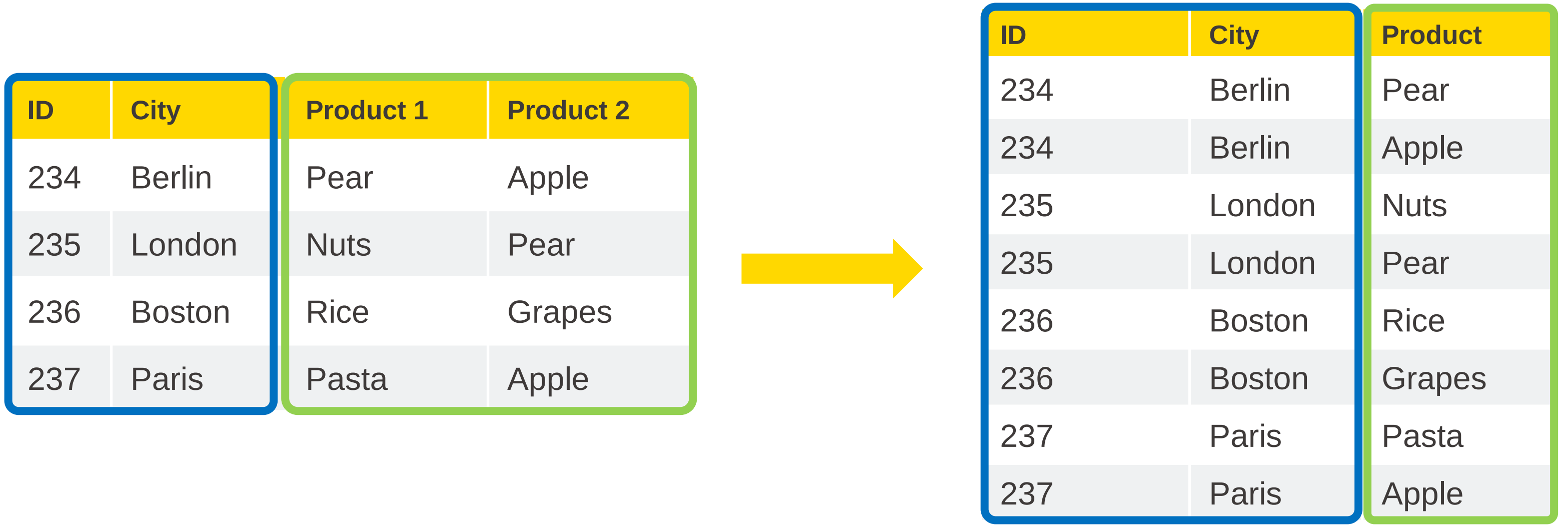
- Filter rows and columns
- Join tables/queries
- Concatenate tables
- Extract samples
- Bin numeric columns
- Sort your data
- Write your own query
- Aggregate your data



Exercise: 03_Data Cleaning (Part 2)

- Explore the data using the Data Explorer node
- Replace numeric outliers in the “Age” column with missing values
- If the age of a customer is missing, replace the birthday with a missing value
  Hint: Use the expression NOT MISSING Age=> $Birthday$
- Impute the missing values in the age column with the column mean
- Remove rows for duplicate CustomerIDs

Data Transformation



ID	City	Product 1	Product 2
234	Berlin	Pear	Apple
235	London	Nuts	Pear
236	Boston	Rice	Grapes
237	Paris	Pasta	Apple

ID	City	Product
234	Berlin	Pear
234	Berlin	Apple
235	London	Nuts
235	London	Pear
236	Boston	Rice
236	Boston	Grapes
237	Paris	Pasta
237	Paris	Apple

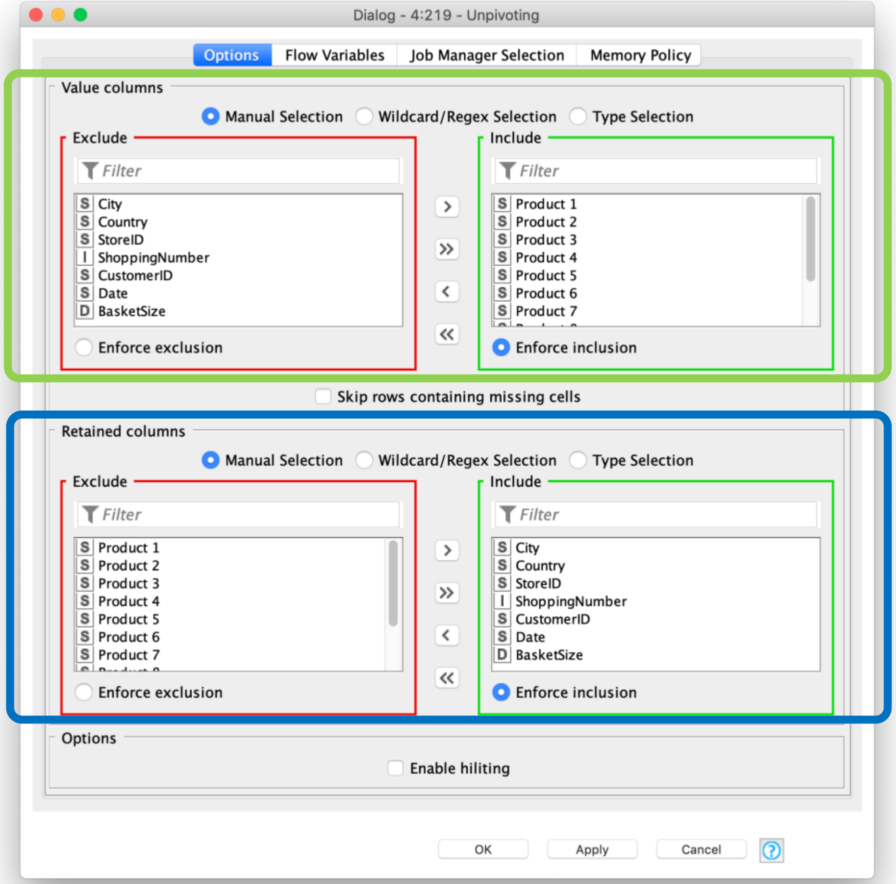
Value Column

Retaining Columns

Solution: Unpivoting Node

Unpivoting

- Rotates the value columns to rows
- Duplicates the remaining columns and appends them to each corresponding row



Dialog - 4:219 - Unpivoting

Value columns

● Manual Selection ○ Wildcard/Regex Selection ○ Type Selection

Exclude

Filter

- S City
- S Country
- S StoreID
- I ShoppingNumber
- S CustomerID
- S Date
- D BasketSize

○ Enforce exclusion

Include

Filter

- S Product 1
- S Product 2
- S Product 3
- S Product 4
- S Product 5
- S Product 6
- S Product 7

● Enforce inclusion

□ Skip rows containing missing cells

Retained columns

● Manual Selection ○ Wildcard/Regex Selection ○ Type Selection

Exclude

Filter

- S Product 1
- S Product 2
- S Product 3
- S Product 4
- S Product 5
- S Product 6
- S Product 7

○ Enforce exclusion

Include

Filter

- S City
- S Country
- S StoreID
- I ShoppingNumber
- S CustomerID
- S Date
- D BasketSize

● Enforce inclusion

Options

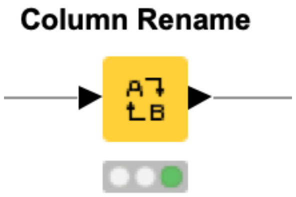
□ Enable hiliting

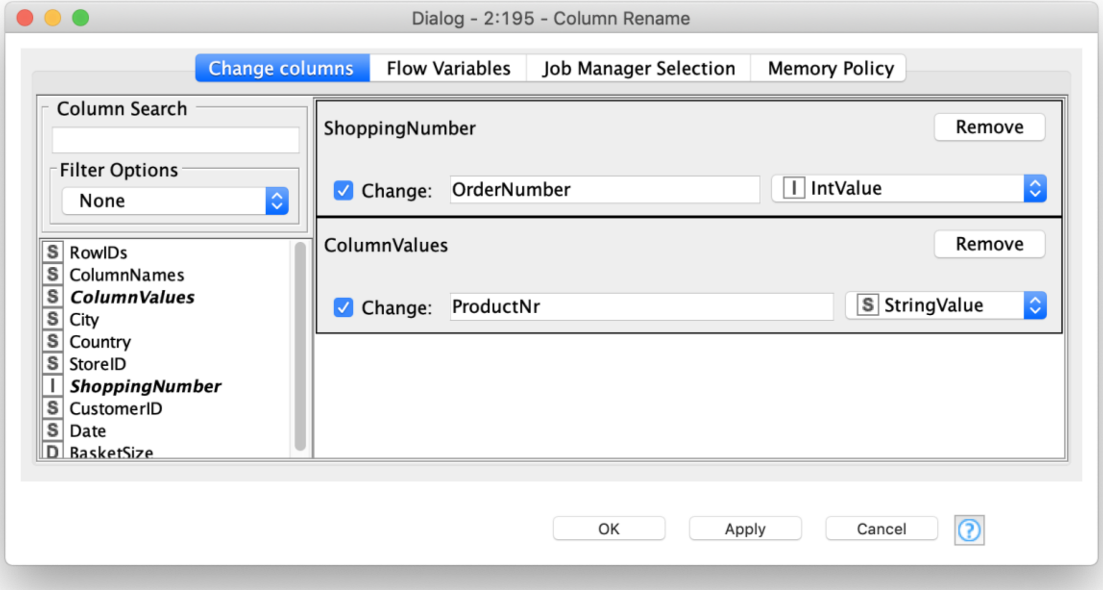
Value Column

Retaining Columns

Column Rename

- Renames column names or changes their types





Dialog - 2:195 - Column Rename

Change columns | Flow Variables | Job Manager Selection | Memory Policy

Column Search

Filter Options: None

- S RowIDs
- S ColumnNames
- S ***ColumnValues***
- S City
- S Country
- S StoreID
- I ***ShoppingNumber***
- S CustomerID
- S Date
- D BasketSize

ShoppingNumber — Remove

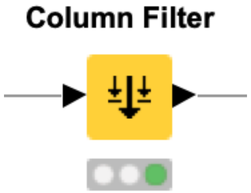
☑ Change: OrderNumber — I IntValue

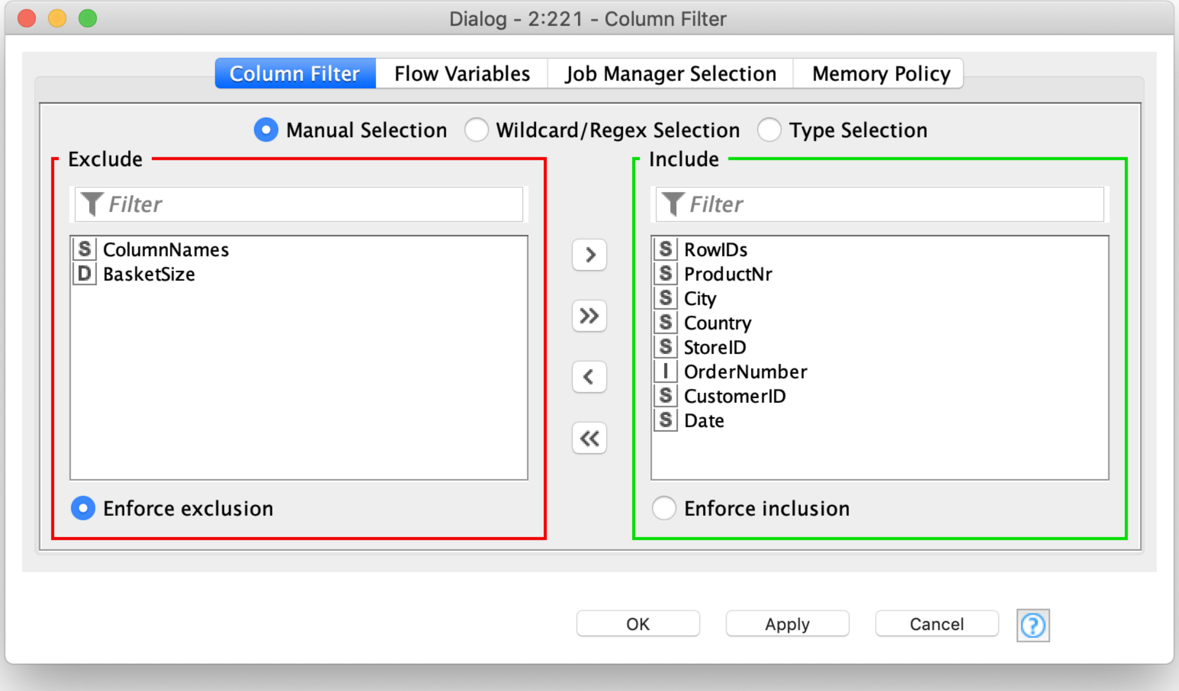
ColumnValues — Remove

☑ Change: ProductNr — S StringValue

Column Filter

- Excludes columns from the table by moving them to the Exclude list





● Manual Selection ○ Wildcard/Regex Selection ○ Type Selection

Exclude:

- S ColumnNames
- D BasketSize

● Enforce exclusion

Include:

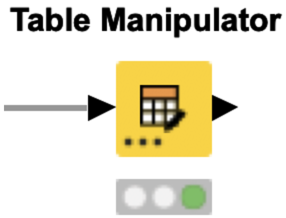
- S RowIDs
- S ProductNr
- S City
- S Country
- S StoreID
- I OrderNumber
- S CustomerID
- S Date

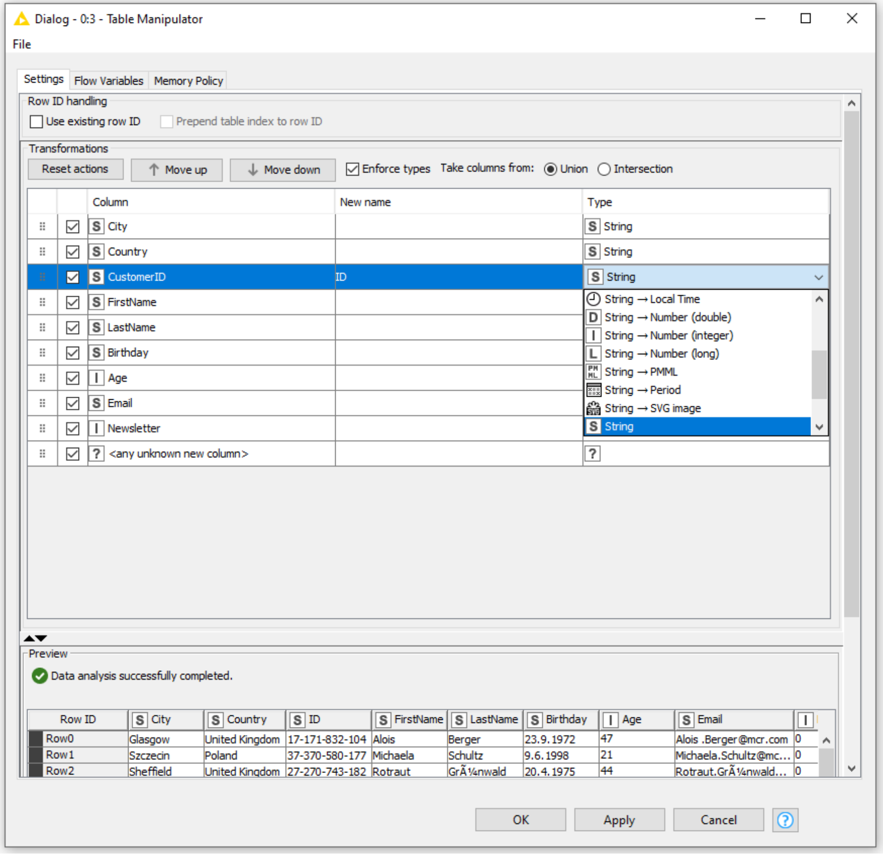
○ Enforce inclusion

Table Manipulator

Allows for

- Concatenation of multiple files/tables
- Column filtering
- Column sorting
- Column renaming
- Column type mapping





Row ID handling

□ Use existing row ID □ Prepend table index to row ID

Transformations

☑ Enforce types Take columns from: ● Union ○ Intersection

	Column	New name	Type
☑	S City		S String
☑	S Country		S String
☑	S CustomerID	ID	S String
☑	S FirstName		
☑	S LastName		
☑	S Birthday		
☑	I Age		
☑	S Email		
☑	I Newsletter		
☑	? <any unknown new column>		?

Preview

Data analysis successfully completed.

Row ID	S City	S Country	S ID	S FirstName	S LastName	S Birthday	I Age	S Email	I
Row0	Glasgow	United Kingdom	17-171-832-104	Alois	Berger	23.9.1972	47	Alois .Berger@mcr.com	0
Row1	Szczecin	Poland	37-370-580-177	Michaela	Schultz	9.6.1998	21	Michaela.Schultz@mc...	0
Row2	Sheffield	United Kingdom	27-270-743-182	Rotraut	GrÃ¼nwald	20.4.1975	44	Rotraut.GrÃ¼nwald...	0

Exercise: 04_Data_Transformation

- Change the structure of the table with the onsite purchases so that each purchased product is in a separate row and not the whole purchase event
  - Unpivot the columns that show the products ordered in one purchase event. Retain other columns in the table.
  - Remove rows that have missing values
  - Rename the "ColumnValues" column to "ProductNr" and "ShoppingNumber" to "OrderNumber" and remove unnecessary columns

Data Aggregation

Data Aggregation

RowID	Group	Value
r1	m	2
r2	f	3
r3	m	1
r4	f	5
r5	f	7
r6	m	5

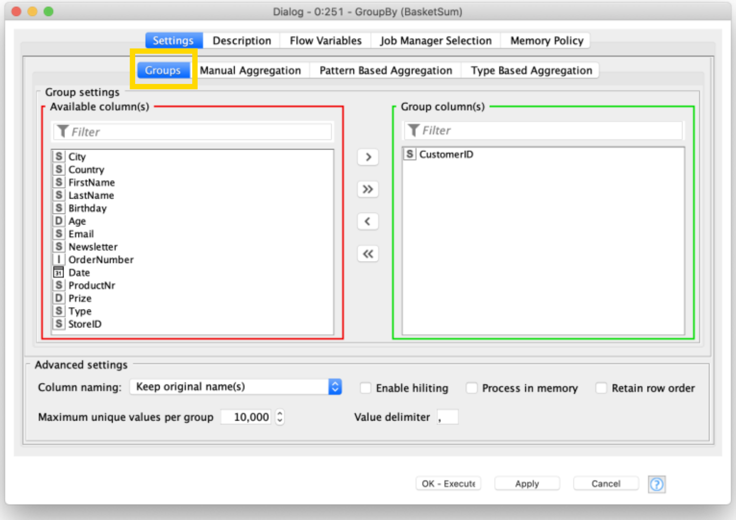
RowID	Group	Sum(Value)
r1+r3+r6	m	8
r2+r4+r5	f	15

aggregated on “group”
by method: sum(“value”)

GroupBy

Aggregate rows to summarize data

- First tab provides grouping options
- Second tab provides control over aggregation details



Dialog - 0:251 - GroupBy (BasketSum)

Settings | Description | Flow Variables | Job Manager Selection | Memory Policy

Groups | Manual Aggregation | Pattern Based Aggregation | Type Based Aggregation

Group settings

Available column(s)

Filter

S	City
S	Country
S	FirstName
S	LastName
S	Birthday
D	Age
S	Email
S	Newsletter
I	OrderNumber
31	Date
S	ProductNr
D	Prize
S	Type
S	StoreID

Group column(s)

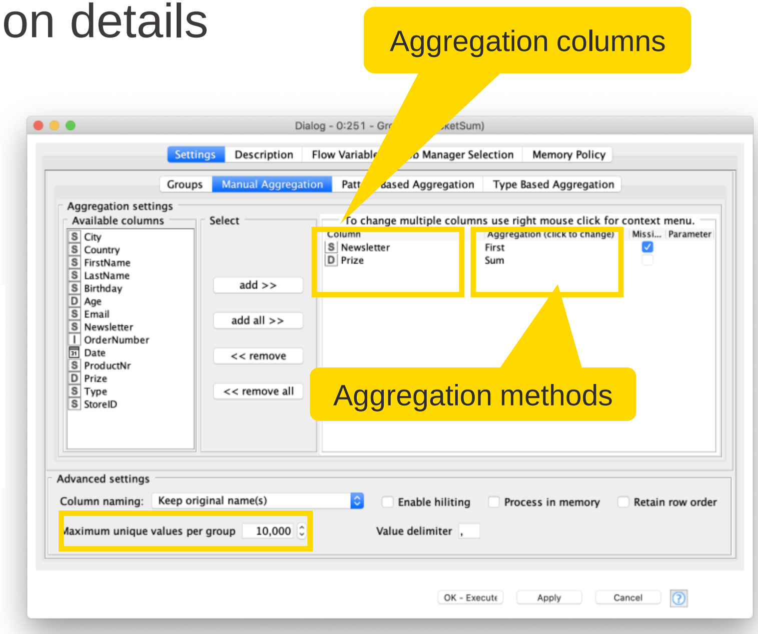
Filter

S	CustomerID

Advanced settings

Column naming: Keep original name(s) □ Enable hiliting □ Process in memory □ Retain row order

Maximum unique values per group 10,000 Value delimiter ,



Aggregation columns

Aggregation methods

Dialog - 0:251 - Gr ketSum)

Settings | Description | Flow Variable o Manager Selection | Memory Policy

Groups | Manual Aggregation | Patt Based Aggregation | Type Based Aggregation

Aggregation settings

Available columns

S	City
S	Country
S	FirstName
S	LastName
S	Birthday
D	Age
S	Email
S	Newsletter
I	OrderNumber
31	Date
S	ProductNr
D	Prize
S	Type
S	StoreID

Select

o change multiple columns use right mouse click for context menu.

Column	Aggregation (click to change)	Missi...	Parameter
S Newsletter	First	☑	
D Prize	Sum	□	

Advanced settings

Column naming: Keep original name(s) □ Enable hiliting □ Process in memory □ Retain row order

Maximum unique values per group 10,000 Value delimiter ,

YouTube KNIME TV video: https://youtu.be/bDwF-TOMtWw

Data Aggregation

Gender	Hair	Age
f	blond	31
m	red	22
f	blond	53
m	brown	16
f	brown	47
f	black	22
m	blond	13
m	red	55

Aggregation: Count

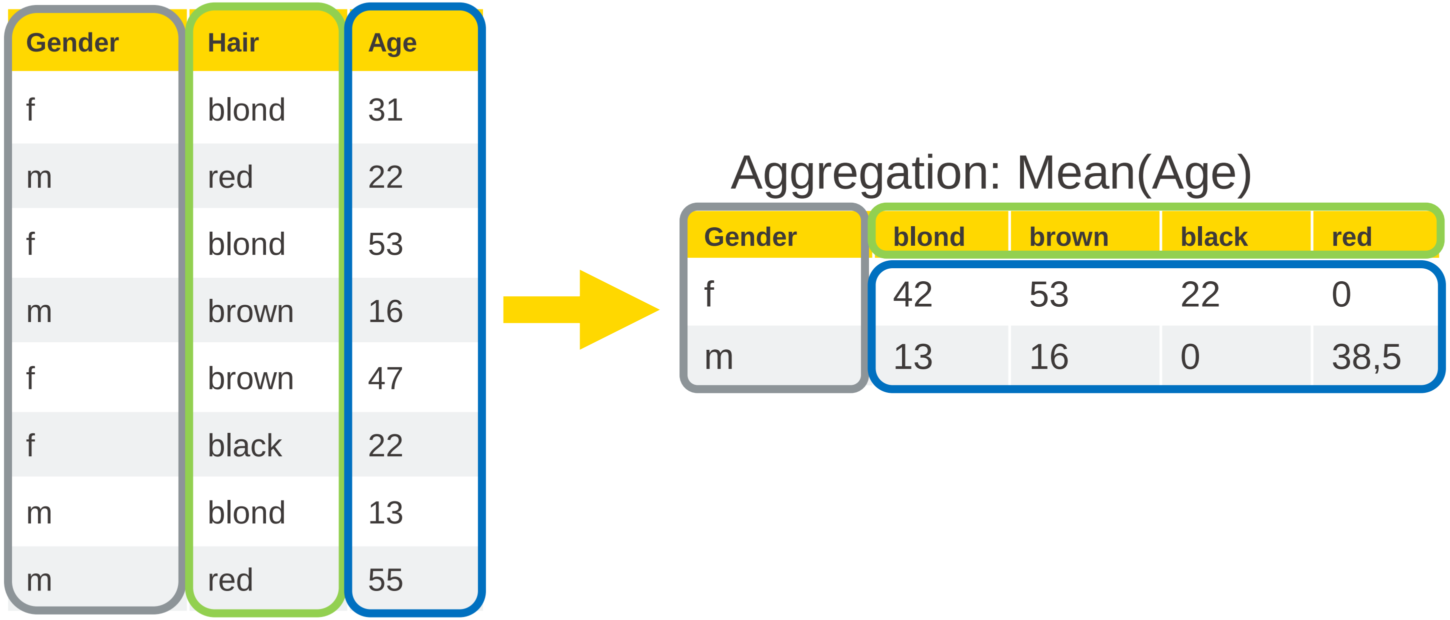
Gender	blond	brown	black	red
f	2	1	1	0
m	1	1	0	2

Aggregation: Mean(Age)

Gender	blond	brown	black	red
f	42	47	22	0
m	13	16	0	38,5

Solution: Pivoting Node

Data Aggregation



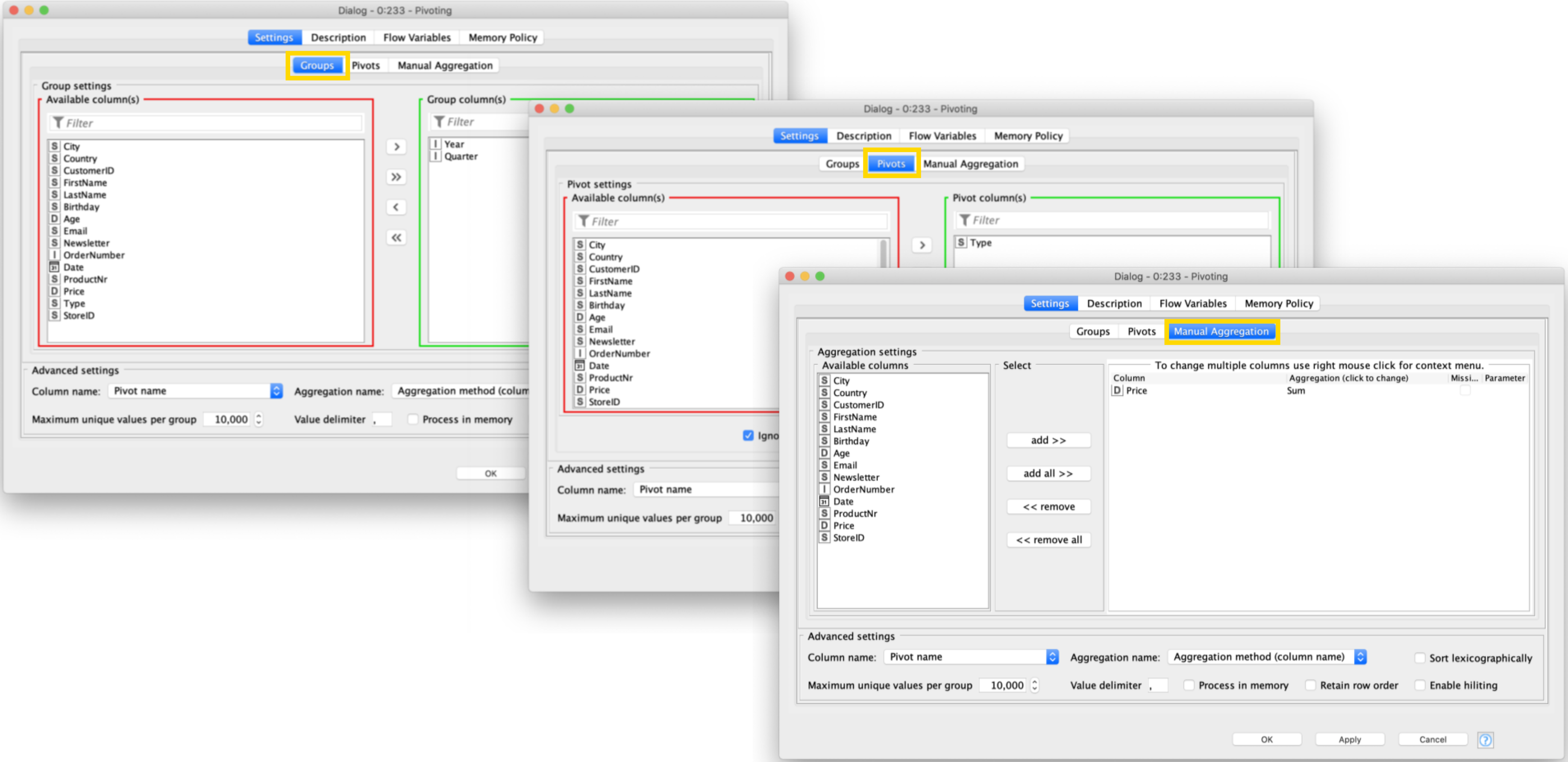
Gender	Hair	Age
f	blond	31
m	red	22
f	blond	53
m	brown	16
f	brown	47
f	black	22
m	blond	13
m	red	55

Aggregation: Mean(Age)

Gender	blond	brown	black	red
f	42	53	22	0
m	13	16	0	38,5

Pivoting Node: Group - Pivot - Aggregate

Pivoting



Dialog - 0:233 - Pivoting

Settings | Description | Flow Variables | Memory Policy

Groups | Pivots | Manual Aggregation

Group settings

Available column(s)

- S City
- S Country
- S CustomerID
- S FirstName
- S LastName
- S Birthday
- D Age
- S Email
- S Newsletter
- I OrderNumber
- Date
- S ProductNr
- D Price
- S Type
- S StoreID

Group column(s)

- I Year
- I Quarter

Advanced settings

Column name: Pivot name Aggregation name: Aggregation method (colum

Maximum unique values per group 10,000 Value delimiter , □ Process in memory

OK

Dialog - 0:233 - Pivoting

Settings | Description | Flow Variables | Memory Policy

Groups | **Pivots** | Manual Aggregation

Pivot settings

Available column(s)

- S City
- S Country
- S CustomerID
- S FirstName
- S LastName
- S Birthday
- D Age
- S Email
- S Newsletter
- I OrderNumber
- Date
- S ProductNr
- D Price
- S StoreID

Pivot column(s)

- S Type

☑ Igno

Advanced settings

Column name: Pivot name

Maximum unique values per group 10,000

Dialog - 0:233 - Pivoting

Settings | Description | Flow Variables | Memory Policy

Groups | Pivots | **Manual Aggregation**

Aggregation settings

Available columns

- S City
- S Country
- S CustomerID
- S FirstName
- S LastName
- S Birthday
- D Age
- S Email
- S Newsletter
- I OrderNumber
- Date
- S ProductNr
- D Price
- S StoreID

Select

add >> | add all >> | << remove | << remove all

To change multiple columns use right mouse click for context menu.

Column	Aggregation (click to change)	Missi...	Parameter
D Price	Sum	□	

Advanced settings

Column name: Pivot name Aggregation name: Aggregation method (column name) □ Sort lexicographically

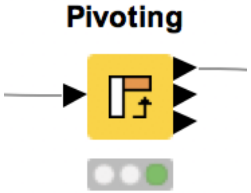
Maximum unique values per group 10,000 Value delimiter , □ Process in memory □ Retain row order □ Enable hiliting

OK | Apply | Cancel

Pivoting

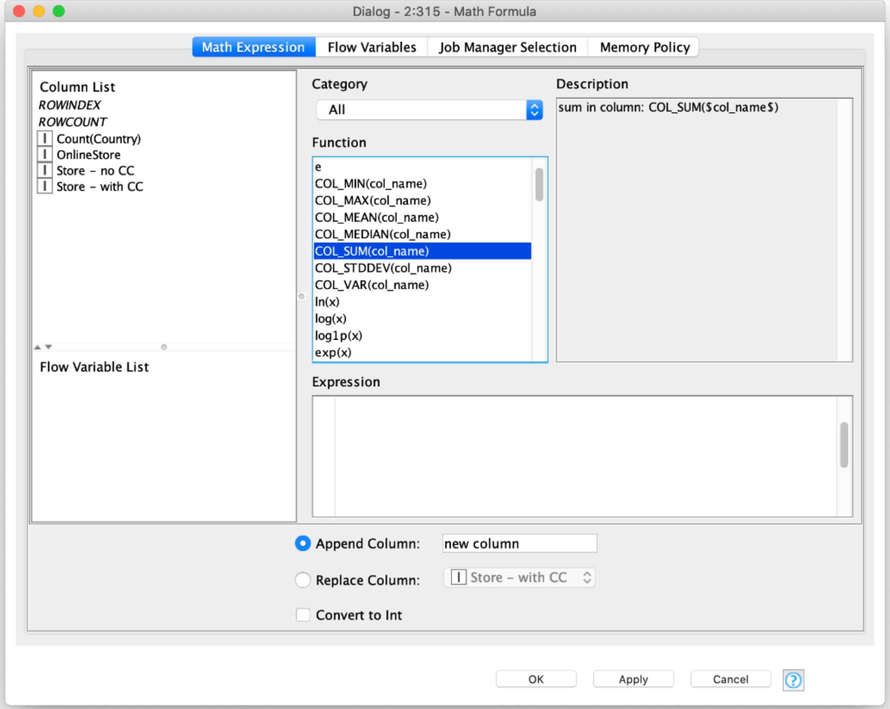
Performs pivoting on selected columns for grouping and pivoting

- Values of group columns become unique rows
- Values of the pivot columns become unique columns for each set of column combinations together with each aggregation
- Many aggregation methods are provided



Math Formula

- Row-wise calculations
- Some col-wise statistics
- Many mathematical functions
- Double-click function, then select col by click



Dialog - 2:315 - Math Formula

Math Expression | Flow Variables | Job Manager Selection | Memory Policy

Column List
ROWINDEX
ROWCOUNT
Count(Country)
OnlineStore
Store – no CC
Store – with CC

Flow Variable List

Category: All

Function
e
COL_MIN(col_name)
COL_MAX(col_name)
COL_MEAN(col_name)
COL_MEDIAN(col_name)
COL_SUM(col_name)
COL_STDDEV(col_name)
COL_VAR(col_name)
ln(x)
log(x)
log1p(x)
exp(x)

Description: sum in column: COL_SUM(col_name)

Expression

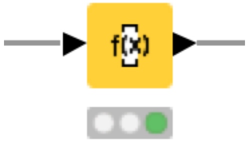
● Append Column: new column

○ Replace Column: Store – with CC

□ Convert to Int

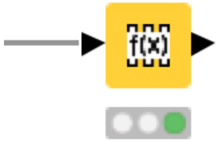
OK | Apply | Cancel

Math Formula

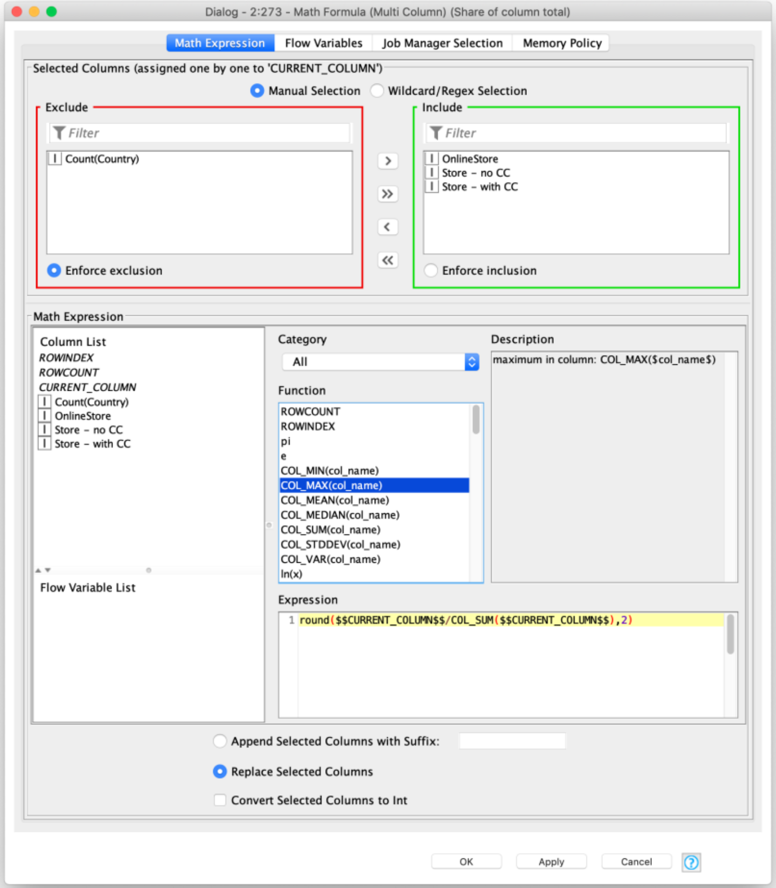


Math Formula (Multi Column)

- Useful if you want to make the same calculations on multiple columns.
- The selected columns from the upper part are called CURRENT_COLUMN in the Column List and Expression dialog.



Math Formula (Multi Column)



Dialog - 2:273 - Math Formula (Multi Column) (Share of column total)

Selected Columns (assigned one by one to 'CURRENT_COLUMN')

● Manual Selection ○ Wildcard/Regex Selection

Exclude

Filter

Count(Country)

● Enforce exclusion

Include

Filter

OnlineStore
Store – no CC
Store – with CC

○ Enforce inclusion

Math Expression

Column List

ROWINDEX
ROWCOUNT
CURRENT_COLUMN
Count(Country)
OnlineStore
Store – no CC
Store – with CC

Flow Variable List

Category: All

Function

ROWCOUNT
ROWINDEX
pi
e
COL_MIN(col_name)
COL_MAX(col_name)
COL_MEAN(col_name)
COL_MEDIAN(col_name)
COL_SUM(col_name)
COL_STDDEV(col_name)
COL_VAR(col_name)
ln(x)

Description

maximum in column: COL_MAX(col_name)

Expression

```
round($$CURRENT_COLUMN$$/COL_SUM($$CURRENT_COLUMN$$),2)
```

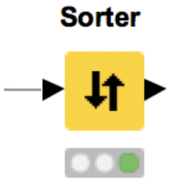
○ Append Selected Columns with Suffix:

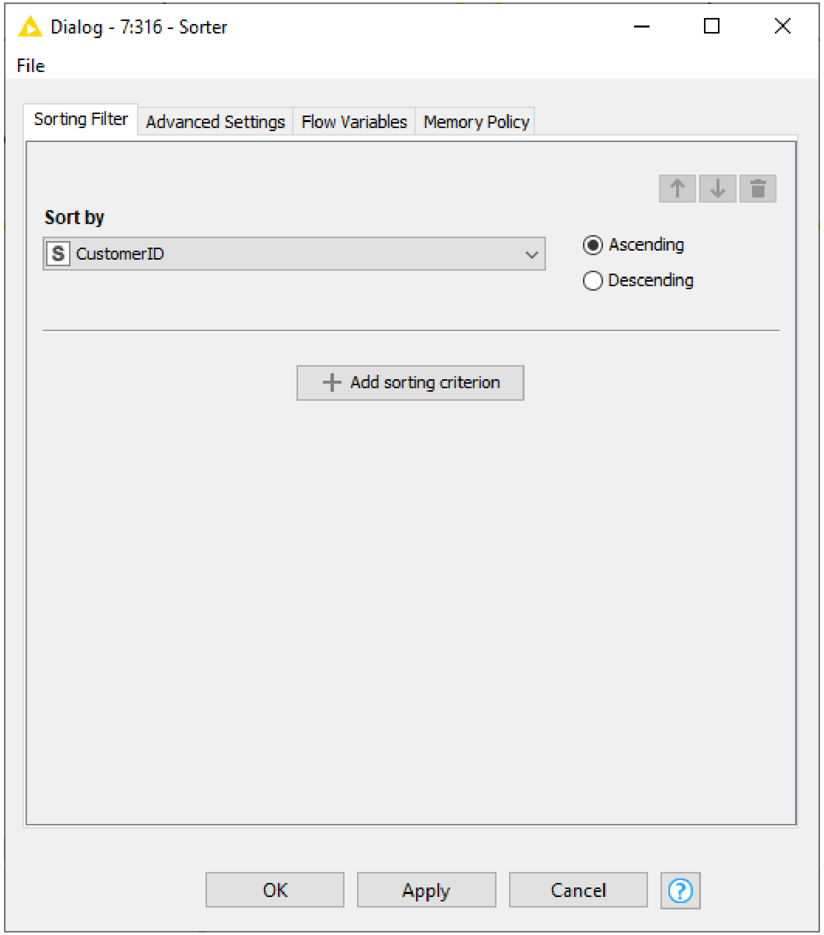
● Replace Selected Columns

□ Convert Selected Columns to Int

Sorter

- Sorts the rows based on the values of the selected column(s), either
  - ascending or
  - descending





Dialog - 7:316 - Sorter

File

Sort by

S CustomerID

● Ascending

○ Descending

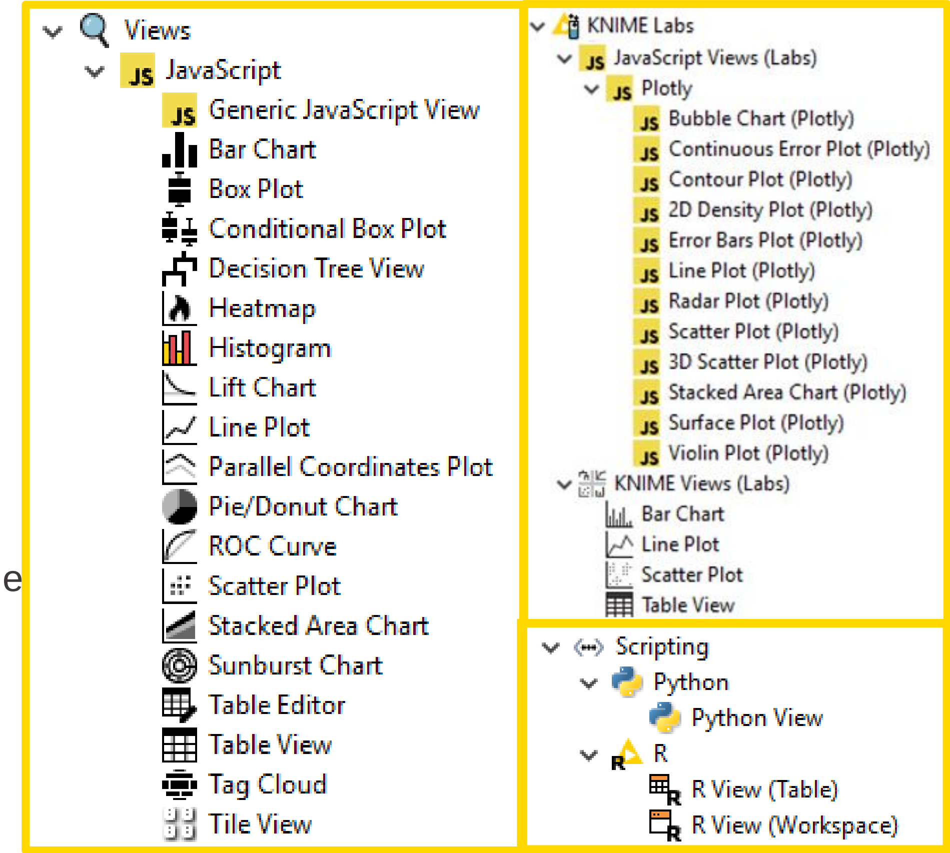
Exercise: 07_Data_Aggregation

- Calculate the total purchase amount by a customer ID both in 2019 and earlier
- Calculate the total purchase amount by quarter and transaction type
- Calculate the numbers of orders by basket size and transaction type (optional)
- Convert the dates of births of the customers to Date&Time and extract the birth year into a separate column (optional)

Data Visualization

Data Visualization

- Large selection of easy to use visualization nodes
  - Web-based and interactive
  - Dedicated nodes,
    no scripting required
- Plotly nodes
  - Similar but integrated from an external library
- New Visualization Nodes in Labs
  - A live preview of the visualization next to the configuration dialog
- R and Python View nodes for highly customizable graphics
  - Require scripting



- Views
  - JavaScript
    - Generic JavaScript View
    - Bar Chart
    - Box Plot
    - Conditional Box Plot
    - Decision Tree View
    - Heatmap
    - Histogram
    - Lift Chart
    - Line Plot
    - Parallel Coordinates Plot
    - Pie/Donut Chart
    - ROC Curve
    - Scatter Plot
    - Stacked Area Chart
    - Sunburst Chart
    - Table Editor
    - Table View
    - Tag Cloud
    - Tile View

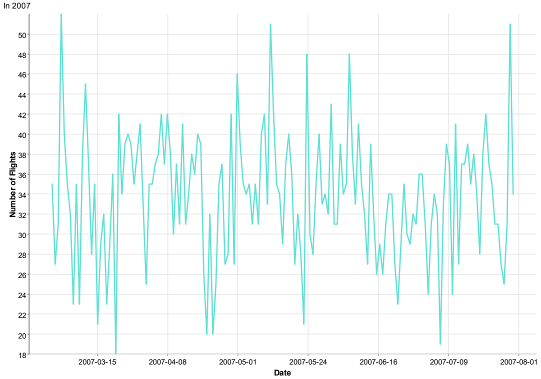
- KNIME Labs
  - JavaScript Views (Labs)
    - Plotly
      - Bubble Chart (Plotly)
      - Continuous Error Plot (Plotly)
      - Contour Plot (Plotly)
      - 2D Density Plot (Plotly)
      - Error Bars Plot (Plotly)
      - Line Plot (Plotly)
      - Radar Plot (Plotly)
      - Scatter Plot (Plotly)
      - 3D Scatter Plot (Plotly)
      - Stacked Area Chart (Plotly)
      - Surface Plot (Plotly)
      - Violin Plot (Plotly)
  - KNIME Views (Labs)
    - Bar Chart
    - Line Plot
    - Scatter Plot
    - Table View

- Scripting
  - Python
    - Python View
  - R
    - R View (Table)
    - R View (Workspace)

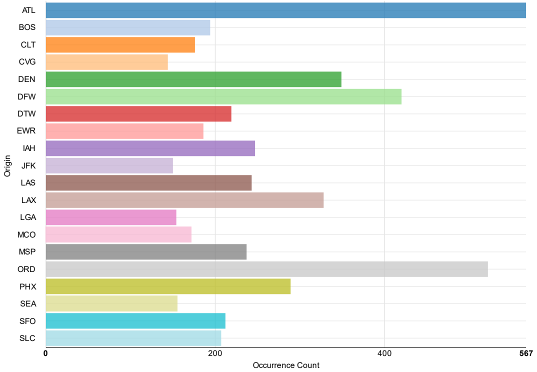
Visualizations Using One Column

Number of Flights by Date

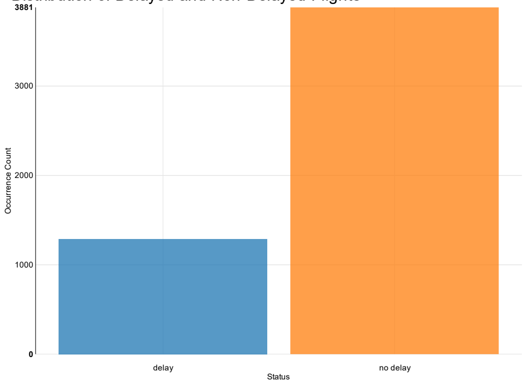
In 2007



Departure Airports

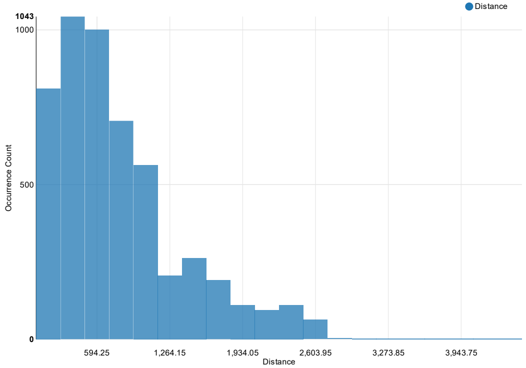


Distribution of Delayed and Non-Delayed Flights



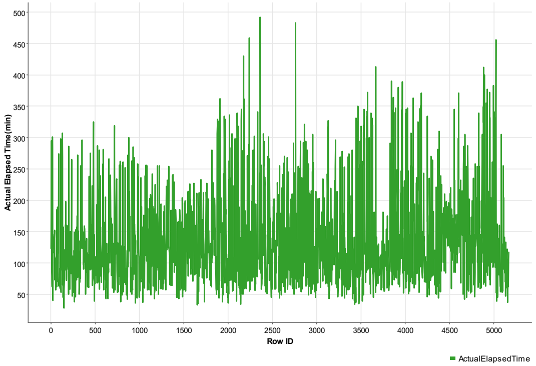
Distribution of Flight Distances

Domestic flights in 2007



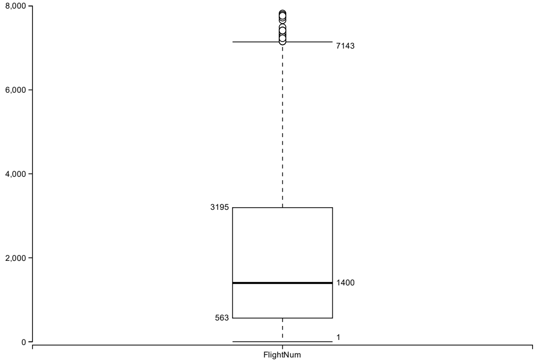
Elapsed Flight Time by Row ID

In 2007



Distribution of Flight Numbers

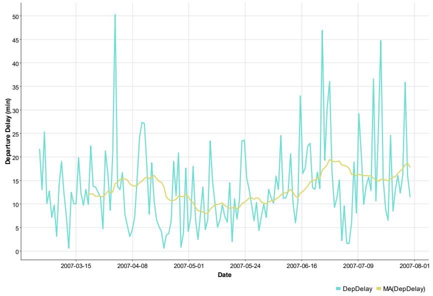
Modest outliers are marked by dots, strong outliers by crosses



Visualizations Using Two Columns

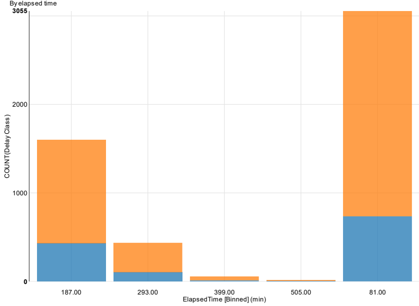
Delay Times by Date

Yellow line shows the moving average over the last 20 days



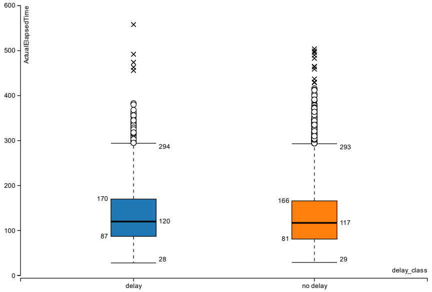
Number of Delayed and Non-Delayed Flights

By elapsed time

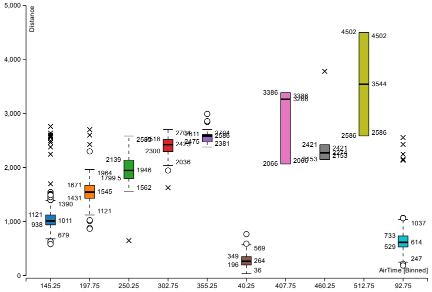


Distribution of Elapsed Time

By delay class

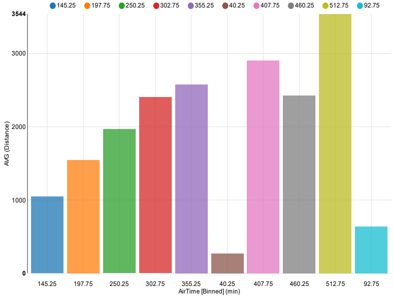


Distribution of Flight Distances by Flight Times



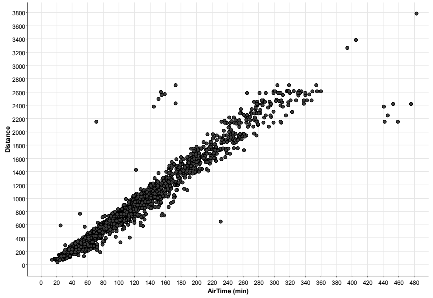
Average Distances by Flight Times

Domestic flights in 2007



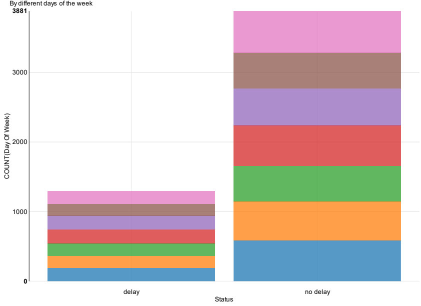
Correlation between Flight Time and Distance

Domestic flights in 2007

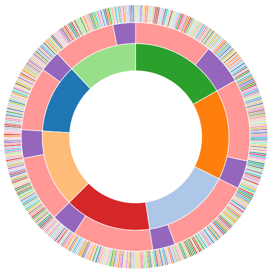


Number of Delayed and Non-Delayed Flights

By different days of the week



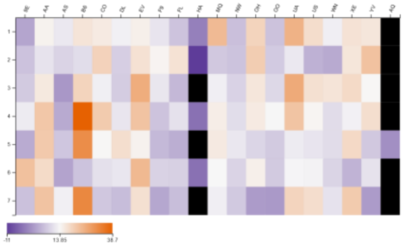
Flights by Day of the Week and Delay Status



Visualizations Using Three Columns

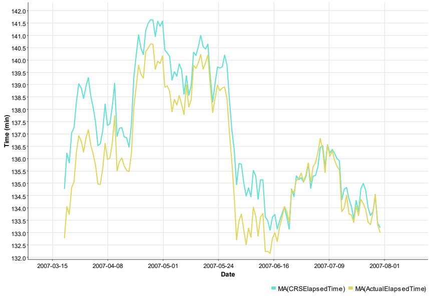
Delay Times by Day of the Week and Carrier

Red indicates large delay



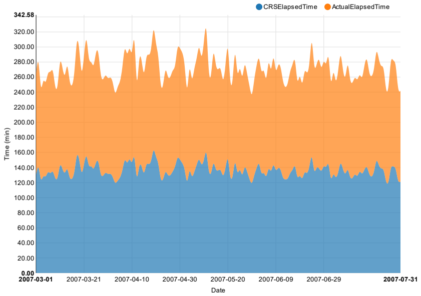
Actual Elapsed Time vs CRS Elapsed Time

Moving average over the last 20 days



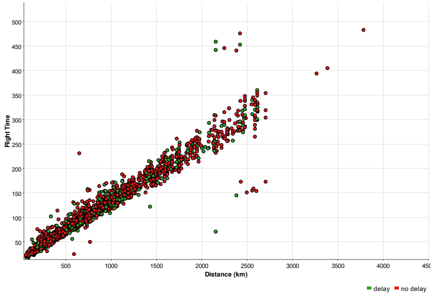
Actual Elapsed Time and CRS Elapsed Time

Over time in 2007



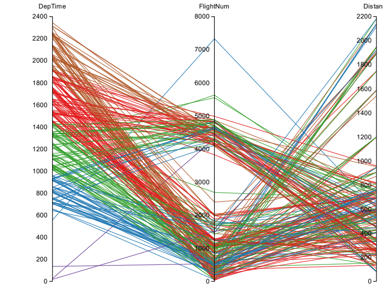
Correlation between Distance and Flight Time

Red indicates a flight on time, green a delayed flight



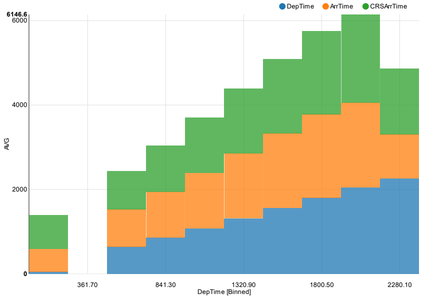
Departure Time vs Flight Number vs Distance

Colors indicate binned departure time



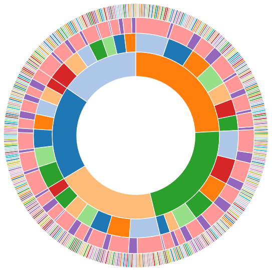
Average Arrival Time and CRS Arrival Time

By binned departure time



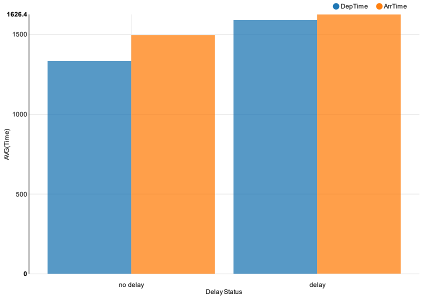
Flights by Delay Status, Month, and Day of the Week

Outest donut indicates row IDs



Average Departure and Arrival Times

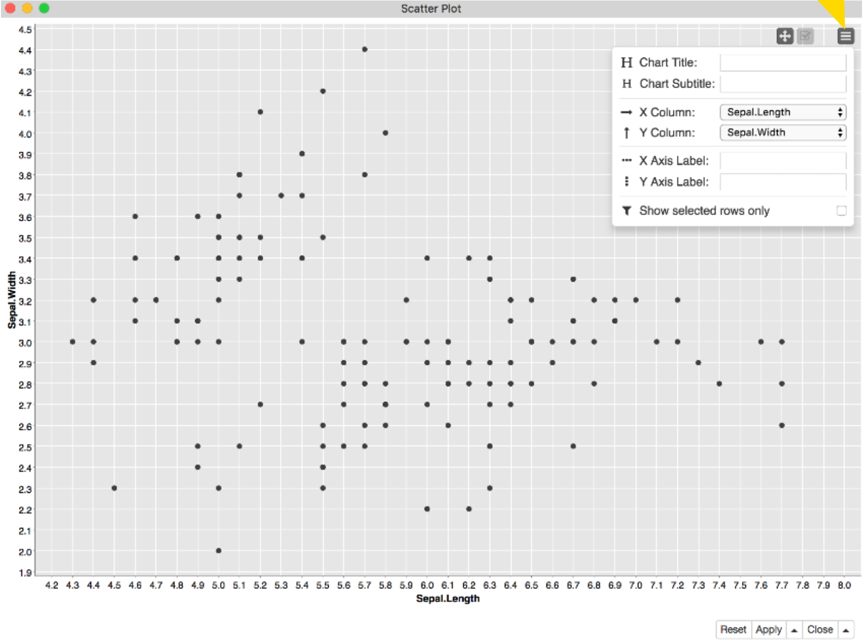
By delay status



Scatter Plot

- Plots different columns on X and Y
- Displays data including color information
- Produces an interactive view and an image

Interactivity options



Scatter Plot

H Chart Title:

H Chart Subtitle:

X Column: Sepal.Length

Y Column: Sepal.Width

X Axis Label:

Y Axis Label:

Show selected rows only ☐

Scatter Plot

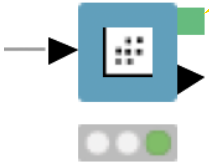
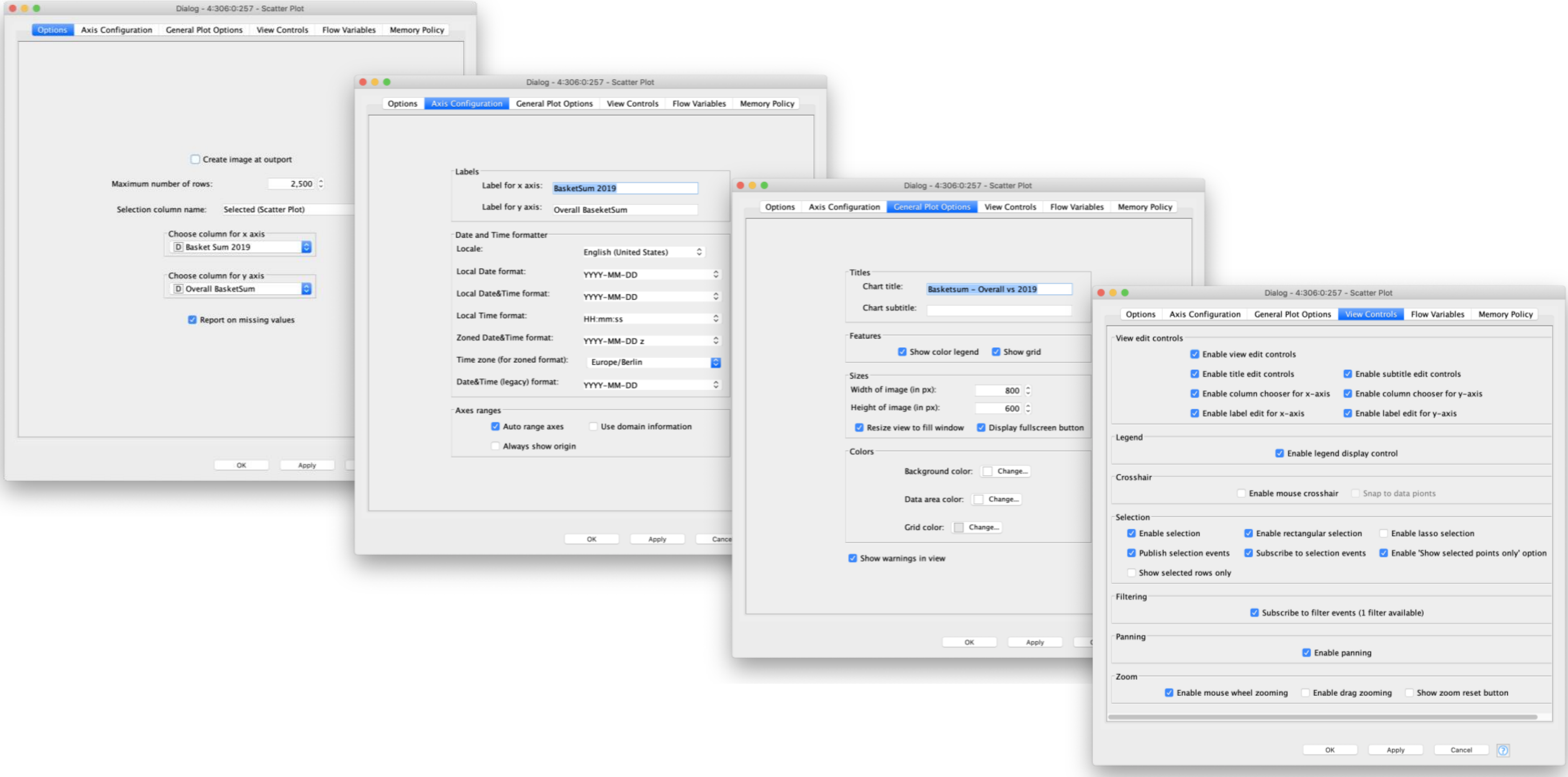


Image outport

Scatter Plot



Dialog - 4:306:0:257 - Scatter Plot
Options | Axis Configuration | General Plot Options | View Controls | Flow Variables | Memory Policy
OK | Apply | Cancel

Options

□ Create image at outport

Maximum number of rows: 2,500

Selection column name: Selected (Scatter Plot)

Choose column for x axis: D Basket Sum 2019

Choose column for y axis: D Overall BasketSum

☑ Report on missing values

Axis Configuration

Labels

Label for x axis: BasketSum 2019

Label for y axis: Overall BaseketSum

Date and Time formatter

Locale: English (United States)

Local Date format: YYYY-MM-DD

Local Date&Time format: YYYY-MM-DD

Local Time format: HH:mm:ss

Zoned Date&Time format: YYYY-MM-DD z

Time zone (for zoned format): Europe/Berlin

Date&Time (legacy) format: YYYY-MM-DD

Axes ranges

☑ Auto range axes □ Use domain information

□ Always show origin

General Plot Options

Titles

Chart title: Basketsum - Overall vs 2019

Chart subtitle:

Features

☑ Show color legend ☑ Show grid

Sizes

Width of image (in px): 800

Height of image (in px): 600

☑ Resize view to fill window ☑ Display fullscreen button

Colors

Background color: Change...

Data area color: Change...

Grid color: Change...

☑ Show warnings in view

View Controls

View edit controls

☑ Enable view edit controls

☑ Enable title edit controls ☑ Enable subtitle edit controls

☑ Enable column chooser for x-axis ☑ Enable column chooser for y-axis

☑ Enable label edit for x-axis ☑ Enable label edit for y-axis

Legend

☑ Enable legend display control

Crosshair

□ Enable mouse crosshair □ Snap to data pionts

Selection

☑ Enable selection ☑ Enable rectangular selection □ Enable lasso selection

☑ Publish selection events ☑ Subscribe to selection events ☑ Enable 'Show selected points only' option

□ Show selected rows only

Filtering

☑ Subscribe to filter events (1 filter available)

Panning

☑ Enable panning

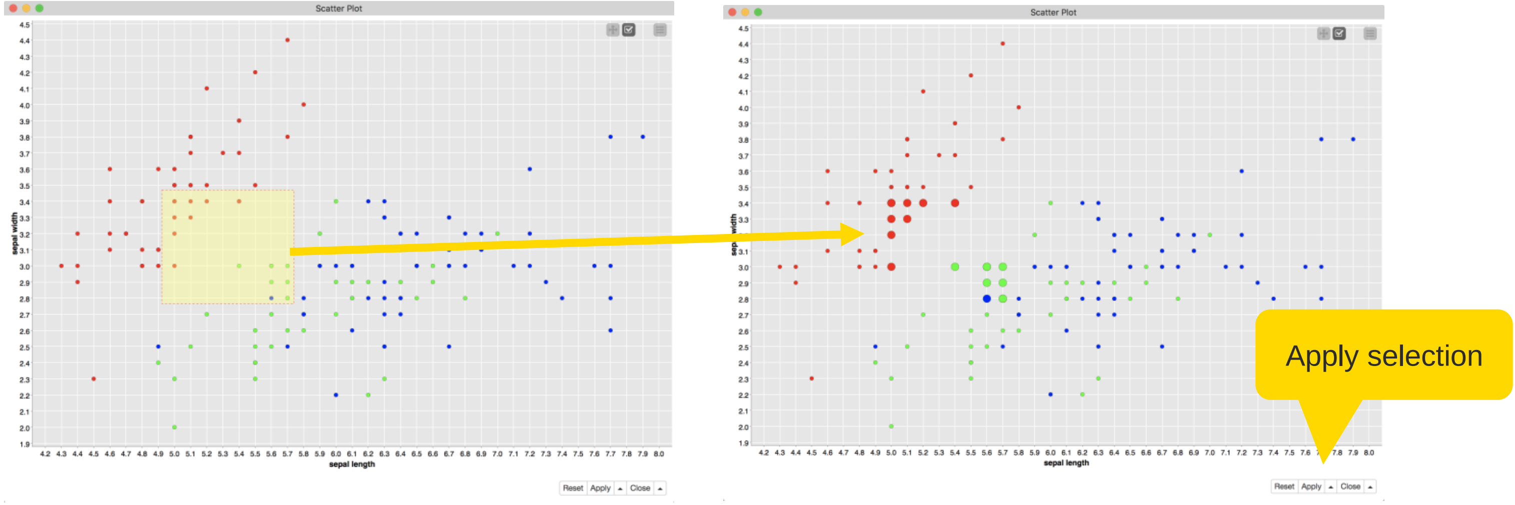
Zoom

☑ Enable mouse wheel zooming □ Enable drag zooming □ Show zoom reset button

Selection and Filtering in JavaScript Views

Interactivity allows you to select data points in views

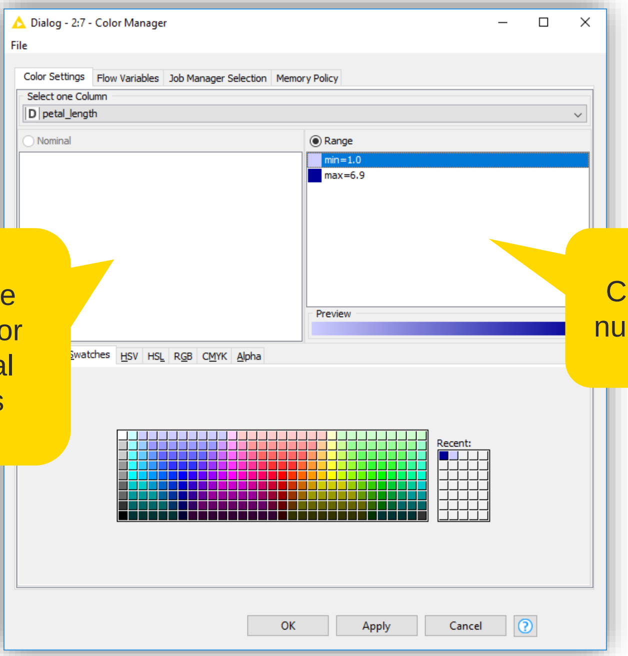
- Selection is propagated to other views
- You can highlight selected rows or filter them
- Click “Apply” to add column to data that indicates selection (true/false) for use in downstream nodes



Apply selection

Color Manager

- Colors by nominal or continuous values
- Syncs colors between views using the color model port and Color Appender node



Dialog - 2:7 - Color Manager

File

Color Settings | Flow Variables | Job Manager Selection | Memory Policy

Select one Column: D petal_length

○ Nominal

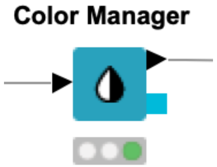
● Range

- min=1.0
- max=6.9

Preview

Discrete colors for nominal values

Color range for numerical values



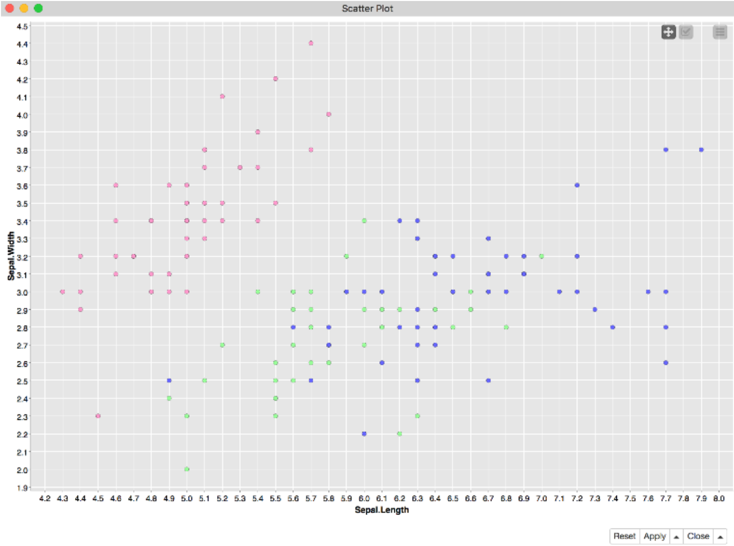
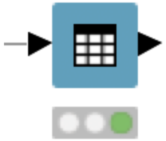
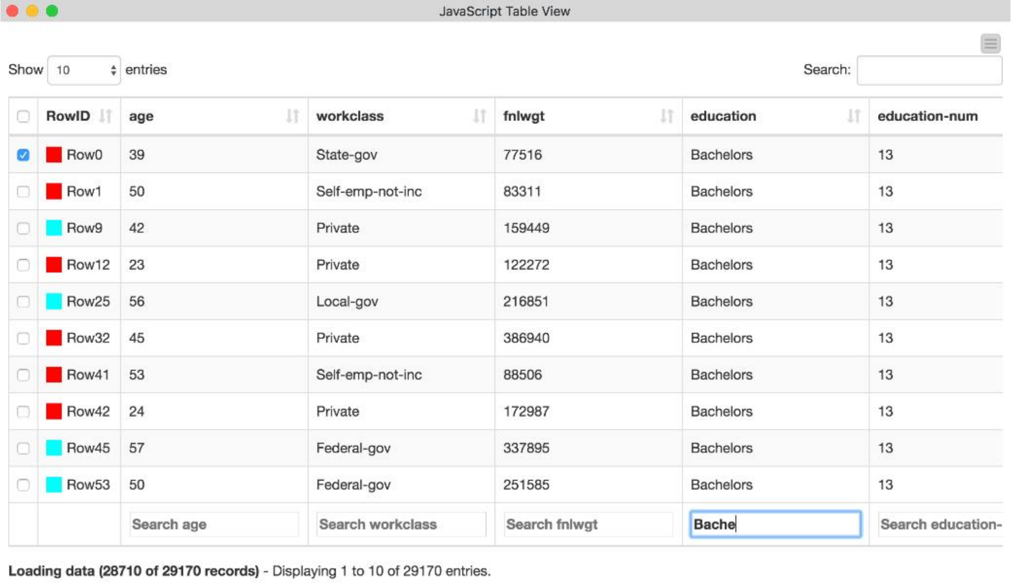


Table View

- Displays data in an HTML table view
- The view offers several interactive features, as well as the possibility to select rows

Table View





JavaScript Table View

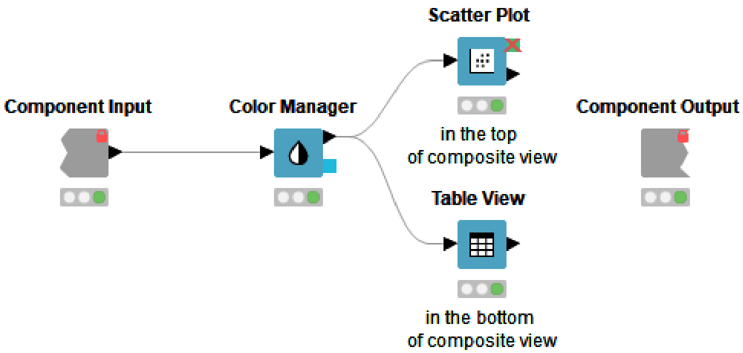
Show 10 entries Search:

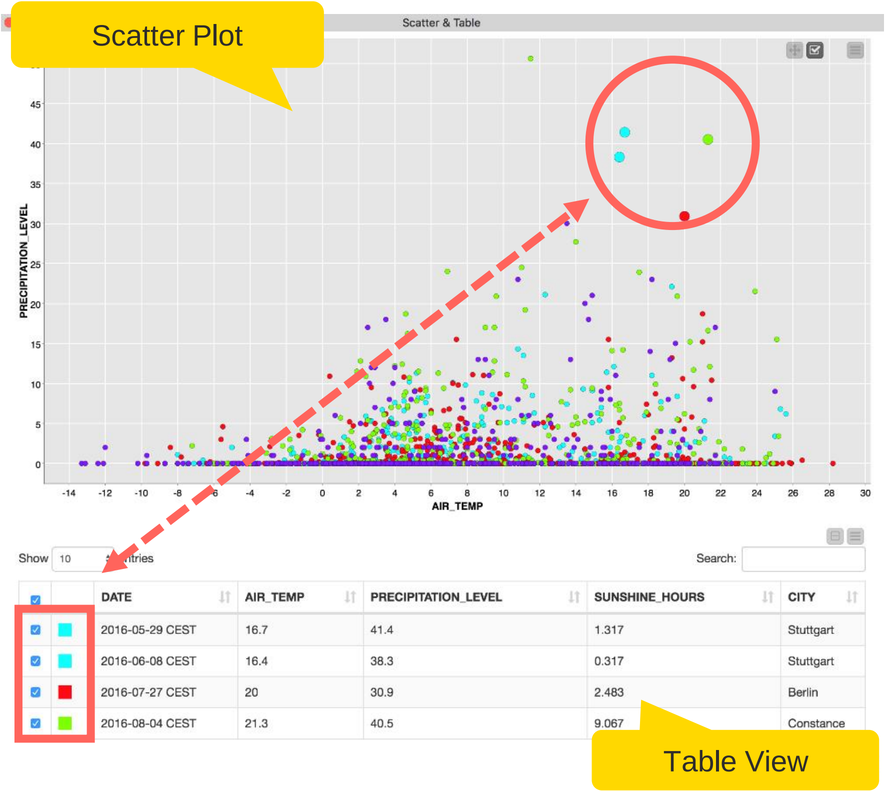
□	RowID	age	workclass	fnlwgt	education	education-num
☑	🟥 Row0	39	State-gov	77516	Bachelors	13
□	🟥 Row1	50	Self-emp-not-inc	83311	Bachelors	13
□	🟦 Row9	42	Private	159449	Bachelors	13
□	🟥 Row12	23	Private	122272	Bachelors	13
□	🟦 Row25	56	Local-gov	216851	Bachelors	13
□	🟥 Row32	45	Private	386940	Bachelors	13
□	🟥 Row41	53	Self-emp-not-inc	88506	Bachelors	13
□	🟥 Row42	24	Private	172987	Bachelors	13
□	🟦 Row45	57	Federal-gov	337895	Bachelors	13
□	🟦 Row53	50	Federal-gov	251585	Bachelors	13
		Search age	Search workclass	Search fnlwgt	Bache	Search education-

Loading data (28710 of 29170 records) - Displaying 1 to 10 of 29170 entries.

Components – Combined Views

- Multiple JavaScript View nodes can be combined in Components
- Selections are transmitted to all other views
- Also for use on the KNIME WebPortal



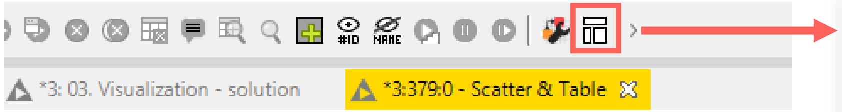


☑		DATE	AIR_TEMP	PRECIPITATION_LEVEL	SUNSHINE_HOURS	CITY
☑	🟦	2016-05-29 CEST	16.7	41.4	1.317	Stuttgart
☑	🟦	2016-06-08 CEST	16.4	38.3	0.317	Stuttgart
☑	🟥	2016-07-27 CEST	20	30.9	2.483	Berlin
☑	🟩	2016-08-04 CEST	21.3	40.5	9.067	Constance

Configure Content and Views Layout

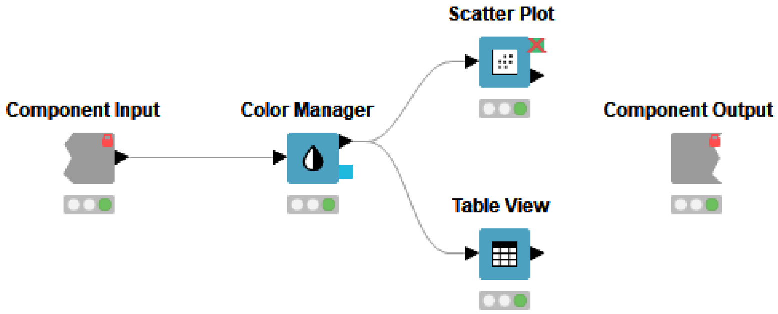
- Click layout button when inside Component to assign views to rows and columns

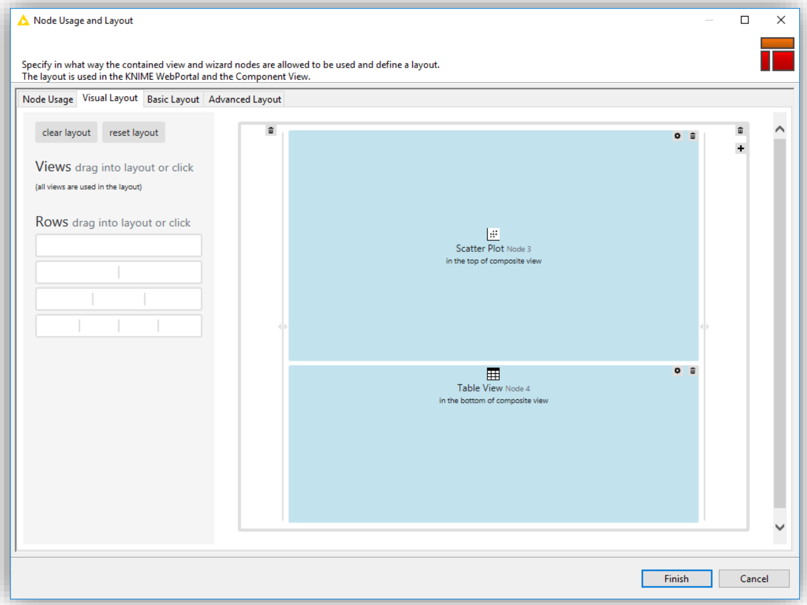
- Add views and rows via drag&drop
- Add columns using **+** buttons



*3: 03. Visualization - solution

*3:379:0 - Scatter & Table





Node Usage and Layout

Specify in what way the contained view and wizard nodes are allowed to be used and define a layout.
The layout is used in the KNIME WebPortal and the Component View.

Views drag into layout or click

(all views are used in the layout)

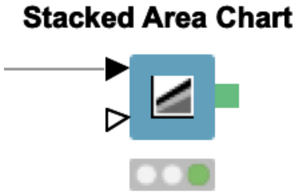
Rows drag into layout or click

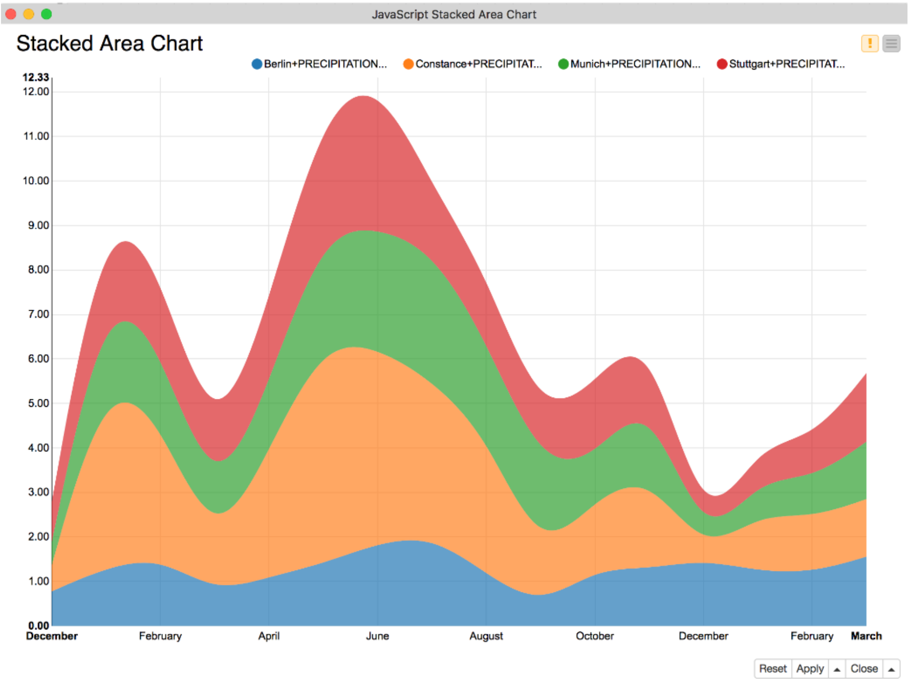
Scatter Plot Node 3
in the top of composite view

Table View Node 4
in the bottom of composite view

Stacked Area Chart

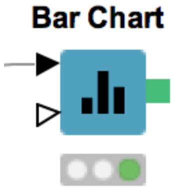
- Visualizes numerical values from multiple columns as stacked areas
- Great for plotting distributions over time

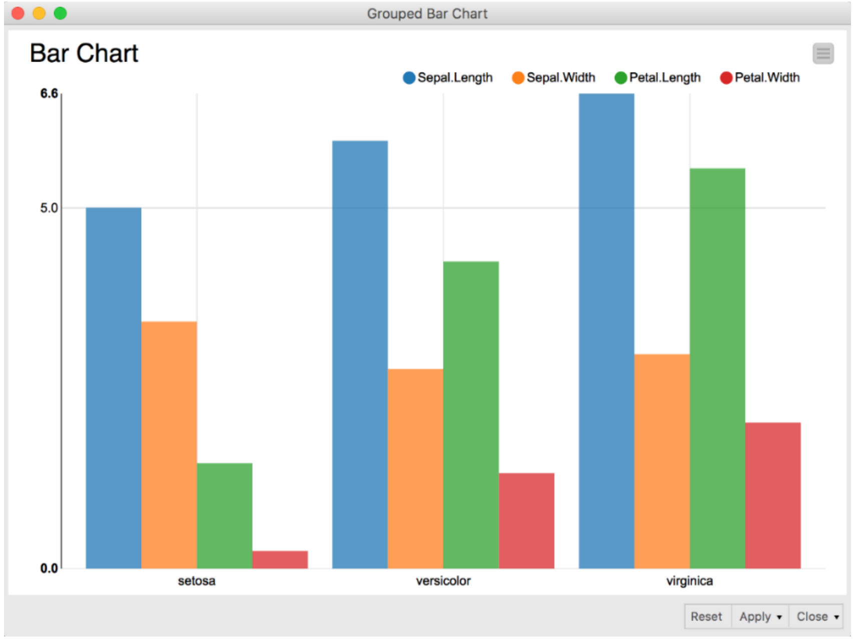




Bar Chart

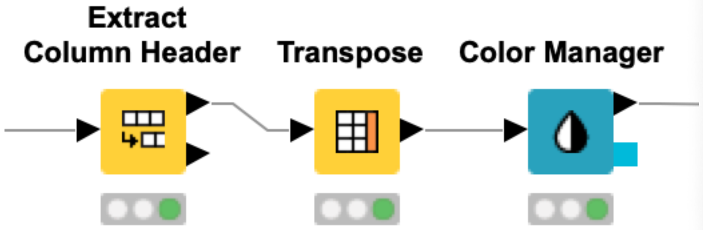
- Shows numerical values across categories
- Vertical or horizontal bars
- Bars can be grouped or stacked





The Optional Color Input Port

- Many of the visualization nodes have an optional port to change the colors
- Expects table with column headers of first table in the first column with assigned colors



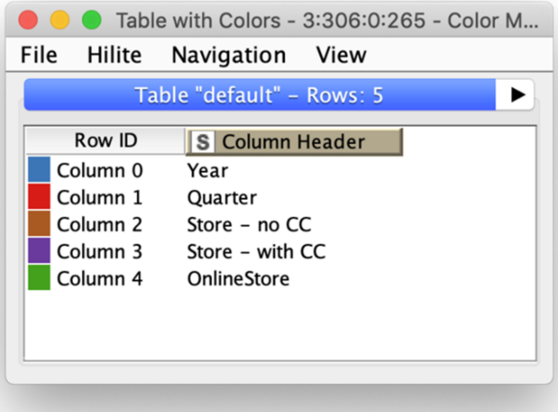
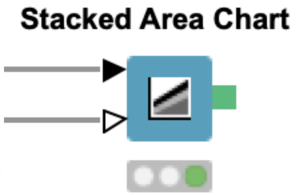


Table with Colors - 3:306:0:265 - Color M...
File Hilite Navigation View
Table "default" – Rows: 5

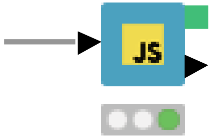
Row ID	S Column Header
🟦 Column 0	Year
🟥 Column 1	Quarter
🟫 Column 2	Store – no CC
🟪 Column 3	Store – with CC
🟩 Column 4	OnlineStore

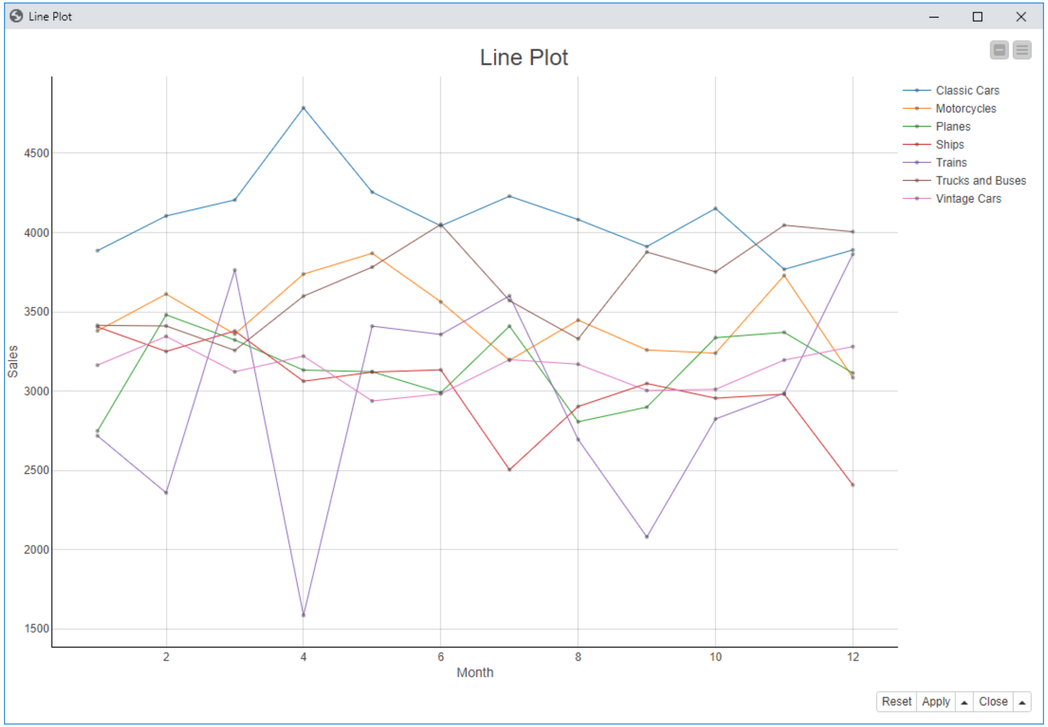


Line Plot

- Plots sequence of values, e.g. over time
- Useful to identify trends, also between groups

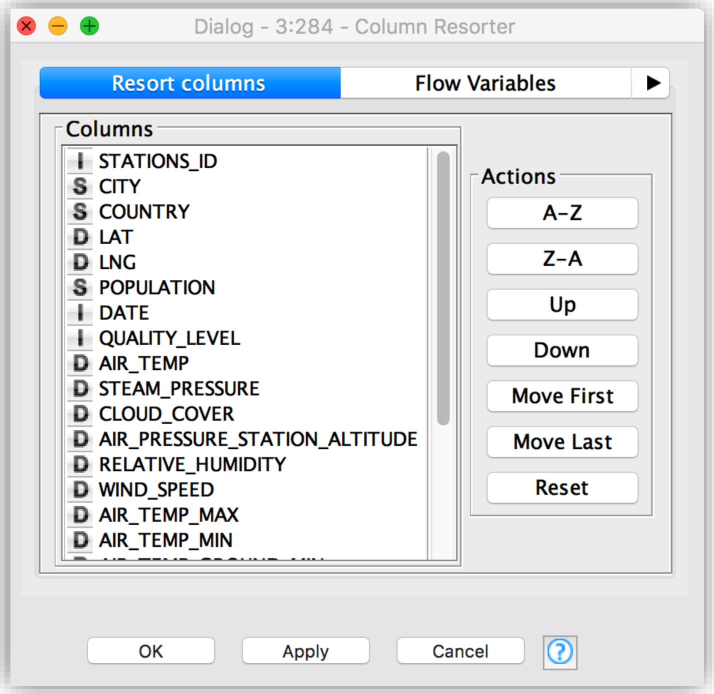
Line Plot (Plotly)





Column Resorter

- Changes the order of the input column based on user defined settings
- Options:
  - Sort alphabetical (A-Z or Z-A)
  - Move the selected columns one step (Up or Down)
  - Move the selected columns to top or end (Move First / Last)

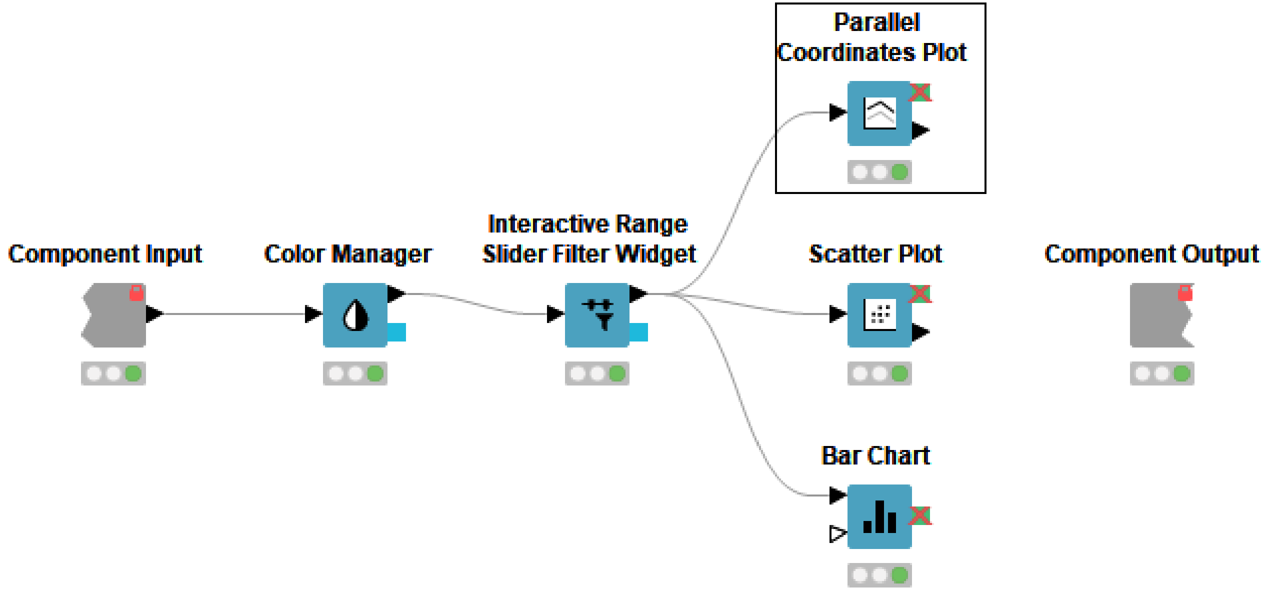


Dialog - 3:284 - Column Resorter

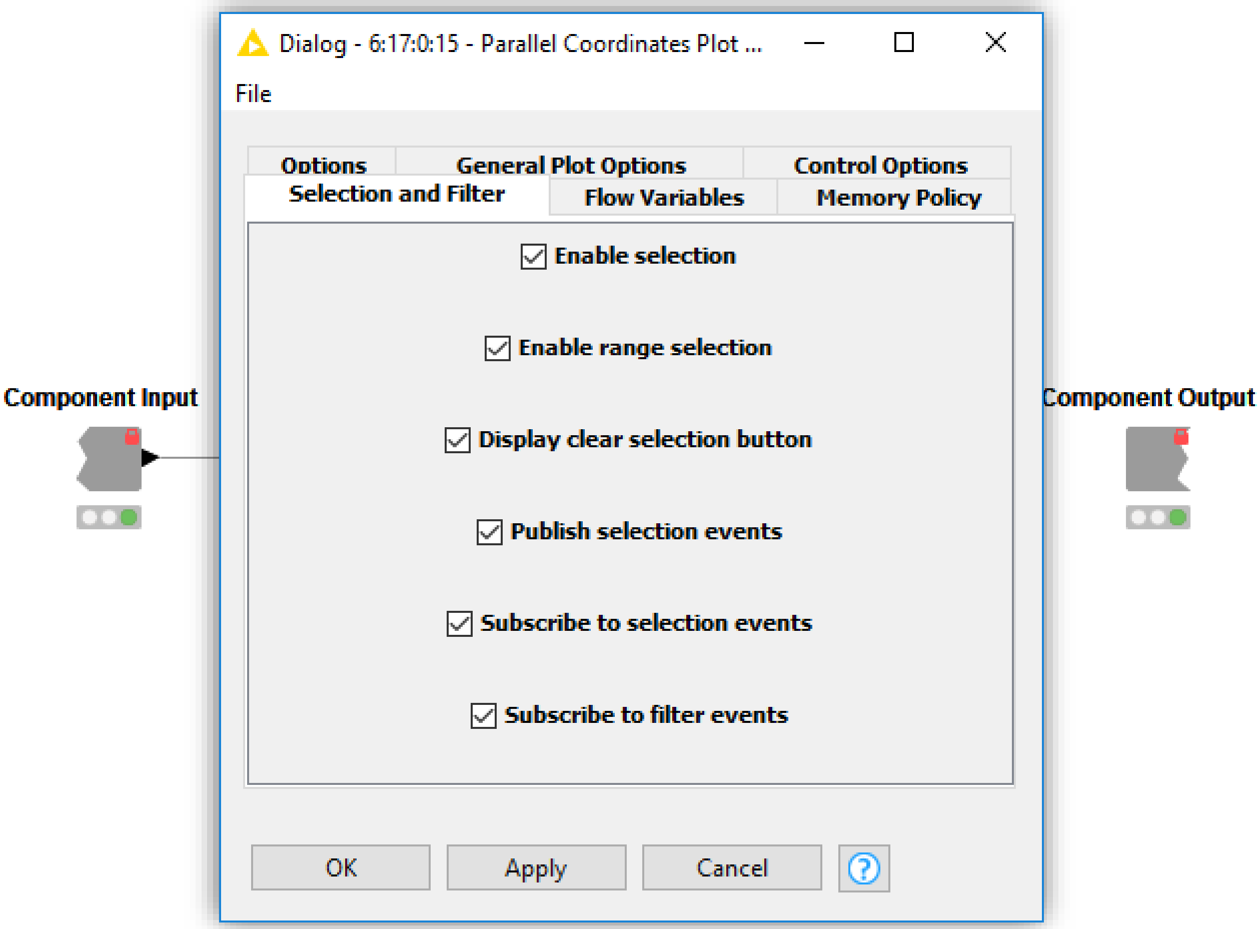
Columns

I STATIONS_ID
S CITY
S COUNTRY
D LAT
D LNG
S POPULATION
I DATE
I QUALITY_LEVEL
D AIR_TEMP
D STEAM_PRESSURE
D CLOUD_COVER
D AIR_PRESSURE_STATION_ALTITUDE
D RELATIVE_HUMIDITY
D WIND_SPEED
D AIR_TEMP_MAX
D AIR_TEMP_MIN

Interactivity across Charts: Selection and Filter Events



Interactivity across Charts: Selection and Filter Events



☑ Enable selection

☑ Enable range selection

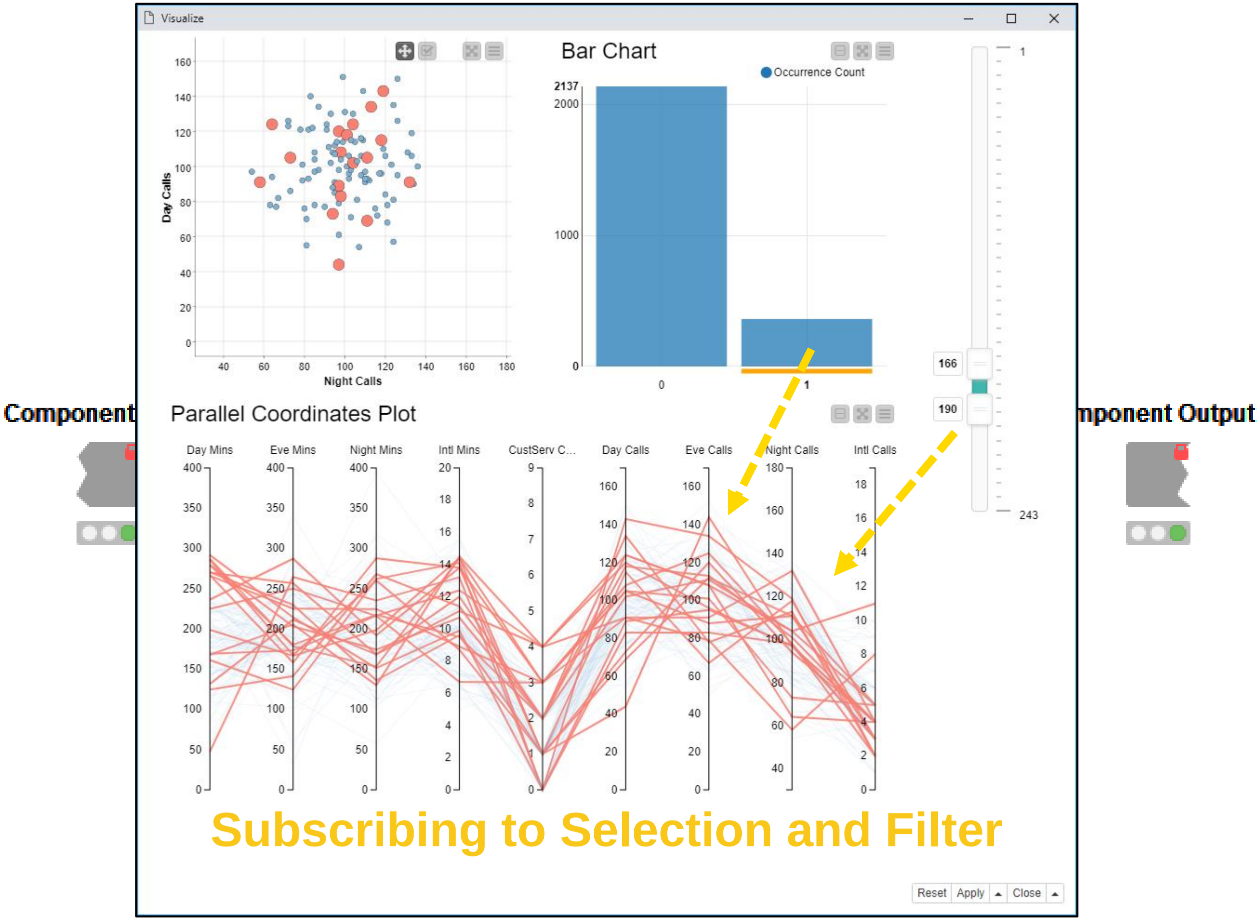
☑ Display clear selection button

☑ Publish selection events

☑ Subscribe to selection events

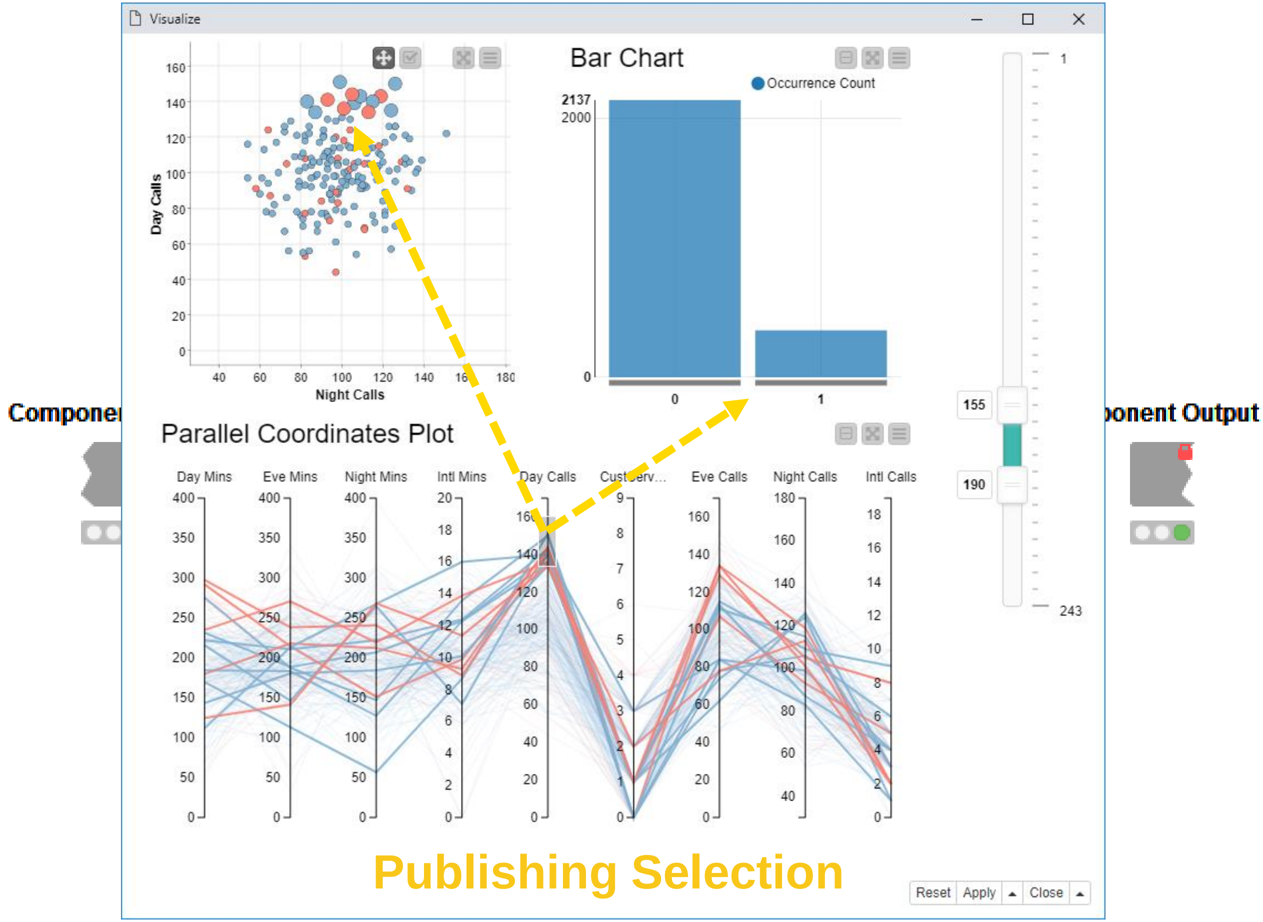
☑ Subscribe to filter events

Interactivity across Charts: Selection and Filter Events



Subscribing to Selection and Filter

Interactivity across Charts: Selection and Filter Events

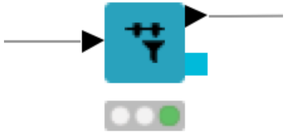


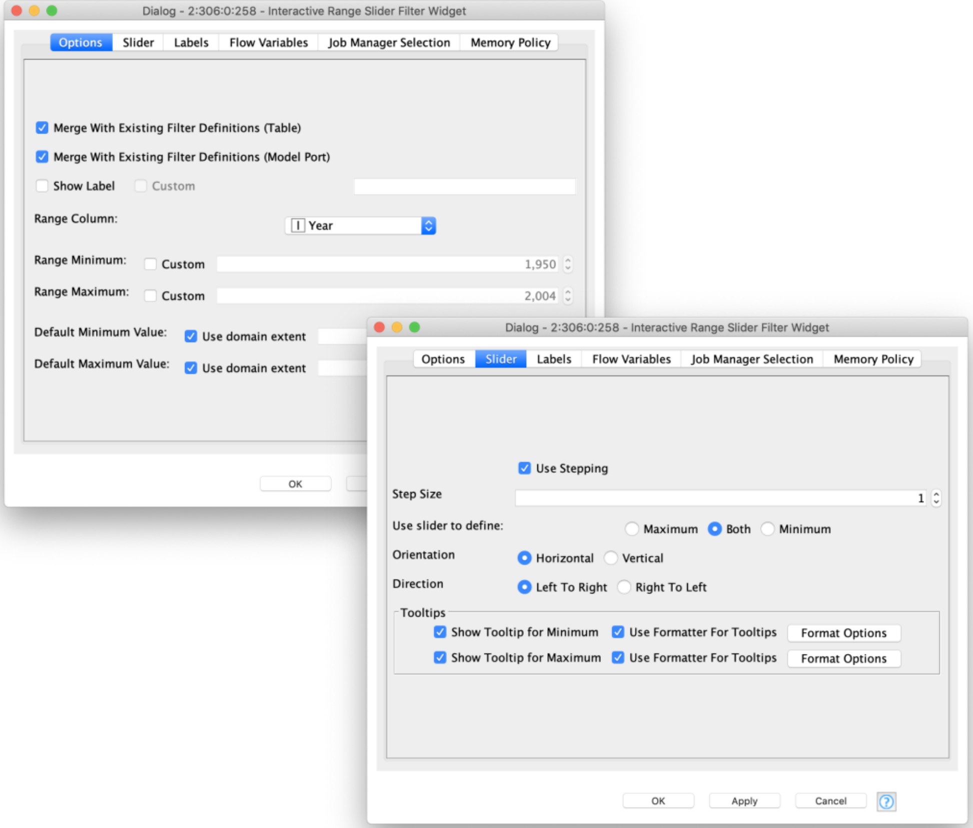
Publishing Selection

Interactive Range Slider Filter Widget

- Slider which can be used to trigger interactive filter events in the view of a component

Interactive Range Slider Filter Widget





Dialog - 2:306:0:258 - Interactive Range Slider Filter Widget

☑ Merge With Existing Filter Definitions (Table)

☑ Merge With Existing Filter Definitions (Model Port)

☐ Show Label ☐ Custom

Range Column: Year

Range Minimum: ☐ Custom 1,950

Range Maximum: ☐ Custom 2,004

Default Minimum Value: ☑ Use domain extent

Default Maximum Value: ☑ Use domain extent

Dialog - 2:306:0:258 - Interactive Range Slider Filter Widget

☑ Use Stepping

Step Size 1

Use slider to define: ○ Maximum ● Both ○ Minimum

Orientation ● Horizontal ○ Vertical

Direction ● Left To Right ○ Right To Left

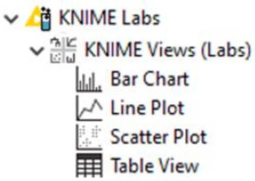
Tooltips

☑ Show Tooltip for Minimum ☑ Use Formatter For Tooltips Format Options

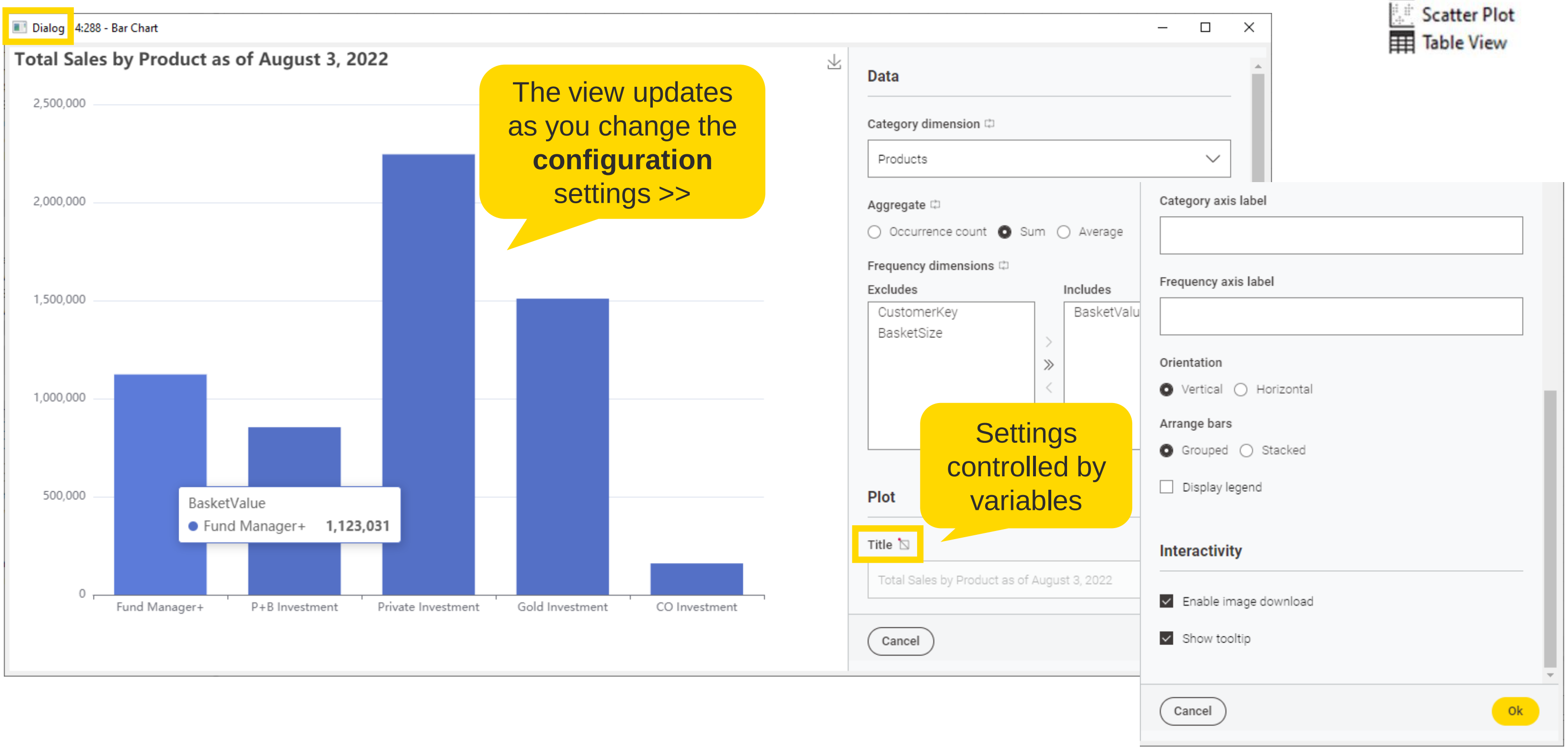
☑ Show Tooltip for Maximum ☑ Use Formatter For Tooltips Format Options

New Visualization Nodes in KNIME (Labs)

- Brand new configuration dialog (available with KNIME 4.6)
  - Explore the visualization as you change the configuration settings



- KNIME Labs
  - KNIME Views (Labs)
    - Bar Chart
    - Line Plot
    - Scatter Plot
    - Table View



Dialog 4:288 - Bar Chart

Total Sales by Product as of August 3, 2022

The view updates as you change the **configuration** settings >>

Data

Category dimension: Products

Aggregate: ○ Occurrence count ● Sum ○ Average

Frequency dimensions

Excludes	Includes
CustomerKey BasketSize	BasketValu

Plot

Title: Total Sales by Product as of August 3, 2022

Settings controlled by variables

Category axis label:

Frequency axis label:

Orientation: ● Vertical ○ Horizontal

Arrange bars: ● Grouped ○ Stacked

□ Display legend

Interactivity

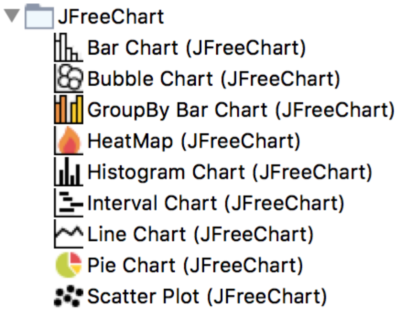
☑ Enable image download

☑ Show tooltip

Legacy View Nodes: JFreeChart & KNIME Views

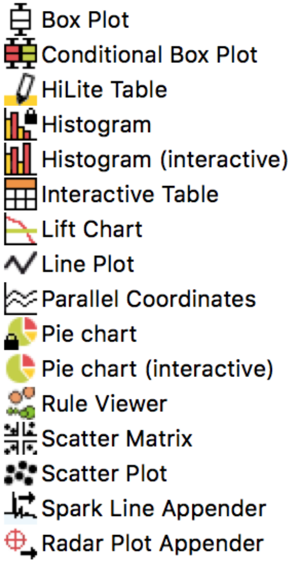
- KNIME provides three types of visualizations
  - **JavaScript Views**
  - JFreeChart
  - KNIME Views

- Active development only for JavaScript Views -> use those!
- JFreeChart and KNIME Views still useful until all plot types are implemented in JS (we're on it)



- JFreeChart
  - Bar Chart (JFreeChart)
  - Bubble Chart (JFreeChart)
  - GroupBy Bar Chart (JFreeChart)
  - HeatMap (JFreeChart)
  - Histogram Chart (JFreeChart)
  - Interval Chart (JFreeChart)
  - Line Chart (JFreeChart)
  - Pie Chart (JFreeChart)
  - Scatter Plot (JFreeChart)

JFreeChart

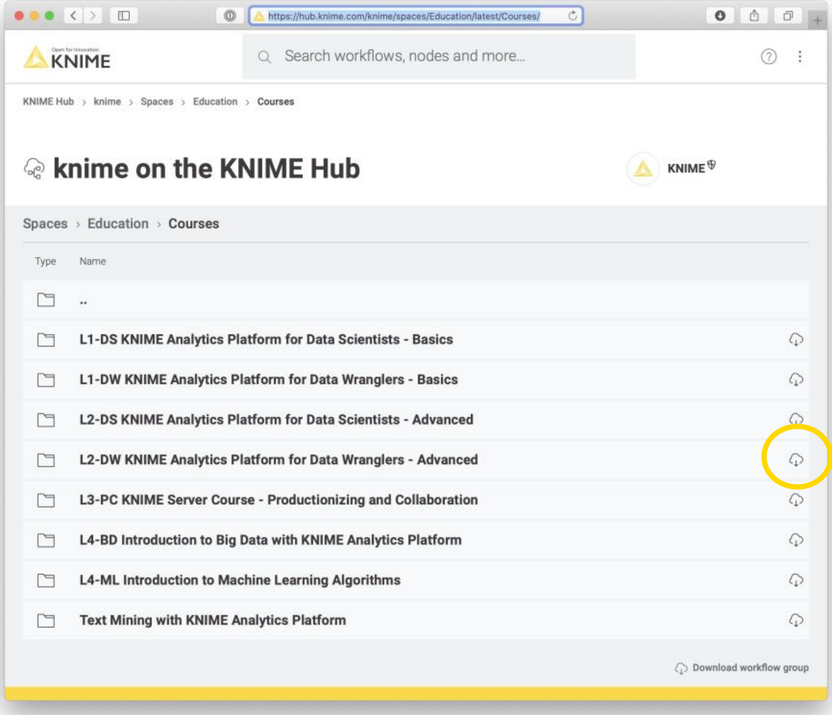


- Box Plot
- Conditional Box Plot
- HiLite Table
- Histogram
- Histogram (interactive)
- Interactive Table
- Lift Chart
- Line Plot
- Parallel Coordinates
- Pie chart
- Pie chart (interactive)
- Rule Viewer
- Scatter Matrix
- Scatter Plot
- Spark Line Appender
- Radar Plot Appender

KNIME Views

Downloading the Exercises

- Download the course material from the KNIME Hub
  https://hub.knime.com/knime/spaces/Education/latest/Courses/



https://hub.knime.com/knime/spaces/Education/latest/Courses/

Search workflows, nodes and more...

KNIME Hub › knime › Spaces › Education › Courses

knime on the KNIME Hub

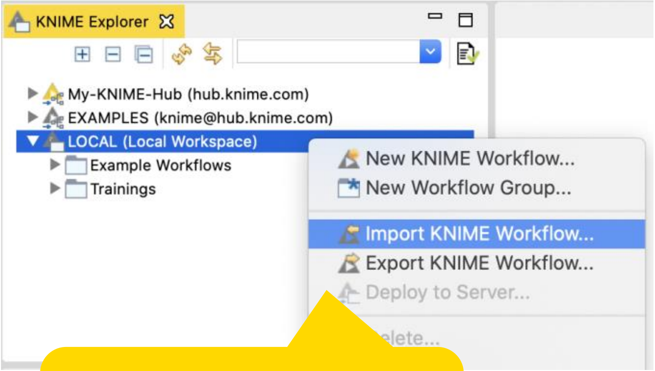
Spaces › Education › Courses

Type	Name
	..
	L1-DS KNIME Analytics Platform for Data Scientists - Basics
	L1-DW KNIME Analytics Platform for Data Wranglers - Basics
	L2-DS KNIME Analytics Platform for Data Scientists - Advanced
	L2-DW KNIME Analytics Platform for Data Wranglers - Advanced
	L3-PC KNIME Server Course - Productionizing and Collaboration
	L4-BD Introduction to Big Data with KNIME Analytics Platform
	L4-ML Introduction to Machine Learning Algorithms
	Text Mining with KNIME Analytics Platform

Download workflow group

Importing the Exercises

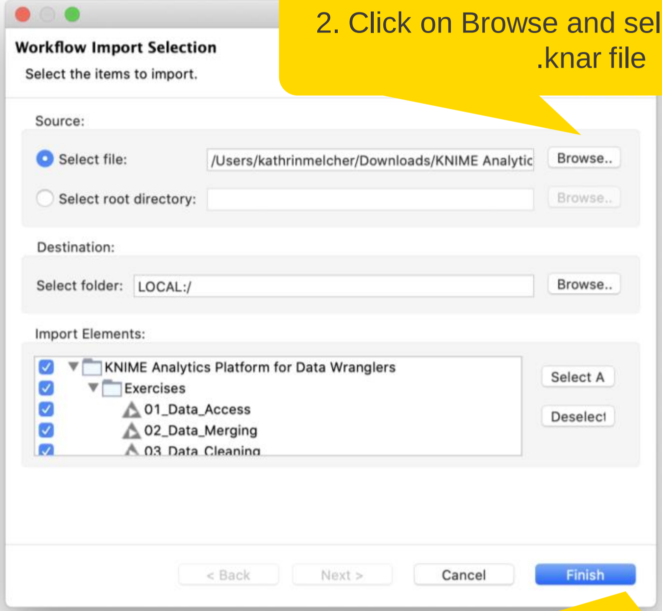
- Import the course material to KNIME Analytics Platform



KNIME Explorer

- My-KNIME-Hub (hub.knime.com)
- EXAMPLES (knime@hub.knime.com)
- LOCAL (Local Workspace)
  - Example Workflows
  - Trainings

1. Right click on LOCAL and select Import KNIME Workflow….



Workflow Import Selection

Select the items to import.

Source:

● Select file: /Users/kathrinmelcher/Downloads/KNIME Analytic [Browse..]

○ Select root directory: [Browse..]

Destination:

Select folder: LOCAL:/ [Browse..]

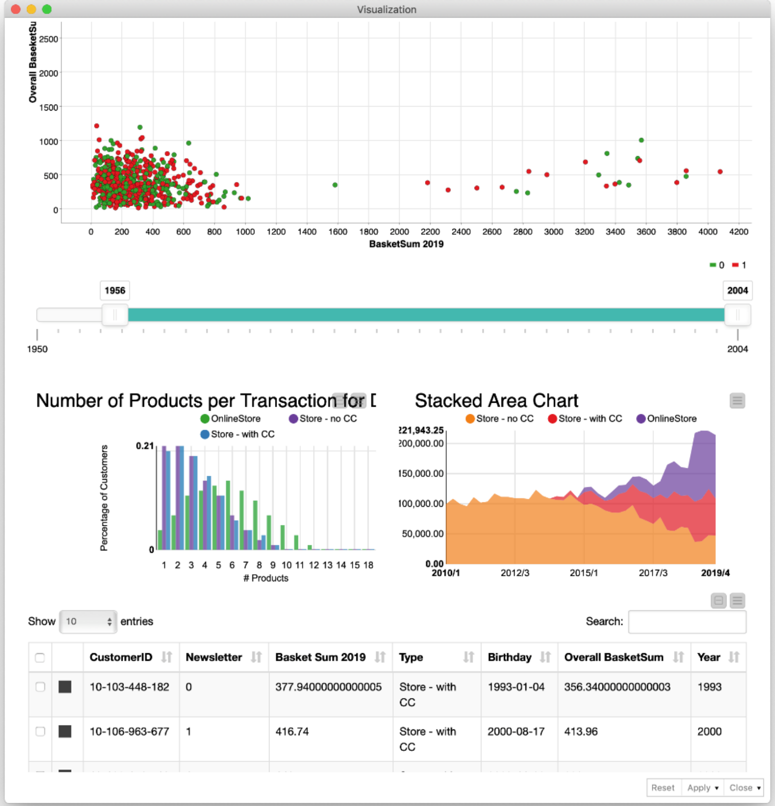
Import Elements:

- ☑ KNIME Analytics Platform for Data Wranglers
  - ☑ Exercises
    - ☑ 01_Data_Access
    - ☑ 02_Data_Merging
    - ☑ 03_Data_Cleaning

2. Click on Browse and select downloaded .knar file

3. Click on Finish

Exercise: 08_Visualization – Goal



Show 10 entries

Search:

□		CustomerID	Newsletter	Basket Sum 2019	Type	Birthday	Overall BasketSum	Year
□	■	10-103-448-182	0	377.94000000000005	Store - with CC	1993-01-04	356.34000000000003	1993
□	■	10-106-963-677	1	416.74	Store - with CC	2000-08-17	413.96	2000

Exercise: 08_Visualization

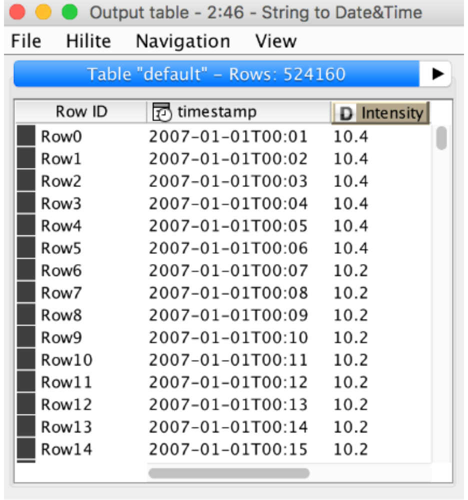
- Create a scatter plot to show the relationship between the total purchase amount in the year 2019 and in the years before
- Visualize the customer data in an interactive table
- Create a stacked area chart to show the development of the total purchase amount over time for each transaction type
- Create a composite view and define the layout

- Optional tasks:
  - Add a range slider to filter the scatter plot by age (optional)
  - Build a bar chart to show the number of products per order for the different transaction types (optional)
  - Change the color for the different transaction types (optional)

Date/Time Data

Date & Time Overview

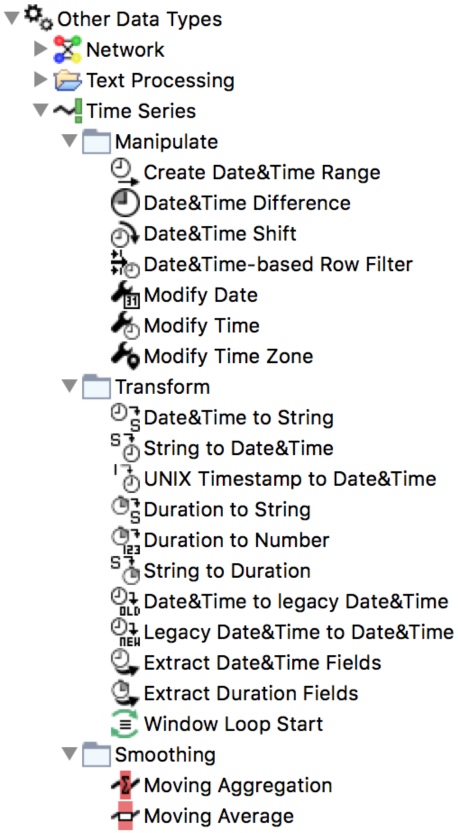
- Dedicated data type for date and time data
- Supported in Date&Time nodes
  - (and others: GroupBy, Pivot, Line Plot)
- Complete re-write in KNIME 3.4



Output table - 2:46 - String to Date&Time

Table "default" – Rows: 524160

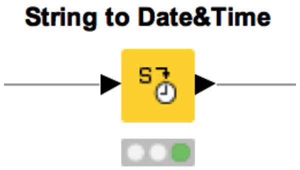
Row ID	timestamp	Intensity
Row0	2007-01-01T00:01	10.4
Row1	2007-01-01T00:02	10.4
Row2	2007-01-01T00:03	10.4
Row3	2007-01-01T00:04	10.4
Row4	2007-01-01T00:05	10.4
Row5	2007-01-01T00:06	10.4
Row6	2007-01-01T00:07	10.2
Row7	2007-01-01T00:08	10.2
Row8	2007-01-01T00:09	10.2
Row9	2007-01-01T00:10	10.2
Row10	2007-01-01T00:11	10.2
Row11	2007-01-01T00:12	10.2
Row12	2007-01-01T00:13	10.2
Row13	2007-01-01T00:14	10.2
Row14	2007-01-01T00:15	10.2

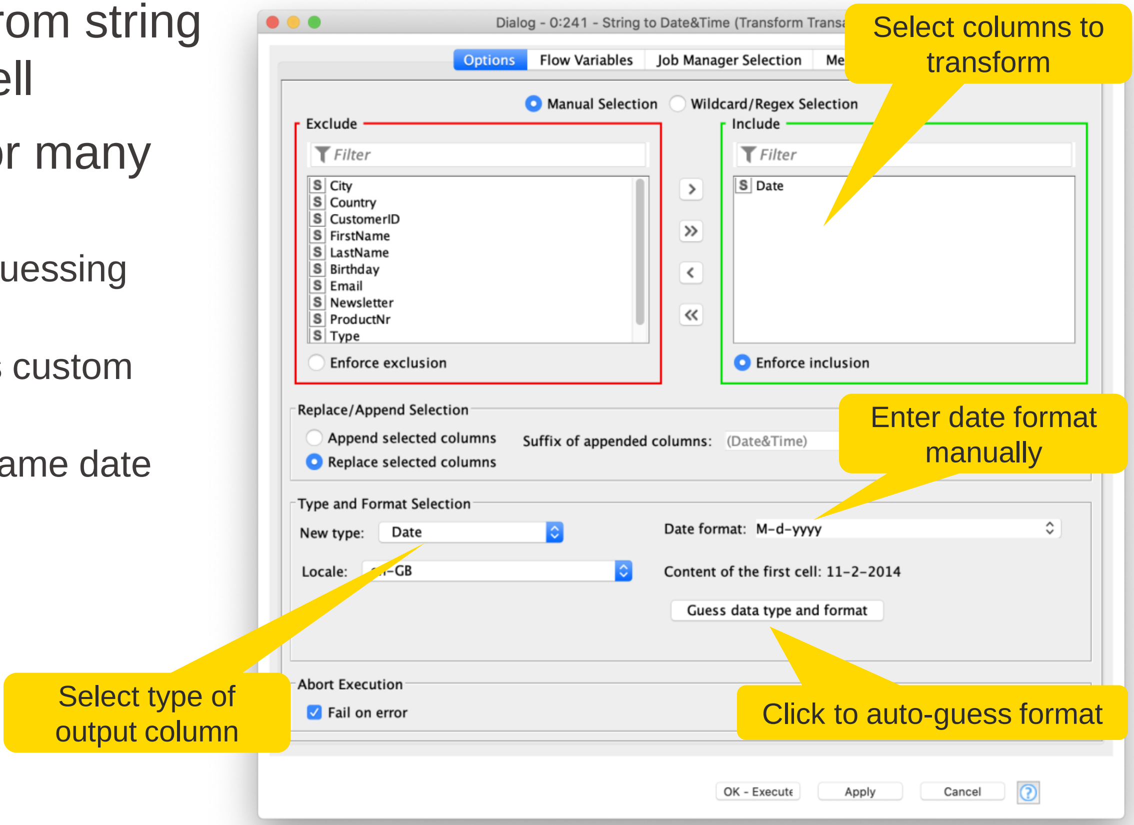


- Other Data Types
  - Network
  - Text Processing
  - Time Series
    - Manipulate
      - Create Date&Time Range
      - Date&Time Difference
      - Date&Time Shift
      - Date&Time-based Row Filter
      - Modify Date
      - Modify Time
      - Modify Time Zone
    - Transform
      - Date&Time to String
      - String to Date&Time
      - UNIX Timestamp to Date&Time
      - Duration to String
      - Duration to Number
      - String to Duration
      - Date&Time to legacy Date&Time
      - Legacy Date&Time to Date&Time
      - Extract Date&Time Fields
      - Extract Duration Fields
      - Window Loop Start
    - Smoothing
      - Moving Aggregation
      - Moving Average

String to Date&Time

- Converts date/time data from string into a native Date&time cell
- Guesses correct format for many types of date formatting
  - Enter format manually if auto-guessing didn‘t work
    - KNIME automatically adds custom formats to auto-guess list
  - Converts multiple columns of same date format in one node




Dialog - 0:241 - String to Date&Time (Transform Trans

Options | Flow Variables | Job Manager Selection | Me

● Manual Selection ○ Wildcard/Regex Selection

Exclude: City, Country, CustomerID, FirstName, LastName, Birthday, Email, Newsletter, ProductNr, Type

○ Enforce exclusion

Include: Date

● Enforce inclusion

Replace/Append Selection

○ Append selected columns

● Replace selected columns

Suffix of appended columns: (Date&Time)

Type and Format Selection

New type: Date

Locale: en-GB

Date format: M-d-yyyy

Content of the first cell: 11-2-2014

Guess data type and format

Abort Execution

☑ Fail on error

OK - Execute | Apply | Cancel

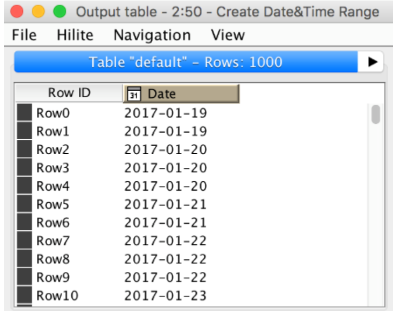
Select columns to transform

Enter date format manually

Select type of output column

Click to auto-guess format

Date&Time – Data Types



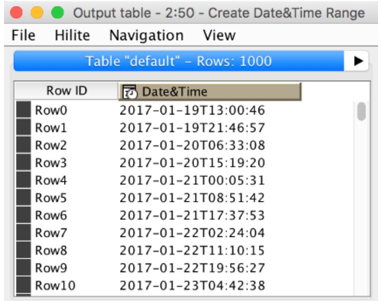
Output table - 2:50 - Create Date&Time Range

File Hilite Navigation View

Table "default" – Rows: 1000

Row ID	Date
Row0	2017-01-19
Row1	2017-01-19
Row2	2017-01-20
Row3	2017-01-20
Row4	2017-01-20
Row5	2017-01-21
Row6	2017-01-21
Row7	2017-01-22
Row8	2017-01-22
Row9	2017-01-22
Row10	2017-01-23

Date



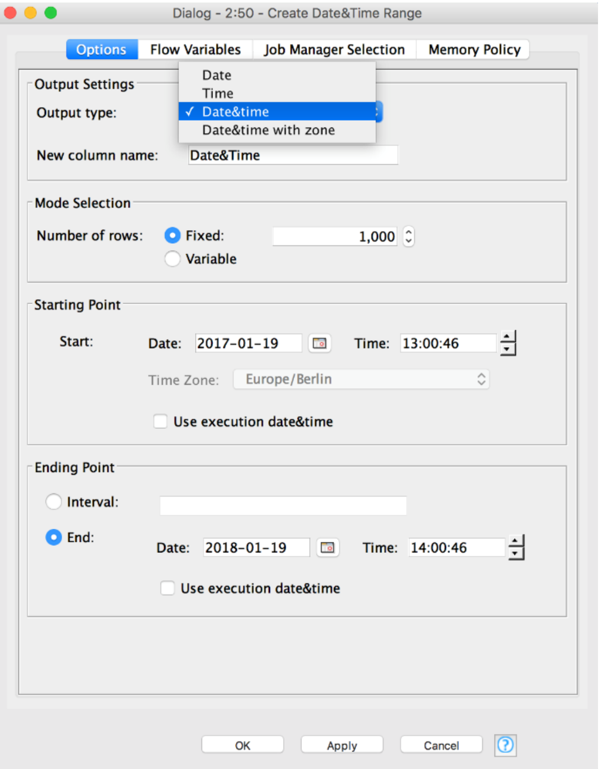
Output table - 2:50 - Create Date&Time Range

File Hilite Navigation View

Table "default" – Rows: 1000

Row ID	Date&Time
Row0	2017-01-19T13:00:46
Row1	2017-01-19T21:46:57
Row2	2017-01-20T06:33:08
Row3	2017-01-20T15:19:20
Row4	2017-01-21T00:05:31
Row5	2017-01-21T08:51:42
Row6	2017-01-21T17:37:53
Row7	2017-01-22T02:24:04
Row8	2017-01-22T11:10:15
Row9	2017-01-22T19:56:27
Row10	2017-01-23T04:42:38

Date & Time



Dialog - 2:50 - Create Date&Time Range

Options | Flow Variables | Job Manager Selection | Memory Policy

Output Settings

Output type: Date / Time / ✓ Date&time / Date&time with zone

New column name: Date&Time

Mode Selection

Number of rows: ● Fixed: 1,000 ○ Variable

Starting Point

Start: Date: 2017-01-19 Time: 13:00:46

Time Zone: Europe/Berlin

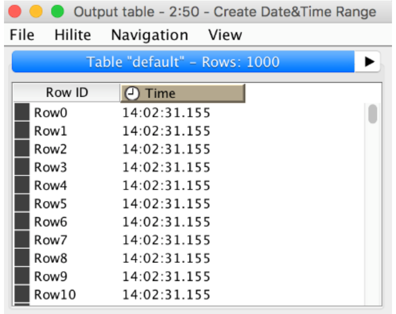
□ Use execution date&time

Ending Point

○ Interval:

● End: Date: 2018-01-19 Time: 14:00:46

□ Use execution date&time



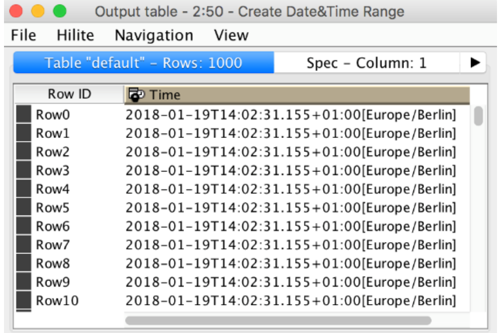
Output table - 2:50 - Create Date&Time Range

File Hilite Navigation View

Table "default" – Rows: 1000

Row ID	Time
Row0	14:02:31.155
Row1	14:02:31.155
Row2	14:02:31.155
Row3	14:02:31.155
Row4	14:02:31.155
Row5	14:02:31.155
Row6	14:02:31.155
Row7	14:02:31.155
Row8	14:02:31.155
Row9	14:02:31.155
Row10	14:02:31.155

Time



Output table - 2:50 - Create Date&Time Range

File Hilite Navigation View

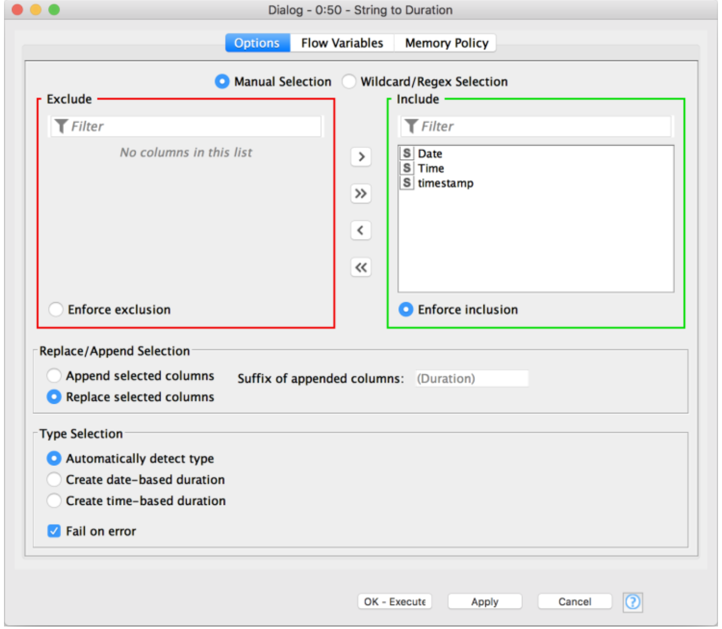
Table "default" – Rows: 1000 | Spec – Column: 1

Row ID	Time
Row0	2018-01-19T14:02:31.155+01:00[Europe/Berlin]
Row1	2018-01-19T14:02:31.155+01:00[Europe/Berlin]
Row2	2018-01-19T14:02:31.155+01:00[Europe/Berlin]
Row3	2018-01-19T14:02:31.155+01:00[Europe/Berlin]
Row4	2018-01-19T14:02:31.155+01:00[Europe/Berlin]
Row5	2018-01-19T14:02:31.155+01:00[Europe/Berlin]
Row6	2018-01-19T14:02:31.155+01:00[Europe/Berlin]
Row7	2018-01-19T14:02:31.155+01:00[Europe/Berlin]
Row8	2018-01-19T14:02:31.155+01:00[Europe/Berlin]
Row9	2018-01-19T14:02:31.155+01:00[Europe/Berlin]
Row10	2018-01-19T14:02:31.155+01:00[Europe/Berlin]

Date & Time +
Time zone

String to Duration

- Takes a string and converts it to a duration cell
  - Three different options to format input strings
- Example: Convert 1 year, 2 months, 3 weeks, and 4 days to duration cell
  - ISO-8601: “P1Y2M3W4D“
  - Short letter: “1y 2M 3w 4d”
  - Long word: “1 year 2 months 3 weeks 4 days”



● Manual Selection ○ Wildcard/Regex Selection

Include: S Date, S Time, S timestamp

○ Enforce exclusion

● Enforce inclusion

Replace/Append Selection

○ Append selected columns — Suffix of appended columns: (Duration)

● Replace selected columns

Type Selection

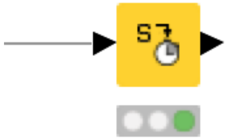
● Automatically detect type

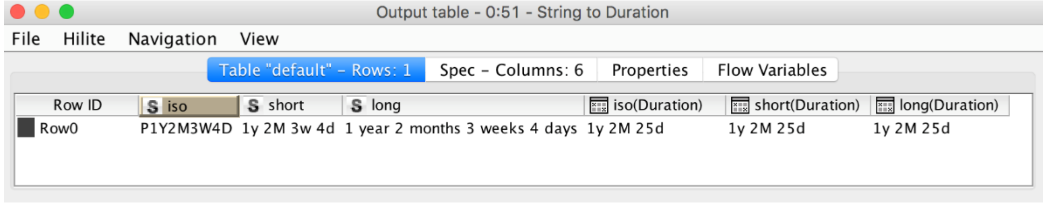
○ Create date-based duration

○ Create time-based duration

☑ Fail on error

String to Duration



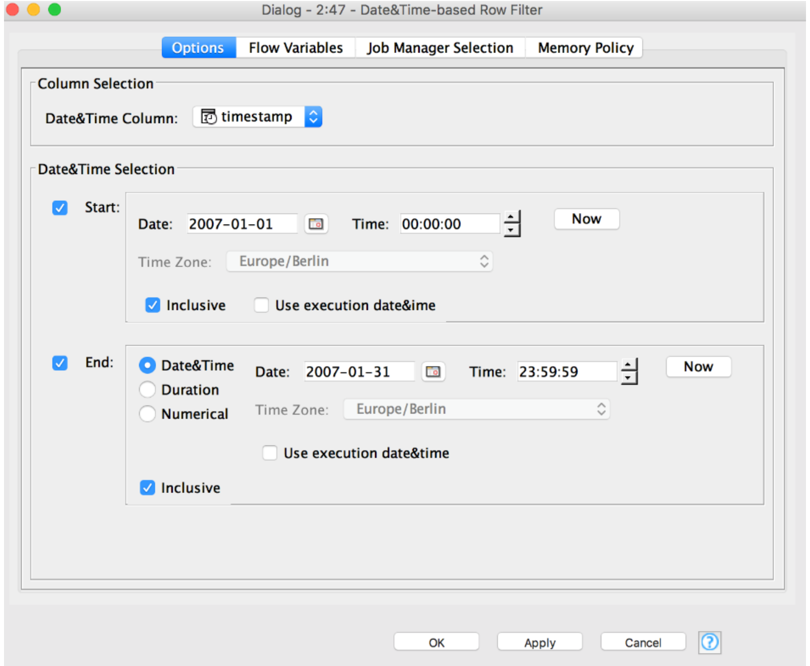


Output table - 0:51 - String to Duration
File
Hilite
Navigation
View
Table "default" – Rows: 1
Spec – Columns: 6
Properties
Flow Variables

Row ID	S iso	S short	S long	iso(Duration)	short(Duration)	long(Duration)
Row0	P1Y2M3W4D	1y 2M 3w 4d	1 year 2 months 3 weeks 4 days	1y 2M 25d	1y 2M 25d	1y 2M 25d

Date&Time-based Row Filter

- Filters rows from a specified time period
- Range can be limited on upper bound, lower bound or both
- Options for end point:
  - Date&Time: Fixed data and time
  - Duration: Duration string (e.g. 2y 3M)
  - Numerical: Select granularity from dropdown and enter number



Dialog - 2:47 - Date&Time-based Row Filter

Column Selection

Date&Time Column: timestamp

Date&Time Selection

☑ Start: Date: 2007-01-01 Time: 00:00:00

Time Zone: Europe/Berlin

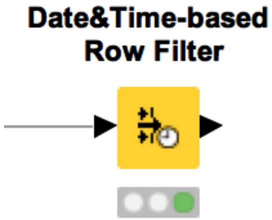
☑ Inclusive ☐ Use execution date&ime

☑ End: ● Date&Time ○ Duration ○ Numerical Date: 2007-01-31 Time: 23:59:59

Time Zone: Europe/Berlin

☐ Use execution date&time

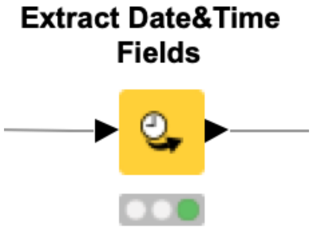
☑ Inclusive

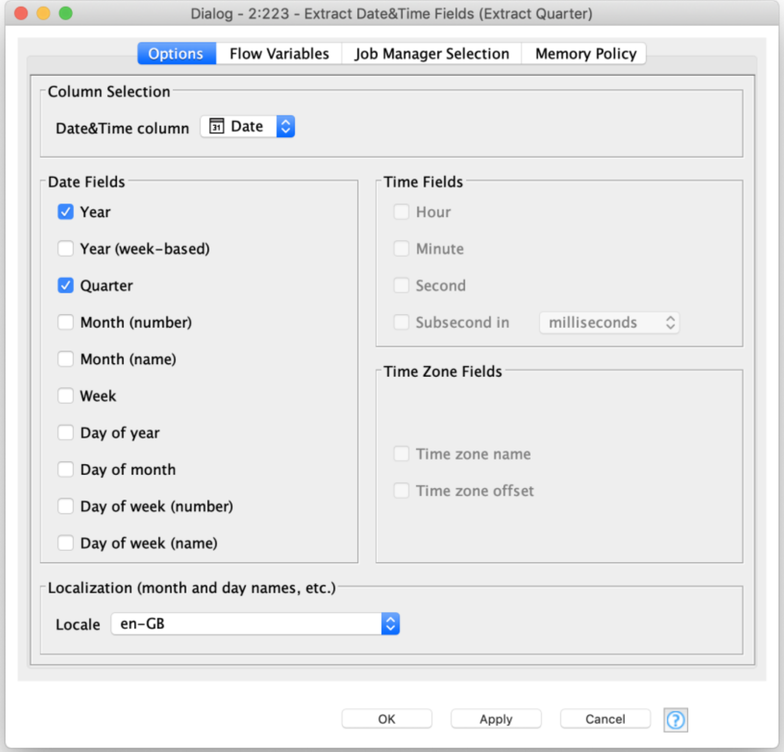


Date&Time-based Row Filter

Extract Date&Time Fields

- Extracts date fields (year, day, month) or time fields (hour, minute, second) from a date&time cell.
- You can pick and choose which fields to include
- Useful when used in combination with data aggregation nodes (groupby, pivot etc.)




Dialog - 2:223 - Extract Date&Time Fields (Extract Quarter)

Options | Flow Variables | Job Manager Selection | Memory Policy

Column Selection

Date&Time column: Date

Date Fields

- ☑ Year
- ☐ Year (week-based)
- ☑ Quarter
- ☐ Month (number)
- ☐ Month (name)
- ☐ Week
- ☐ Day of year
- ☐ Day of month
- ☐ Day of week (number)
- ☐ Day of week (name)

Time Fields

- ☐ Hour
- ☐ Minute
- ☐ Second
- ☐ Subsecond in milliseconds

Time Zone Fields

- ☐ Time zone name
- ☐ Time zone offset

Localization (month and day names, etc.)

Locale: en-GB

Exercise: 06_DateTime_Manipulation

- Convert order dates from string to Date&Time
- Extract the product purchases that were submitted in 2019
- Extract the remaining product purchases into a separate table
- Extract quarter and year of each product purchase into separate columns

Data Export & Reporting

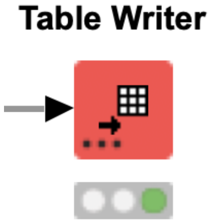
Exporting Data

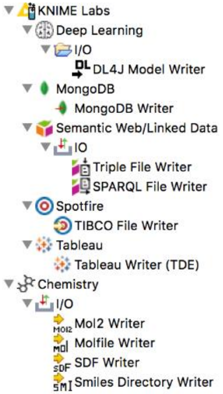
After an analysis is completed, what next?

- Write results to a file
- Upload results to a Cloud Storage
- Create/update a database
- Generate a rich report using BIRT
- Send your data to Tableau, Spotfire, PowerBI to create a report
- Deploy via KNIME WebPortal
- Deploy your model as RESTful web service

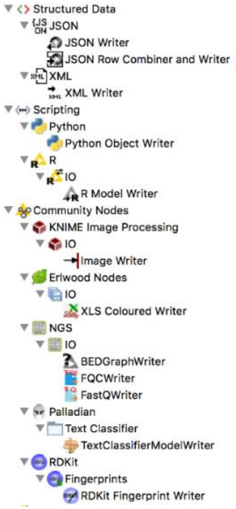
Data Export Nodes

- Typically characterized by:
  - Magenta color
  - 1 input port, no output ports
  - Create file on file system or write to database

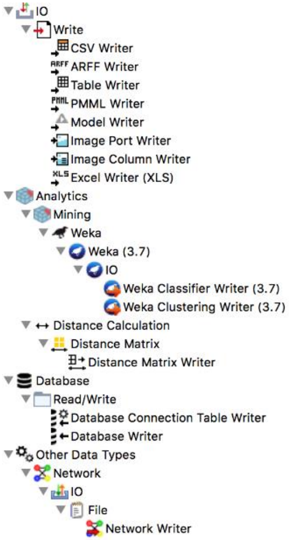




- KNIME Labs
  - Deep Learning
    - I/O
      - DL4J Model Writer
  - MongoDB
    - MongoDB Writer
  - Semantic Web/Linked Data
    - IO
      - Triple File Writer
      - SPARQL File Writer
  - Spotfire
    - TIBCO File Writer
  - Tableau
    - Tableau Writer (TDE)
- Chemistry
  - I/O
    - Mol2 Writer
    - Molfile Writer
    - SDF Writer
    - Smiles Directory Writer



- Structured Data
  - JSON
    - JSON Writer
    - JSON Row Combiner and Writer
  - XML
    - XML Writer
- Scripting
  - Python
    - Python Object Writer
  - R
    - IO
      - R Model Writer
- Community Nodes
  - KNIME Image Processing
    - IO
      - Image Writer
  - Erlwood Nodes
    - IO
      - XLS Coloured Writer
  - NGS
    - IO
      - BEDGraphWriter
      - FQCWriter
      - FastQWriter
  - Palladian
    - Text Classifier
      - TextClassifierModelWriter
  - RDKit
    - Fingerprints
      - RDKit Fingerprint Writer

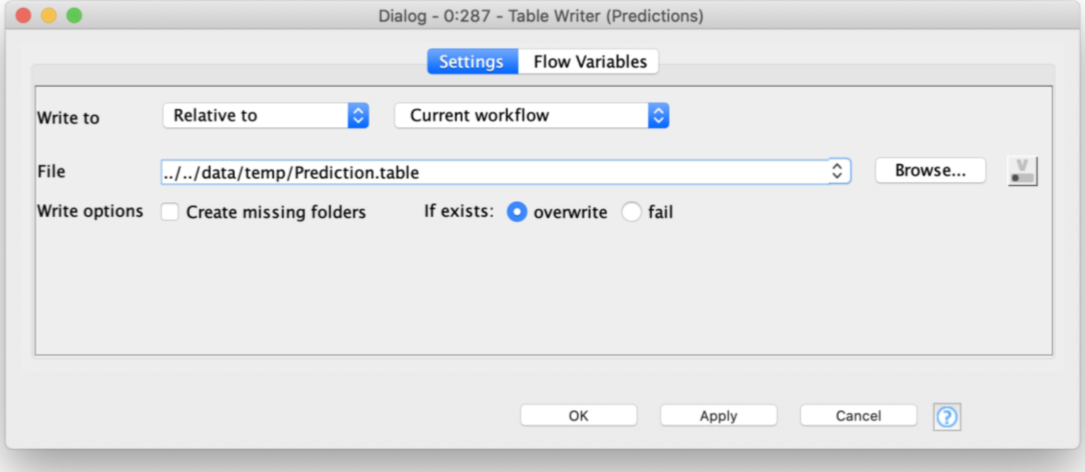


- IO
  - Write
    - CSV Writer
    - ARFF Writer
    - Table Writer
    - PMML Writer
    - Model Writer
    - Image Port Writer
    - Image Column Writer
    - Excel Writer (XLS)
- Analytics
  - Mining
    - Weka
      - Weka (3.7)
        - IO
          - Weka Classifier Writer (3.7)
          - Weka Clustering Writer (3.7)
  - Distance Calculation
    - Distance Matrix
      - Distance Matrix Writer
- Database
  - Read/Write
    - Database Connection Table Writer
    - Database Writer
- Other Data Types
  - Network
    - IO
      - File
        - Network Writer

https://www.youtube.com/watch?v=Og7VZOJhsOc&feature=youtu.be

Table Writer

Table Writer



Dialog - 0:287 - Table Writer (Predictions)

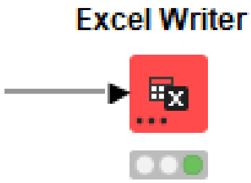
Write to: Relative to | Current workflow

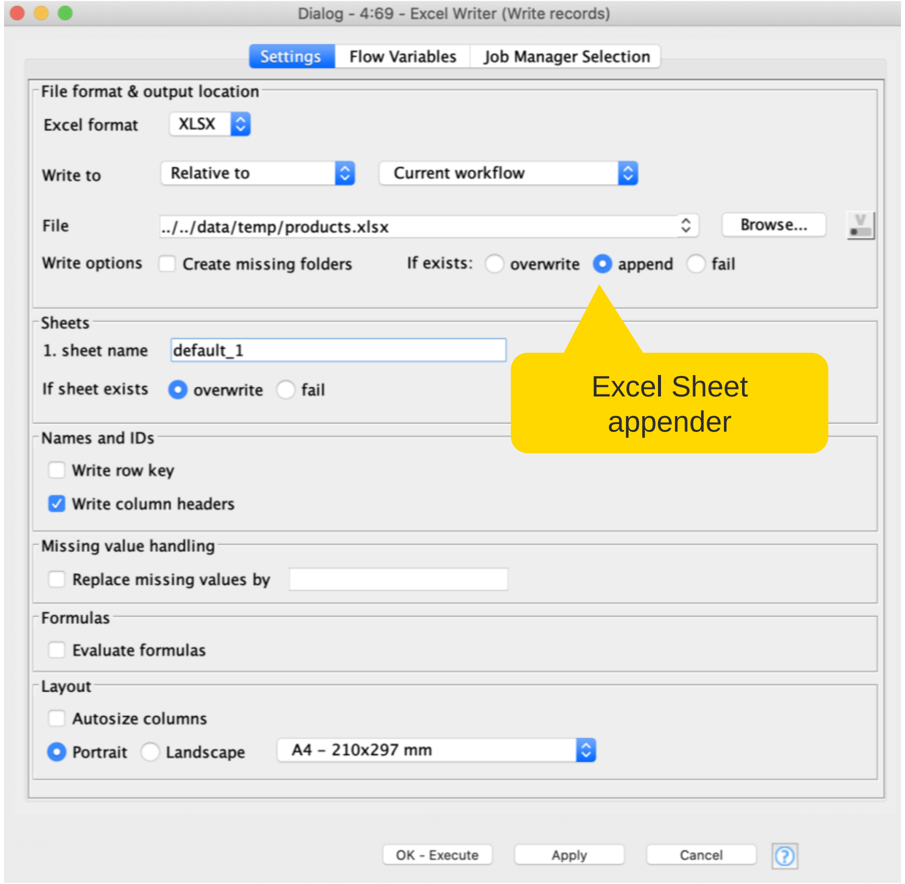
File: ../../data/temp/Prediction.table

Write options: □ Create missing folders   If exists: ● overwrite ○ fail

Excel Writer

- Writes the input table into a spreadsheet of an Excel file
- Select append, to append a spreadsheet to an existing Excel File and define the name of the new sheet





Dialog - 4:69 - Excel Writer (Write records)

File format & output location

Excel format: XLSX

Write to: Relative to | Current workflow

File: ../../data/temp/products.xlsx

Write options: □ Create missing folders   If exists: ○ overwrite ● append ○ fail

Sheets

1. sheet name: default_1

If sheet exists: ● overwrite ○ fail

Names and IDs

□ Write row key

☑ Write column headers

Missing value handling

□ Replace missing values by

Formulas

□ Evaluate formulas

Layout

□ Autosize columns

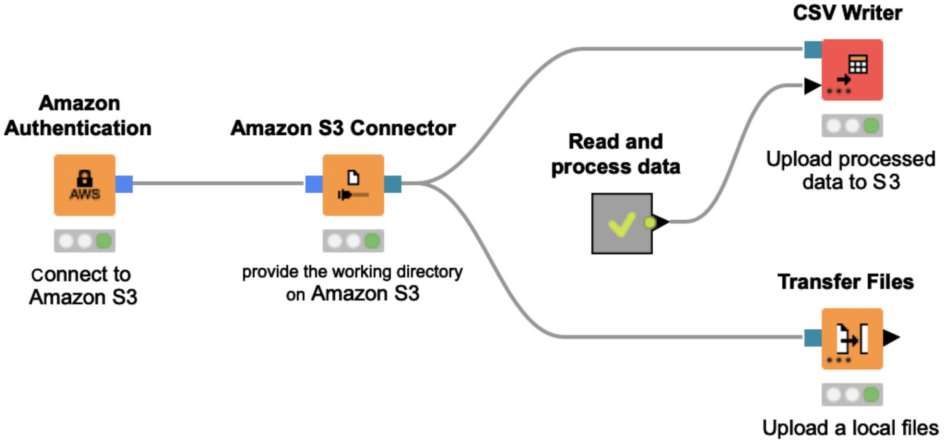
● Portrait ○ Landscape   A4 – 210x297 mm

Excel Sheet appender

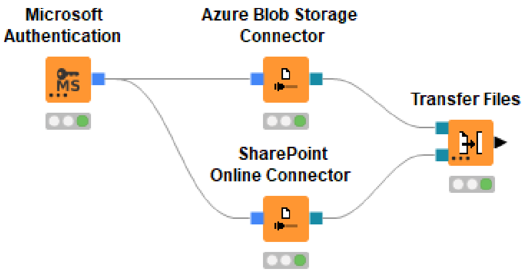
Write Files to a Remote File System

- The new file handling framework makes it easy to upload data to remote file systems
  - Write processed data directly with a writer node
  - Upload local files with the Transfer Files node

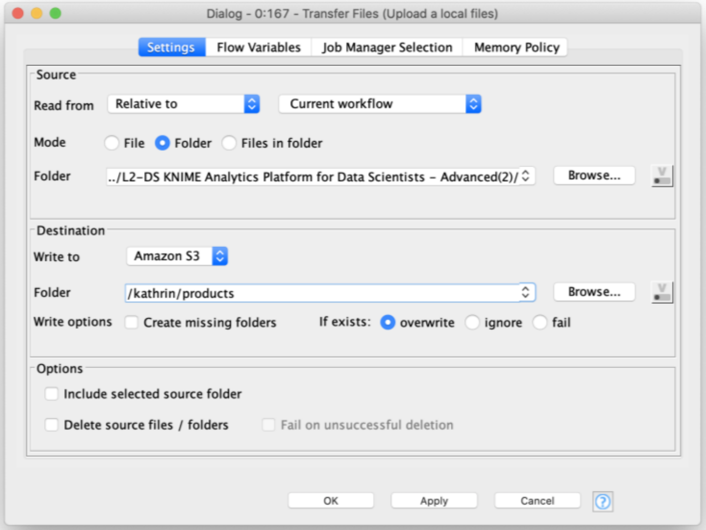
- Supported file systems
  - Microsoft Azure
  - Google
  - Amazon
  - Databricks
  - BigData file systems (hdfs, httpFS, …)
  - On-premise (e.g. ssh, ftp, …)



Full Flexibility with the Transfer Files node



Same cloud environment



Dialog - 0:167 - Transfer Files (Upload a local files)

Source

Read from: Relative to | Current workflow

Mode: ○ File ● Folder ○ Files in folder

Folder: ../L2-DS KNIME Analytics Platform for Data Scientists – Advanced(2)/

Destination

Write to: Amazon S3

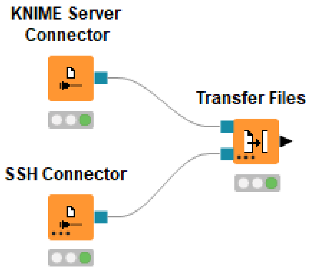
Folder: /kathrin/products

Write options: □ Create missing folders If exists: ● overwrite ○ ignore ○ fail

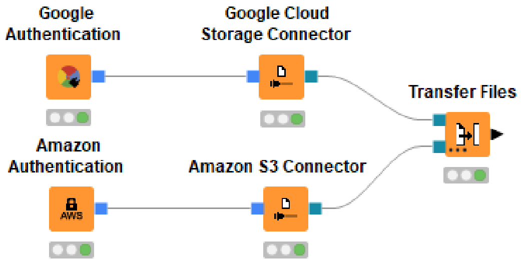
Options

□ Include selected source folder

□ Delete source files / folders □ Fail on unsuccessful deletion



On-premise

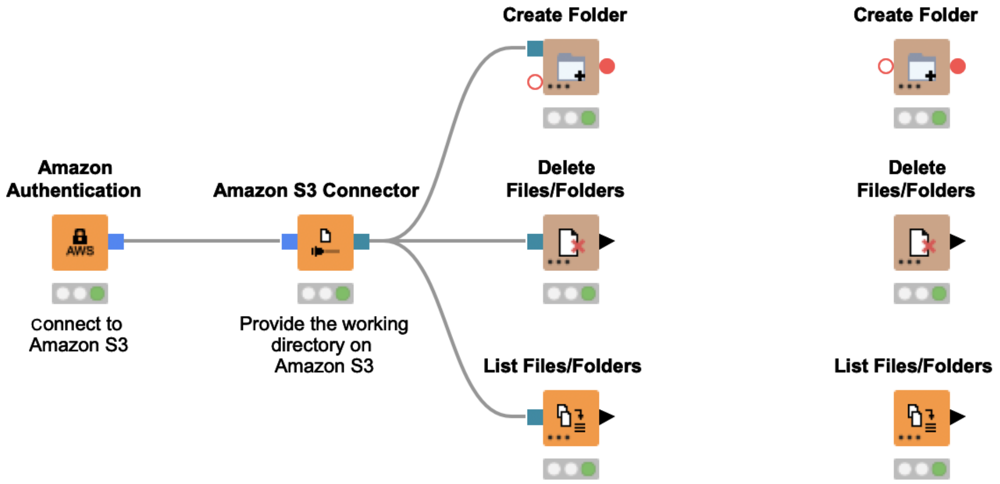


Cross cloud environments

Other Utility Nodes

Can be used with local and remote file systems

- Create a folder
- Delete files or folders
- List all files in a folder



connect to Amazon S3

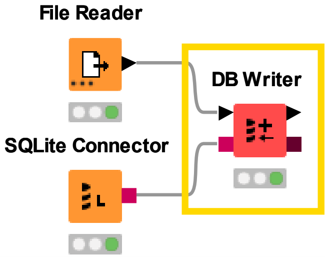
Provide the working directory on Amazon S3

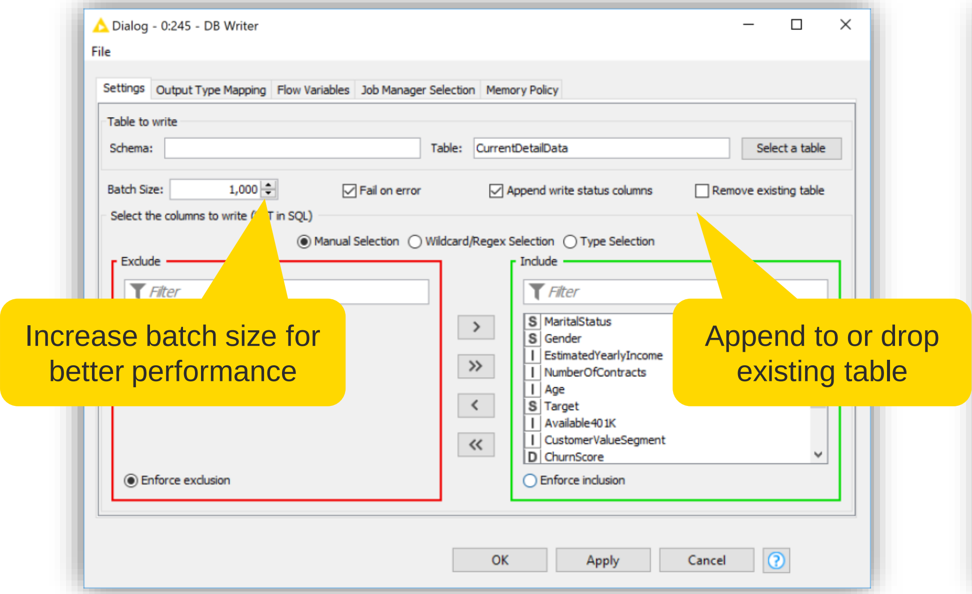
- Further information about file handling

https://docs.knime.com/latest/analytics_platform_file_handling_guide/index.html

DB Writer

- Writes data from a KNIME data table **directly** into a database table





Dialog - 0:245 - DB Writer

File

Settings | Output Type Mapping | Flow Variables | Job Manager Selection | Memory Policy

Table to write

Schema:
Table: CurrentDetailData
Select a table

Batch Size: 1,000
☑ Fail on error
☑ Append write status columns
□ Remove existing table

Select the columns to write (T in SQL)

● Manual Selection ○ Wildcard/Regex Selection ○ Type Selection

Exclude

Filter

● Enforce exclusion

Include

Filter

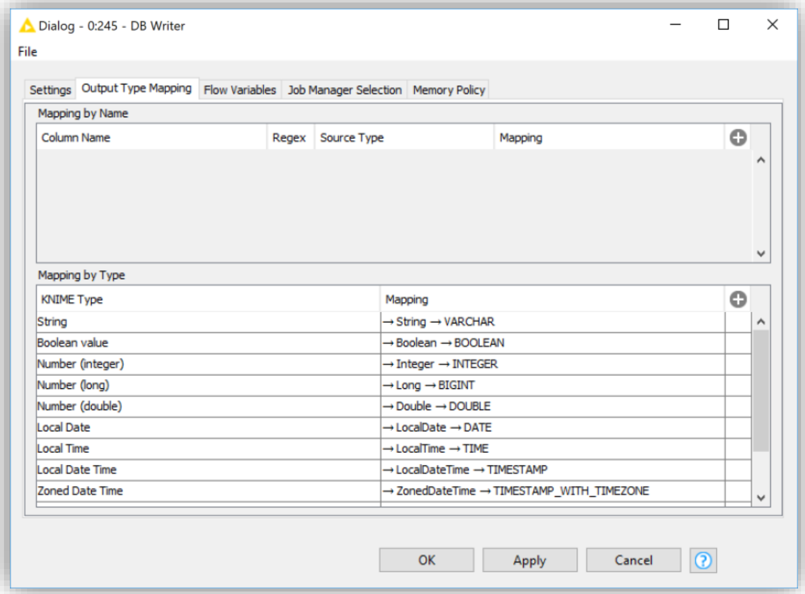
S	MaritalStatus
S	Gender
I	EstimatedYearlyIncome
I	NumberOfContracts
I	Age
S	Target
I	Available401K
S	CustomerValueSegment
D	ChurnScore

○ Enforce inclusion

OK | Apply | Cancel

Increase batch size for better performance

Append to or drop existing table



Dialog - 0:245 - DB Writer

File

Settings | Output Type Mapping | Flow Variables | Job Manager Selection | Memory Policy

Mapping by Name

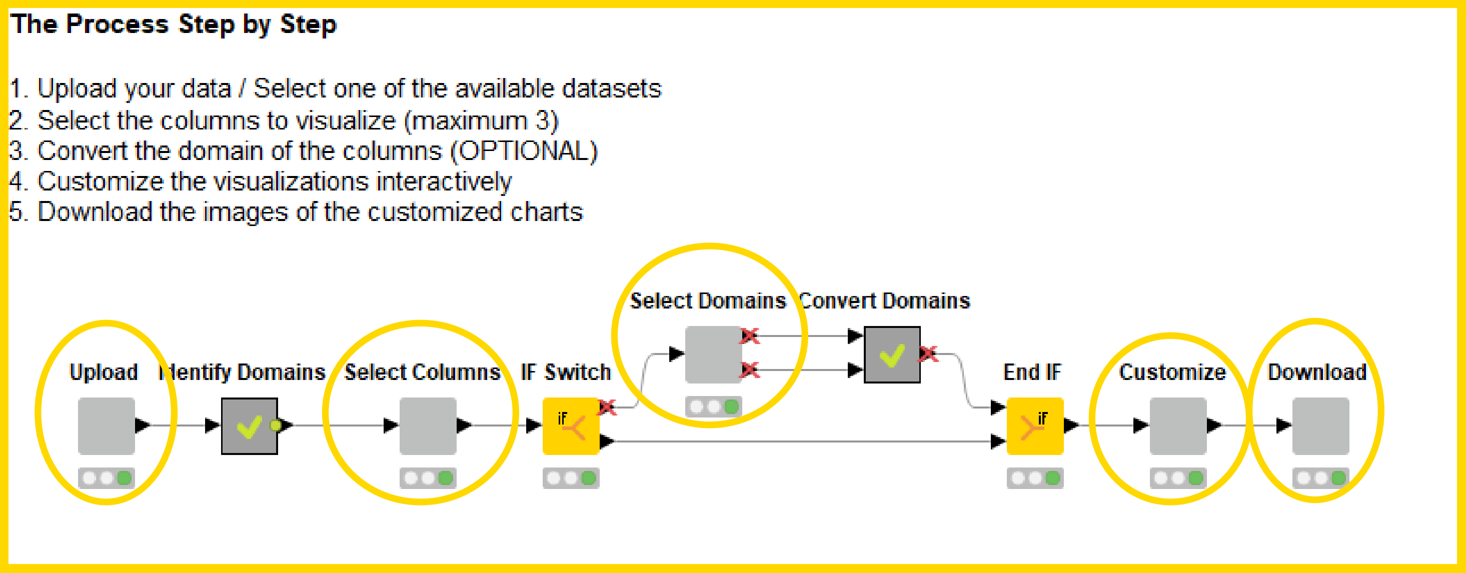
Column Name	Regex	Source Type	Mapping

Mapping by Type

KNIME Type	Mapping
String	→ String → VARCHAR
Boolean value	→ Boolean → BOOLEAN
Number (integer)	→ Integer → INTEGER
Number (long)	→ Long → BIGINT
Number (double)	→ Double → DOUBLE
Local Date	→ LocalDate → DATE
Local Time	→ LocalTime → TIME
Local Date Time	→ LocalDateTime → TIMESTAMP
Zoned Date Time	→ ZonedDateTime → TIMESTAMP_WITH_TIMEZONE

OK | Apply | Cancel

Creating a Dashboard on KNIME WebPortal



The Process Step by Step

1. Upload your data / Select one of the available datasets
2. Select the columns to visualize (maximum 3)
3. Convert the domain of the columns (OPTIONAL)
4. Customize the visualizations interactively
5. Download the images of the customized charts

Step 1
Upload File

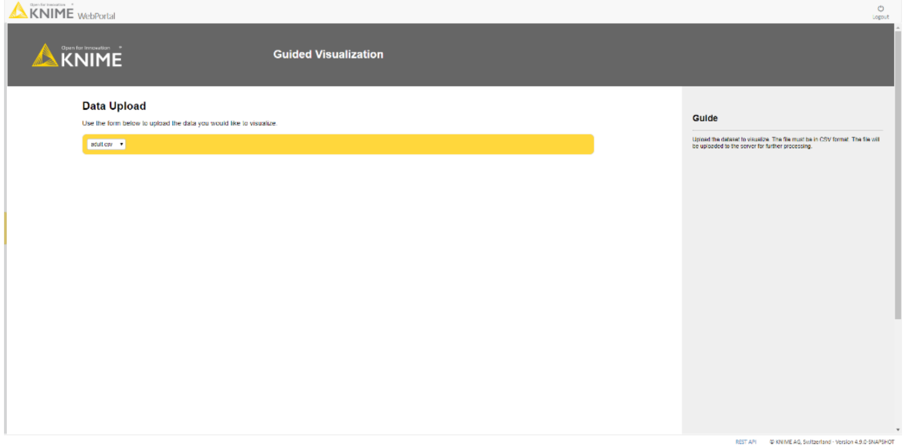
Step 2
Select Columns

Step 3
Customize Column Domains

Step 4
Interactive View

Step 5
Download Image

Workflow on KNIME WebPortal



Guided Visualization

Data Upload

Use the form below to upload the data you would like to visualize.

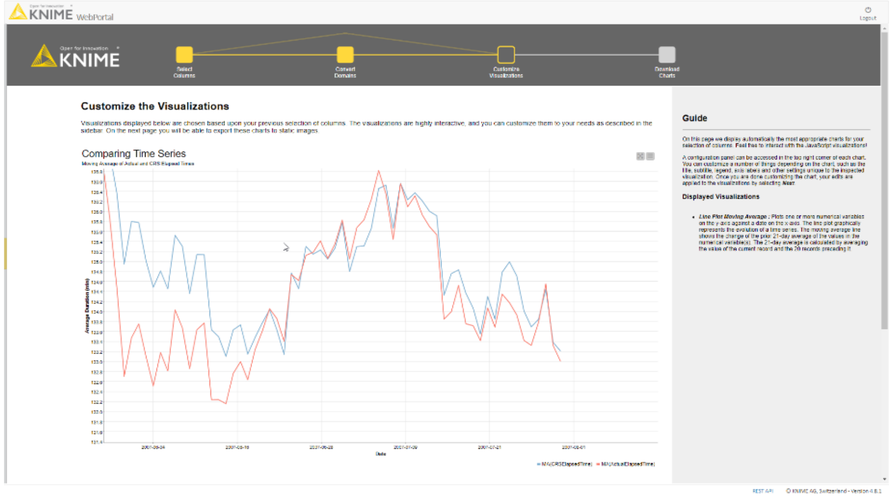
adult.csv ▾

Guide

Upload the dataset to visualize. The file must be in CSV format. The file will be uploaded to the server for further processing.

REST API © KNIME AG, Switzerland - Version 4.9.0-SNAPSHOT

WebPortal Page (Step 1)
Upload File



Customize the Visualizations

Visualizations displayed below are chosen based upon your previous selection of columns. The visualizations are highly interactive, and you can customize them to your needs as described in the sidebar. On the next page you will be able to export these charts to static images.

Guide

On this page we display automatically the most appropriate charts for your selection of columns. Feel free to interact with the JavaScript visualizations!

A configuration panel can be accessed in the top right corner of each chart. You can customize a number of things depending on the chart, such as the title, subtitle, legend, axis labels and other settings unique to the inspected visualization. Once you are done customizing the chart, your edits are applied to the visualizations by selecting *Next*.

Displayed Visualizations

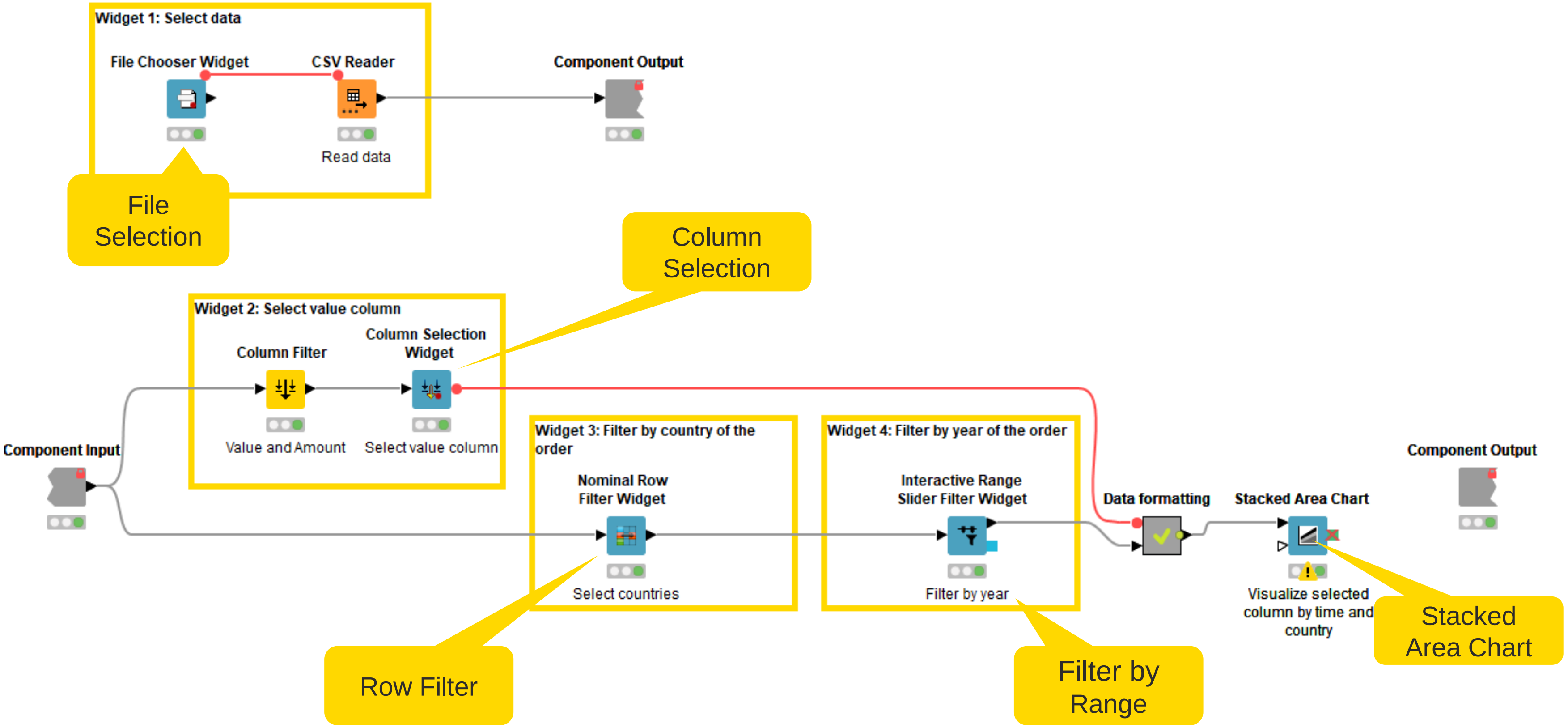
- ***Line Plot Moving Average :*** Plots one or more numerical variables on the y-axis against a date on the x-axis. The line plot graphically represents the evolution of a time series. The moving average line shows the change of the prior 21-day average of the values in the numerical variable(s). The 21-day average is calculated by averaging the value of the current record and the 20 records preceding it.

REST API © KNIME AG, Switzerland - Version 4.8.1

WebPortal Page (Step 4)
Interactive View

Available in KNIME Server

Components to Produce Dashboard on Web Page



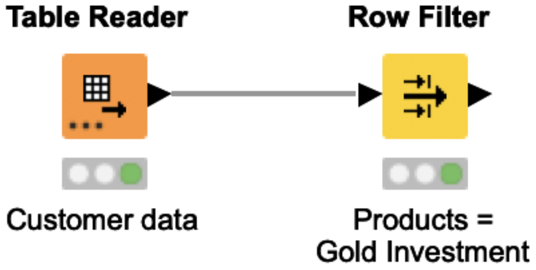
Exercise: 09_Deployment

- Write the clean customer data to an Excel file into the folder "data/temp"
- Write the full transaction data to the "Transactions.sqlite" database

Flow Variables

Flow Variables: Usage Example

- Each month you need to produce a sales report for the most popular product

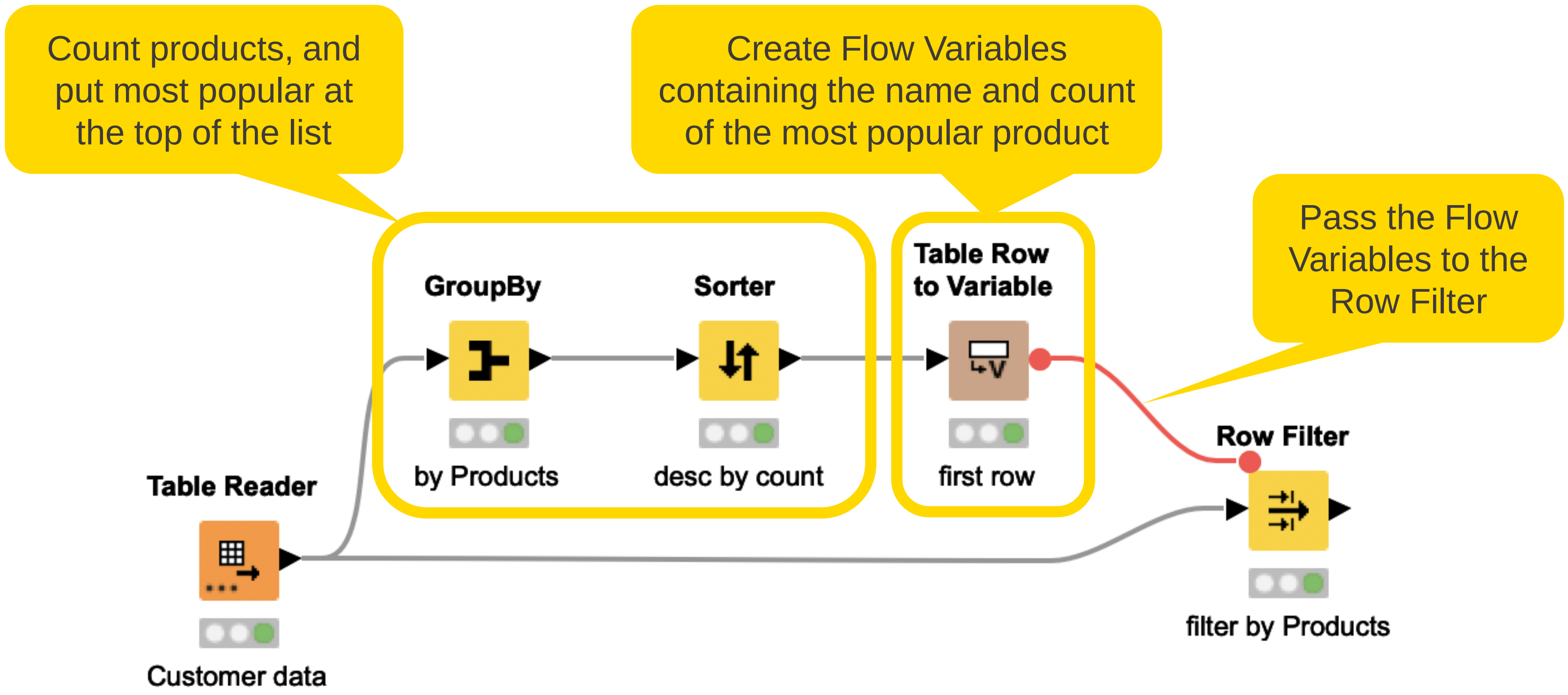


Filter only rows where Products = Gold Investment

Flow Variables: Usage Example

- Each month I need to launch the Analytics Platform, aggregate the data to identify the most popular product, and update the Row Filter accordingly
- Or do I? Maybe Flow Variables can help…

Automatically Filter by Most Popular Product



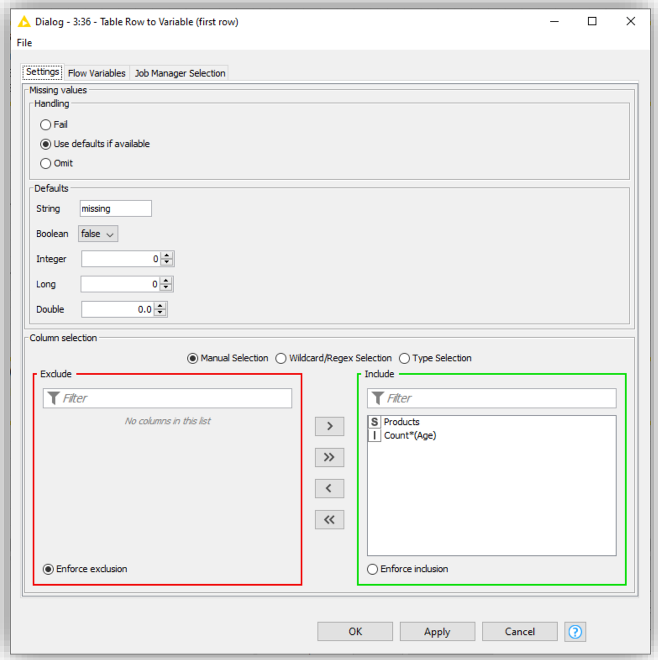
Count products, and put most popular at the top of the list

Create Flow Variables containing the name and count of the most popular product

Pass the Flow Variables to the Row Filter

Table Row to Variable

- Takes a table as input and converts the first row to Flow Variables
  - Column names -> Flow Variable names
  - Column values -> Flow Variable values
- Only the first row is transformed, additional rows are discarded



Dialog - 3:36 - Table Row to Variable (first row)

Missing values

Handling

○ Fail
● Use defaults if available
○ Omit

Defaults

String: missing
Boolean: false
Integer: 0
Long: 0
Double: 0.0

Column selection

● Manual Selection ○ Wildcard/Regex Selection ○ Type Selection

Exclude

Filter

No columns in this list

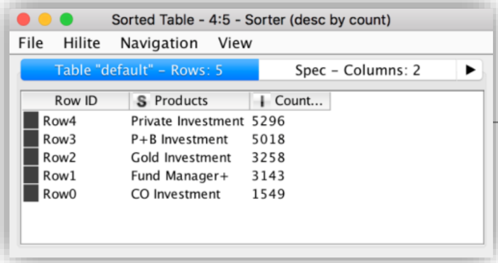
● Enforce exclusion

Include

Filter

S Products
I Count*(Age)

○ Enforce inclusion



Sorted Table - 4:5 - Sorter (desc by count)

Row ID	S Products	I Count...
Row4	Private Investment	5296
Row3	P+B Investment	5018
Row2	Gold Investment	3258
Row1	Fund Manager+	3143
Row0	CO Investment	1549

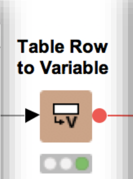
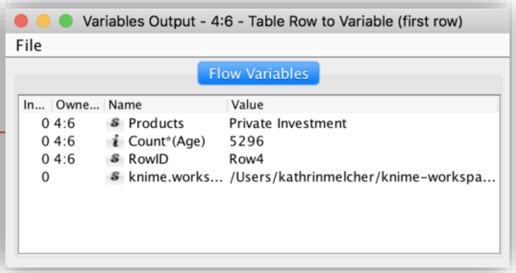


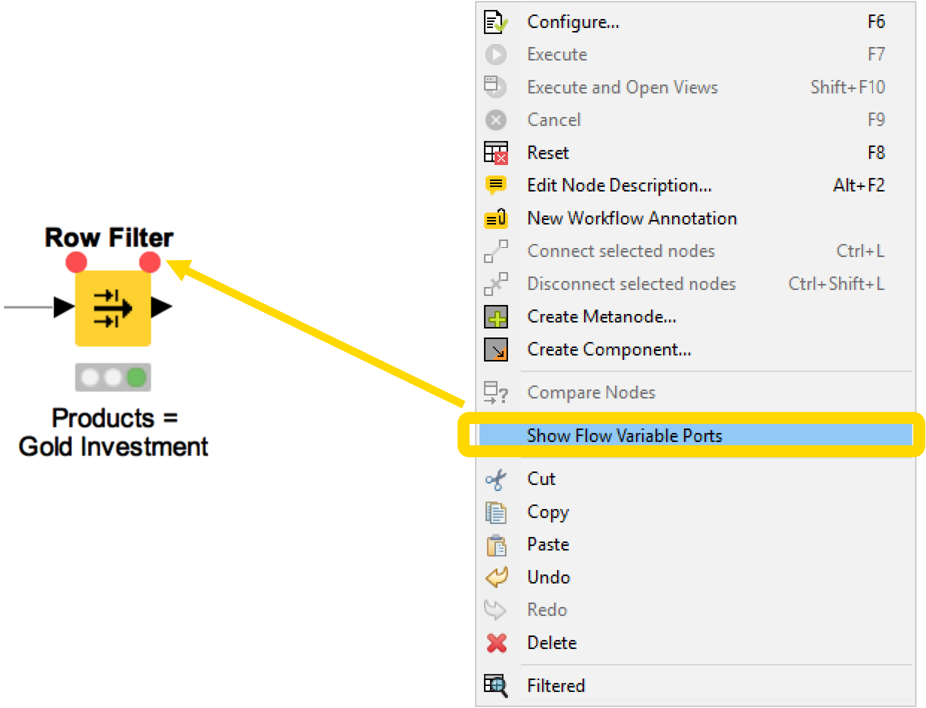
Table Row to Variable



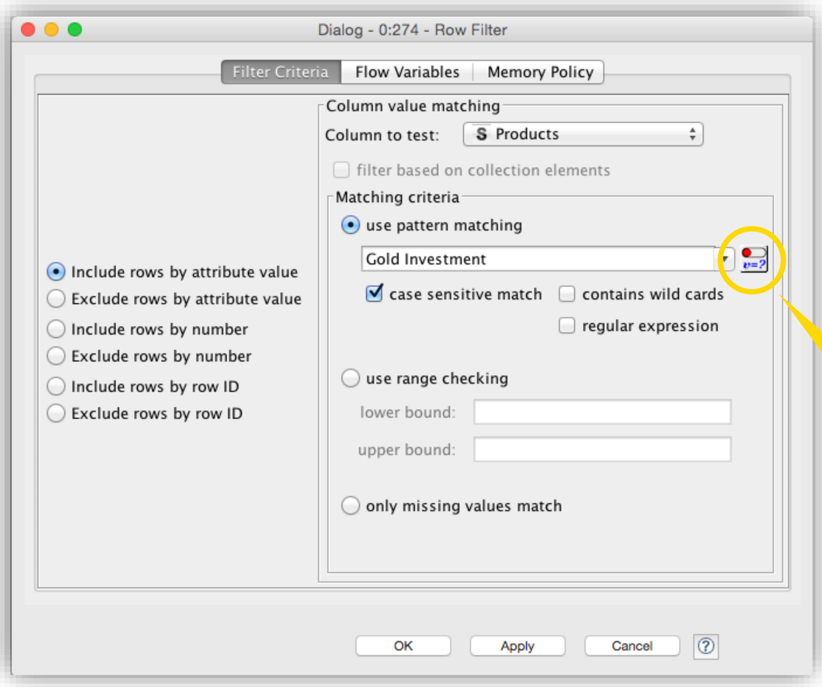
Variables Output - 4:6 - Table Row to Variable (first row)

In...	Owne...	Name	Value
0	4:6	Products	Private Investment
0	4:6	Count*(Age)	5296
0	4:6	RowID	Row4
0		knime.works...	/Users/kathrinmelcher/knime-workspa...

Flow Variable Ports



Apply a Flow Variable (Button)



Dialog - 0:274 - Row Filter

Filter Criteria | Flow Variables | Memory Policy

- ● Include rows by attribute value
- ○ Exclude rows by attribute value
- ○ Include rows by number
- ○ Exclude rows by number
- ○ Include rows by row ID
- ○ Exclude rows by row ID

Column value matching

Column to test: Products

□ filter based on collection elements

Matching criteria

● use pattern matching

Gold Investment

☑ case sensitive match □ contains wild cards

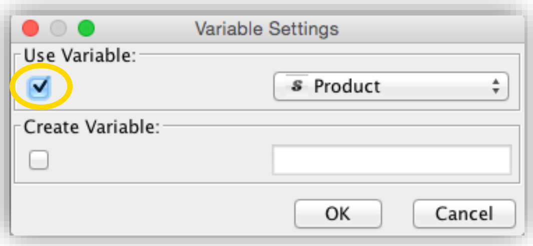
□ regular expression

○ use range checking

lower bound:

upper bound:

○ only missing values match



Variable Settings

Use Variable:

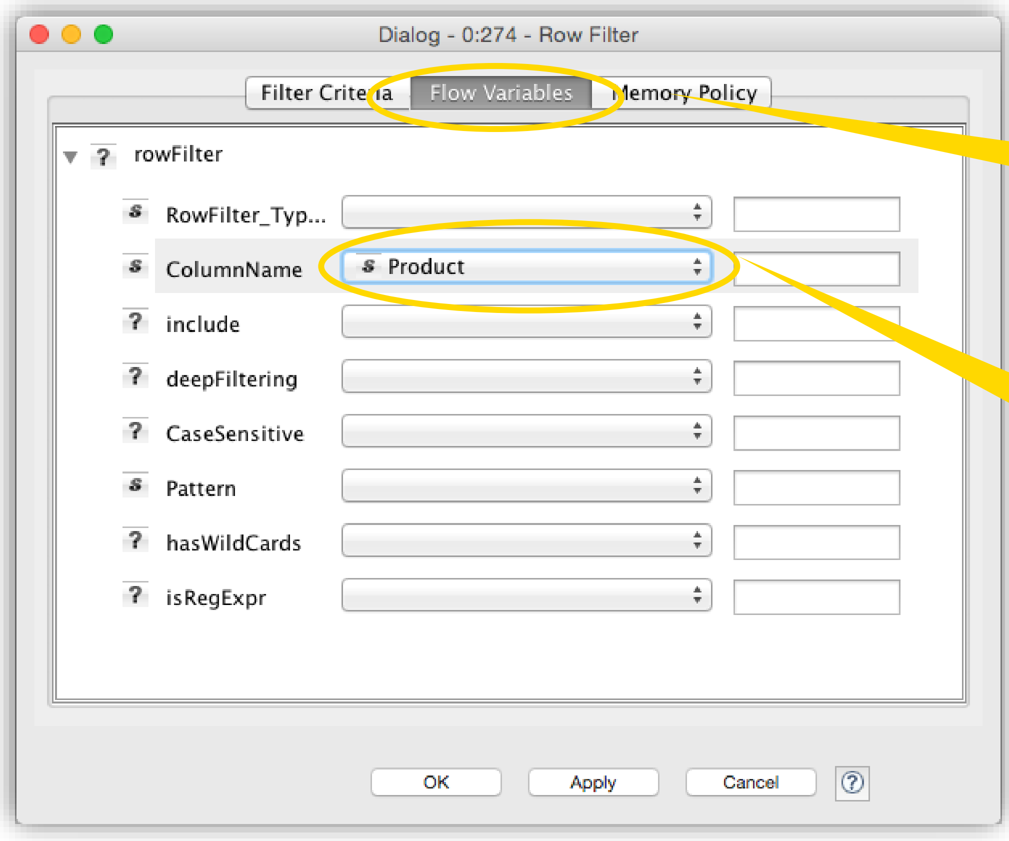
☑ Product

Create Variable:

□

The Flow Variable button

Apply a Flow Variable (Advanced)



Dialog - 0:274 - Row Filter

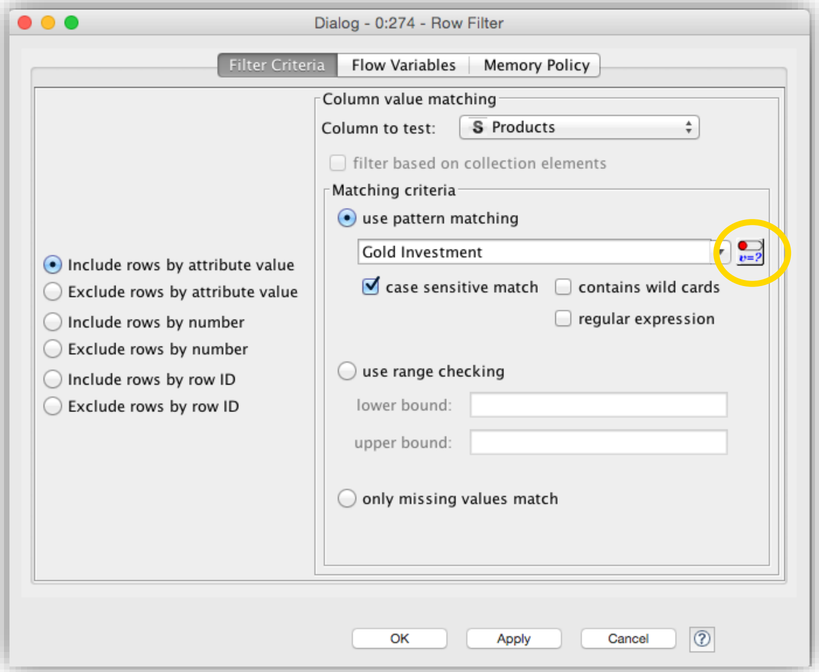
rowFilter

- RowFilter_Typ...
- ColumnName: Product
- include
- deepFiltering
- CaseSensitive
- Pattern
- hasWildCards
- isRegExpr

The Flow Variables tab

List of available Flow Variables

Create a Flow Variable (Button)



Dialog - 0:274 - Row Filter

Filter Criteria | Flow Variables | Memory Policy

- ● Include rows by attribute value
- ○ Exclude rows by attribute value
- ○ Include rows by number
- ○ Exclude rows by number
- ○ Include rows by row ID
- ○ Exclude rows by row ID

Column value matching

Column to test: S Products

□ filter based on collection elements

Matching criteria

● use pattern matching

Gold Investment

☑ case sensitive match □ contains wild cards

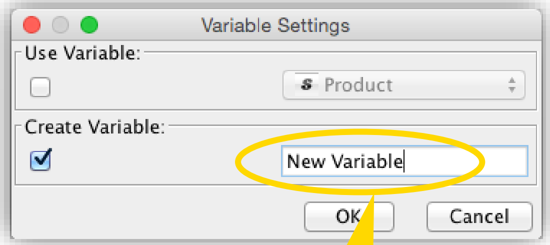
□ regular expression

○ use range checking

lower bound:

upper bound:

○ only missing values match



Variable Settings

Use Variable:

□ S Product

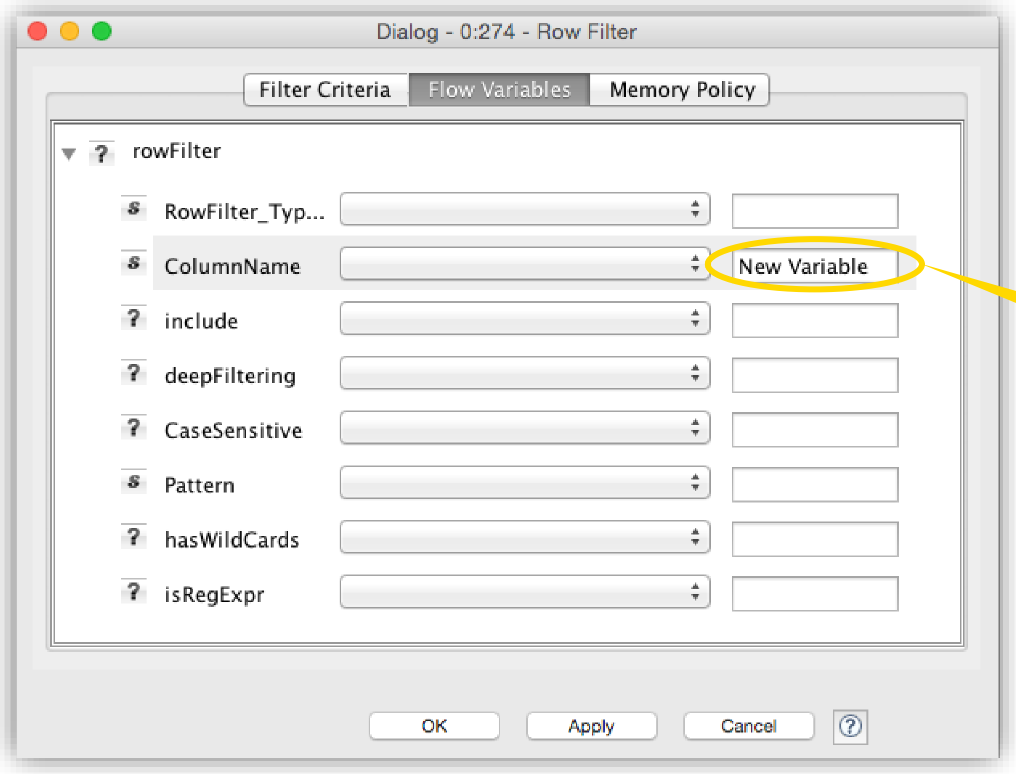
Create Variable:

☑ New Variable

Name of the new Flow Variable

Create a Flow Variable (Advanced)

- Converting a setting value into a Flow Variable



Dialog - 0:274 - Row Filter

rowFilter		
RowFilter_Typ...		
ColumnName		New Variable
include		
deepFiltering		
CaseSensitive		
Pattern		
hasWildCards		
isRegExpr		

Name of the new Flow Variable

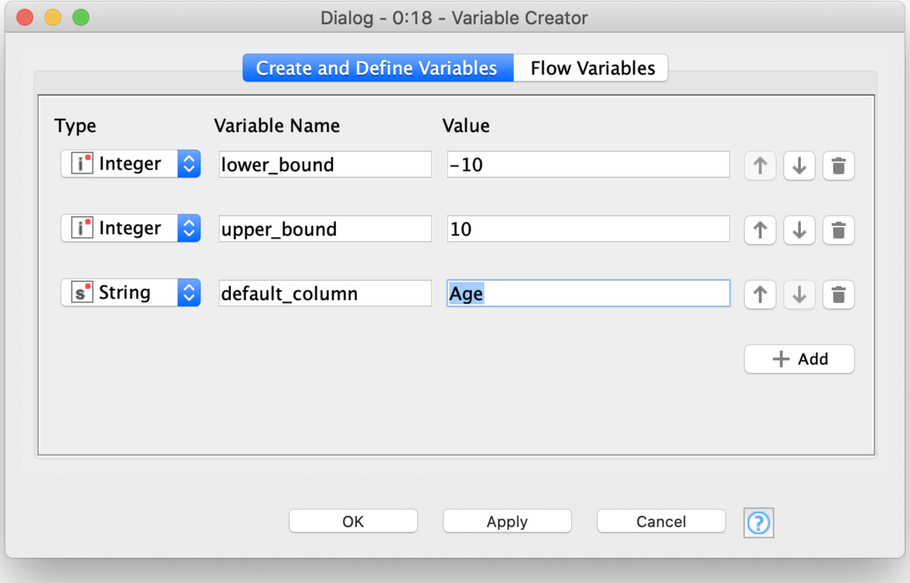
Exercise: 10_Flow_Variables

- Activity I: Filter the customer data to
  - Customers of the "Gold Investment" product
  - Customers of the most common product in the data

Variable Creator

- Allows to create flow variables of different types
- Click on “+ Add” to add a new variable and define a custom
  - Variable Name
  - Variable Value

Variable Creator

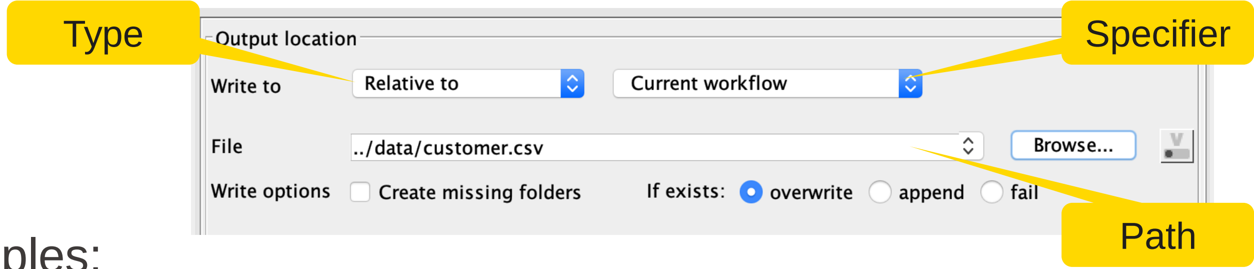


Dialog - 0:18 - Variable Creator

Type	Variable Name	Value
Integer	lower_bound	-10
Integer	upper_bound	10
String	default_column	Age

Path Variables

- Special flow variable type to point to a file or folder
  - E.g. to control output location of a file
- A path type consists of three parts:
  - **Type**: Specifies the file system type e.g. local, relative, mountpoint, custome_url or connected.
  - **Specifier**: Optional string with additional file system specific information e.g. relative to which location (knime.workflow)
  - **Path**: Specifies the location within the file system



Output location

Write to: Relative to | Current workflow

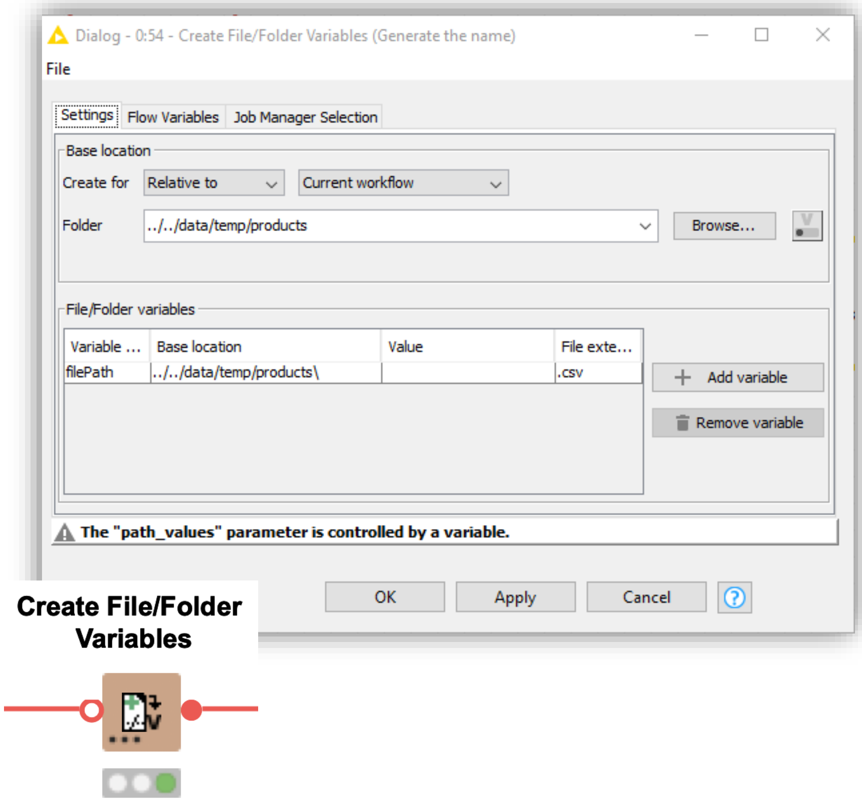
File: ../data/customer.csv

Write options: ☐ Create missing folders If exists: ● overwrite ○ append ○ fail

- Examples:
  - (LOCAL, , C:\Users\username\Desktop)
  - (RELATIVE, knime.workflow, file1.csv)
  - (MOUNTPOINT, MOUNTPOINT_NAME, /path/to/file1.csv)
  - (CONNECTED, amazon-s3:eu-west-1, /mybucket/file1.csv)

Create File/Folder Variables

- Creates one or multiple path flow variable(s) pointing to files / folders
- Inputs:
  - Base location
  - Flow variable name(s)
  - Value (file name or path relative to base location)
  - File extension (optional)
- Output variables can be used to control the output location in writer nodes.



Dialog - 0:54 - Create File/Folder Variables (Generate the name)

File

Settings | Flow Variables | Job Manager Selection

Base location

Create for: Relative to — Current workflow

Folder: ../../data/temp/products — Browse...

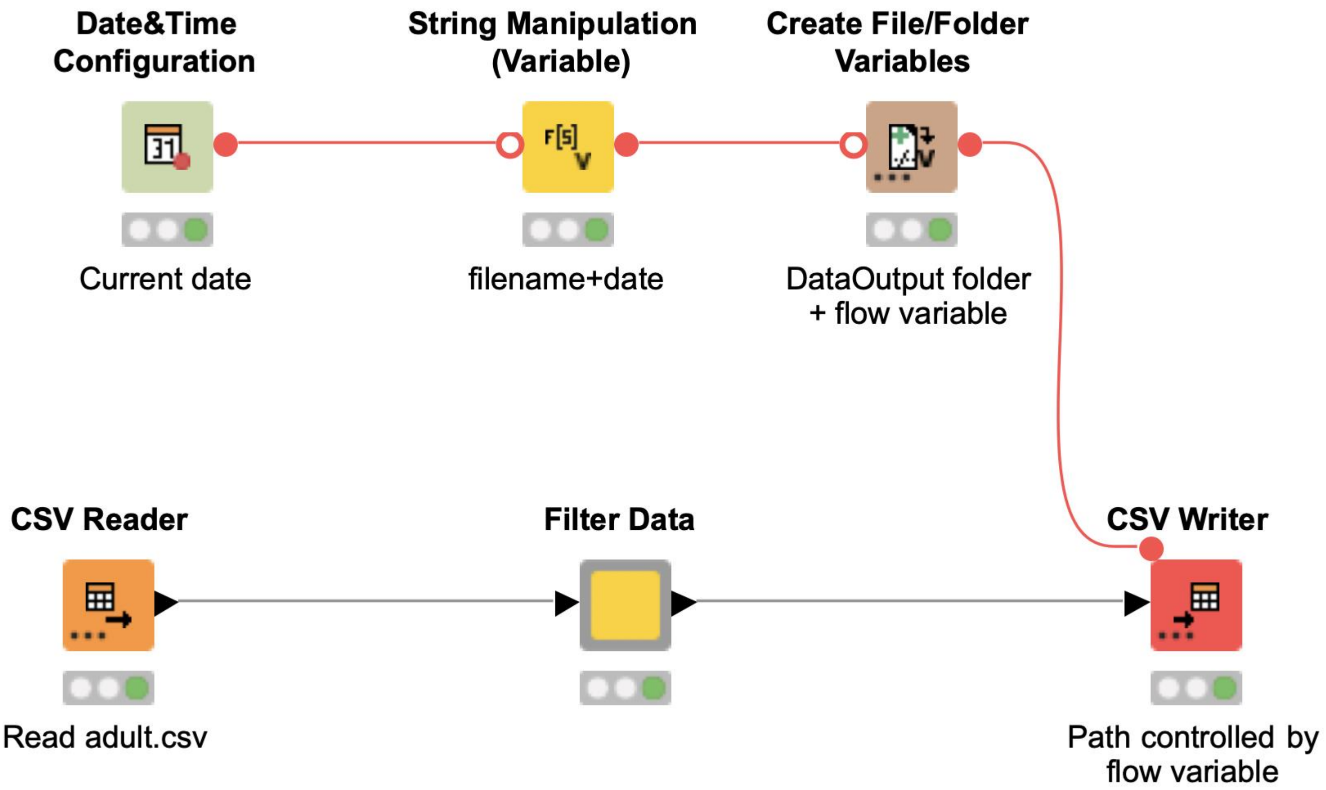
File/Folder variables

Variable ...	Base location	Value	File exte...
filePath	../../data/temp/products\		.csv

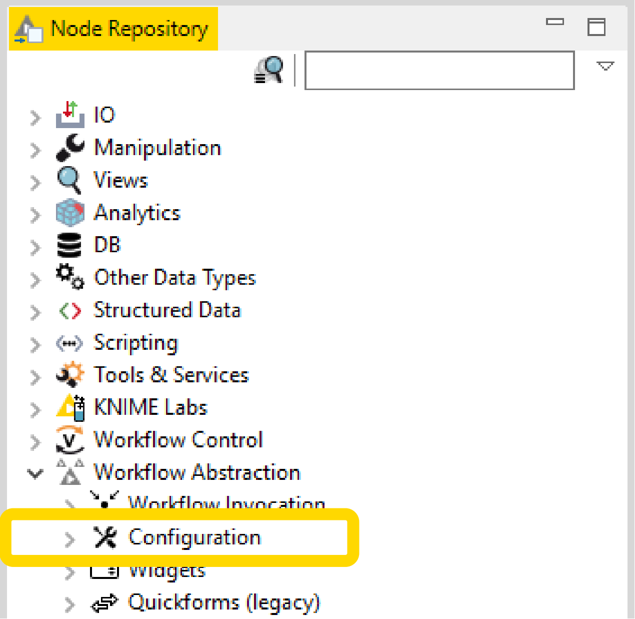
The "path_values" parameter is controlled by a variable.

Create File/Folder Variables

Example: Add Execution Date to File Name

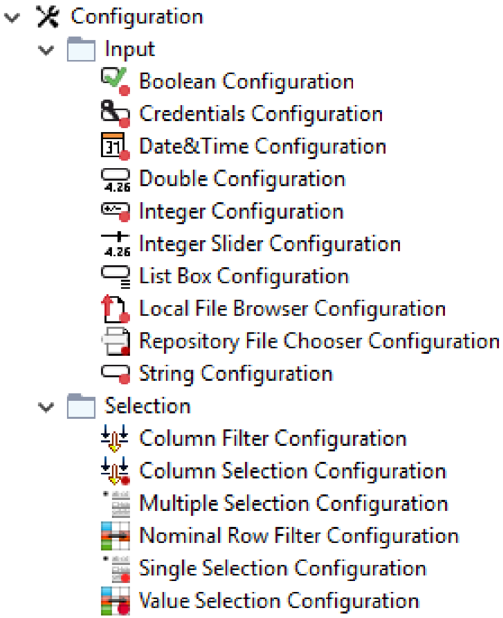

Date&Time Configuration
String Manipulation (Variable)
Create File/Folder Variables
Current date
filename+date
DataOutput folder + flow variable
CSV Reader
Filter Data
CSV Writer
Read adult.csv
Path controlled by flow variable

Configuration Nodes for Variable Creation and Output



Node Repository

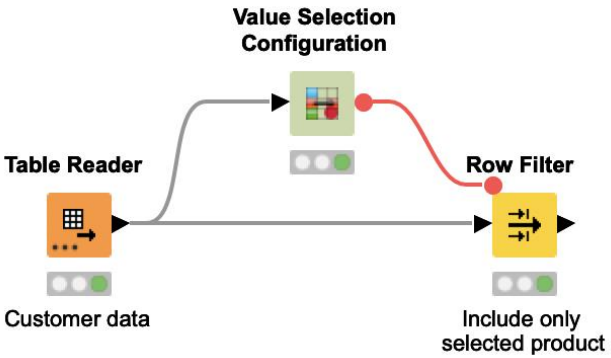
- IO
- Manipulation
- Views
- Analytics
- DB
- Other Data Types
- Structured Data
- Scripting
- Tools & Services
- KNIME Labs
- Workflow Control
- Workflow Abstraction
  - Workflow Invocation
  - Configuration
  - Widgets
  - Quickforms (legacy)

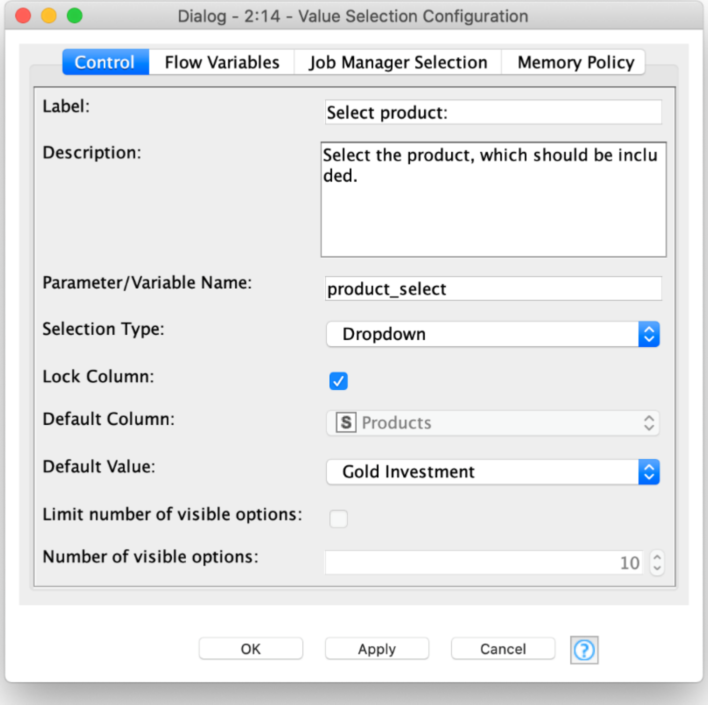


- Configuration
  - Input
    - Boolean Configuration
    - Credentials Configuration
    - Date&Time Configuration
    - Double Configuration
    - Integer Configuration
    - Integer Slider Configuration
    - List Box Configuration
    - Local File Browser Configuration
    - Repository File Chooser Configuration
    - String Configuration
  - Selection
    - Column Filter Configuration
    - Column Selection Configuration
    - Multiple Selection Configuration
    - Nominal Row Filter Configuration
    - Single Selection Configuration
    - Value Selection Configuration

Configuration Node Configuration

Use Configuration nodes to create Flow Variables





Label:	Select product:
Description:	Select the product, which should be inclu ded.
Parameter/Variable Name:	product_select
Selection Type:	Dropdown
Lock Column:	☑
Default Column:	S Products
Default Value:	Gold Investment
Limit number of visible options:	□
Number of visible options:	10

Simple Configuration of Component

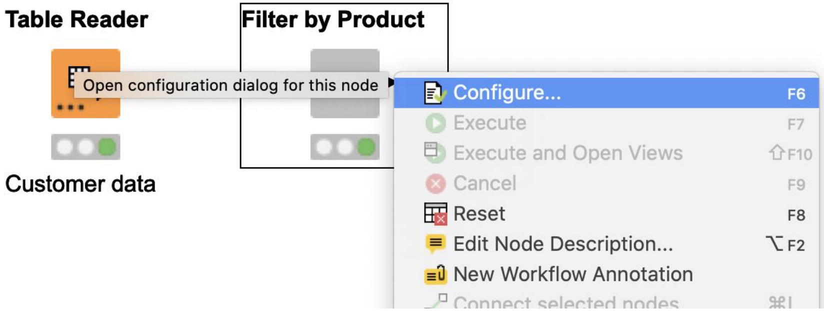
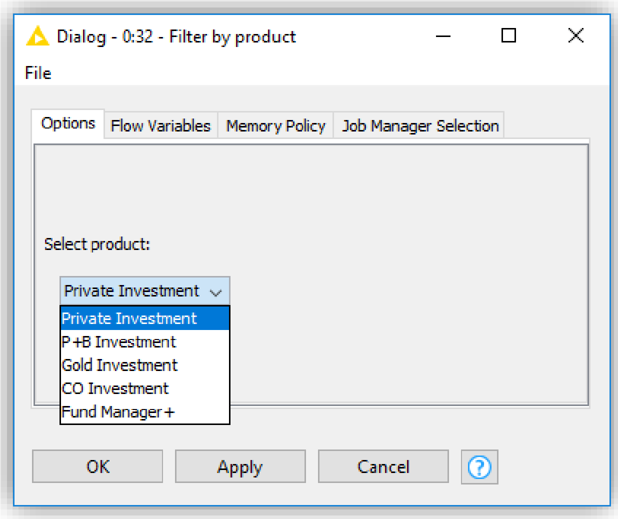


Table Reader
Customer data
Filter by Product
Open configuration dialog for this node
Configure... F6
Execute F7
Execute and Open Views ⇧F10
Cancel F9
Reset F8
Edit Node Description... ⌥F2
New Workflow Annotation
Connect selected nodes ⌘L



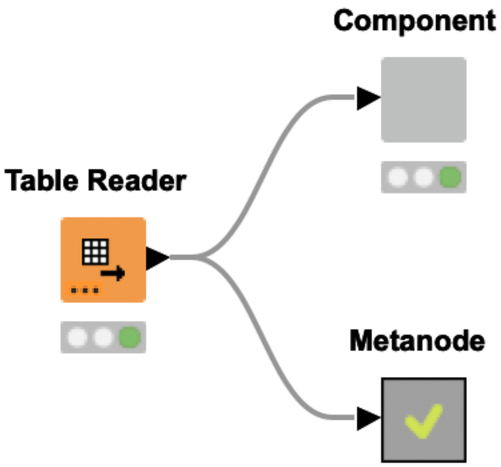
Select product: Private Investment

- Private Investment
- P+B Investment
- Gold Investment
- CO Investment
- Fund Manager+

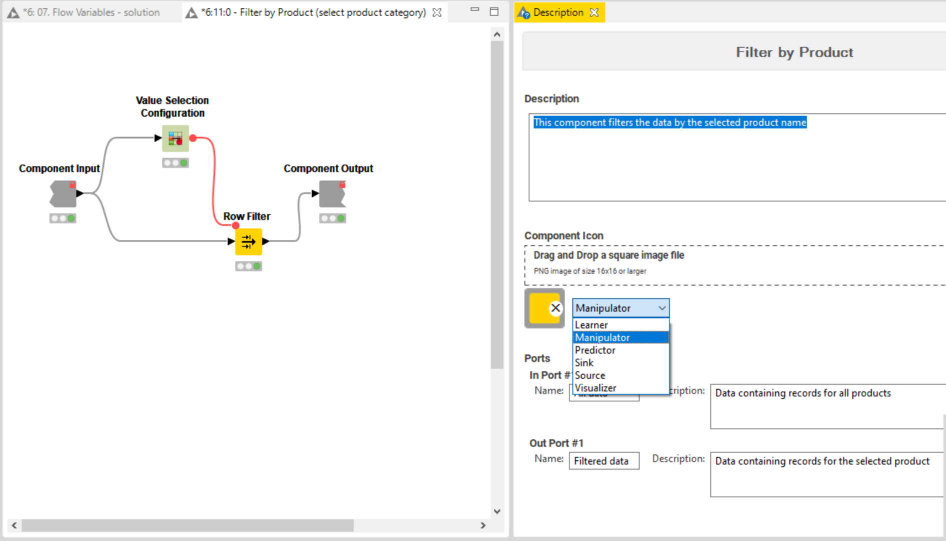
- Double click a component to configure it
- For use on the WebPortal, replace Configuration nodes with Widget nodes

Components

- Encapsulate a functionality for reuse and sharing
- Main features:
  - Local Flow Variable scope
  - Configurable via Configuration nodes
- Key to advanced functionality in KNIME products
  - Component corresponds to a KNIME WebPortal page
  - Configurations on a WebPortal page are defined using Widg
  - Can be shared via KNIME Hub



Component Description



*6: 07. Flow Variables - solution
*6:11:0 - Filter by Product (select product category)
Value Selection Configuration
Component Input
Component Output
Row Filter
Description

Filter by Product

Description

This component filters the data by the selected product name

Component Icon

Drag and Drop a square image file

PNG image of size 16x16 or larger

Manipulator

Learner
Manipulator
Predictor
Sink
Source
Visualizer

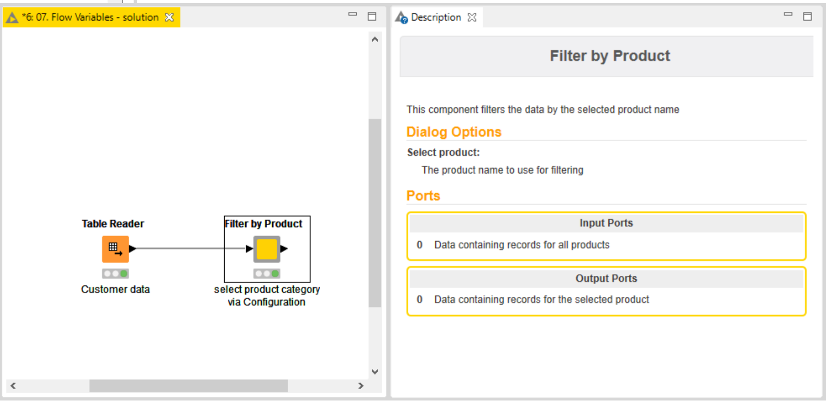
Ports

In Port #

Name: | Description: Data containing records for all products

Out Port #1

Name: Filtered data | Description: Data containing records for the selected product



*6: 07. Flow Variables - solution
Table Reader
Customer data
Filter by Product
select product category via Configuration
Description

Filter by Product

This component filters the data by the selected product name

Dialog Options

Select product:

The product name to use for filtering

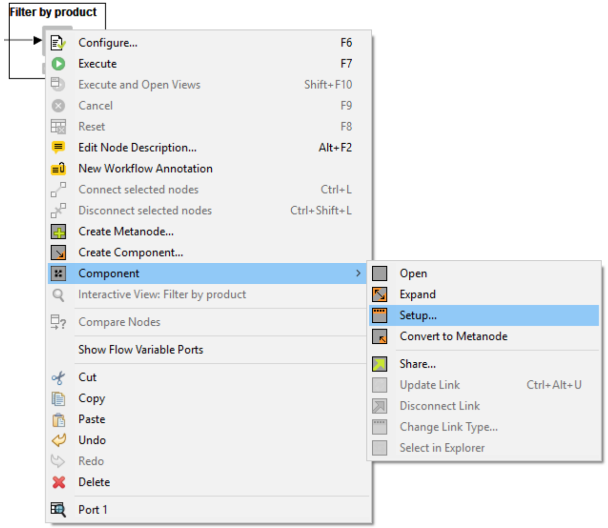
Ports

Input Ports	
0	Data containing records for all products

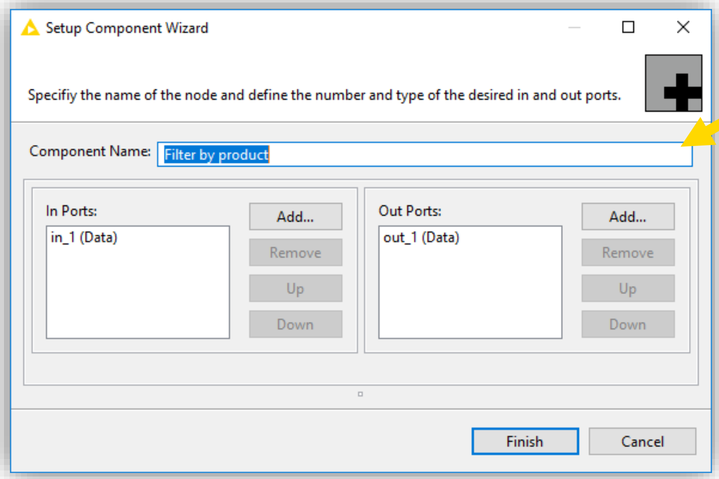
Output Ports	
0	Data containing records for the selected product

- Double click a component to configure it
- For use on the WebPortal, replace Configuration nodes with Widget nodes

Configure Component Ports



- Add input and output ports to metanodes/components
- Remove ports to adapt to changes after creation of the metanode/component

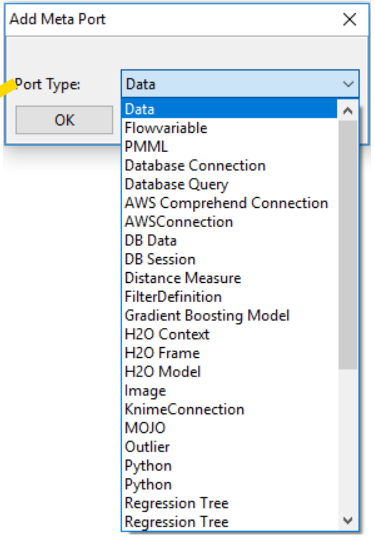

Setup Component Wizard

Specifiy the name of the node and define the number and type of the desired in and out ports.

Component Name: Filter by product

In Ports: in_1 (Data)

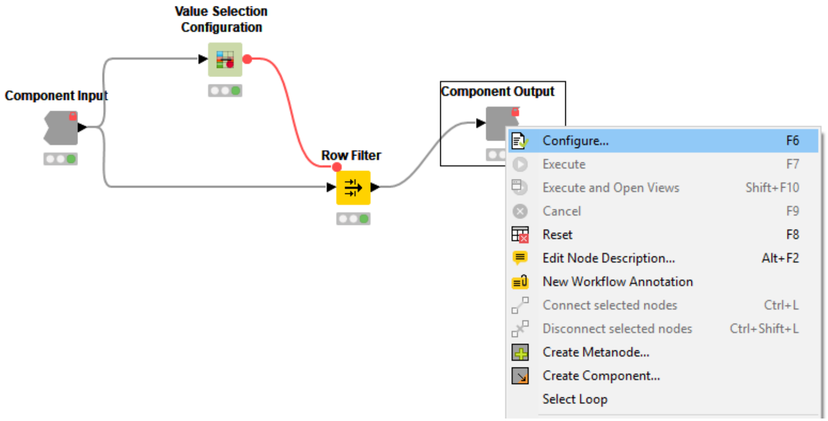
Out Ports: out_1 (Data)

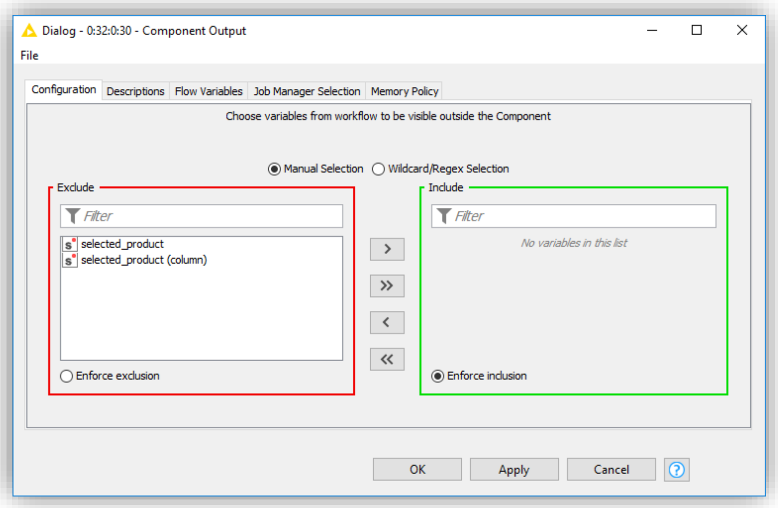

Add Meta Port

Port Type: Data

Passing Variables from Components

- Flow Variables are -by default - only available locally inside the component
- Configure the component input/output to pass Flow Variables from/to outside the component


Value Selection Configuration
Component Input
Row Filter
Component Output
Configure... F6
Execute F7
Execute and Open Views Shift+F10
Cancel F9
Reset F8
Edit Node Description... Alt+F2
New Workflow Annotation
Connect selected nodes Ctrl+L
Disconnect selected nodes Ctrl+Shift+L
Create Metanode...
Create Component...
Select Loop


Dialog - 0:32:0:30 - Component Output

File

Configuration | Descriptions | Flow Variables | Job Manager Selection | Memory Policy

Choose variables from workflow to be visible outside the Component

● Manual Selection ○ Wildcard/Regex Selection

Exclude

Filter

selected_product
selected_product (column)

○ Enforce exclusion

Include

Filter

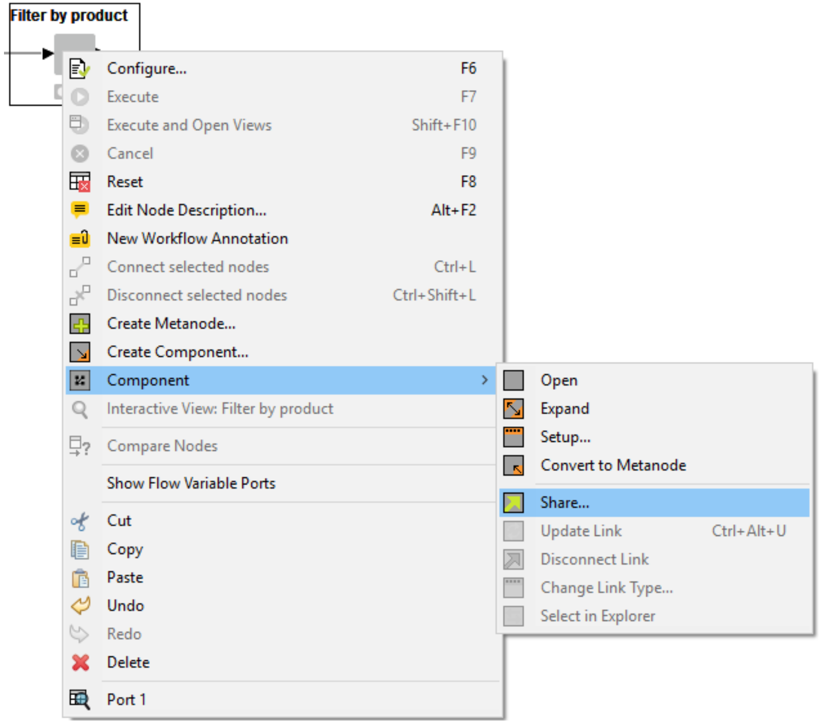
No variables in this list

● Enforce inclusion

OK | Apply | Cancel | ?

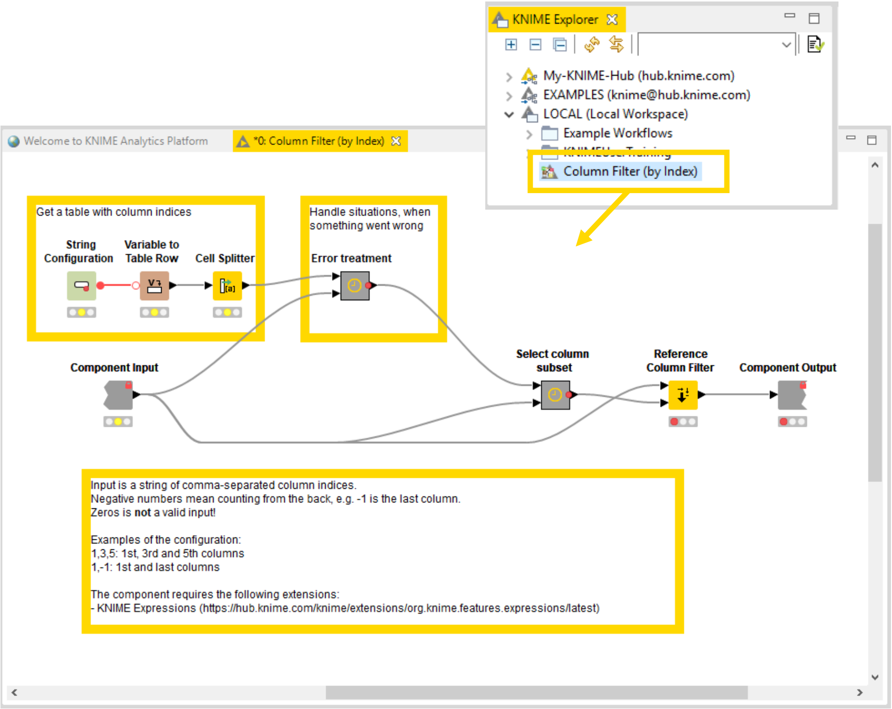
What is a Shared Component?

- Components can be saved in your KNIME workspace for later reuse
- To do this, simply right-click any component and select “Share…”
- Shared components are read-only instances of a component
- Public Shared Components are available on the EXAMPLES Server and on the KNIME Hub



How can you Edit a Shared Component?

- Components can be edited using the Component Editor, similar to workflows
- To edit a component using the Component Editor, double-click the component in its location in the KNIME Explorer
- To ensure components are executable when opened in the Component Editor, choose the option to “Include input data with component” when sharing it



Get a table with column indices

Handle situations, when something went wrong

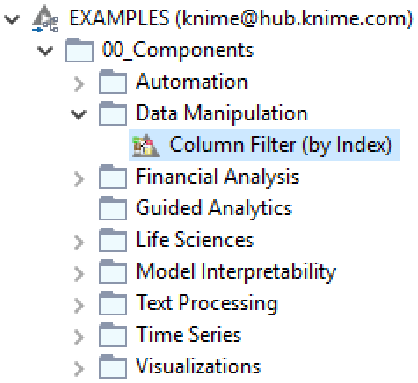
Input is a string of comma-separated column indices.
Negative numbers mean counting from the back, e.g. -1 is the last column.
Zeros is **not** a valid input!

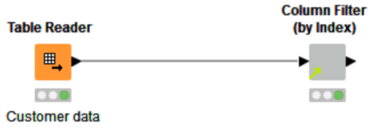
Examples of the configuration:
1,3,5: 1st, 3rd and 5th columns
1,-1: 1st and last columns

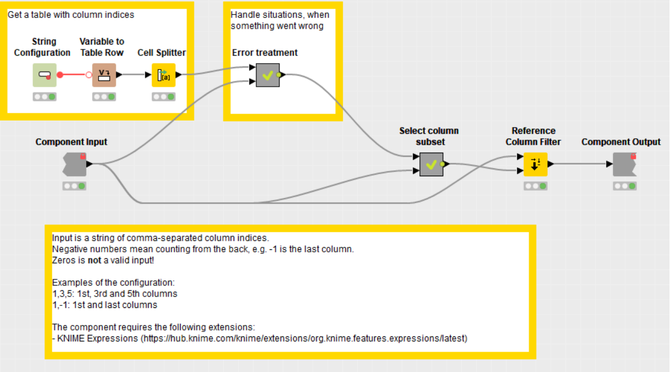
The component requires the following extensions:
- KNIME Expressions (https://hub.knime.com/knime/extensions/org.knime.features.expressions/latest)

How can you Use a Shared Component?

- To use a Shared Component, drag and drop it to your workflow editor
- Instances of Shared Components can be updated either manually or when the workflow is opened
- A Shared Component can also be unlinked from its original location, which makes it editable in the workflow directly
- Update Shared Components by overwriting them


EXAMPLES (knime@hub.knime.com)
00_Components
Automation
Data Manipulation
Column Filter (by Index)
Financial Analysis
Guided Analytics
Life Sciences
Model Interpretability
Text Processing
Time Series
Visualizations





Get a table with column indices

Handle situations, when something went wrong

Input is a string of comma-separated column indices.
Negative numbers mean counting from the back, e.g. -1 is the last column.
Zeros is **not** a valid input!

Examples of the configuration:
1,3,5: 1st, 3rd and 5th columns
1,-1: 1st and last columns

The component requires the following extensions:
- KNIME Expressions (https://hub.knime.com/knime/extensions/org.knime.features.expressions/latest)

Exercise: 10_Flow_Variables

Start with exercise: *Flow Variables, Activity II*

- Create a component that allows a user to choose an investment product and filter the data by that product

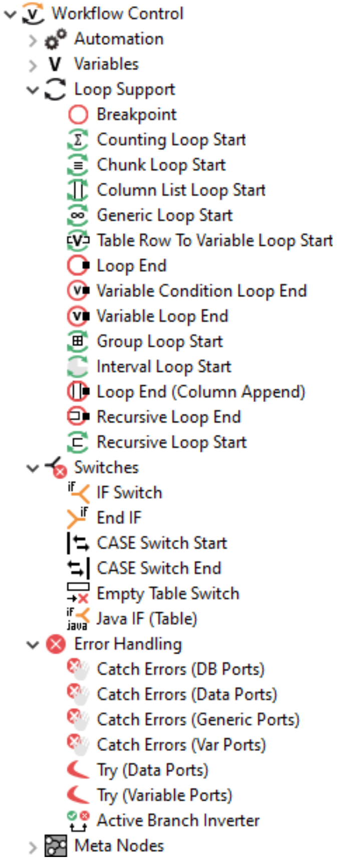
Optional Exercises

- Activity III: Create a path variable that automatically has the current execution date in the file name and write the filtered table into a CSV file
- Activity IV: Create a component that allows to select multiple products out of all available products, using a flow variable of type array

Workflow Control
Loops, Switches, Try-Catch

Workflow Control Structures

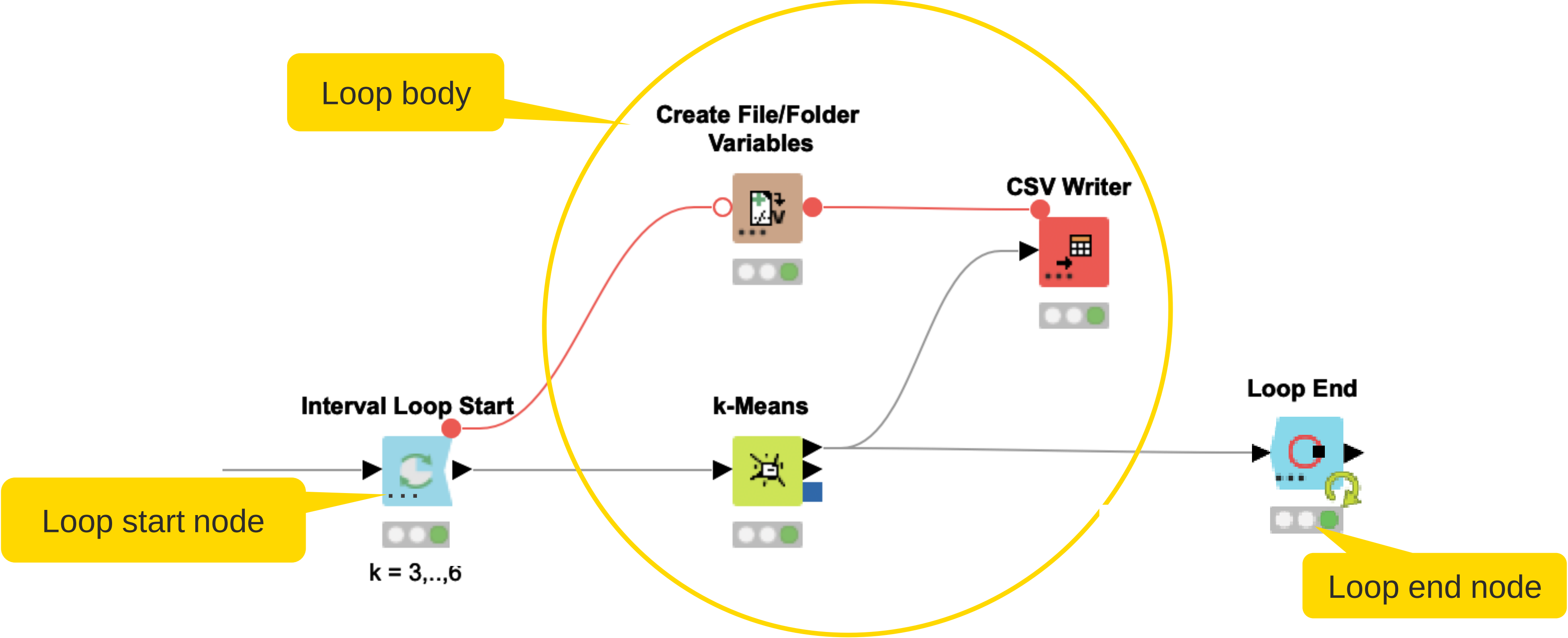
- Loops
  - Iterate over a workflow snippet with variable inputs.
- Switches
  - Direct the path of a workflow by selectively executing one or more workflow branches.
- Try-Catch
  - Handle workflow branches that may fail in execution - when you don’t know about this before executing



- Workflow Control
  - Automation
  - Variables
  - Loop Support
    - Breakpoint
    - Counting Loop Start
    - Chunk Loop Start
    - Column List Loop Start
    - Generic Loop Start
    - Table Row To Variable Loop Start
    - Loop End
    - Variable Condition Loop End
    - Variable Loop End
    - Group Loop Start
    - Interval Loop Start
    - Loop End (Column Append)
    - Recursive Loop End
    - Recursive Loop Start
  - Switches
    - IF Switch
    - End IF
    - CASE Switch Start
    - CASE Switch End
    - Empty Table Switch
    - Java IF (Table)
  - Error Handling
    - Catch Errors (DB Ports)
    - Catch Errors (Data Ports)
    - Catch Errors (Generic Ports)
    - Catch Errors (Var Ports)
    - Try (Data Ports)
    - Try (Variable Ports)
    - Active Branch Inverter
  - Meta Nodes

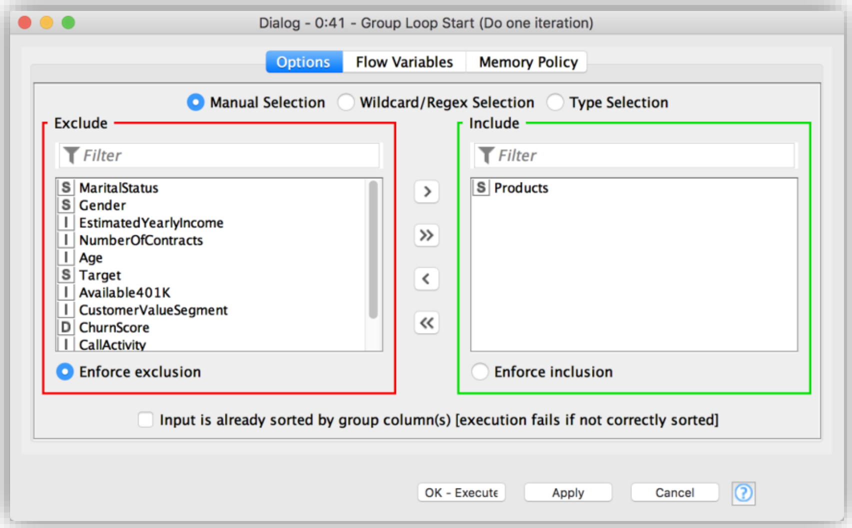
The Loop Block

- A loop block is defined by the appropriate loop start and loop end nodes.
- Loop body = the nodes in between (including the side branches).



Group Loop Start

- Similar to GroupBy except without aggregation tab.
- Each iteration of the loop passes the next group of rows.
- You can implement an aggregation task. It can be anything from a complex calculation to updating a database.



Dialog - 0:41 - Group Loop Start (Do one iteration)

● Manual Selection ○ Wildcard/Regex Selection ○ Type Selection

Exclude

S MaritalStatus
S Gender
I EstimatedYearlyIncome
I NumberOfContracts
I Age
S Target
I Available401K
I CustomerValueSegment
D ChurnScore
I CallActivity

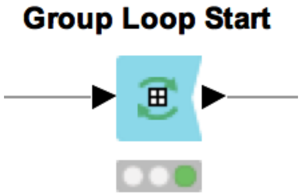
● Enforce exclusion

Include

S Products

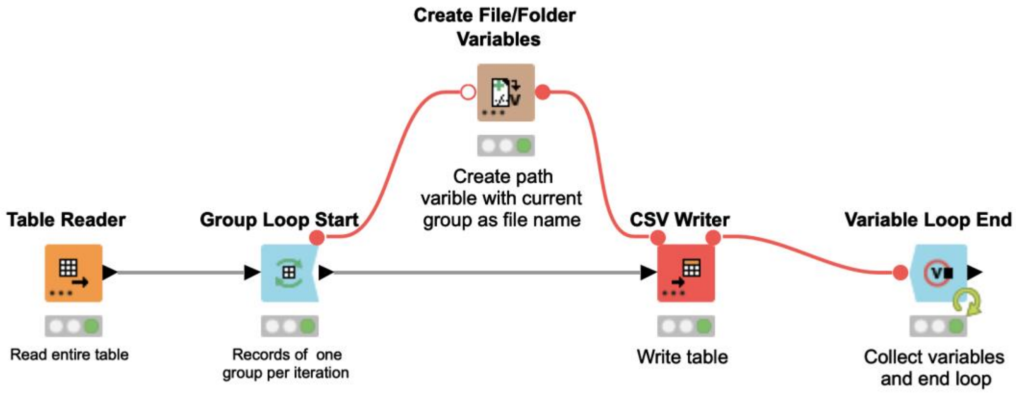
○ Enforce inclusion

□ Input is already sorted by group column(s) [execution fails if not correctly sorted]



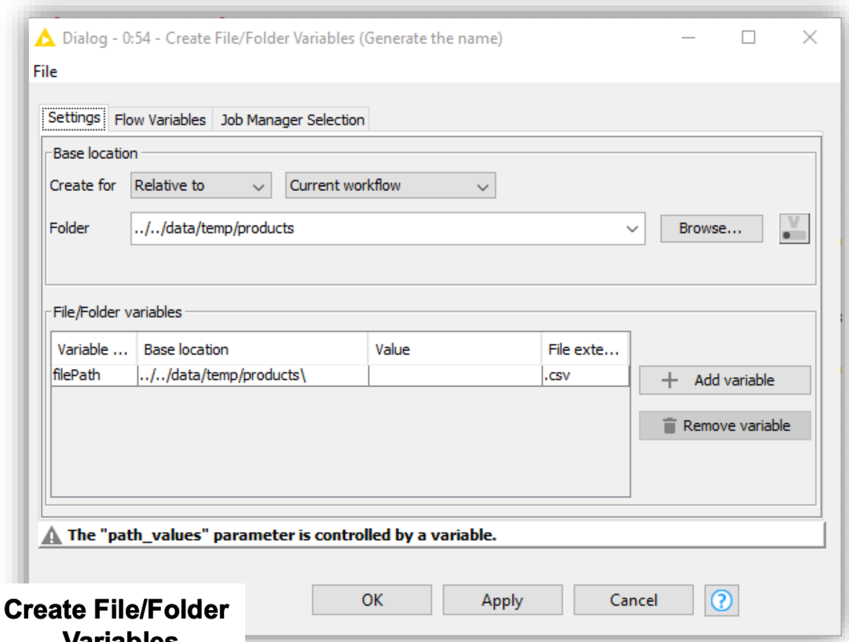
Example: Writing Aggregated Files

- Group Loop Start → Variable Loop End
- Group data by specific column values
- Iterate over all groups of data
- Create an appropriate file name
- Write grouped data to tables with new file name


Create File/Folder Variables
Create path varible with current group as file name
Table Reader
Read entire table
Group Loop Start
Records of one group per iteration
CSV Writer
Write table
Variable Loop End
Collect variables and end loop

Create File/Folder Variables

- Creates one or multiple path flow variable(s) pointing to files / folders
- Inputs:
  - Base location
  - Flow variable name(s)
  - Value (file name or path relative to base location)
  - File extension (optional)
- Output variables can be used to control the output location in writer nodes.



Dialog - 0:54 - Create File/Folder Variables (Generate the name)

File

Base location

Create for: Relative to — Current workflow

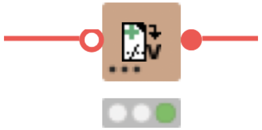
Folder: ../../data/temp/products

File/Folder variables

Variable ...	Base location	Value	File exte...
filePath	../../data/temp/products\		.csv

The "path_values" parameter is controlled by a variable.

Create File/Folder Variables



Example: Writing Multiple Excel Sheets

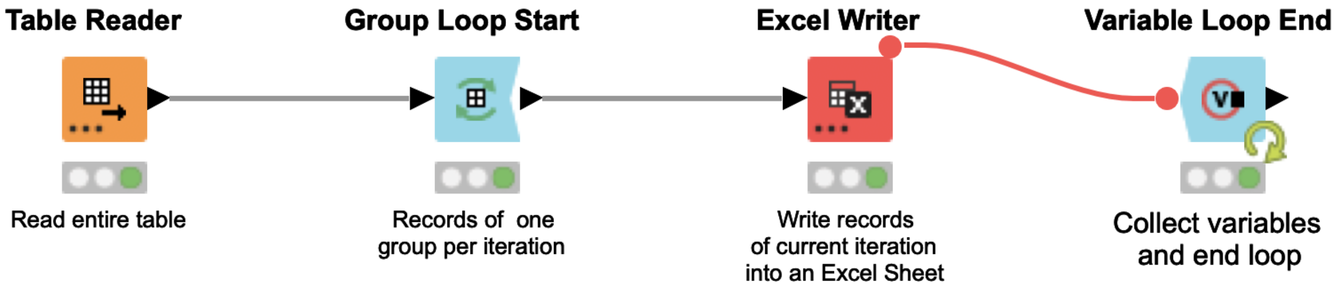


Table Reader
Read entire table
Group Loop Start
Records of one group per iteration
Excel Writer
Write records of current iteration into an Excel Sheet
Variable Loop End
Collect variables and end loop

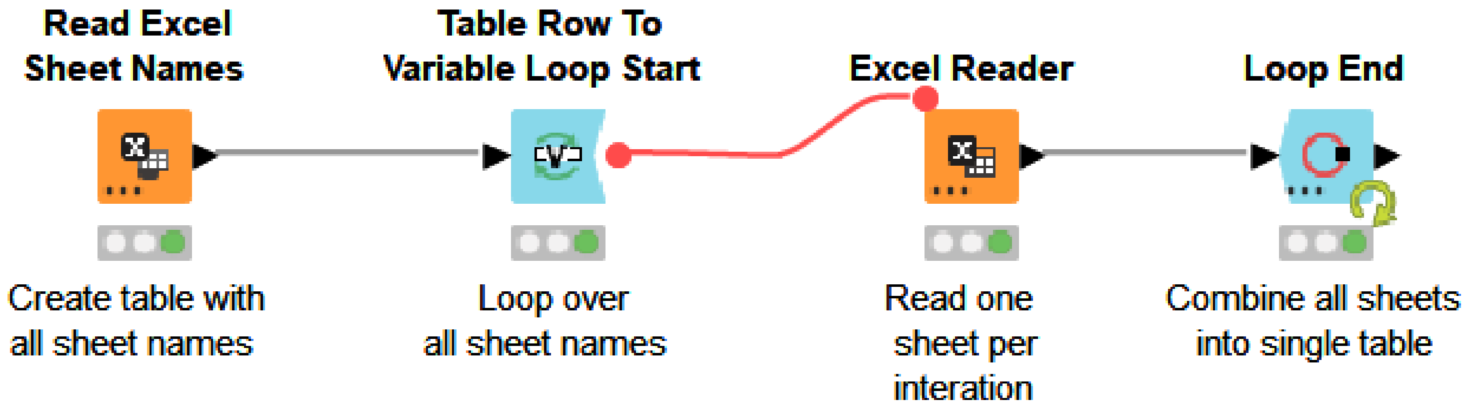
Workflow Control Exercise, Activity I

Goal: Build a loop that will create an Excel file with separate Excel sheets for the records of different products.

- Read the table CurrentDetailData.table (Table Reader node)
- Start a loop that handles the records for the different products in separate iterations (Group Loop Start node)
- For each product write one Excel sheet into a single Excel file (Excel Writer node)
- Close and execute the loop (Variable Loop End node)

Example: Reading Many Excel Sheets

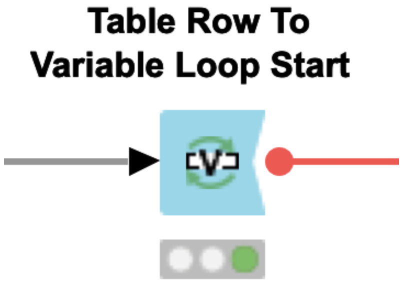
- List all sheet names of an Excel file
- Convert sheet name into a flow variable (1 sheet name per iteration)
- In each iteration, read the spreadsheet with the current sheet name
- Close the loop and collect the results

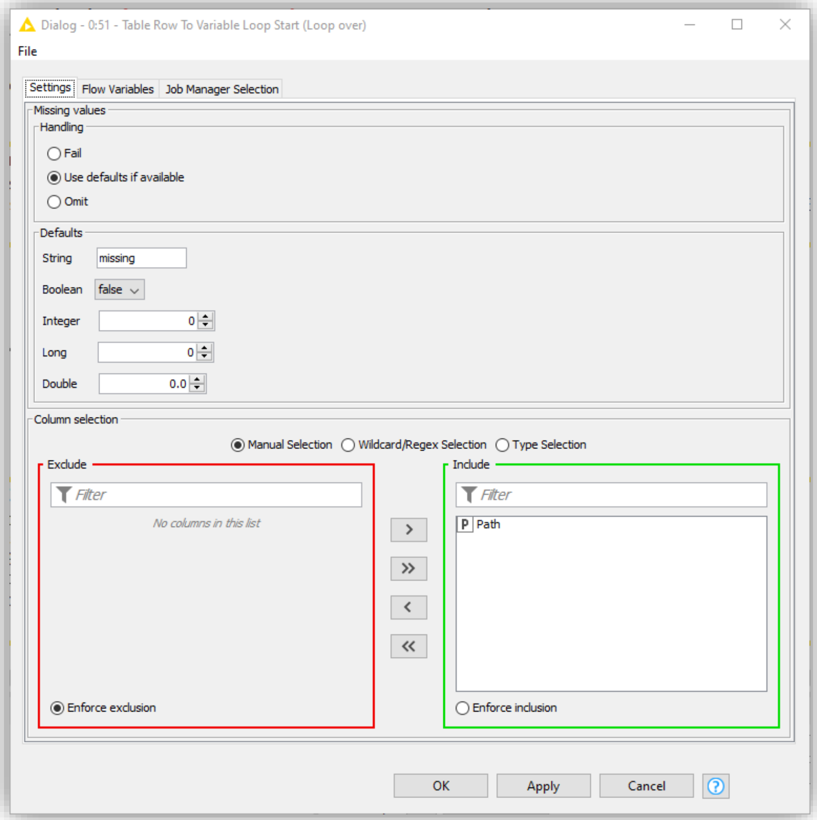


Read Excel Sheet Names
Table Row To Variable Loop Start
Excel Reader
Loop End
Create table with all sheet names
Loop over all sheet names
Read one sheet per interation
Combine all sheets into single table

Table Row to Variable Loop Start

- Similar to the
  Table Row to Variable node
- Each iteration of the loop converts the next row of the input table into Flow Variables
- Injects variables into other nodes to re-execute subflows with a progression of settings




Dialog - 0:51 - Table Row To Variable Loop Start (Loop over)

Missing values

Handling

○ Fail
● Use defaults if available
○ Omit

Defaults

String: missing
Boolean: false
Integer: 0
Long: 0
Double: 0.0

Column selection

● Manual Selection ○ Wildcard/Regex Selection ○ Type Selection

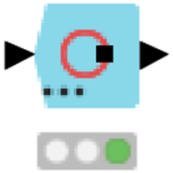
Exclude
Filter
No columns in this list
● Enforce exclusion

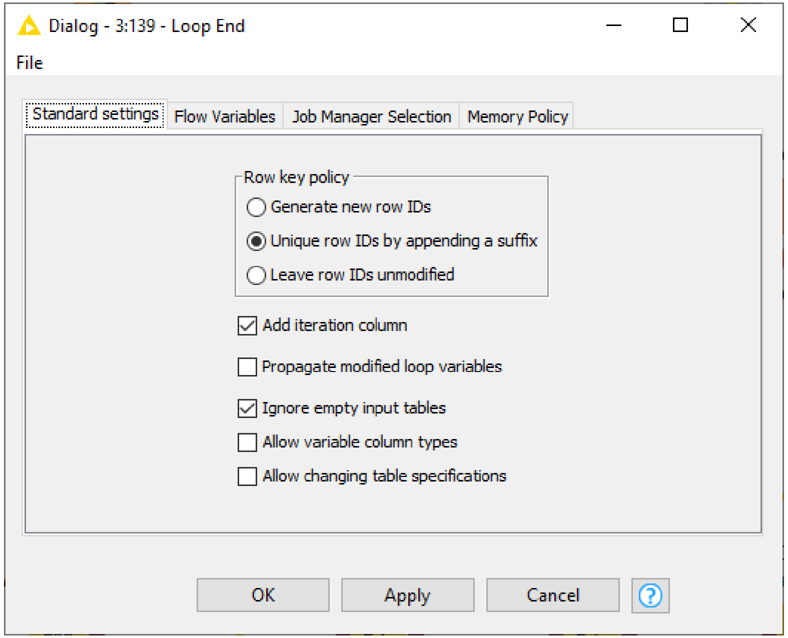
Include
Filter
P Path
○ Enforce inclusion

Loop End

- Can be used to end of a loop
- Collects the results of the different iterations by row-wise concatenation of the incoming tables
- Provides options to:
  - Add a column with the iteration number
  - Propagate modified loop variables
  - Allow variable column types
  - Allow changing table specifications

Loop End





Row key policy

○ Generate new row IDs

● Unique row IDs by appending a suffix

○ Leave row IDs unmodified

☑ Add iteration column

□ Propagate modified loop variables

☑ Ignore empty input tables

□ Allow variable column types

□ Allow changing table specifications

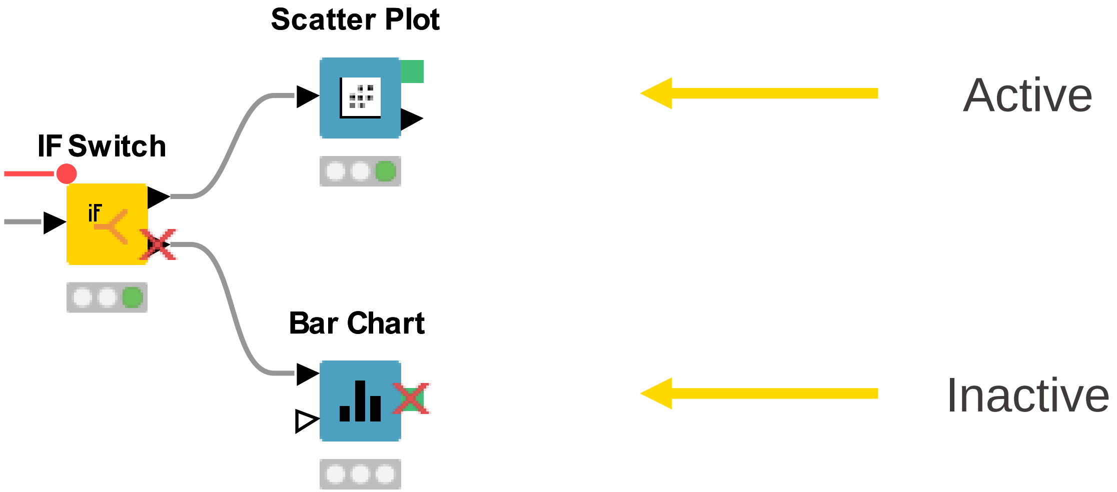
Workflow Control Exercise, Activity II

Goal: Create a loop that reads and concatenates all the sheets in an Excel file.

- Create a table that contains all sheet names of the Excel file created in Activity I (Read Excel Sheet Names node)
- Start a loop that iterates over the sheet names (Table Row to Variable Loop Start node)
- Read the Excel sheet with the sheet name in the current iteration (Excel Reader node)
- Close the loop and concatenate the tables from the different iterations (Loop End node)

Switches

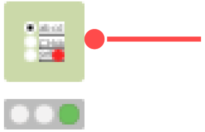
- A switch allows you to selectively activate branches of a workflow
- Inactive branches are marked with a red x on their output ports. Inactive nodes propagate down stream.

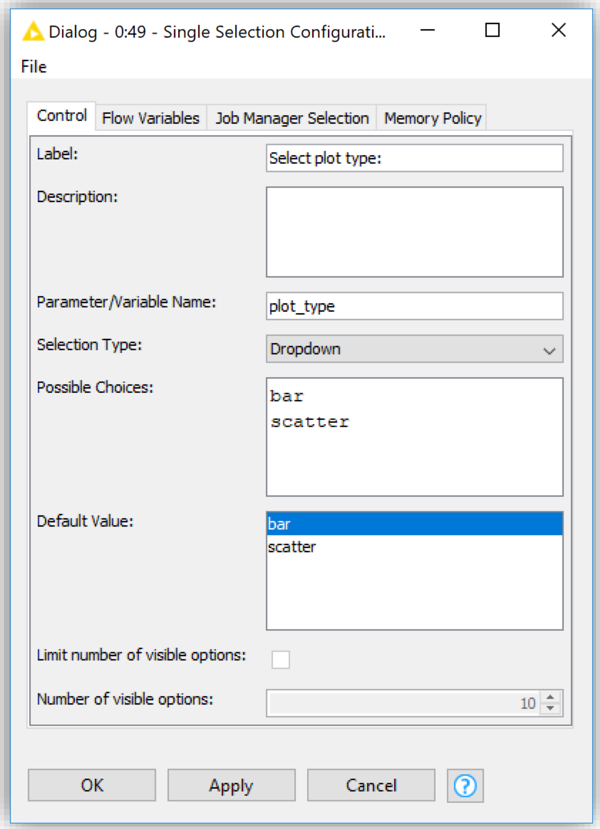


Single Selection Configuration

- Configuration: Select single value from list of Strings
- Returns selection as string type Flow Variable
- Choose between different layout options (dropdown, radio buttons...)

Single Selection Configuration





Label: Select plot type:

Description:

Parameter/Variable Name: plot_type

Selection Type: Dropdown

Possible Choices:

```
bar
scatter
```

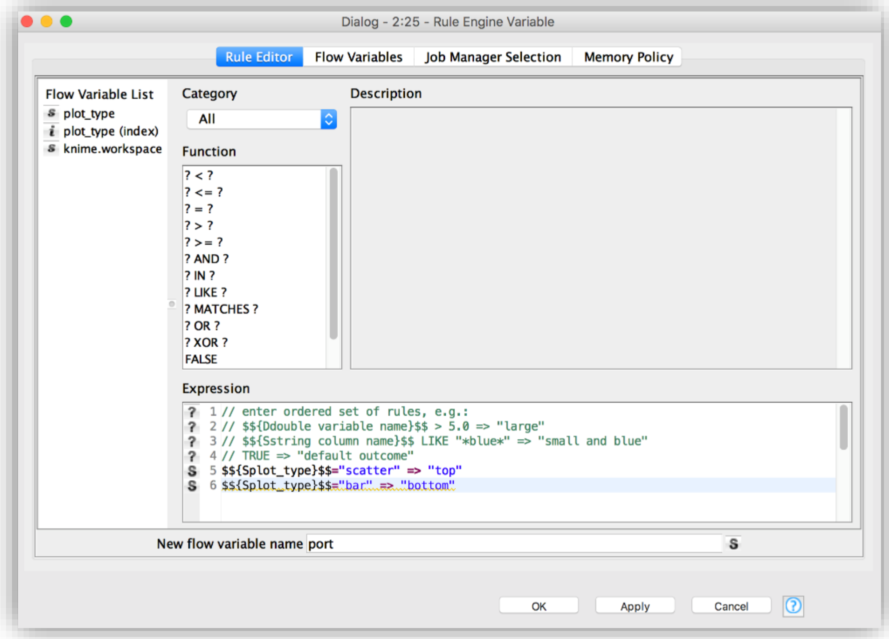
Default Value: bar
scatter

Limit number of visible options: □

Number of visible options: 10

Rule Engine/Rule Engine Variable

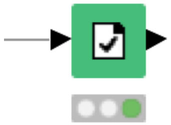
- Defines custom logic for using simple rules.
- Rules like: <Antecedent> => <Consequence>
  - (1=1 => “true”)
- May be used in Flow Variables or tables
- Easiest way to encode logic for switches



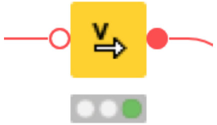
```
// enter ordered set of rules, e.g.:
// $${Ddouble variable name}$$ > 5.0 => "large"
// $${Sstring column name}$$ LIKE "*blue*" => "small and blue"
// TRUE => "default outcome"
$${Splot_type}$$="scatter" => "top"
$${Splot_type}$$="bar" => "bottom"
```

New flow variable name: port

Rule Engine

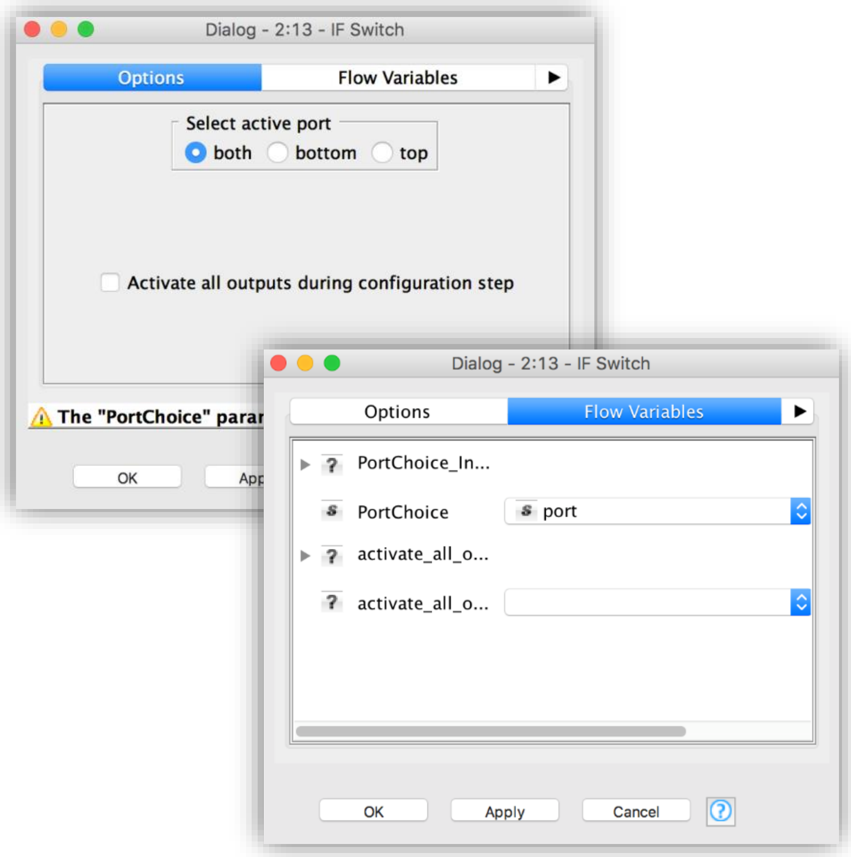


Rule Engine Variable



If Switch

- Controls which branches of your workflow are active programmatically
- Controlled with a Flow Variable, setting the value to the literal Strings: “top”, “bottom”, “both”
- May be used in Flow Variables or tables (different nodes)



Dialog - 2:13 - IF Switch

Select active port

● both ○ bottom ○ top

□ Activate all outputs during configuration step

⚠ The "PortChoice" parar

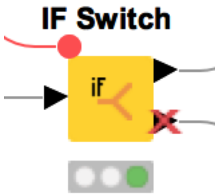
Dialog - 2:13 - IF Switch

PortChoice_In...

PortChoice: port

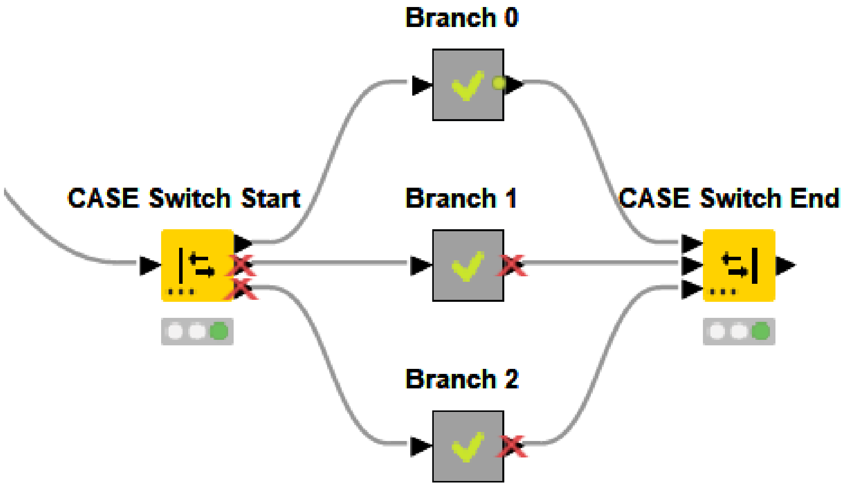
activate_all_o...

activate_all_o...



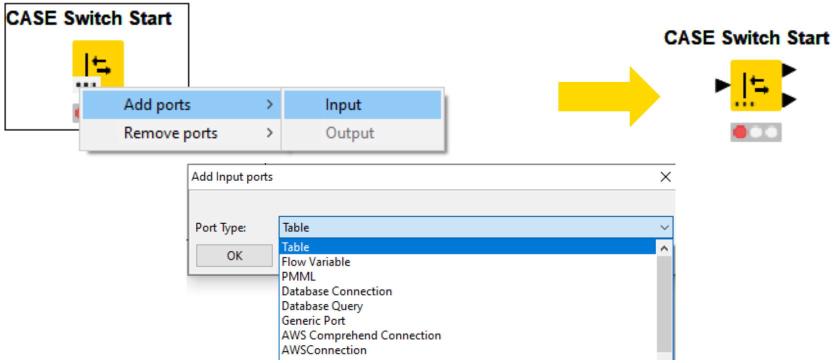
Case Switch Start & End

- Similar to If-Switch: Takes data from single input port and passes it to the active output port
- Nodes connected to inactive branches are not executed

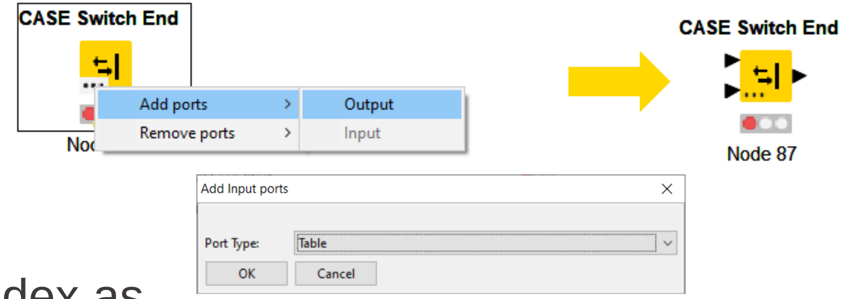


Case Switch Start & End

- Case Switch Start
  - Add an input port with a specific type (e.g., Data)
  - → Two output ports are also added
  - → Additional output ports can be added
- Case Switch End
  - Add an output port with a specific type (e.g., Data)
  - → Two input ports are also added
  - → Additional input ports can be added



Port Type: Table



Port Type: Table

- Configure via node dialog, or pass port index as Flow Variable
  - From the top, 0, 1, 2, ... (however many ports there are)

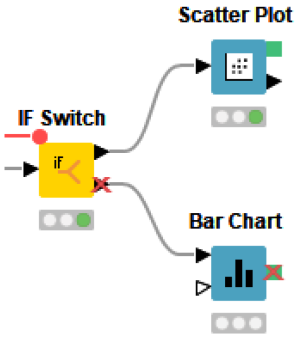
The Difference between Loops and Switches

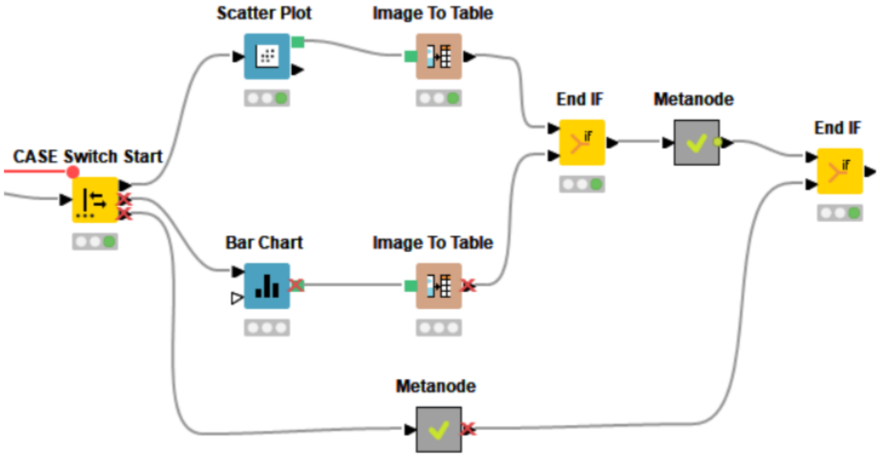
Loops

- The Loop Start is connected to the Loop End node; they form a pair.
- A loop iterates over a workflow part.

Switches

- A Switch Start can be used without a corresponding Switch End.
  They can also be combined.

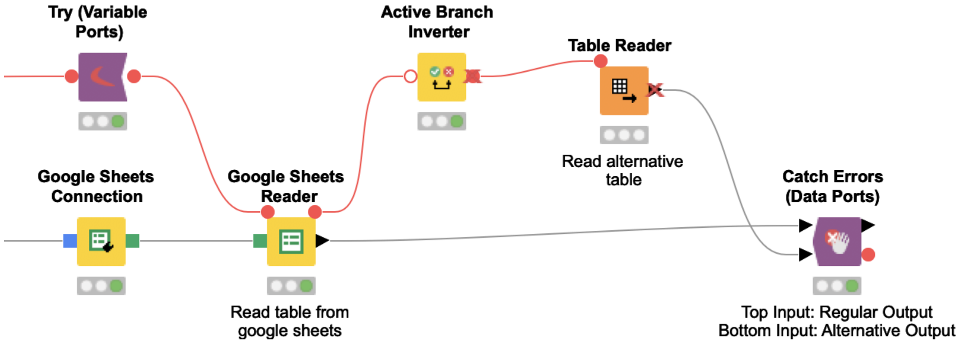




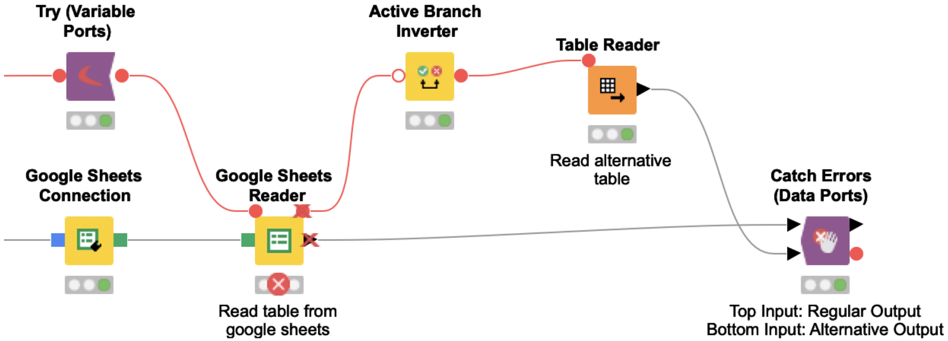
Try-Catch

- A way to catch errors in workflows
- Useful when it is hard to know if a node will execute (for example,
  when reading from a Google Sheet)
- KNIME tries to execute the nodes, but if it fails will fall back to
  an alternative branch

Regular Execution



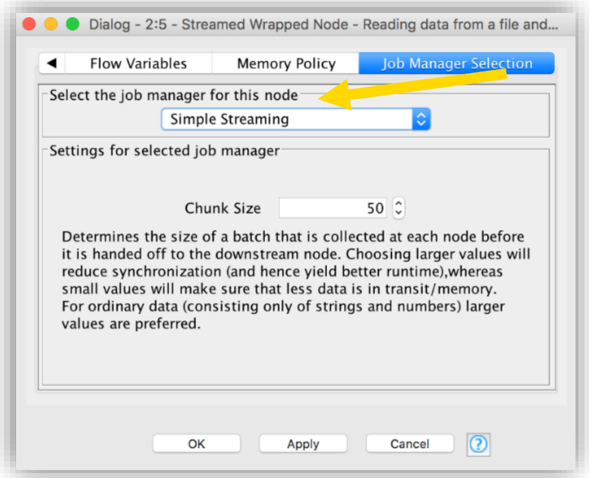
Alternative Execution



Streaming

- Standard execution: Node by node. The node processes all data, finishes, then passes the data to the next node, etc.
- Streaming: Nodes executed concurrently, each nodes passes data to the next as soon as it is available, i.e. before node is fully executed
  - Faster execution, esp. for reading/preprocessing data
- Install KNIME Streaming Execution (Beta) extension
- Create Component -> Configure -> Job Manager Selection -> Simple Streaming
  - Not available for all nodes (show in node repository)
  - Can only execute entire metanode, not individual nodes
  - Intermediate results not available since nothing is cached

Streaming



Dialog - 2:5 - Streamed Wrapped Node - Reading data from a file and...

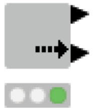
Select the job manager for this node: Simple Streaming

Settings for selected job manager

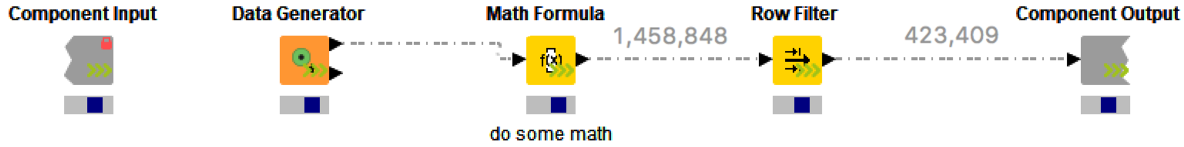
Chunk Size: 50

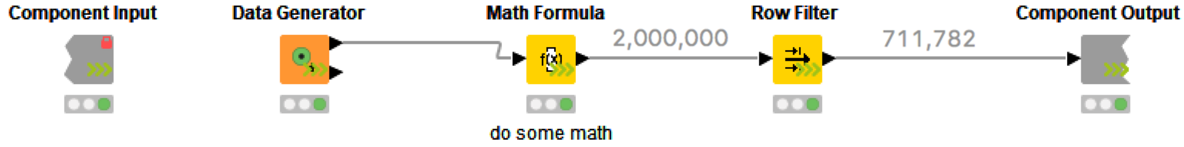
Determines the size of a batch that is collected at each node before it is handed off to the downstream node. Choosing larger values will reduce synchronization (and hence yield better runtime),whereas small values will make sure that less data is in transit/memory. For ordinary data (consisting only of strings and numbers) larger values are preferred.

Streamed Component - Reading data from a file and process



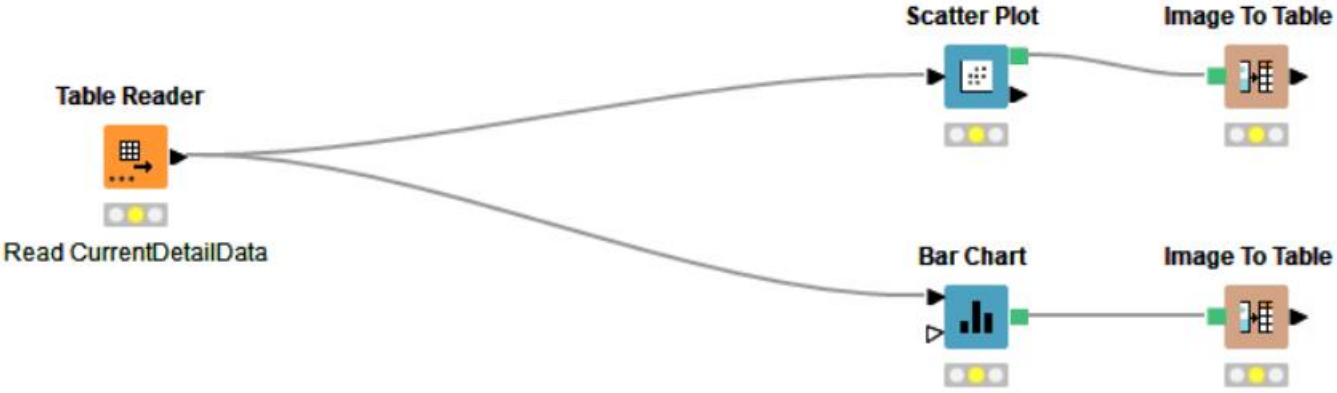
Sub Workflow as Component.
To open it:
- right click > Component > Open
- Ctrl + Double Click





Workflow Control Exercise, Activity III

- Extend the workflow below with a switch that only creates one type of visualization
  - Create a Single Selection Configuration node with the possible values "scatter" and "bar"
  - Use the CASE Switch Data (Start) that activates the top or the middle branch depending on the selection scatter/bar (Use the "...(index)" flow variable to define the active port)
  - Combine the outputs of the two branches with the CASE Switch Data (End) node



Introduction to Data Science

Churn Prediction

CRM System
Data about your customer

- Demographics
- Behavior
- Revenues

Model

- Churn Prediction
- Upselling Likelihood
- Product Propensity /NBO
- Campaign Management
- Customer Segmentation
- …

Customer Segmentation

CRM System

Data about your customer

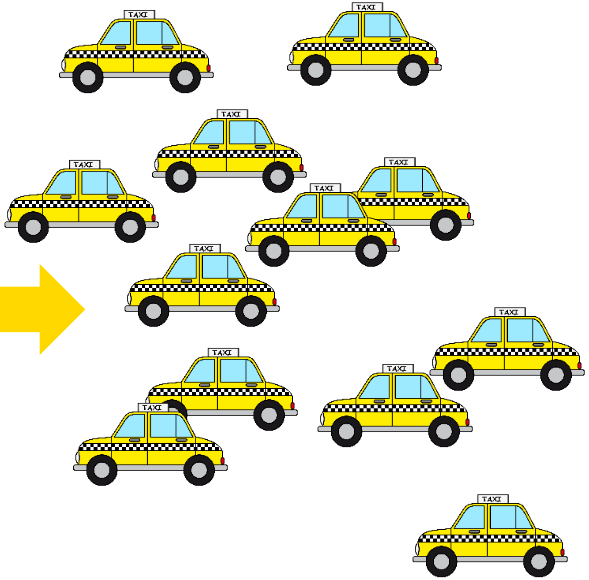
- Demographics
- Behavior
- Revenues

Model

- Churn Prediction
- Upselling Likelihood
- Product Propensity /NBO
- Campaign Management
- Customer Segmentation
- …

Demand Prediction

How many taxis do I need in NYC on Wednesday at 12:00?



Recommendation Engines / Market Basket Analysis

Model

$$Rule:\ X \Rightarrow Y$$

$$Support = \frac{frq(X,Y)}{N}$$

$$Confidence = \frac{frq(X,Y)}{frq(X)}$$

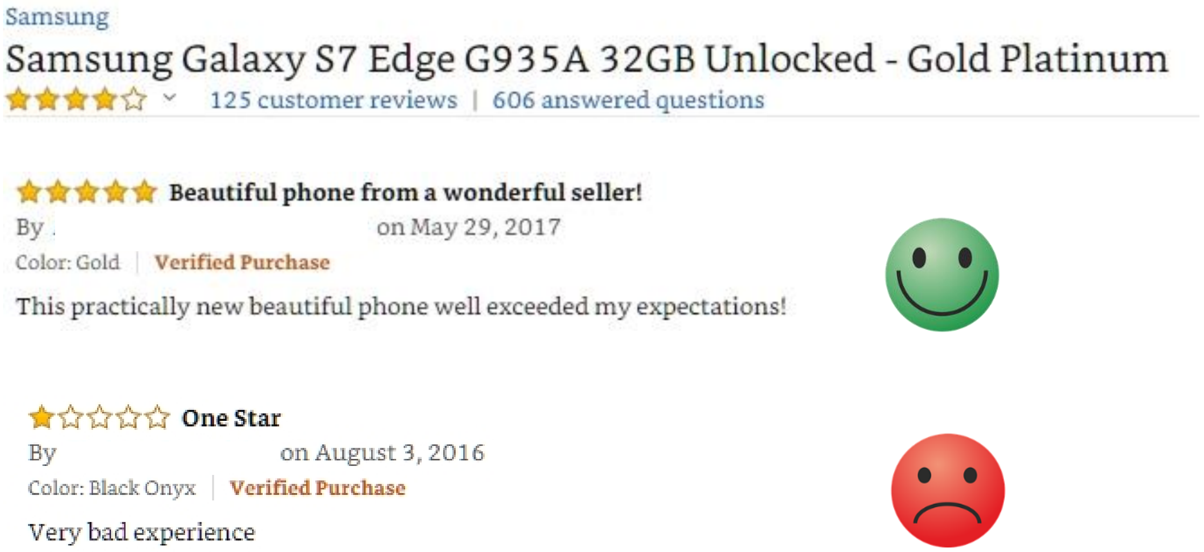
$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

Recommendation

IF

=>

Sentiment Analysis



Samsung

Samsung Galaxy S7 Edge G935A 32GB Unlocked - Gold Platinum

★★★★☆ 125 customer reviews | 606 answered questions

★★★★★ **Beautiful phone from a wonderful seller!**

By on May 29, 2017

Color: Gold | Verified Purchase

This practically new beautiful phone well exceeded my expectations!

🙂

★☆☆☆☆ **One Star**

By on August 3, 2016

Color: Black Onyx | Verified Purchase

Very bad experience

☹

Today's Challenge – Analyze Some Customer Data

Initial situation:

- Customer data in two datasets:
  - Phone usage
  - Contract information
- Column “Phone” is in both datasets
- Column “churn” encodes whether a customer is happy

Goal:

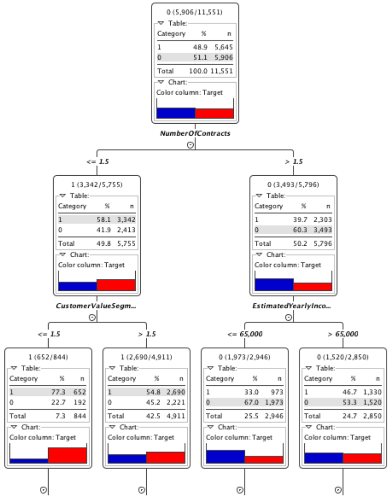
- Find rules that describe happy and unhappy customers by
  - training a decision tree model
  - calculating aggregations (optional)

Classification

Predict *nominal* outcomes on existing data (supervised)

- Applications
  - Churn analysis (yes/no)
  - Chemical activity (active/inactive)
  - Spam detection (spam/not spam)
  - Optical character recognition (A-Z)

- Methods
  - Decision Trees
  - Neural Networks
  - Naïve Bayes
  - Logistic Regression



0 (5,906/11,551)

Table:

Category	%	n
1	48.9	5,645
0	51.1	5,906
Total	100.0	11,551

Chart:
Color column: Target

NumberOfContracts

<= 1.5

1 (3,342/5,755)

Table:

Category	%	n
1	58.1	3,342
0	41.9	2,413
Total	49.8	5,755

Chart:
Color column: Target

CustomerValueSegm...

<= 1.5

1 (652/844)

Table:

Category	%	n
1	77.3	652
0	22.7	192
Total	7.3	844

Chart:
Color column: Target

> 1.5

1 (2,690/4,911)

Table:

Category	%	n
1	54.8	2,690
0	45.2	2,221
Total	42.5	4,911

Chart:
Color column: Target

> 1.5

0 (3,493/5,796)

Table:

Category	%	n
1	39.7	2,303
0	60.3	3,493
Total	50.2	5,796

Chart:
Color column: Target

EstimatedYearlyInco...

<= 65,000

0 (1,973/2,946)

Table:

Category	%	n
1	33.0	973
0	67.0	1,973
Total	25.5	2,946

Chart:
Color column: Target

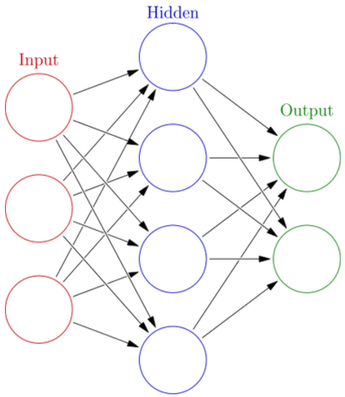
> 65,000

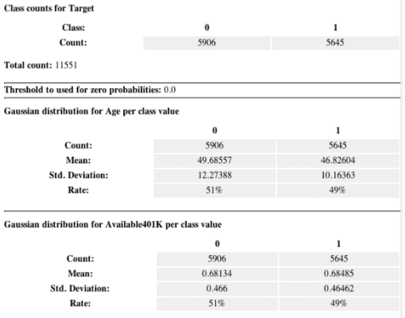
0 (1,520/2,850)

Table:

Category	%	n
1	46.7	1,330
0	53.3	1,520
Total	24.7	2,850

Chart:
Color column: Target





Class counts for Target

Class:	0	1
Count:	5906	5645

Total count: 11551

Threshold to used for zero probabilities: 0.0

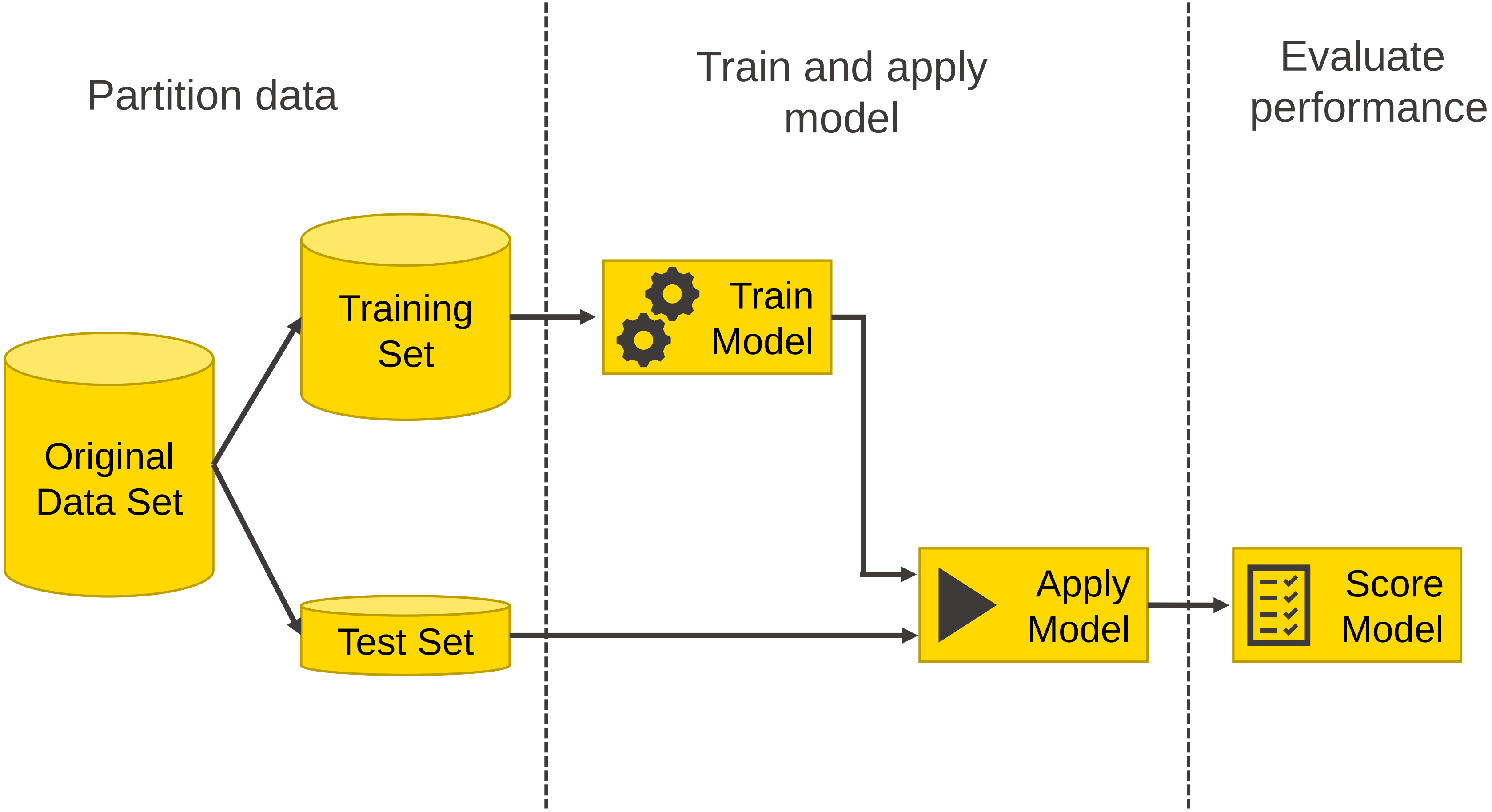
Gaussian distribution for Age per class value

	0	1
Count:	5906	5645
Mean:	49.68557	46.82604
Std. Deviation:	12.27388	10.16363
Rate:	51%	49%

Gaussian distribution for Available401K per class value

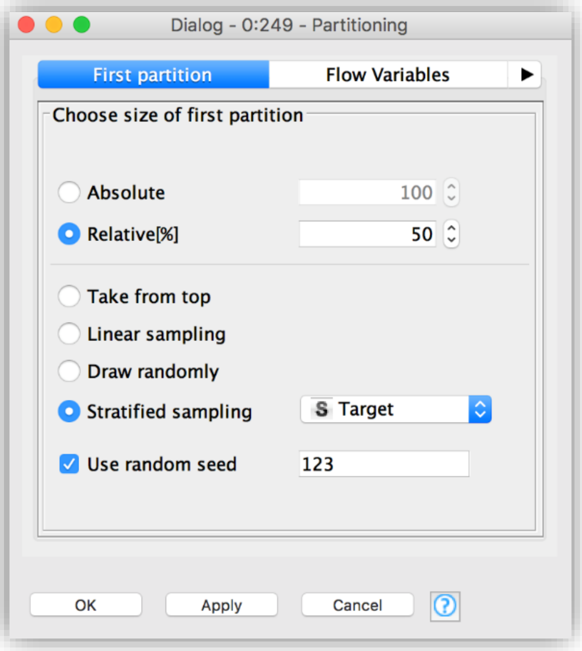
	0	1
Count:	5906	5645
Mean:	0.68134	0.68485
Std. Deviation:	0.466	0.46462
Rate:	51%	49%

Data Mining: Process Overview



Partitioning

- Use to split data into training and evaluation sets
  - Partition by count (e.g. 10 rows) or fraction (e.g. 10%)
  - Sample by a variety of methods; random, linear, stratified



Dialog - 0:249 - Partitioning

Choose size of first partition

○ Absolute 100

● Relative[%] 50

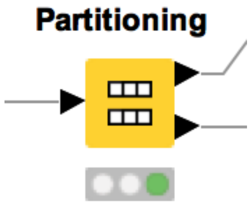
○ Take from top

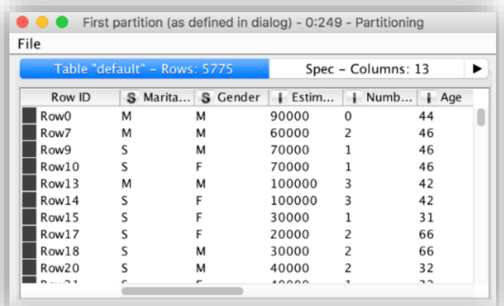
○ Linear sampling

○ Draw randomly

● Stratified sampling S Target

☑ Use random seed 123

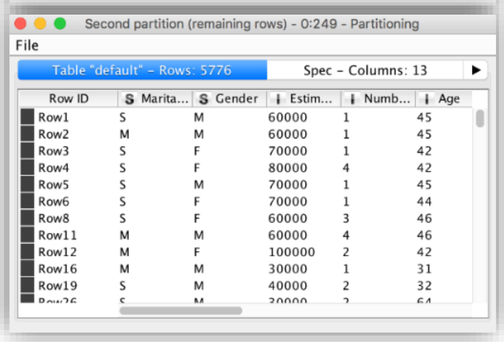




First partition (as defined in dialog) - 0:249 - Partitioning

Table "default" – Rows: 5775 | Spec – Columns: 13

Row ID	S Marita...	S Gender	I Estim...	I Numb...	I Age
Row0	M	M	90000	0	44
Row7	M	M	60000	2	46
Row9	S	M	70000	1	46
Row10	S	F	70000	1	46
Row13	M	M	100000	3	42
Row14	S	F	100000	3	42
Row15	S	F	30000	1	31
Row17	S	F	20000	2	66
Row18	S	M	30000	2	66
Row20	S	M	40000	2	32



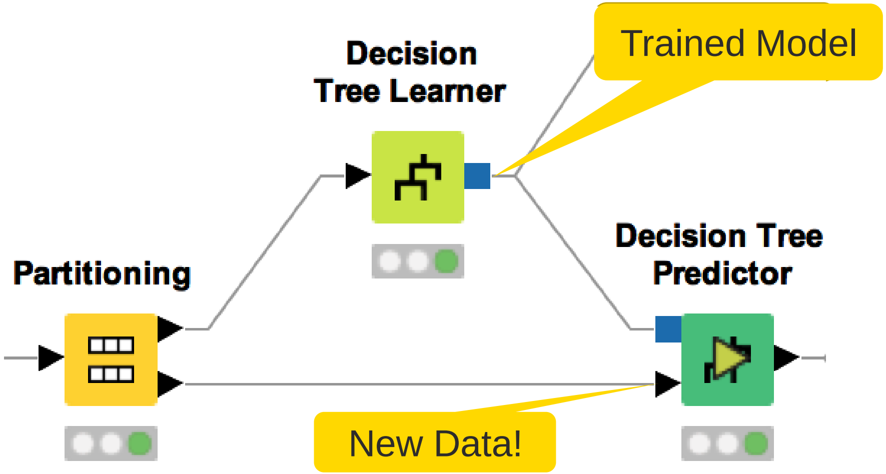
Second partition (remaining rows) - 0:249 - Partitioning

Table "default" – Rows: 5776 | Spec – Columns: 13

Row ID	S Marita...	S Gender	I Estim...	I Numb...	I Age
Row1	S	M	60000	1	45
Row2	M	M	60000	1	45
Row3	S	F	70000	1	42
Row4	S	F	80000	4	42
Row5	S	M	70000	1	45
Row6	S	F	70000	1	44
Row8	S	F	60000	3	46
Row11	M	M	60000	4	46
Row12	M	F	100000	2	42
Row16	M	M	30000	1	31
Row19	S	M	40000	2	32

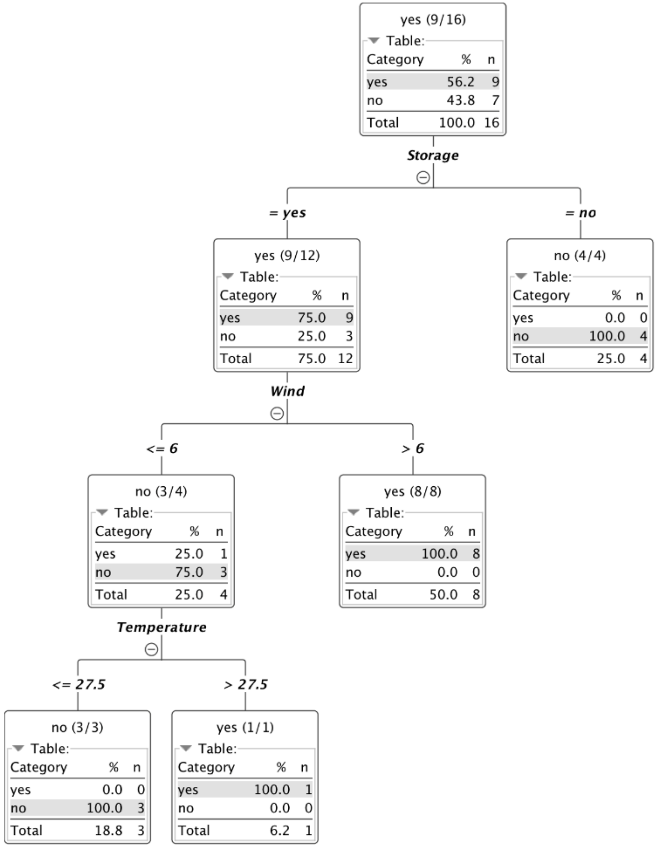
Learner-Predictor Motif

- Most data mining approaches in KNIME use a Learner-predictor motif.
- The Learner node trains the model with its input data.
- The Predictor node applies the model to a different subset of data.



Goal: A Decision Tree

Outlook	Wind	Temp	Storage	Sailing
sunny	3	30	yes	yes
sunny	3	25	yes	no
rain	12	15	yes	yes
overcast	15	2	no	no
rain	16	25	yes	yes
sunny	14	18	yes	yes
rain	3	5	no	no
sunny	9	20	yes	yes
overcast	14	5	no	no
sunny	1	7	no	no
rain	4	25	yes	no
rain	14	24	yes	yes
sunny	11	20	yes	yes
sunny	2	18	yes	no
overcast	8	22	yes	yes
overcast	13	24	yes	yes



yes (9/16)

Table:

Category	%	n
yes	56.2	9
no	43.8	7
Total	100.0	16

Storage

= yes

yes (9/12)

Table:

Category	%	n
yes	75.0	9
no	25.0	3
Total	75.0	12

= no

no (4/4)

Table:

Category	%	n
yes	0.0	0
no	100.0	4
Total	25.0	4

Wind

<= 6

no (3/4)

Table:

Category	%	n
yes	25.0	1
no	75.0	3
Total	25.0	4

> 6

yes (8/8)

Table:

Category	%	n
yes	100.0	8
no	0.0	0
Total	50.0	8

Temperature

<= 27.5

no (3/3)

Table:

Category	%	n
yes	0.0	0
no	100.0	3
Total	18.8	3

> 27.5

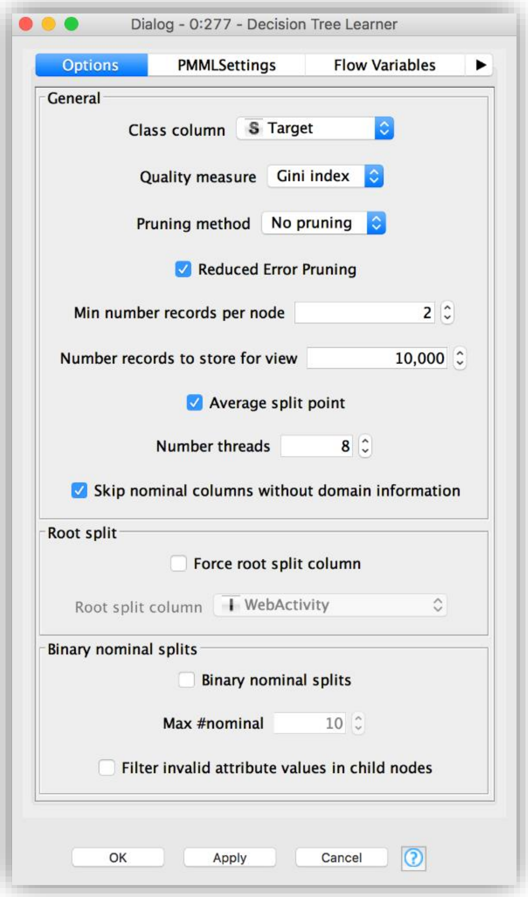
yes (1/1)

Table:

Category	%	n
yes	100.0	1
no	0.0	0
Total	6.2	1

Decision Tree Learner

Decision
Tree Learner


Dialog - 0:277 - Decision Tree Learner

General

Class column: S Target

Quality measure: Gini index

Pruning method: No pruning

☑ Reduced Error Pruning

Min number records per node: 2

Number records to store for view: 10,000

☑ Average split point

Number threads: 8

☑ Skip nominal columns without domain information

Root split

□ Force root split column

Root split column: WebActivity

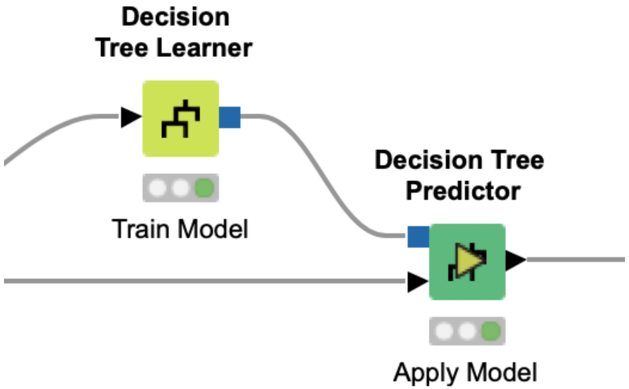
Binary nominal splits

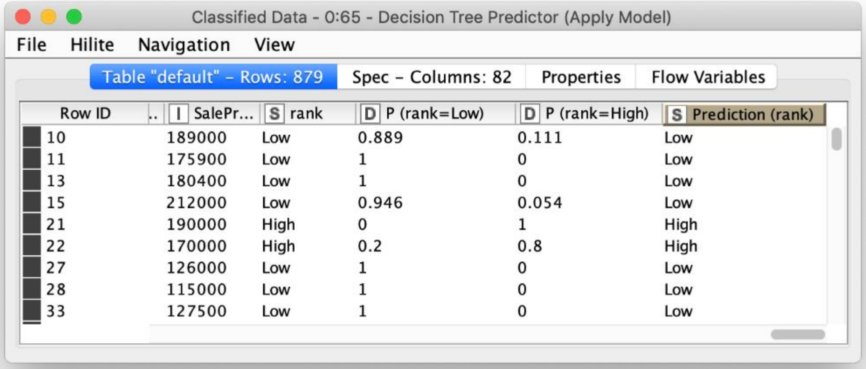
□ Binary nominal splits

Max #nominal: 10

□ Filter invalid attribute values in child nodes

Applying the Model – What are the Outputs?

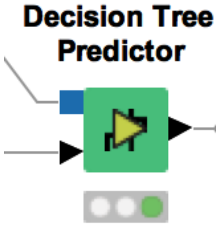


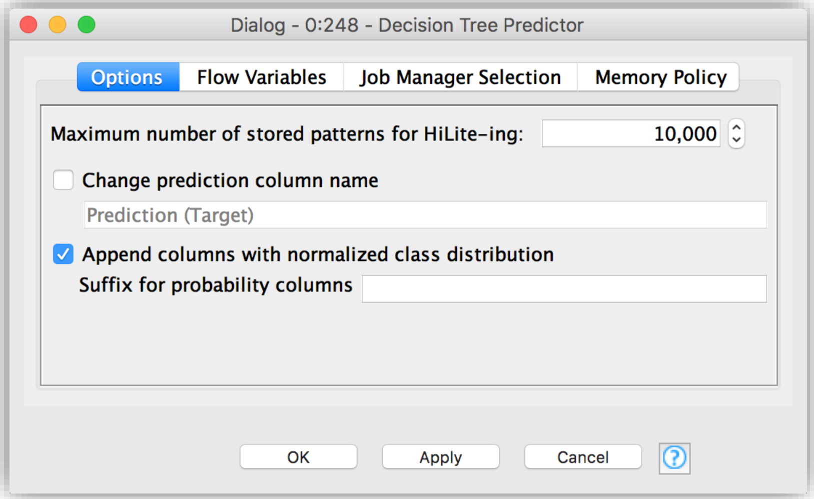

Classified Data - 0:65 - Decision Tree Predictor (Apply Model)
File
Hilite
Navigation
View
Table "default" – Rows: 879
Spec – Columns: 82
Properties
Flow Variables

Row ID	..	I SalePr...	S rank	D P (rank=Low)	D P (rank=High)	S Prediction (rank)
10		189000	Low	0.889	0.111	Low
11		175900	Low	1	0	Low
13		180400	Low	1	0	Low
15		212000	Low	0.946	0.054	Low
21		190000	High	0	1	High
22		170000	High	0.2	0.8	High
27		126000	Low	1	0	Low
28		115000	Low	1	0	Low
33		127500	Low	1	0	Low

Decision Tree Predictor

- Takes a decision tree model and apply it to new data
- Check the box to append class probabilities





Maximum number of stored patterns for HiLite-ing: 10,000

□ Change prediction column name

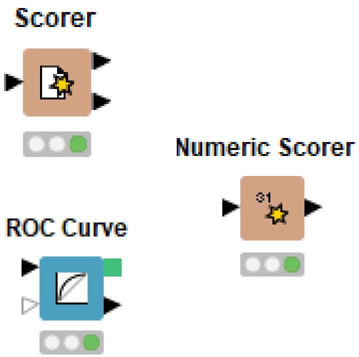
Prediction (Target)

☑ Append columns with normalized class distribution

Suffix for probability columns

Evaluation Metrics

- Why evaluation metrics?
  - Quantify the power of the model as a classifier/predictor
  - Compare model configurations and/or models, and select the best performing one
  - Obtain the expected performance of the model for new data
- Different model evaluation techniques are available for
  - Classification/regression models
  - Imbalanced/balanced target class distributions



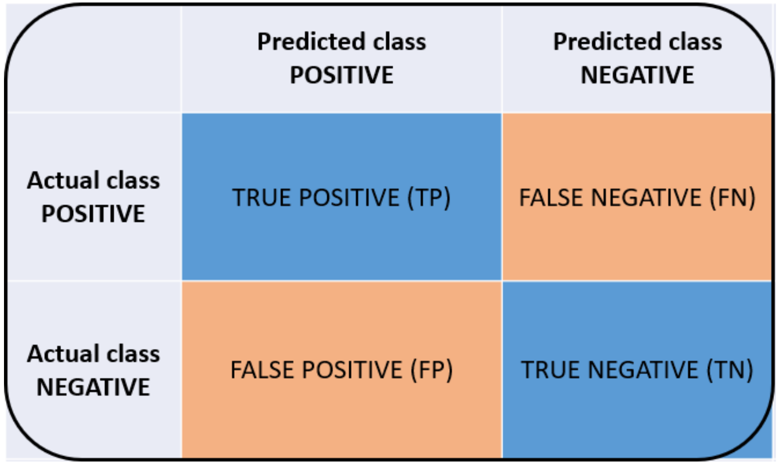
Overall Accuracy

$$\boldsymbol{Overall\ Accuracy = \frac{\#\ Correct\ Classifications}{\#\ All\ Events}}$$

- The proportion of correct classifications
- Downsides:
  - Only considers the performance in general and not for the different target classes
  - Therefore, not informative when the target class distribution is unbalanced

Confusion Matrix

Arbitrarily define one target class as POSITIVE and the remaining class(es) as NEGATIVE



	Predicted class POSITIVE	Predicted class NEGATIVE
Actual class POSITIVE	TRUE POSITIVE (TP)	FALSE NEGATIVE (FN)
Actual class NEGATIVE	FALSE POSITIVE (FP)	TRUE NEGATIVE (TN)

TRUE POSITIVE (**TP**): Actual and predicted class is positive

TRUE NEGATIVE (**TN**): Actual and predicted class is negative

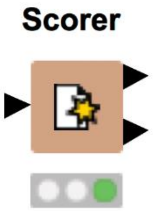
FALSE NEGATIVE (**FN**): Actual class is positive and predicted negative

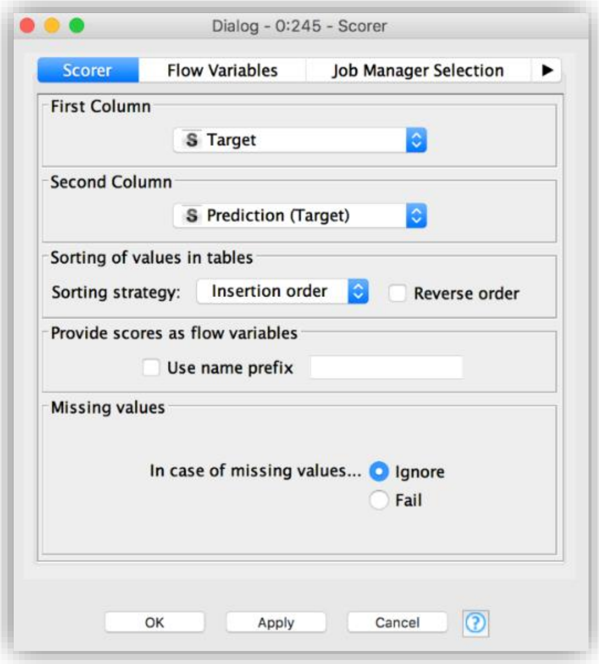
FALSE POSITIVE (**FP**): Actual class is negative and predicted positive

- Use these four statistics to calculate other evaluation metrics, such as *overall accuracy*, *true positive rate,* and *false positive rate*

Scorer

Compare predicted results to known truth in order to evaluate model quality




Dialog - 0:245 - Scorer

First Column

S Target

Second Column

S Prediction (Target)

Sorting of values in tables

Sorting strategy: Insertion order □ Reverse order

Provide scores as flow variables

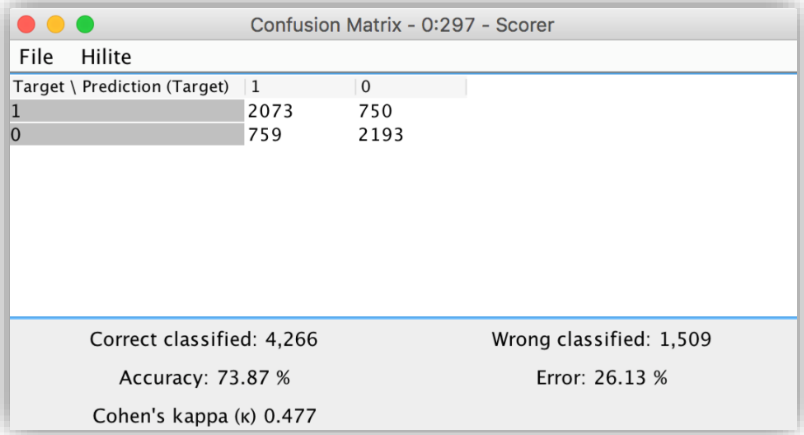
□ Use name prefix

Missing values

In case of missing values... ● Ignore ○ Fail

Scorer

Confusion matrix shows the distribution of model errors



Target \ Prediction (Target)	1	0
1	2073	750
0	759	2193

Correct classified: 4,266

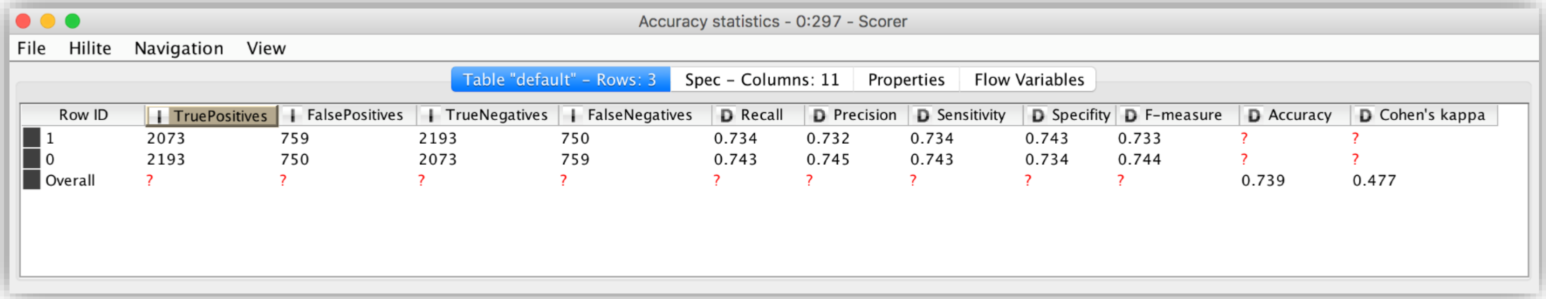
Wrong classified: 1,509

Accuracy: 73.87 %

Error: 26.13 %

Cohen's kappa (κ) 0.477

An accuracy statistics table provides a detailed analysis of model quality



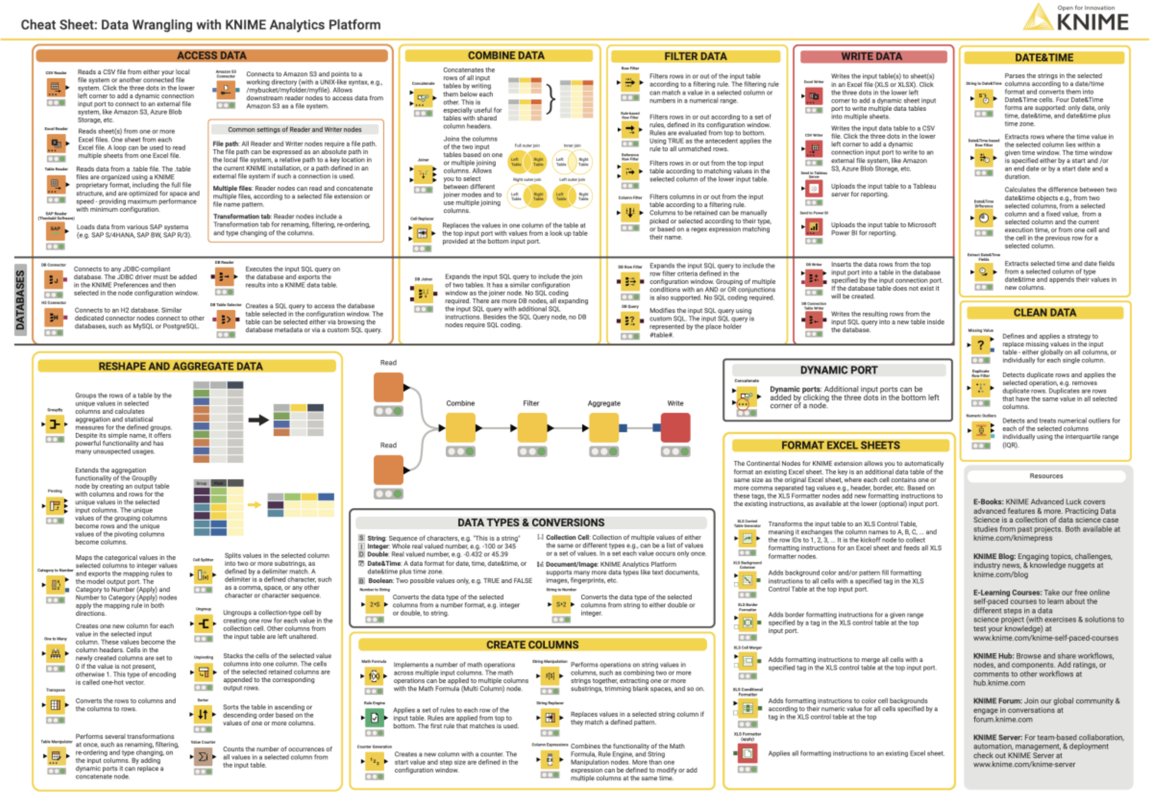
Accuracy statistics - 0:297 - Scorer
File Hilite Navigation View
Table "default" - Rows: 3
Spec - Columns: 11
Properties
Flow Variables

Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specifity	F-measure	Accuracy	Cohen's kappa
1	2073	759	2193	750	0.734	0.732	0.734	0.743	0.733	?	?
0	2193	750	2073	759	0.743	0.745	0.743	0.734	0.744	?	?
Overall	?	?	?	?	?	?	?	?	?	0.739	0.477

Exercise: 13_Training_a_Churn_Prediction_Model

- Read the CallsData.xls and ContractData.csv files
- Join the two data tables based on the columns “Area Code” and “Phone”
- Change the data type of the columns “Area Code” and “Churn” to string
- Partition the data into a training set and a test set
- Train a decision tree to detect customers that are likely to churn
- Apply the mode to the test set and evaluate the model performance

Data Wrangler Cheat Sheet



Cheat Sheet: Data Wrangling with KNIME Analytics Platform

Open for Innovation KNIME

ACCESS DATA

CSV Reader
Reads a CSV file from either your local file system or another connected file system. Click the three dots in the lower left corner to add a dynamic connection input port to connect to an external file system, like Amazon S3, Azure Blob Storage, etc.

Excel Reader
Reads sheet(s) from one or more Excel files. One sheet from each Excel file. A loop can be used to read multiple sheets from one Excel file.

Table Reader
Reads data from a .table file. The .table files are organized using a KNIME proprietary format, including the full file structure, and are optimized for space and speed - providing maximum performance with minimum configuration.

SAP Reader (Theobald Software)
Loads data from various SAP systems (e.g. SAP S/4HANA, SAP BW, SAP R/3).

Amazon S3 Connector
Connects to Amazon S3 and points to a working directory (with a UNIX-like syntax, e.g., /mybucket/myfolder/myfile). Allows downstream reader nodes to access data from Amazon S3 as a file system.

Common settings of Reader and Writer nodes

File path: All Reader and Writer nodes require a file path. The file path can be expressed as an absolute path in the local file system, a relative path to a key location in the current KNIME installation, or a path defined in an external file system if such a connection is used.

Multiple files: Reader nodes can read and concatenate multiple files, according to a selected file extension or file name pattern.

Transformation tab: Reader nodes include a Transformation tab for renaming, filtering, re-ordering, and type changing of the columns.

COMBINE DATA

Concatenate
Concatenates the rows of all input tables by writing them below each other. This is especially useful for tables with shared column headers.

Joiner
Joins the columns of the two input tables based on one or multiple joining columns. Allows you to select between different joiner modes and to use multiple joining columns.

Cell Replacer
Replaces the values in one column of the table at the top input port with values from a look up table provided at the bottom input port.

FILTER DATA

Row Filter
Filters rows in or out of the input table according to a filtering rule. The filtering rule can match a value in a selected column or numbers in a numerical range.

Rule-based Row Filter
Filters rows in or out according to a set of rules, defined in its configuration window. Rules are evaluated from top to bottom. Using TRUE as the antecedent applies the rule to all unmatched rows.

Reference Row Filter
Filters rows in or out from the top input table according to matching values in the selected column of the lower input table.

Column Filter
Filters columns in or out from the input table according to a filtering rule. Columns to be retained can be manually picked or selected according to their type, or based on a regex expression matching their name.

WRITE DATA

Excel Writer
Writes the input table(s) to sheet(s) in an Excel file (XLS or XLSX). Click the three dots in the lower left corner to add a dynamic sheet input port to write multiple data tables into multiple sheets.

CSV Writer
Writes the input data table to a CSV file. Click the three dots in the lower left corner to add a dynamic connection input port to write to an external file system, like Amazon S3, Azure Blob Storage, etc.

Send to Tableau Server
Uploads the input table to a Tableau server for reporting.

Send to Power BI
Uploads the input table to Microsoft Power BI for reporting.

DATE&TIME

String to Date&Time
Parses the strings in the selected columns according to a date/time format and converts them into Date&Time cells. Four Date&Time forms are supported: only date, only time, date&time, and date&time plus time zone.

Date&Time-based Row Filter
Extracts rows where the time value in the selected column lies within a given time window. The time window is specified either by a start and /or an end date or by a start date and a duration.

Date&Time Difference
Calculates the difference between two date&time objects e.g., from two selected columns, from a selected column and a fixed value, from a selected column and the current execution time, or from one cell and the cell in the previous row for a selected column.

Extract Date&Time Fields
Extracts selected time and date fields from a selected column of type date&time and appends their values in new columns.

DATABASES

DB Connector
Connects to any JDBC-compliant database. The JDBC driver must be added in the KNIME Preferences and then selected in the node configuration window.

H2 Connector
Connects to an H2 database. Similar dedicated connector nodes connect to other databases, such as MySQL or PostgreSQL.

DB Reader
Executes the input SQL query on the database and exports the results into a KNIME data table.

DB Table Selector
Creates a SQL query to access the database table selected in the configuration window. The table can be selected either via browsing the database metadata or via a custom SQL query.

DB Joiner
Expands the input SQL query to include the join of two tables. It has a similar configuration window as the joiner node. No SQL coding required. There are more DB nodes, all expanding the input SQL query with additional SQL instructions. Besides the SQL Query node, no DB nodes require SQL coding.

DB Row Filter
Expands the input SQL query to include the row filter criteria defined in the configuration window. Grouping of multiple conditions with an AND or OR conjunctions is also supported. No SQL coding required.

DB Query
Modifies the input SQL query using custom SQL. The input SQL query is represented by the place holder #table#.

DB Writer
Inserts the data rows from the top input port into a table in the database specified by the input connection port. If the database table does not exist it will be created.

DB Connection Table Writer
Writes the resulting rows from the input SQL query into a new table inside the database.

CLEAN DATA

Missing Value
Defines and applies a strategy to replace missing values in the input table - either globally on all columns, or individually for each single column.

Duplicate Row Filter
Detects duplicate rows and applies the selected operation, e.g. removes duplicate rows. Duplicates are rows that have the same value in all selected columns.

Numeric Outliers
Detects and treats numerical outliers for each of the selected columns individually using the interquartile range (IQR).

RESHAPE AND AGGREGATE DATA

GroupBy
Groups the rows of a table by the unique values in selected columns and calculates aggregation and statistical measures for the defined groups. Despite its simple name, it offers powerful functionality and has many unsuspected usages.

Pivoting
Extends the aggregation functionality of the GroupBy node by creating an output table with columns and rows for the unique values in the selected input columns. The unique values of the grouping columns become rows and the unique values of the pivoting columns become columns.

Category to Number
Maps the categorical values in the selected columns to integer values and exports the mapping rules to the model output port. The Category to Number (Apply) and Number to Category (Apply) nodes apply the mapping rule in both directions.

One to Many
Creates one new column for each value in the selected input column. These values become the column headers. Cells in the newly created columns are set to 0 if the value is not present, otherwise 1. This type of encoding is called one-hot vector.

Transpose
Converts the rows to columns and the columns to rows.

Table Manipulator
Performs several transformations at once, such as renaming, filtering, re-ordering and type changing, on the input columns. By adding dynamic ports it can replace a concatenate node.

Cell Splitter
Splits values in the selected column into two or more substrings, as defined by a delimiter match. A delimiter is a defined character, such as a comma, space, or any other character or character sequence.

Ungroup
Ungroups a collection-type cell by creating one row for each value in the collection cell. Other columns from the input table are left unaltered.

Unpivoting
Stacks the cells of the selected value columns into one column. The cells of the selected retained columns are appended to the corresponding output rows.

Sorter
Sorts the table in ascending or descending order based on the values of one or more columns.

Value Counter
Counts the number of occurrences of all values in a selected column from the input table.

DYNAMIC PORT

Concatenate
Dynamic ports: Additional input ports can be added by clicking the three dots in the bottom left corner of a node.

DATA TYPES & CONVERSIONS

S **String**: Sequence of characters, e.g. "This is a string"
I **Integer**: Whole real valued number, e.g. -100 or 345
D **Double**: Real valued number, e.g. -0.432 or 45.39
Date&Time: A data format for date, time, date&time, or date&time plus time zone.
B **Boolean**: Two possible values only, e.g. TRUE and FALSE

[ ] **Collection Cell**: Collection of multiple values of either the same or different types e.g., can be a list of values or a set of values. In a set each value occurs only once.

Document/Image: KNIME Analytics Platform supports many more data types like text documents, images, fingerprints, etc.

Number to String
Converts the data type of the selected columns from a number format, e.g. integer or double, to string.

String to Number
Converts the data type of the selected columns from string to either double or integer.

CREATE COLUMNS

Math Formula
Implements a number of math operations across multiple input columns. The math operations can be applied to multiple columns with the Math Formula (Multi Column) node.

Rule Engine
Applies a set of rules to each row of the input table. Rules are applied from top to bottom. The first rule that matches is used.

Counter Generation
Creates a new column with a counter. The start value and step size are defined in the configuration window.

String Manipulation
Performs operations on string values in columns, such as combining two or more strings together, extracting one or more substrings, trimming blank spaces, and so on.

String Replacer
Replaces values in a selected string column if they match a defined pattern.

Column Expressions
Combines the functionality of the Math Formula, Rule Engine, and String Manipulation nodes. More than one expression can be defined to modify or add multiple columns at the same time.

FORMAT EXCEL SHEETS

The Continental Nodes for KNIME extension allows you to automatically format an existing Excel sheet. The key is an additional data table of the same size as the original Excel sheet, where each cell contains one or more comma separated tag values e.g., header, border, etc. Based on these tags, the XLS Formatter nodes add new formatting instructions to the existing instructions, as available at the lower (optional) input port.

XLS Control Table Generator
Transforms the input table to an XLS Control Table, meaning it exchanges the column names to A, B, C, ... and the row IDs to 1, 2, 3, ... It is the kickoff node to collect formatting instructions for an Excel sheet and feeds all XLS formatter nodes.

XLS Background Colorizer
Adds background color and/or pattern fill formatting instructions to all cells with a specified tag in the XLS Control Table at the top input port.

XLS Border Formatter
Adds border formatting instructions for a given range specified by a tag in the XLS control table at the top input port.

XLS Cell Merger
Adds formatting instructions to merge all cells with a specified tag in the XLS control table at the top input port.

XLS Conditional Formatter
Adds formatting instructions to color cell backgrounds according to their numeric value for all cells specified by a tag in the XLS control table at the top

XLS Formatter (apply)
Applies all formatting instructions to an existing Excel sheet.

Resources

E-Books: KNIME Advanced Luck covers advanced features & more. Practicing Data Science is a collection of data science case studies from past projects. Both available at knime.com/knimepress

KNIME Blog: Engaging topics, challenges, industry news, & knowledge nuggets at knime.com/blog

E-Learning Courses: Take our free online self-paced courses to learn about the different steps in a data science project (with exercises & solutions to test your knowledge) at www.knime.com/knime-self-paced-courses

KNIME Hub: Browse and share workflows, nodes, and components. Add ratings, or comments to other workflows at hub.knime.com

KNIME Forum: Join our global community & engage in conversations at forum.knime.com

KNIME Server: For team-based collaboration, automation, management, & deployment check out KNIME Server at www.knime.com/knime-server

https://www.knime.com/sites/default/files/2021-07/cheat-sheet-data-wrangling.pdf

Open for Innovation
KNIME

Thank you!

education@knime.com

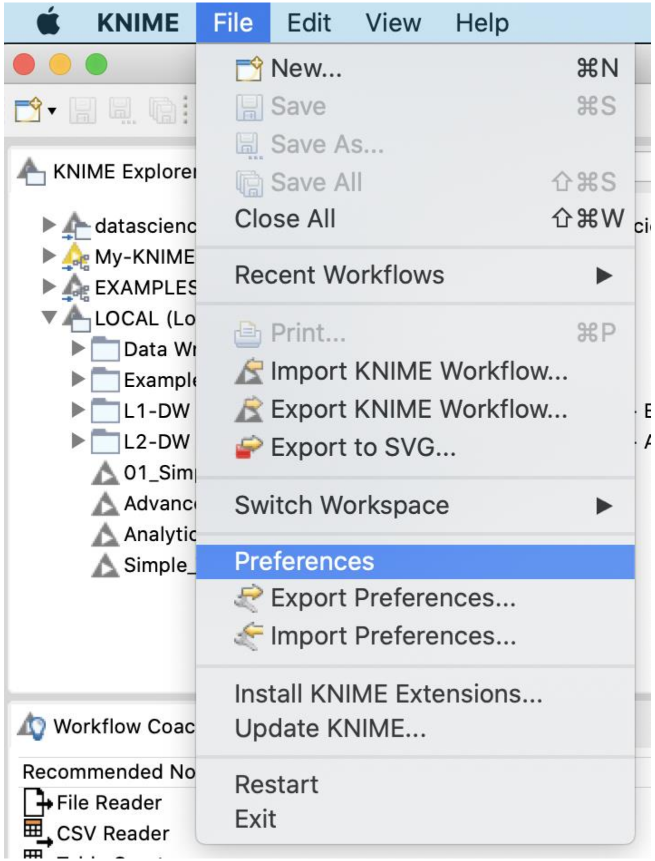
Attachment: How to use a local update site to install extensions

Adding a Local Update Site

- Download the update site as zip
  - KNIME update site as zip
  - Previous versions of the KNIME update site as zip
  - Community update sites as zip

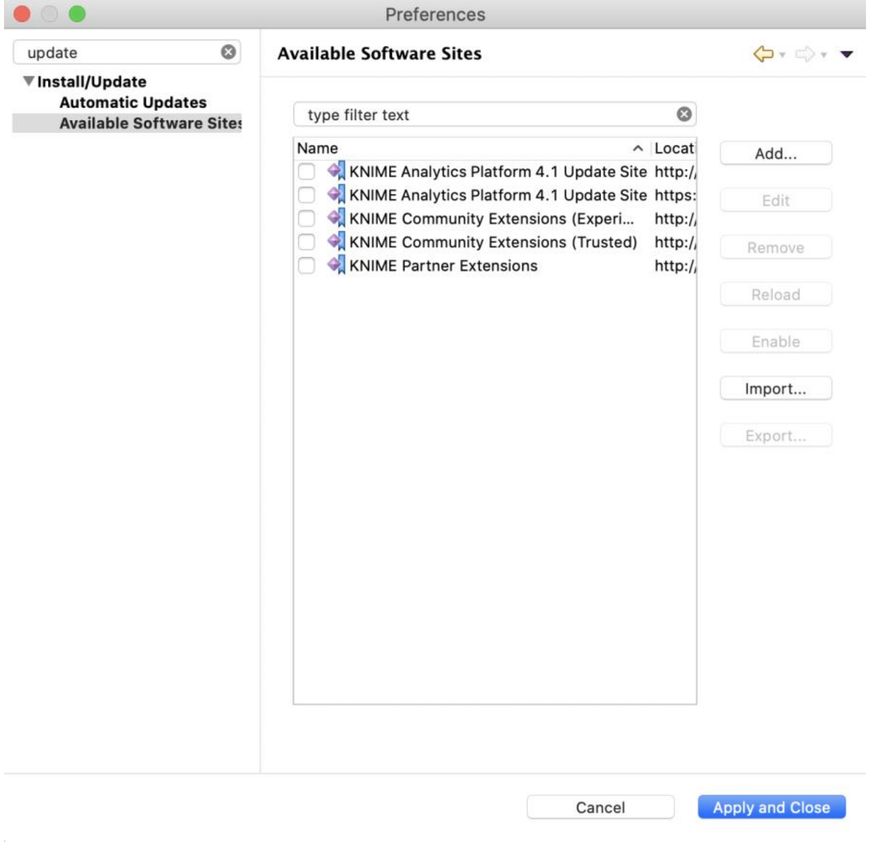
Adding a Local Update Site

- Open KNIME Analytics Platform and go to the preference page by clicking on
- File -> Preferences



Adding a Local Update Site

1. Search for update (upper left search bar) and go to Available Software sites.
2. Uncheck all existing software sites.
3. Click on Add.. on the upper right.



Preferences

update

Install/Update
Automatic Updates
Available Software Sites

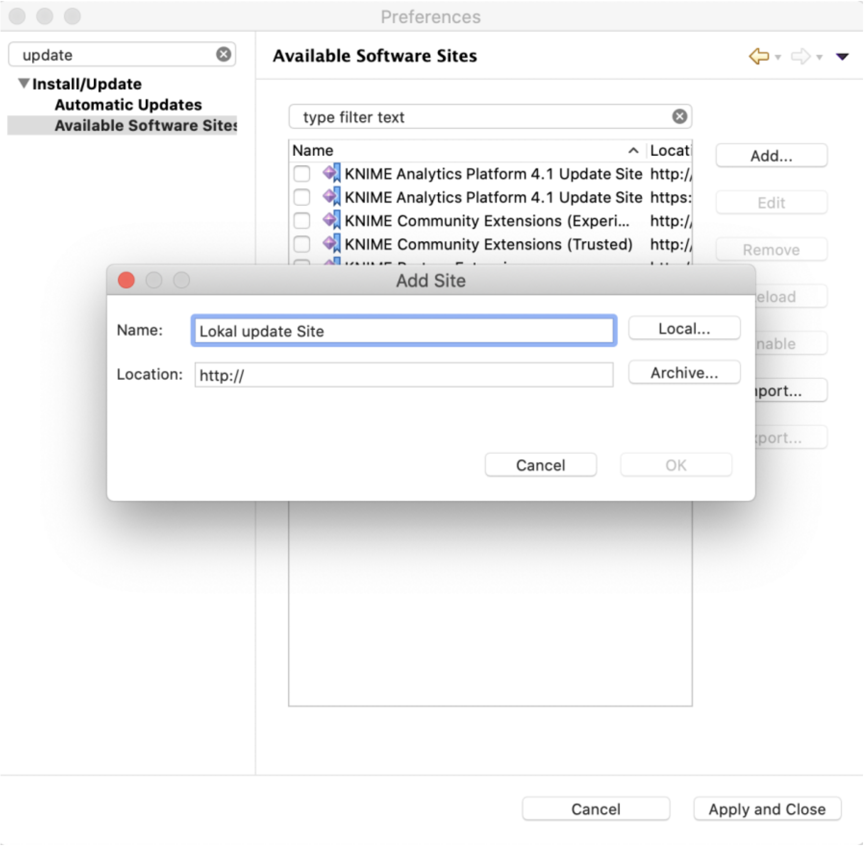
Available Software Sites

type filter text

	Name	Locat
□	KNIME Analytics Platform 4.1 Update Site	http:/
□	KNIME Analytics Platform 4.1 Update Site	https:
□	KNIME Community Extensions (Experi...	http:/
□	KNIME Community Extensions (Trusted)	http:/
□	KNIME Partner Extensions	http:/

Adding a Local Update Site

1. Define a name
2. Click on Archive and select the folder you've just downloaded
3. Click OK
4. Click Apply and Close



Preferences
update
Install/Update
Automatic Updates
Available Software Sites
Available Software Sites
type filter text
Name
Locat
KNIME Analytics Platform 4.1 Update Site http:/
KNIME Analytics Platform 4.1 Update Site https:
KNIME Community Extensions (Experi... http:/
KNIME Community Extensions (Trusted) http:/
Add...
Edit
Remove
eload
nable
port...
port...
Add Site
Name: Lokal update Site
Location: http://
Local...
Archive...
Cancel
OK
Cancel
Apply and Close