

[L1-DW] KNIME Analytics Platform for Data Wranglers: Basics

KNIME AG

April 17, 2023



Structure of the course

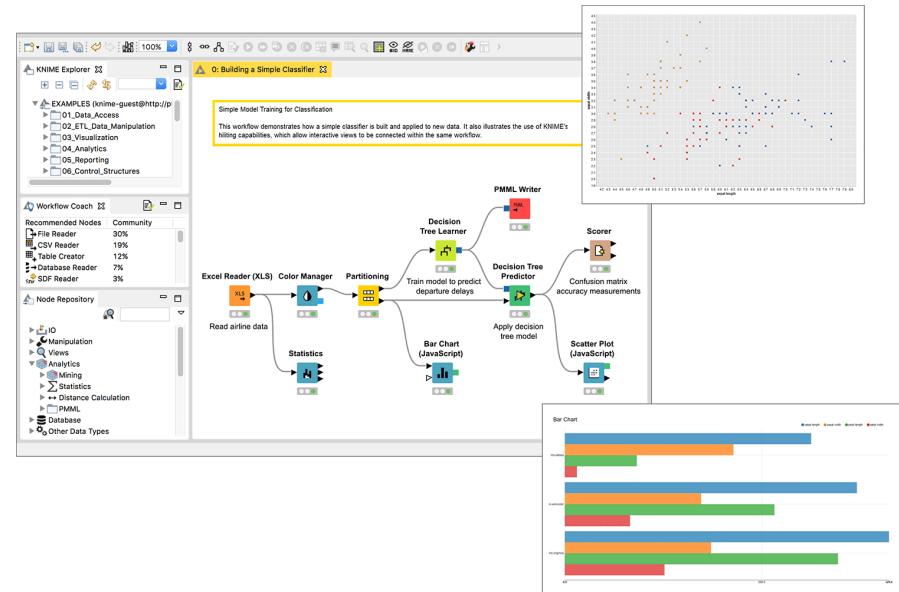
- This course covers the following topics
 - Introduction to KNIME Analytics Platform
 - Importing Data
 - Data Merging
 - Data Cleaning & Transformation
 - Data Aggregation
 - Data Visualization
 - Q&A and Summary
- Structure of the course
 - Introduction to the topic and workflow demo
 - Hands-on exercise
 - Solution walk-through

Overview KNIME Analytics Platform



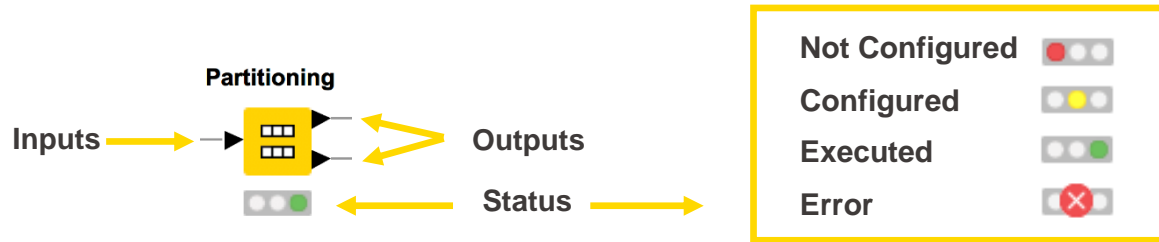
What is KNIME Analytics Platform?

- **Open source**
- Visual programming paradigm
- Data analysis, manipulation, visualization, and reporting
- Multiple extensions:
 - Text Mining
 - Network Mining
 - Cheminformatics
 - Geospatial Analytics
 - Many integrations, such as Java, R, Python, Weka, Keras, Plotly, H2O, etc.

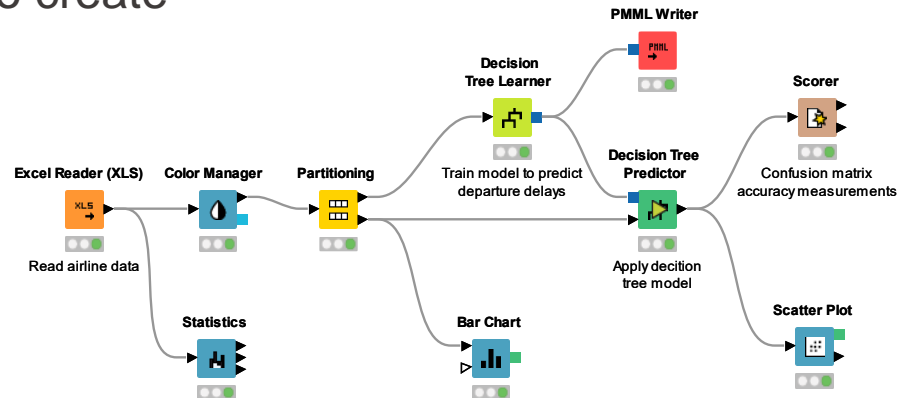


Visual KNIME Workflows

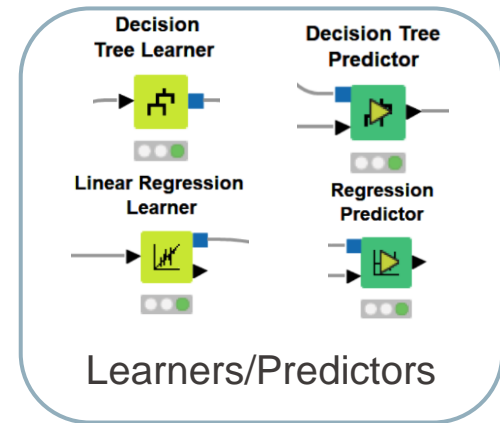
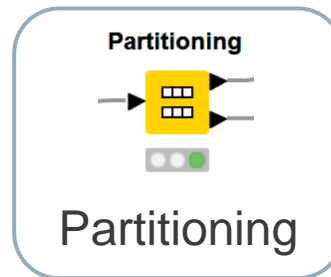
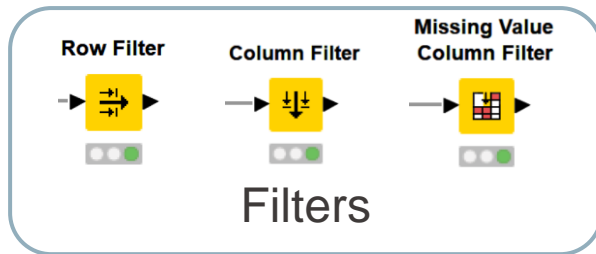
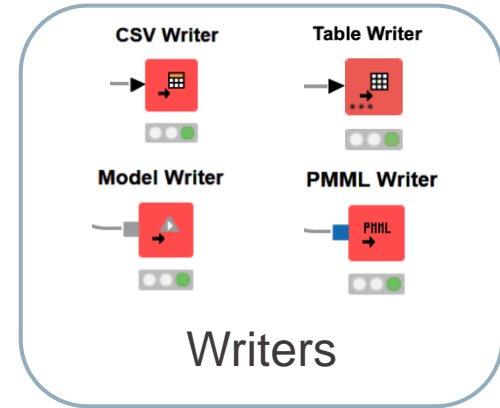
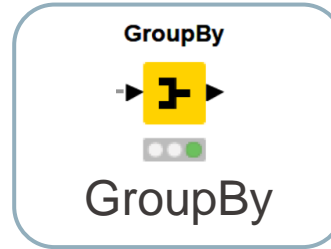
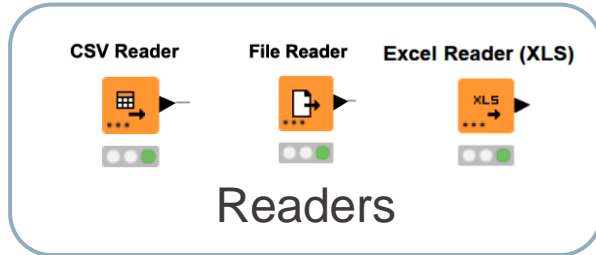
NODES perform tasks on data



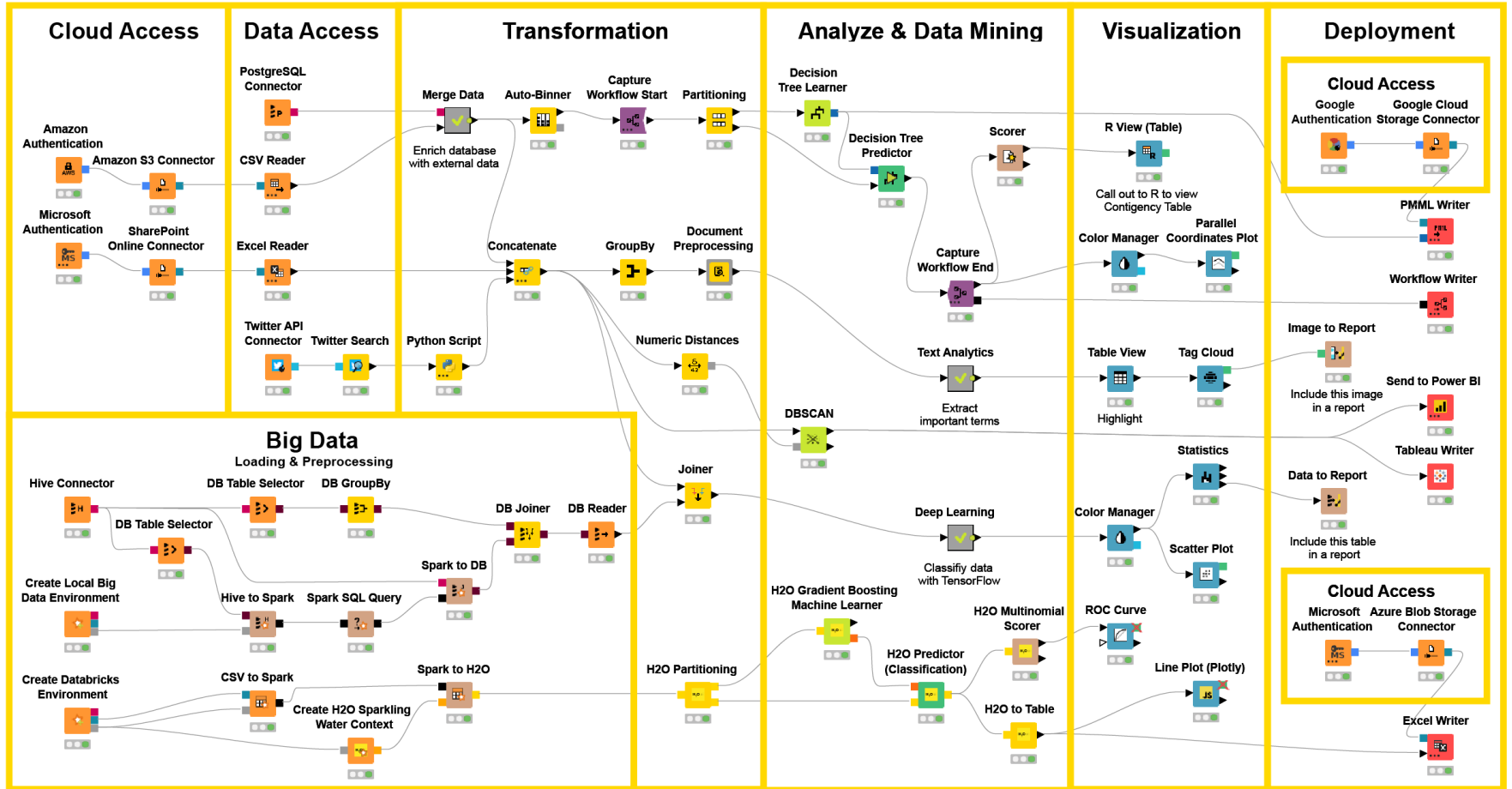
Nodes are combined to create **WORKFLOWS**



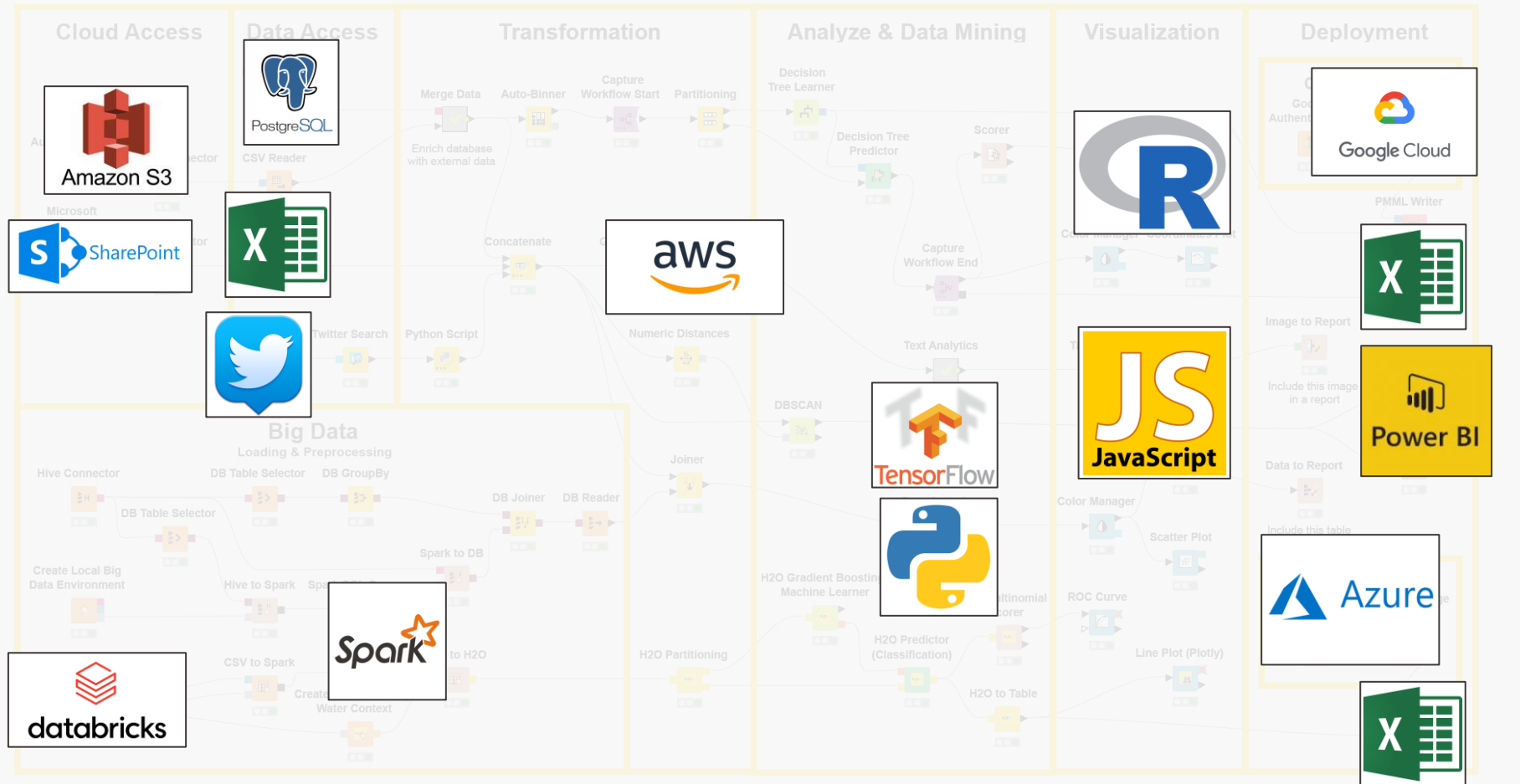
Frequently Used Nodes



Covering all Stages of the Data Science Life Cycle



Technologies & Languages under the Hood



Let's Start KNIME & Build a Workflow



Installation

<https://www.knime.com/downloads>

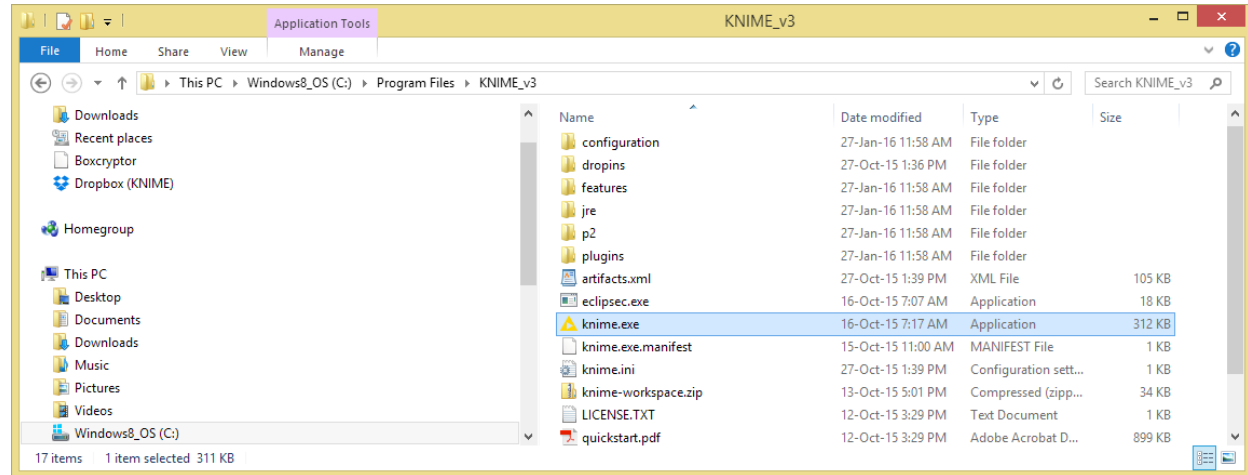
- Select the KNIME Analytics Platform version for your computer:
 - Mac
 - Windows – 32 or 64 bit
 - Linux
- Download the archive and extract the file, or download the installer package and run it

Start KNIME Analytics Platform

- Use the shortcut created by the installer

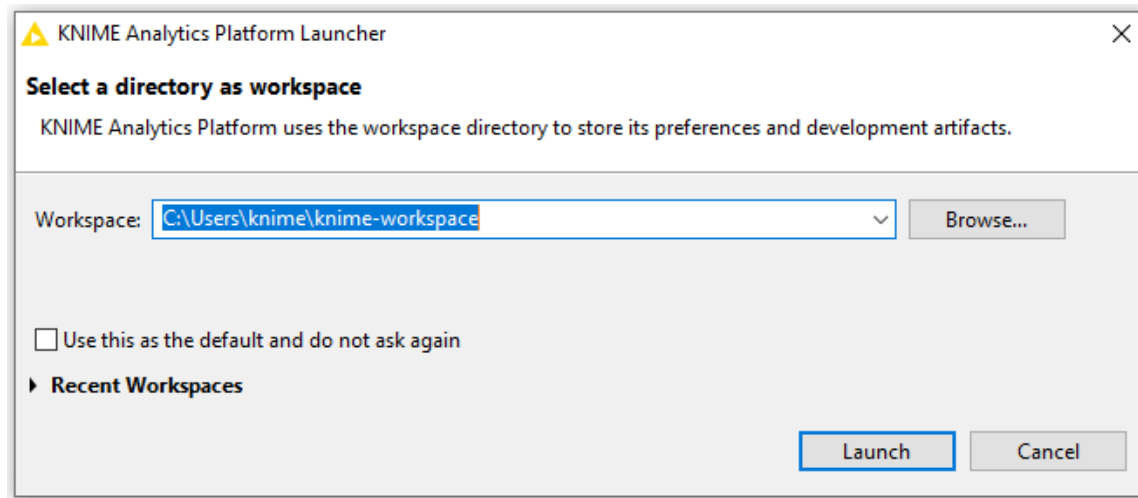


- Or go to the installation directory and launch KNIME via the knime.exe



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)



The KNIME Analytics Platform Workbench

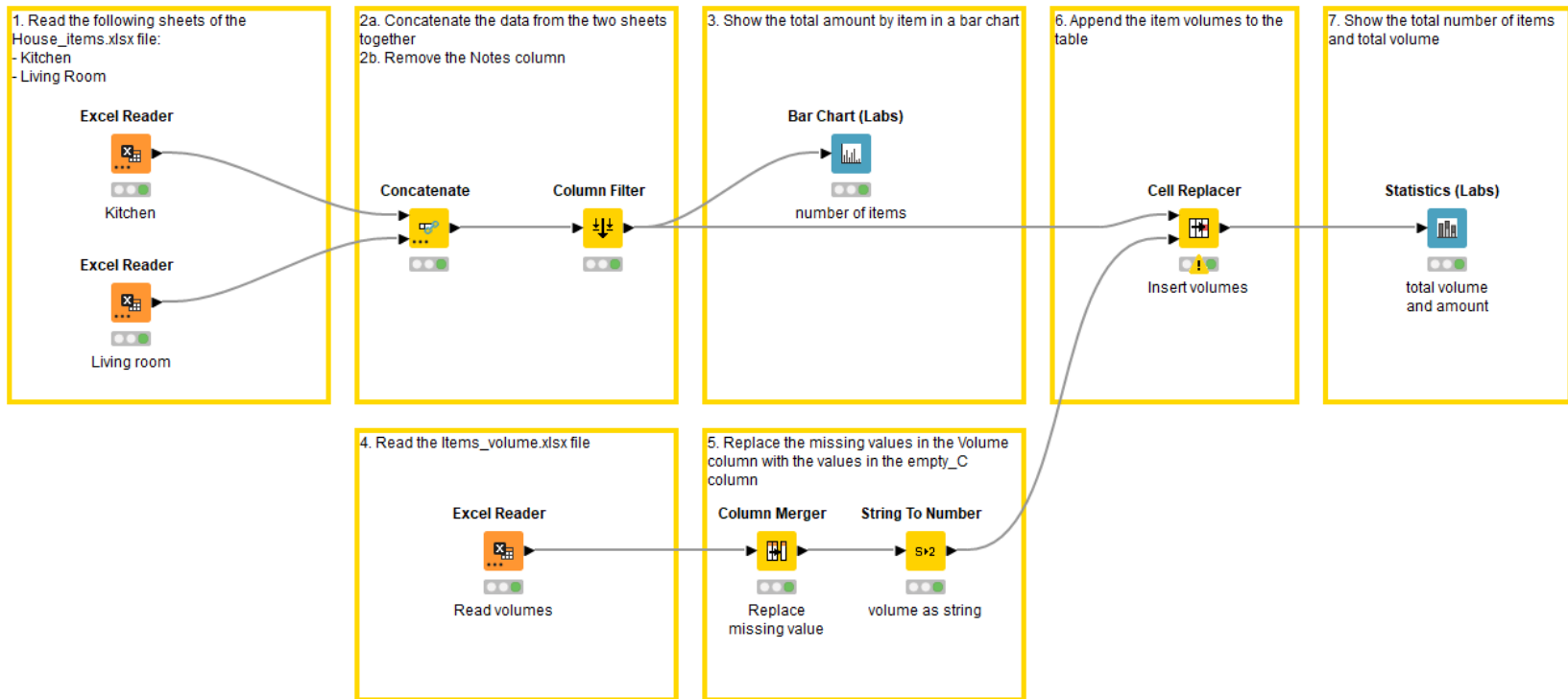
The screenshot displays the KNIME Analytics Platform Workbench interface. The main workspace shows a workflow titled "My first Workflow" with the following steps: File Reader (read adult.csv) -> Row Filter (keep only records born in the US) -> Column Filter (remove gender) -> Table Writer (write table). The interface is divided into several panels:

- KNIME Explorer:** Located on the top left, it shows a tree view of the project structure, including "My-KNIME-Hub", "EXAMPLES", and "LOCAL (Local Workspace)".
- Workflow Coach:** Located on the middle left, it displays a list of recommended nodes with their usage percentages, such as "GroupBy" (12%), "Joiner" (9%), and "Column Filter" (8%).
- Node Repository:** Located on the bottom left, it provides a comprehensive list of nodes categorized by function, such as "IO", "Manipulation", "Views", and "Analytics".
- Node Description:** Located on the top right, it provides detailed information about the selected "Row Filter" node, including its purpose and configuration options.
- Outline:** Located at the bottom left, it shows a small thumbnail of the current workflow.
- Console & Node Monitor:** Located at the bottom right, it displays the execution status of the selected node and its output. The "Row Filter" node is shown as "EXECUTED".

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

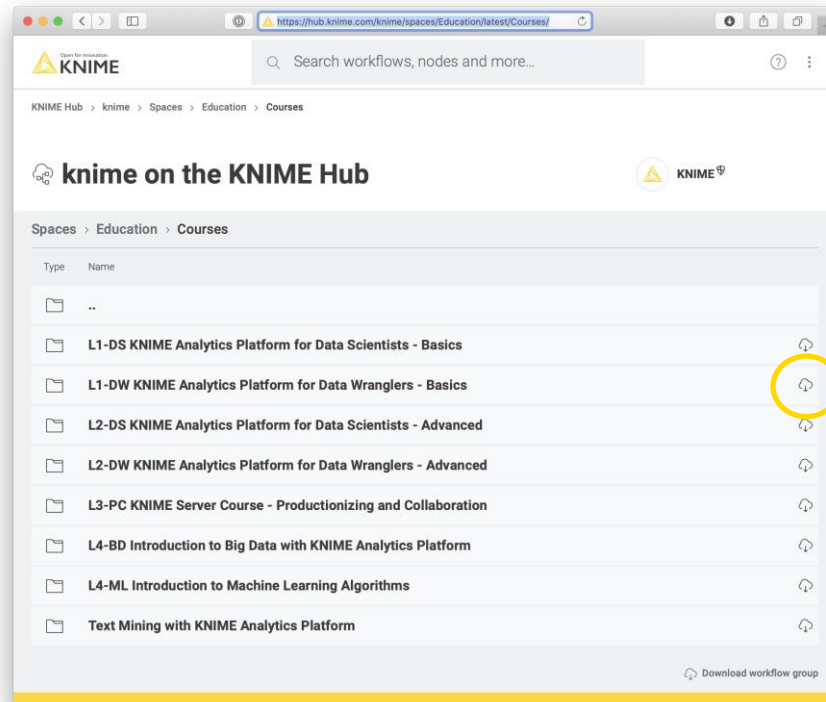
Move to new apartment -workflow

- Let's complete this together!



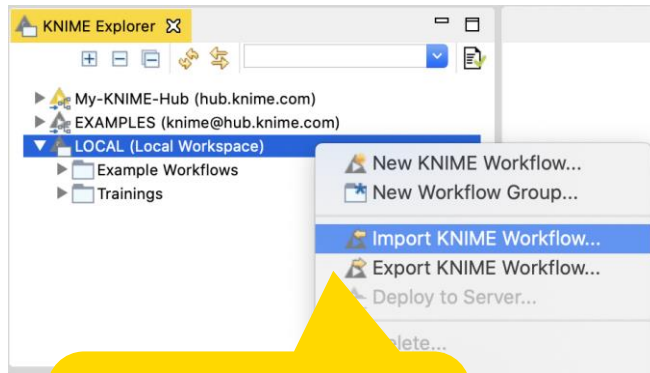
Downloading Exercises

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

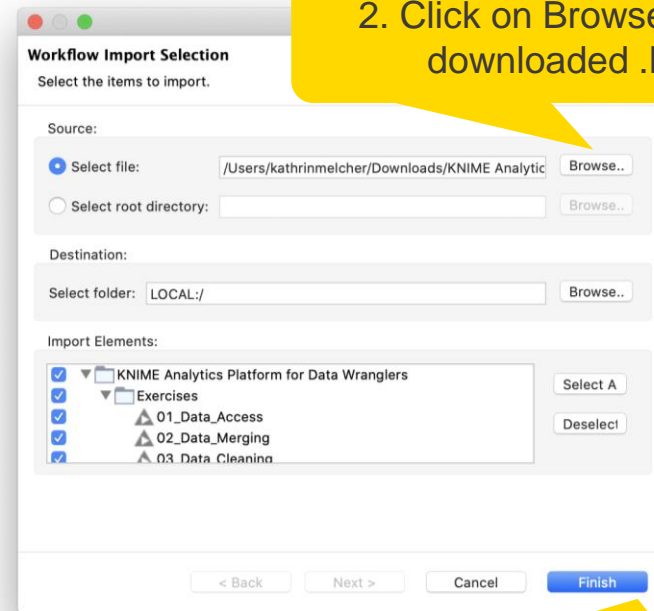


Importing Exercises

- Import the course material to KNIME Analytics Platform



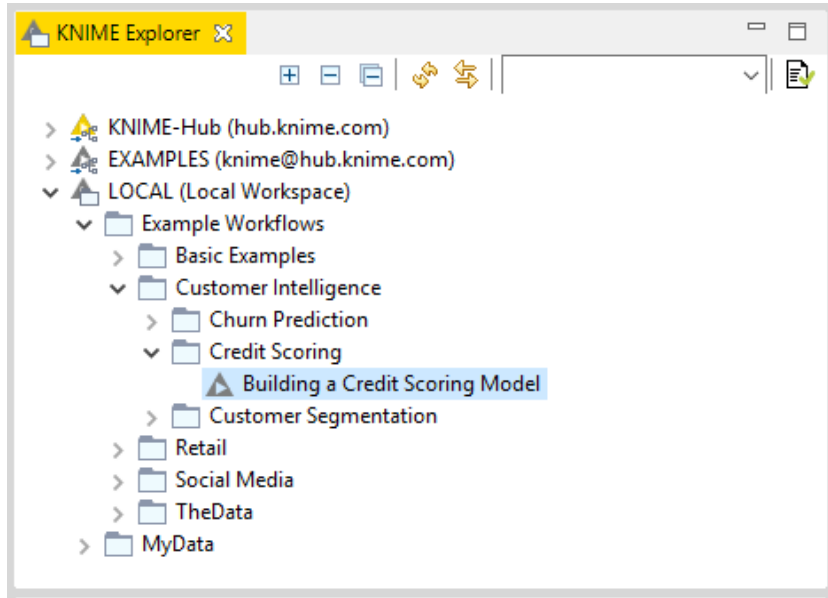
1. Right click on LOCAL and select Import KNIME Workflow....



2. Click on Browse and select downloaded .knar file

3. Click on Finish

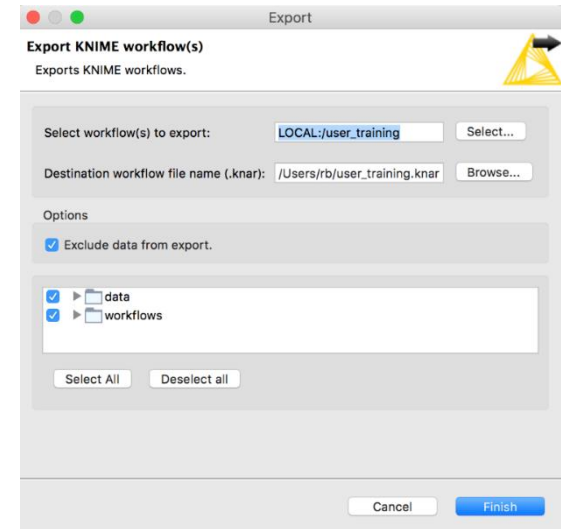
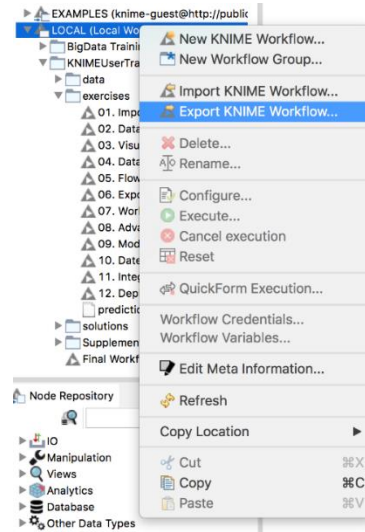
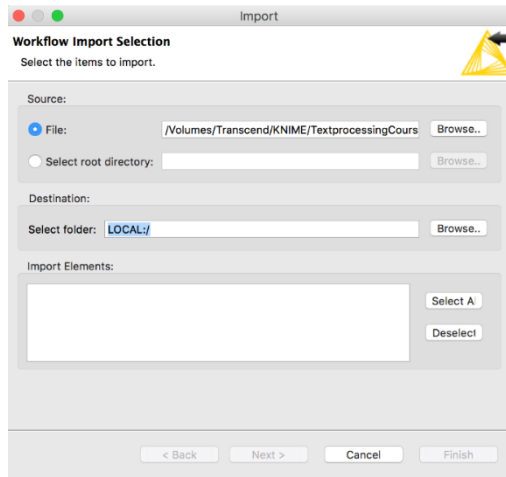
KNIME Explorer



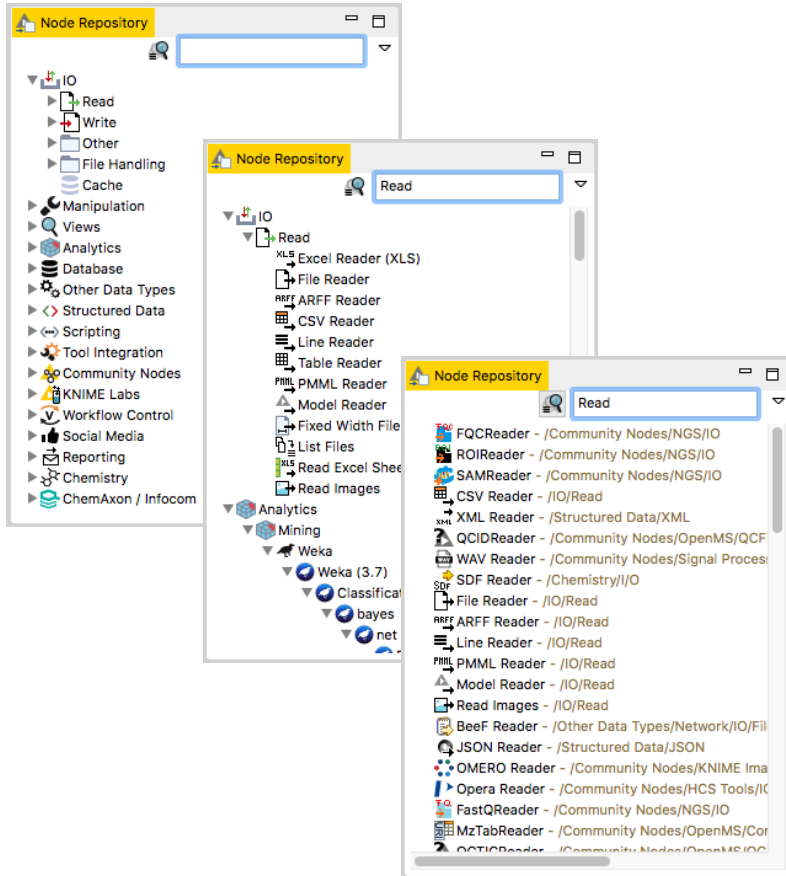
- In LOCAL you can access your own workflow projects.
- Other mountpoints allow you to connect to
 - Examples
 - KNIME Community Hub
 - KNIME Business Hub
- The Explorer toolbar on the top has a search box and buttons to
 - ↕ select the workflow displayed in the active editor
 - 🔄 refresh the view
- 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

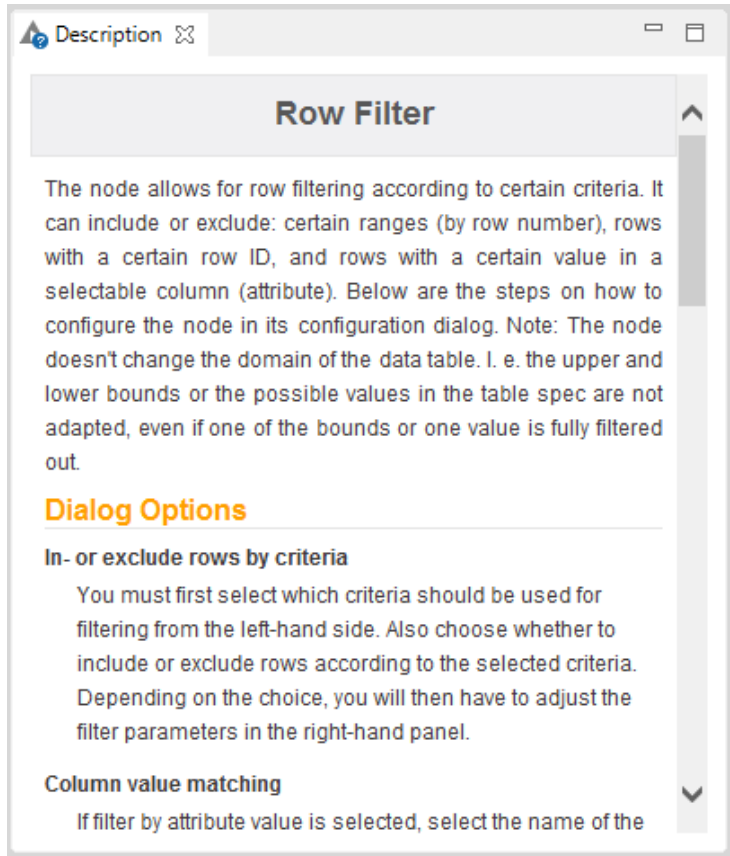


Node Repository



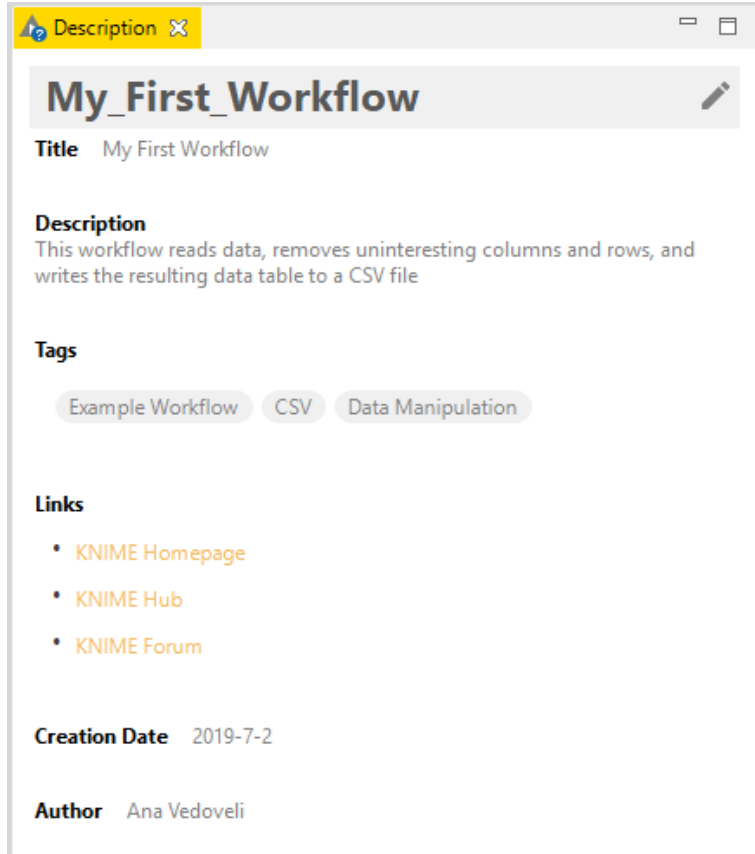
- The Node Repository lists all KNIME nodes
- The search box has 2 modes
 -  Standard Search – exact match of node name
 -  Fuzzy Search – finds the most similar node name

Description



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

Workflow Description



The screenshot shows a web browser window with the title 'Description'. The main content area displays the workflow details for 'My_First_Workflow'. The title is 'My First Workflow'. The description states: 'This workflow reads data, removes uninteresting columns and rows, and writes the resulting data table to a CSV file'. The tags are 'Example Workflow', 'CSV', and 'Data Manipulation'. The links section includes 'KNIME Homepage', 'KNIME Hub', and 'KNIME Forum'. The creation date is '2019-7-2' and the author is 'Ana Vedoveli'.

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](#)

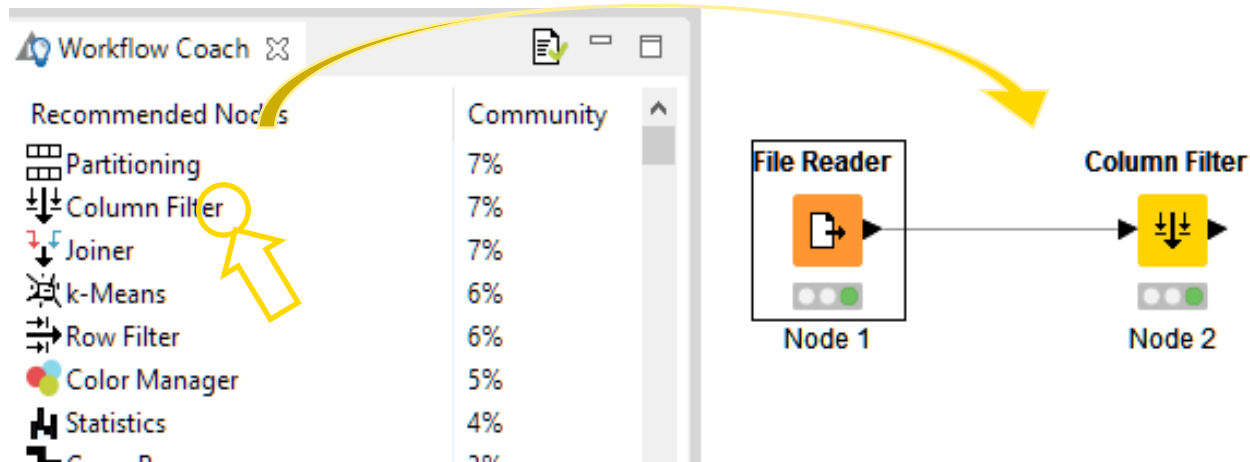
Creation Date 2019-7-2

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



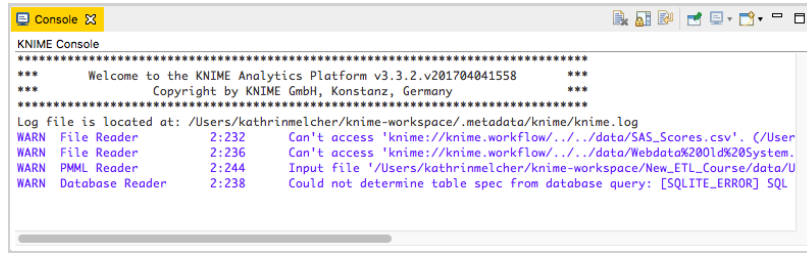
Node Monitor

- 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.

The screenshot displays the KNIME Node Monitor window for the node "Get Customers from Database (0:1207)". The node state is "EXECUTED". The "Port Output" is set to "Port 0". A context menu is open, listing options: "Show Output Table" (checked), "Show Variables", "Show Configuration", "Show Entire Configuration", "Show Node Timing Information", and "Show Graph Annotations". Below the menu is a table of customer data.

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

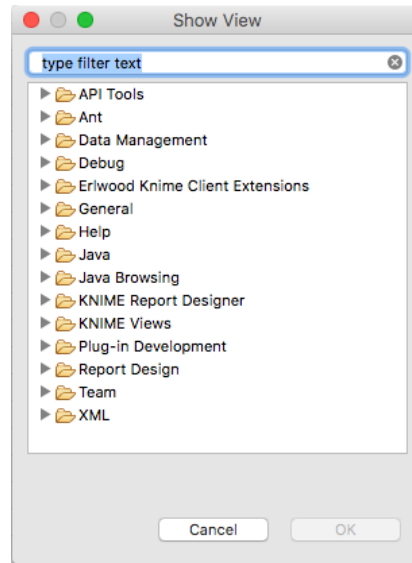
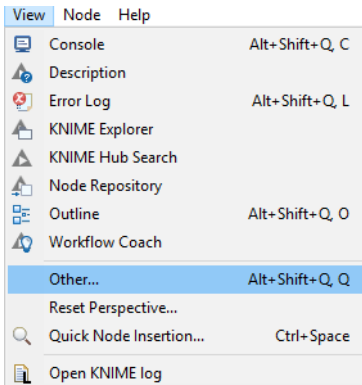
Console and Other Views



```
KNIME Console
*****
*** 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%201d%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
```

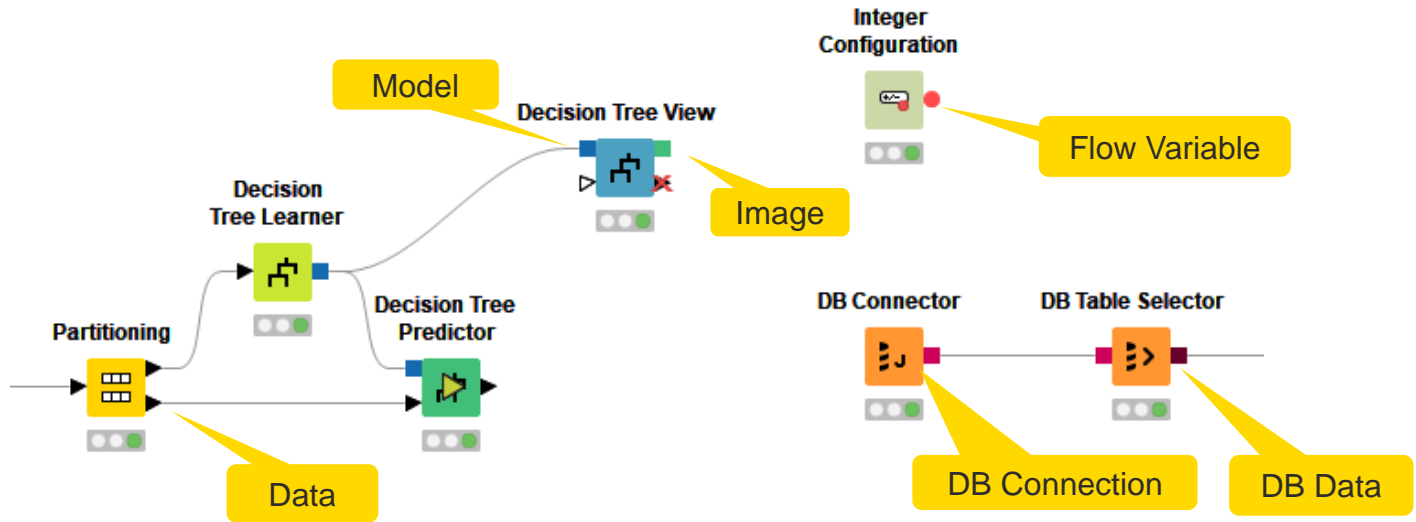
- Console view prints out error and warning messages about what is going on under the hood

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



Inserting and Connecting Nodes

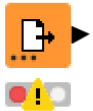
- Insert nodes into workspace by dragging them from the Node Repository or by double-clicking in the Node Repository
- Connect nodes by left-clicking the output port of Node A and dragging the cursor to the (matching) input port of Node B
- Common port types:



More on Nodes...

- A node can have 4 states:

File Reader



Not Configured:

The node is waiting for configuration or incoming data.

File Reader



Configured:

The node has been configured correctly and can be executed.

File Reader



Executed:

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

File Reader



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

Dialog - 0:1 - File Reader

File

Settings Transformation Advanced Settings Limit Rows Encoding Flow Variables Memory Policy

Input location

Read from: Relative to Current workflow

Mode: File Files in folder

File: .../data/CustomerInfoSystem1.csv

Reader options

Format

Autodetect 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

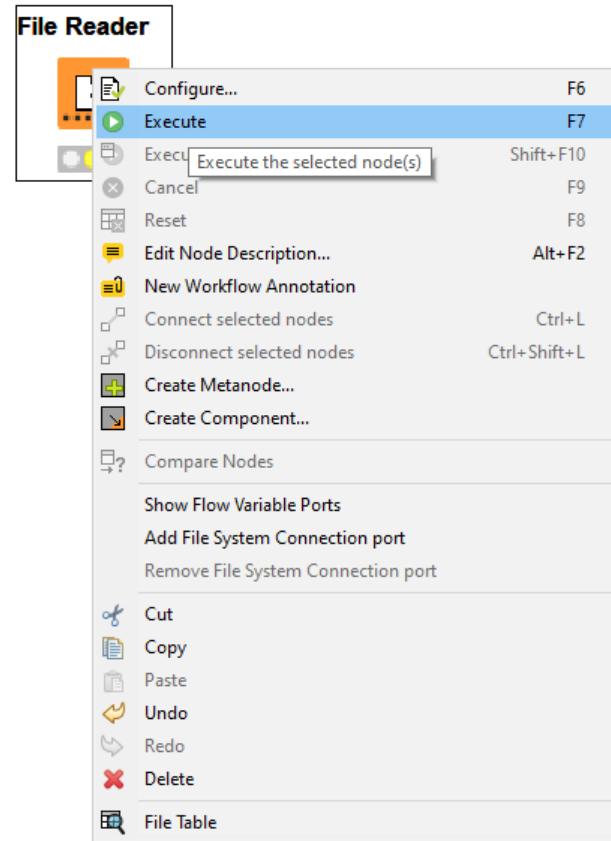
Preview

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@mcr.com
Row2	Sheffield	United Kingdom	27-270-743-182	Rotraut	GrÄ14nwald	20.4.1975	44	Rotraut.GrÄ14nwald@mcr.com
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.com
Row6	Valencia	Spain	65-655-257-939	Hanspeter	Storch	25.1.1998	21	Hanspeter_Storch@mcr.com

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

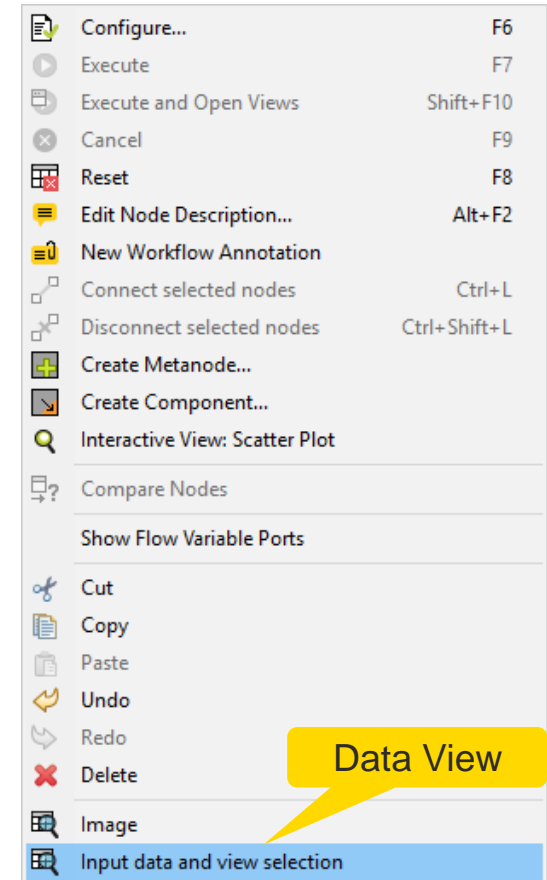
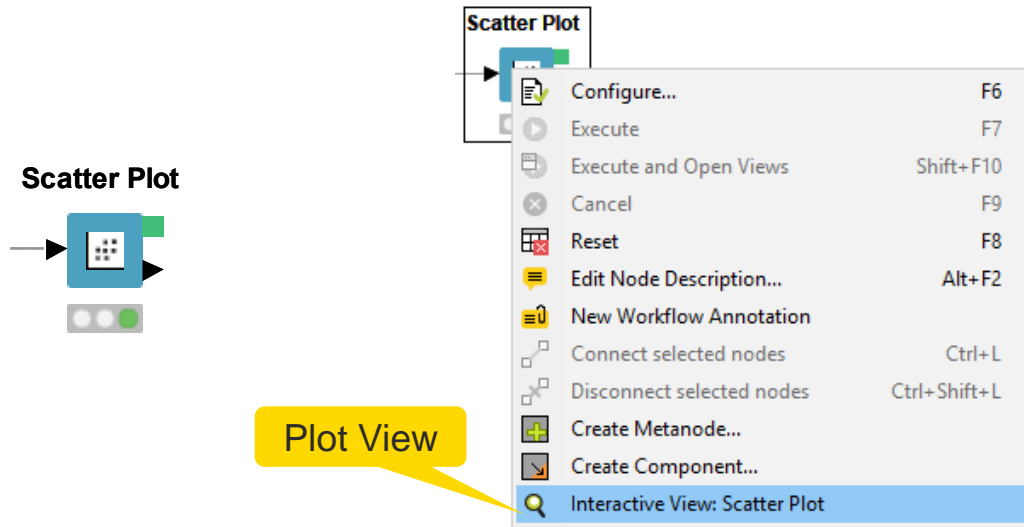


- The buttons in the toolbar can be used for the active workflow. The most important buttons are:

-  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



KNIME File Extensions

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

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

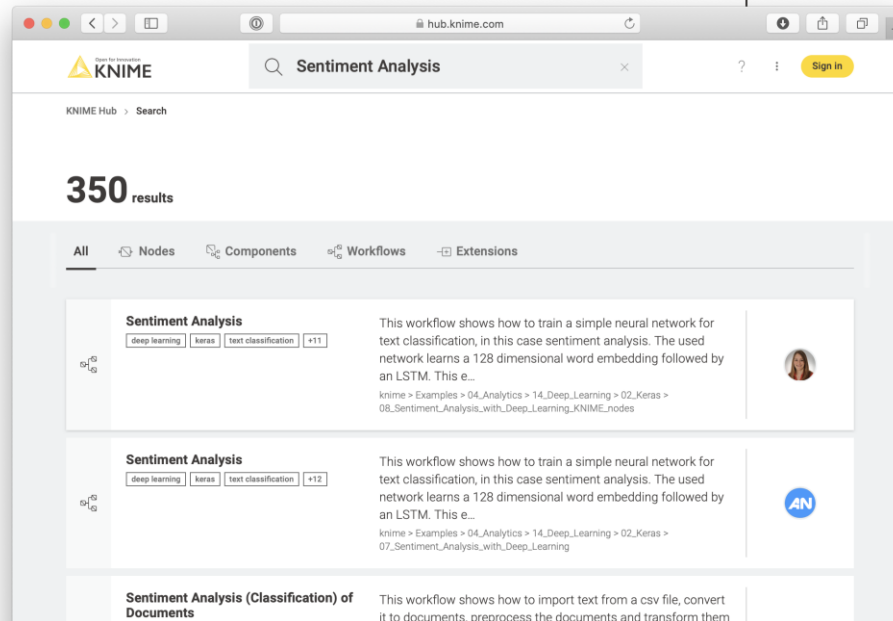
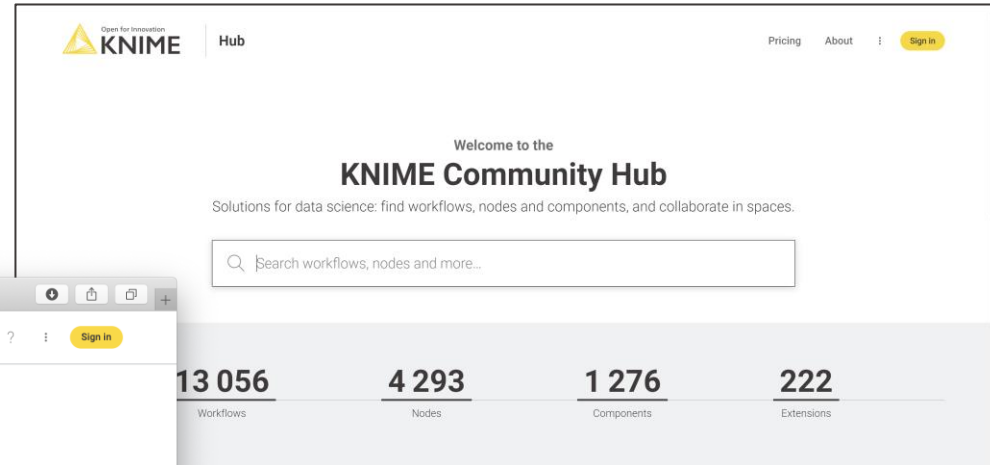


- ***.knar** for KNIME Archive Files



Getting Started: KNIME Community Hub

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



<https://hub.knime.com>

Getting Started: KNIME Examples

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

The screenshot displays the KNIME Explorer interface. On the left, a file browser shows a tree structure of example workflows, including 'My-KNIME-Hub (hub.knime.com)', 'EXAMPLES (knime@hub.knime.com)', and 'LOCAL (Local Workspace)'. The main workspace shows a workflow titled 'Simple Preprocessing Example' with a yellow box highlighting its description: 'Shows the use of different filter nodes.' The workflow diagram includes nodes for 'File Reader', 'Row Filter', 'Column Filter', 'Reference Column Filter', 'Numeric Binner', 'Nominal Value Row Filter', 'Reference Row Filter', 'Row Filter', and 'Concatenate'. A yellow banner at the top of the workspace reads: 'This is a temporary copy of "knime://EXAMPLES/Users/knime/Examples/02_ETL_Data_Manipulation/00_Basic_Examples/01_Example_for_Standard_Preprocessing". Use "Save As..." to save a permanent copy of the workflow to your local workspace, or a mounted KNIME Server.' On the right, a smaller window shows a list of repositories: 'My-KNIME-Hub (hub.knime.com)', 'EXAMPLES (knime@hub.knime.com)', 'LOCAL (Local Workspace)', 'Example Workflows', and 'MyData'. A tooltip over the 'EXAMPLES' entry says 'Double-click to see the examples'.

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"

The screenshot displays the KNIME Analytics Platform (Nightly Build) interface. The main window is titled "Basic Customer Segmentation Use Case". The interface is divided into several sections:

- Repository:** A sidebar on the left containing a search bar and categorized nodes under "IO", "Manipulation", and "Views".
- Workflow:** The central workspace shows a workflow starting with two "Excel Reader" nodes (one for "ContractData.xls" and one for "ContractData.csv") feeding into a "Metanode". The "Metanode" then feeds into a "k-Means" node, which is configured for "10 clusters on all numerical inputs". The output of the "k-Means" node is split into two "Denormalizer (PMML)" nodes. The top one is labeled "Input data with assigned cluster" and the bottom one is labeled "Cluster centers". Both denormalizers have a "Back to original data range" option.
- Table View:** Below the workflow, a table view shows the first 100 rows of a file with 16 columns. The table is titled "1: File Table" and "Flow Variables".

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

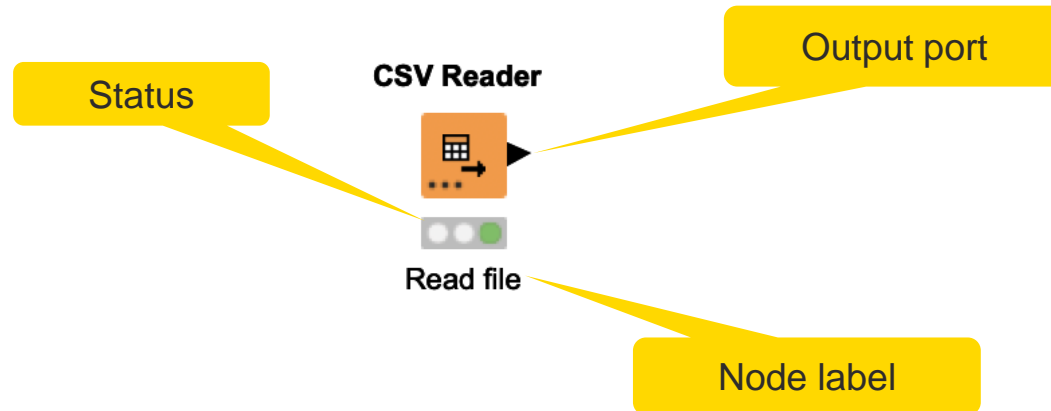
Importing Data



Data Source Nodes

Typically characterized by:

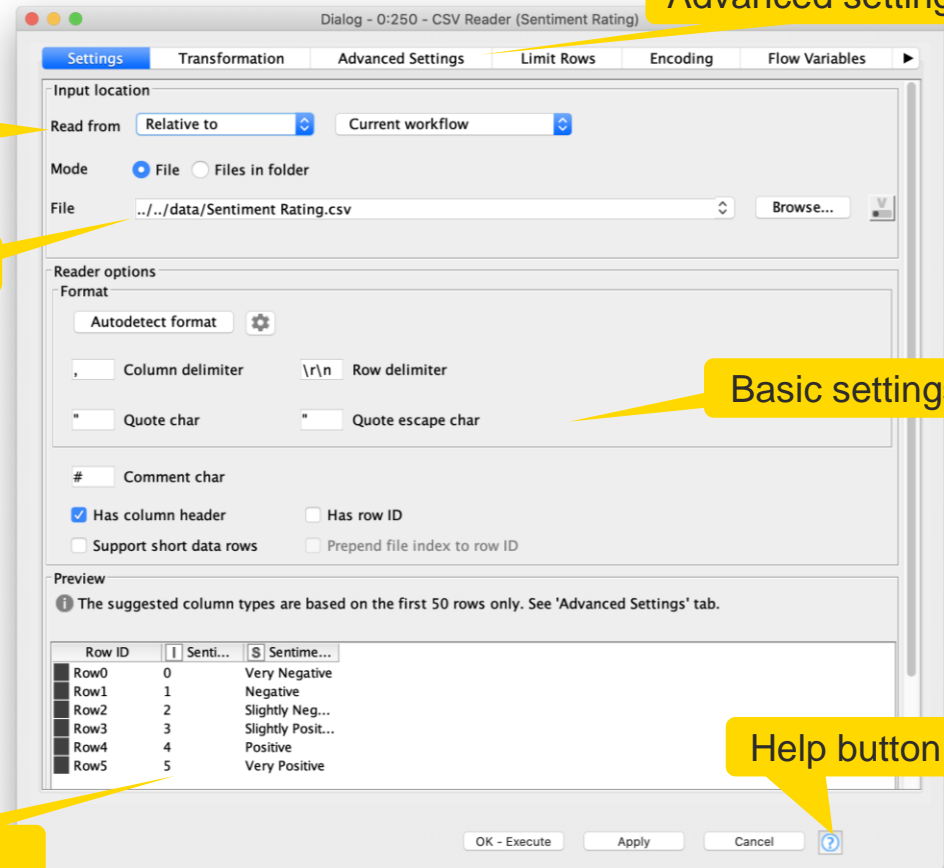
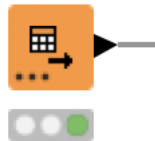
- Orange color
- By default no input ports, 1-2 output ports
- New file handling with KNIME 4.3.
 - Consistent user experience across all nodes and file systems
 - Managing of various file systems within the same workflow
 - Performance improvements



CSV Reader

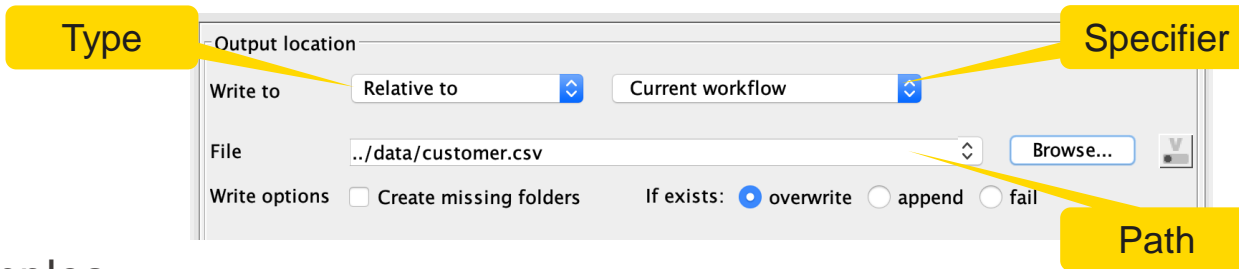
- Reads either one or multiple .csv and .txt files
- Further tabs to
 - limit the rows
 - select encoding

CSV Reader



Common Settings: File Path

- A path consists of three parts:
 - **Type:** Specifies the file system type e.g. local, relative, mountpoint, custom_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



- 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:

Mode: File Files in folder

File:

Relative to ...

Read from:

File:

Mountpoint

Read from:

File:

Custom URL

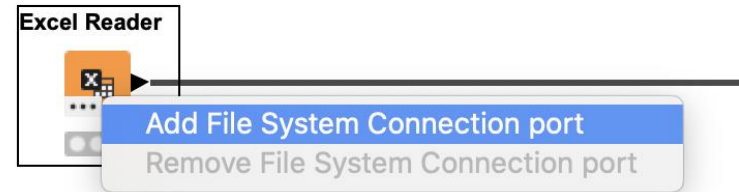
Read from:

URL:

Common Settings: Connecting to other File Systems

- Add file system connection port to connect to another file system

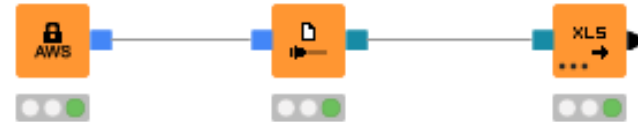
- Click on the three dots on the lower left to add or remove a dynamic port.



- Supported file systems

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

Amazon Authentication Amazon S3 Connector Excel Reader (XLS)



Input location

Read from

Mode File Files in folder

File

Common Settings: Read Single or Multiple Files

Single file

Input location

Read from

Mode File Files in folder

File

Files in a folder

Input location

Read from

Mode File Files in folder Include subfolders

Folder

Selected 22 of 22 files

Filter options

File filter options

File extension(s)

Case sensitive

File name

Case sensitive Wildcard Regular expression

Include hidden files

Folder filter options

Folder name

Case sensitive Wildcard Regular expression

Include hidden folders

- Option to include subfolder
- Option to define filter criterions

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)

Dialog - 0:2 - CSV Reader

File Settings **Transformation** Advanced Settings Limit Rows Encoding Flow Variables Memory Policy

Transformations

Reset actions ↑ Move up ↓ Move down Enforce types Take columns from: Union Intersection

	Column	New name	Type
::	<input checked="" type="checkbox"/> S City		S String
::	<input checked="" type="checkbox"/> S Country		S String
::	<input checked="" type="checkbox"/> S CustomerID	ID	S String
::	<input checked="" type="checkbox"/> S FirstName		Local Time
::	<input checked="" type="checkbox"/> S LastName		D Number (double)
::	<input checked="" type="checkbox"/> S Birthday		I Number (integer)
::	<input checked="" type="checkbox"/> I Age		L Number (long)
::	<input checked="" type="checkbox"/> S Email		PMML
::	<input checked="" type="checkbox"/> I Newsletter		Period
::	<input checked="" type="checkbox"/> ? <any unknown new column>		SVG image
			S String
			?

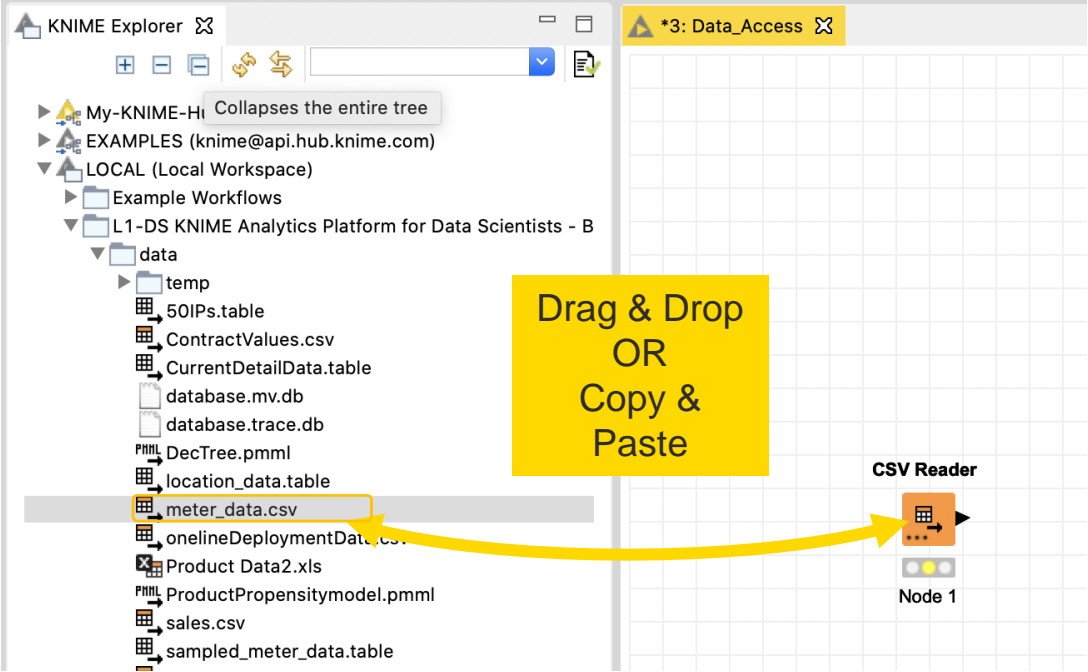
Preview

i 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 Newswle...
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@mcr.com	0
Row2	Sheffield	United Kingdom	27-270-743-182	Rotraut	GrÄ¼nwald	20.4.1975	44	Rotraut.GrÄ¼nwald@mcr.com	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

OK Apply Cancel ?

Alternative Faster Way ...



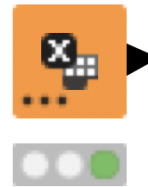
Excel Reader (XLS)

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

Excel Reader



Read Excel Sheet Names



Excel Reader

Excel Reader



File system

Path

Sheet specific settings

Preview

A screenshot of the 'Excel Reader' dialog box in KNIME. The dialog has a title bar 'Dialog - 0:1 - Excel Reader' and a 'File' menu. It contains several tabs: 'Settings', 'Transformation', 'Advanced Settings', 'Flow Variables', and 'Memory Policy'. The 'Settings' tab is active and shows the following options:

- Input location:** 'Read from' is set to 'Relative to' and 'Current workflow'. 'Mode' is set to 'File'. The 'File' field contains '..../data/Product Data2.xls'.
- Sheet selection:** 'Select first sheet with data' is selected. The sheet name is 'Product Data.xls_defa...'.
- Column header:** 'Table contains column names in row number' is checked, with '1' selected.
- Row ID:** 'Generate row IDs' is selected, and 'Table contains row IDs in column' is set to 'A'.
- Sheet area:** 'Read entire data of the sheet' is selected.

At the bottom, there are 'Preview' and 'File Content' tabs. The 'Preview' tab is active and shows a table with the following data:

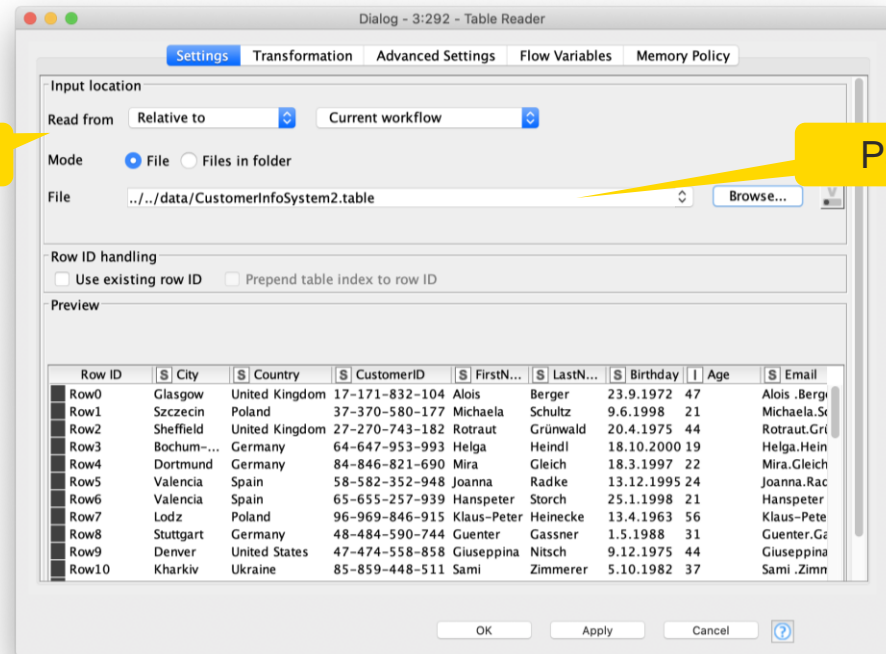
Row ID	Custom...	Products
Row0	11000	Private Investment
Row1	11001	Private Investment
Row2	11002	Private Investment
Row3	11003	Private Investment
Row4	11004	Private Investment

Buttons for 'OK', 'Apply', 'Cancel', and a help icon are at the bottom right.

Table Reader

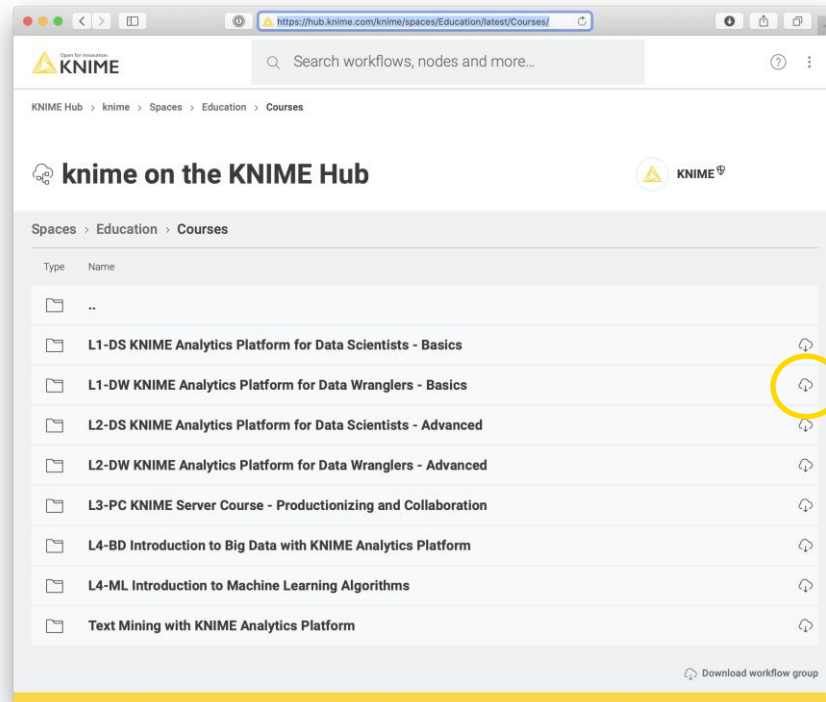
- Reads tables from the native KNIME Format
- Maximum performance, minimum configuration

Table Reader



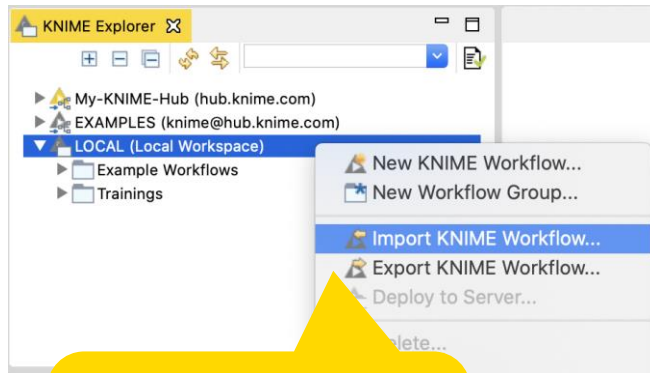
Downloading Exercises

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

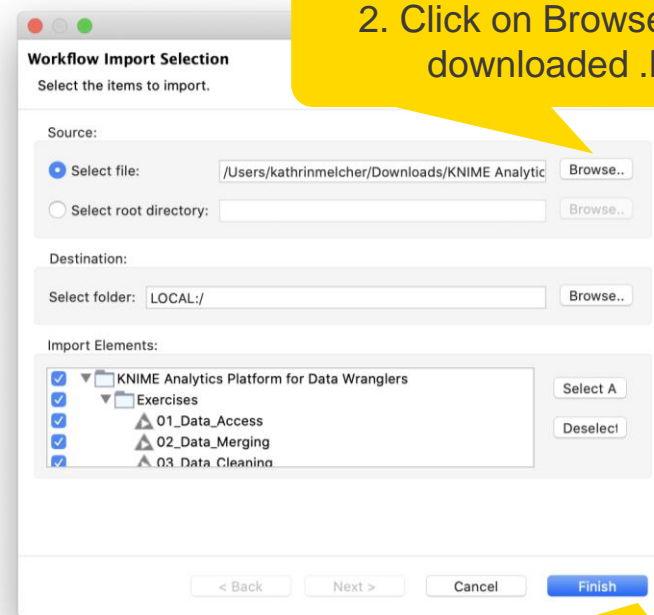


Importing Exercises

- Import the course material to KNIME Analytics Platform



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



2. Click on Browse and select downloaded .knar file

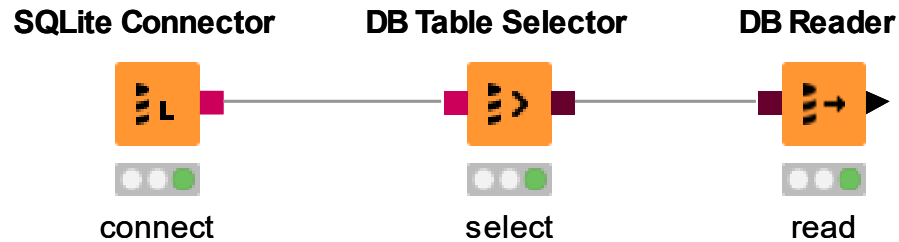
3. Click on Finish

Accessing Databases



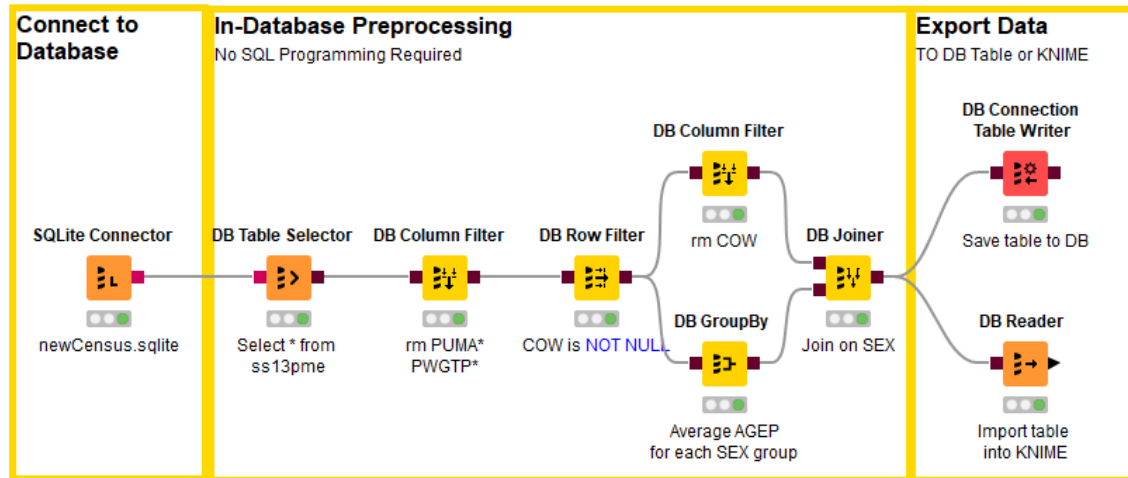
Database Connectivity

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

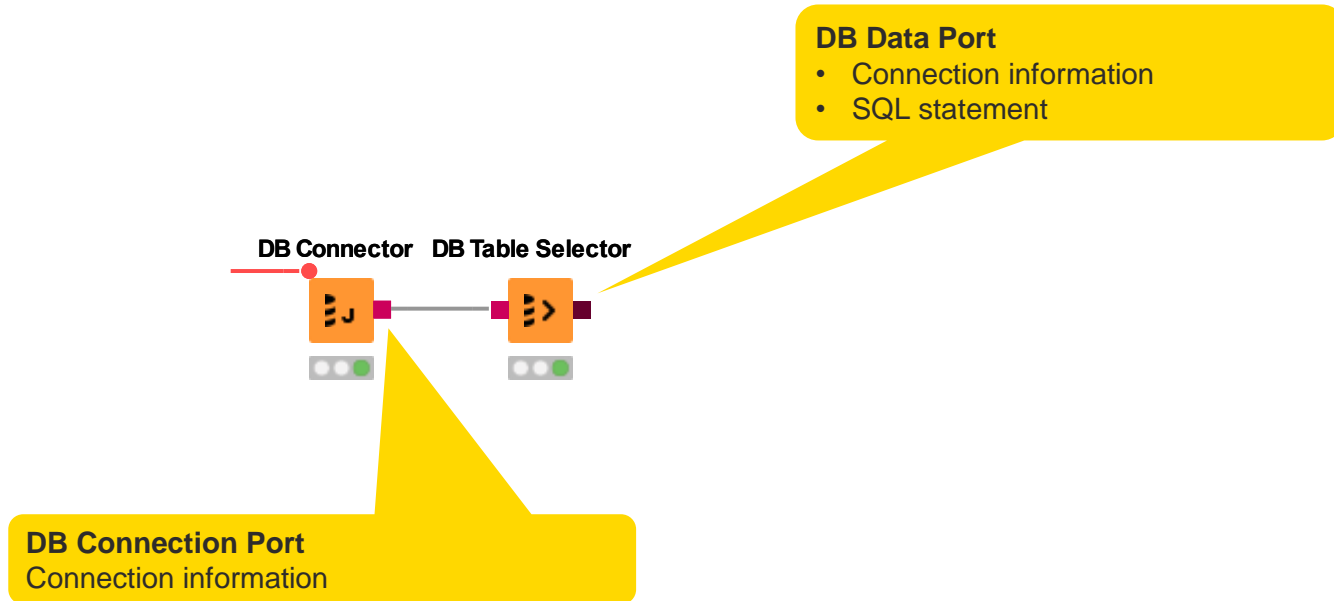


Database Extension

- 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

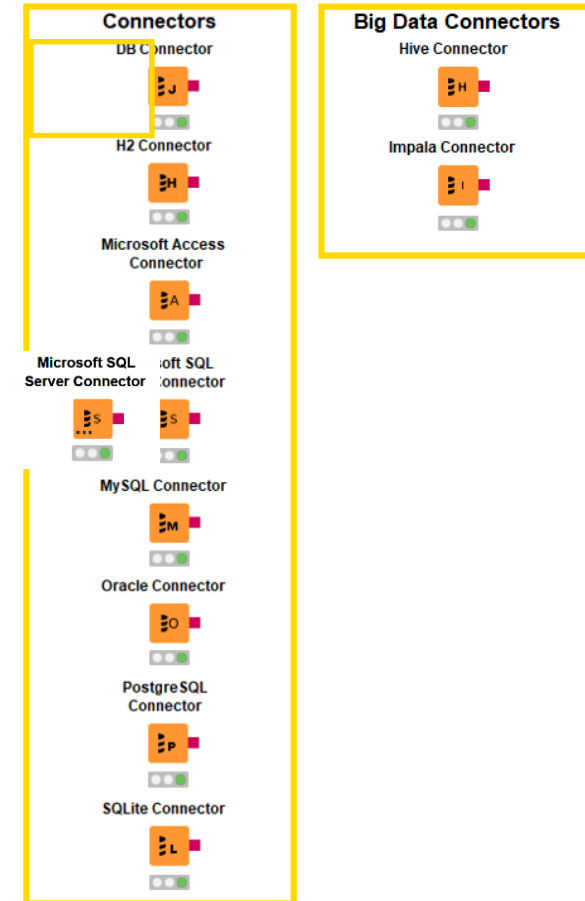


Database Port Types



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

Open KNIME
and go to File -
> Preferences

Preferences

type filter text

- ▶ General
- ▶ Help
- ▶ Install/Update
- ▶ Java
- ▼ KNIME
 - Customization Profiles
 - Data Storage
 - Databases**
 - Databases (legacy)
 - H2O
 - JavaScript Views
 - KNIME Explorer
 - ▶ KNIME GUI
 - Kerberos
 - Master Key
 - Meta Info Preferences
 - Preferred Renderers
 - Python
 - Python Deep Learning
 - R
 - TensorFlow
 - ▶ Textprocessing
 - ▶ Workflow Coach
- ▶ Run/Debug
- ▶ Team

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
------	---------	---------

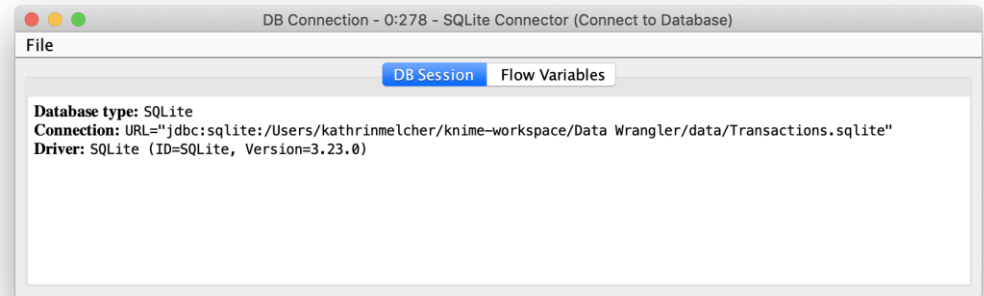
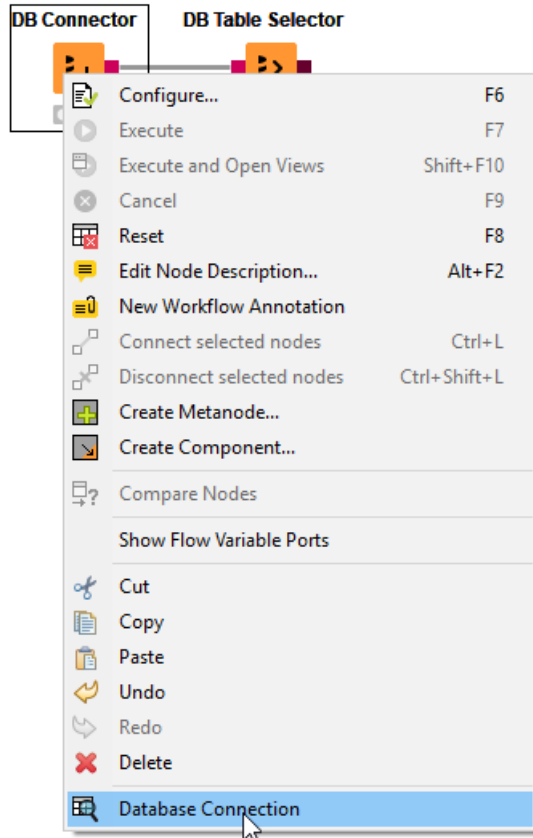
Buttons: Edit, Add, Remove, Up, Down

Buttons: Restore Defaults, Apply

Buttons: Cancel, Apply and Close

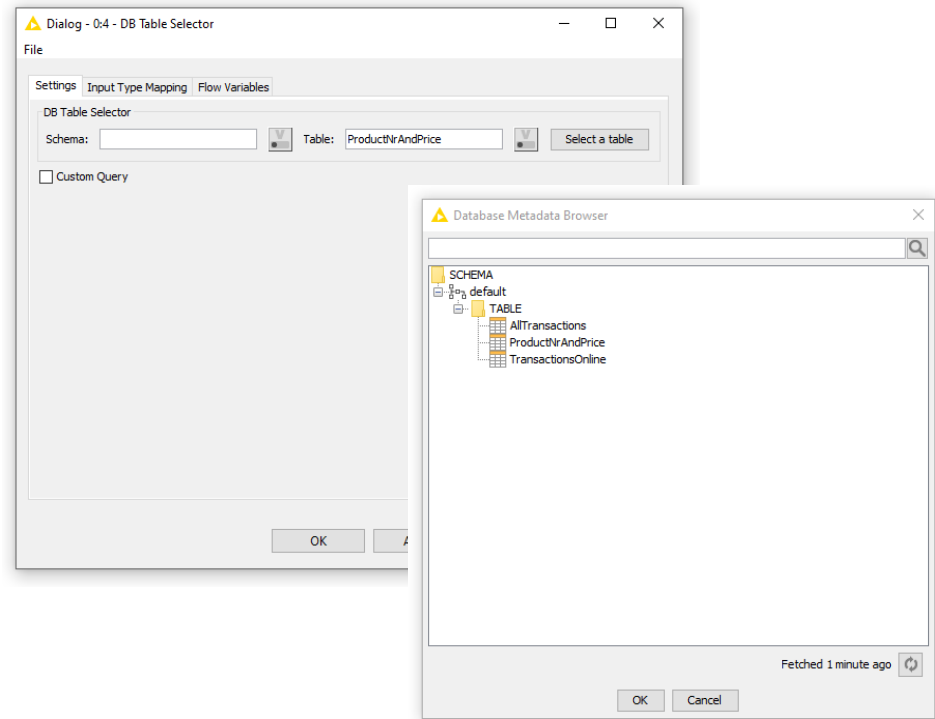
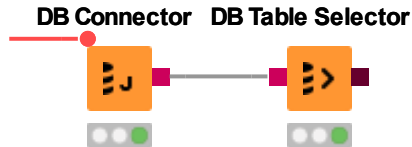
Register single jar file JDBC drivers and new JDBC drivers with companion files

Database JDBC Connection Port View



DB Table Selector

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



Database Connection Port View

The screenshot shows the context menu for the DB Table Selector node. The menu items are:

- Configure... (F6)
- Execute (F7)
- Execute and Open Views (↑F10)
- Cancel (F9)
- Reset (F8)
- Edit Node Description... (⌘F2)
- New Workflow Annotation
- Connect selected nodes (⌘L)
- Disconnect selected nodes (↑⌘L)
- Create Metanode...
- Create Component...
- Compare Nodes
- Show Flow Variable Ports
- Cut
- Copy
- Paste
- Undo
- Redo
- Delete
- DB Data (highlighted)

The top screenshot shows the DB Data window with the Table Preview tab selected. It displays a table with 17 rows and 3 columns: D, Prize, and ProductNr.

Row ID	D	Prize	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	

The bottom screenshot shows the DB Data window with the DB Query tab selected. The SQL query is:

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

A "Copy SQL to clipboard" button is visible at the bottom right of the window.

DB Reader

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

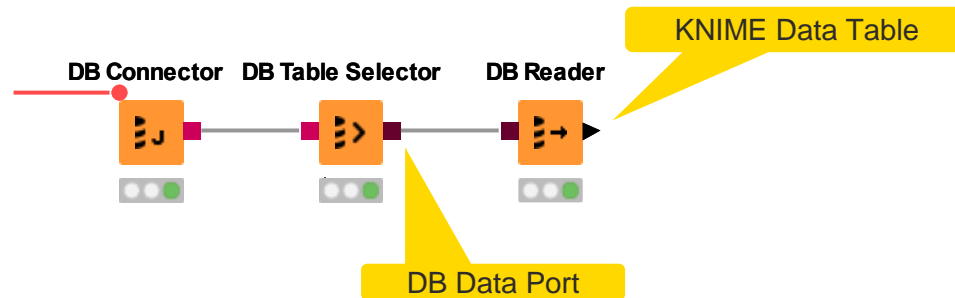
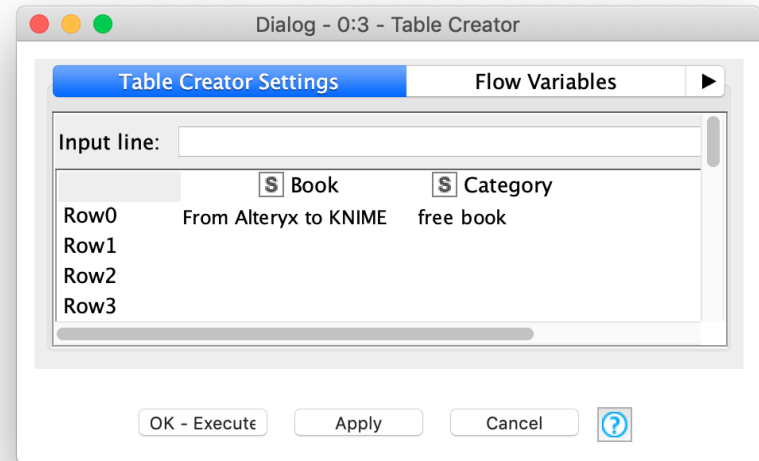
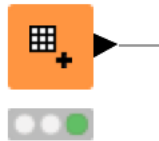


Table Creator

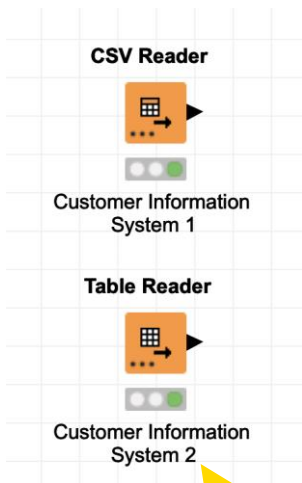
- Allows you to create data tables manually
- Data can be entered in a spreadsheet – like the table in the configuration dialog

Table Creator



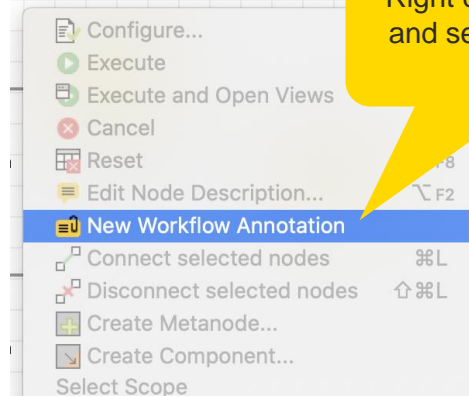
Comments & Annotations

Comments

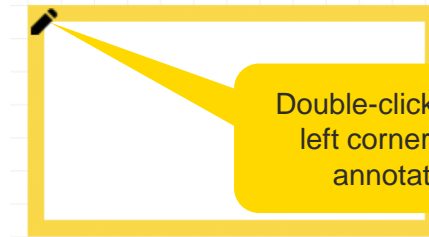


Double-click to change the node label

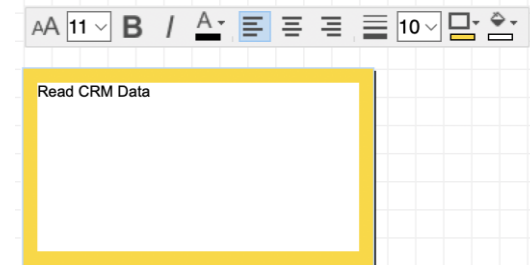
Annotations



Right click in the workflow and select New Workflow Annotation



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



Exercise: 01_Data_Access

Open the workflow 01_Data_Access and read the following data:

- 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

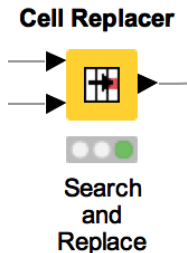
Try to use workflow relative-paths

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)
 - 2-S Number To String
 - S-2 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
 - f-F 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)

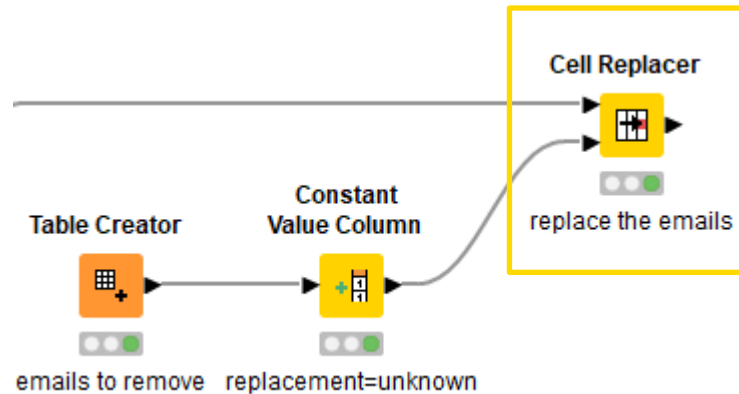
- Row
 - Filter
 - Filter Apply
 - Filter Apply Row Splitter
 - Filter Definition Merger
 - HiLite Row Splitter
 - Nominal Value Row Filter
 - Numeric Row Splitter
 - Reference Row Filter
 - Reference Row Splitter
 - Row Filter
 - Row Splitter
 - Rule-based Row Filter
 - Rule-based Row Filter (Dictionary)
 - Rule-based Row Splitter
 - Rule-based Row Splitter (Dictionary)
 - 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)
 - 2-S Number To String (PMML)
 - Numeric Binner (PMML)
 - One to Many (PMML)
 - Ruleset Editor
 - Ruleset Predictor
 - Ruleset to Table
 - S-2 String To Number (PMML)
 - XML To PMML
 - Cell To PMML
 - PMML To Cell

Cell Replacer

Replaces the content of a column based on a lookup

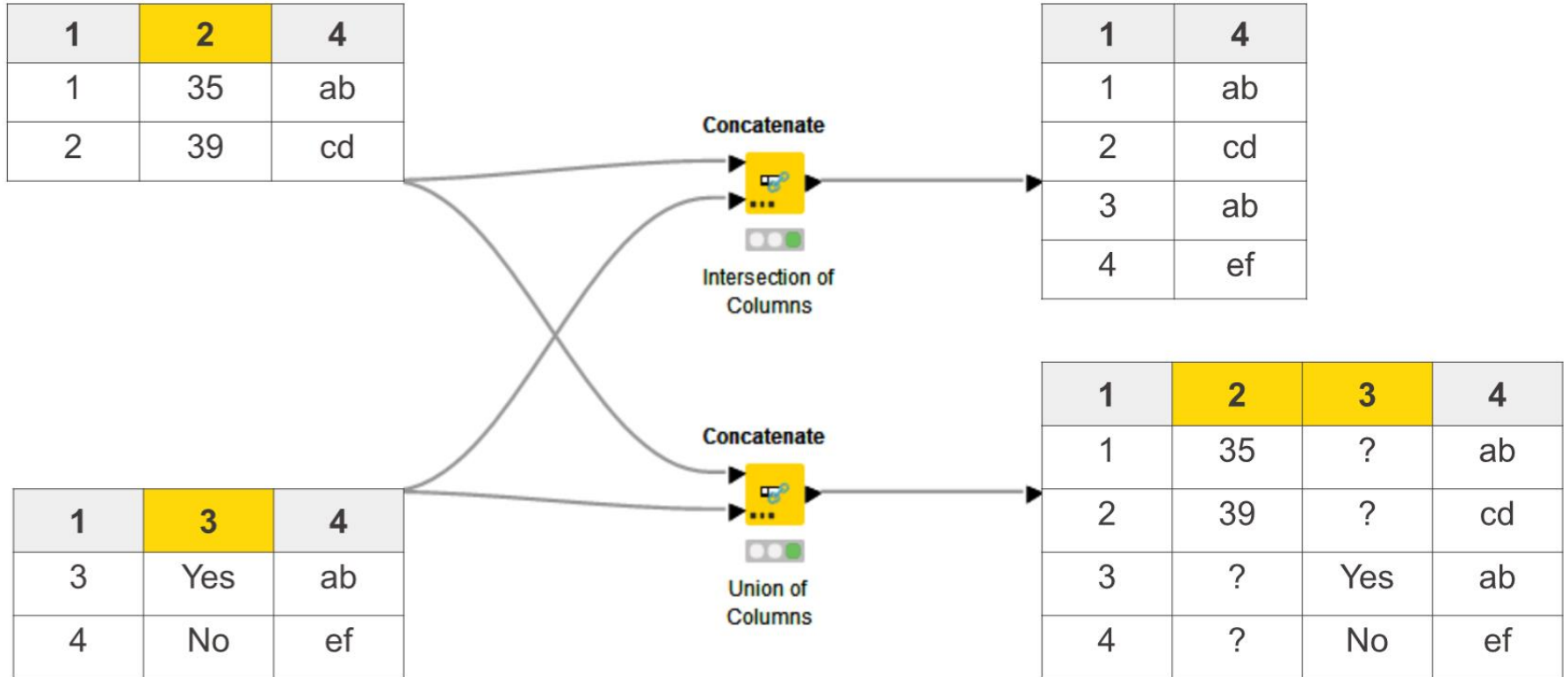
- The top port references the table you want to search
- Bottom port holds the lookup table (search keys and replacement values)



The screenshot shows the configuration dialog for the Cell Replacer node. The tabs at the top are Options, Flow Variables, Job Manager Selection, and Memory Policy. The configuration is as follows:

- Input table:** Target column is set to "Email".
- Dictionary table:** Input (Lookup) is set to "email" and Output (Replacement) is set to "replacement".
- Dictionary matching behaviour (only applicable for Strings):** Matching behaviour is set to "Exact" (selected), with "Substring", "Wildcard", and "Regex" also available. The "Case sensitive" checkbox is checked.
- (Additional) Result Columns:** The "Append result as new column" checkbox is unchecked, with "Replacement" in the adjacent text field. The "Create additional 'found' / 'not found' column" checkbox is also unchecked, with "Found" set to "found" and "Not Found" set to "not found".
- If no element matches use:** "Input" is selected (radio button), with "Missing" also available.
- Metadata in Output:** The "Copy metadata from replacement column" checkbox is checked.

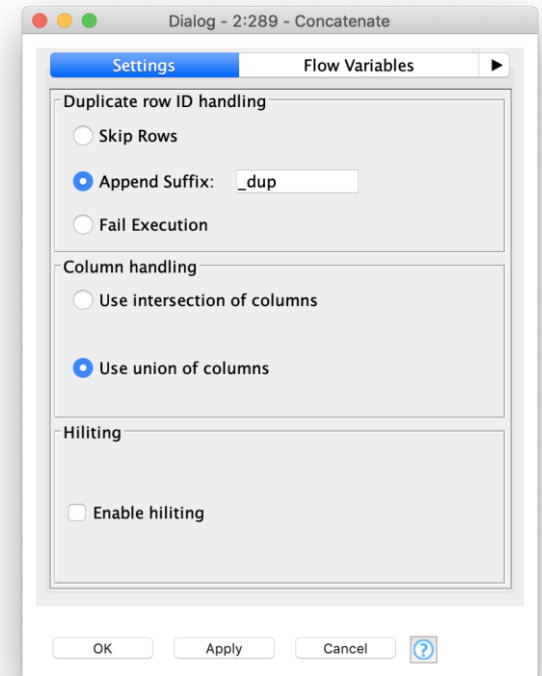
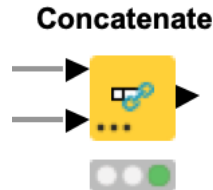
Concatenate two tables



Concatenate

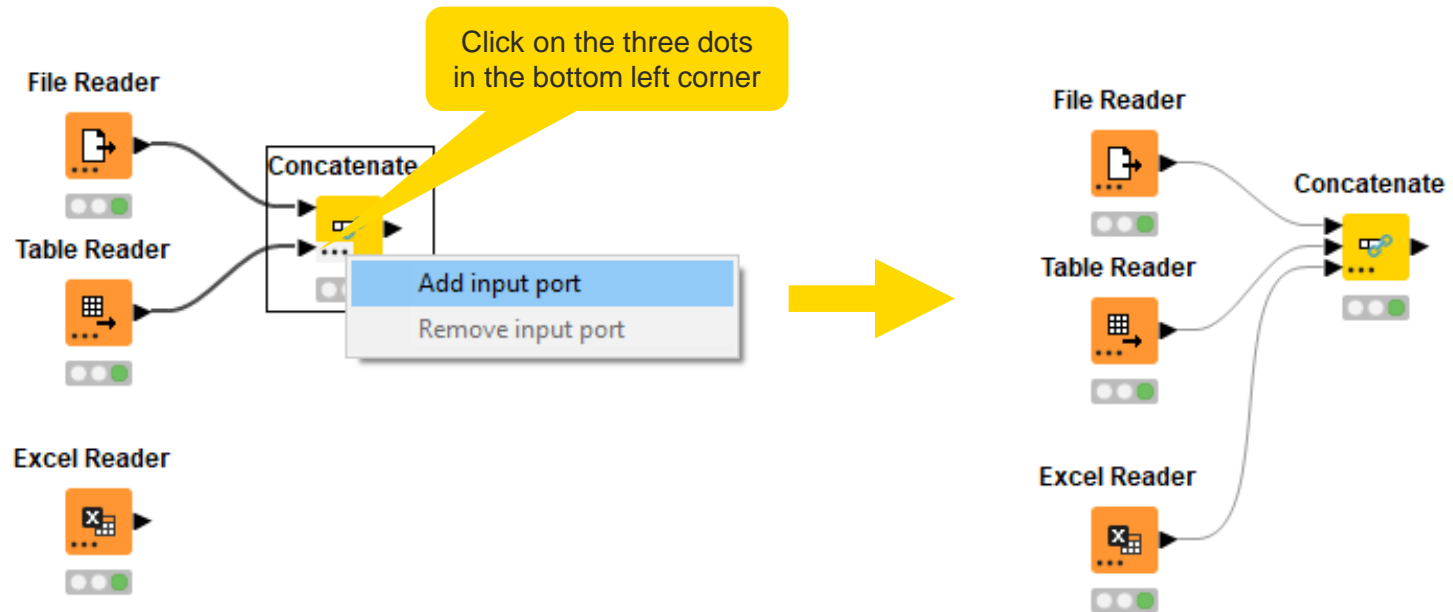
Combine rows from two tables with shared columns

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



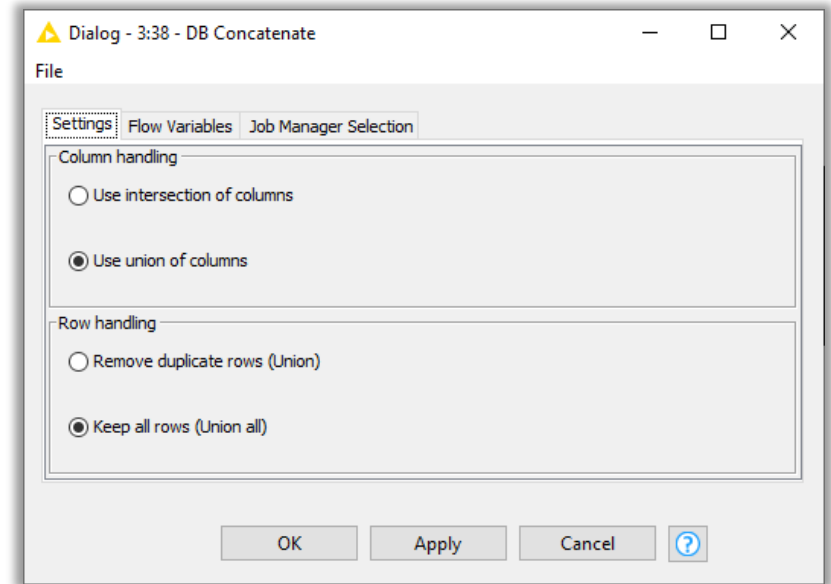
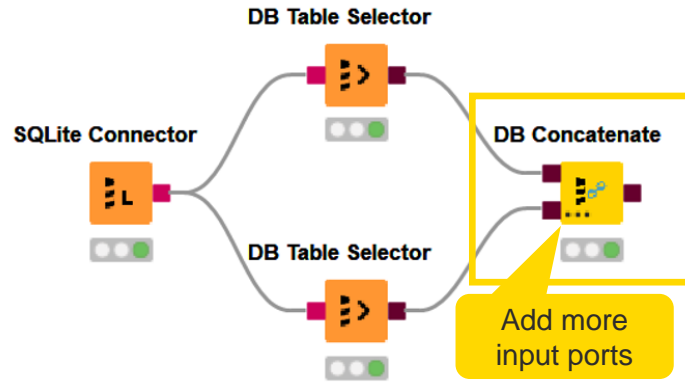
Dynamic Ports

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



DB Concatenate

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



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	2019-10-13	#29999

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

Join by CustomerKey

Inner Join

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

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

Join by CustomerKey

Full Outer Join

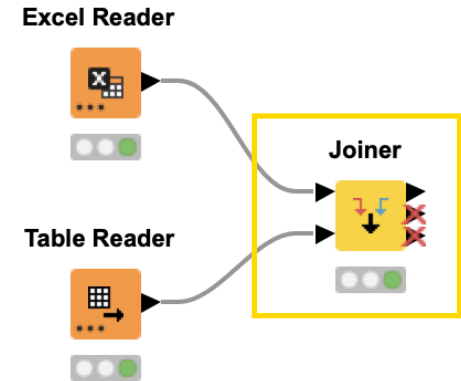
CustomerKey	OrderDate	OrderID	DoB	City	Gender
17	?	?	1974-02-23	Berlin	F
65	?	?	2001-05-25	Stuttgart	F
35	?	?	1988-08-05	Cologne	M
15	2019-10-07	#28985	1983-07-20	Hamburg	M
10	2019-10-13	#29999	1993-01-13	Berlin	M
22	2019-09-23	#23444	?	?	?
24	2019-09-30	#23457	?	?	?

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

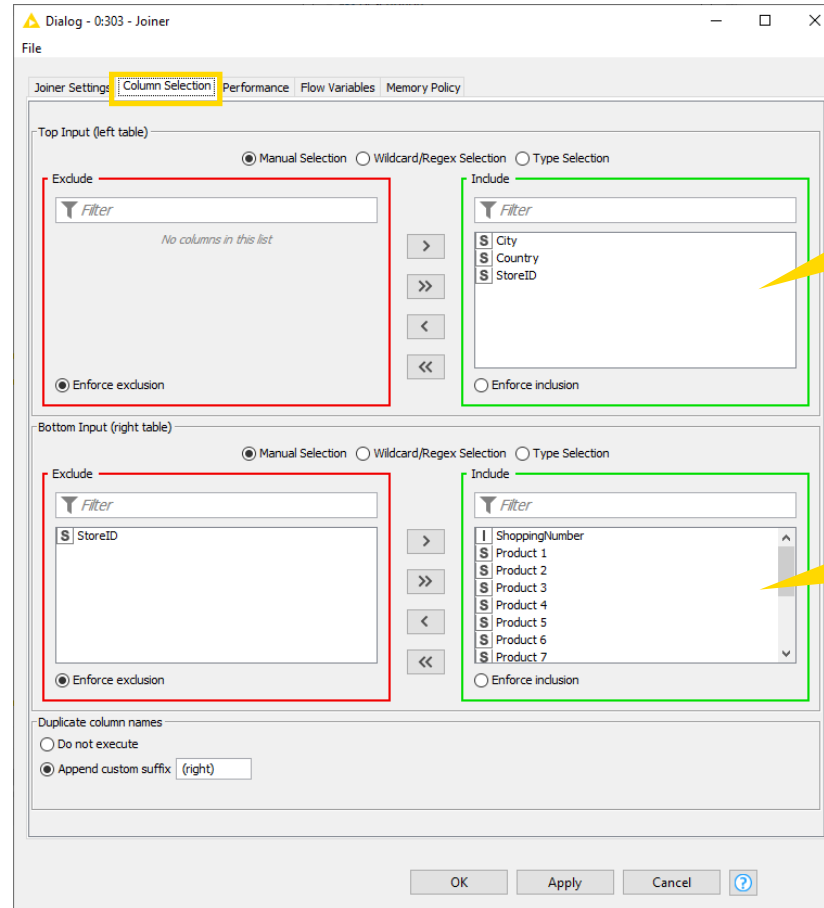
Values to join on.
Multiple joining columns
are allowed

Select the rows which
should be included in the
joined table

Activate this checkbox to
activate the bottom
output ports

The screenshot shows the 'Joiner' dialog box in KNIME. The 'Joiner Settings' tab is active. The 'Join columns' section is highlighted with a yellow box. It shows 'Match' set to 'all of the following' and 'Top Input (left table)' and 'Bottom Input (right table)' both set to 'StoreID'. Below this, there are checkboxes for 'Include in output': 'Matching rows' (checked), 'Left unmatched rows' (unchecked), and 'Right unmatched rows' (checked). A Venn diagram labeled 'Right outer join' is shown. The 'Output options' section has 'Split join result into multiple tables' (unchecked), 'Merge join columns' (unchecked), and 'Hilting enabled' (unchecked). The 'Row Keys' section has 'Concatenate original row keys with separator' (checked) and 'Assign new row keys sequentially' (unchecked). The dialog has 'OK', 'Apply', 'Cancel', and a help icon at the bottom.

Joiner Configuration – Column Selection

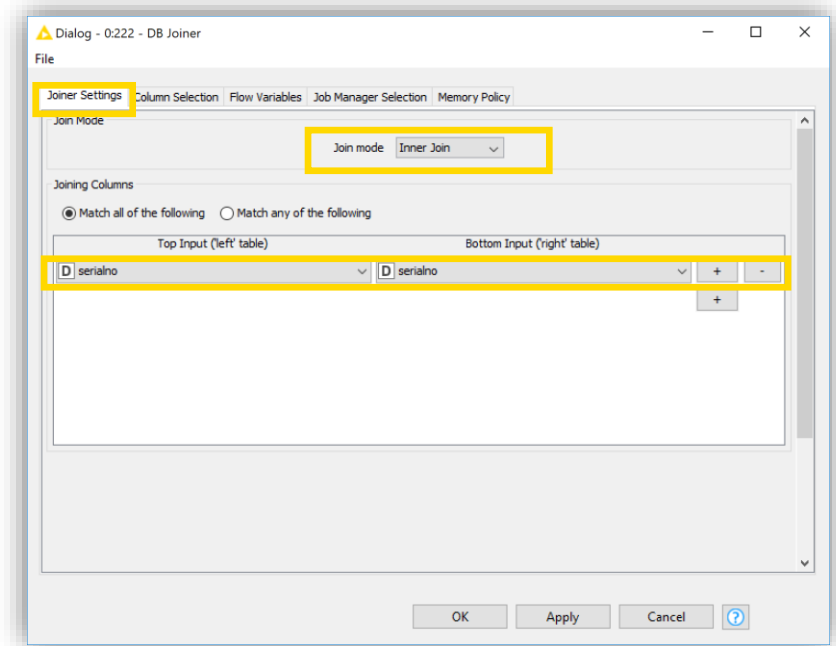
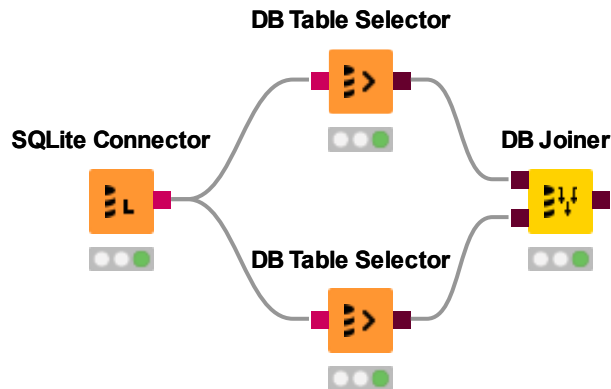


Columns from top table for joined table

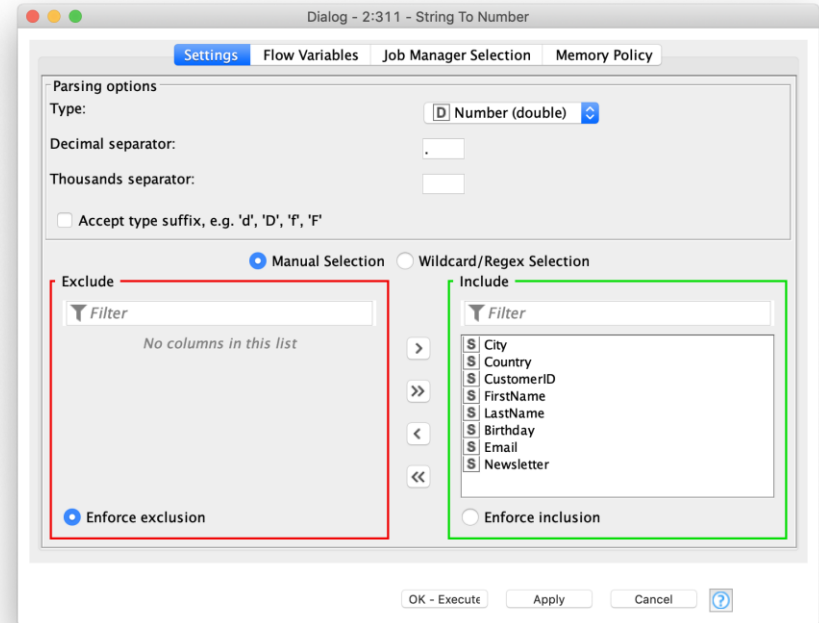
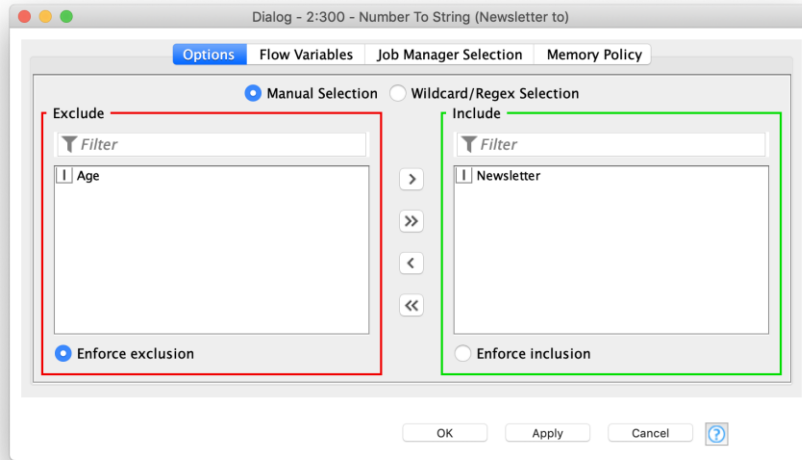
Columns from lower table for joined table

DB Joiner

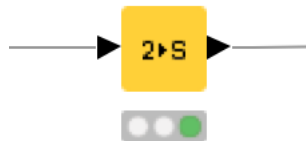
- In-database joiner
- Creates the SQL statement to join two tables stored in the same database
- No coding required



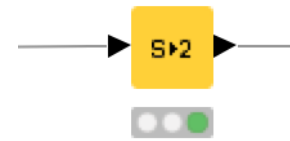
Type Conversion



Number To String

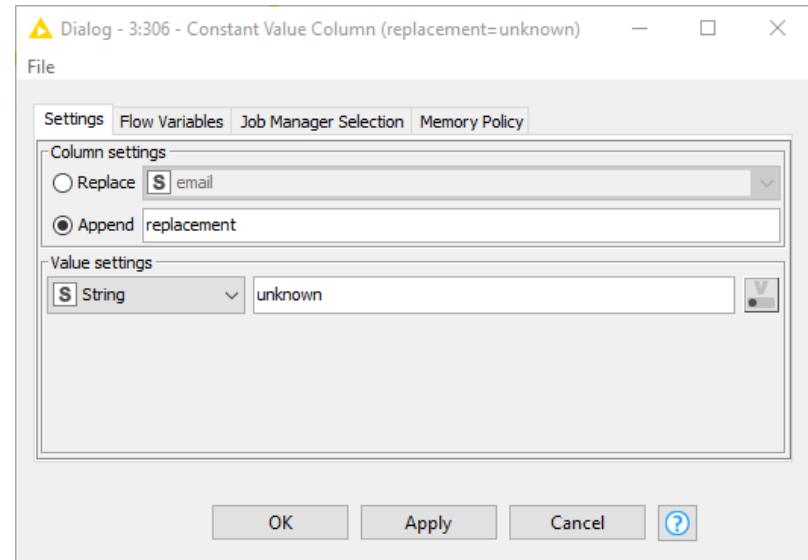
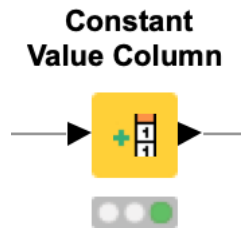


String To Number



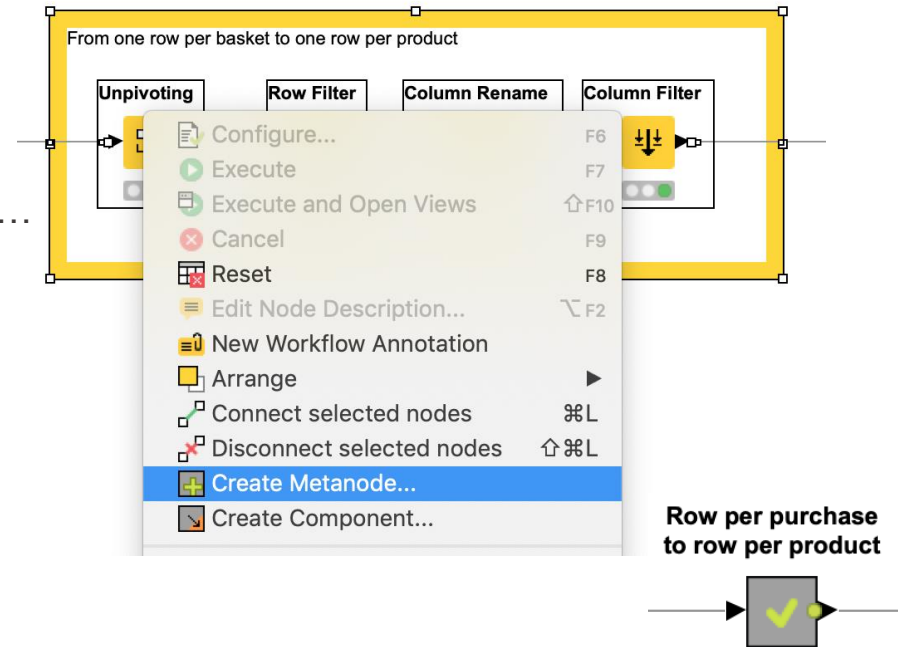
Constant Value Column

- Adds or replaces a column with a single constant value
- Can be used to add an empty column



Workflow Organization – Good Practices

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



Exercise: 02_Data Merging

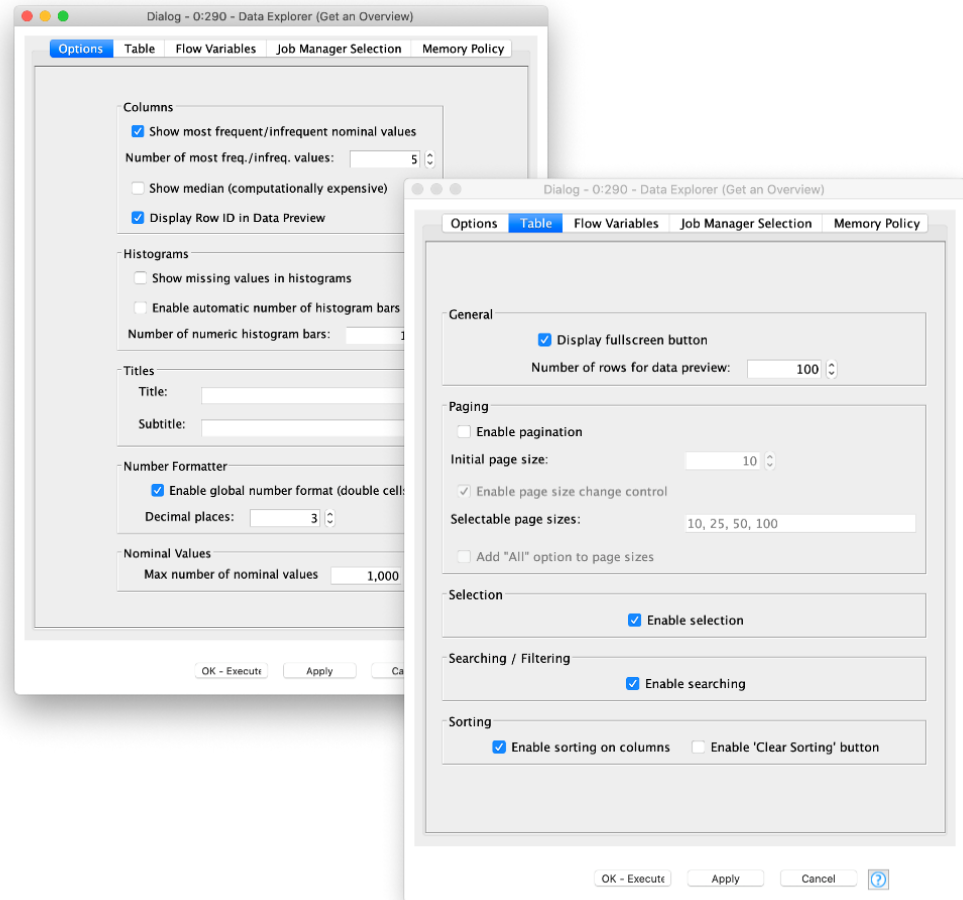
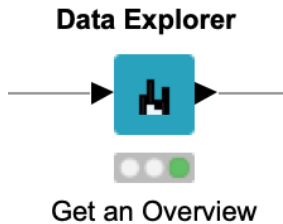
- **Concatenate** the customer information from the two systems
- Perform lookup operation to manipulate selected values in the data:
 - Read a table with a few emails
 - Add a constant value column to the table, with the value “unknown”
 - Replace the provided email addresses in the customer information table with “unknown” using Cell Replacer
- 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

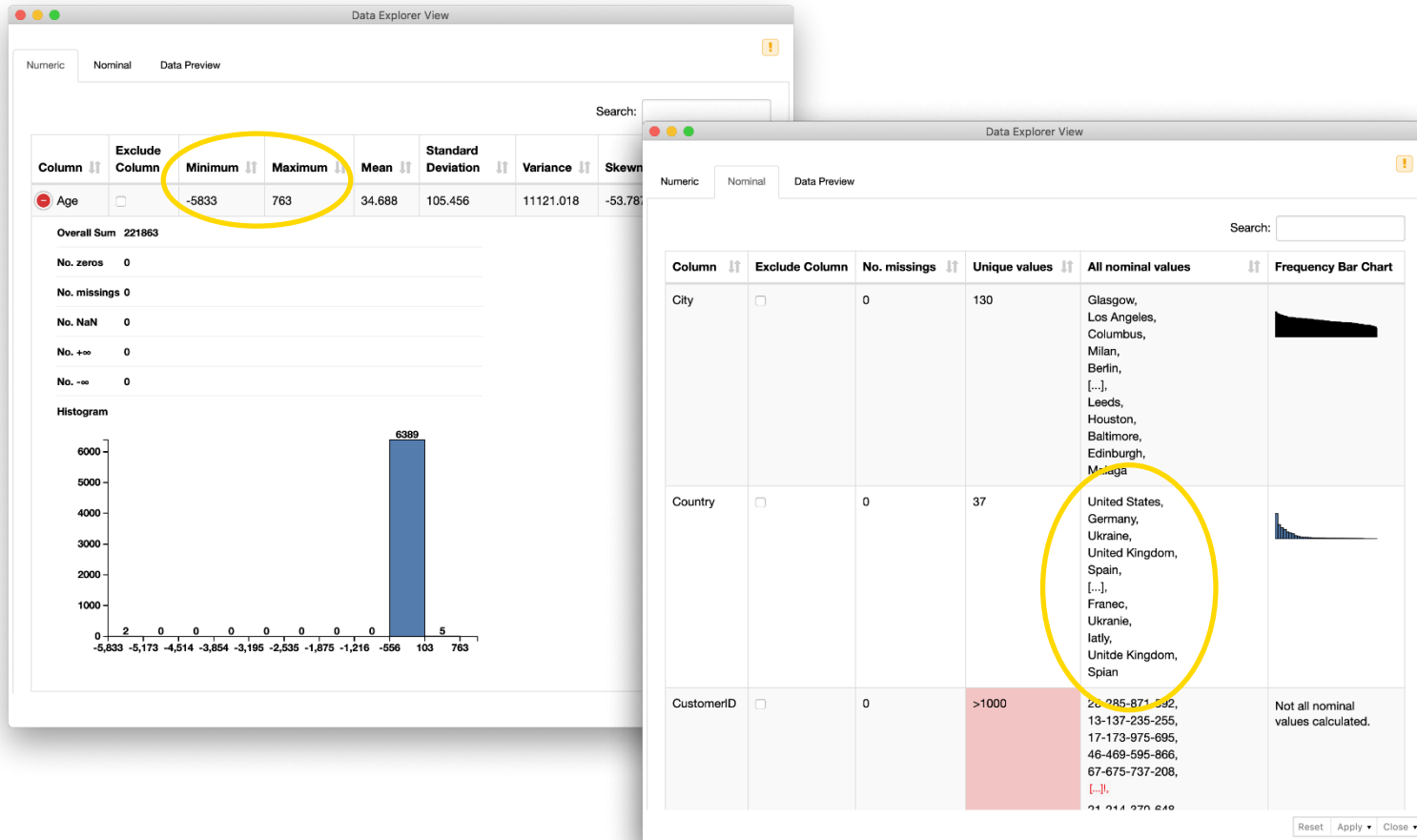


Data Explorer

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

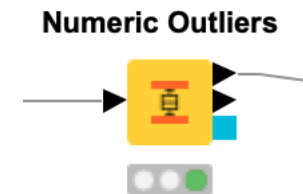
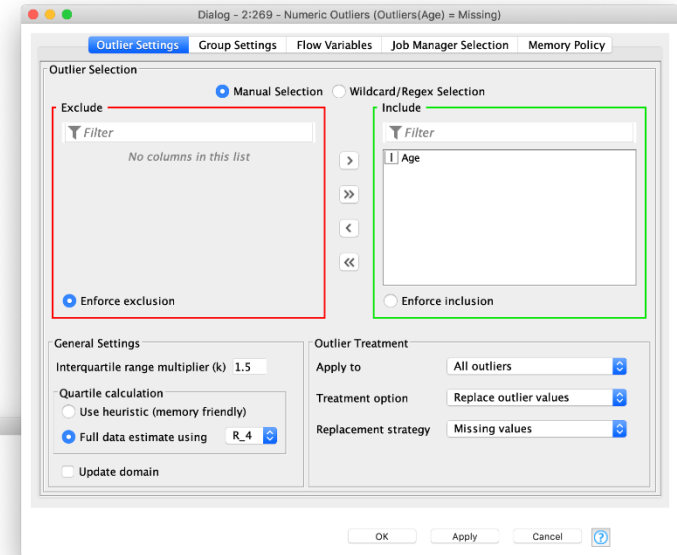
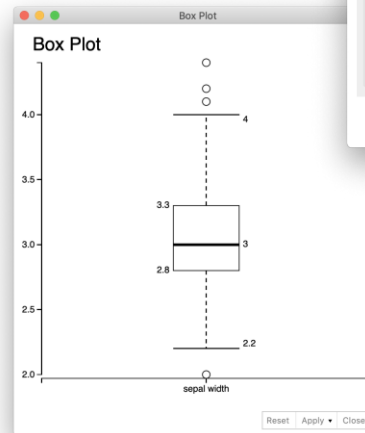


Customer Data Output



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 borders correspond to the whiskers of a box plot

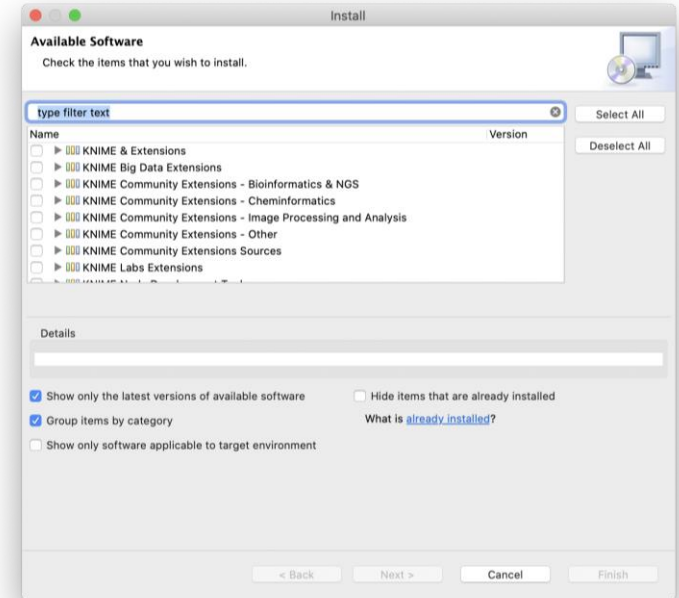
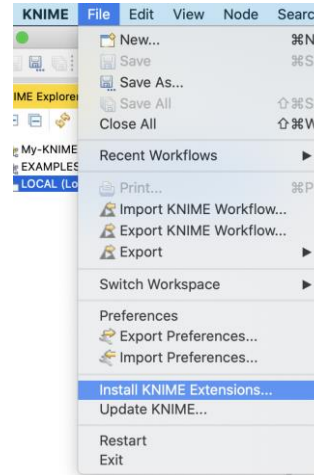


Calculating the Quantiles - Example

- The quantiles Q_1 , Q_2 , and Q_3 are the three cut points that divide a dataset into four equal-sized groups, after sorting them.
- Example: 3, 6, 7, 8, 8, 10, 13, 15, 16, 20
 - Q_1 : Rank = $10 \times (1/4) = 2.5$, rounds up to 3 $\Rightarrow Q_1 = 7$
 - Q_2 : Rank = $10 \times (2/4) = 5$, $\Rightarrow Q_2 = (8+10)/2=9$
 - Q_3 : Rank = $10 \times (3/4) = 7.5$, rounds up to 8 $\Rightarrow Q_3 = 15$

More Nodes – Installing Extensions

- Go to File -> Install Extensions
- Select the following extensions:
 - KNIME JavaScript (Labs)
(in the folder KNIME Labs Extensions)
 - KNIME Data Generation
(in the folder KNIME & Extensions)



Alternative Way via the KNIME Community Hub

- Drag & Drop extension from the KNIME Community Hub into KNIME Analytics Platform.

The image shows two screenshots illustrating the process of installing an extension from the KNIME Community Hub into the KNIME Analytics Platform.

Left Screenshot (KNIME Community Hub): The browser window shows the KNIME Community Hub page for the "KNIME Data Generation" extension. The extension is highlighted with a dashed box, and a yellow arrow points from it to the right screenshot.

Right Screenshot (KNIME Analytics Platform): The KNIME Analytics Platform interface shows a workflow titled "My first Workflow" with the following nodes: File Reader (read adult.csv), Row Filter (keep only records born in the US), Column Filter (remove gender), and Table Writer (Write table). A yellow box highlights the Row Filter node, and a yellow arrow points from the extension in the left screenshot to this node. The Row Filter node's configuration dialog is open, showing the "Row Filter" settings.

Row Filter Configuration Dialog:

The node allows for row filtering according to certain criteria. It can include or exclude certain ranges (by row number), 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

ID	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relational
Row1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband
Row2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-fa
Row3	53	Private	224721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband
Row5	37	Private	234652	Masters	14	Married-civ-spouse	Exec-managerial	Wife
Row7	52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband

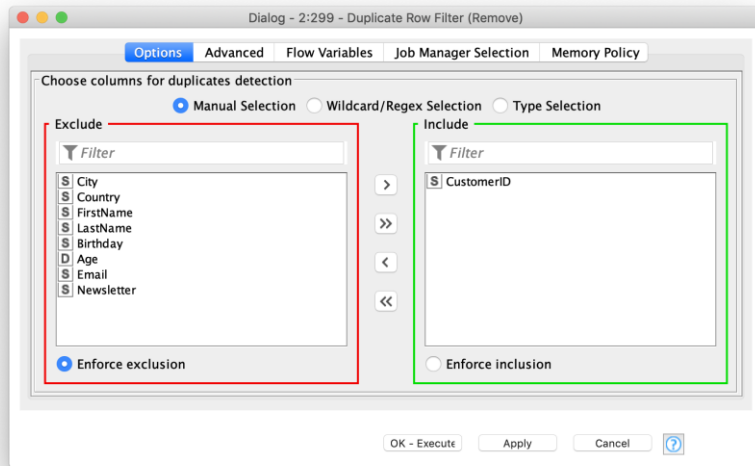
Duplicate and Missing Value Handling



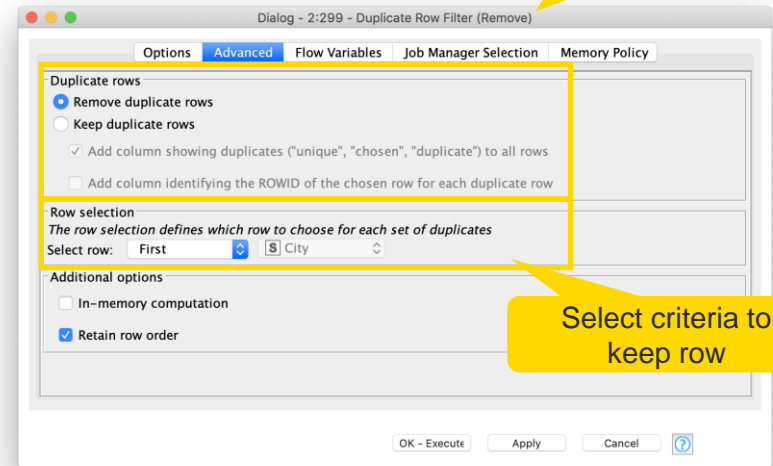
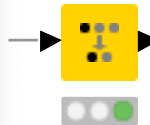
Duplicate Row Filter

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



Duplicate Row Filter

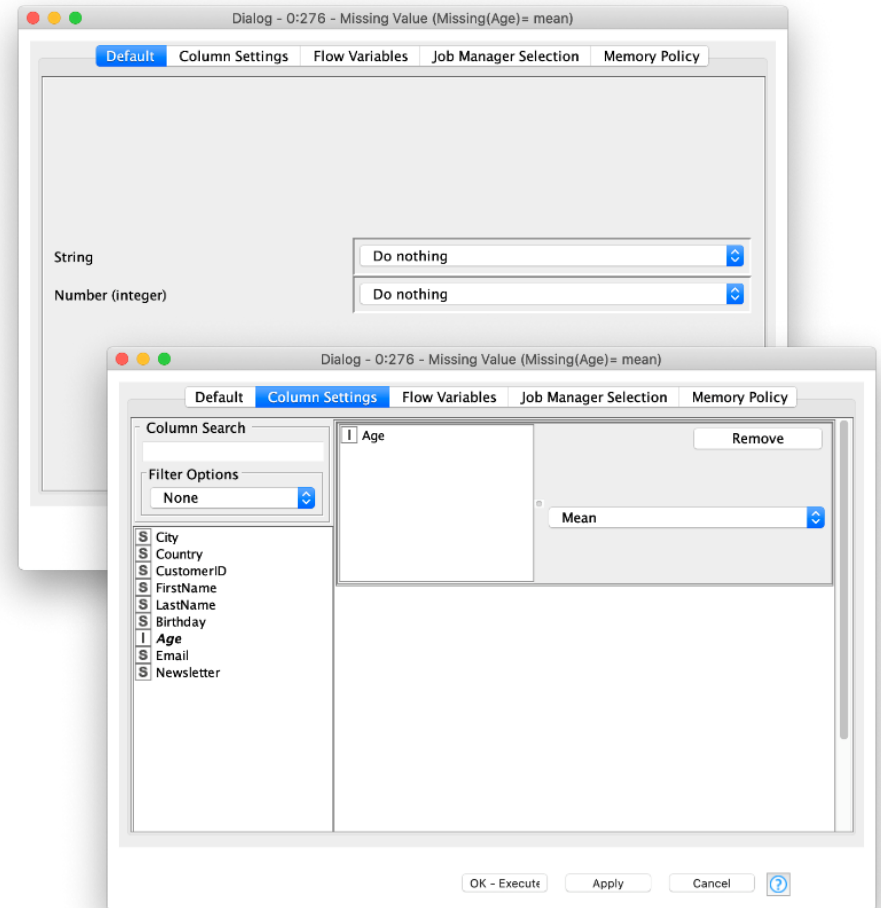
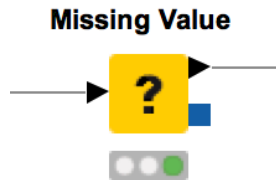


Flag or Remove Duplicates

Select criteria to keep row

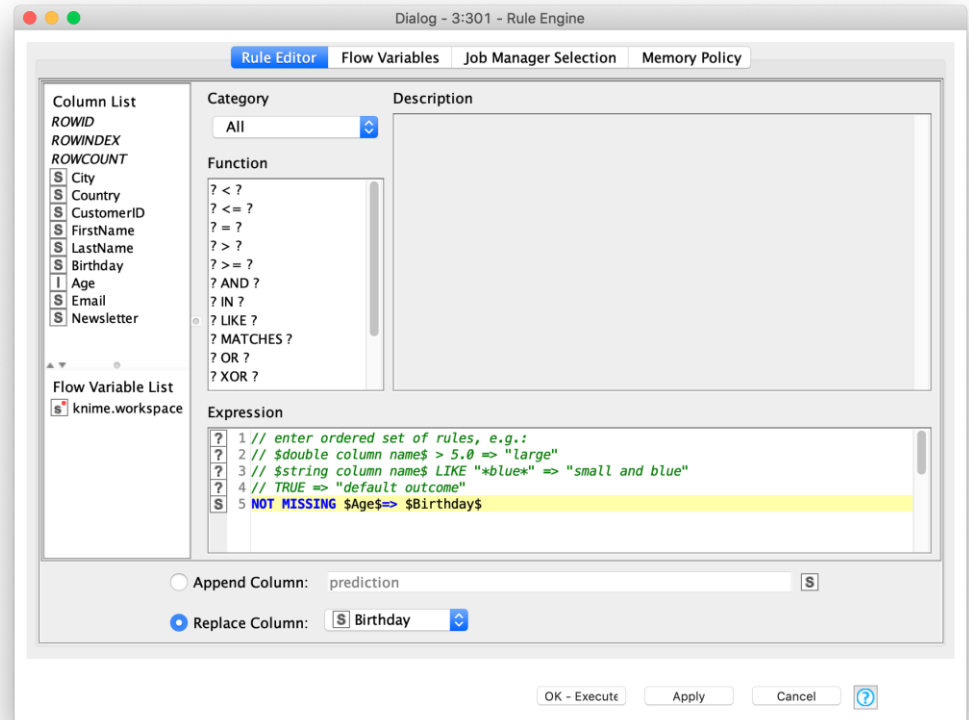
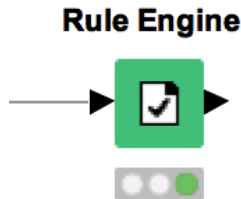
Missing Value

- 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

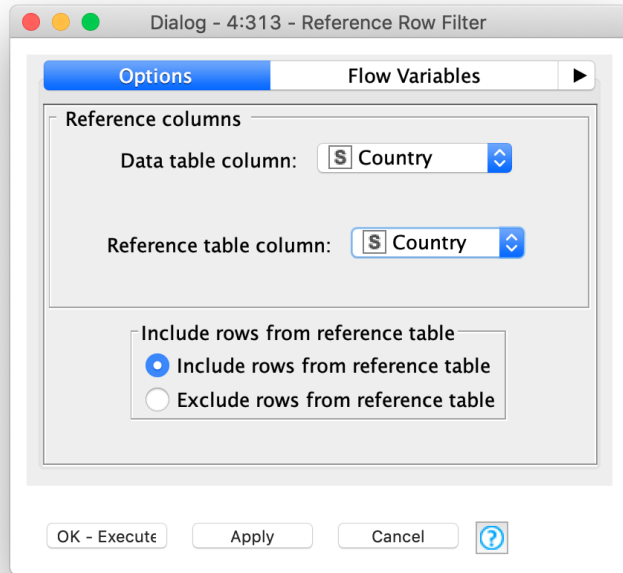


Rule Engine

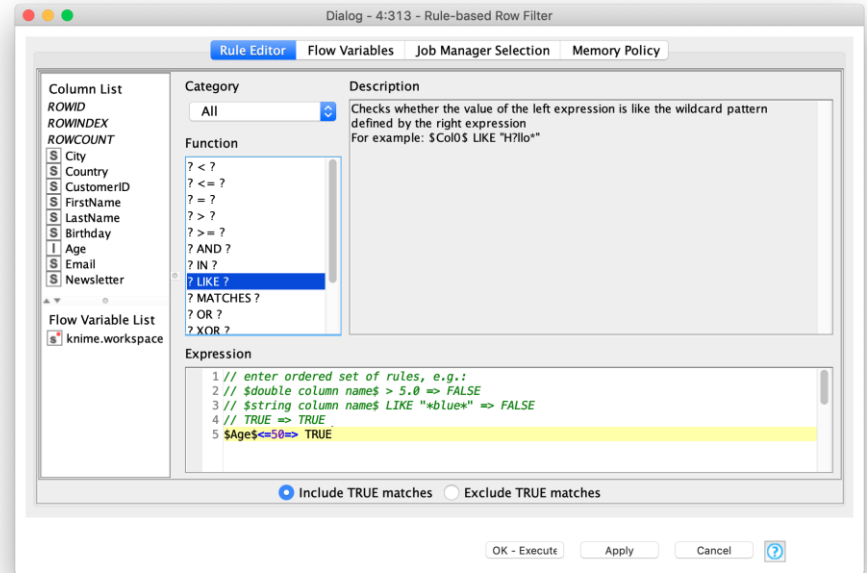
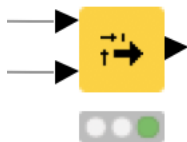
- Defines custom logic to use simple rules
- Rules like: **<Antecedent/Condition> => <Consequence>**
 - (1=1 => “true”)
- Tries to match rules to each row of the input table



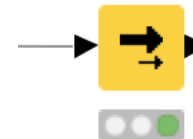
Other Options to Filter Rows



Reference Row Filter



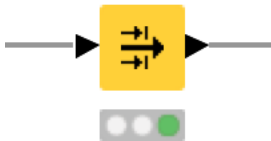
Rule-based Row Filter



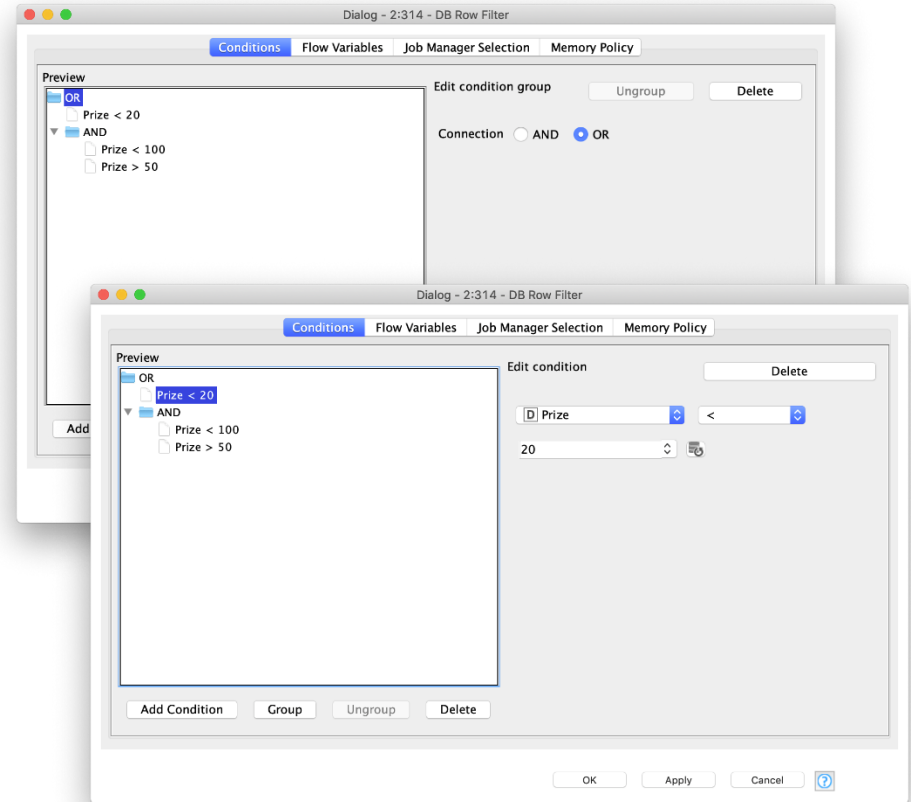
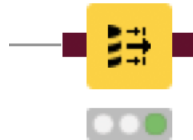
In-Database Row Filtering – DB Row Filter

- 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

Row Filter (Labs)

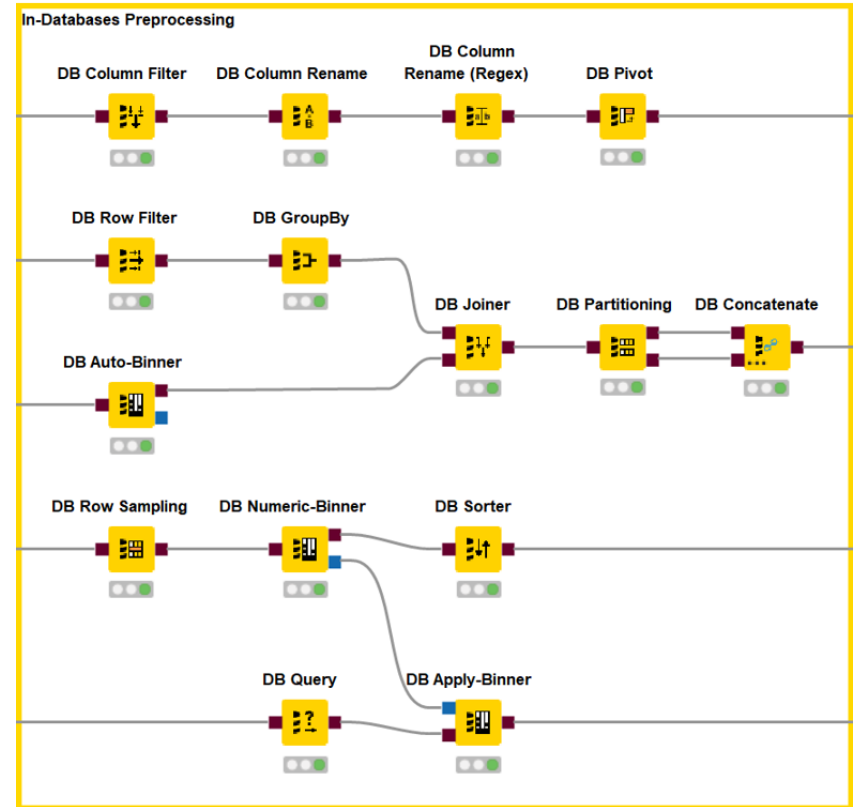


DB Row Filter



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



Standardization



Standardize Table Format

Onsite transaction data

Read table - 0:1 - Table Reader

Table "default" - Rows: 20000 Spec - Columns: 16 Properties Flow Variables

Row ID	Shoppi...	Product 1	Product 2	Product 3	Product 4	Product 5	Prod...
Row900	480105758	V-220-2011	A-43-2005	W-52-2005	?	?	?
Row901	480105759	Q-166-2005	I-184-2010	?	?	?	?
Row902	480105760	V-181-1983	W-52-2018	?	?	?	?
Row903	480105761	N-160-1983	I-184-1989	Z-280-2003	C-151-2015	A-91-2002	?
Row904	480105762	H-172-1991	V-265-1991	H-235-1996	H-124-1998	Q-43-2015	?

Product Nr & Price

KNIME data table - 0:4 - DB Reader

Table "database" - Rows: 50700 Spec - Columns: 2 Properties Flow Variables

Row ID	Price	ProductNr
Row45899	66.99	C-061-2010
Row45900	19.99	V-061-2010
Row45901	110.59	C-061-2010
Row45902	63.99	N-061-2010
Row45903	54.99	Q-064-2010
Row45904	48.99	W-064-2010

Online transaction data

KNIME data table - 0:6 - DB Reader

Table "database" - Rows: 11679 Spec - Columns: 5 Properties Flow Variables

Row ID	OrderN...	Date	Customer...	ProductNr	Price
Row0	23893756	8-28-2015	69-695-442-229	I-165-2017	12.99
Row1	23893756	8-28-2015	69-695-442-229	B-172-2005	58.99
Row2	23893756	8-28-2015	69-695-442-229	W-181-2003	75.59
Row3	23893756	8-28-2015	69-695-442-229	F-055-2017	111.99
Row4	23893756	8-28-2015	69-695-442-229	F-289-2008	37.59
Row5	23893756	8-28-2015	69-695-442-229	F-289-1990	47.99

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

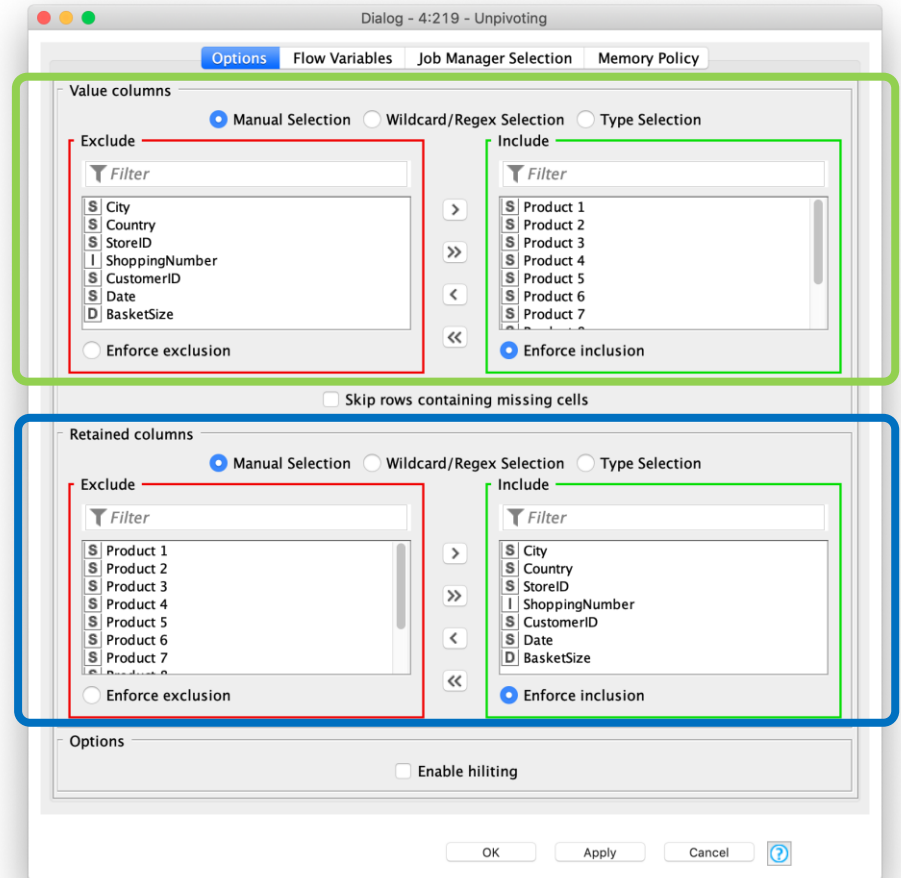
Value Column

Retaining Columns

Solution: Unpivoting Node

Unpivoting

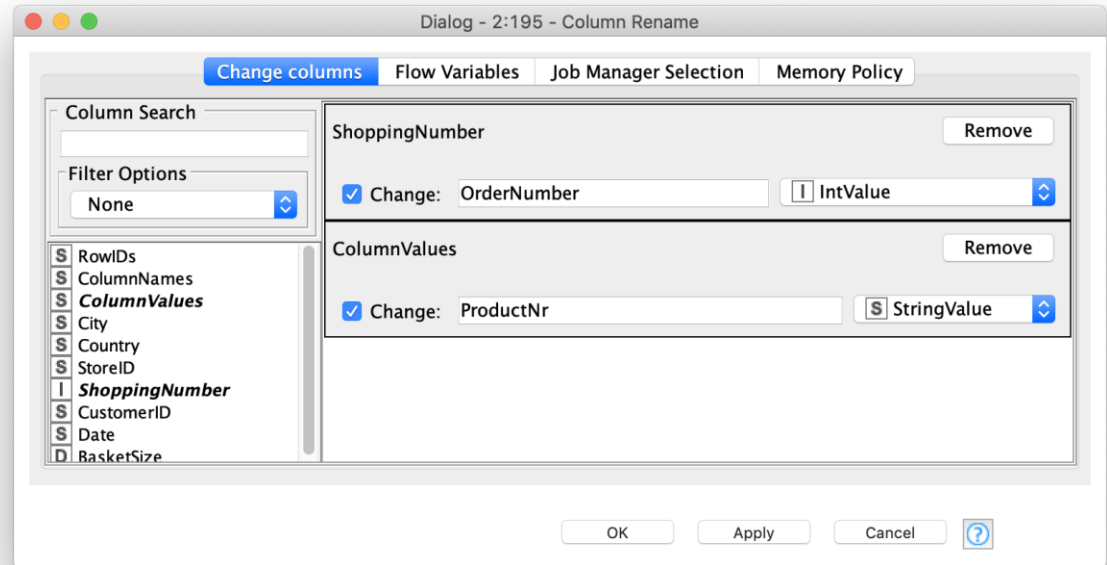
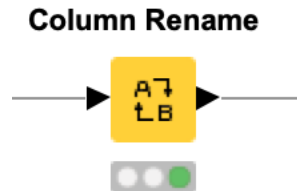
- Rotates the value columns to rows
- Duplicates the remaining columns and appends them to each corresponding row



Value Column
Retaining Columns

Column Rename

- Renames column names or changes their types



Column Filter

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

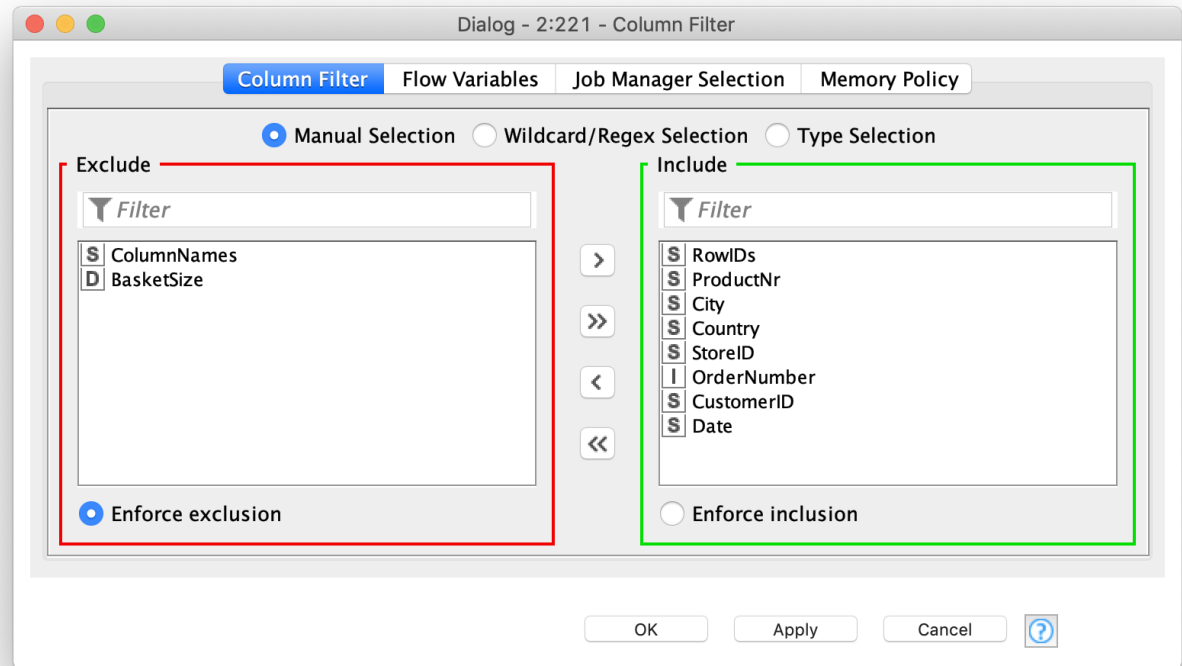
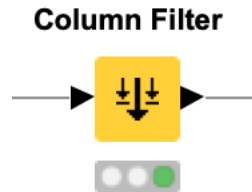
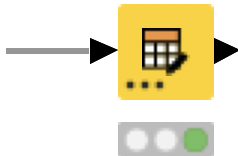


Table Manipulator

Allows for

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

Table Manipulator



Dialog - 0:3 - Table Manipulator

File

Settings Flow Variables Memory Policy

Row ID handling
 Use existing row ID Prepend table index to row ID

Transformations
Reset actions ↑ Move up ↓ Move down 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		String → Local Time
☑	S LastName		D String → Number (double)
☑	S Birthday		I String → Number (integer)
☑	I Age		L String → Number (long)
☑	I Email		HL String → PMML
☑	I Newsletter		String → Period
☑	I ?		String → SVG image
☑	? <any unknown new column>		S String
☑			? ?

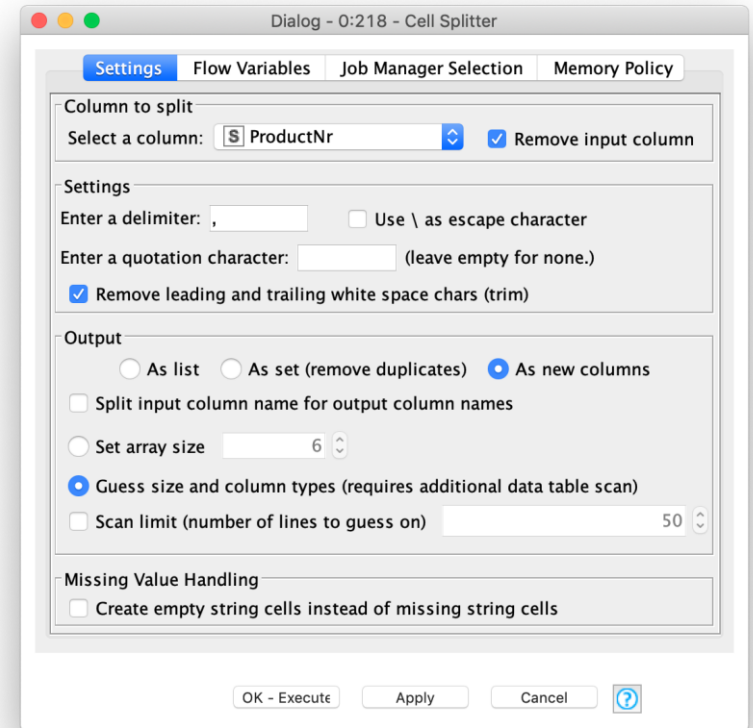
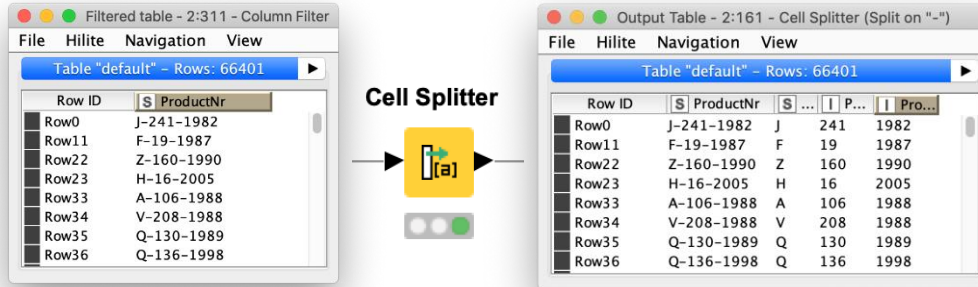
Preview
✔ Data analysis successfully completed.

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@mcrc.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	Gräßlwald	20.4.1975	44	Rotraut.Gräßlwald...	0

OK Apply Cancel ?

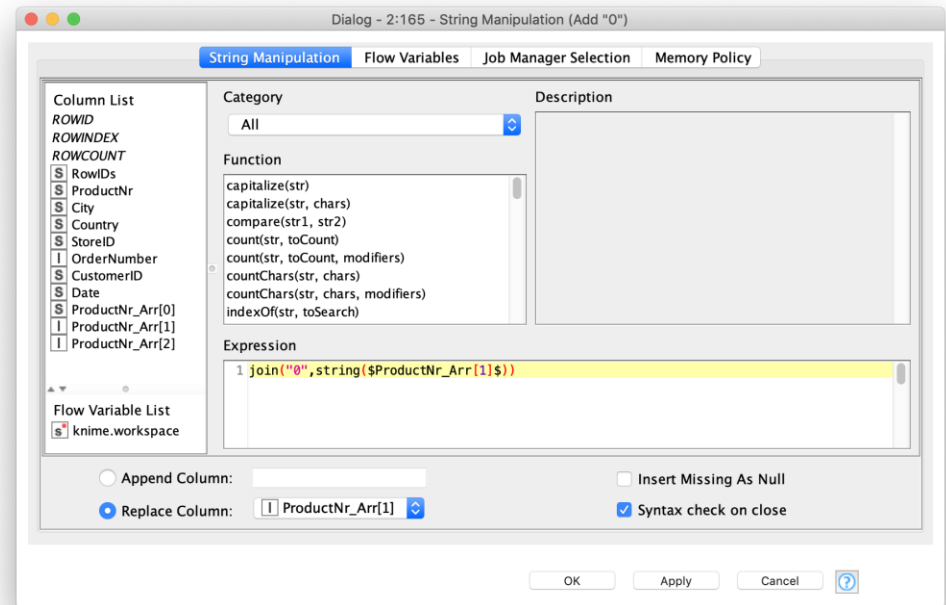
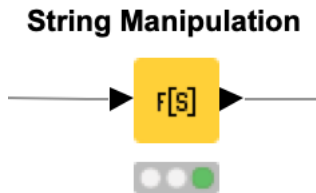
Cell Splitter

Splits the content of one column into many columns based on a delimiter

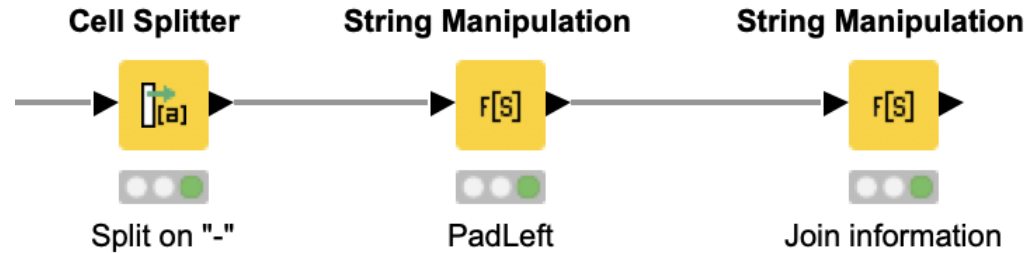


String Manipulation

- Modifies existing strings or creates new columns
 - Cleans up capitalization
 - Joins string values
 - Pads strings, e.g. padLeft
 - Replaces string values



Standardization of Product Numbers



Exercise: 03_Data Cleaning (Part 1 & 2)

- Explore the data using the Data Explorer node
- Replace numeric outliers in the “Age” column with missing values

Exercise: 03_Data Cleaning (Part 3 & 4)

- Replace the birthday with a missing value if age is missing
Hint: Use the expression NOT MISSING \$Age\$=> \$Birthday\$
- Replace the missing values in the age column with the column mean
- Remove rows for duplicate CustomerIDs
- Correct the spelling mistakes in the “Country” column (Optional)
 - Extract the values with spelling mistakes
 - Manually define the correct spelling for the lookup table
 - Create the lookup table automatically, using a similarity search

Exercise: 04_Data_Transformation

- List purchases of different products in rows instead of columns
 - Unpivot the columns that show the products ordered in one purchase event. Retain other columns in the table.
 - Remove rows with missing values
 - Rename the "ColumnValues" column to "ProductNr" and "ShoppingNumber" to "OrderNumber" and remove unnecessary columns
 - Standardize the product numbers (Optional)

Exercise: 05_Data_Manipulation (optional)

- Join the price to the onsite product purchase data
- Add the transaction types to each product purchase
 - "Store - no CC" if the customer ID is not available in the onsite transaction
 - "Store - CC" if the customer ID is available in the onsite transaction
 - "OnlineStore" for the orders coming from the online store
- Concatenate the online and onsite purchases
- Join the customer information to each transaction

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



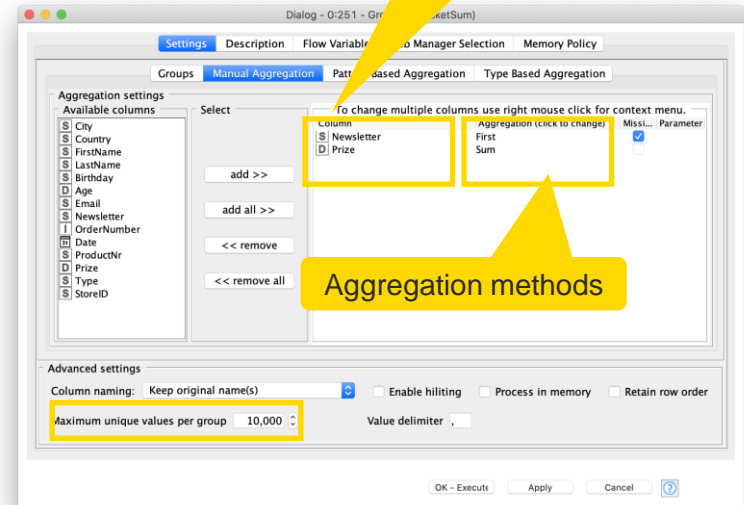
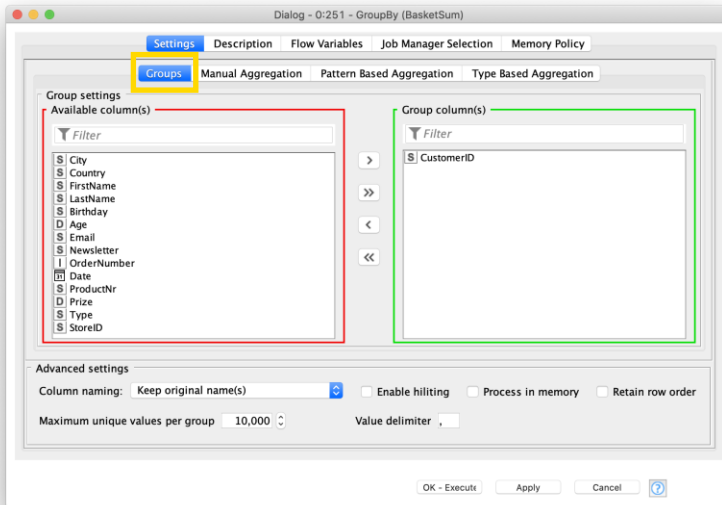
RowID	Group	Sum(Value)
r1+r3+r6	m	8
r2+r4+r5	f	15

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



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



Original Data Table:

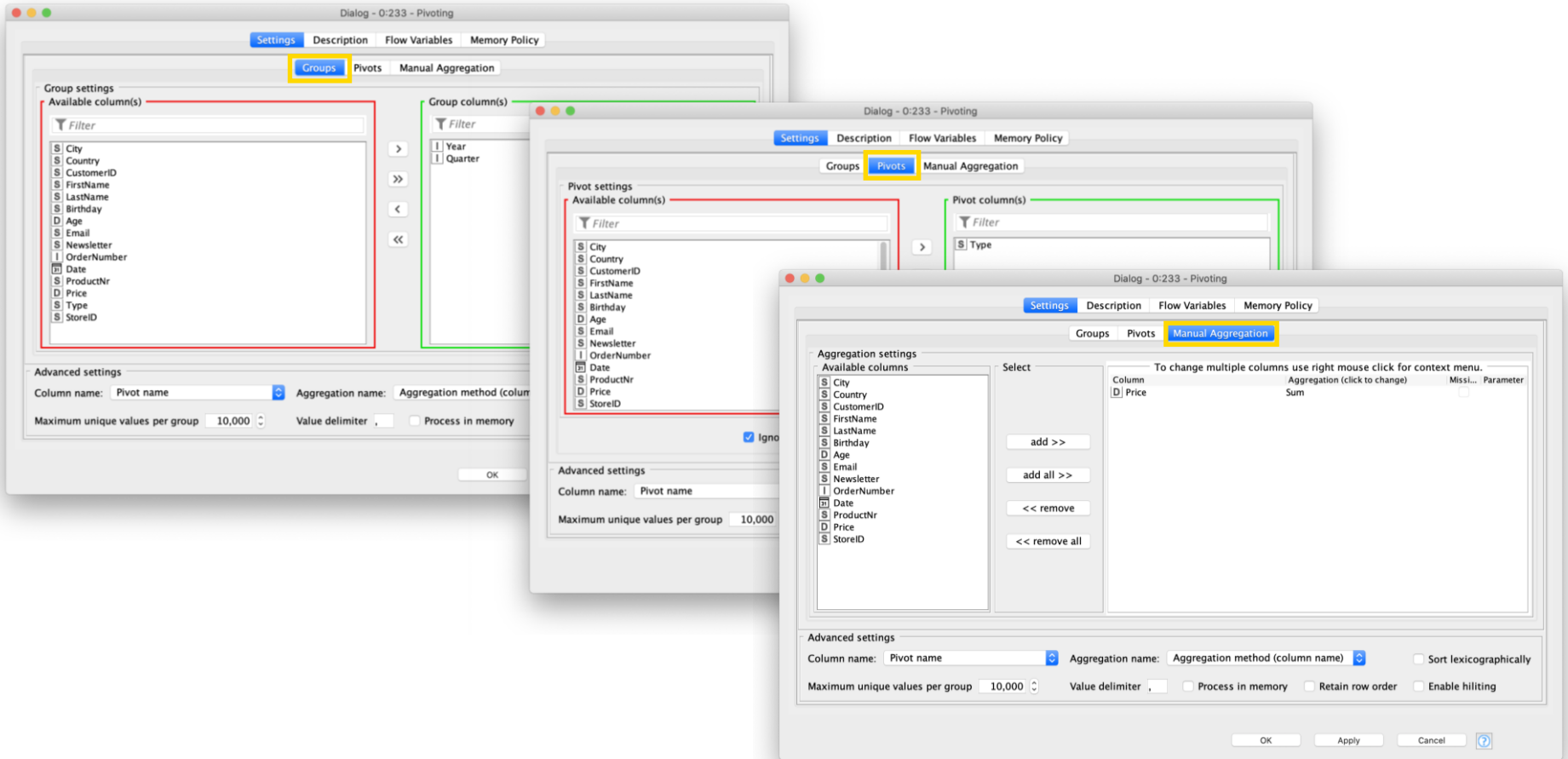
Gender	Hair	Age
f	blond	31
m	red	22
f	blond	53
m	brown	16
f	brown	47
f	black	22
m	blond	13
m	red	55

Aggregation: Mean(Age)

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

Pivoting Node: Group - Pivot - Aggregate

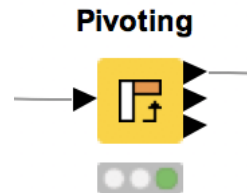
Pivoting



Pivoting

Performs pivoting on selected columns for grouping and pivoting

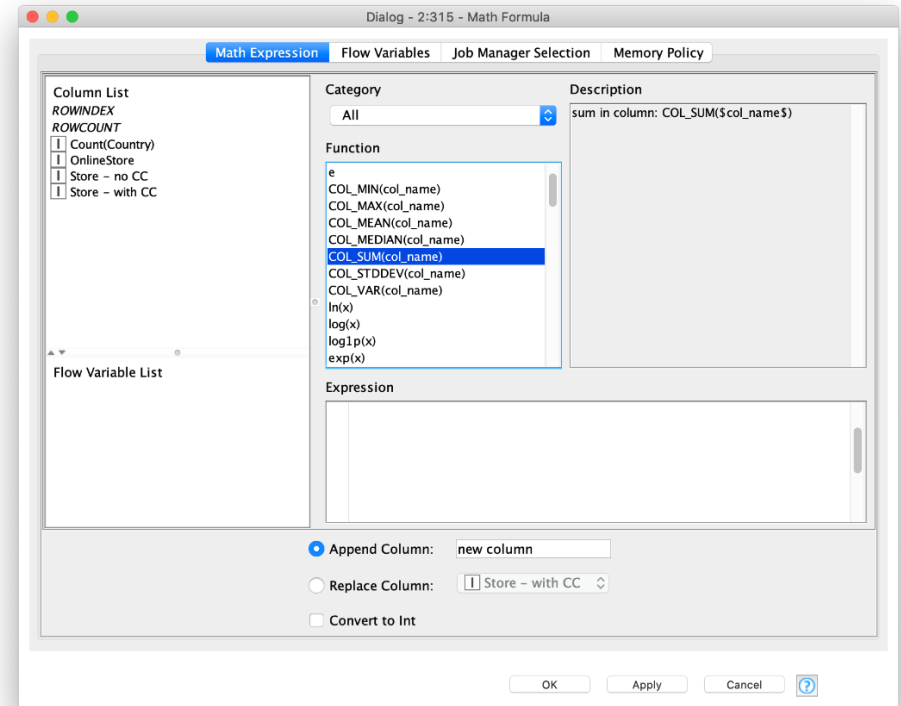
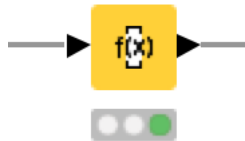
- Values of group columns become unique rows
- Values of the pivot columns become unique columns for each set of column combinations together with each aggregation
- Many aggregation methods are provided



Math Formula

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

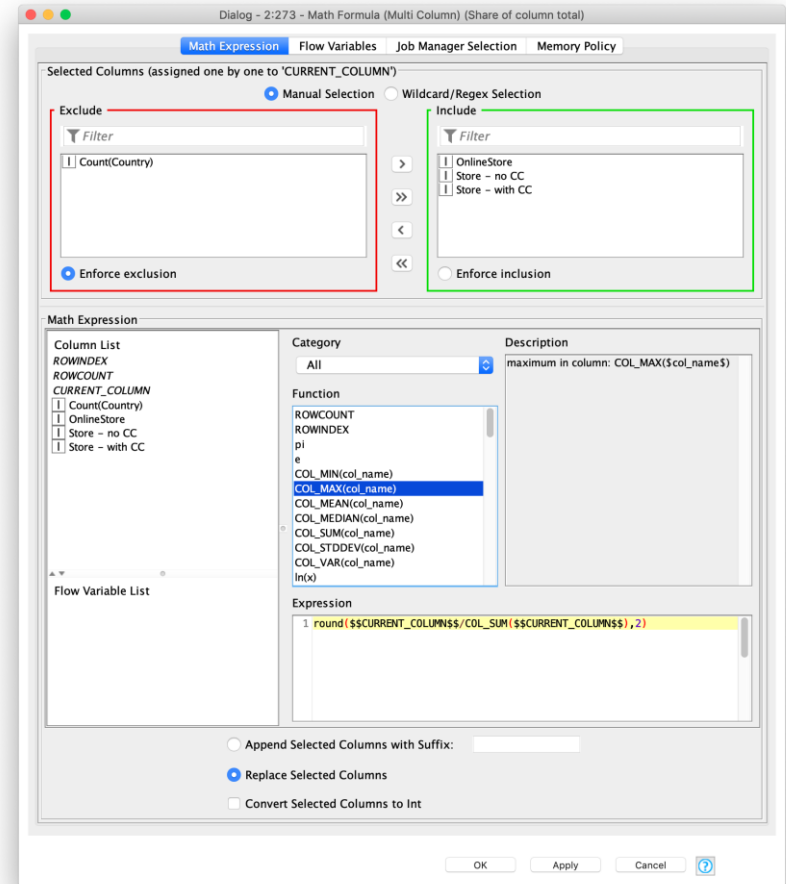
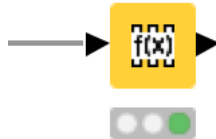
Math Formula



Math Formula (Multi Column)

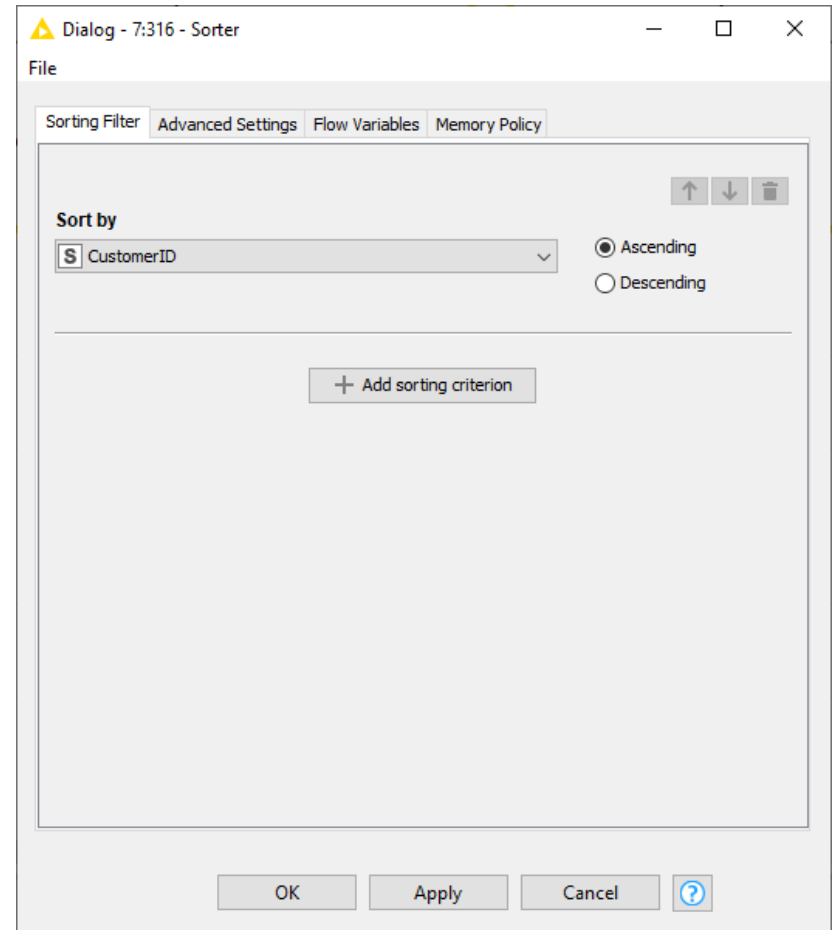
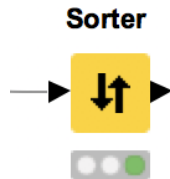
- Useful if you want to make the same calculations on multiple columns.
- The selected columns from the upper part are called `CURRENT_COLUMN` in the Column List and Expression dialog.

Math Formula (Multi Column)



Sorter

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



Exercise: 06_Data_Aggregation

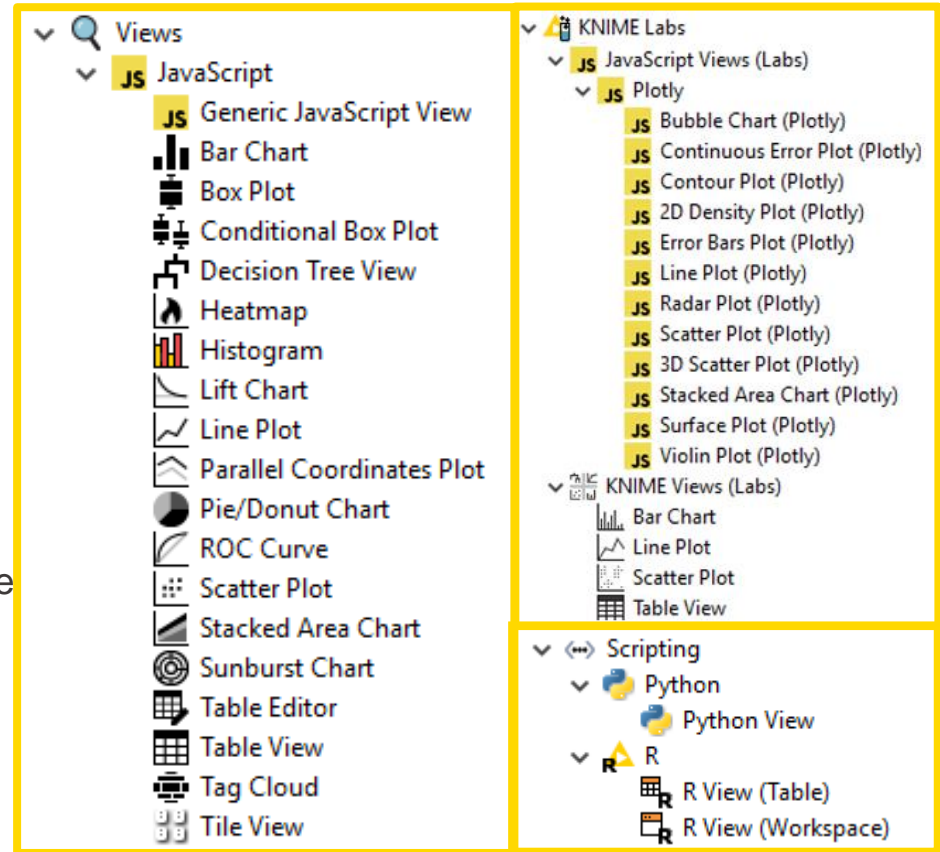
- Calculate the total purchase amount by a customer ID both in 2019 and overall
- Calculate the total purchase amount by quarter and transaction type
- Calculate the numbers of orders by basket size and transaction type (optional)

Data Visualization



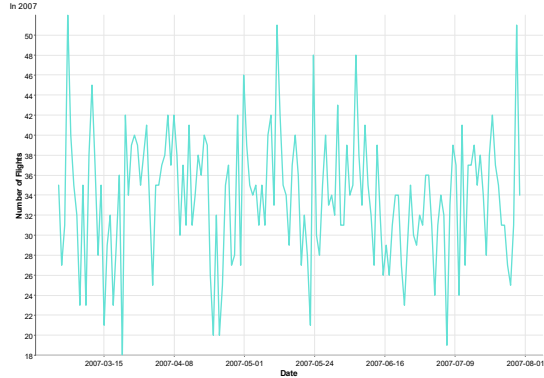
Data Visualization

- 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

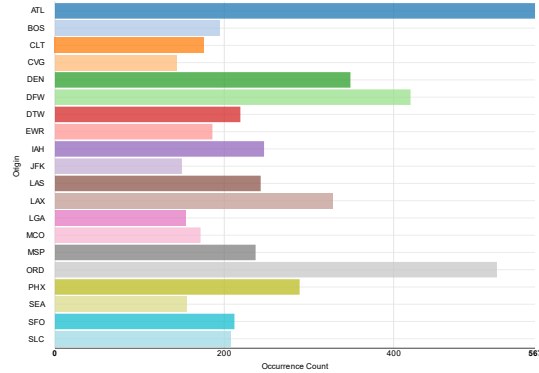


Visualizations Using One Column

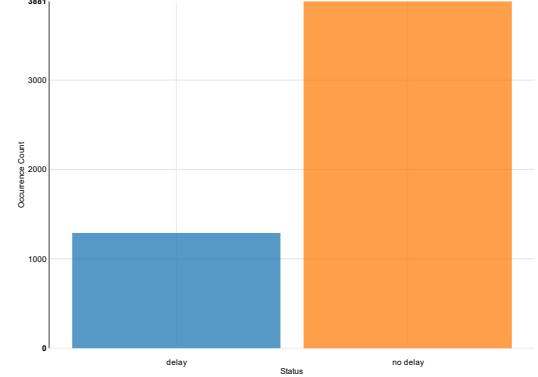
Number of Flights by Date



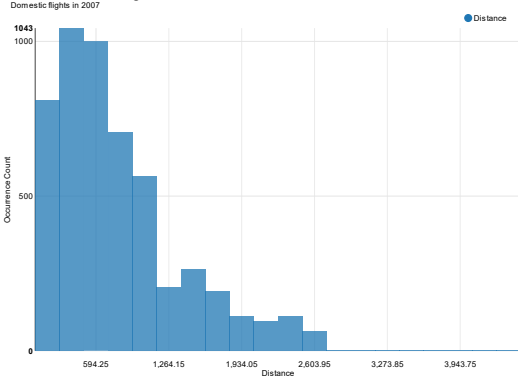
Departure Airports



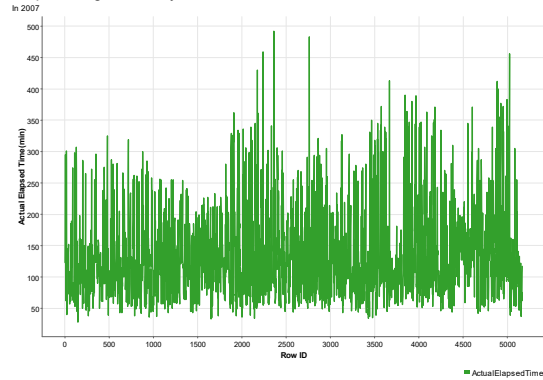
Distribution of Delayed and Non-Delayed Flights



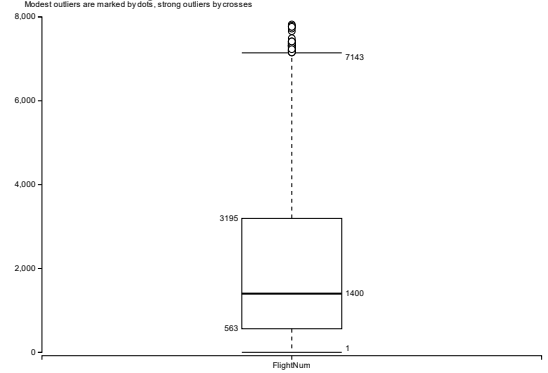
Distribution of Flight Distances



Elapsed Flight Time by Row ID

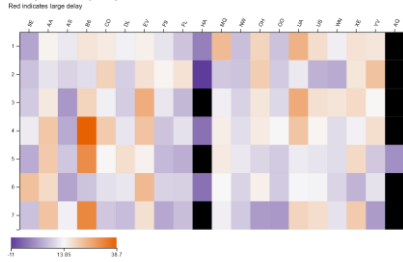


Distribution of Flight Numbers

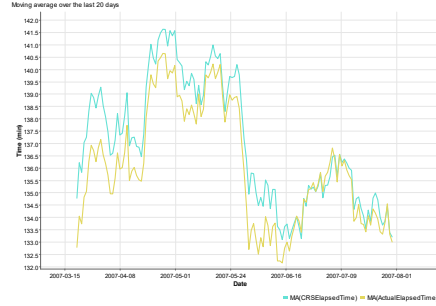


Visualizations Using Three Columns

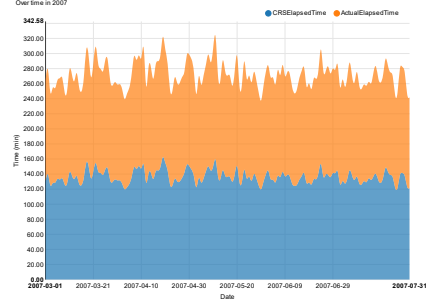
Delay Times by Day of the Week and Carrier



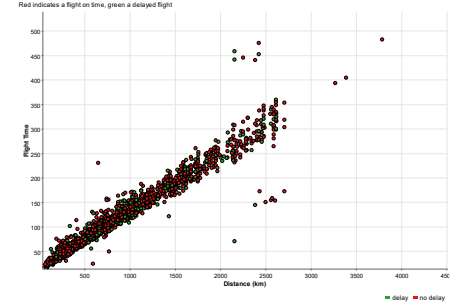
Actual Elapsed Time vs CRS Elapsed Time



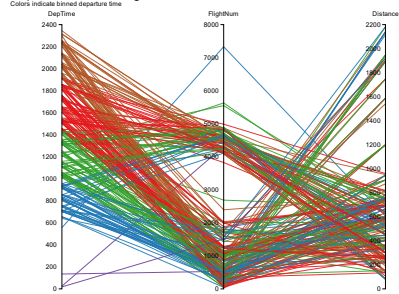
Actual Elapsed Time and CRS Elapsed Time



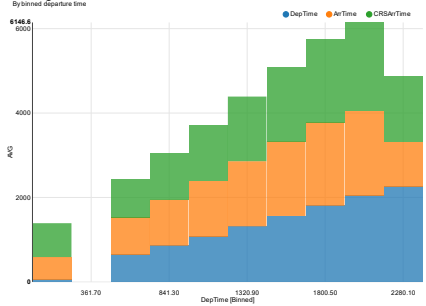
Correlation between Distance and Flight Time



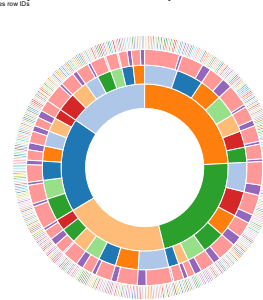
Departure Time vs Flight Number vs Distance



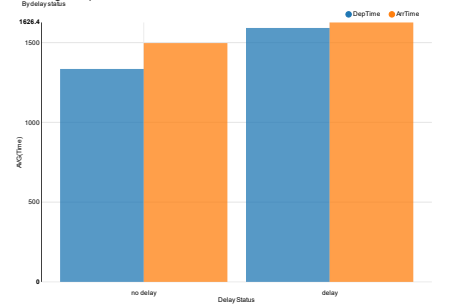
Average Arrival Time and CRS Arrival Time



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

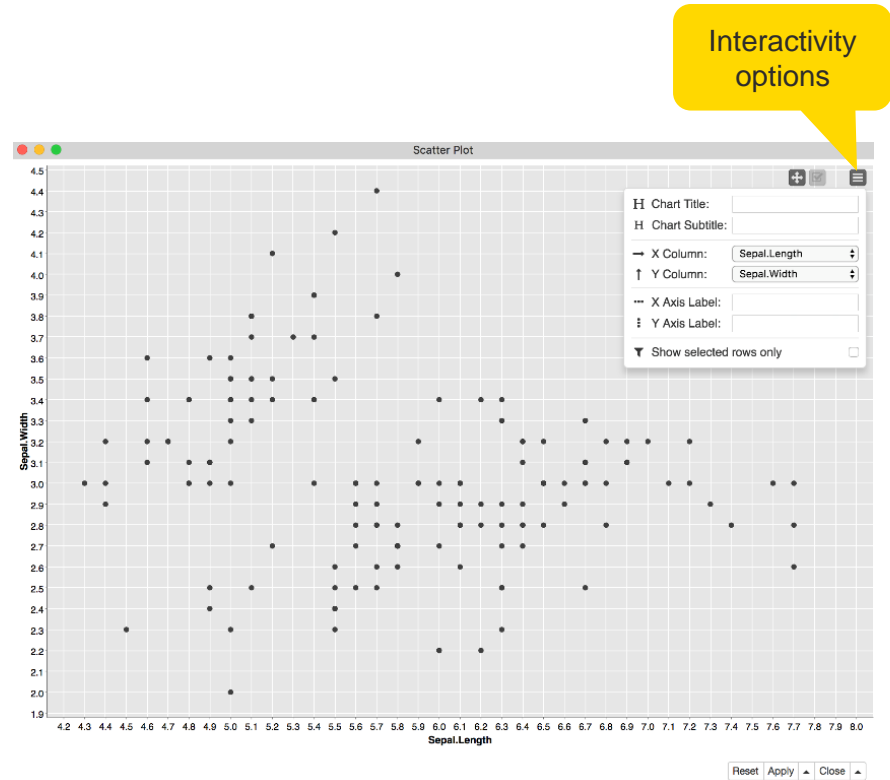
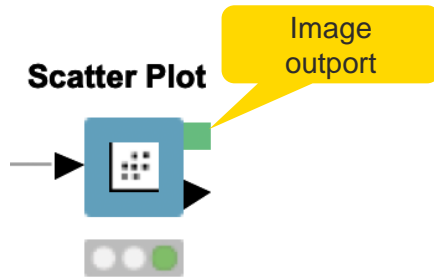


Average Departure and Arrival Times

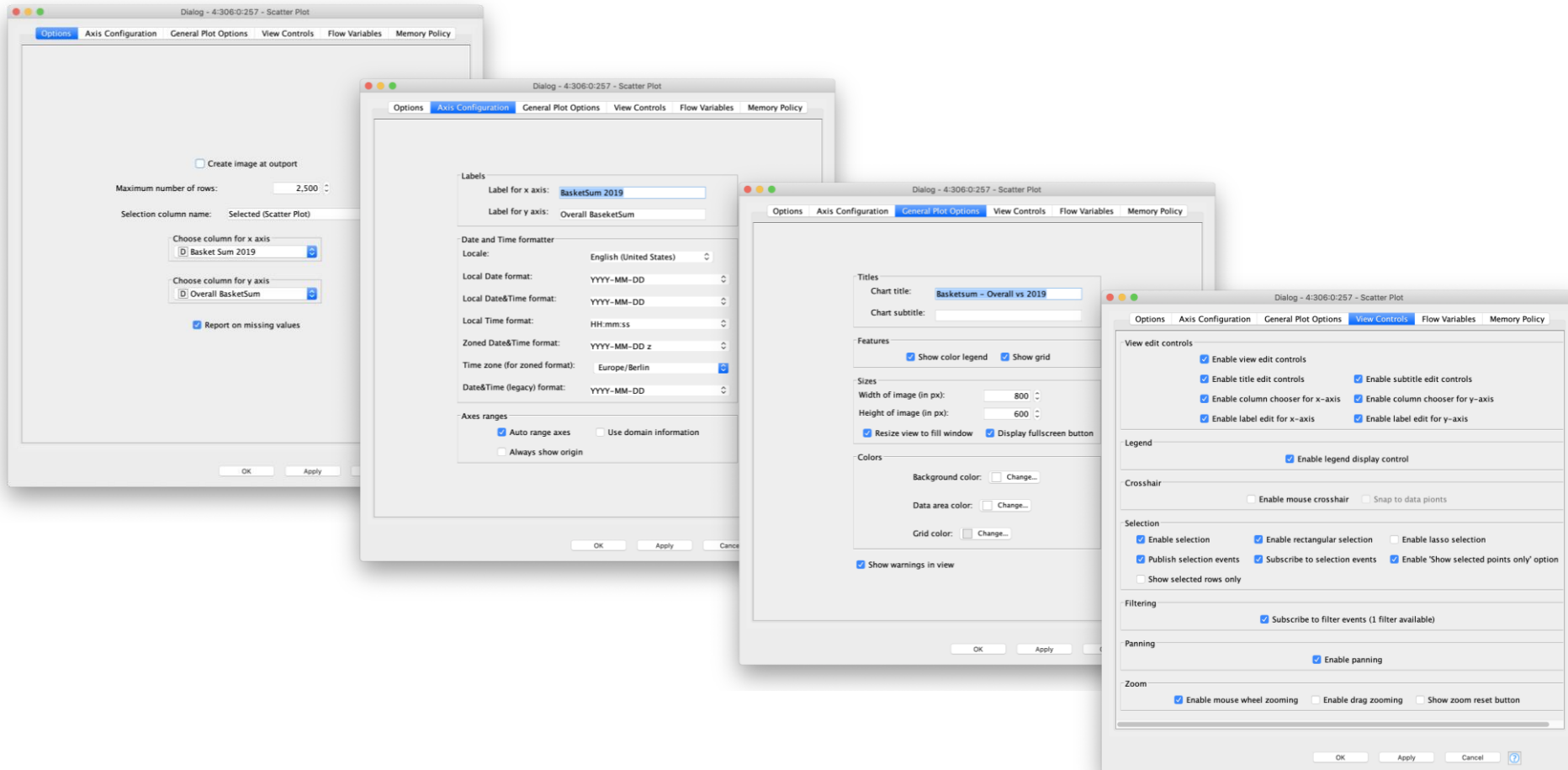


Scatter Plot

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



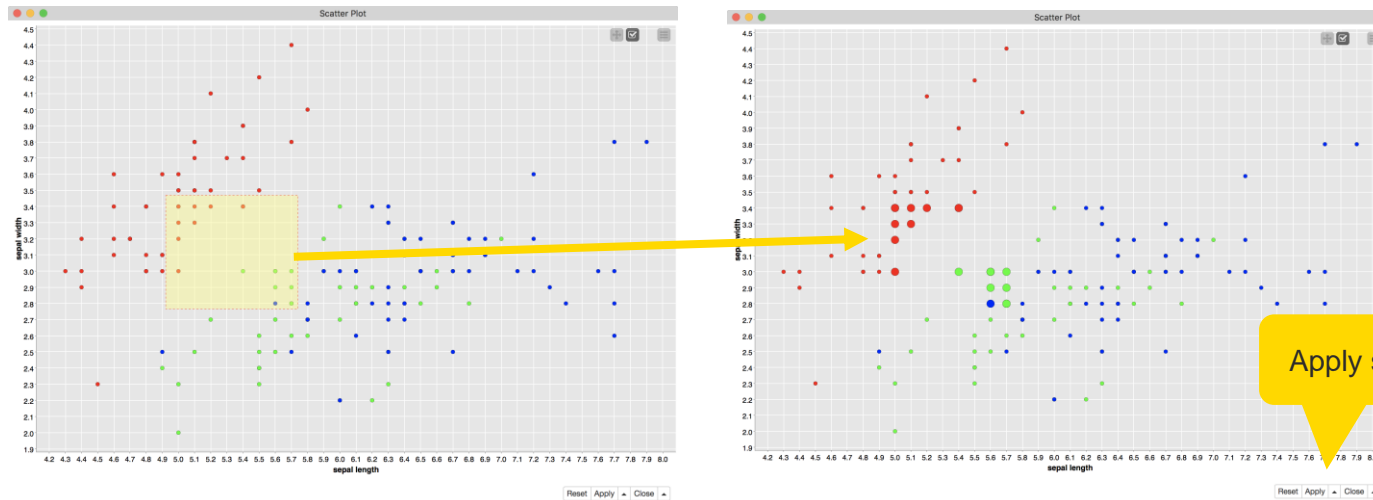
Scatter Plot



Selection and Filtering in JavaScript Views

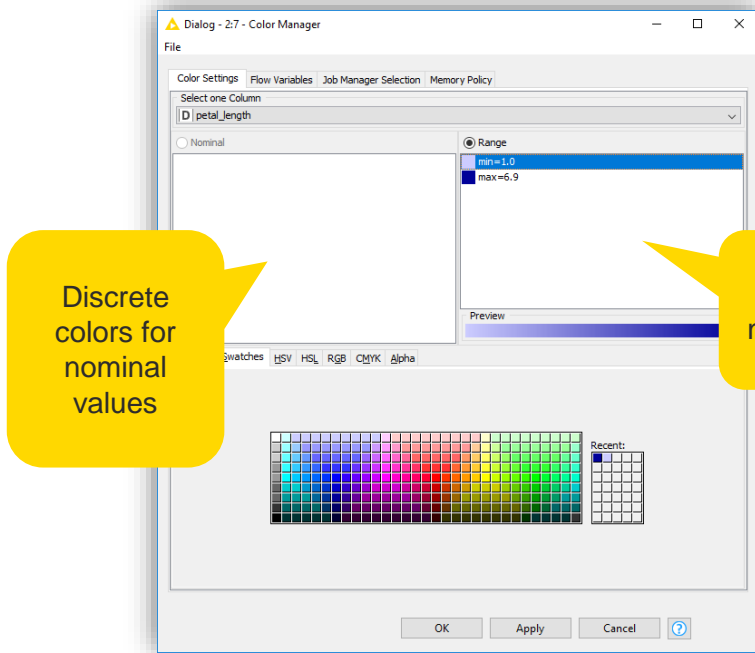
Interactivity allows you to select data points in views

- 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

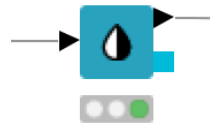


Color Manager

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



Color Manager



Color range for numerical values

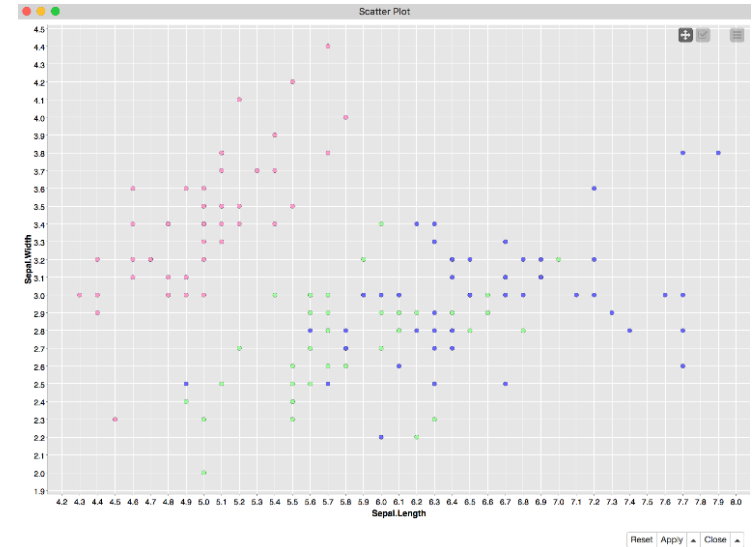
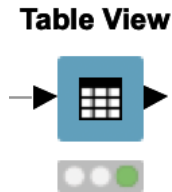


Table View

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



JavaScript Table View

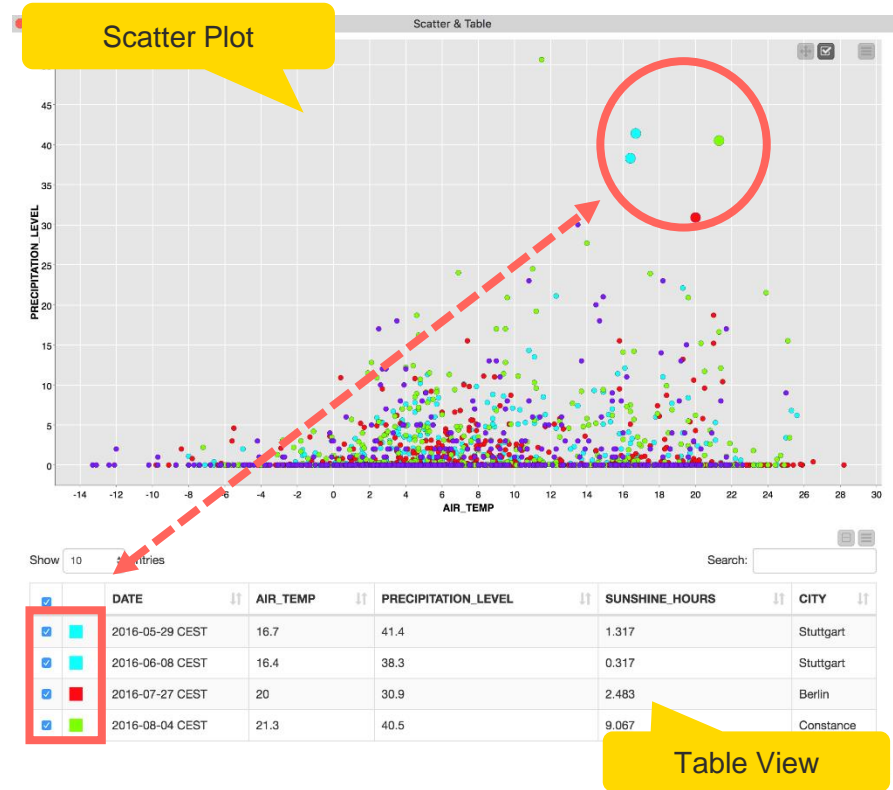
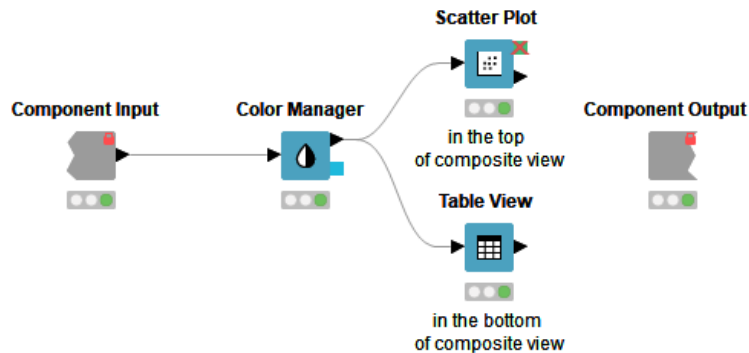
Show entries Search:

<input type="checkbox"/>	RowID	age	workclass	fnlwgt	education	education-num
<input checked="" type="checkbox"/>	Row0	39	State-gov	77516	Bachelors	13
<input type="checkbox"/>	Row1	50	Self-emp-not-inc	83311	Bachelors	13
<input type="checkbox"/>	Row9	42	Private	159449	Bachelors	13
<input type="checkbox"/>	Row12	23	Private	122272	Bachelors	13
<input type="checkbox"/>	Row25	56	Local-gov	216851	Bachelors	13
<input type="checkbox"/>	Row32	45	Private	386940	Bachelors	13
<input type="checkbox"/>	Row41	53	Self-emp-not-inc	88506	Bachelors	13
<input type="checkbox"/>	Row42	24	Private	172987	Bachelors	13
<input type="checkbox"/>	Row45	57	Federal-gov	337895	Bachelors	13
<input type="checkbox"/>	Row53	50	Federal-gov	251585	Bachelors	13
		<input type="text" value="Search age"/>	<input type="text" value="Search workclass"/>	<input type="text" value="Search fnlwgt"/>	<input type="text" value="Bachel"/>	<input type="text" value="Search education-"/>

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

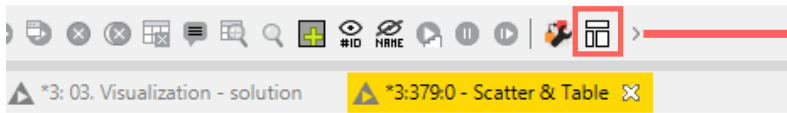
Components – Combined Views

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

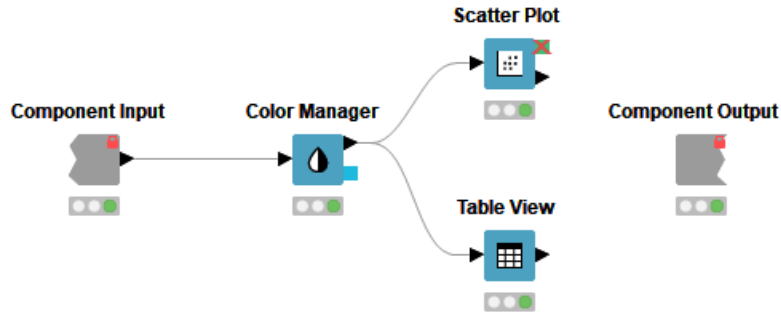
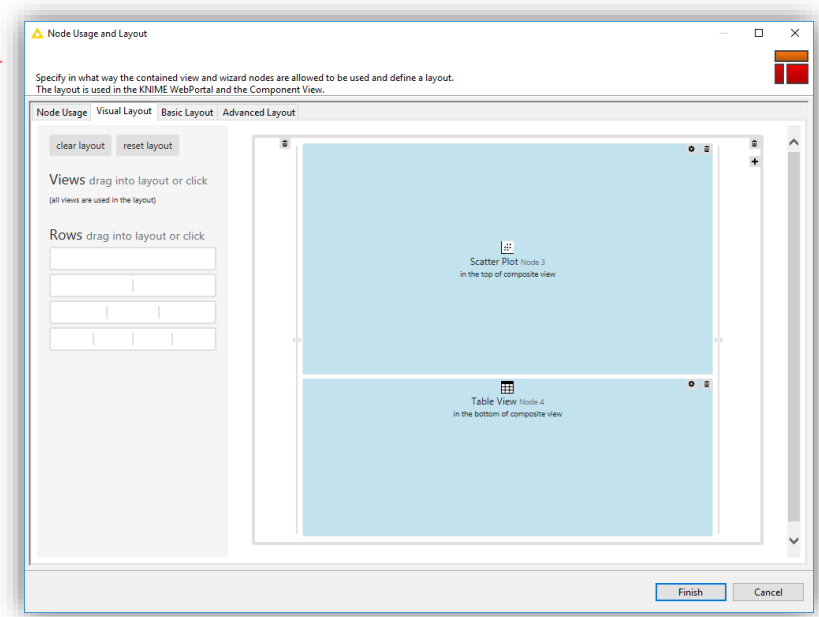


Configure Content and Views Layout

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

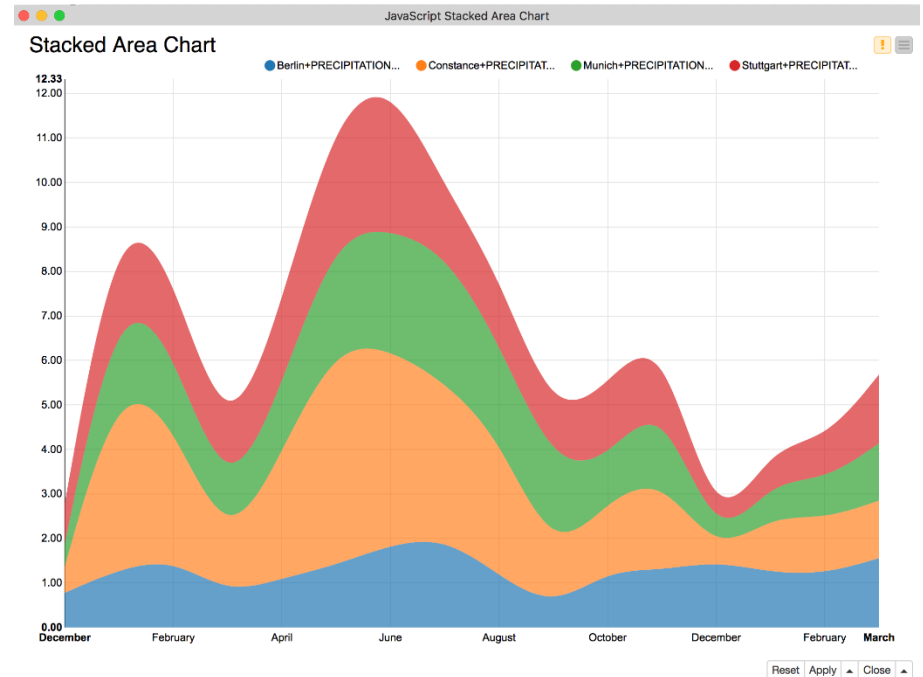
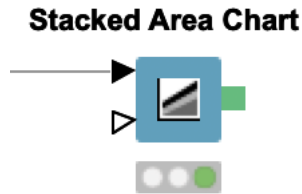


- Add views and rows via drag&drop
- Add columns using + buttons



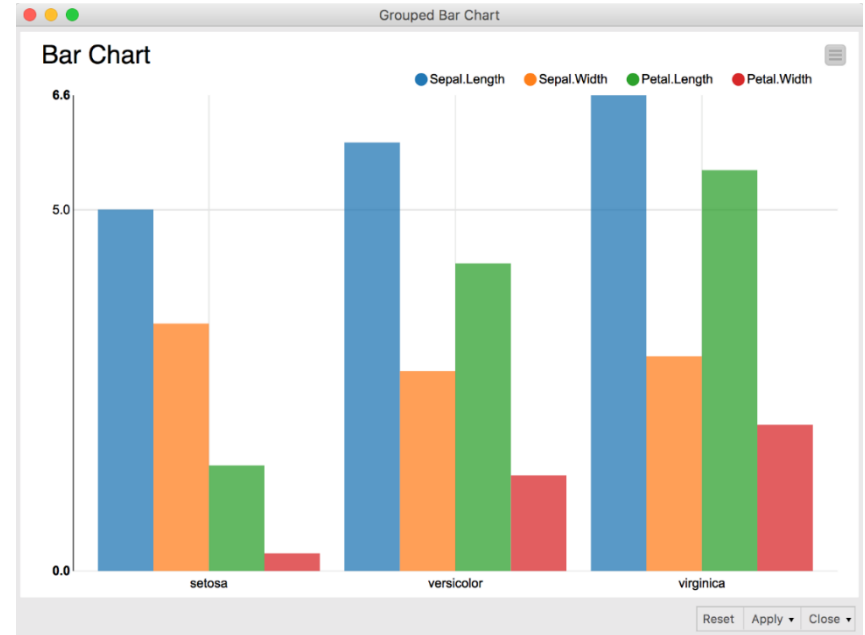
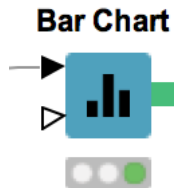
Stacked Area Chart

- 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

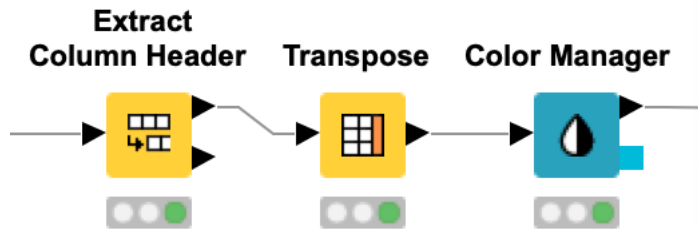


Table with Colors - 3:306:0:265 - Color M...

File Hilite Navigation View

Table "default" - Rows: 5

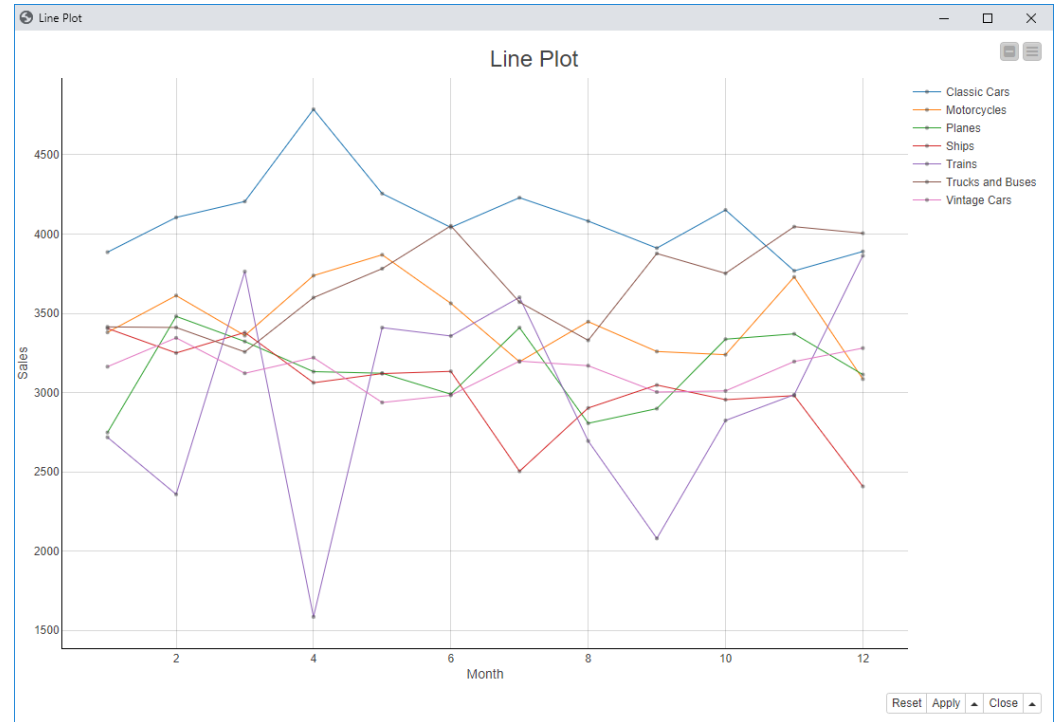
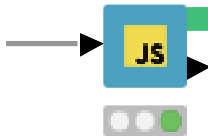
Row ID	Column Header
Column 0	Year
Column 1	Quarter
Column 2	Store - no CC
Column 3	Store - with CC
Column 4	OnlineStore



Line Plot

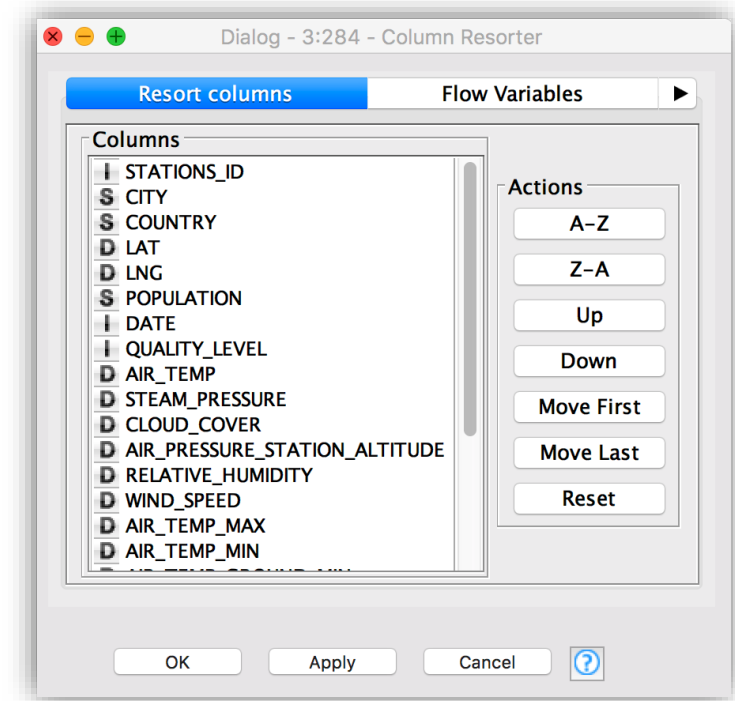
- Plots sequence of values, e.g. over time
- Useful to identify trends, also between groups

Line Plot (Plotly)

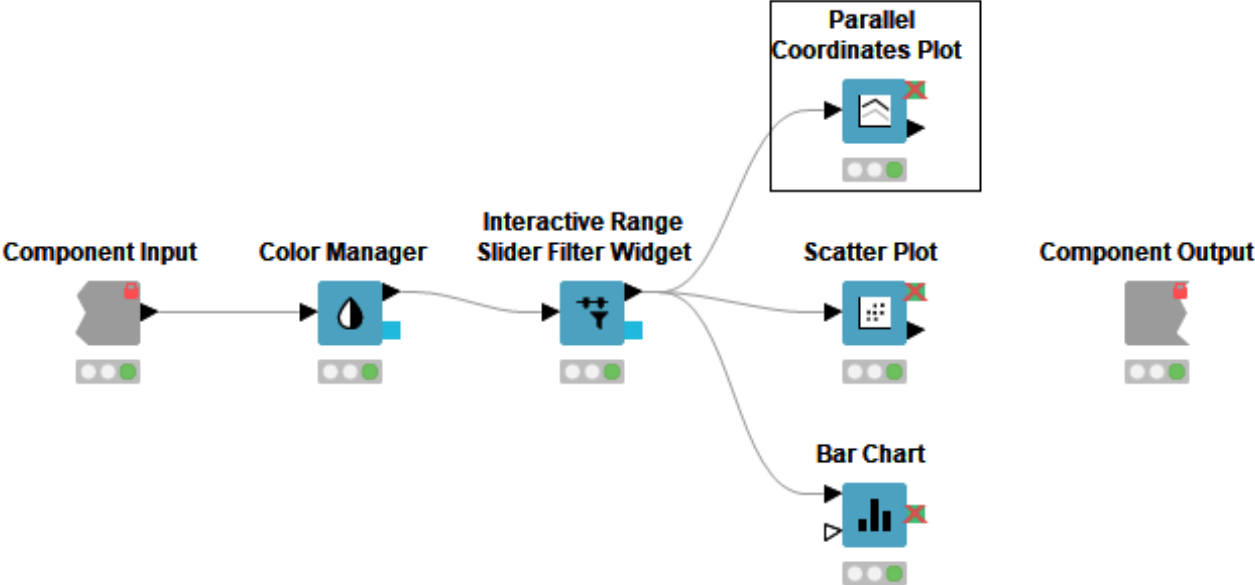


Column Resorter

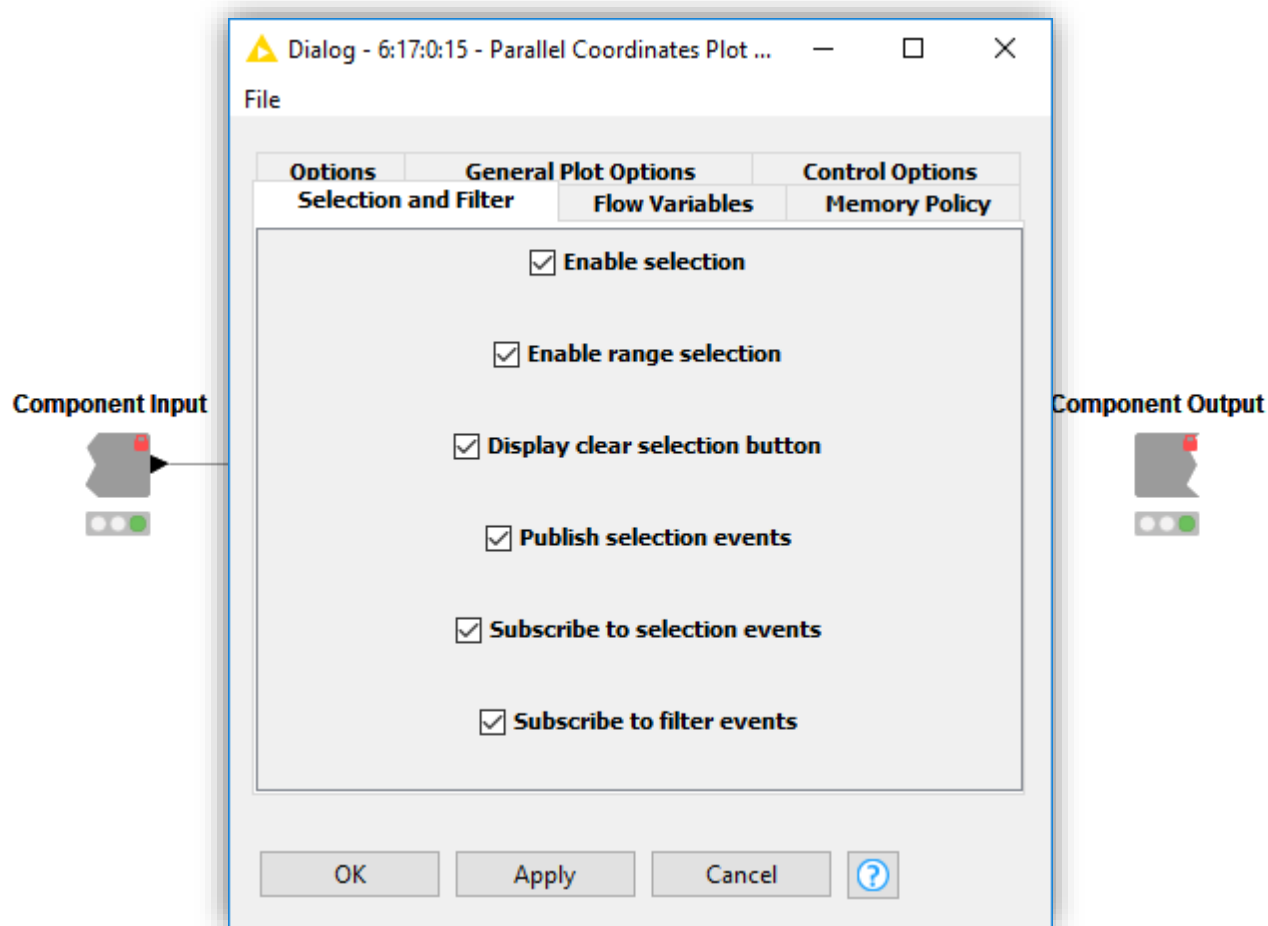
- Changes the order of the input column based on user defined settings
- Options:
 - Sort alphabetical (A-Z or Z-A)
 - Move the selected columns one step (Up or Down)
 - Move the selected columns to top or end (Move First / Last)



Interactivity across Charts: Selection and Filter Events



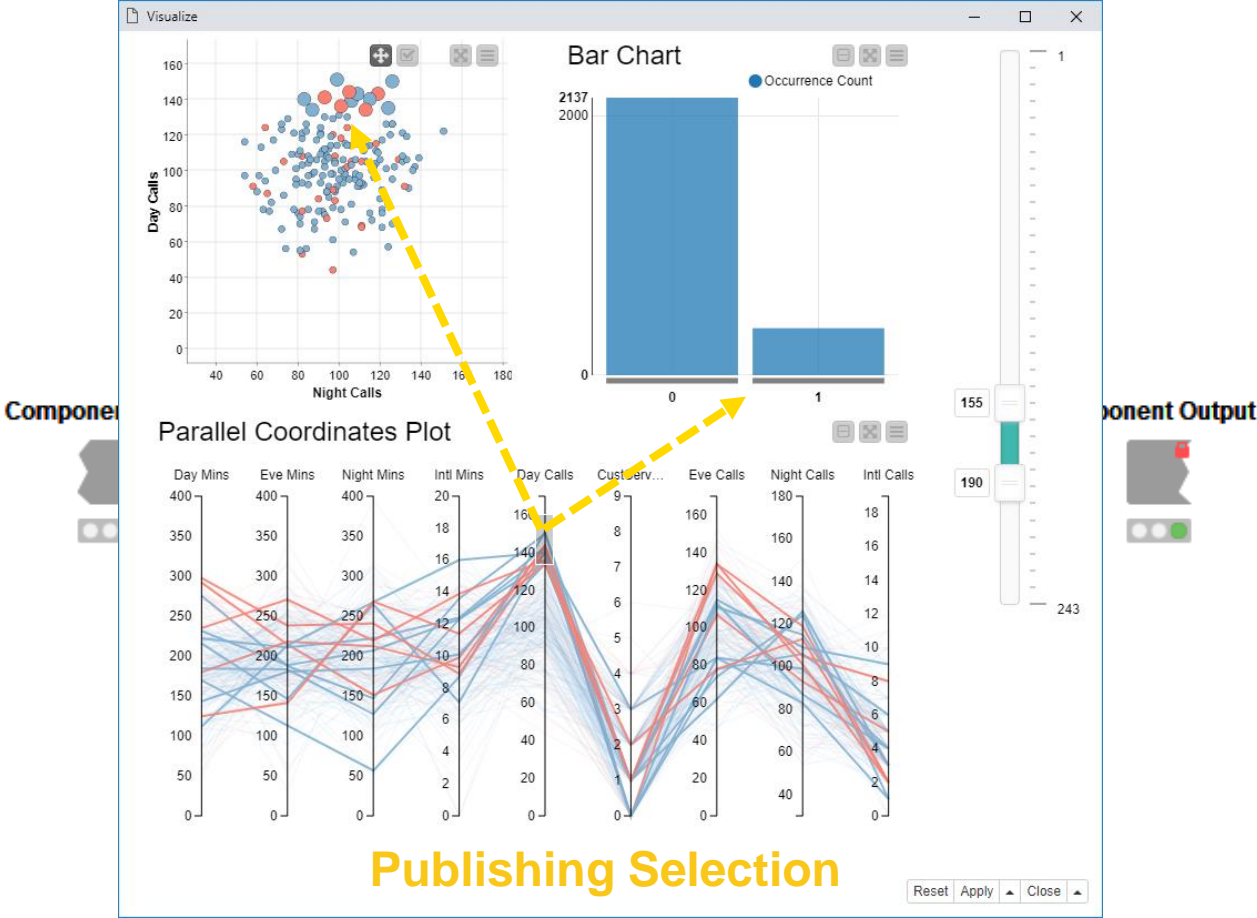
Interactivity across Charts: Selection and Filter Events



Interactivity across Charts: Selection and Filter Events



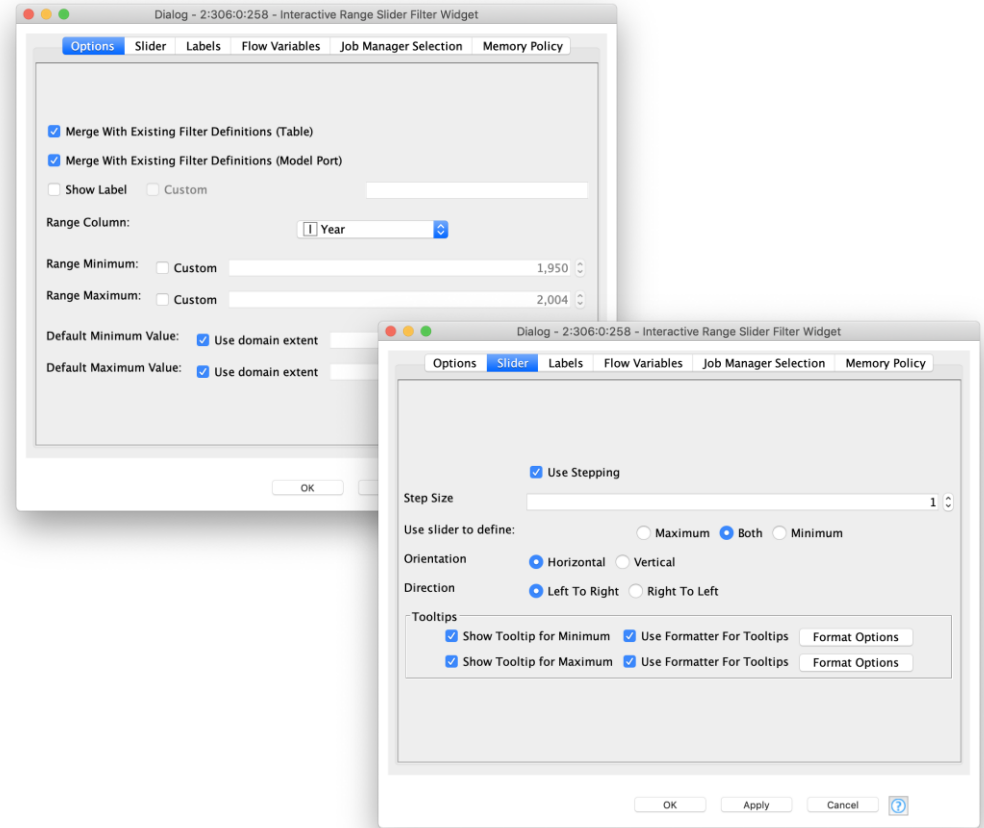
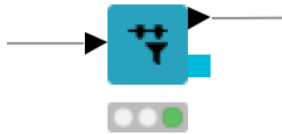
Interactivity across Charts: Selection and Filter Events



Interactive Range Slider Filter Widget

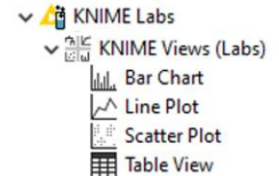
- Slider which can be used to trigger interactive filter events in the view of a component

Interactive Range Slider Filter Widget



New Visualization Nodes in KNIME (Labs)

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



Dialog 4288 - Bar Chart

Total Sales by Product as of August 3, 2022

The view updates as you change the configuration settings >>

Settings controlled by variables

Data

Category dimension
Products

Aggregate
 Occurrence count Sum Average

Frequency dimensions

Excludes
CustomerKey
BasketSize

Includes
BasketValue

Plot

Title
Total Sales by Product as of August 3, 2022

Orientation
 Vertical Horizontal

Arrange bars
 Grouped Stacked

Display legend

Interactivity
 Enable image download
 Show tooltip

Product	Sales
Fund Manager+	~1,100,000
P+B Investment	~800,000
Private Investment	~2,200,000
Gold Investment	~1,500,000
CO Investment	~100,000

BasketValue
Fund Manager+ 1,123,031

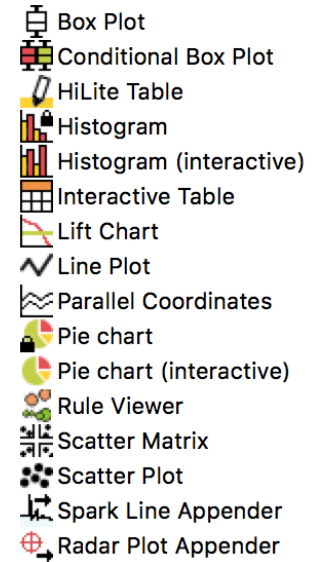
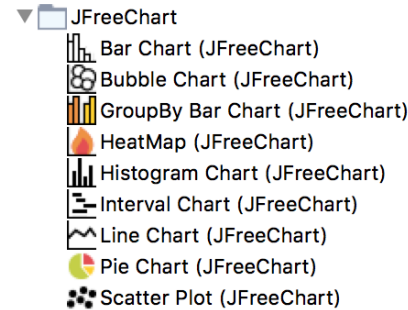
Cancel

Cancel

Ok

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)



Confirmation of Attendance and Survey

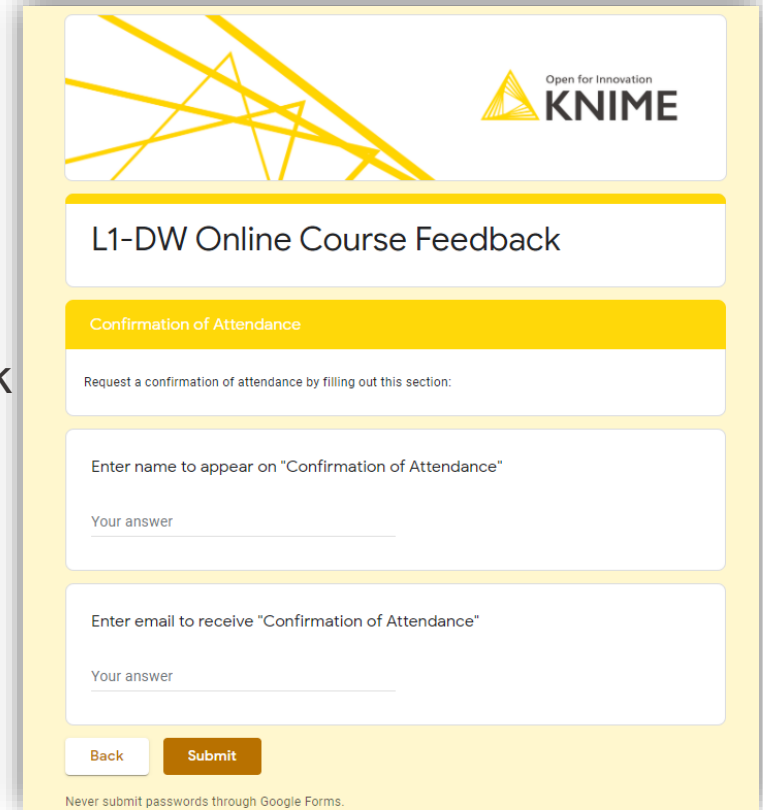
- If you would like to get a “Confirmation of Attendance” please click on the link below*

[Confirmation of Attendance and Survey](#)

- The link also takes you to our course feedback survey. Filling it in is optional but highly appreciated!

Thank you!

*Please send your request within the next 3 days



Open for Innovation
KNIME

L1-DW Online Course Feedback

Confirmation of Attendance

Request a confirmation of attendance by filling out this section:

Enter name to appear on "Confirmation of Attendance"

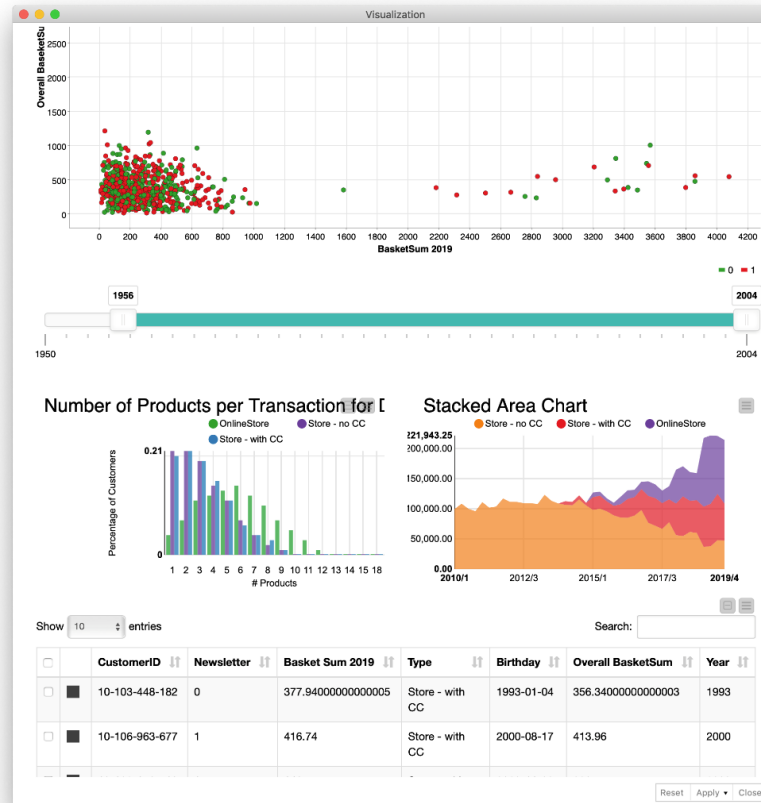
Your answer

Enter email to receive "Confirmation of Attendance"

Your answer

Never submit passwords through Google Forms.

Exercise: 07_Visualization – Goal



Exercise: 07_Visualization

- Create a scatter plot to show the relationship between the total purchase amount in 2019 and overall
- 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 and design a composite view

- Optional tasks:
 - Add a range slider to filter the scatter plot by age
 - Build a bar chart to show the number of products per order for the different transaction types
 - Customize the colors in the stacked area chart

Data Wrangler Cheat Sheet

Cheat Sheet: Data Wrangling with KNIME Analytics Platform



ACCESS DATA

CSV 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 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.

JSON 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.

Parquet 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.

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

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

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

SQL Executes a SQL query by accessing the database metadata or via a custom SQL query.

COMBINE DATA

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

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

Join 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.

FILTER DATA

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.

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

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

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.

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 conjunction is also supported. No SQL coding required. The input SQL query is represented by the place holder #filter#.

WRITE DATA

Excel 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 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.

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

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

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

Database Writes the resulting rows from the input SQL query into a new table inside the database.

DATABASES

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

RESHAPE AND AGGREGATE DATA

Reshape 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 uses.

Reshape Extends the aggregation functionality of the Reshape 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 preceding columns become columns.

Reshape Maps the original values in the selected columns to integer values in the model output part. The Category to Number (Apply) and Number to Category (Apply) nodes apply the mapping rule in both directions.

Reshape Creates one new column for each value in the selected columns. 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.

Reshape Converts the rows to columns and the columns to rows.

Reshape Performs several transformations on rows, such as reordering, filtering, re-ordering and type changing, on the input columns. By adding dynamic ports it can replace a concatenate node.

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

Split 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.

Split Backs the cells of the selected value column into rows. The cells of the selected retained columns are appended to the corresponding output rows.

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

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

DATA TYPES & CONVERSIONS

String: Sequence of characters, e.g. "This is a string"

Integer: Whole real-valued number, e.g. 100 or 345

Double: Real-valued number, e.g. 0.42 or 45.39

Date/Time: A data format for date, time, date/time, or date/time plus time zone

Boolean: Two possible values only, e.g. TRUE and FALSE

Convert Converts the data type of the selected columns from a number (integer, e.g. integer or double, to string.

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

COLLECTOR 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 where each value occurs only once.

DOCUMENTING KNIME

Analytics Platform supports many ways to document your models, reports, integrations, etc.

CREATE COLUMNS

Math 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.

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

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

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

String 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.

DYNAMIC PORT

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

FORMAT EXCEL SHEETS

The Conditional 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.

Format 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 default node to select formatting instructions for all cells with a specified tag in the XLS Control Table at the top input port.

Format 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.

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

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

Format 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.

Format Applies all formatting instructions to an existing Excel sheet.

DATETIME

Parse Parses the strings in the selected column according to a date-time format and converts them into Date/Time cells. Four Date/Time zones are supported: only date, only time, date/time, and date/time plus time zone.

Parse Filters 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.

Parse Calculates the difference between two datetime 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 of a selected column.

Parse Extracts selected time and date fields from a selected column of type datetime and appends their values in new columns.

CLEAN DATA

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

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

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

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

© 2023 KNIME AG. All rights reserved.

155

Stay Connected with KNIME



Blog:
knime.com/blog



KNIME Self-Paced Courses:
knime.com/knime-self-paced-courses



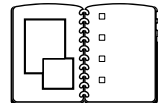
Forum:
forum.knime.com



Email:
education@knime.com



KNIME Hub:
hub.knime.com



Medium Journal:
medium.com/low-code-for-advanced-data-science

**Follow us on
social media:**



Thank you!

education@knime.com

April 17, 2023



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

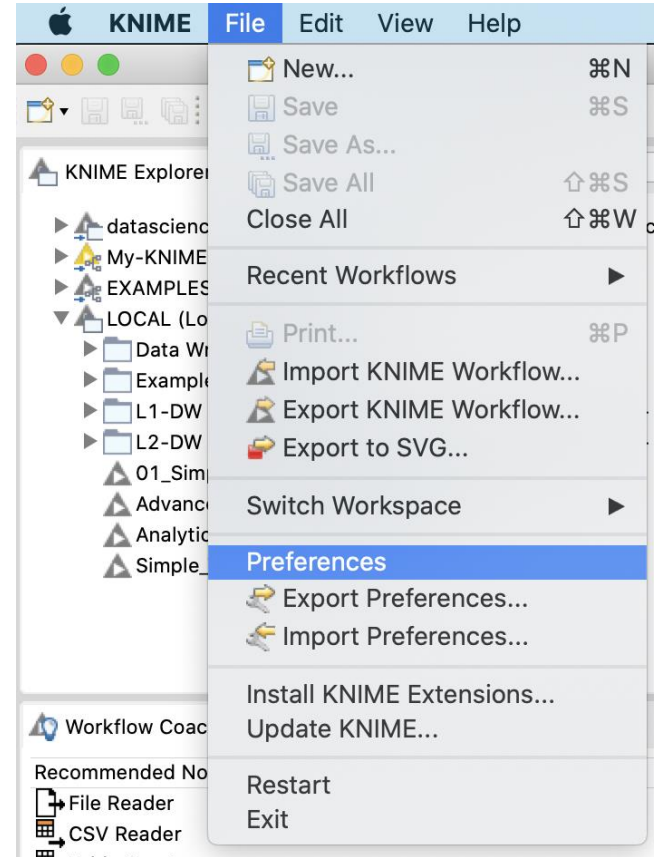


Adding a Local Update Site

- 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

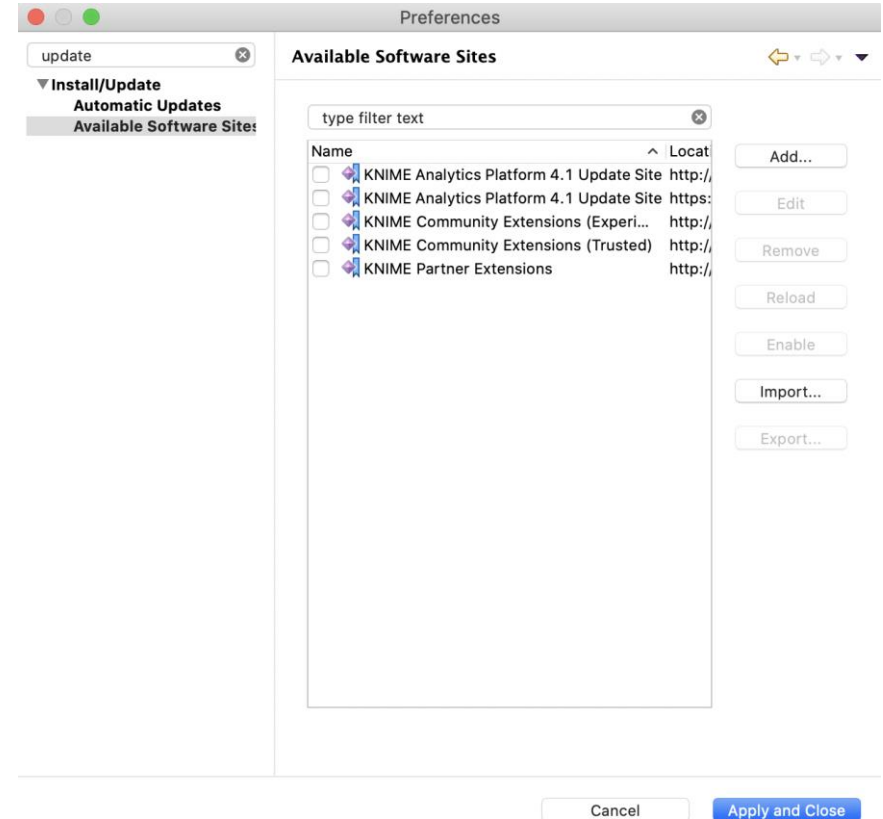
Adding a Local Update Site

- 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.



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

