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

KNIME AG

April 17, 2023

Structure of the course

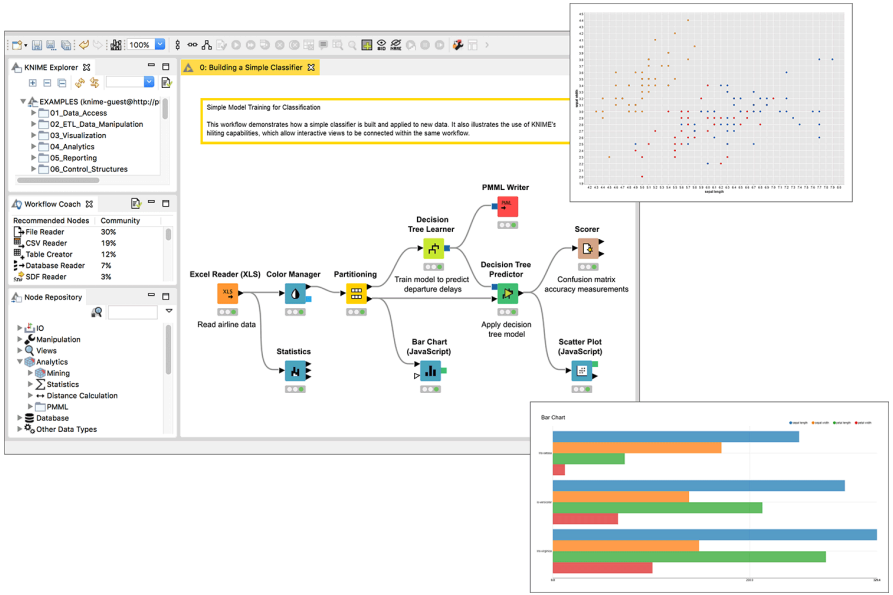
- This course covers the following topics
  - Introduction to KNIME Analytics Platform
  - Importing Data
  - Data Merging
  - Data Cleaning &Transformation
  - Data Aggregation
  - Data Visualization
  - Q&A and Summary
- Structure of the course
  - Introduction to the topic and workflow demo
  - Hands-on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

Overview
KNIME Analytics Platform

What is KNIME Analytics Platform?

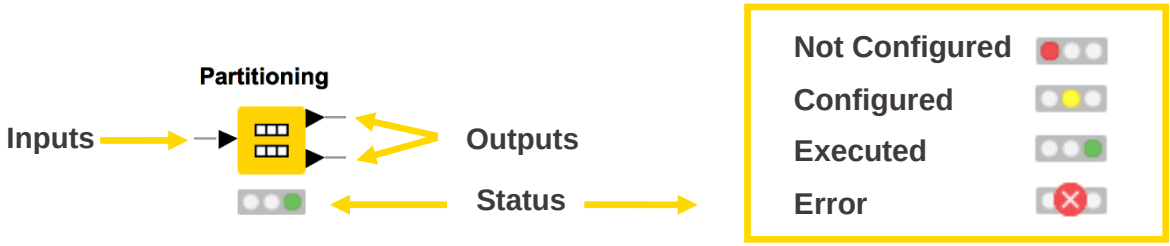
- **Open source**
- Visual programming paradigm
- Data analysis, manipulation, visualization, and reporting

- Multiple extensions:
  - Text Mining
  - Network Mining
  - Cheminformatics
  - Geospatial Analytics
  - Many integrations,
    such as Java, R, Python,
    Weka, Keras, Plotly, H2O, etc.

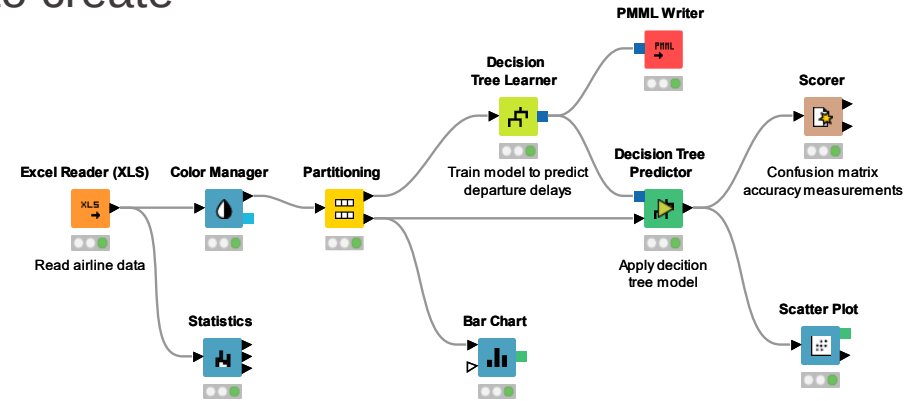


Visual KNIME Workflows

NODES perform tasks on data

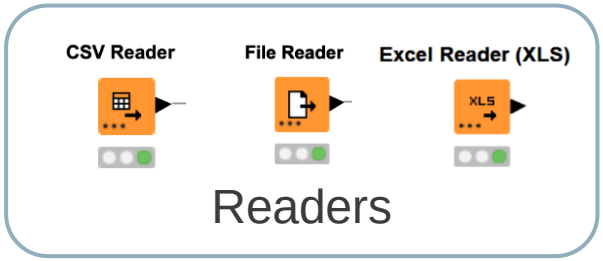


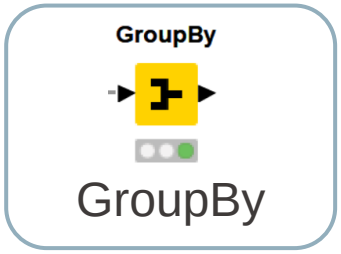
Nodes are combined to create **WORKFLOWS**

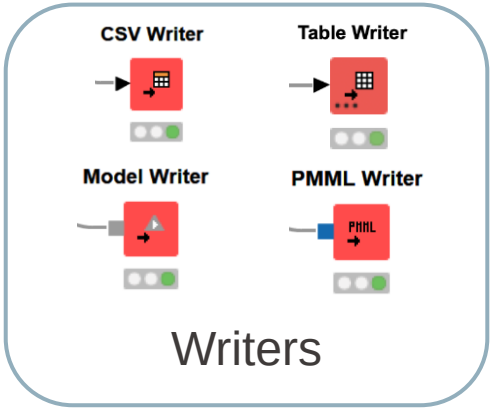


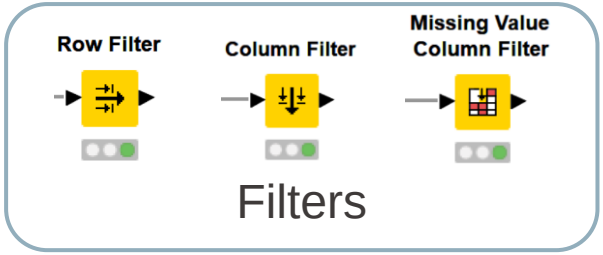
Open for Innovation KNIME

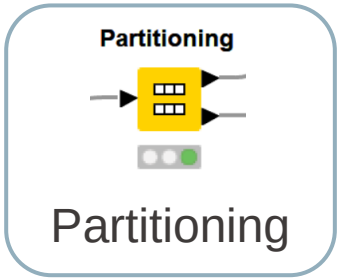
Frequently Used Nodes

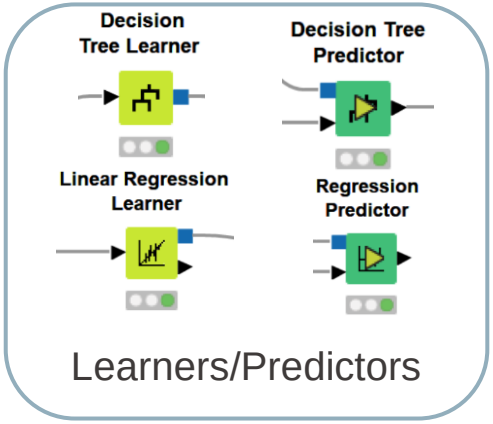




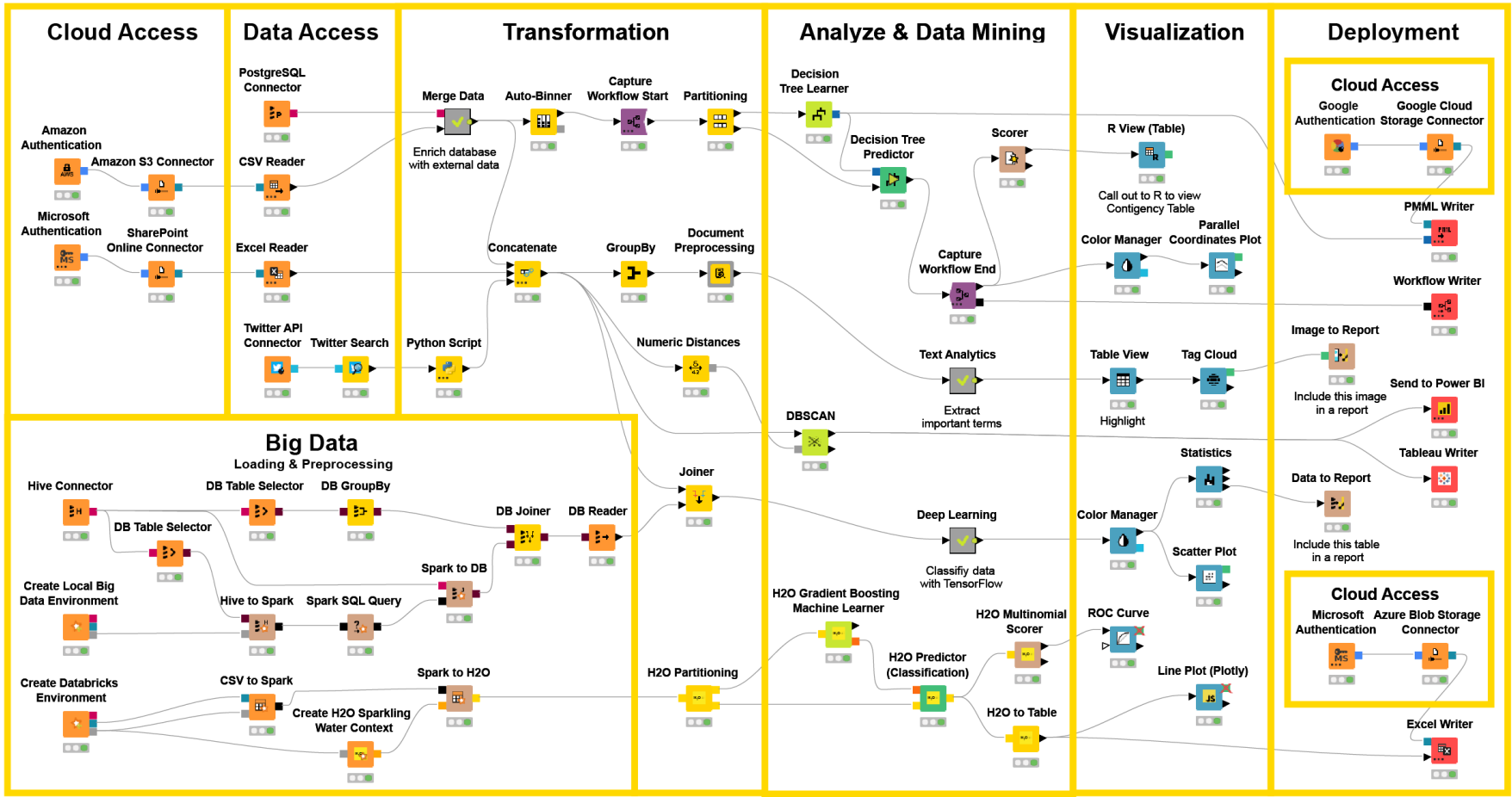




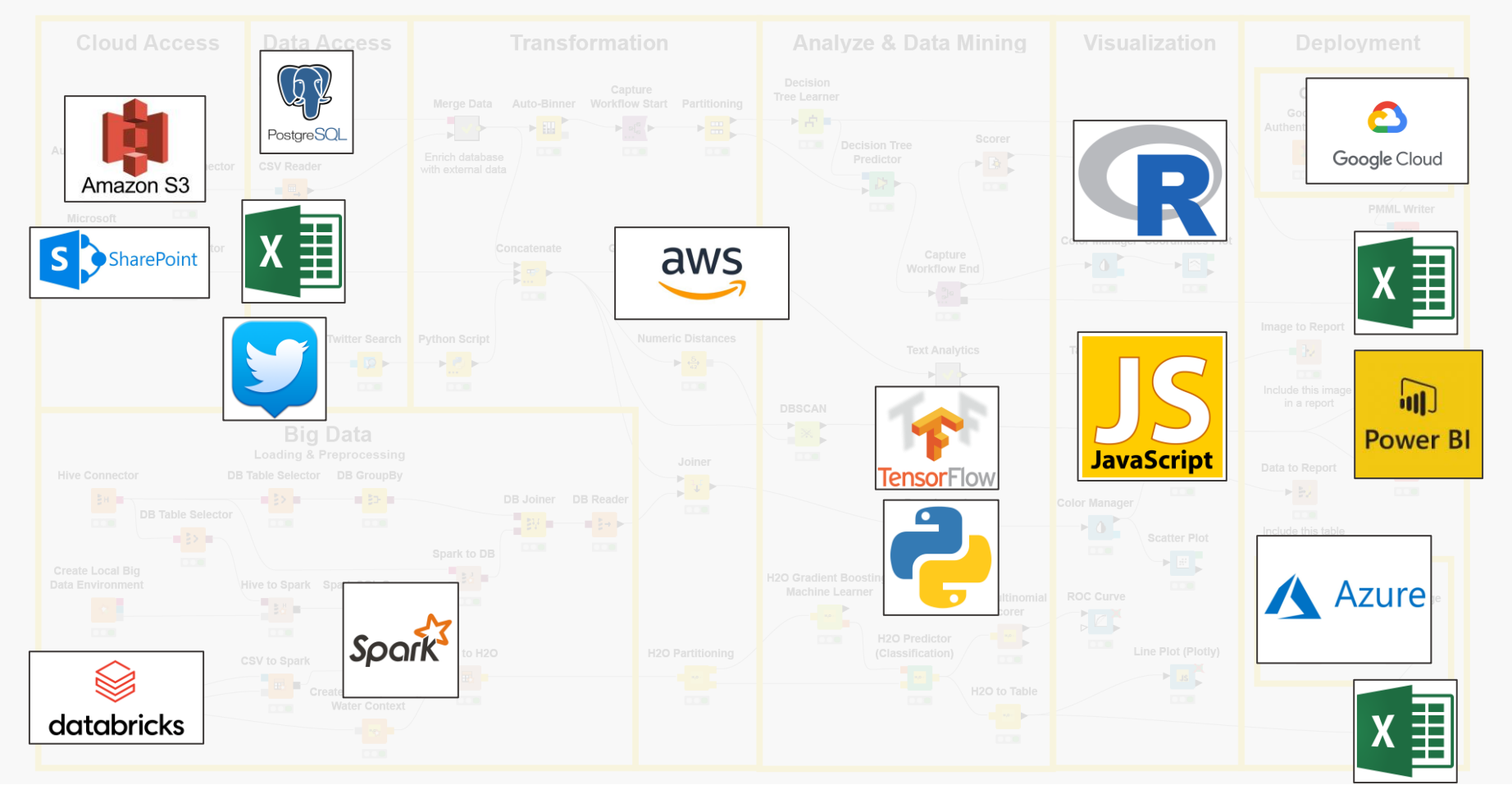




Covering all Stages of the Data Science Life Cycle



Technologies & Languages under the Hood



Open for Innovation KNIME

Let’s Start KNIME & Build a Workflow

Installation

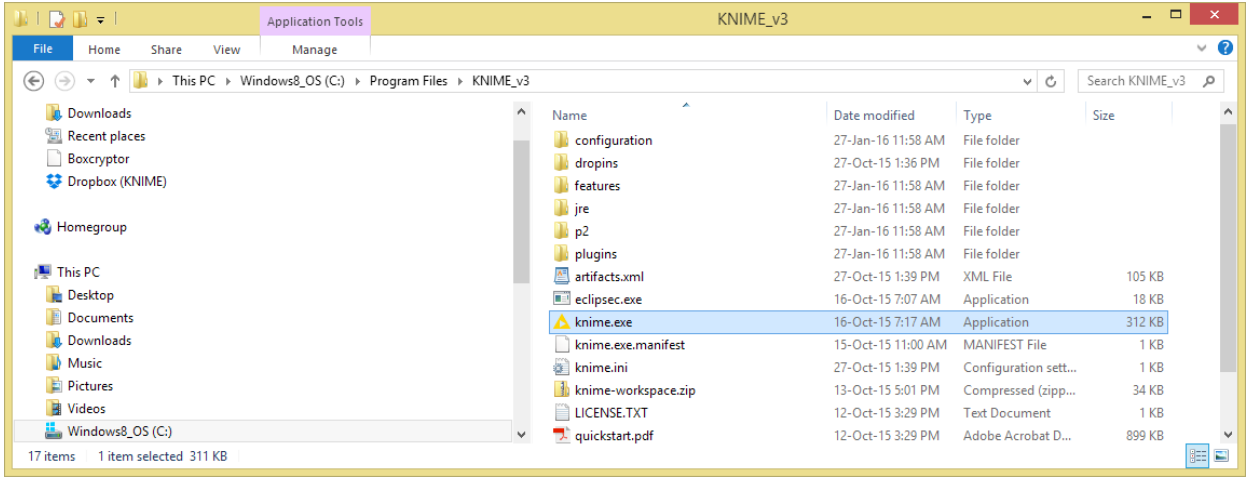
https://www.knime.com/downloads

- 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

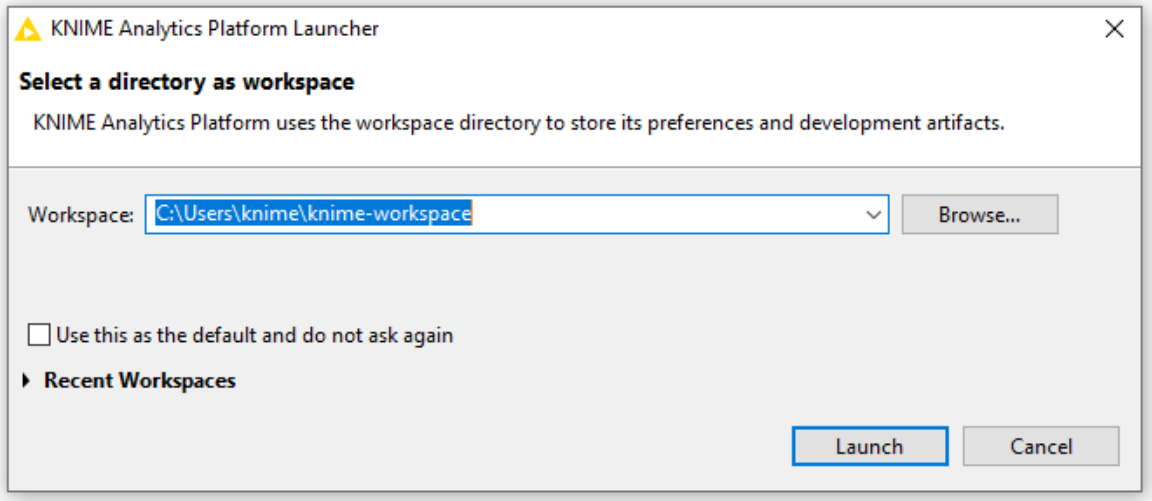
- Use the shortcut created by the installer

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

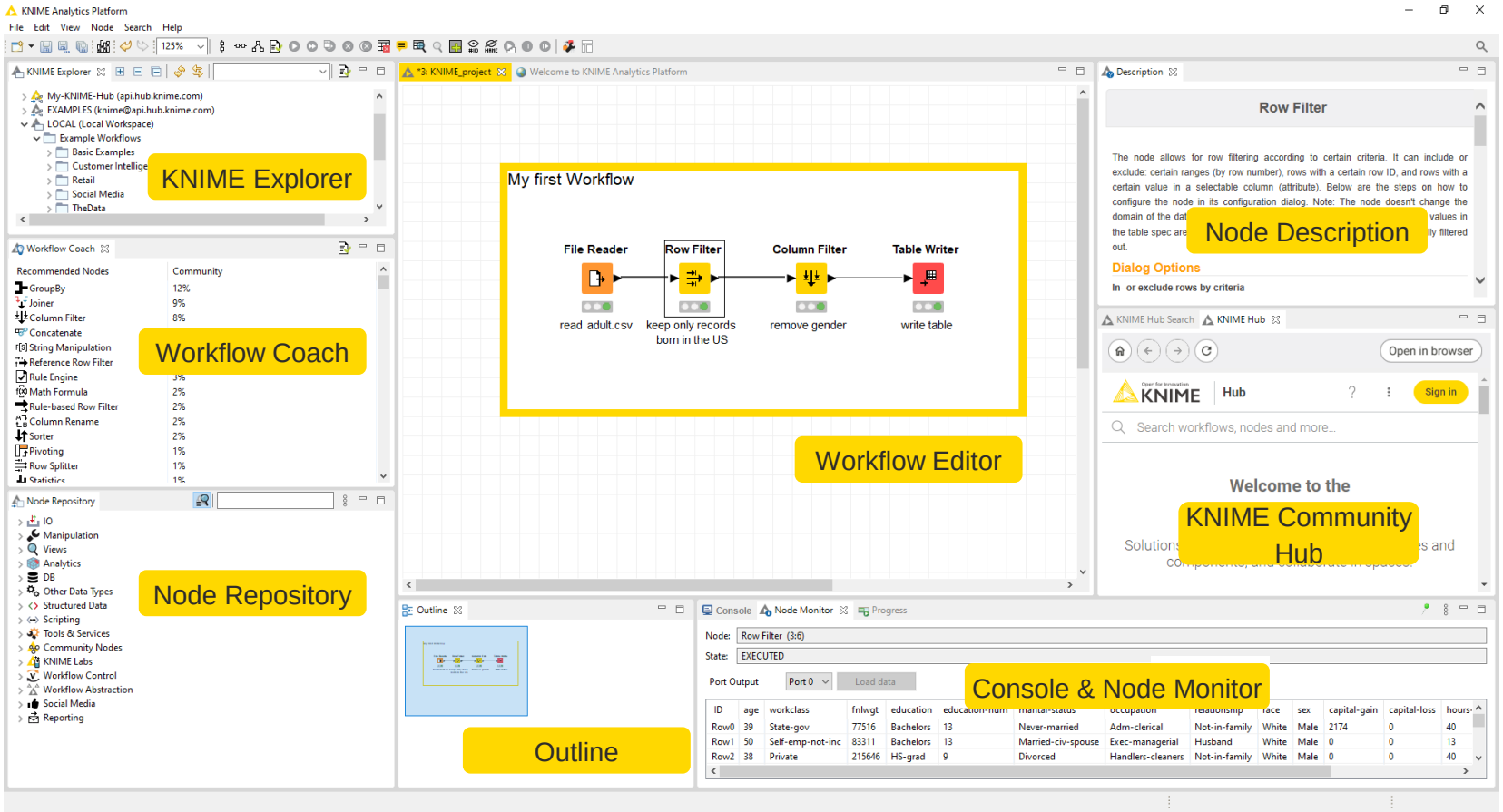


The KNIME Workspace

- 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)

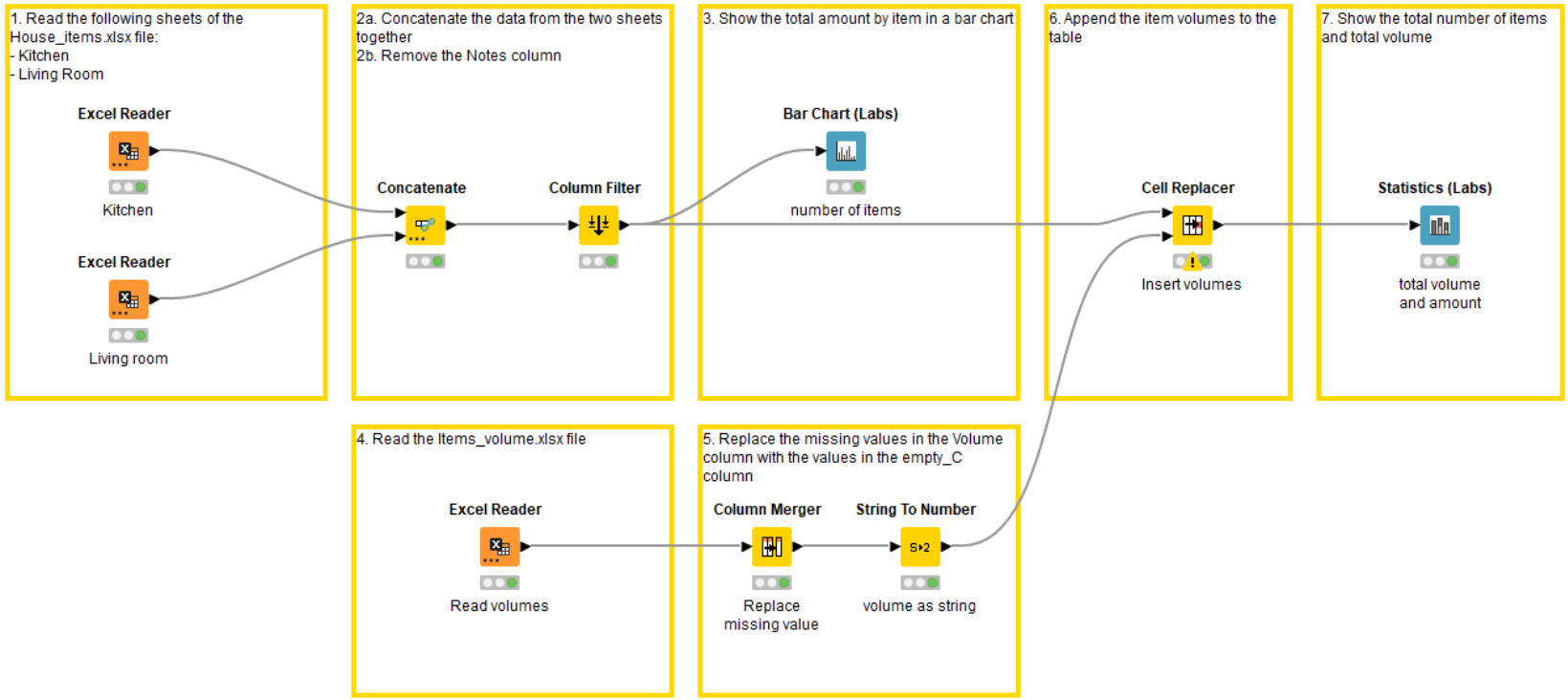


The KNIME Analytics Platform Workbench



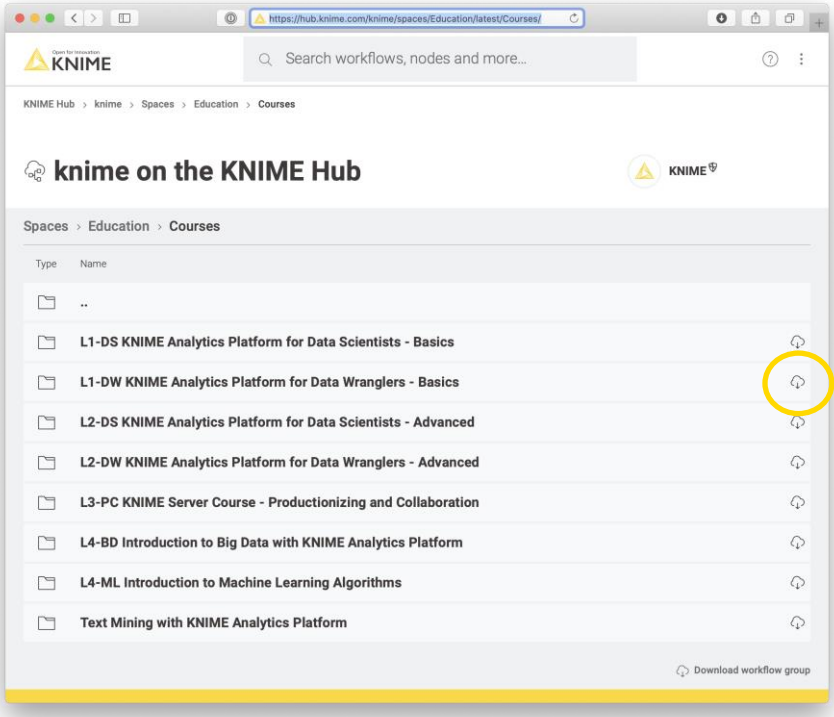
Move to new apartment -workflow

- Let’s complete this together!



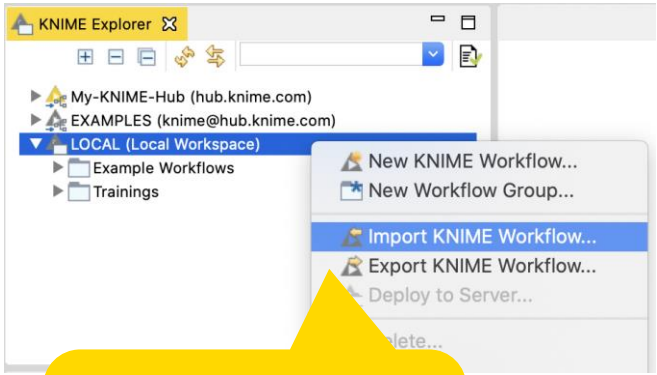
Downloading Exercises

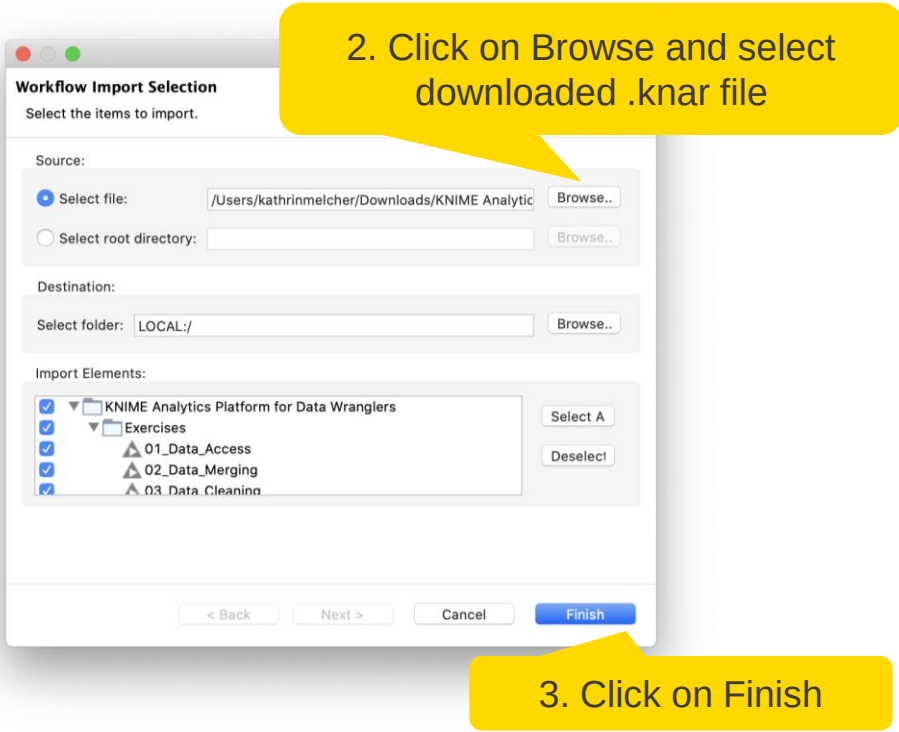
- Download the course material from the KNIME Community Hub https://hub.knime.com/knime/spaces/Education/latest/Courses/



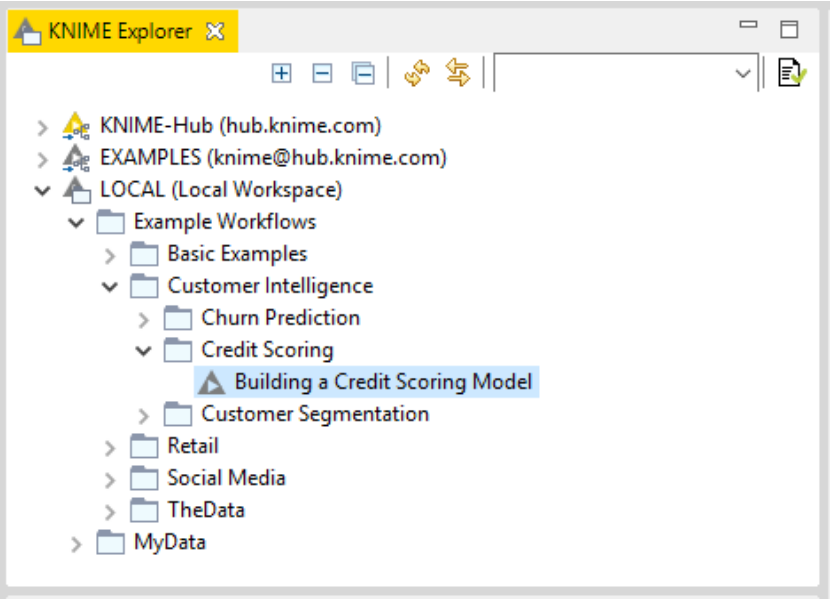
Importing Exercises

- Import the course material to KNIME Analytics Platform





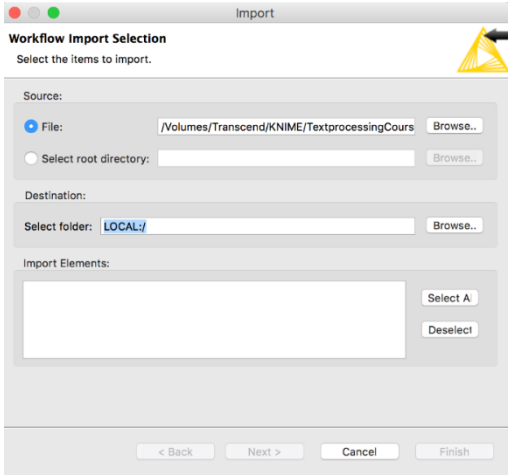
KNIME Explorer

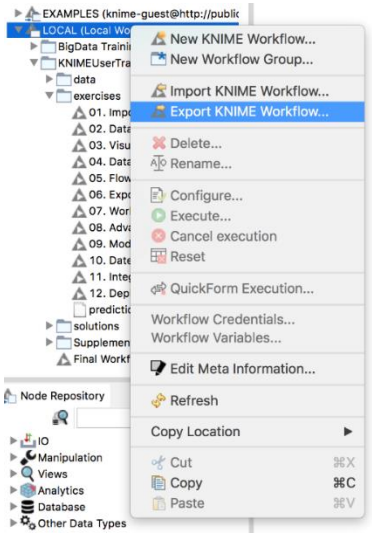


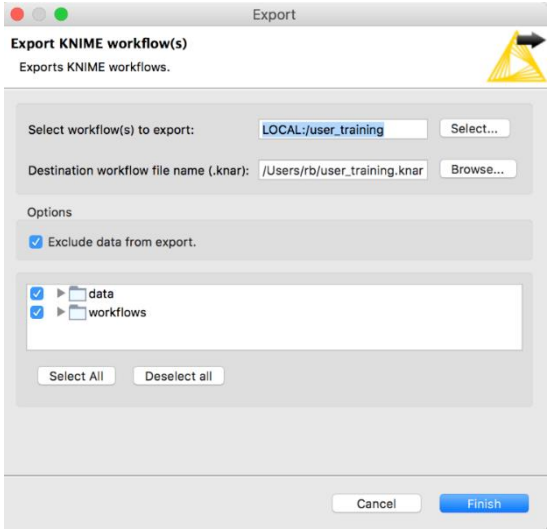
- In LOCAL you can access your own workflow projects.
- Other mountpoints allow you to connect to
  - Examples
  - KNIME Community Hub
  - KNIME Business Hub
- The Explorer toolbar on the top has a search box and buttons to
  - select the workflow displayed in the active editor
  - refresh the view
-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

Creating New Workflows, Importing, and Exporting

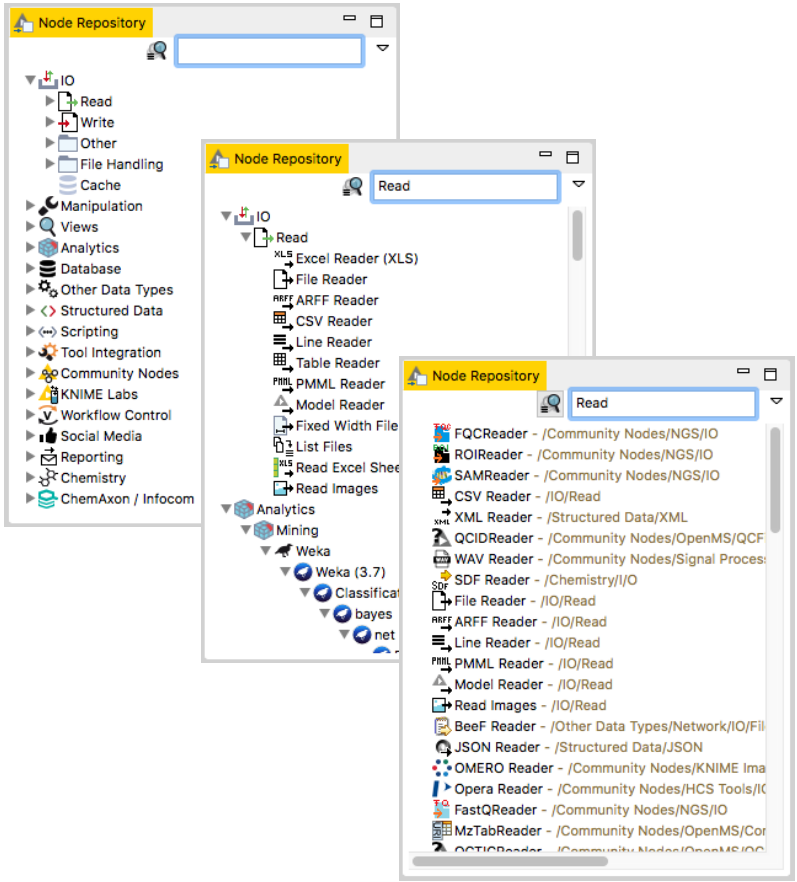
- Right-click inside the KNIME Explorer to create a new workflow or a workflow group, or to import a workflow
- Right-click the workflow or workflow group to export





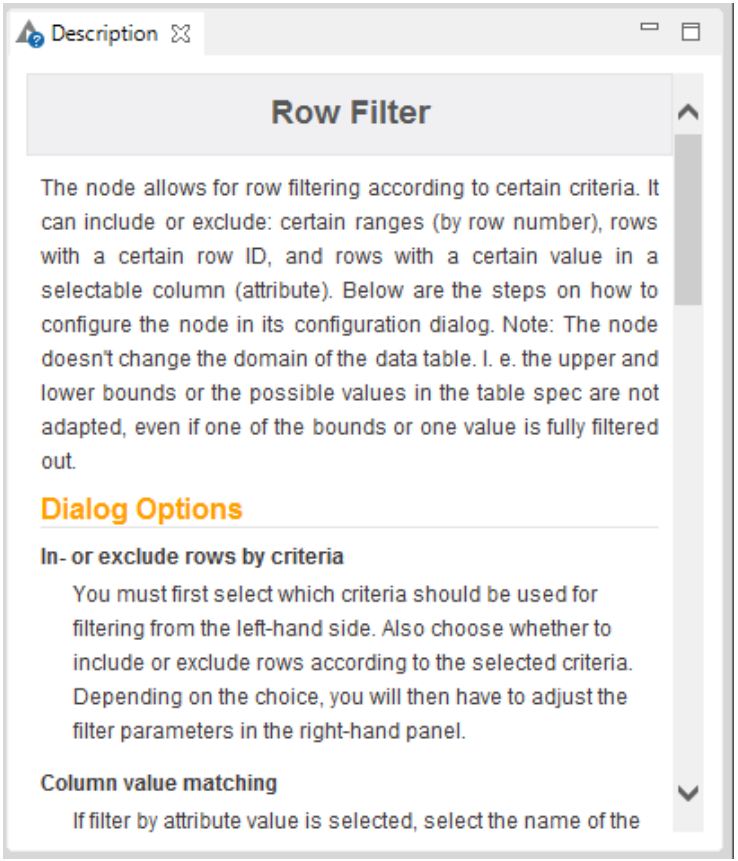


Node Repository



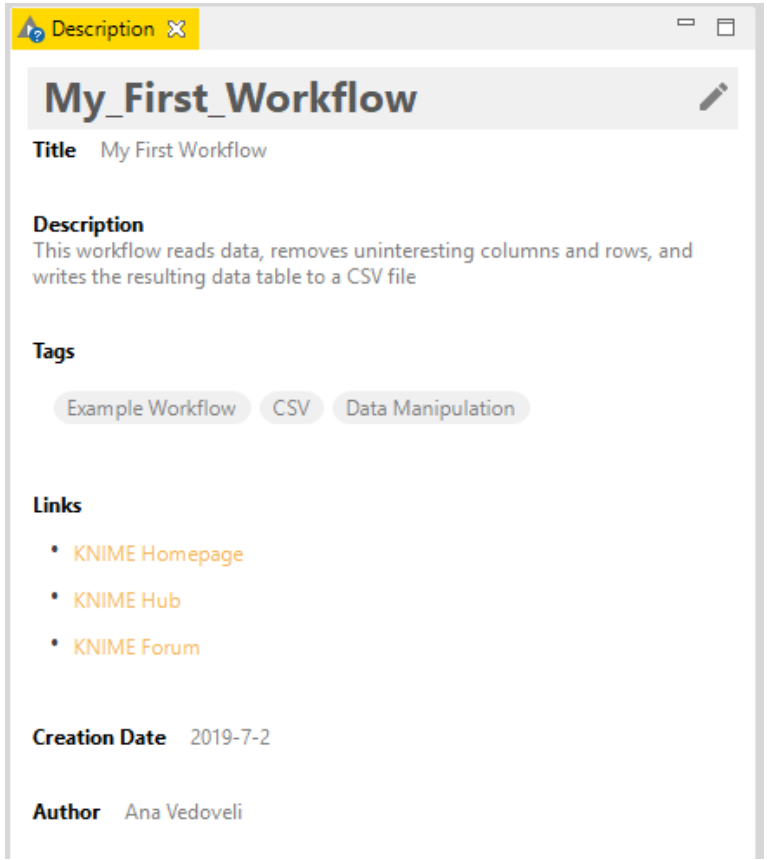
- The Node Repository lists all KNIME nodes
- The search box has 2 modes
  - Standard Search – exact match of node name
  - Fuzzy Search – finds the most similar node name

Description



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

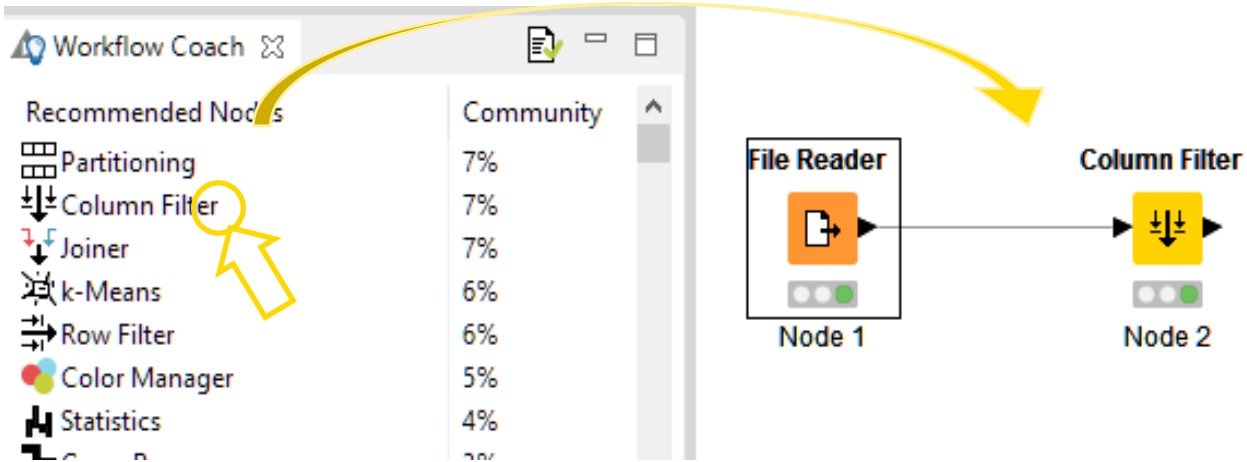
Workflow Description



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

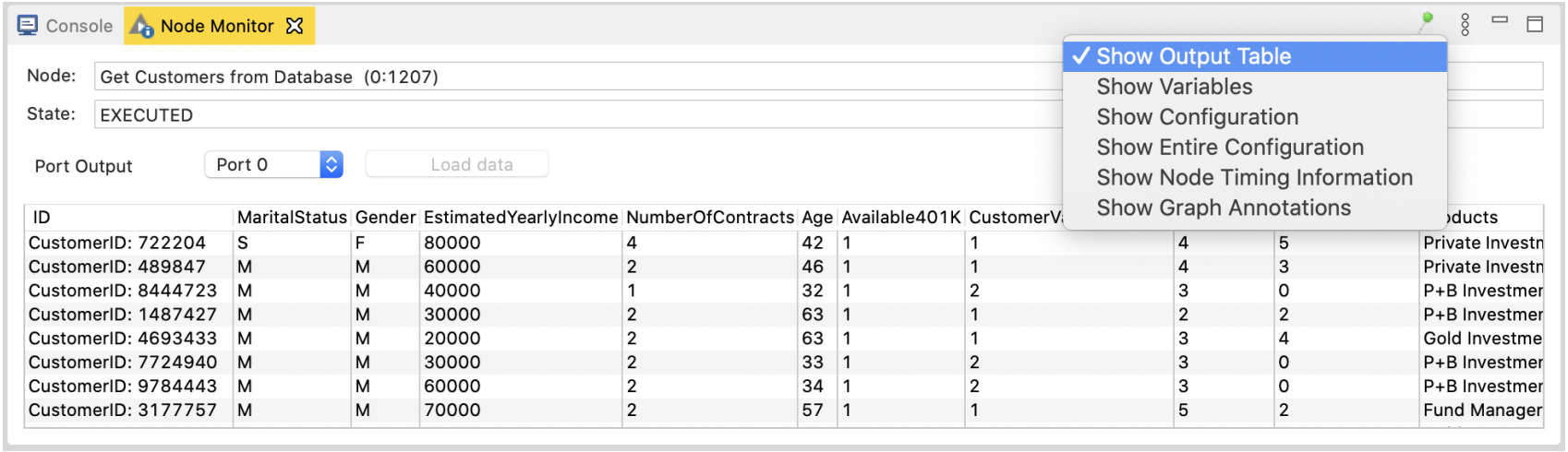
Workflow Coach

- 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



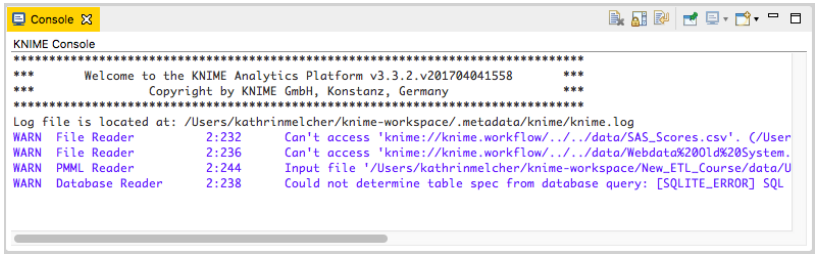
Node Monitor

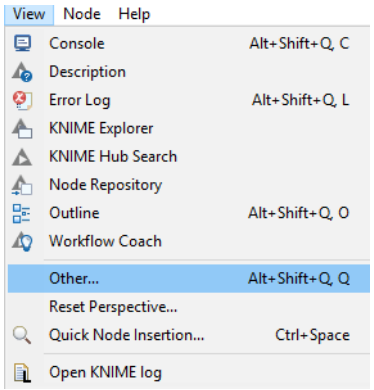
- By default the Node Monitor shows you the output table of the node selected in the workflow editor
- Click on the three dots on the upper right to show the flow variables, configuration, etc.

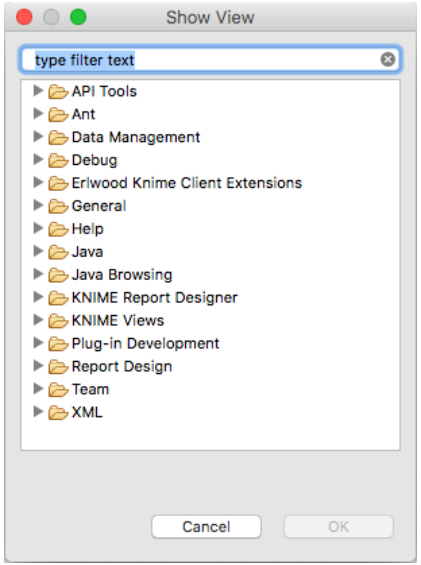


ID	MaritalStatus	Gender	EstimatedYearlyIncome	NumberOfContracts	Age	Available401K	CustomerV			oducts
CustomerID: 722204	S	F	80000	4	42	1	1	4	5	Private Investn
CustomerID: 489847	M	M	60000	2	46	1	1	4	3	Private Investn
CustomerID: 8444723	M	M	40000	1	32	1	2	3	0	P+B Investmer
CustomerID: 1487427	M	M	30000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

Console and Other Views



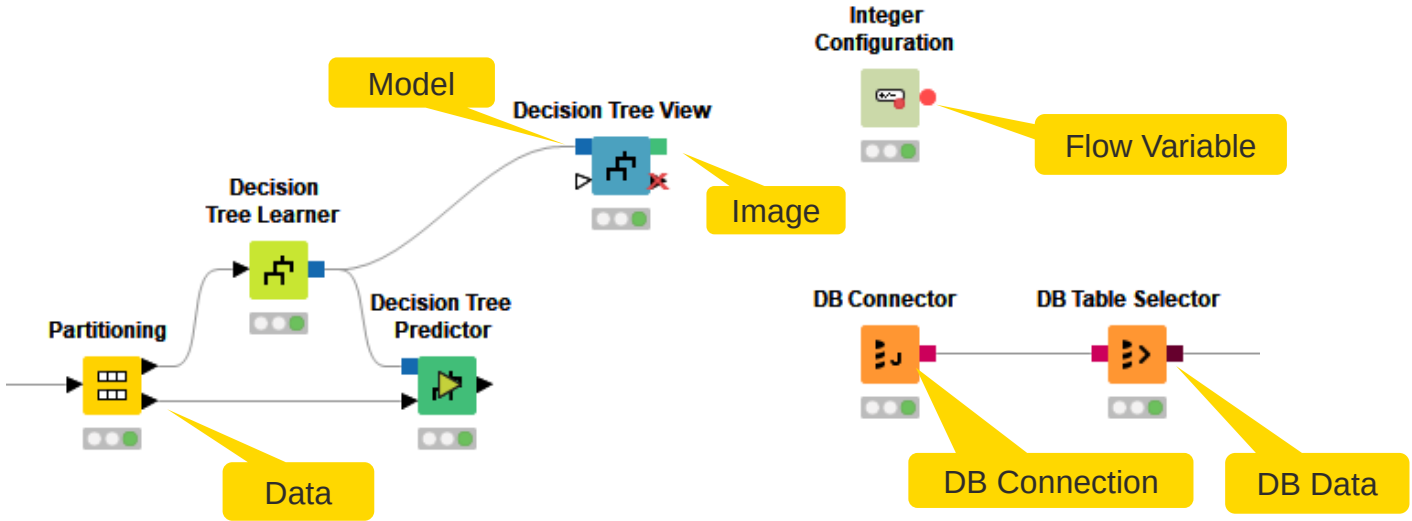




- Console view prints out error and warning messages about what is going on under the hood
- Click View and select Other… to add different views
  - Node Monitor, Licenses, etc.

Inserting and Connecting Nodes

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

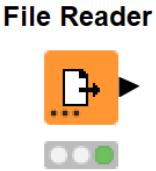


More on Nodes…

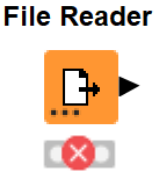
- A node can have 4 states:

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

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



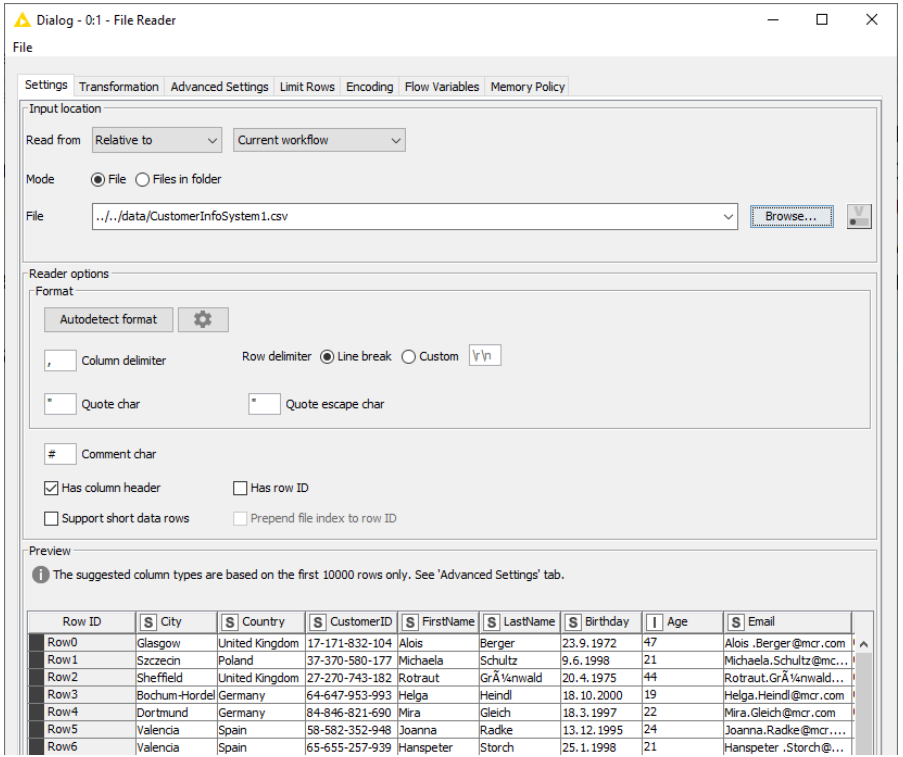
Executed:
The node has been successfully executed. Results may be viewed and used in downstream nodes.



Error:
The node has encountered an error during execution.

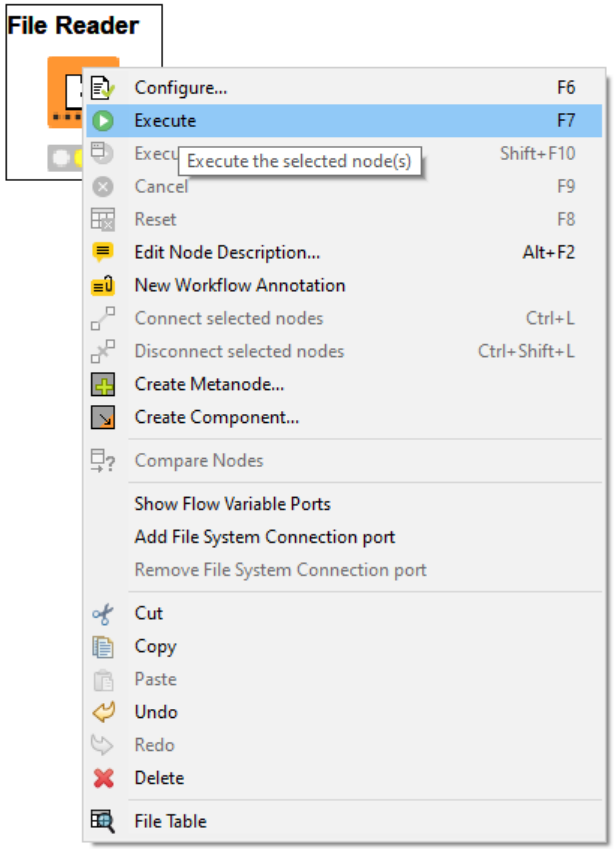
Node Configuration

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



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

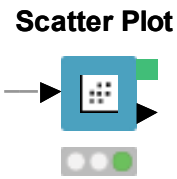
Execute selected nodes and open first view

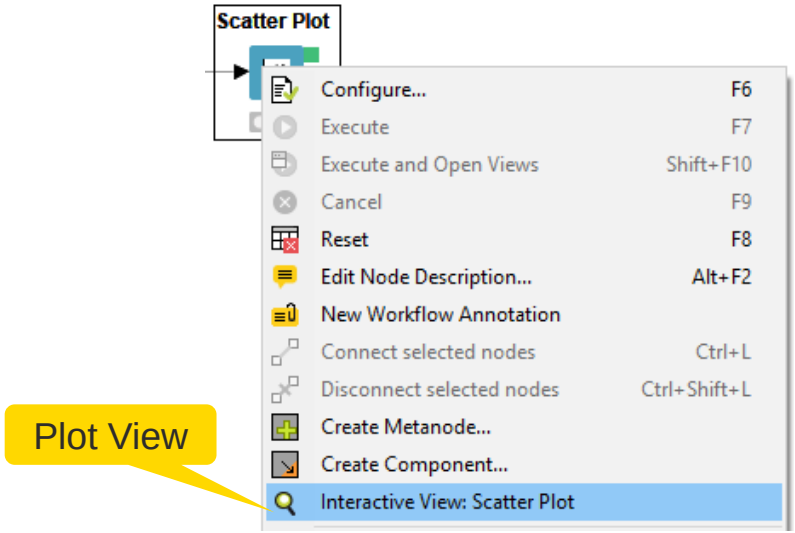
Cancel all selected, running nodes (F9)

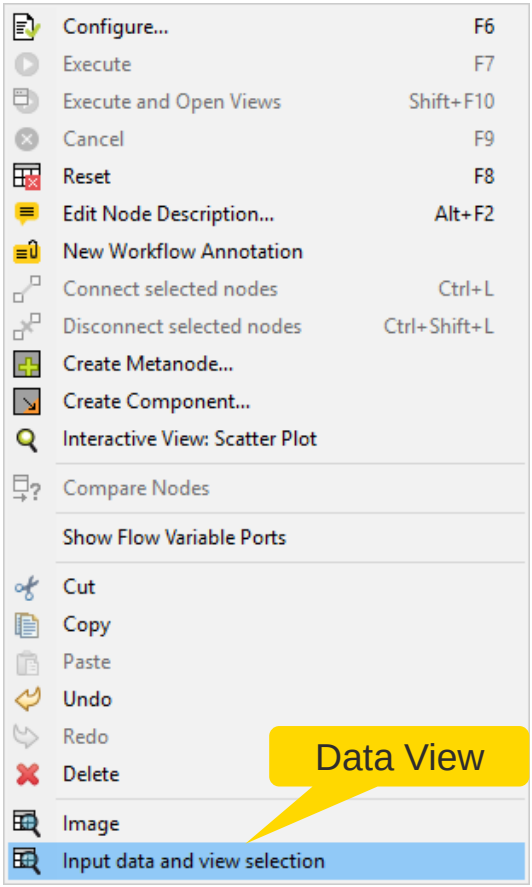
Cancel all running nodes

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

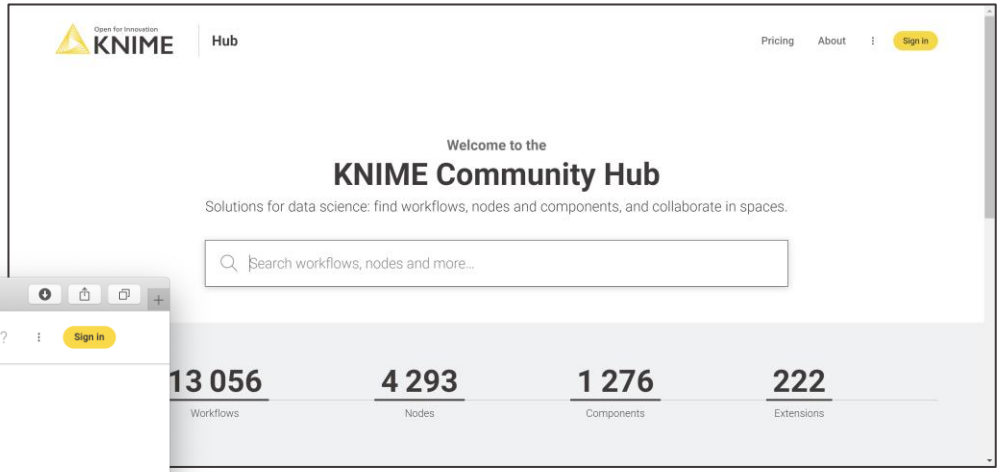
Dedicated file extensions for workflows and workflow groups associated with KNIME Analytics Platform

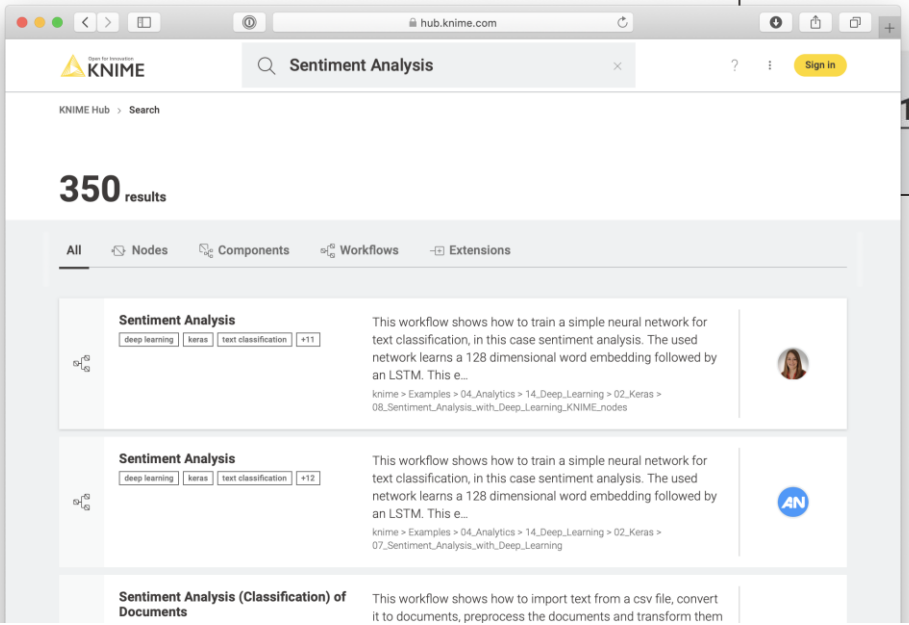
- ***.knwf** for KNIME Workflow Files

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

Getting Started: KNIME Community Hub

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

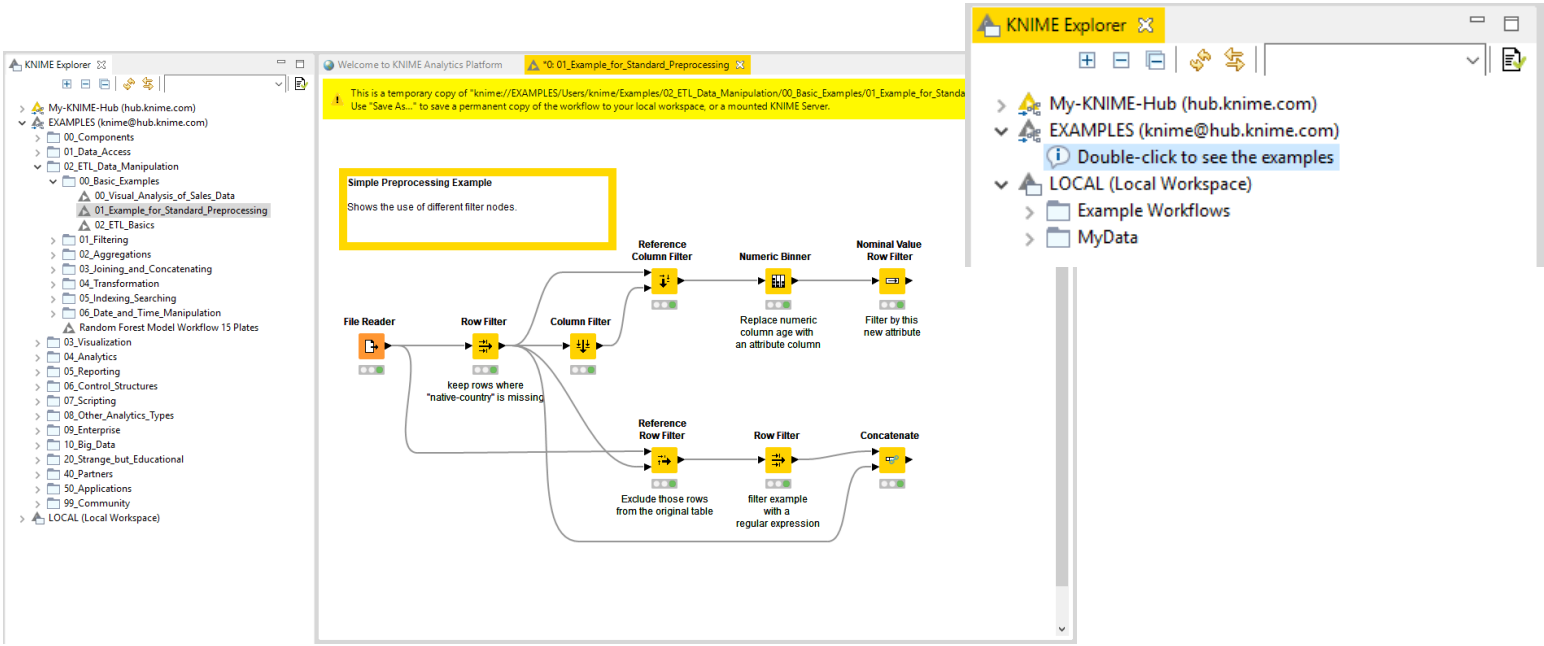




https://hub.knime.com

Getting Started: KNIME Examples

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

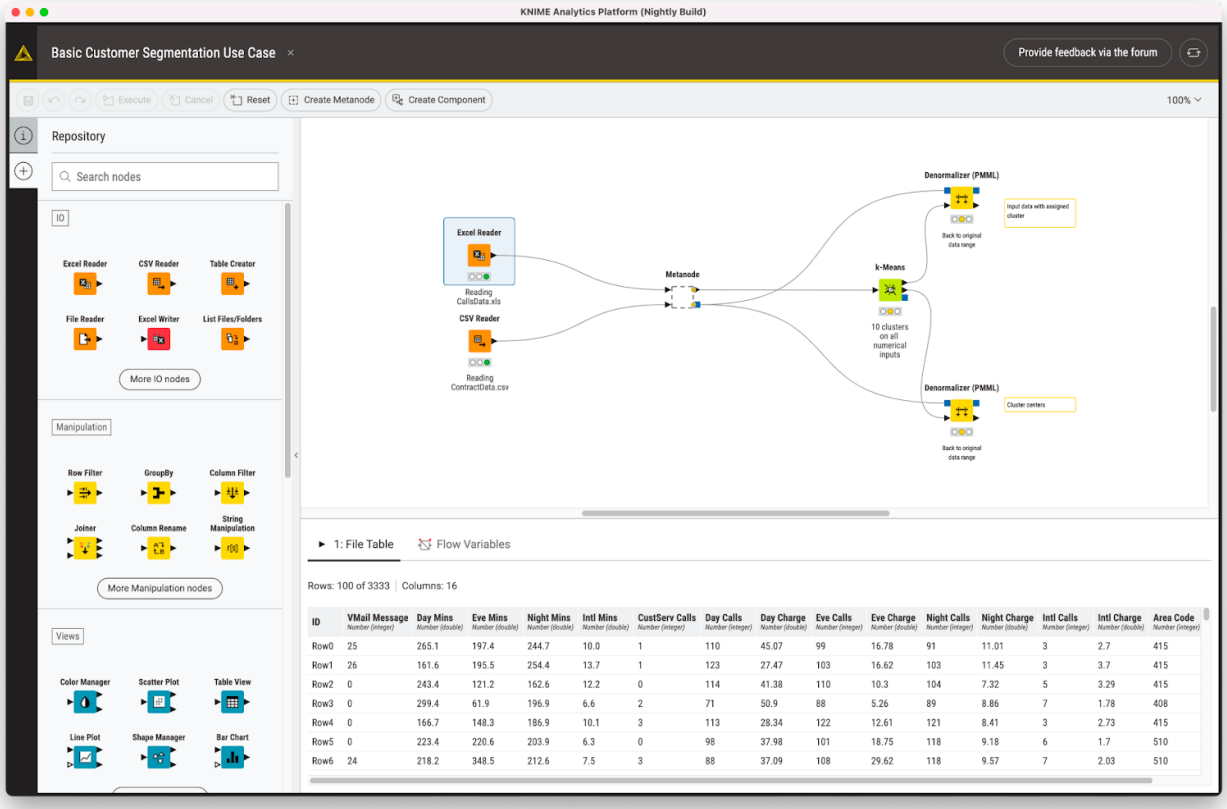


Hot Keys (for Future Reference)

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"



ID	VMail Message	Day Mins	Eve Mins	Night Mins	Intl Mins	CustServ Calls	Day Calls	Day Charge	Eve Calls	Eve Charge	Night Calls	Night Charge	Intl Calls	Intl Charge	Area Code
	Number (integer)	Number (double)	Number (double)	Number (double)	Number (double)	Number (integer)	Number (integer)	Number (double)	Number (integer)	Number (double)	Number (integer)	Number (double)	Number (integer)	Number (double)	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

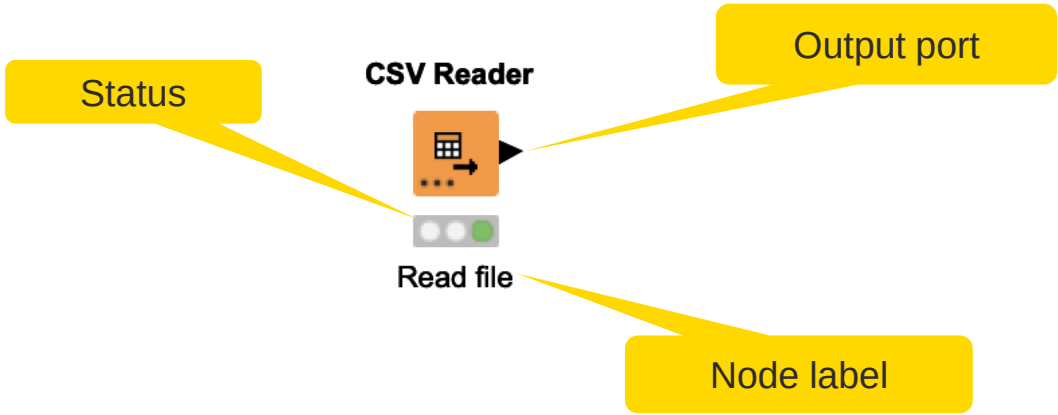
- Analyze data from a retail company, which has an online shop and stores
- Data:
  - Customer information from two different systems (.csv, .table)
  - Purchases from the online store (sqlite database)
  - List of product numbers and prices (sqlite database)
  - Purchases from the stores (.table)
 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

Importing Data

Data Source Nodes

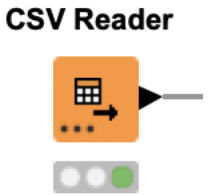
Typically characterized by:

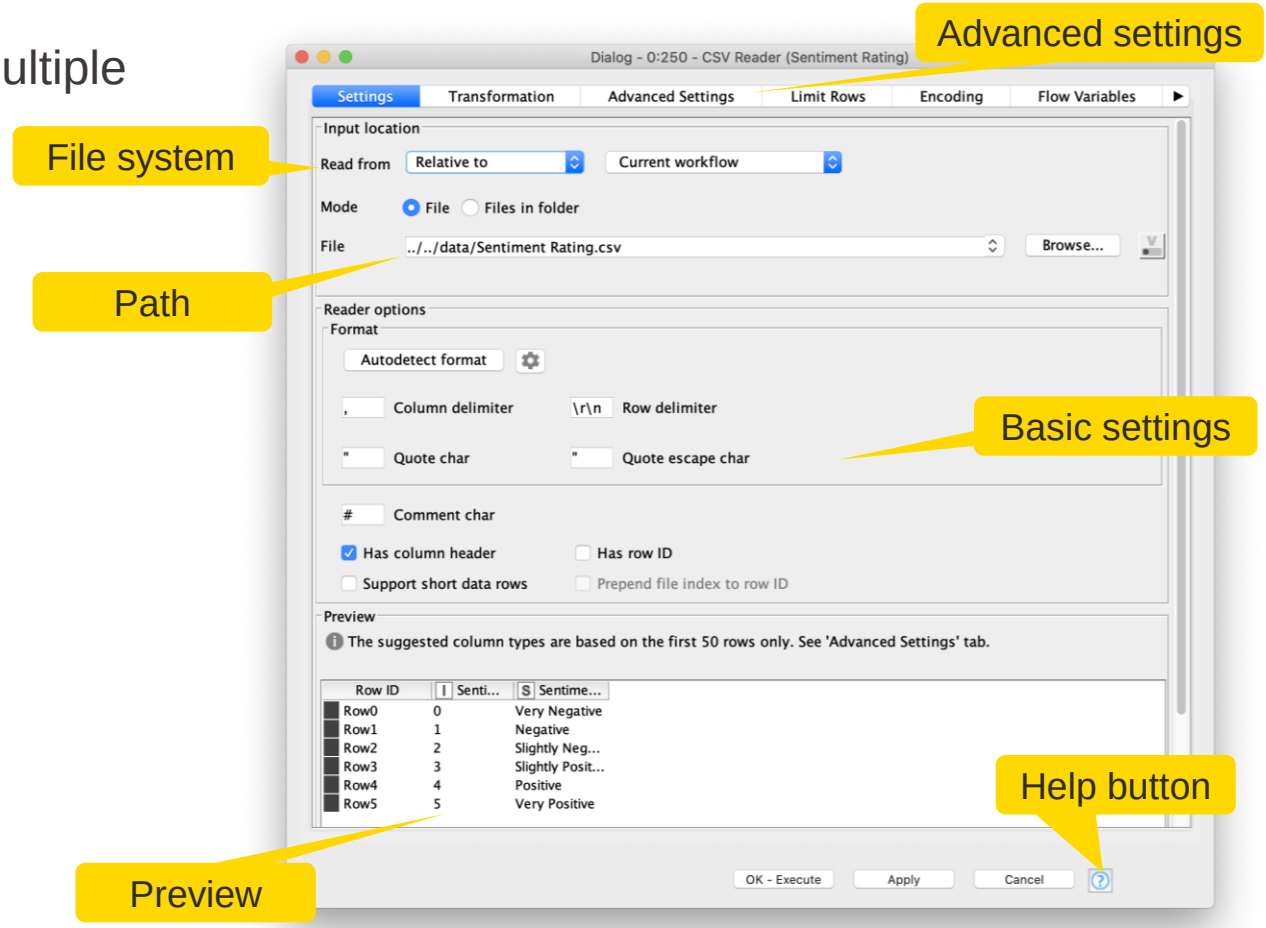
- Orange color
- By default no input ports, 1-2 output ports
- New file handling with KNIME 4.3.
  - Consistent user experience across all nodes and file systems
  - Managing of various file systems within the same workflow
  - Performance improvements



CSV Reader

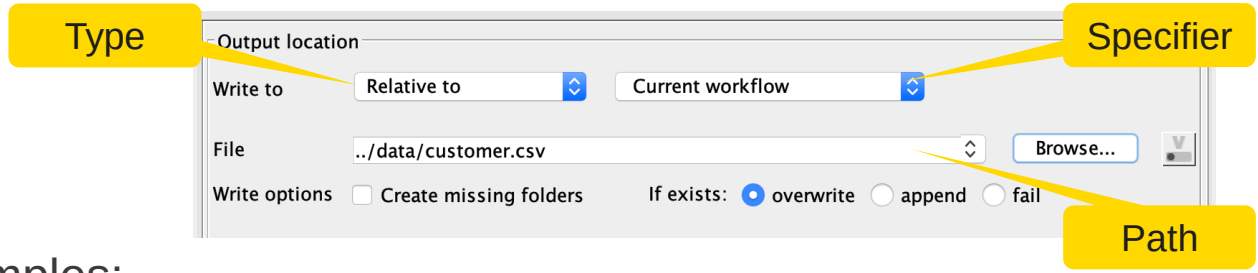
- Reads either one or multiple .csv and .txt files
- Further tabs to
  - limit the rows
  - select encoding





Common Settings: File Path

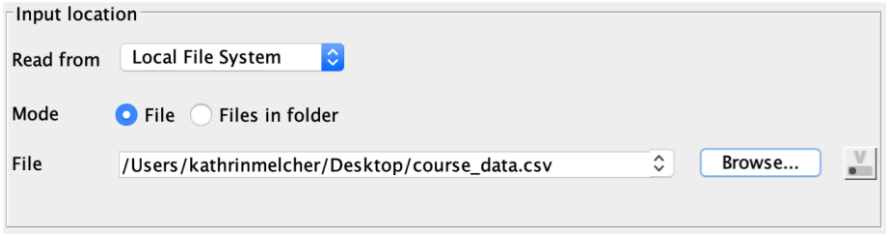
- A path consists of three parts:
  - **Type**: Specifies the file system type e.g. local, relative, mountpoint, custome_url or connected.
  - **Specifier**: Optional string with additional file system specific information e.g. relative to which location (knime.workflow)
  - **Path**: Specifies the location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



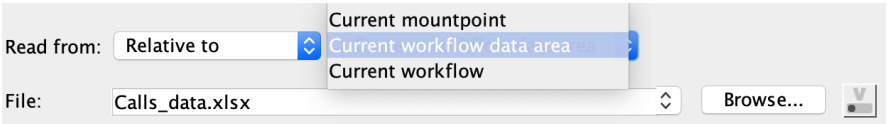
- Examples:
  - (LOCAL, , C:\Users\username\Desktop)
  - (RELATIVE, knime.workflow, file1.csv)
  - (MOUNTPOINT, MOUNTPOINT_NAME, /path/to/file1.csv)
  - (CONNECTED, amazon-s3:eu-west-1, /mybucket/file1.csv)

Common Settings: Four Default File Systems

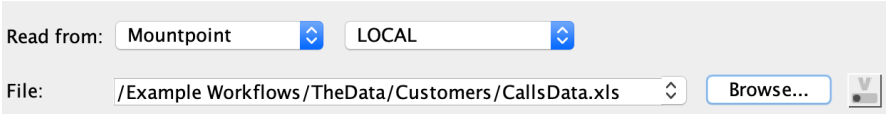
- **Local File System**



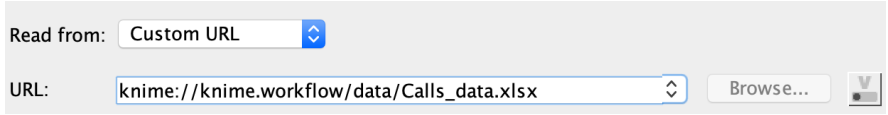
- **Relative to …**



- **Mountpoint**

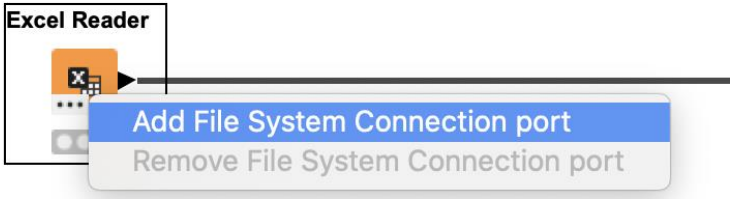


- **Custom URL**

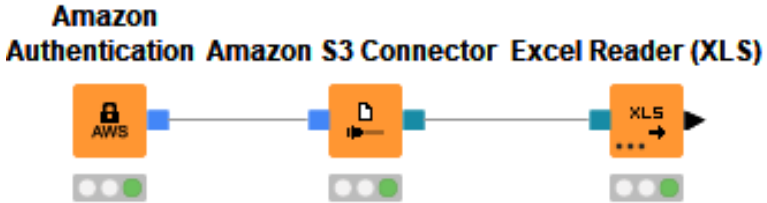


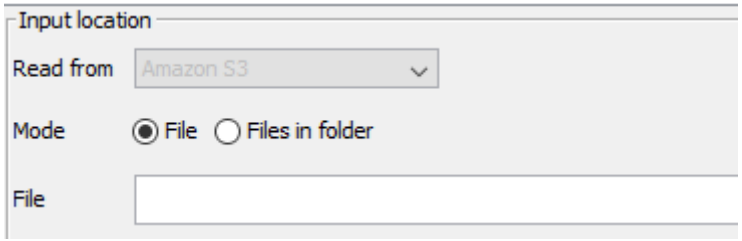
Common Settings: Connecting to other File Systems

- 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, …)

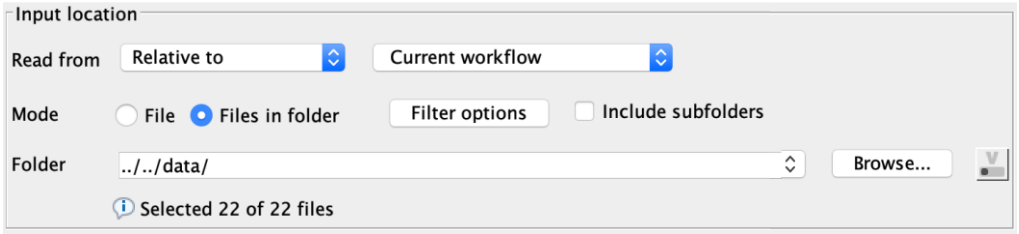




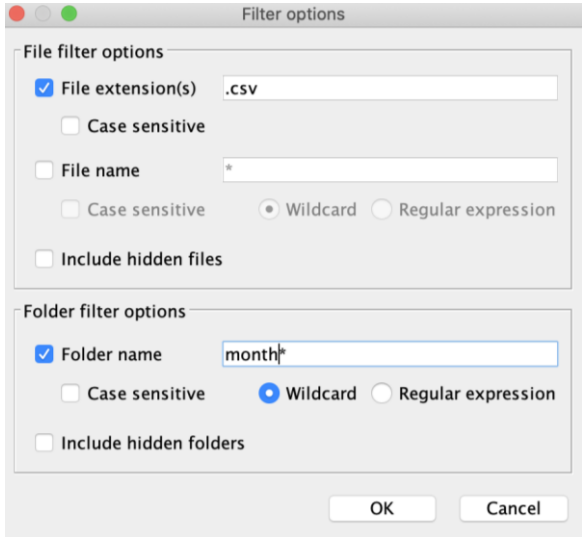
Common Settings: Read Single or Multiple Files

- Single file

- Files in a folder

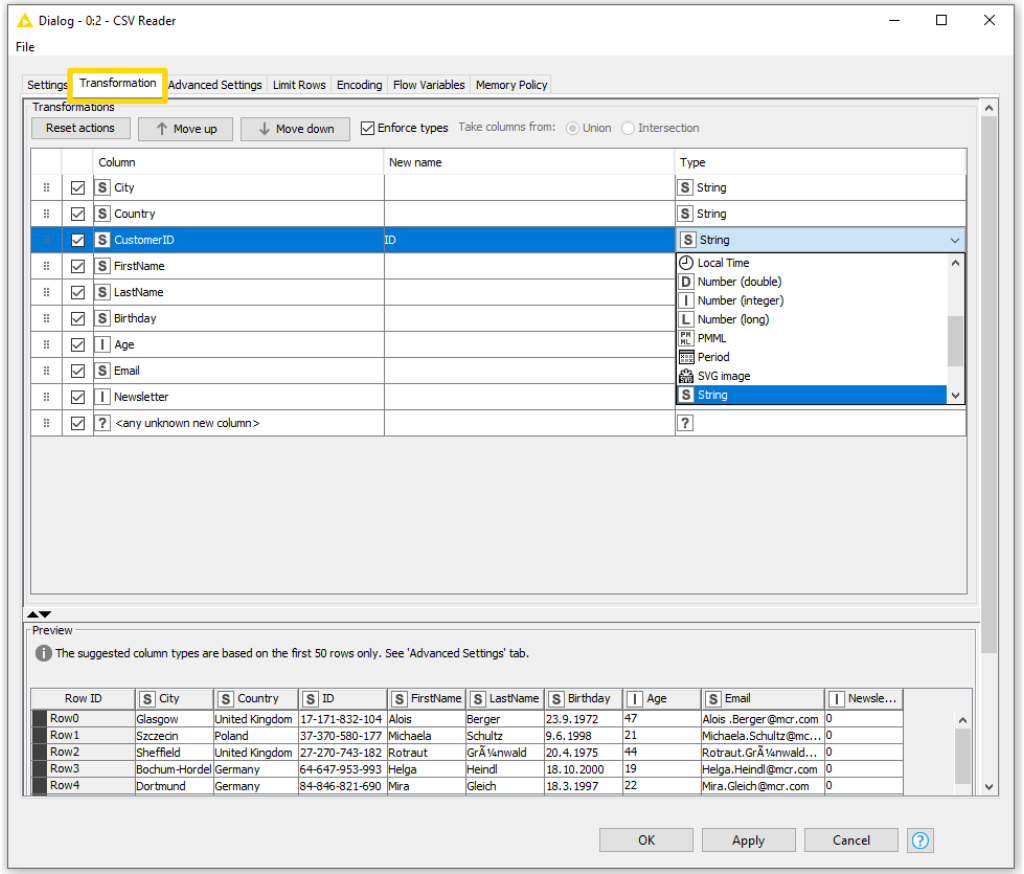


  - Option to include subfolder
  - Option to define filter criterions

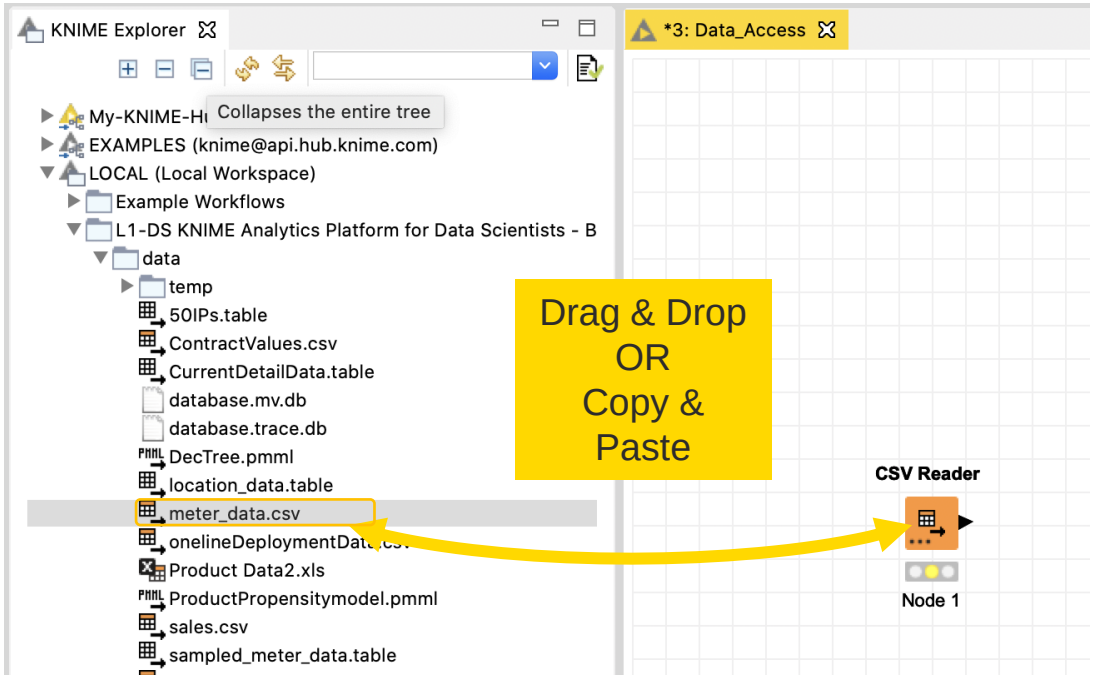


Common Settings: Transformation Tab

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

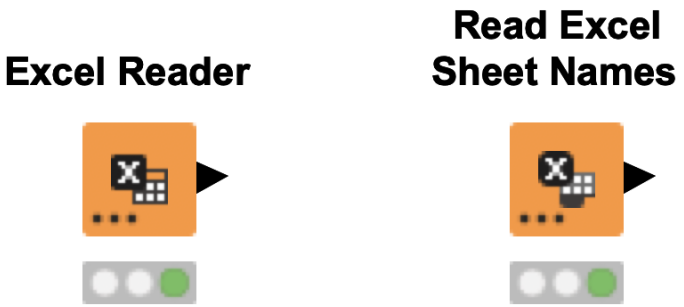


Alternative Faster Way …



Excel Reader (XLS)

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



Excel Reader

Excel Reader

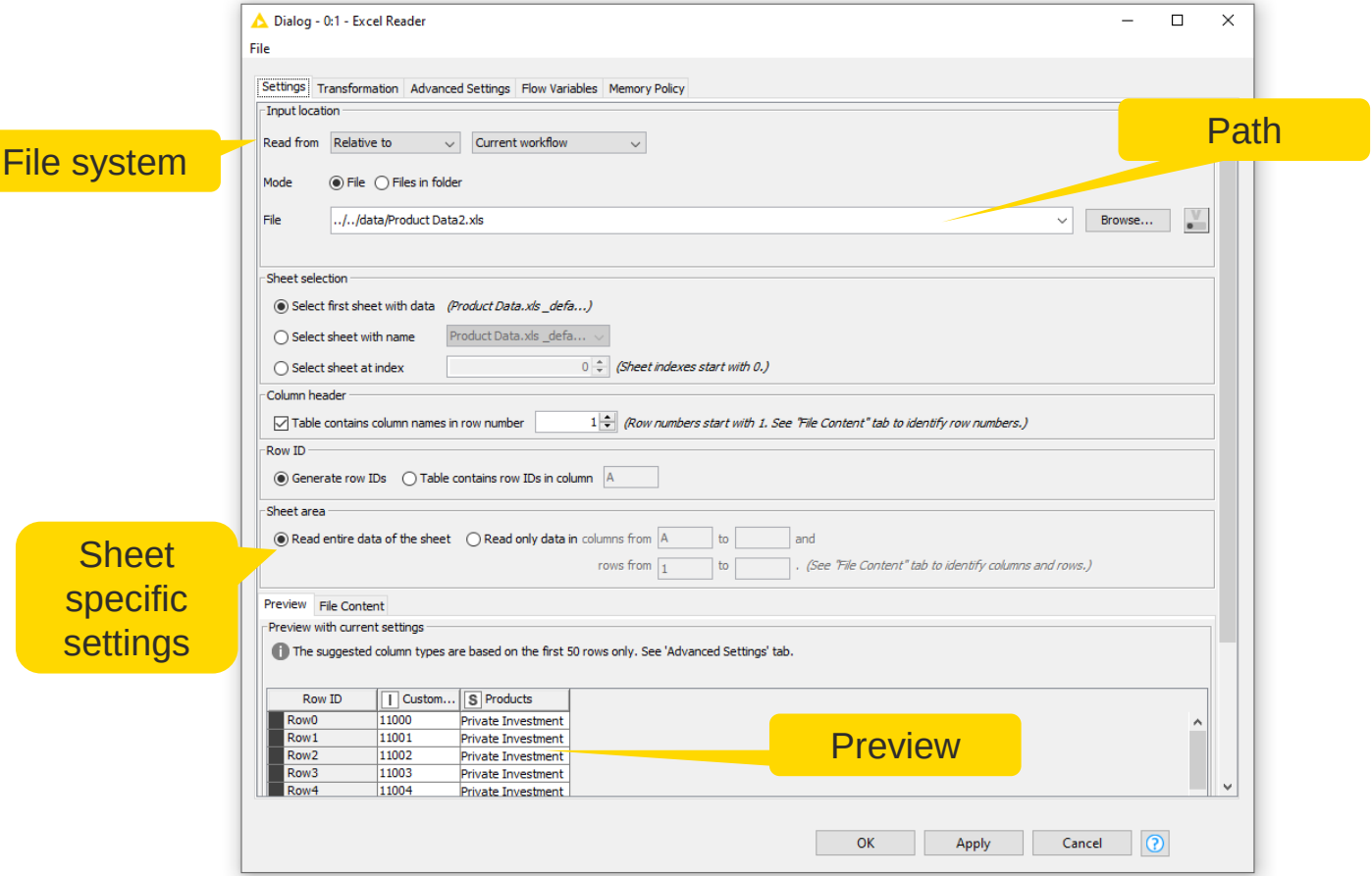
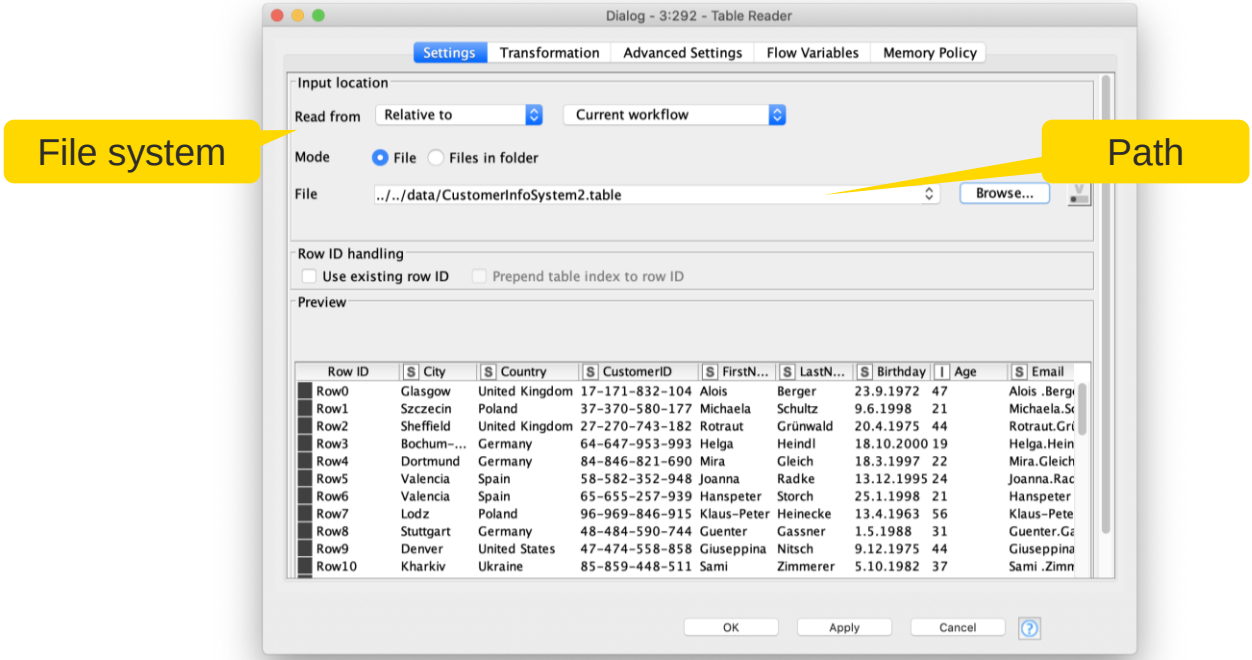


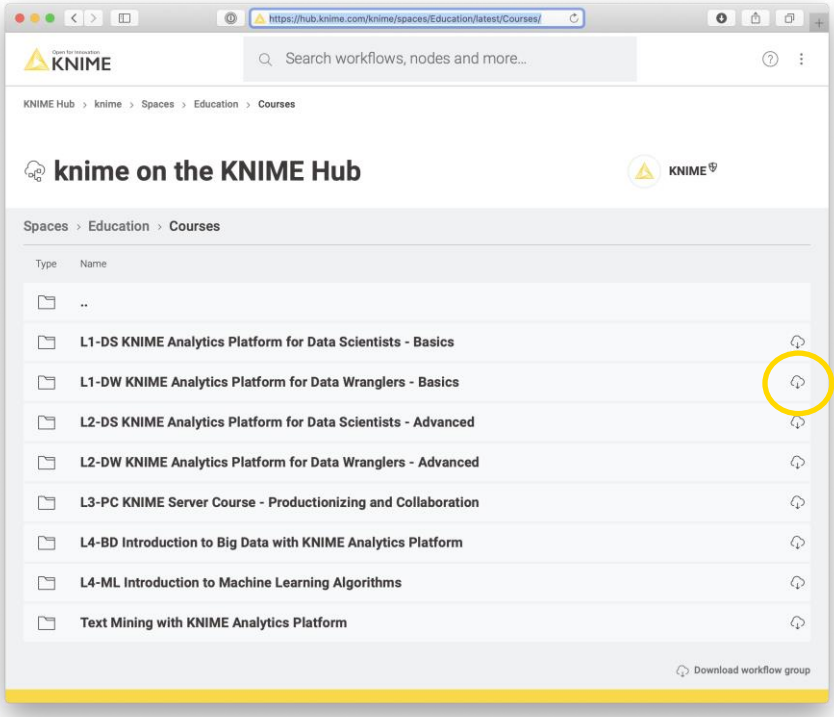
Table Reader

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



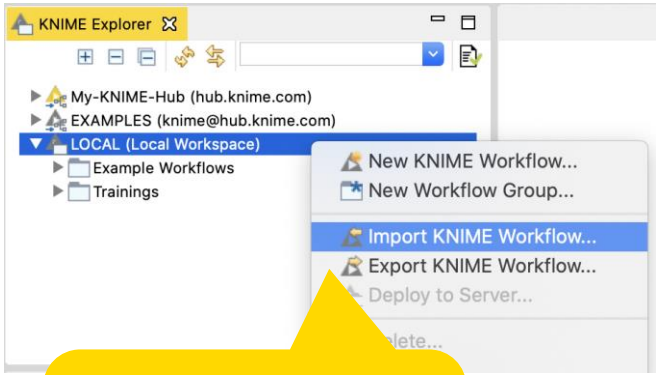
Downloading Exercises

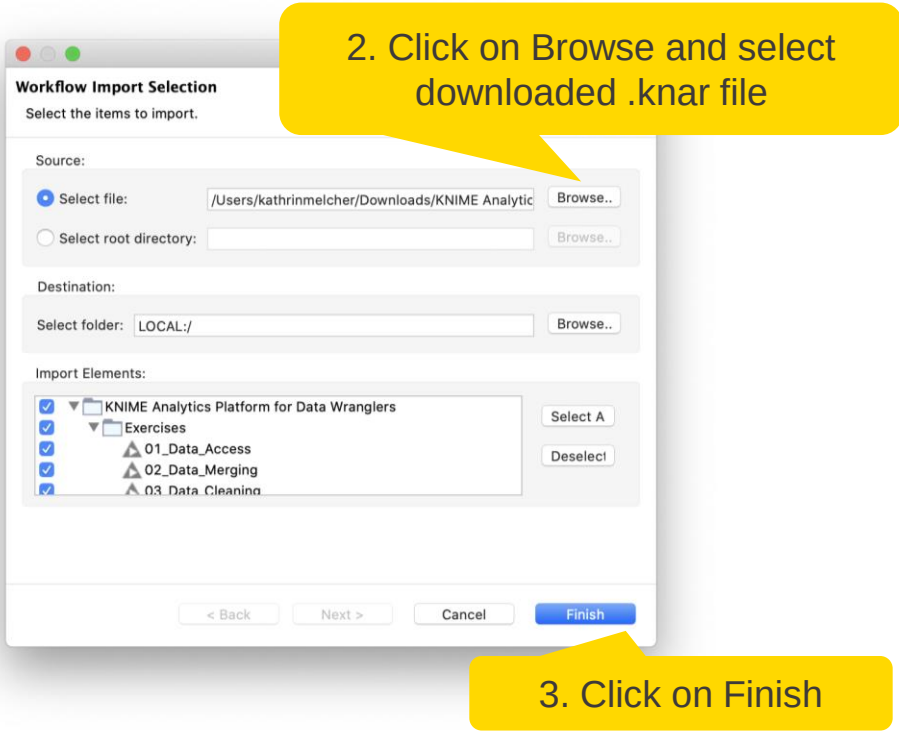
- Download the course material from the KNIME Community Hub https://hub.knime.com/knime/spaces/Education/latest/Courses/



Importing Exercises

- Import the course material to KNIME Analytics Platform

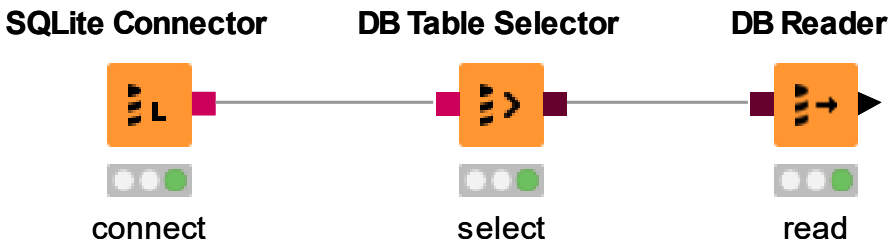




Accessing Databases

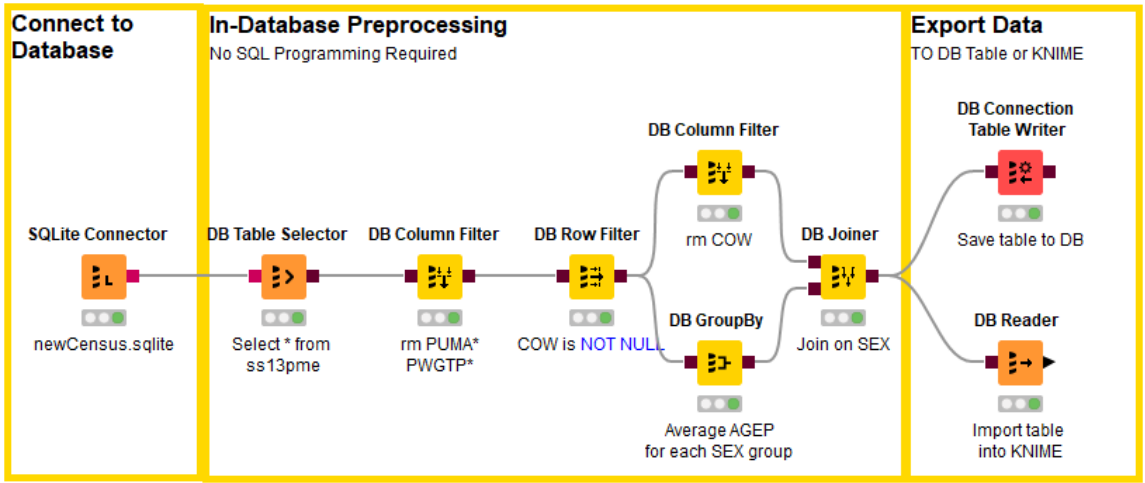
Database Connectivity

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

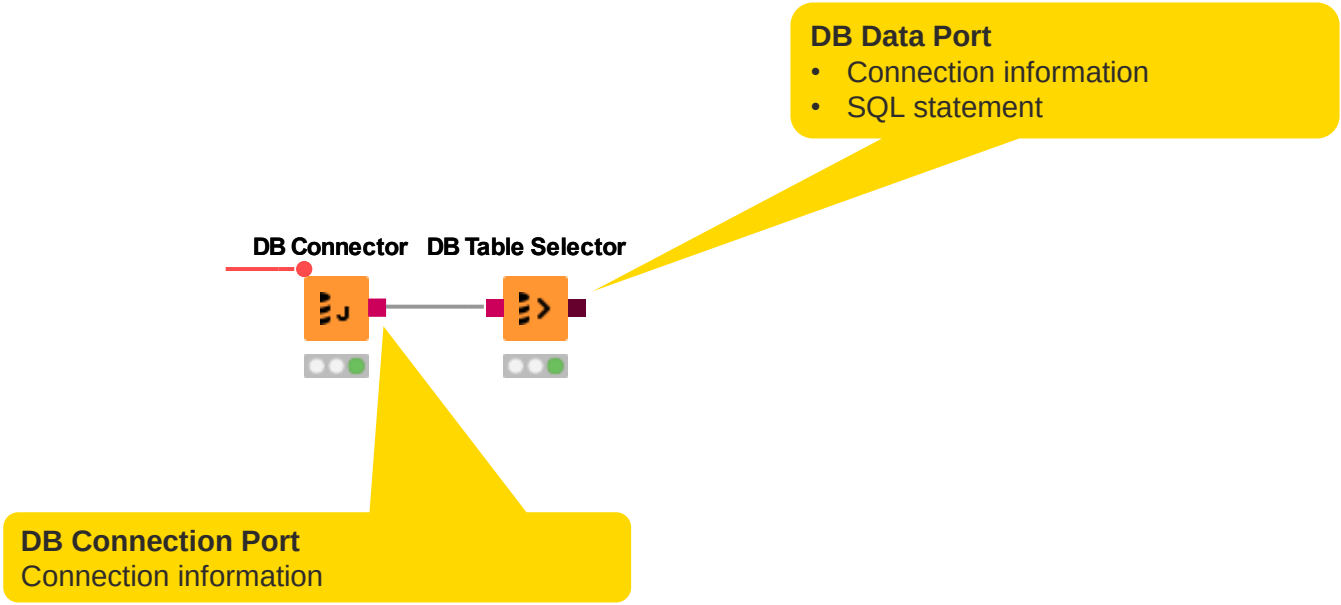


Database Extension

- 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

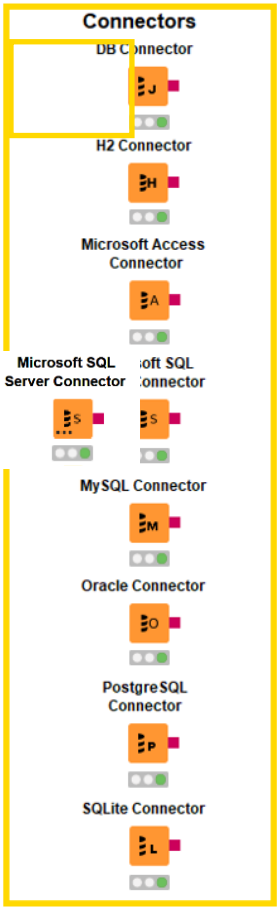


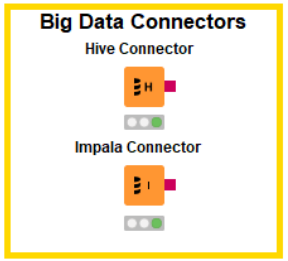
Database Port Types



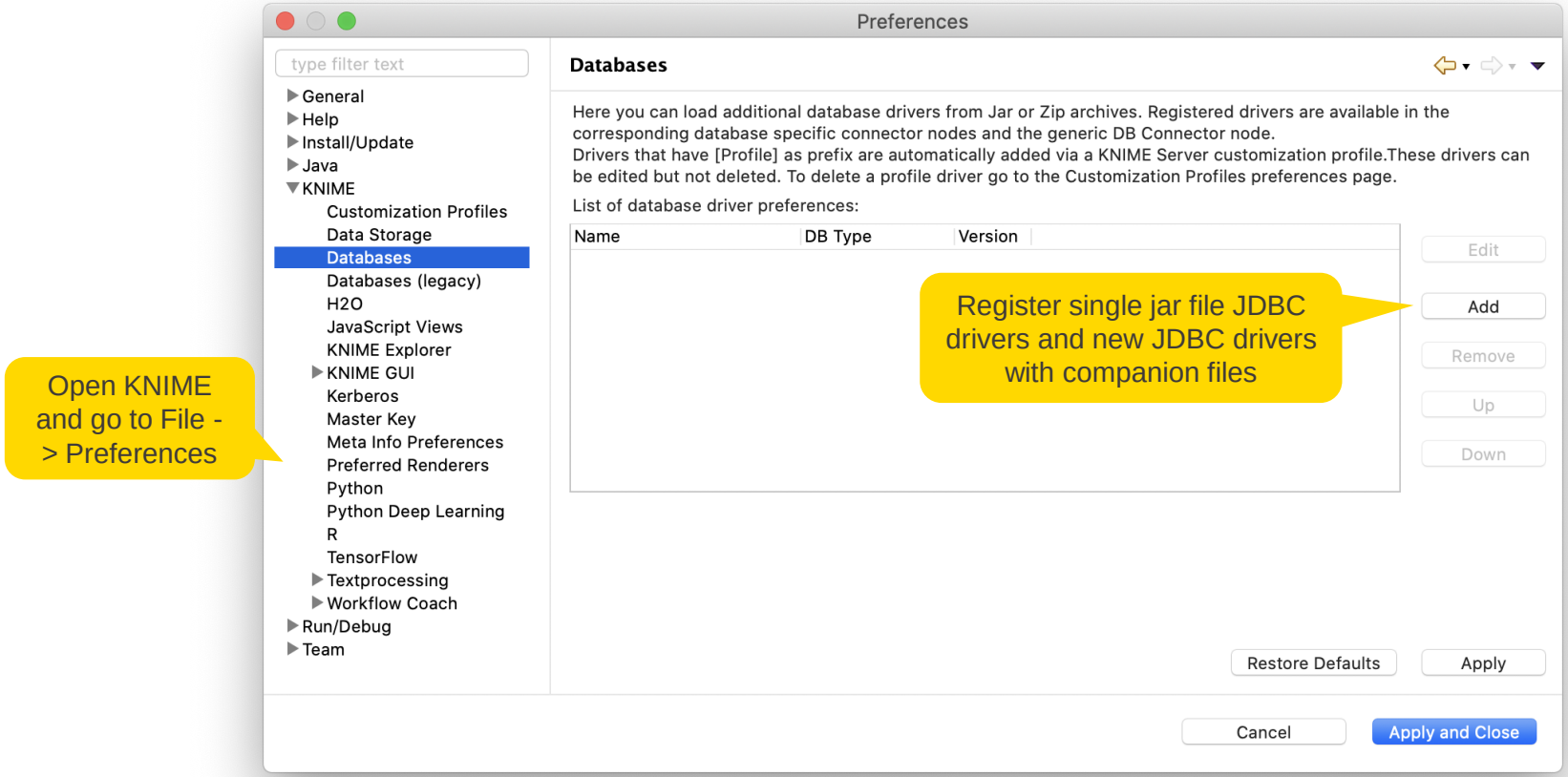
Database Connectors

- 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

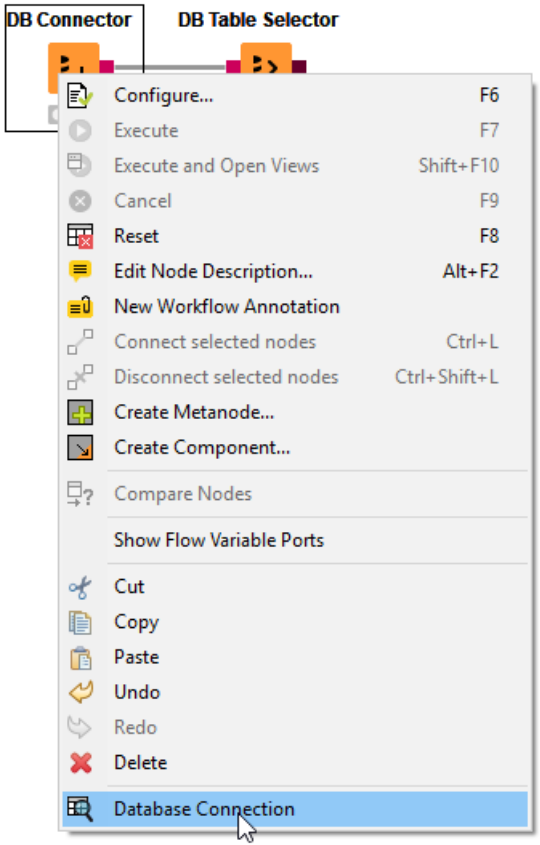


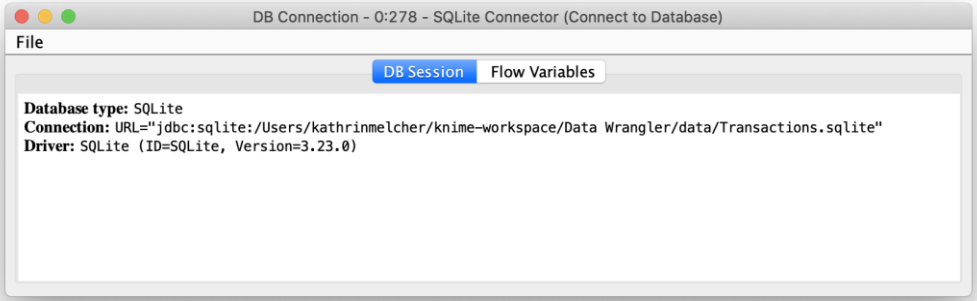


Register JDBC Driver



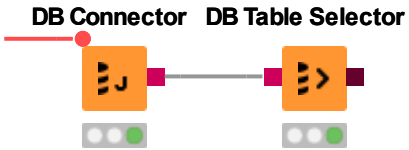
Database JDBC Connection Port View

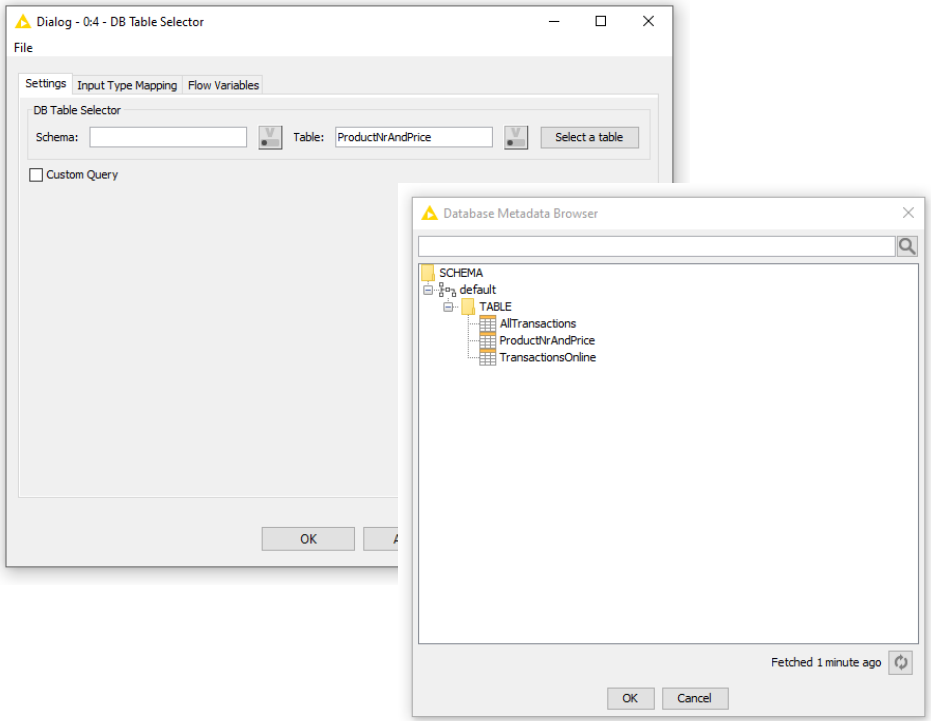




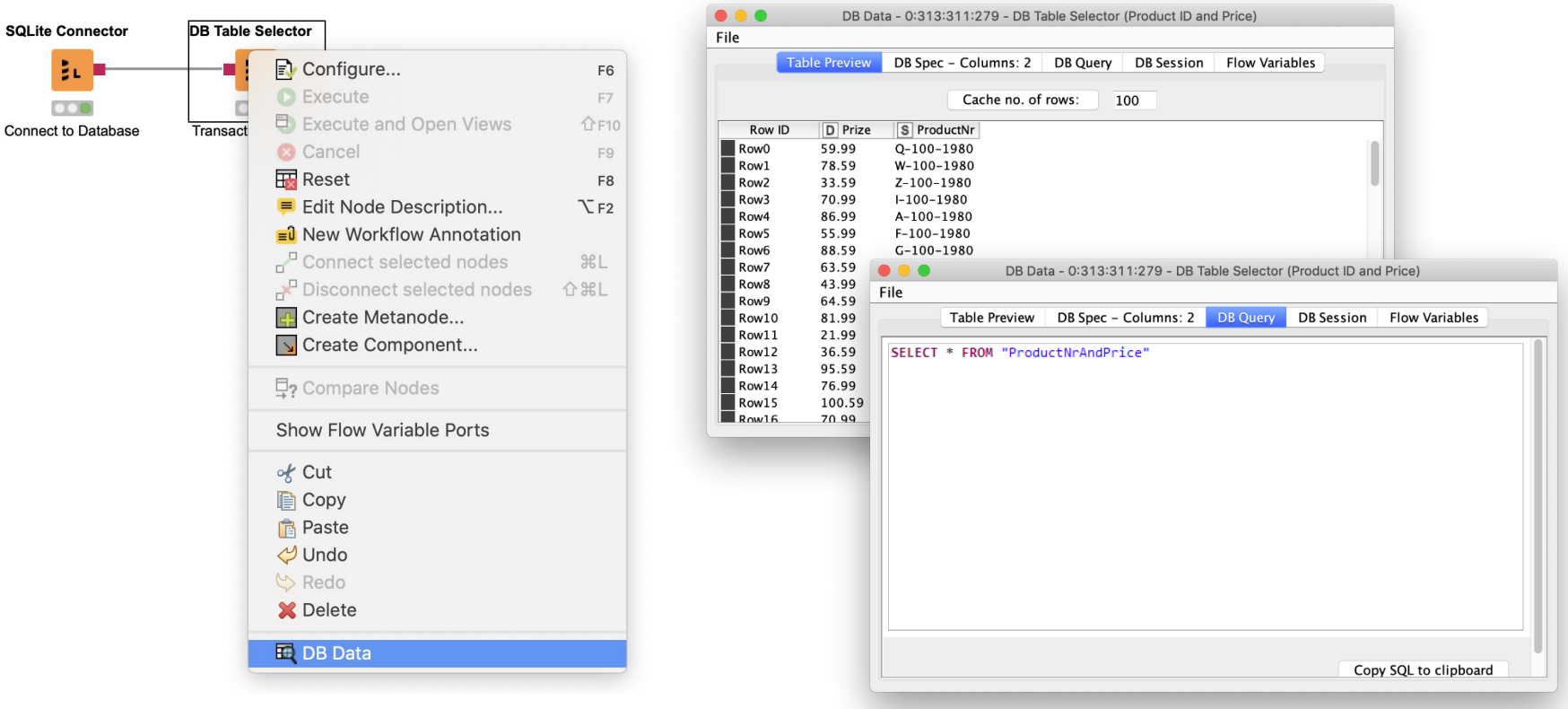
DB Table Selector

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





Database Connection Port View



DB Reader

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

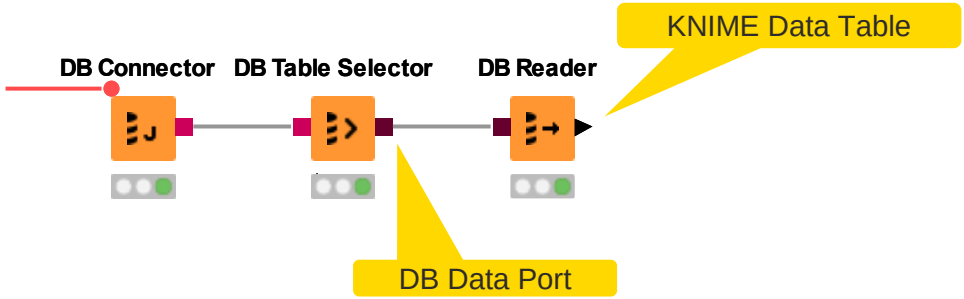
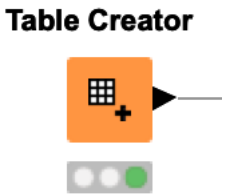
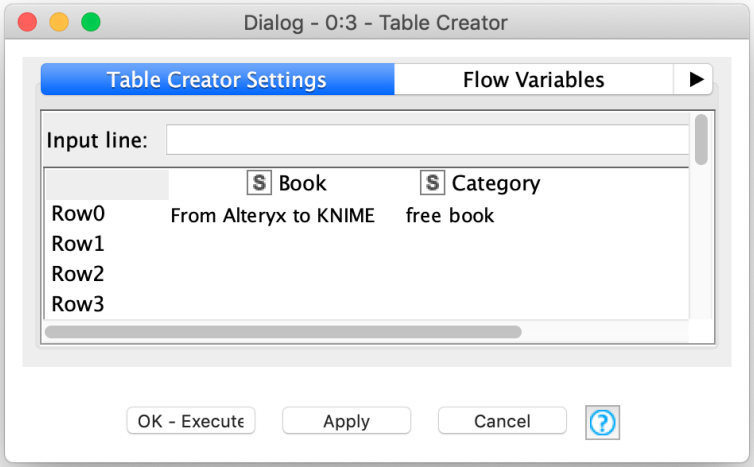


Table Creator

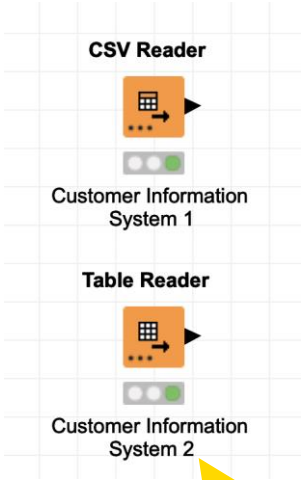
- Allows you to create data tables manually
- Data can be entered in a spreadsheet – like the table in the configuration dialog





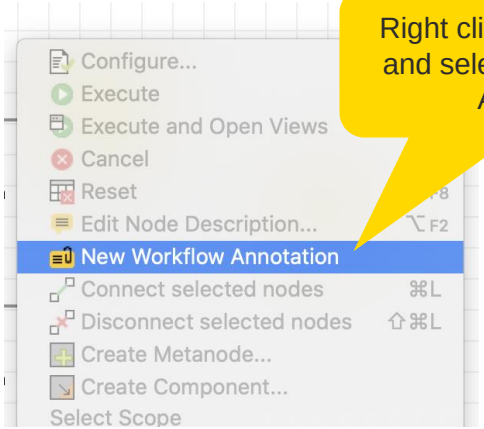
Comments & Annotations

Comments

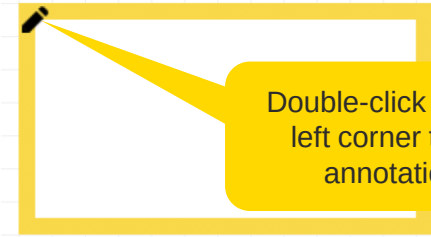


Double-click to change the node label

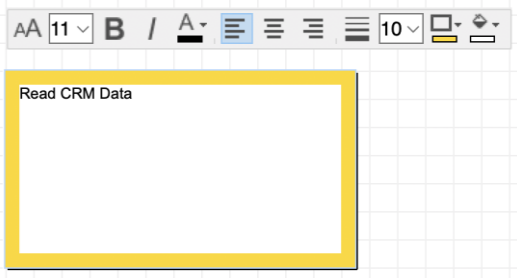
Annotations



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:

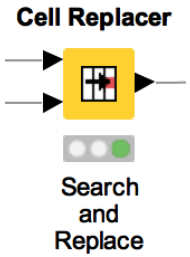
- Customer information
  - CustomerInfoSystem1.csv
  - CustomerInfoSystem2.table
- Online shop transactions, and product number & price information
  - TransactionOnline from Transations.sqlite
  - ProductNrAndPrice from Transations.sqlite
- Store transactions and information
  - Store.xlsx
  - TransactionsStore.table

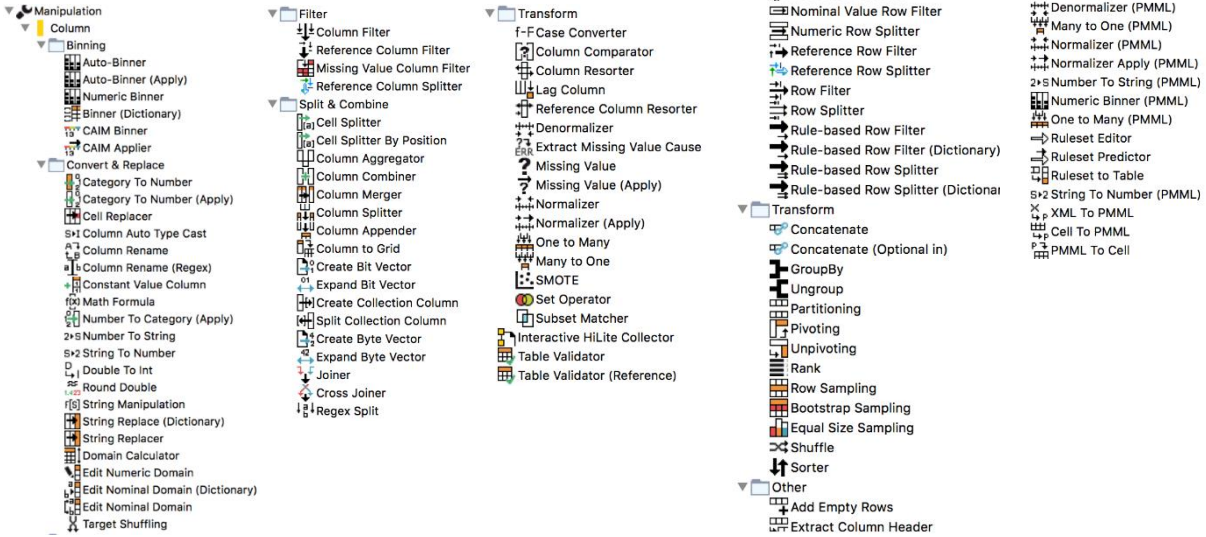
Try to use workflow relative-paths

Data Merging

Data Manipulation Nodes

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

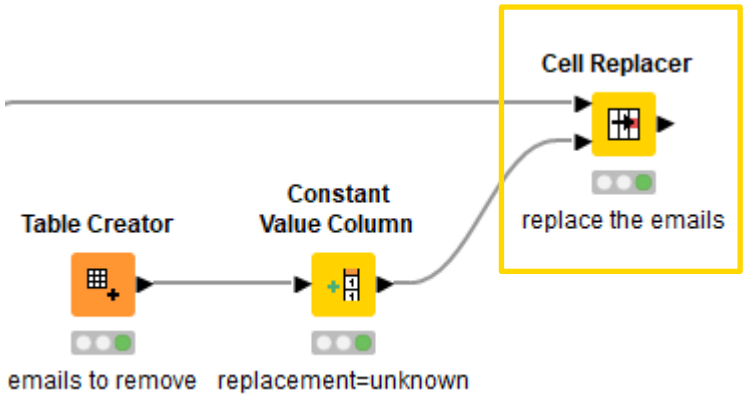


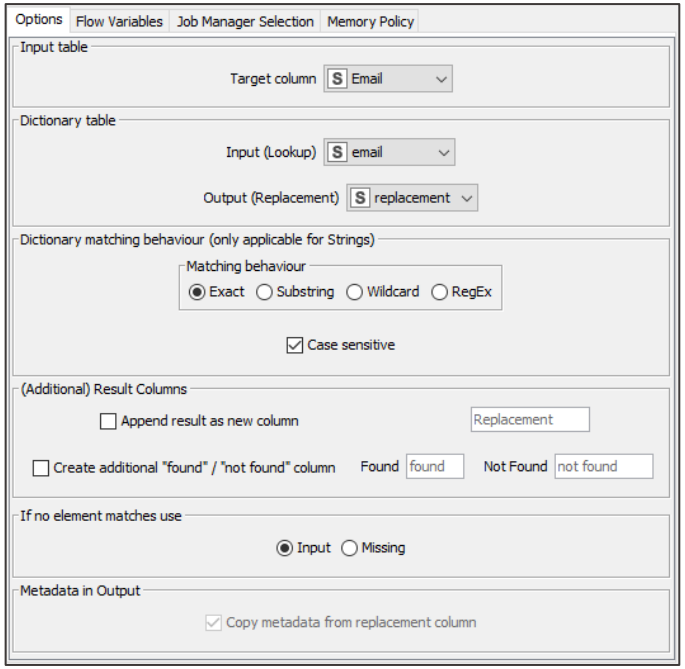


Cell Replacer

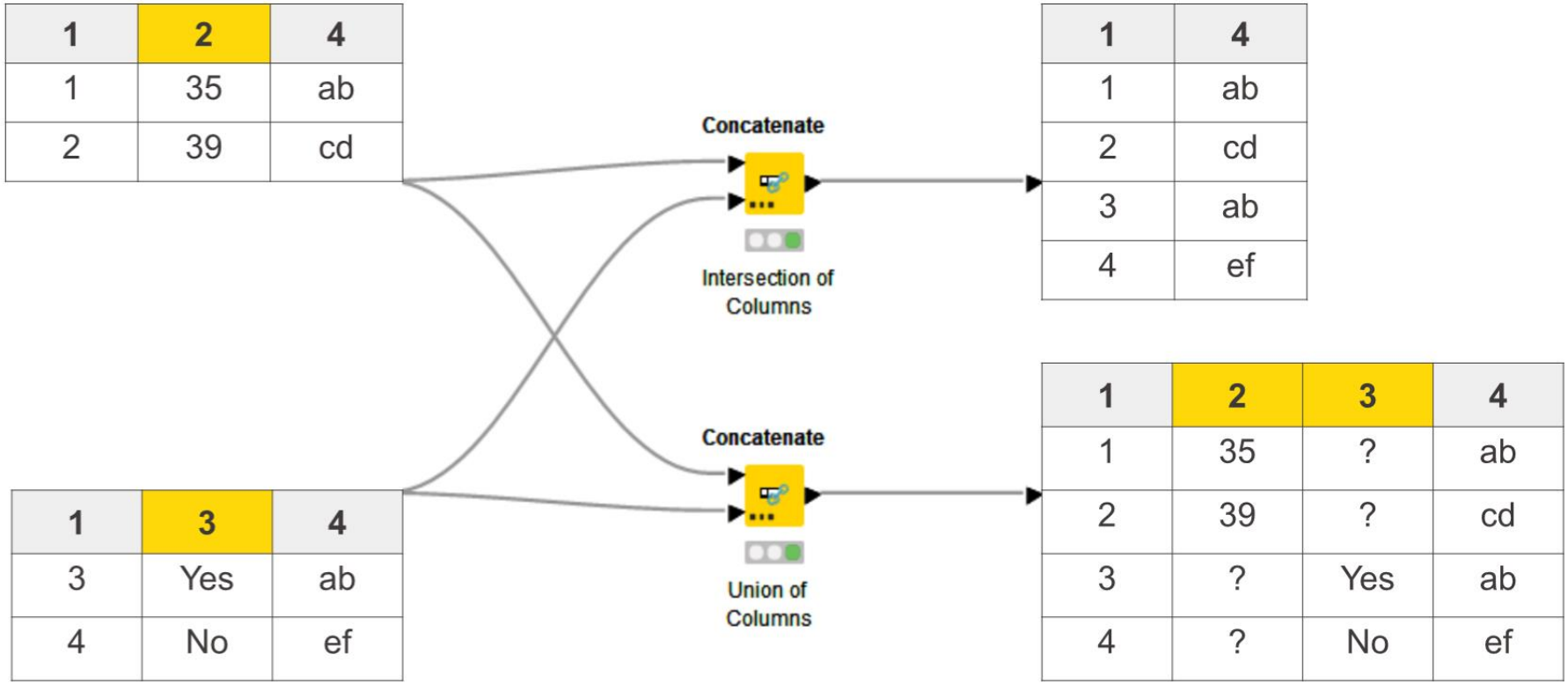
Replaces the content of a column based on a lookup

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





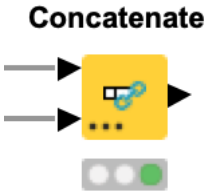
Concatenate two tables

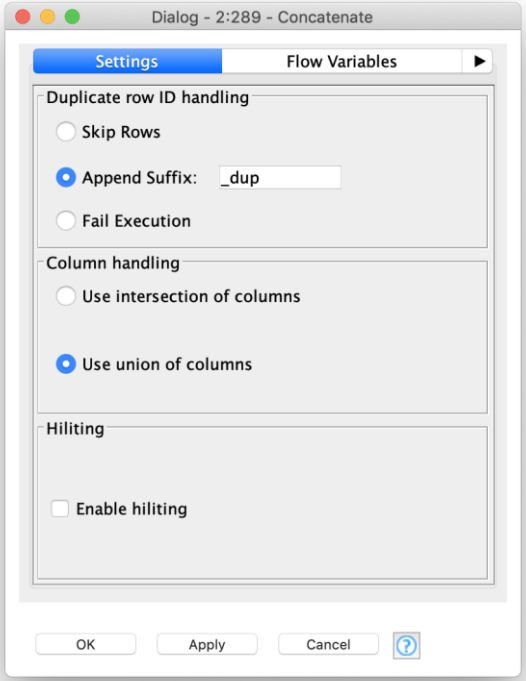


Concatenate

Combine rows from two tables with shared columns

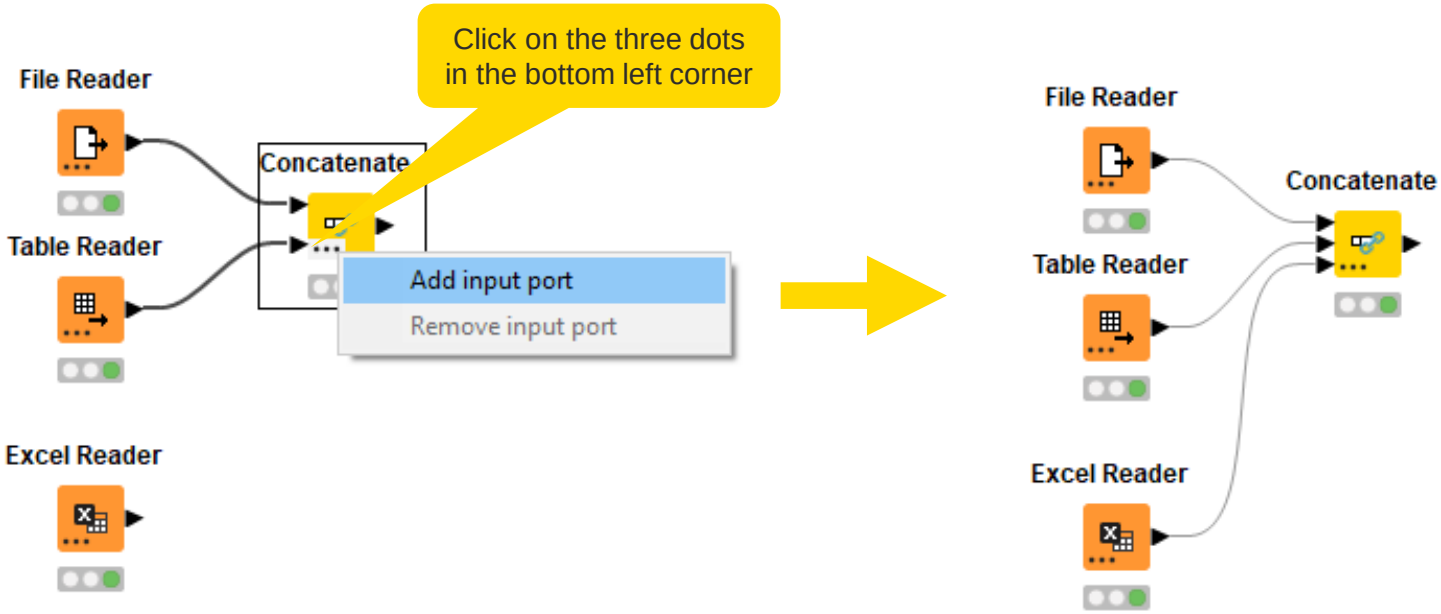
- Handles duplicate row keys gracefully
- Take the union or intersection of columns





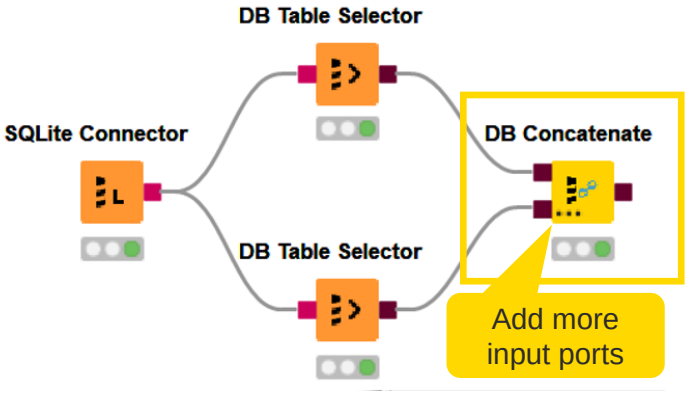
Dynamic Ports

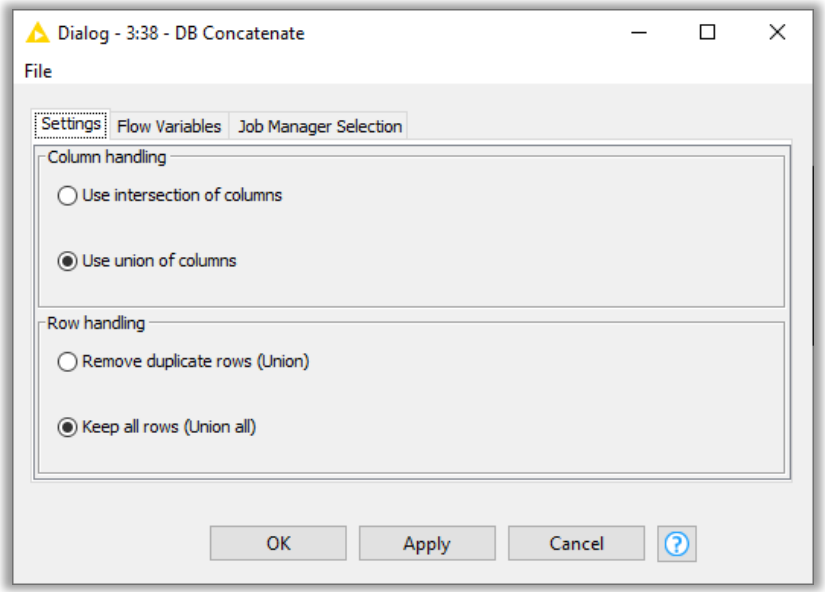
Add and remove node ports based on your needs, e.g. in order to concatenate three or more tables



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	2091-10-13	#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	2091-10-13	#29999	1993-01-13	Berlin	M

Left Outer Join

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

Right Outer Join

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

Joining Columns of Data

Left Table

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

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

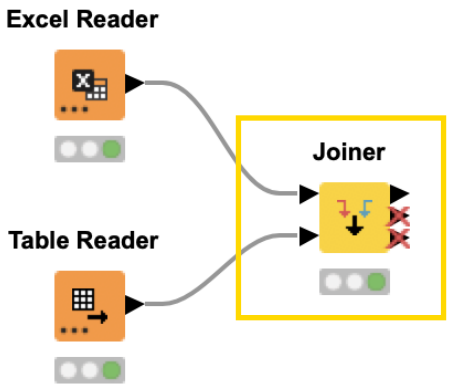
CustomerKey	OrderDate	OrderID	DoB	City	Gender
17	?	?	1974-02-23	Berlin	F
65	?	?	2001-05-25	Stuttgart	F
35	?	?	1988-08-05	Cologne	M
15	2019-10-07	#28985	1983-07-20	Hamburg	M
10	2091-10-13	#29999	1993-01-13	Berlin	M
22	2019-09-23	#23444	?	?	?
24	2019-09-30	#23457	?	?	?

Missing values in the left table

Missing values in the right table

Joiner

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



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

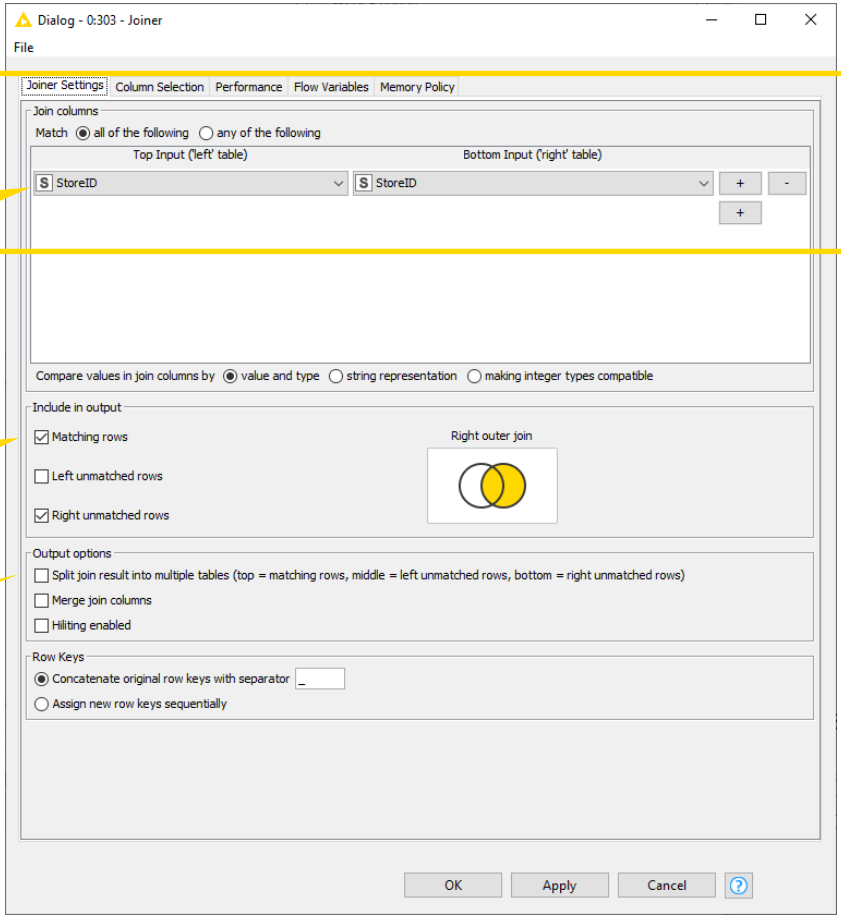
- By default, the two bottom output ports are deactivated

Joiner Configuration – Linking Rows

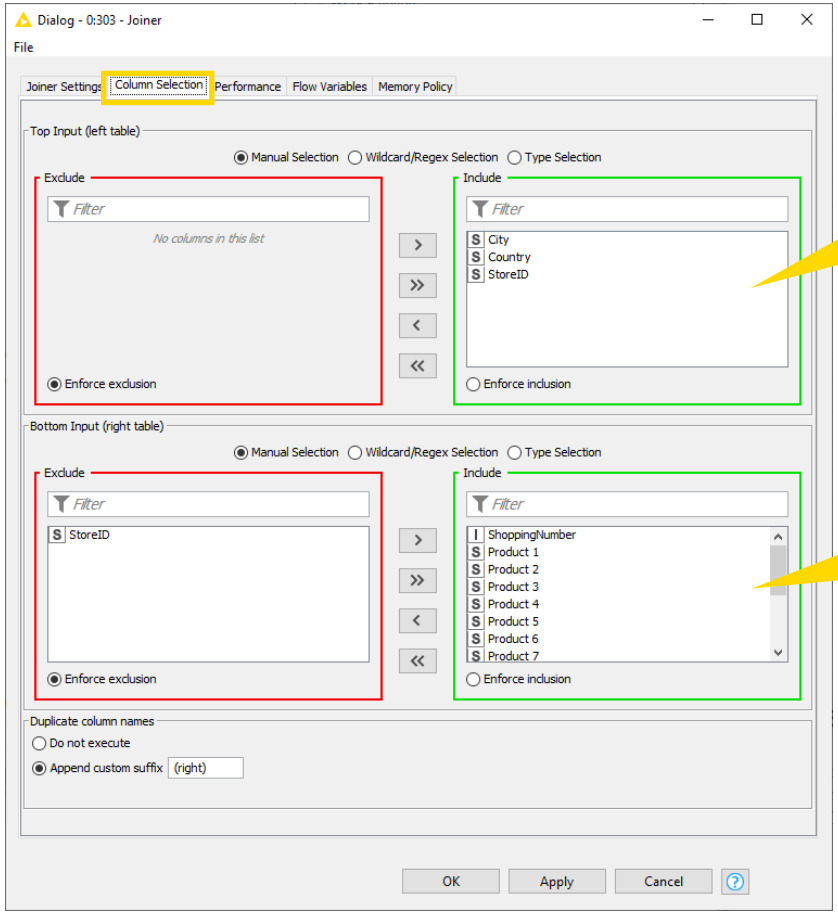
Values to join on. Multiple joining columns are allowed

Select the rows which should be included in the joined table

Activate this checkbox to activate the bottom output ports



Joiner Configuration – Column Selection

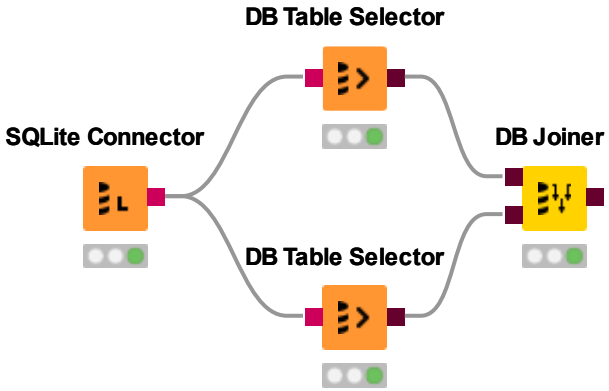


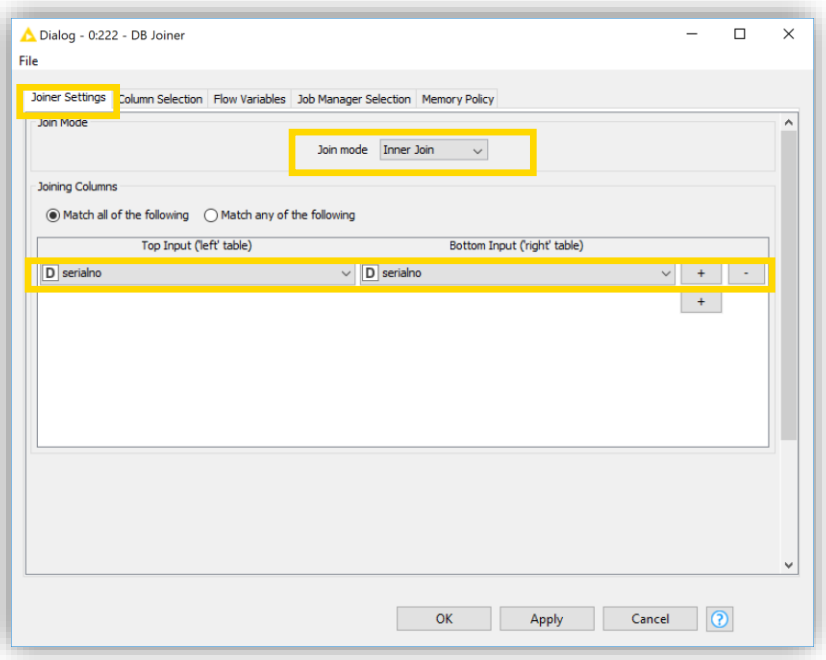
Columns from top table for joined table

Columns from lower table for joined table

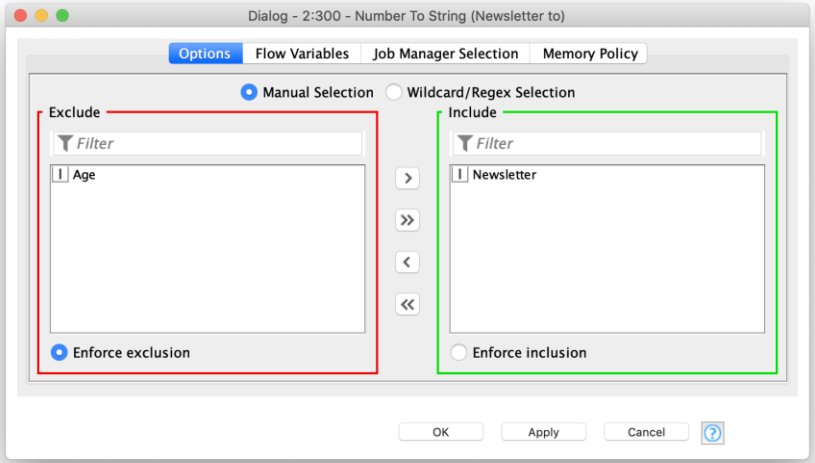
DB Joiner

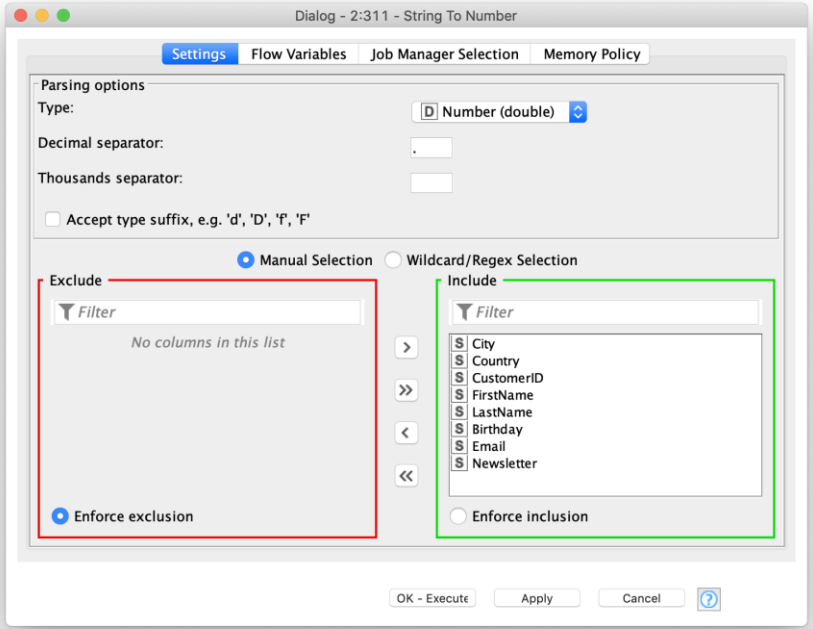
- In-database joiner
- Creates the SQL statement to join two tables stored in the same database
- No coding required



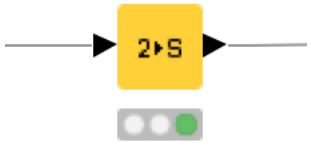


Type Conversion

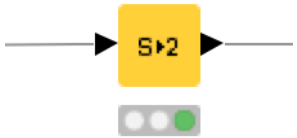




Number To String

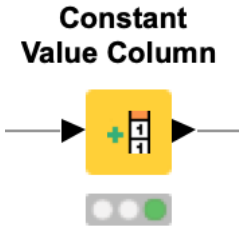


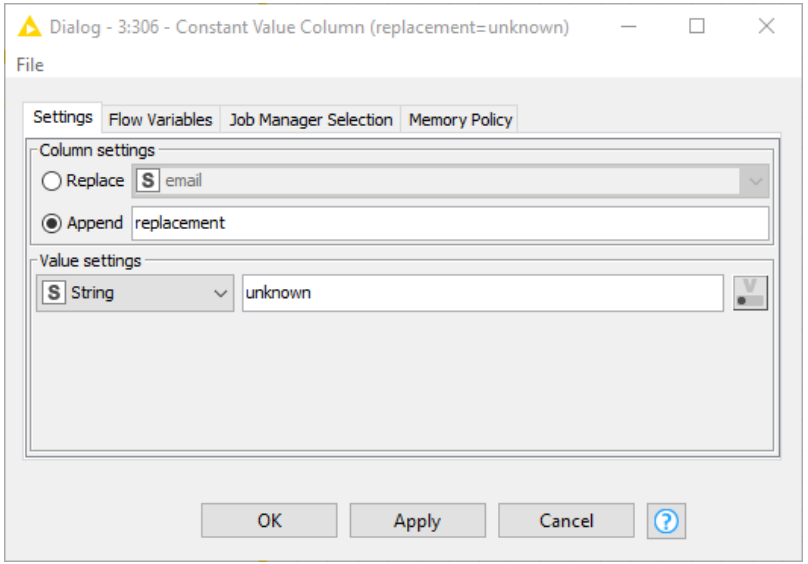
String To Number



Constant Value Column

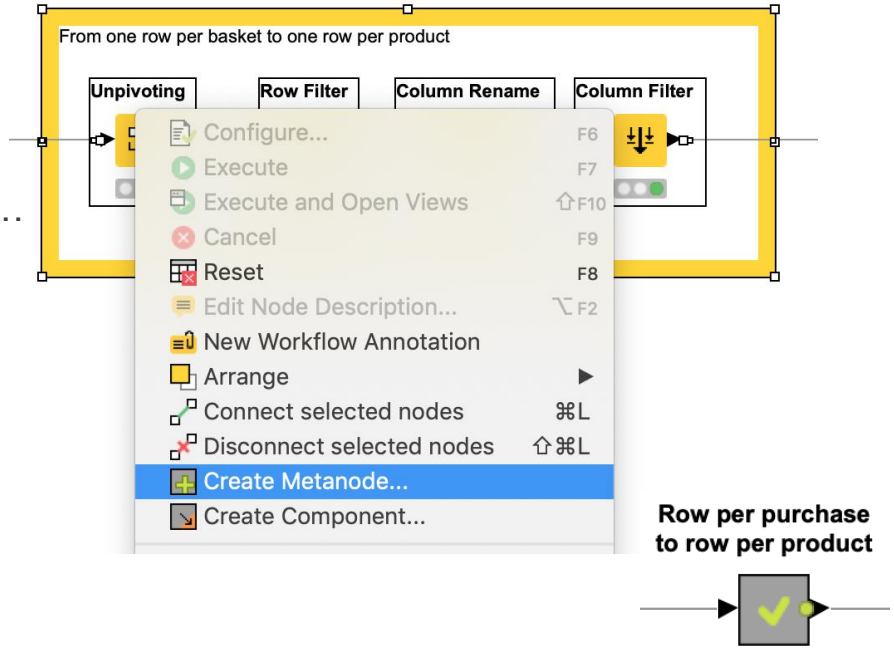
- Adds or replaces a column with a single constant value
- Can be used to add an empty column





Workflow Organization – Good Practices

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



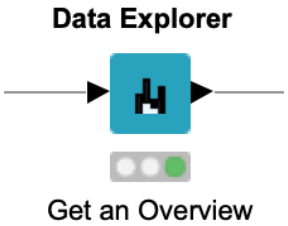
Exercise: 02_Data Merging

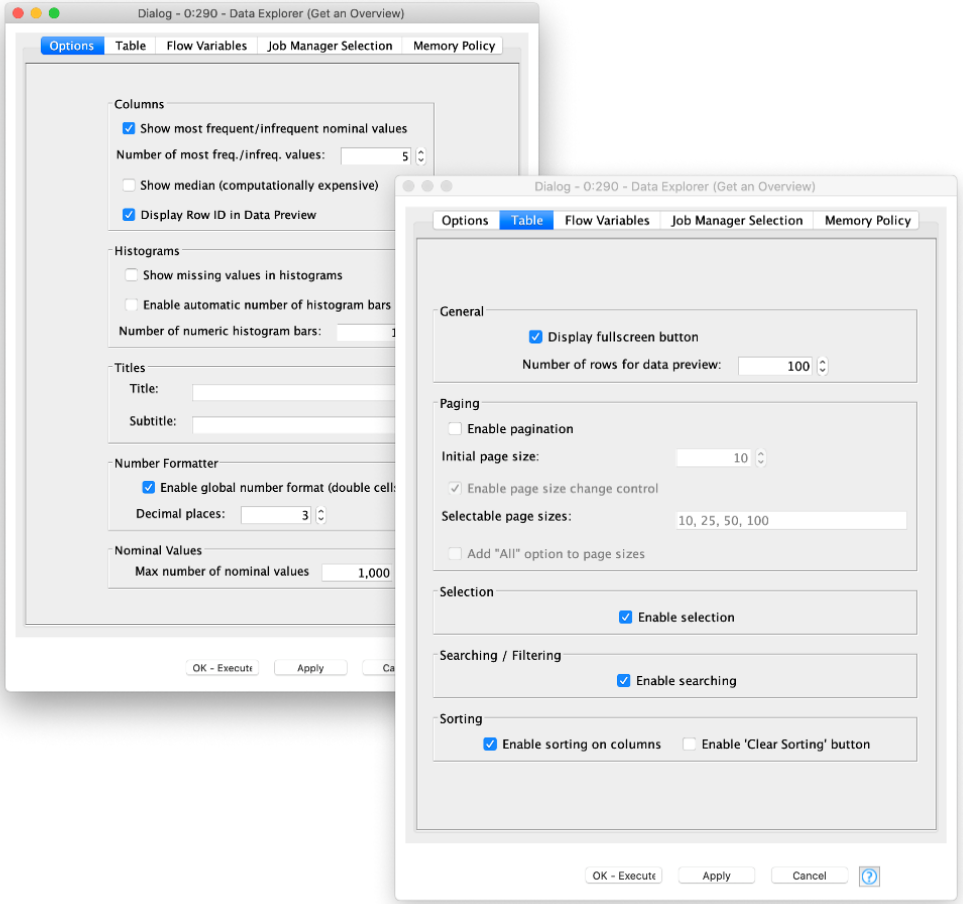
- **Concatenate** the customer information from the two systems
- Perform lookup operation to manipulate selected values in the data:
  - Read a table with a few emails
  - Add a constant value column to the table, with the value “unknown”
  - Replace the provided email addresses in the customer information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

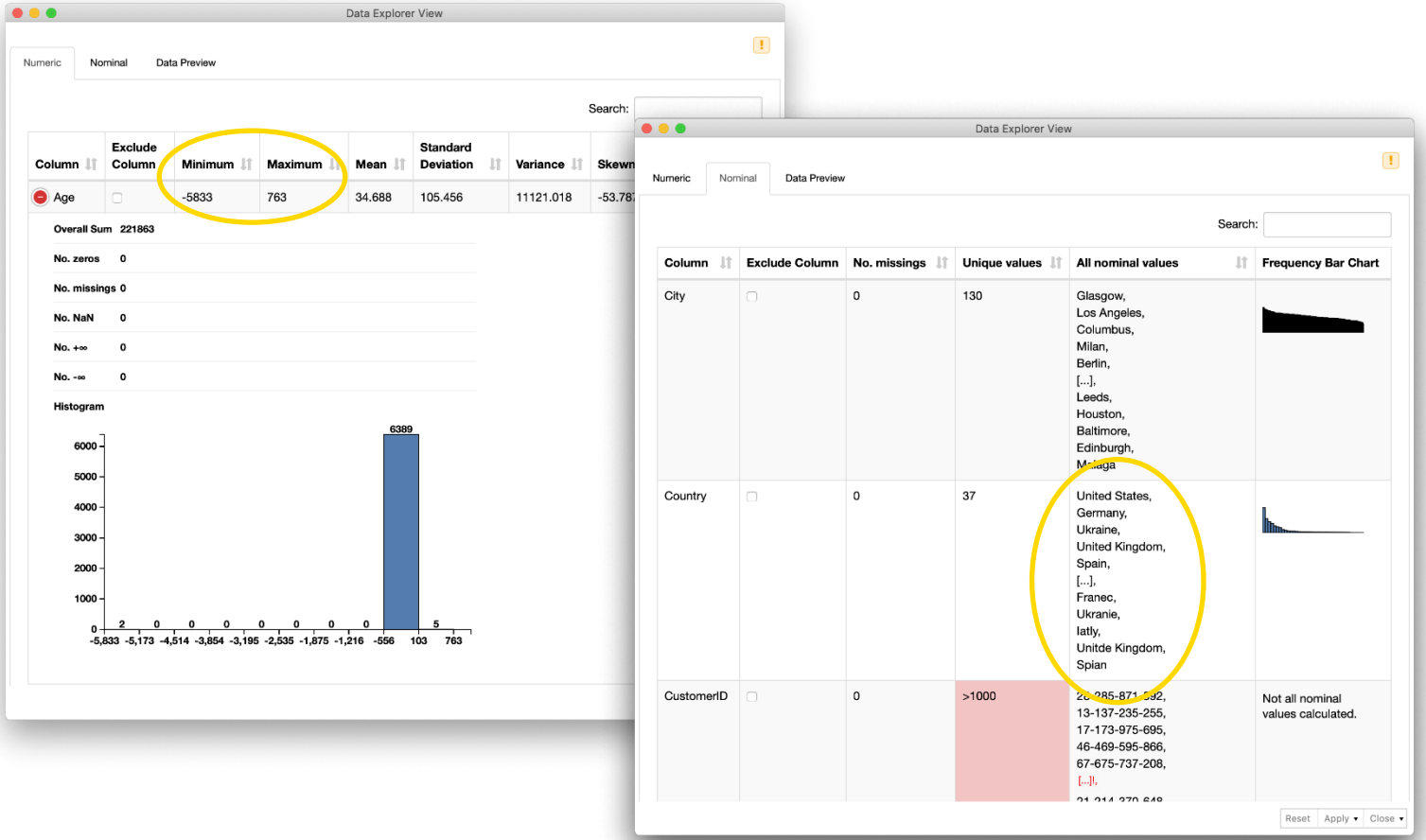
Data Explorer

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





Customer Data Output



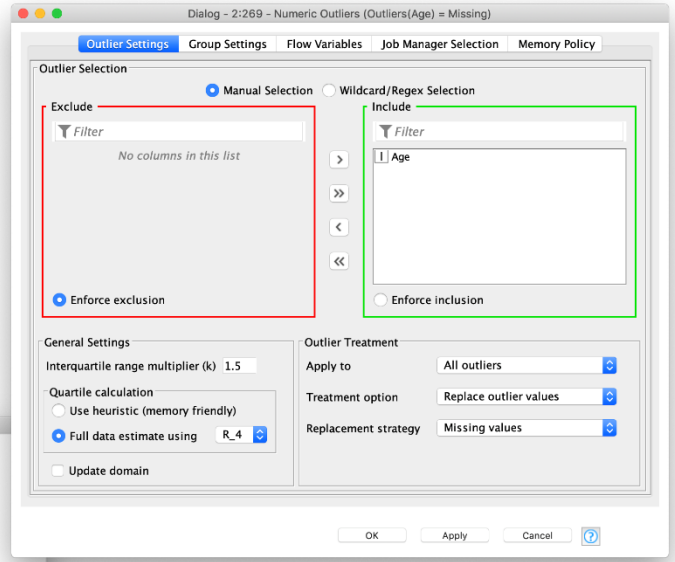
Numeric Outlier

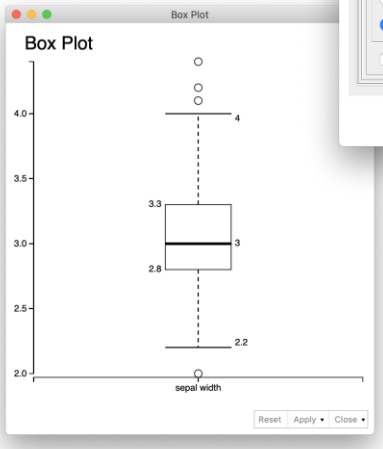
- Detects and treats outliers
- x is a numeric outlier if

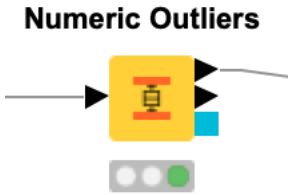
  $x < Q_1 - k * IQR$

  $x > Q_3 + k * IQR$

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





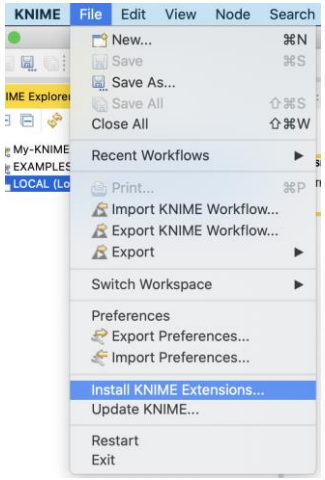


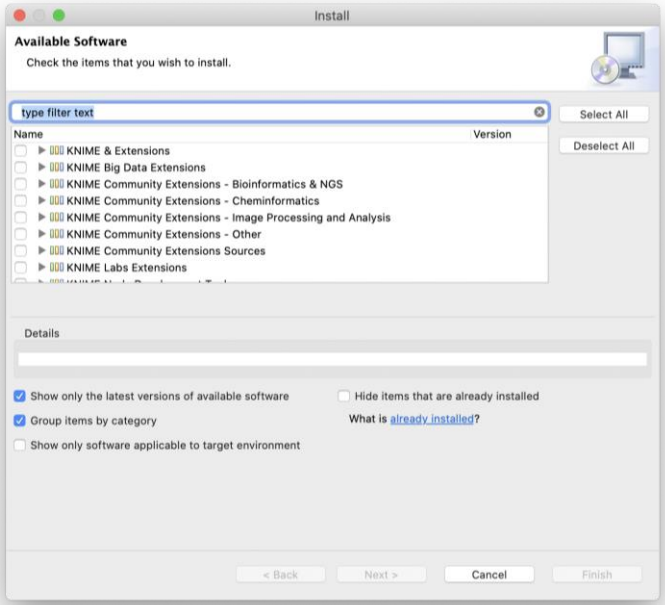
Calculating the Quantiles - Example

- 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×(1/4) = 2.5, rounds up to 3 => Q_1 =7
  - Q_2 : Rank =10×(2/4) = 5, => Q_2 = (8+10)/2=9
  - Q_3 : Rank =10×(3/4) = 7.5, rounds up to 8 => Q_3 =15

More Nodes – Installing Extensions

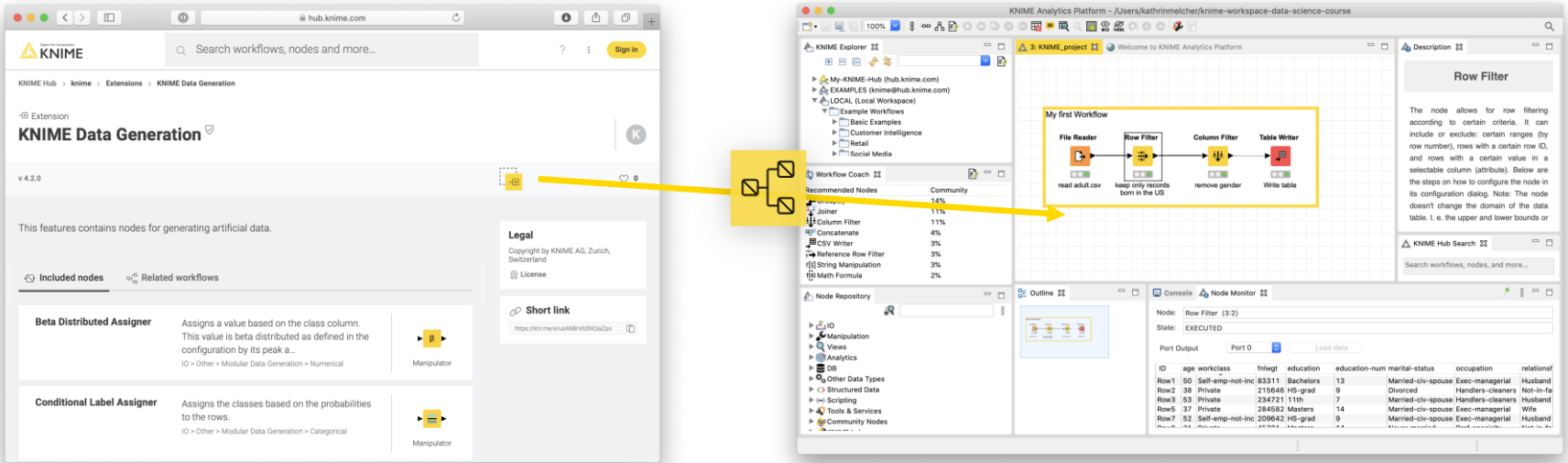
- Go to File -> Install Extensions
- Select the following extensions:
  - KNIME JavaScript (Labs)
    (in the folder KNIME Labs Extensions)
  - KNIME Data Generation
    (in the folder KNIME & Extensions)





Alternative Way via the KNIME Community Hub

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

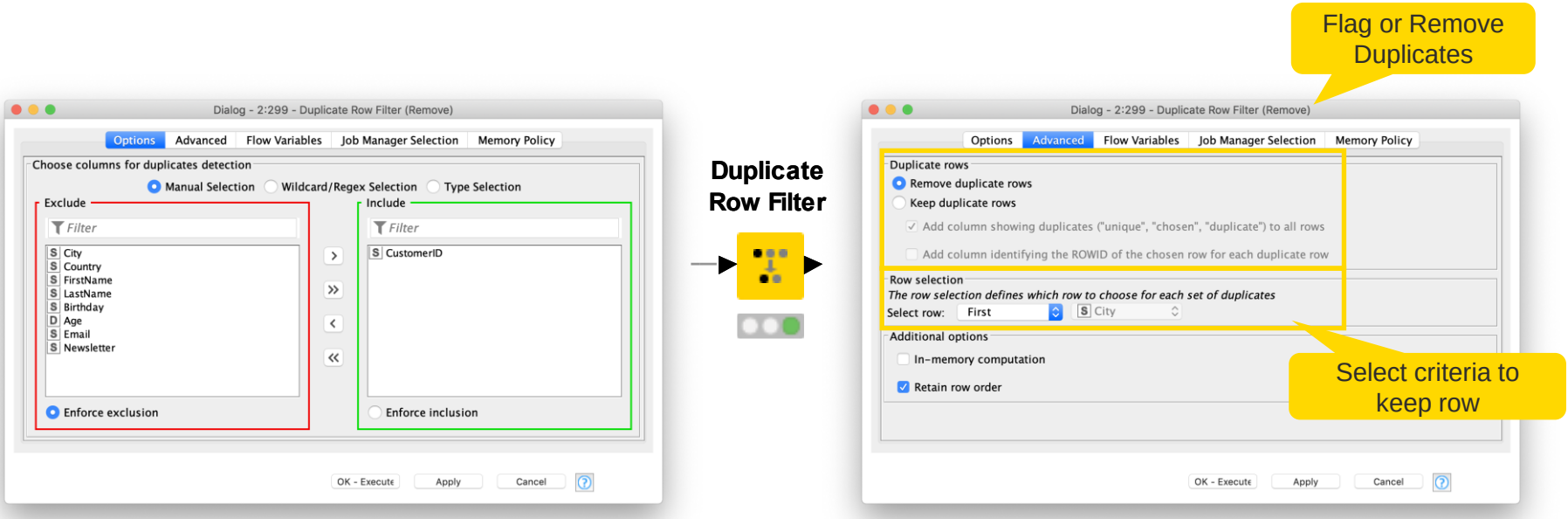


Duplicate and Missing Value Handling

Duplicate Row Filter

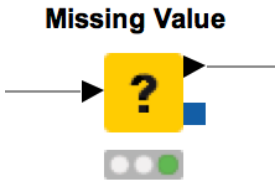
Detects duplicate rows and apply a selected treatment

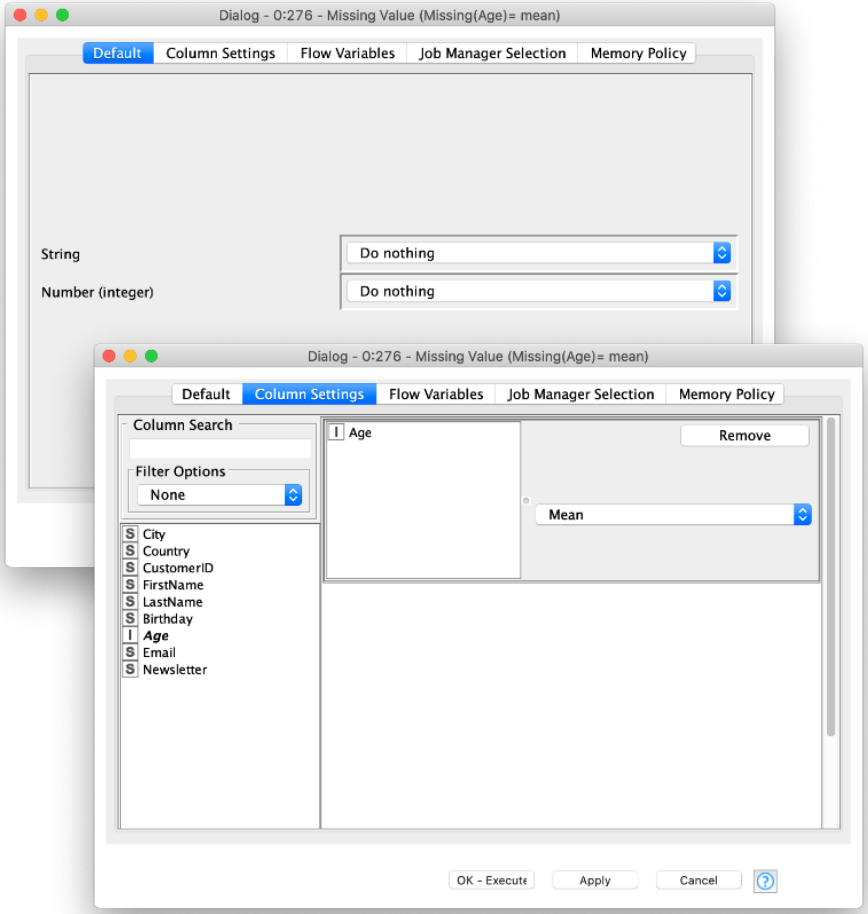
- First tab provides the option to select columns for duplicate detection
- Second tab provides options for treating duplicated values



Missing Value

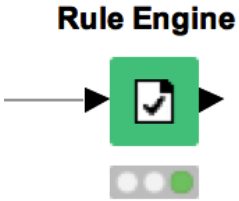
- Defines how to handle missing values for all columns of a given type
  - Affects all columns that are not explicitly mentioned in the second tab
- Defines how to handle missing values for each available column

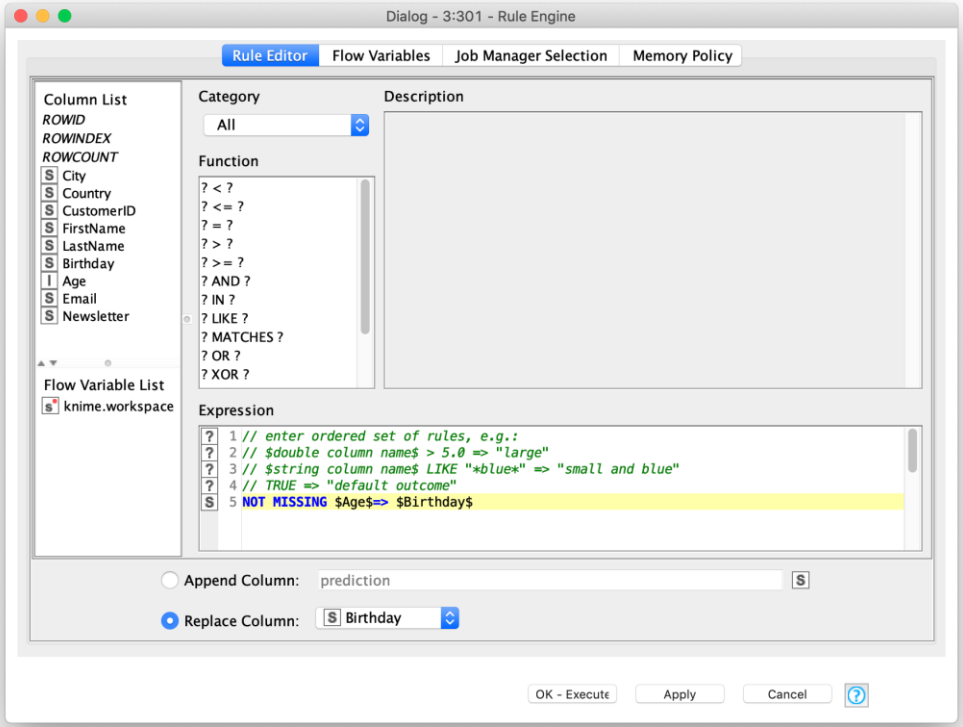




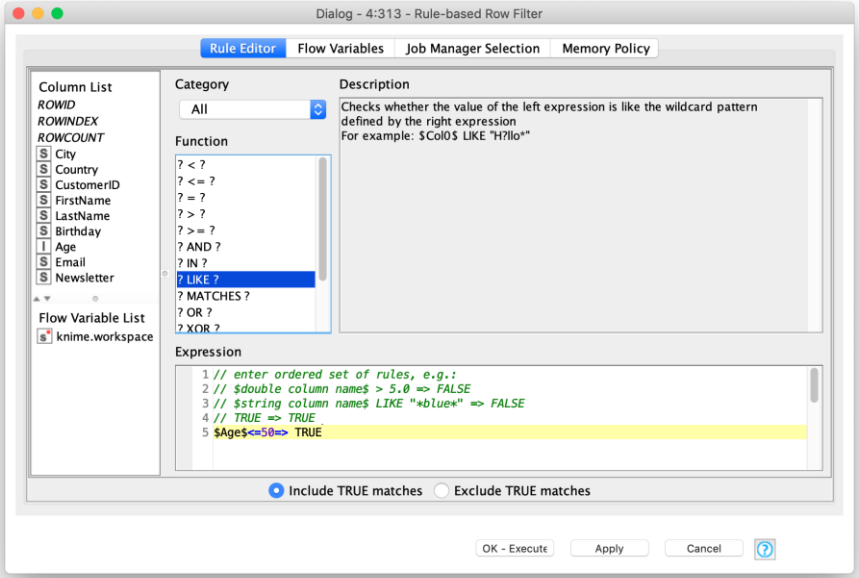
Rule Engine

- 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

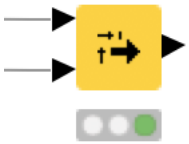




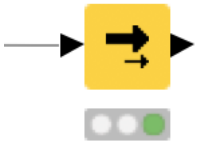
Other Options to Filter Rows



Reference Row Filter

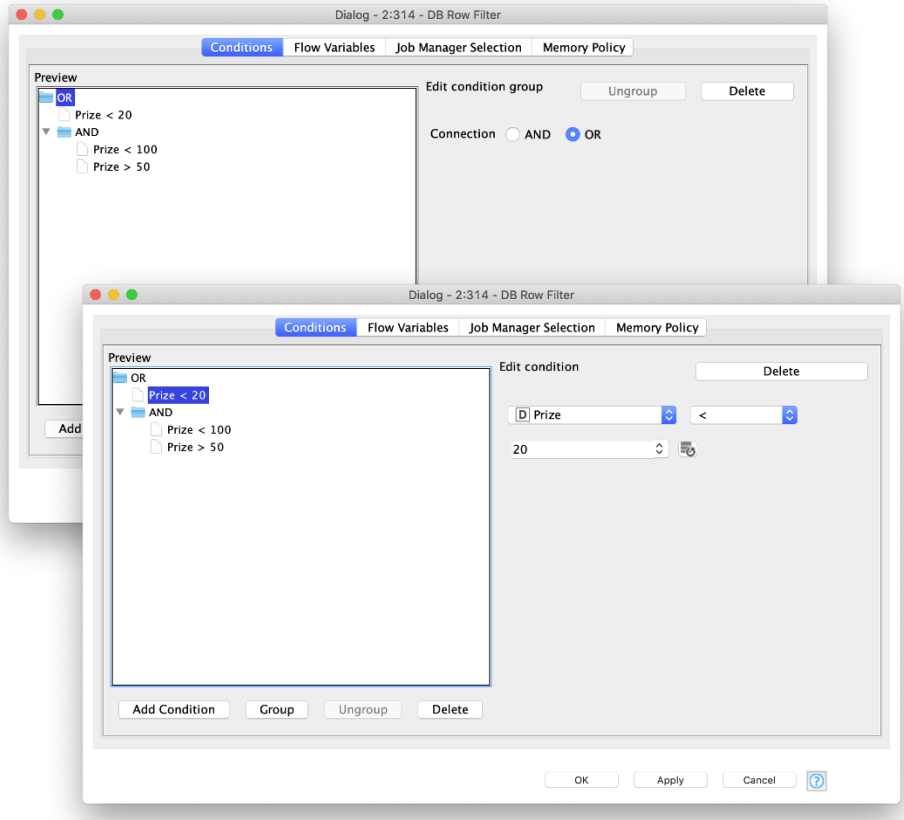


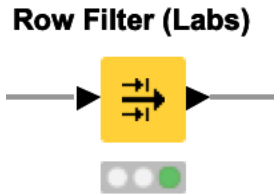
Rule-based Row Filter

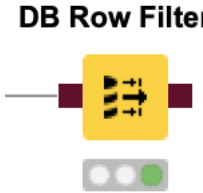


In-Database Row Filtering – DB Row Filter

- 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

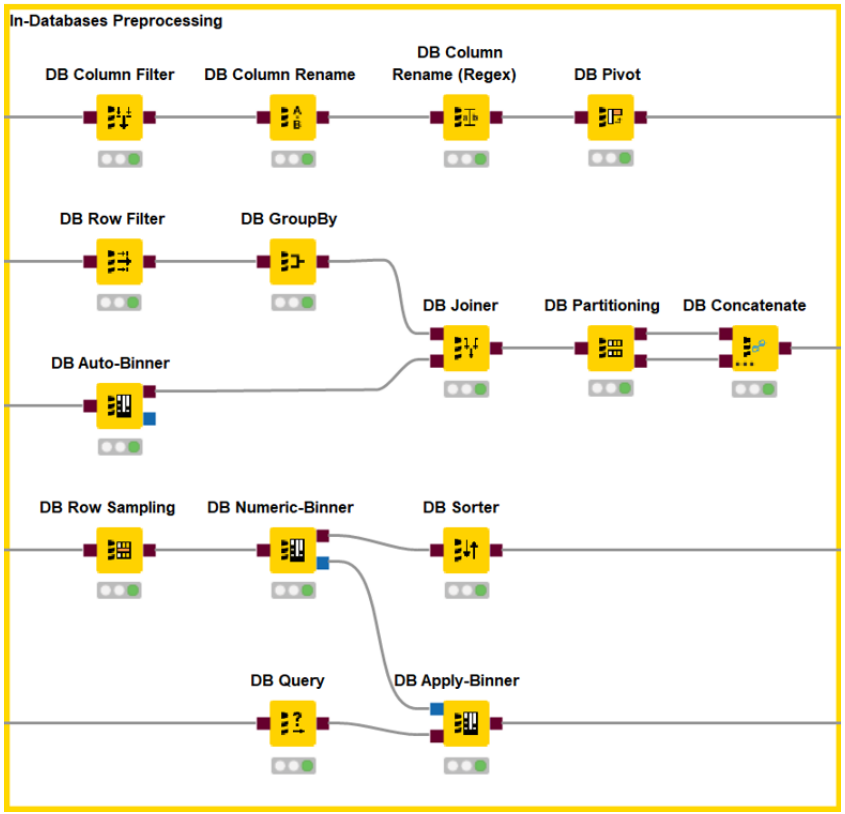






Query Nodes

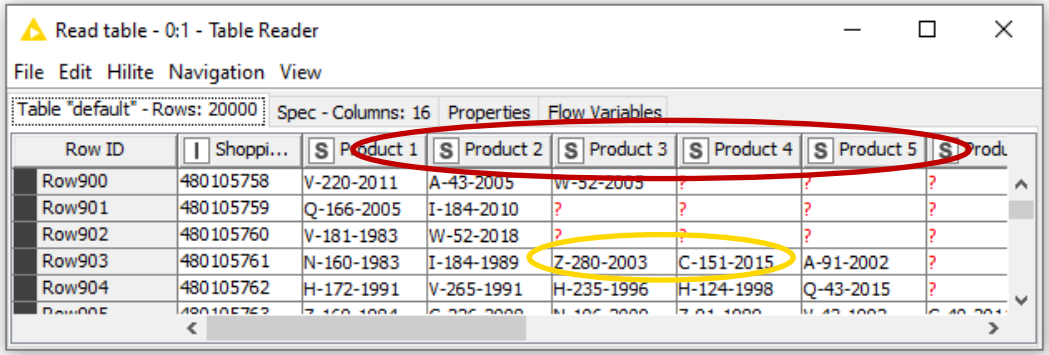
- Filter rows and columns
- Join tables/queries
- Concatenate tables
- Extract samples
- Bin numeric columns
- Sort your data
- Write your own query
- Aggregate your data



Standardization

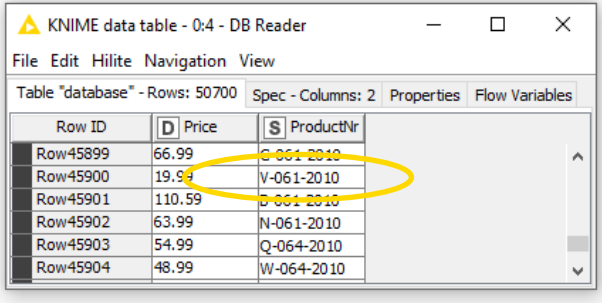
Standardize Table Format

Onsite transaction data



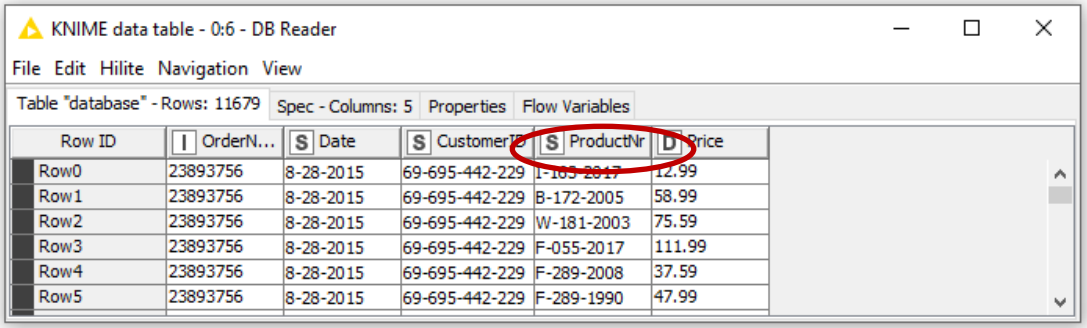
Row ID	Shoppi...	Product 1	Product 2	Product 3	Product 4	Product 5	Produ
Row900	480105758	V-220-2011	A-43-2005	W-52-2005	?	?	?
Row901	480105759	Q-166-2005	I-184-2010	?	?	?	?
Row902	480105760	V-181-1983	W-52-2018	?	?	?	?
Row903	480105761	N-160-1983	I-184-1989	Z-280-2003	C-151-2015	A-91-2002	?
Row904	480105762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

Product Nr & Price



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

Online transaction data



Row ID	OrderN...	Date	CustomerID	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
234	Berlin
235	London
236	Boston
237	Paris

Product 1	Product 2
Pear	Apple
Nuts	Pear
Rice	Grapes
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

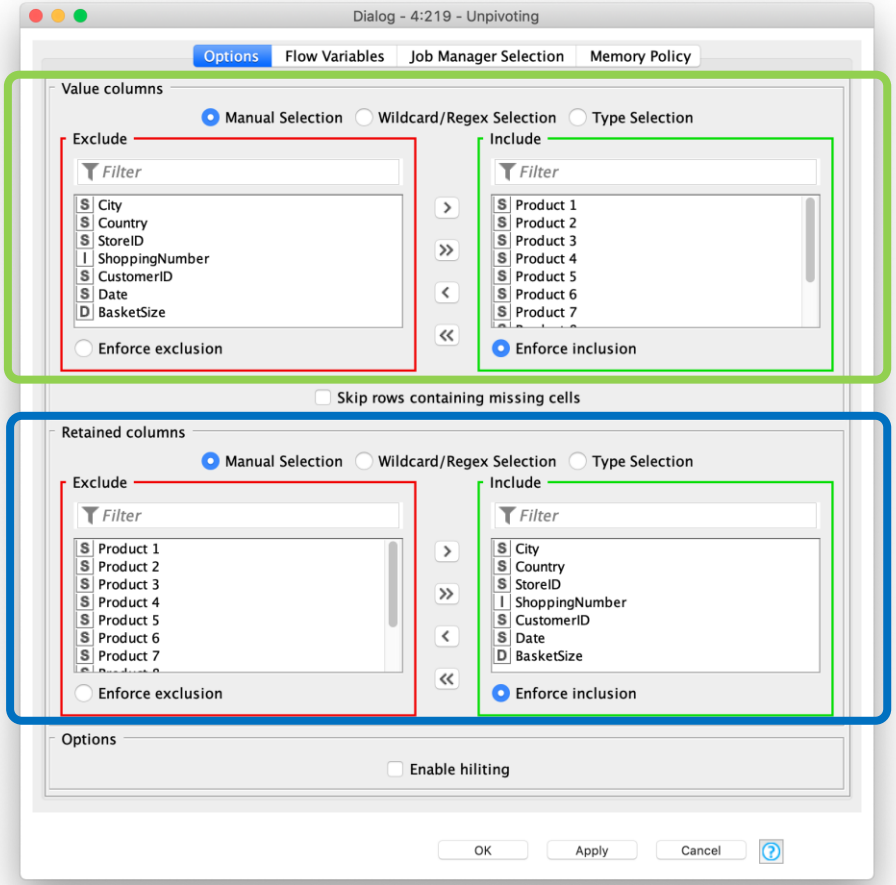
Value Column

Retaining Columns

Solution: Unpivoting Node

Unpivoting

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

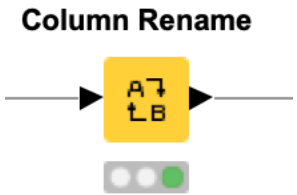


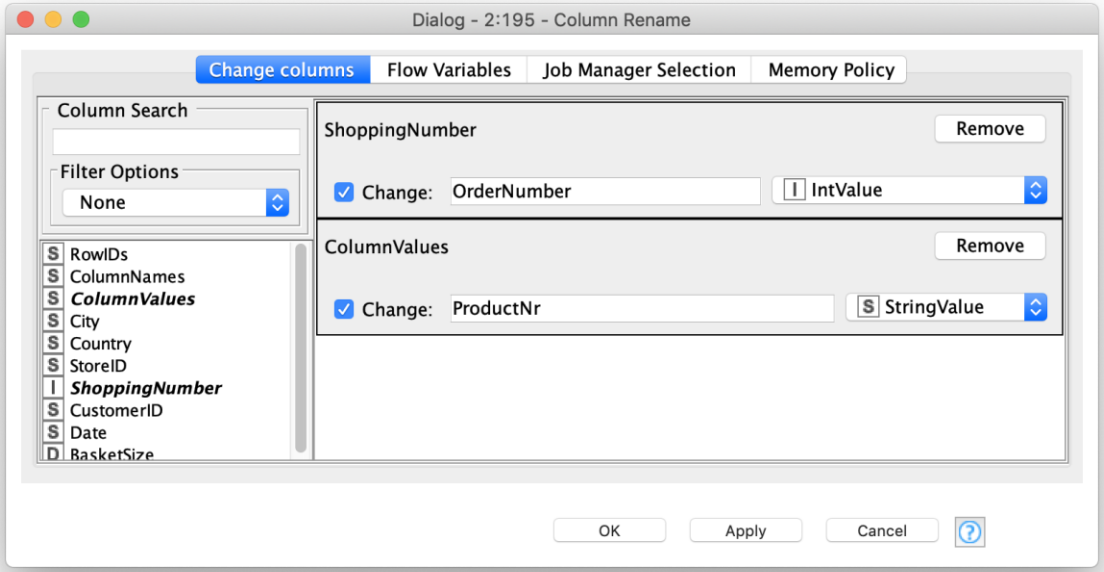
Value Column

Retaining Columns

Column Rename

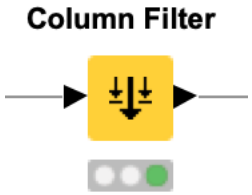
- Renames column names or changes their types





Column Filter

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



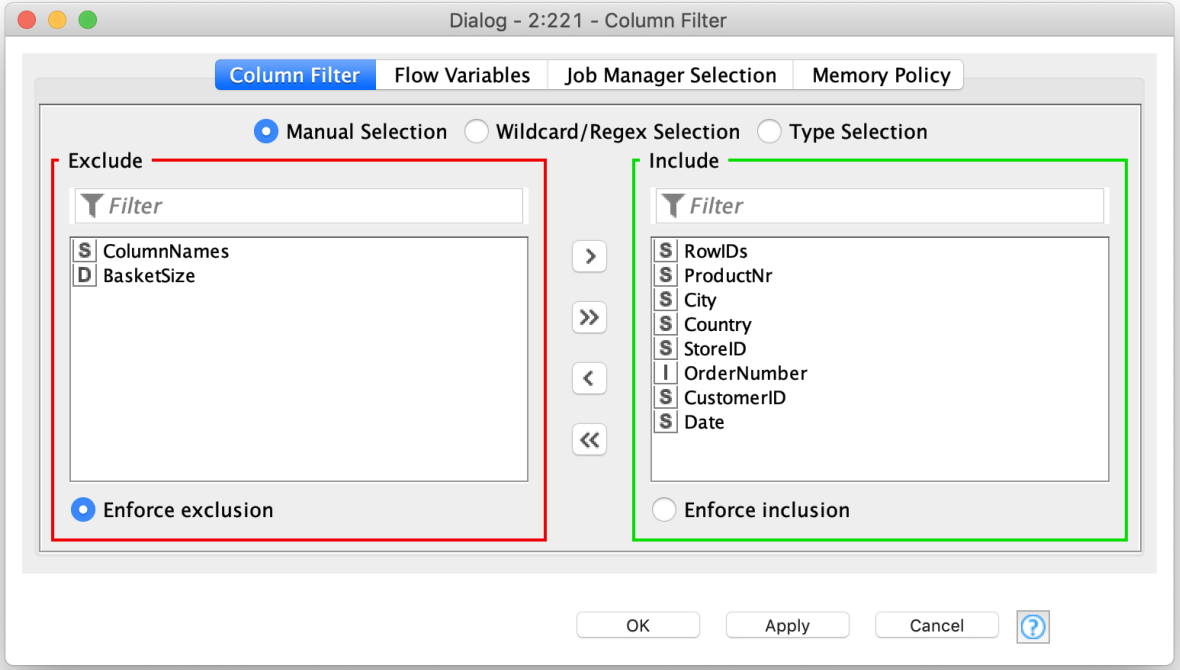
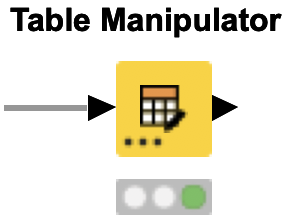
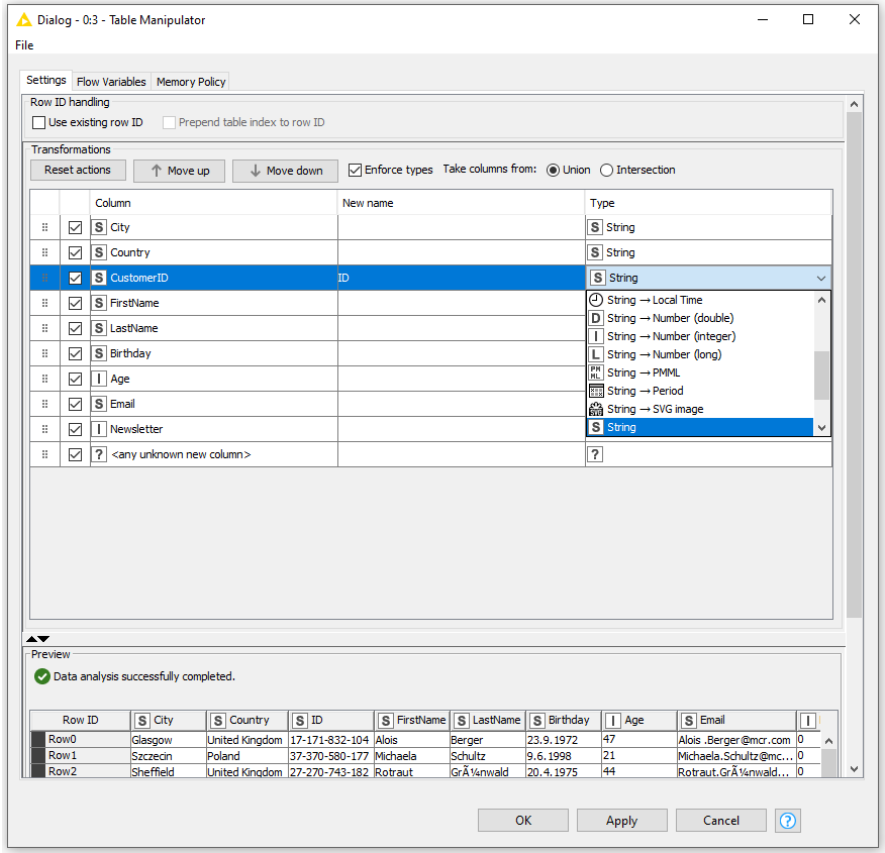


Table Manipulator

Allows for

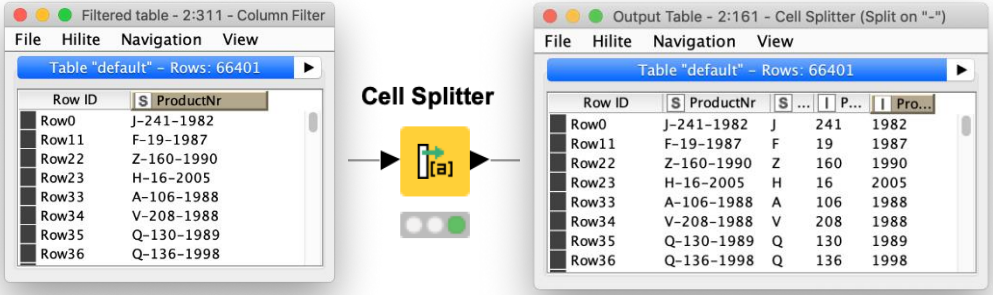
- Concatenation of multiple files/tables
- Column filtering
- Column sorting
- Column renaming
- Column type mapping

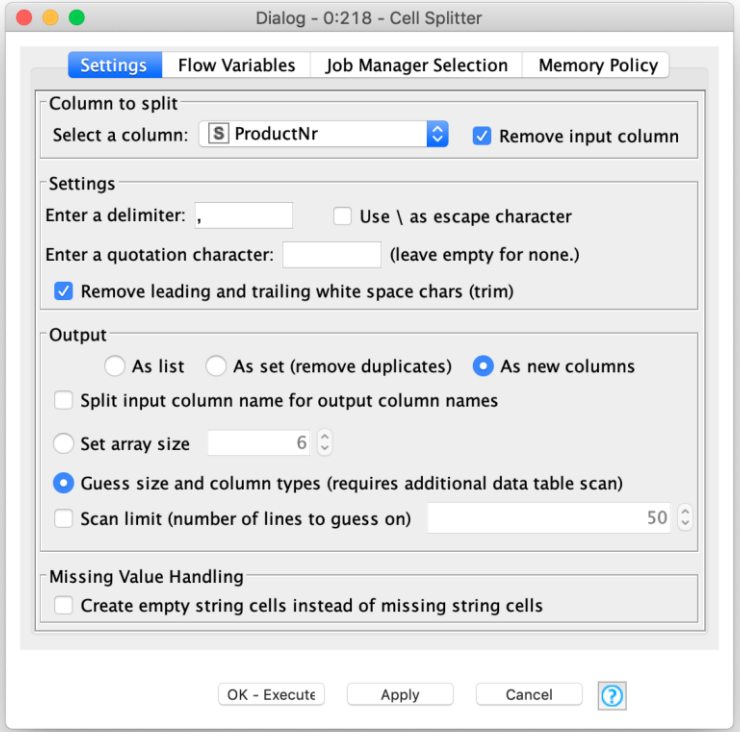




Cell Splitter

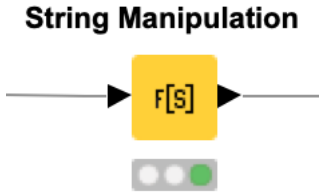
Splits the content of one column into many columns based on a delimiter

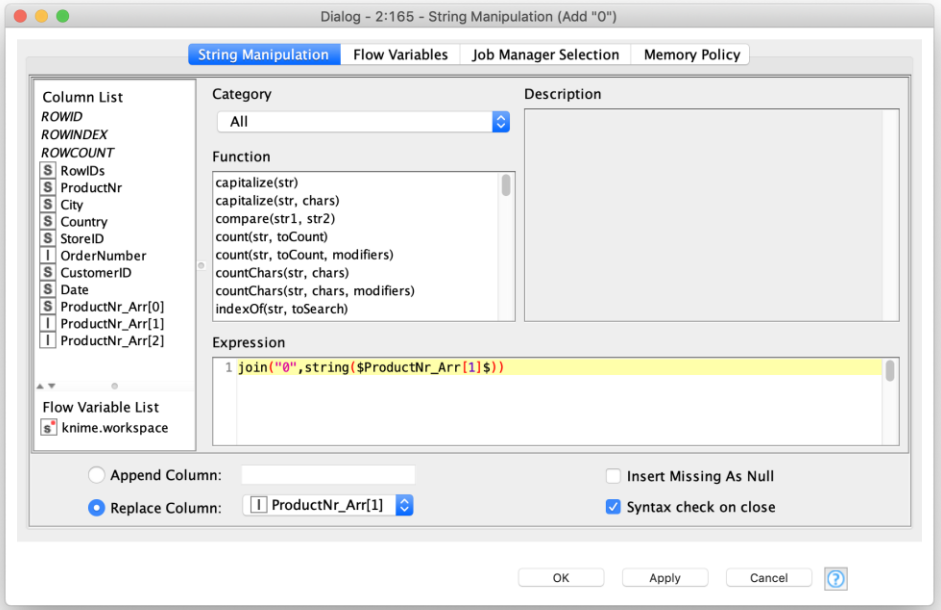




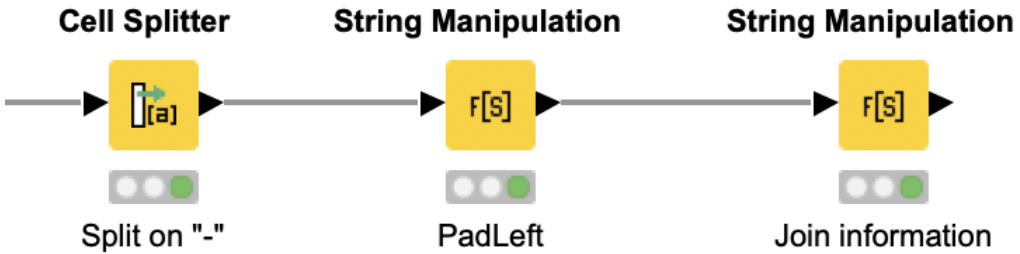
String Manipulation

- 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

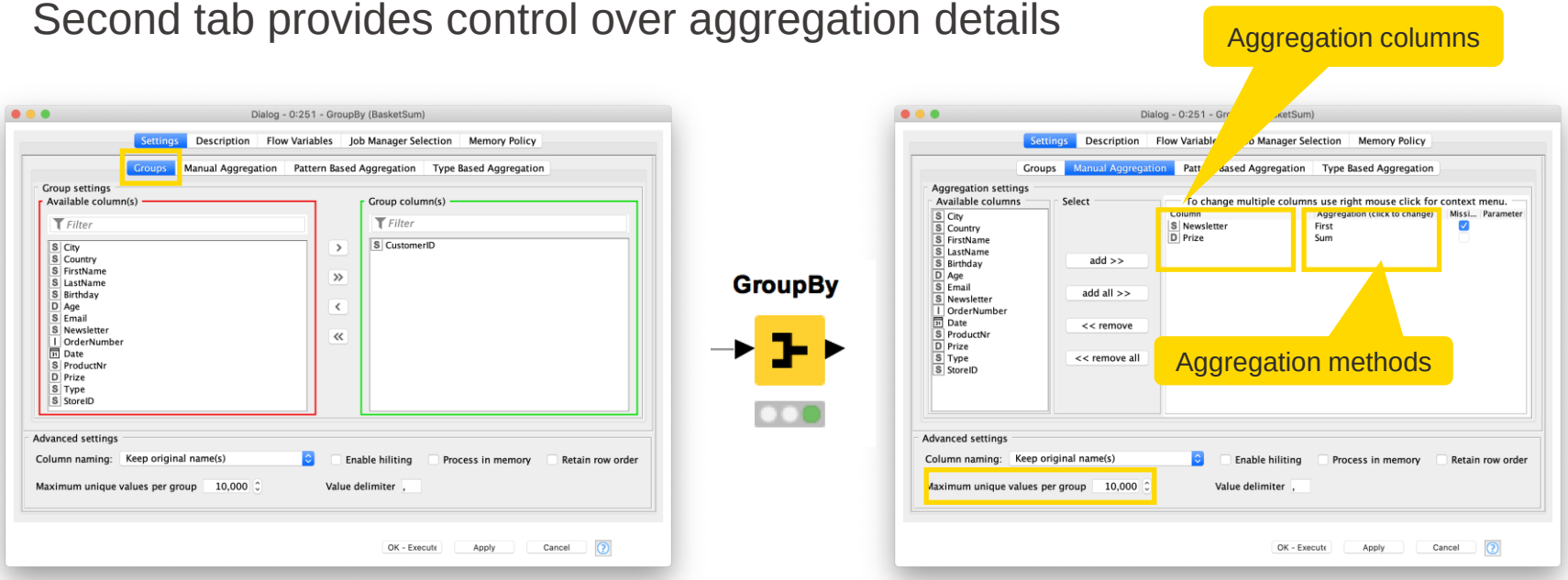
RowID	Group	Sum(Value)
r1+r3+r6	m	8
r2+r4+r5	f	15

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

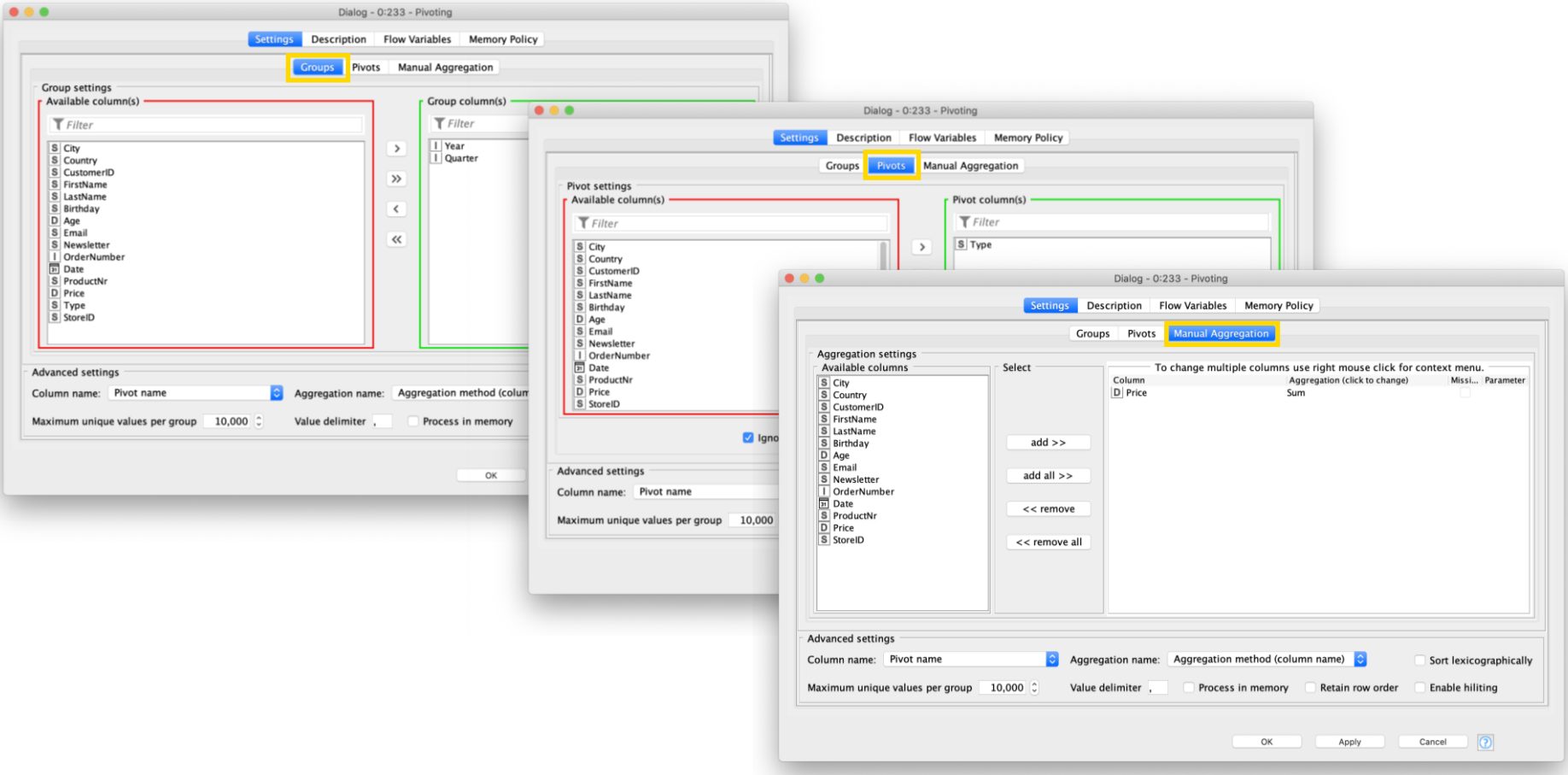
Gender	Hair	Age
f	blond	31
m	red	22
f	blond	53
m	brown	16
f	brown	47
f	black	22
m	blond	13
m	red	55

Aggregation: Mean(Age)

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

Pivoting Node: Group - Pivot - Aggregate

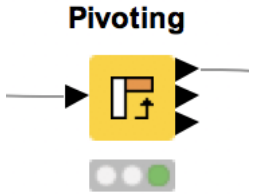
Pivoting



Pivoting

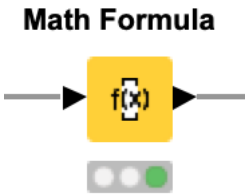
Performs pivoting on selected columns for grouping and pivoting

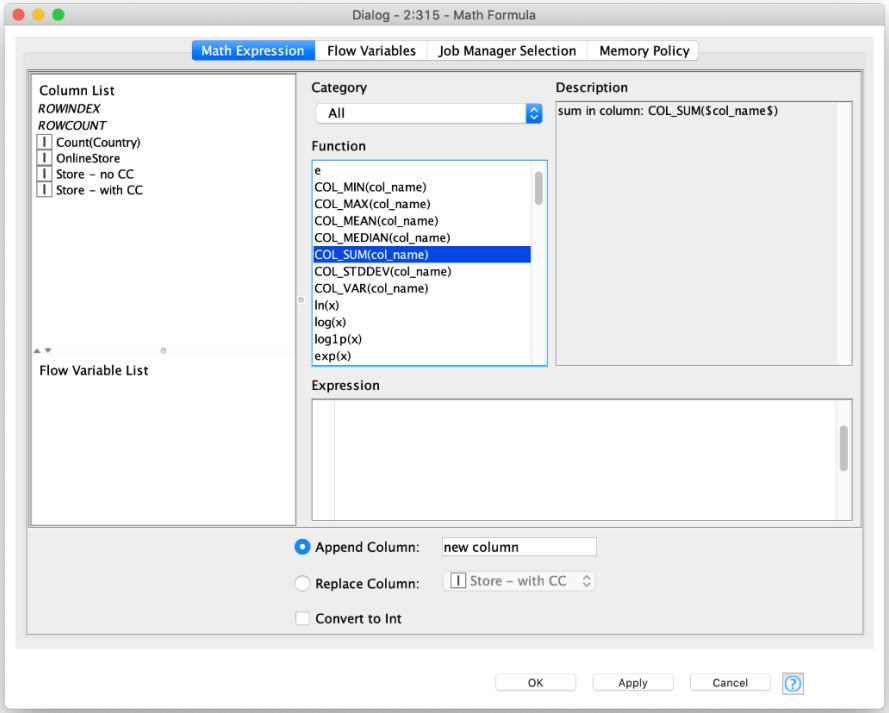
- Values of group columns become unique rows
- Values of the pivot columns become unique columns for each set of column combinations together with each aggregation
- Many aggregation methods are provided



Math Formula

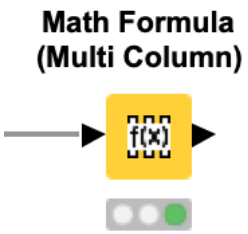
- Row-wise calculations
- Some col-wise statistics
- Many mathematical functions
- Double-click function, then select col by click

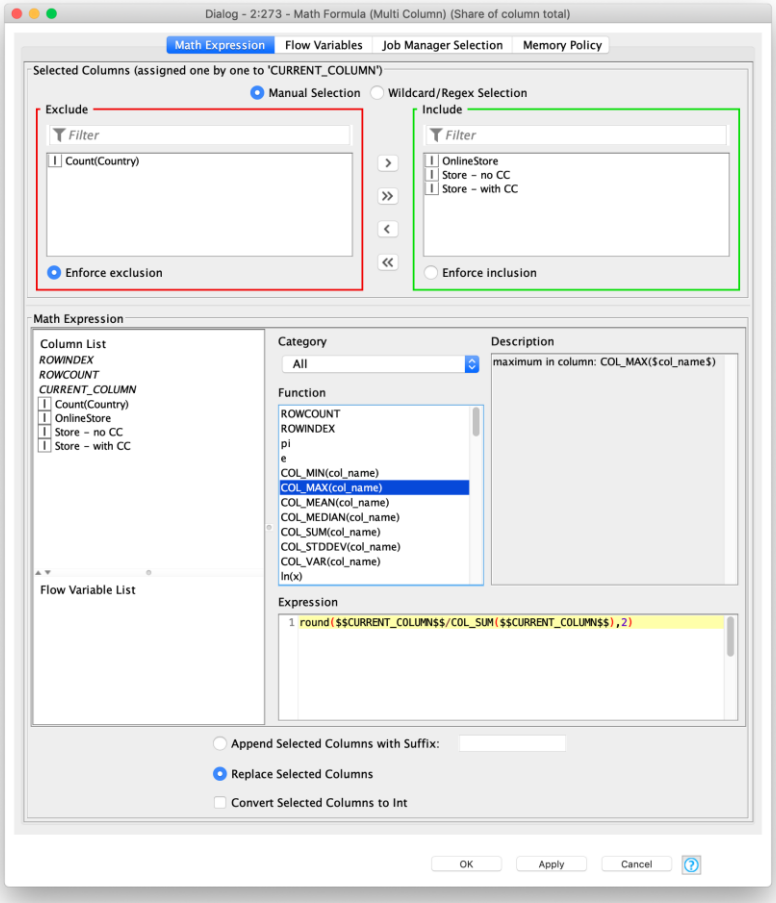




Math Formula (Multi Column)

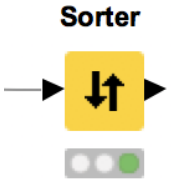
- Useful if you want to make the same calculations on multiple columns.
- The selected columns from the upper part are called CURRENT_COLUMN in the Column List and Expression dialog.

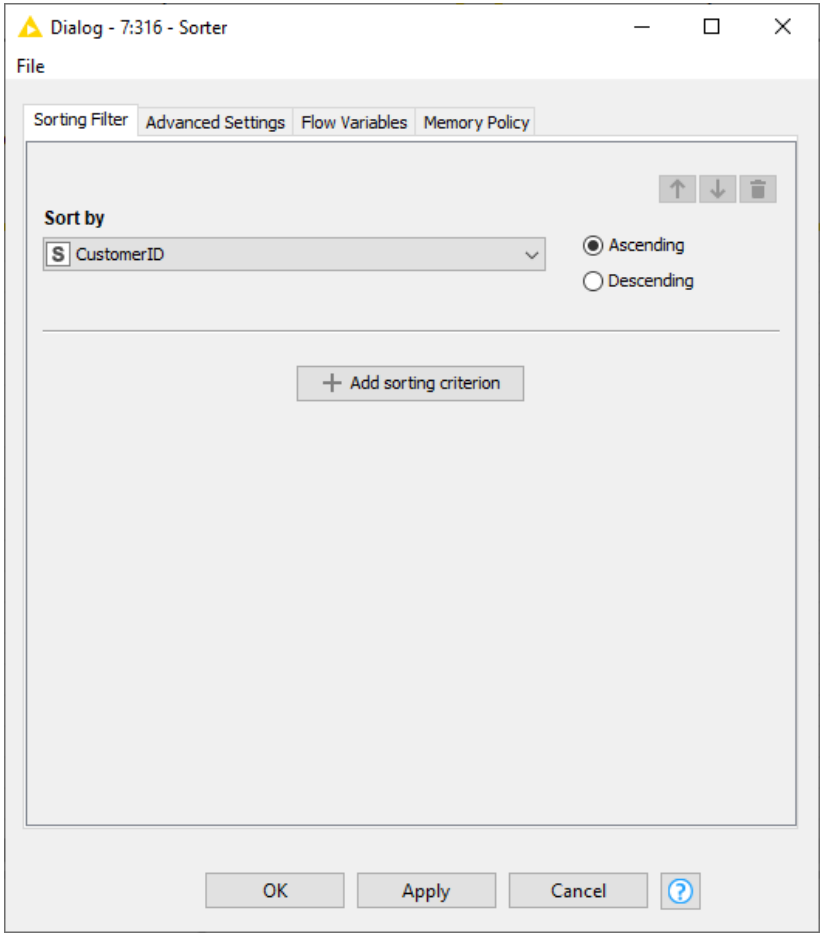




Sorter

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





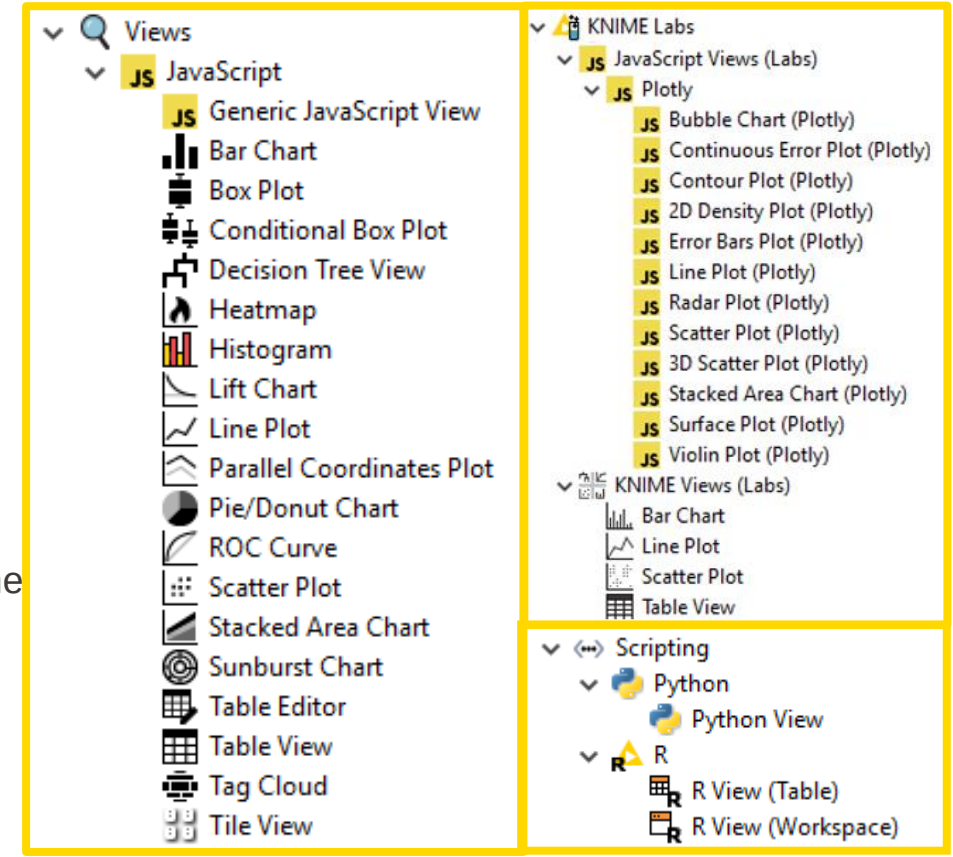
Exercise: 06_Data_Aggregation

- 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)

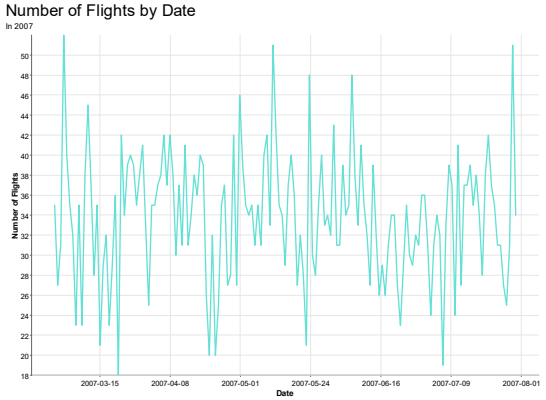
Data Visualization

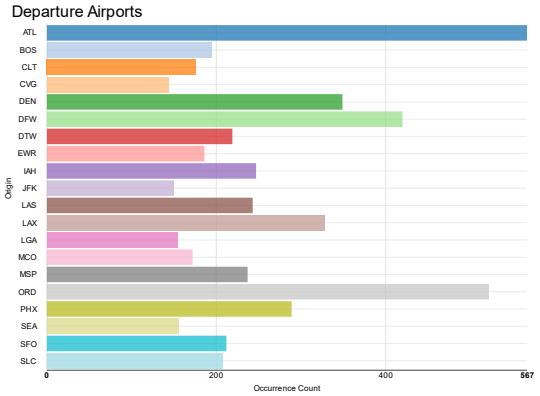
Data Visualization

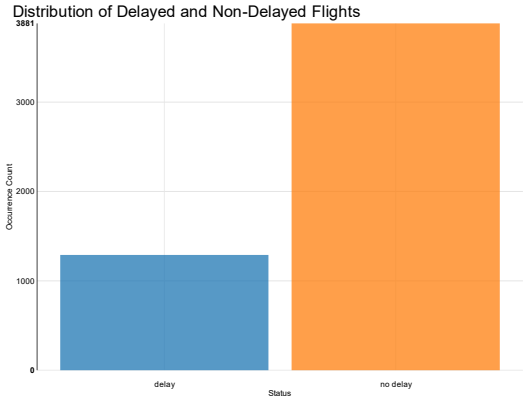
- Large selection of easy to use visualization nodes
  - Web-based and interactive
  - Dedicated nodes, no scripting required
- Plotly nodes
  - Similar but integrated from an external library
- New Visualization Nodes in Labs
  - A live preview of the visualization next to the configuration dialog
-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

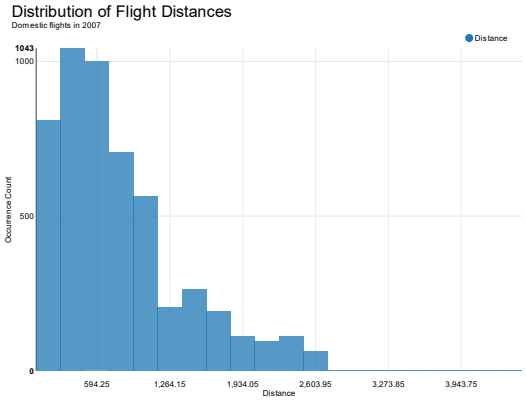


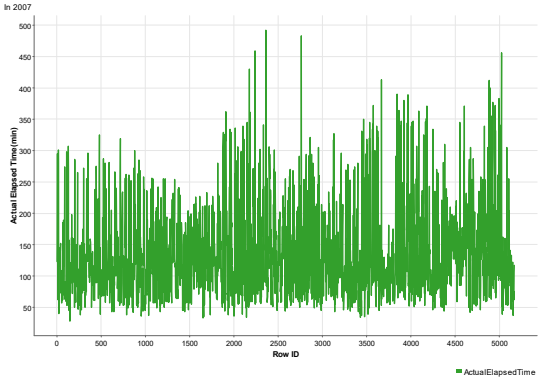
Visualizations Using One Column

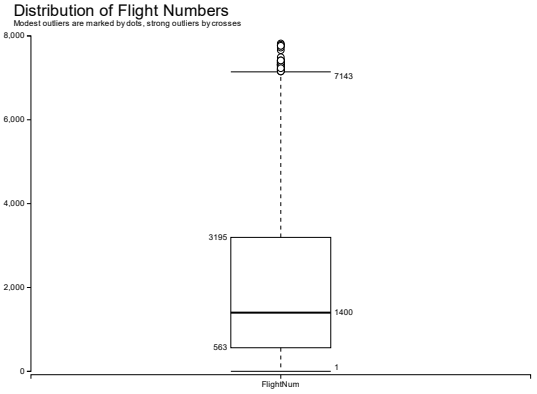




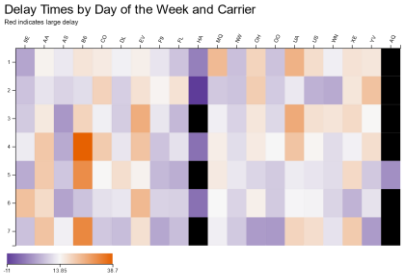


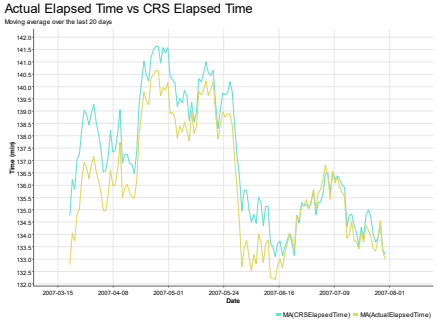


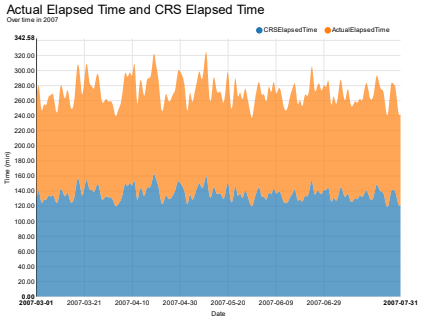


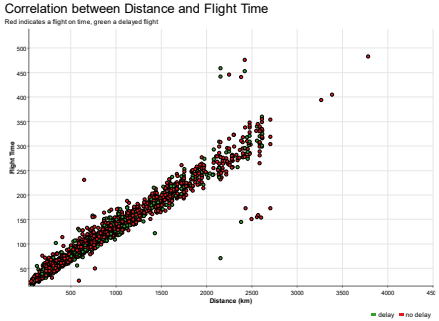


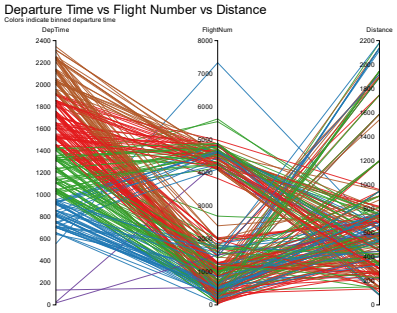
Visualizations Using Three Columns

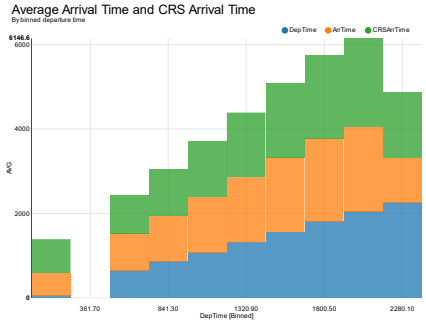


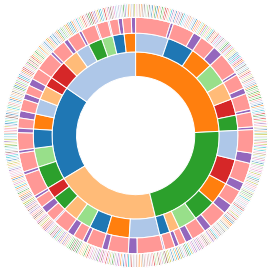


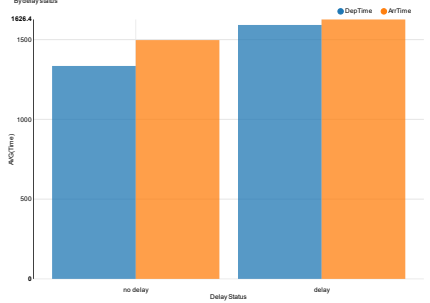






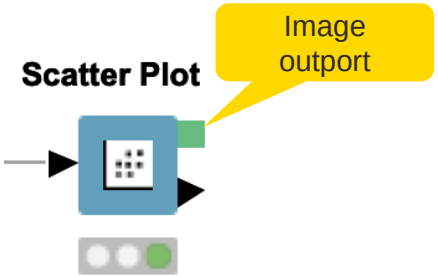


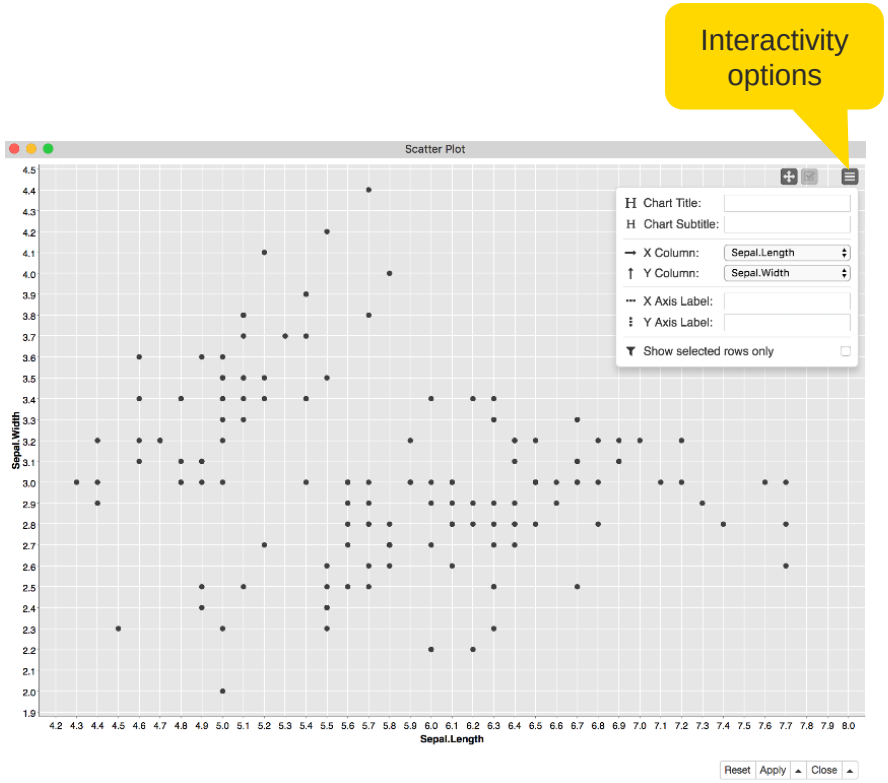




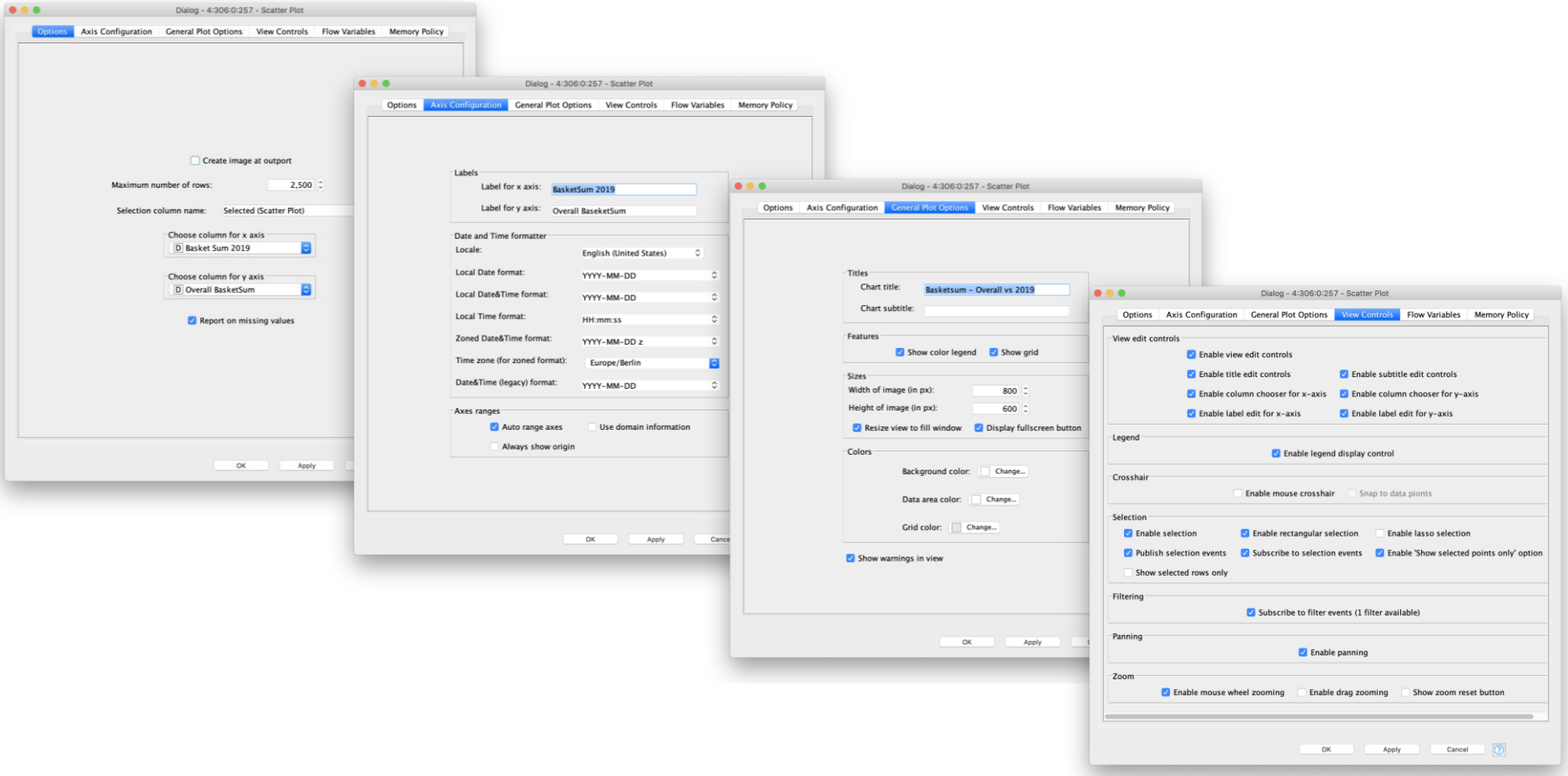
Scatter Plot

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





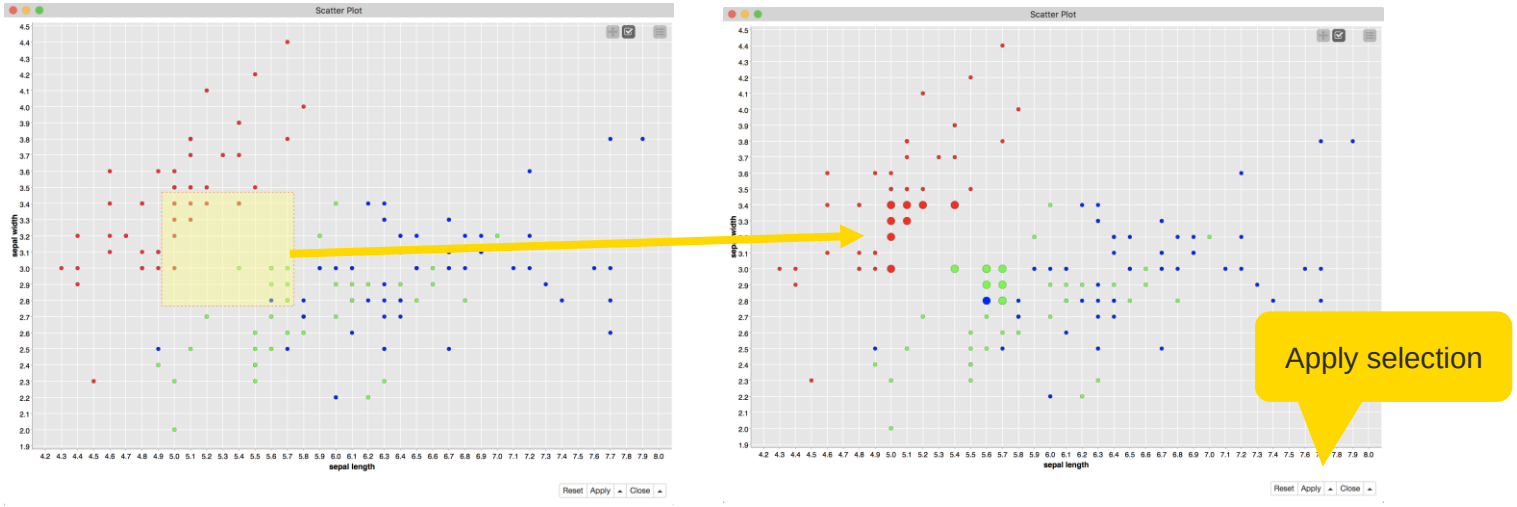
Scatter Plot



Selection and Filtering in JavaScript Views

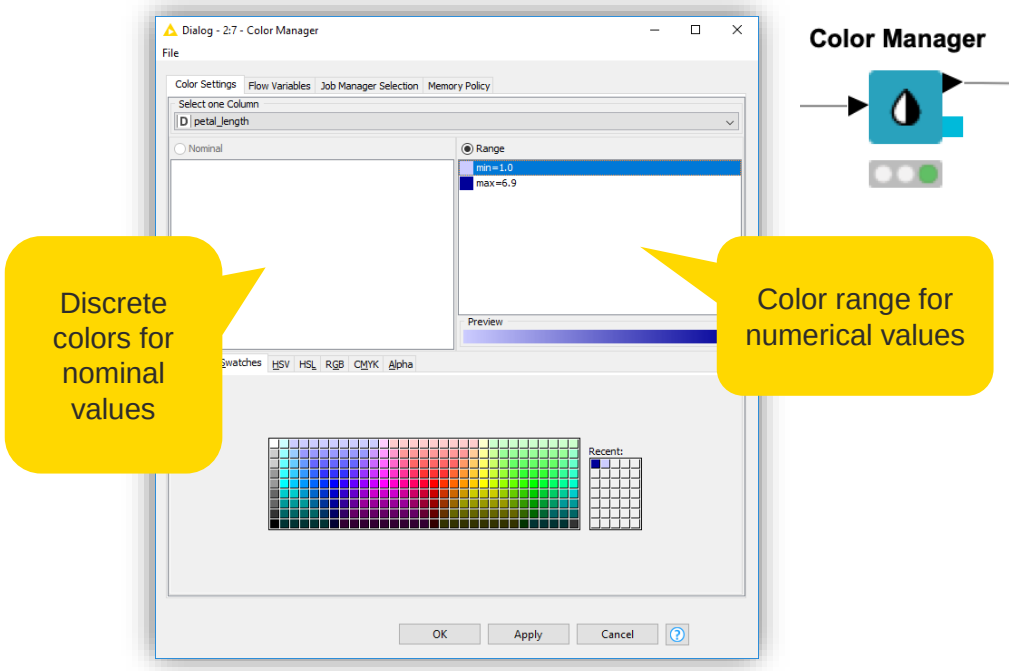
Interactivity allows you to select data points in views

- 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



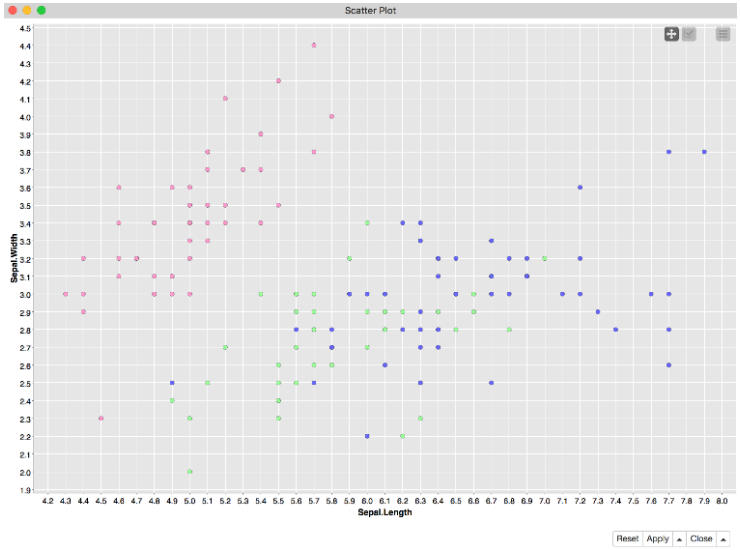
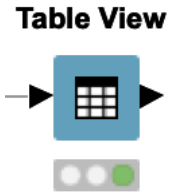
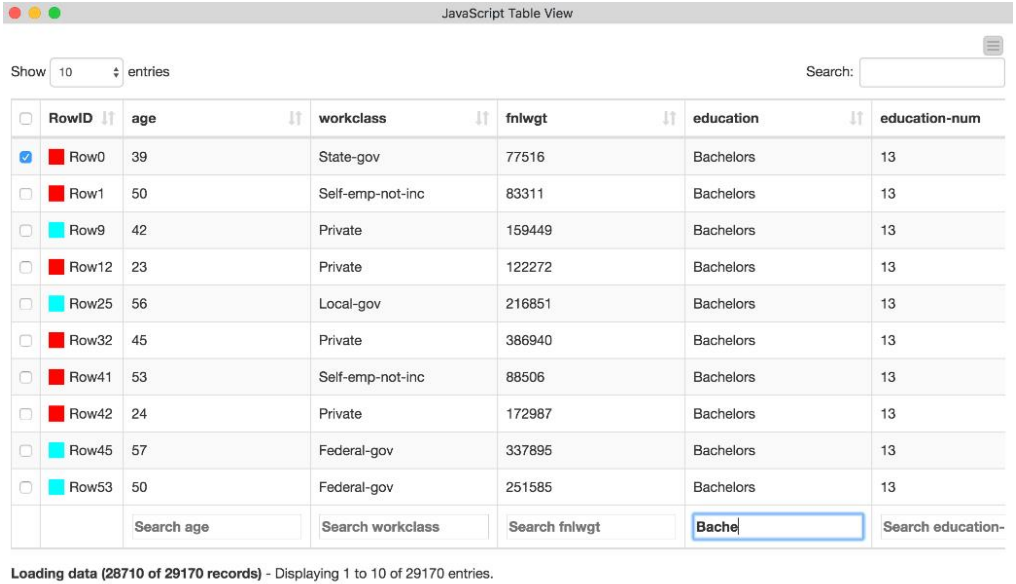


Table View

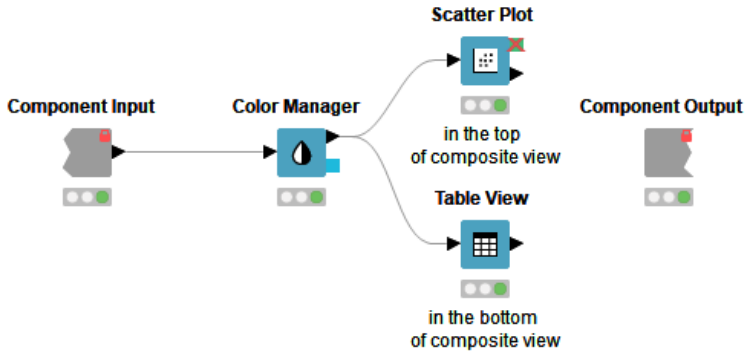
- Displays data in an HTML table view
- The view offers several interactive features, as well as the possibility to select rows

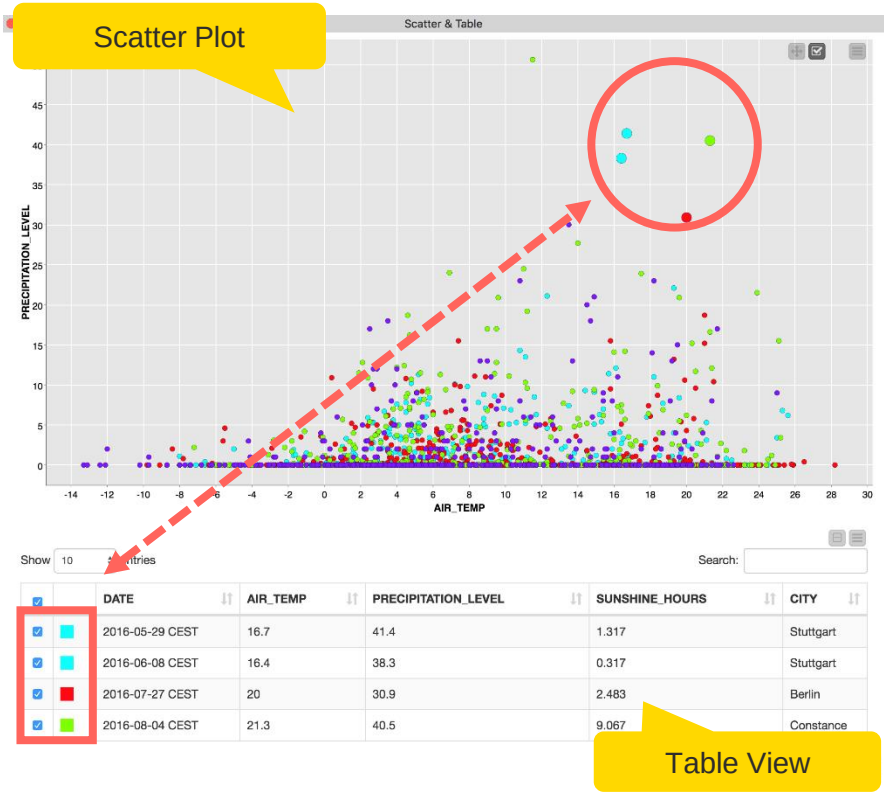




Components – Combined Views

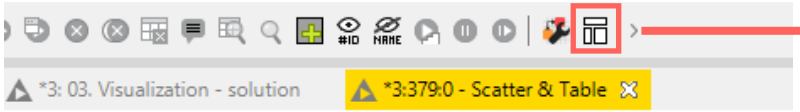
- Multiple JavaScript View nodes can be combined in Components
- Selections are transmitted to all other views
- Also for use on the KNIME KNIME Business Hub

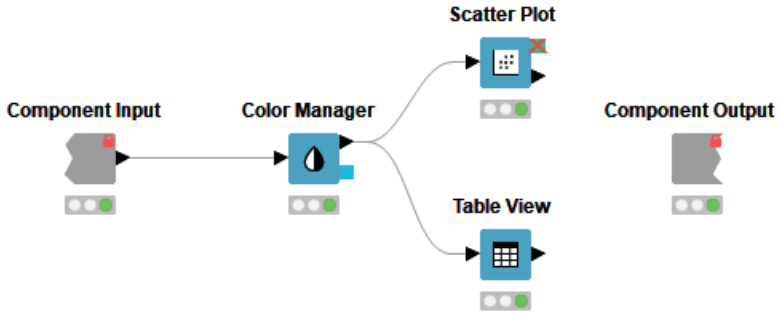




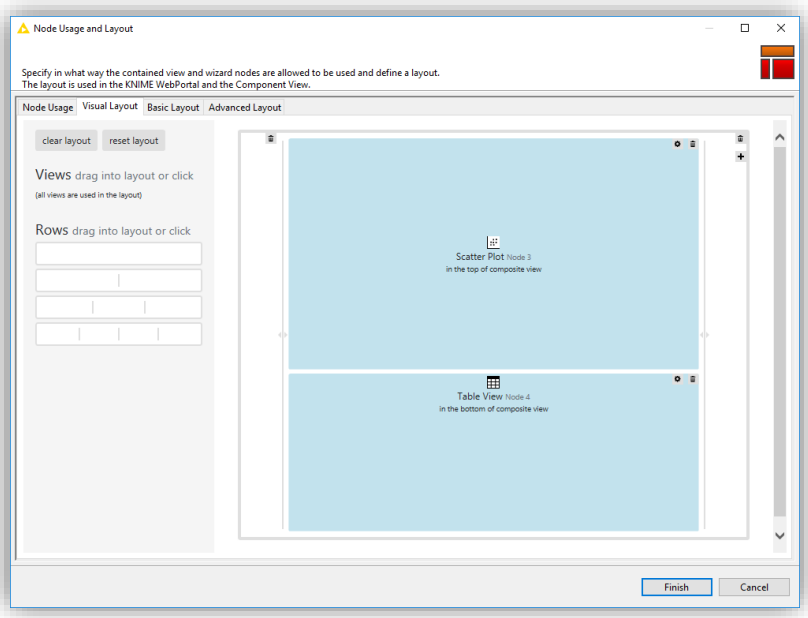
Configure Content and Views Layout

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



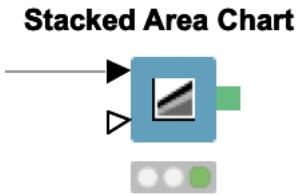


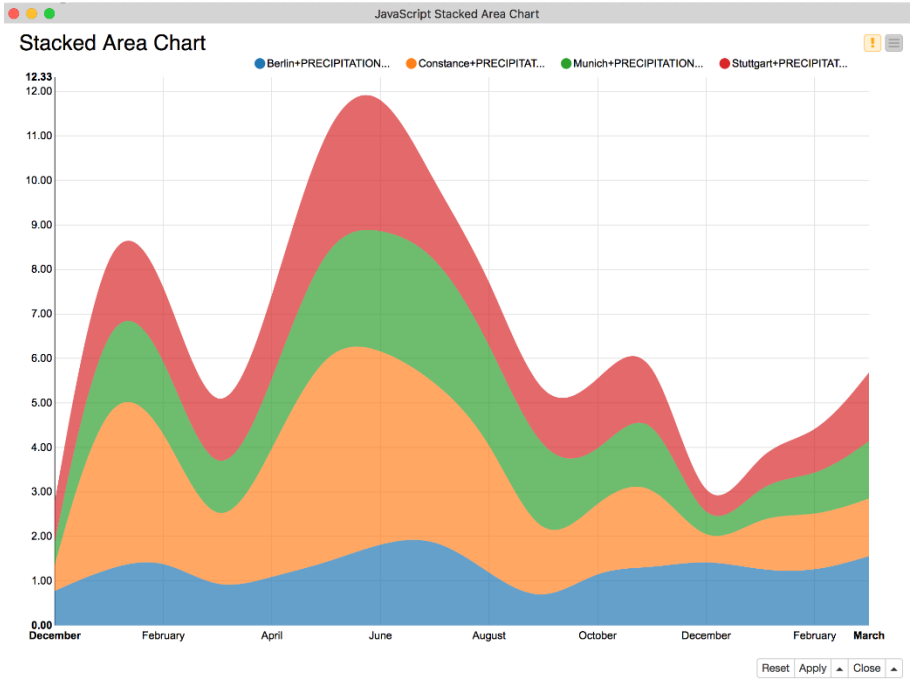
- Add views and rows via drag&drop
- Add columns using **+** buttons



Stacked Area Chart

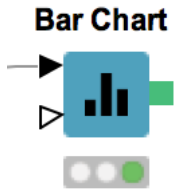
- Visualizes numerical values from multiple columns as stacked areas
- Great for plotting distributions over time

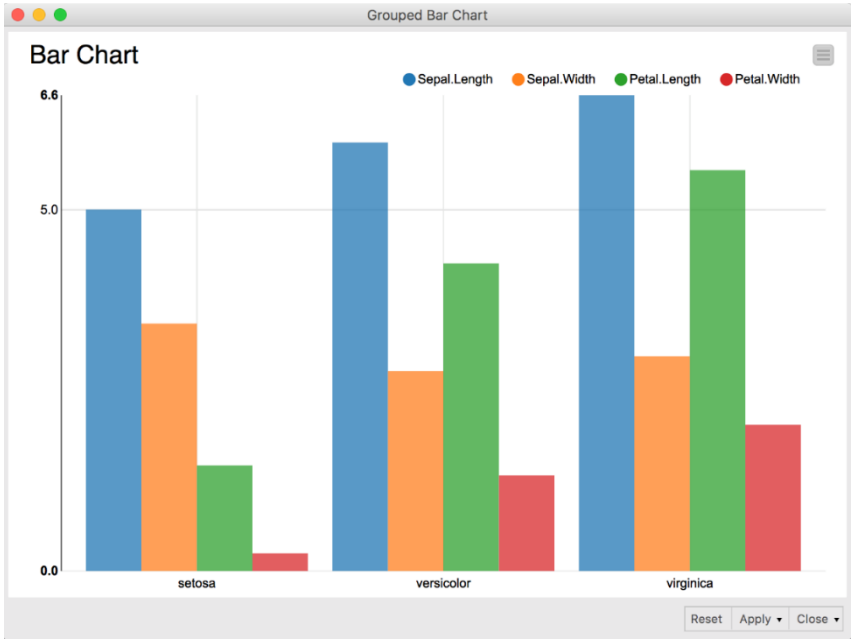




Bar Chart

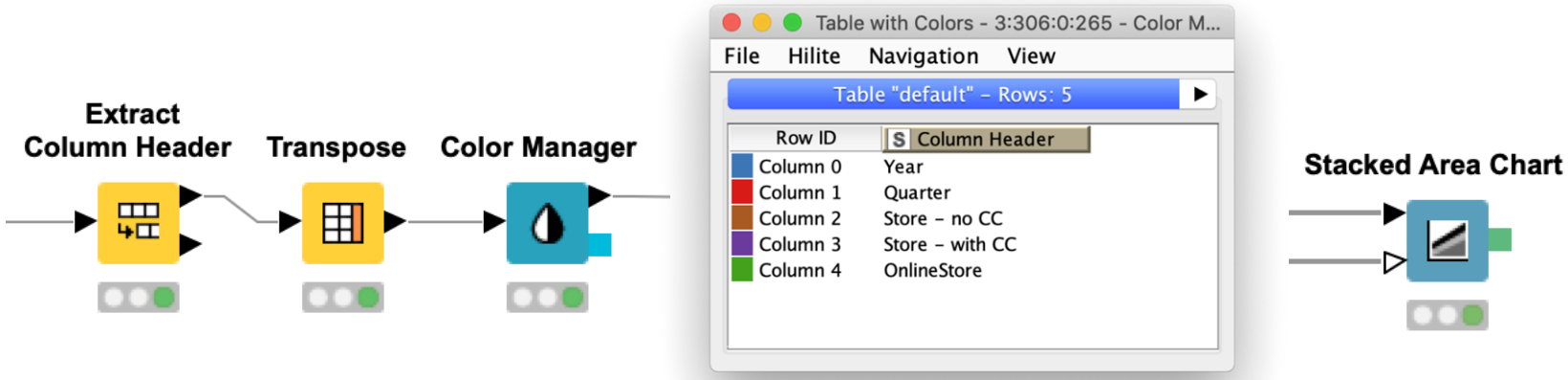
- Shows numerical values across categories
- Vertical or horizontal bars
- Bars can be grouped or stacked





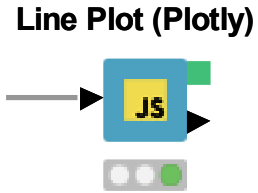
The Optional Color Input Port

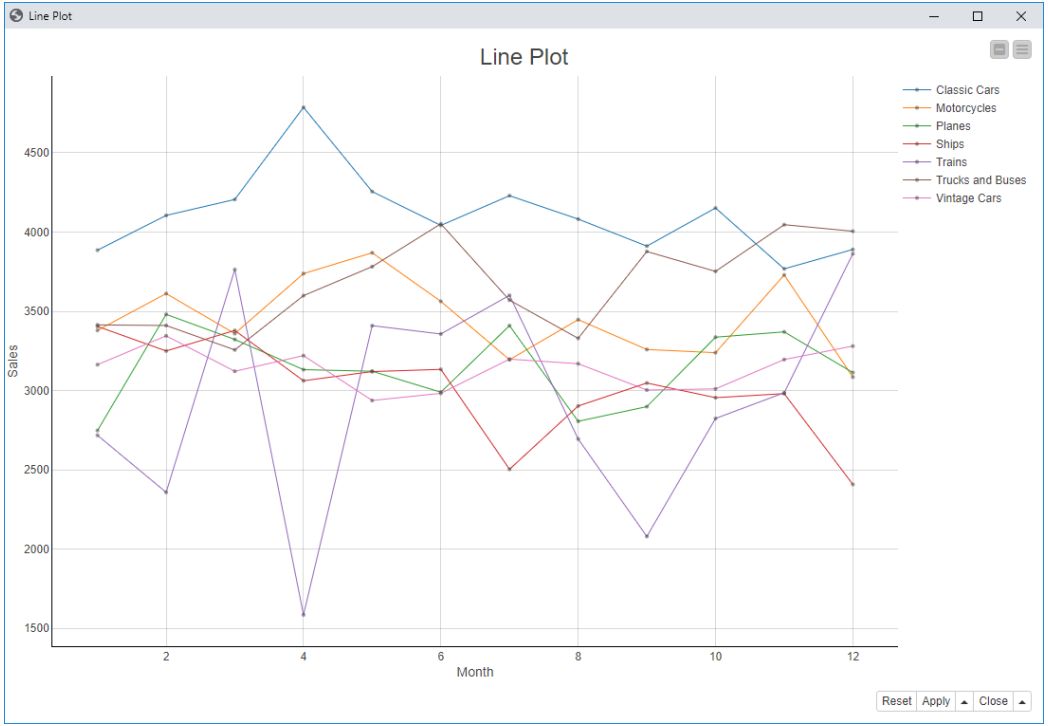
- Many of the visualization nodes have an optional port to change the colors
- Expects table with column headers of first table in the first column with assigned colors



Line Plot

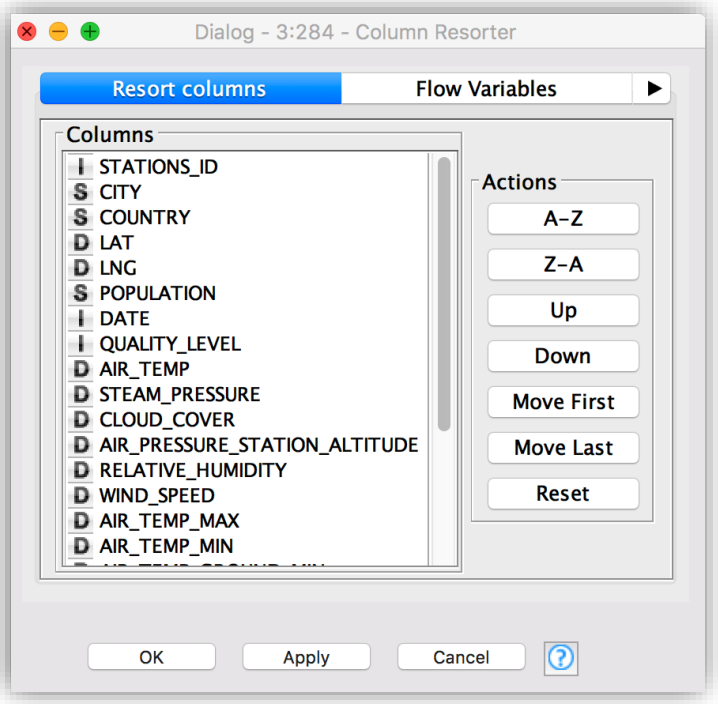
- Plots sequence of values, e.g. over time
- Useful to identify trends, also between groups



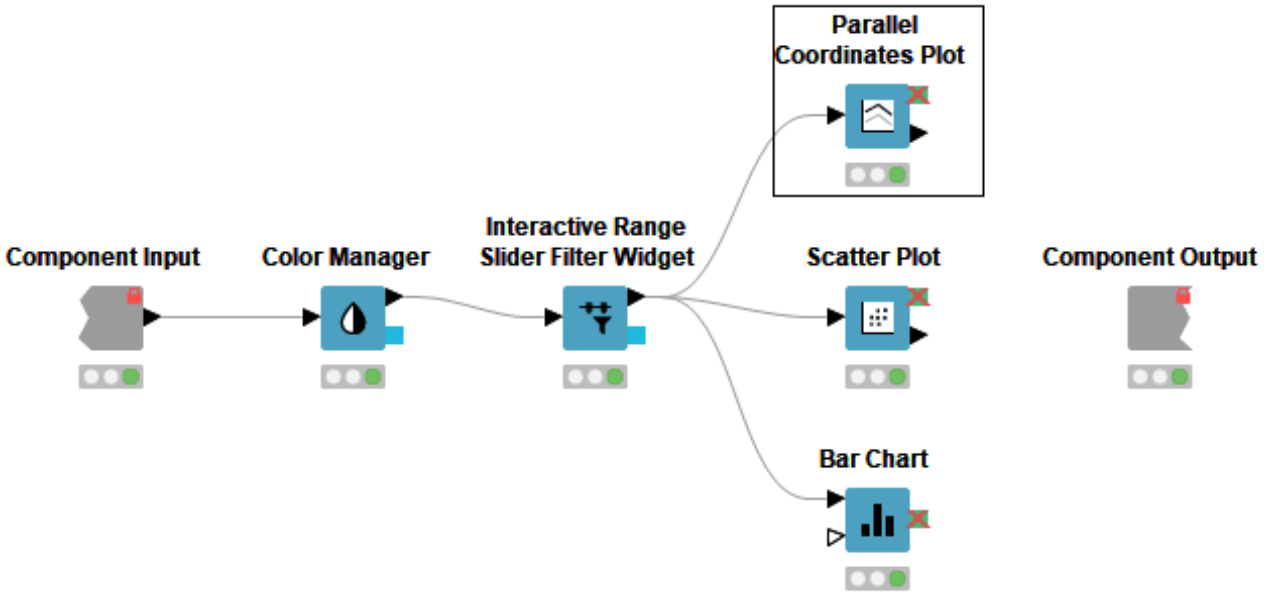


Column Resorter

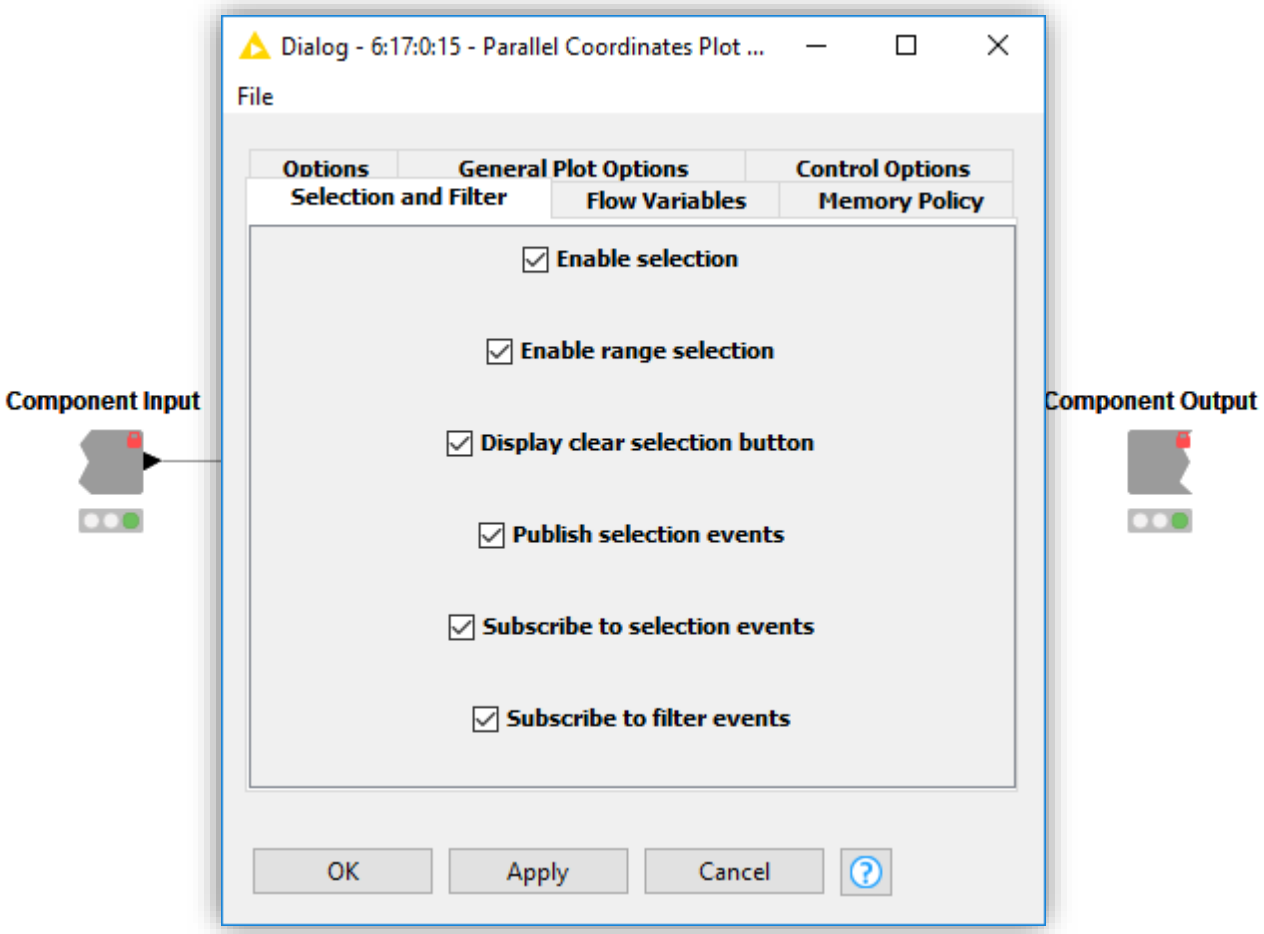
- Changes the order of the input column based on user defined settings
- Options:
  - Sort alphabetical (A-Z or Z-A)
  - Move the selected columns one step (Up or Down)
  - Move the selected columns to top or end (Move First / Last)



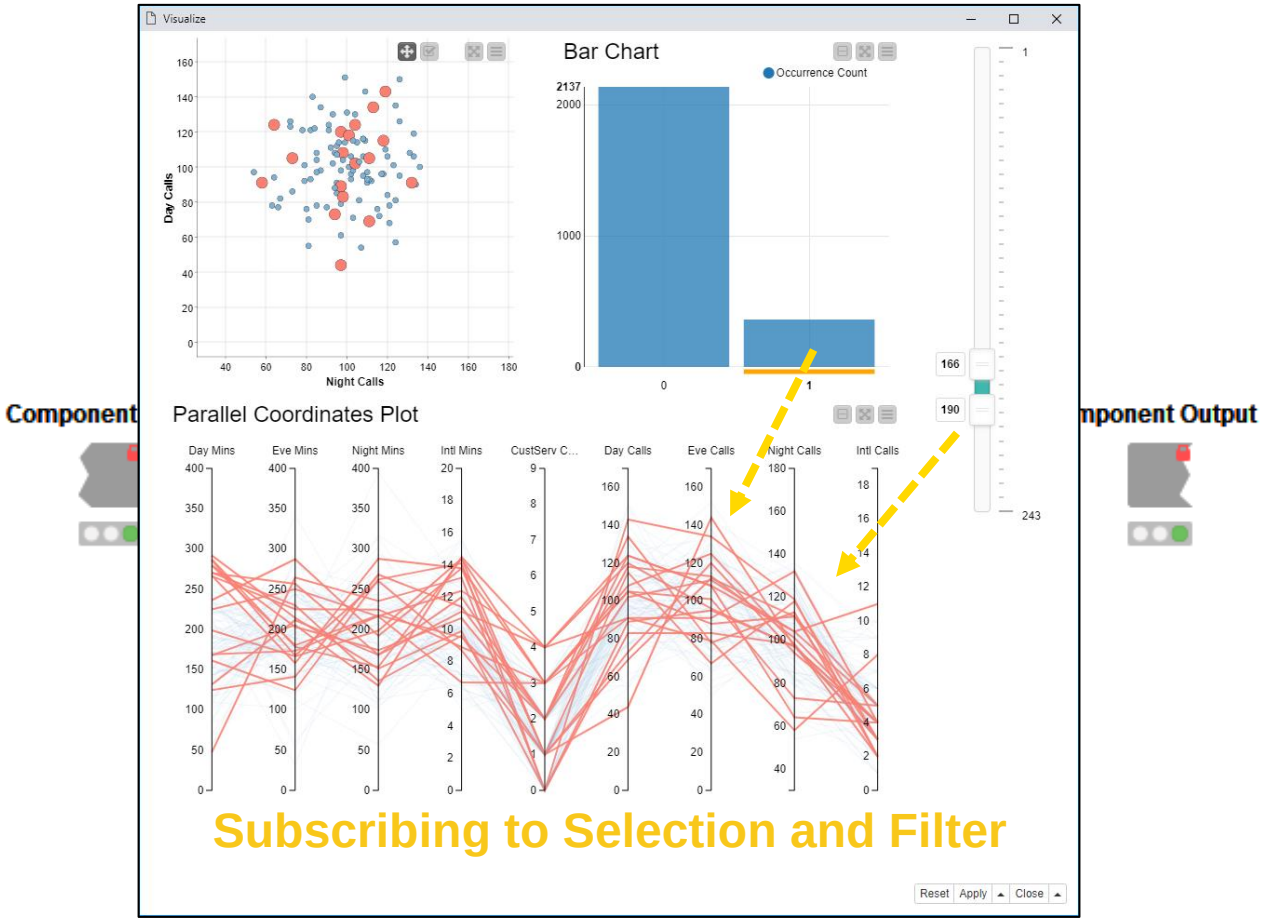
Interactivity across Charts: Selection and Filter Events



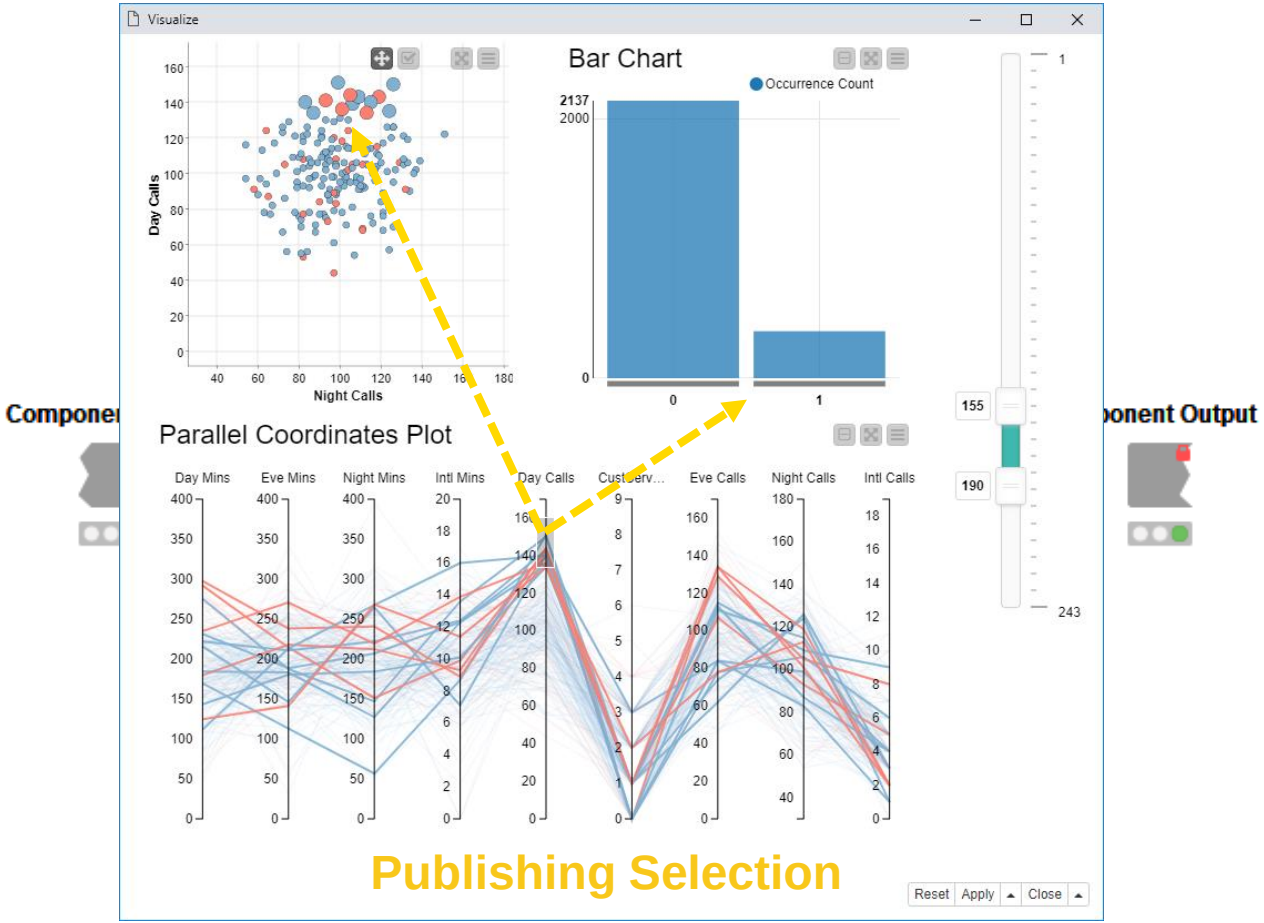
Interactivity across Charts: Selection and Filter Events



Interactivity across Charts: Selection and Filter Events

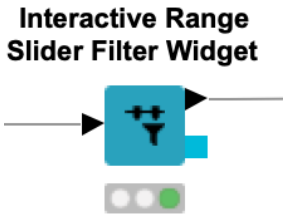


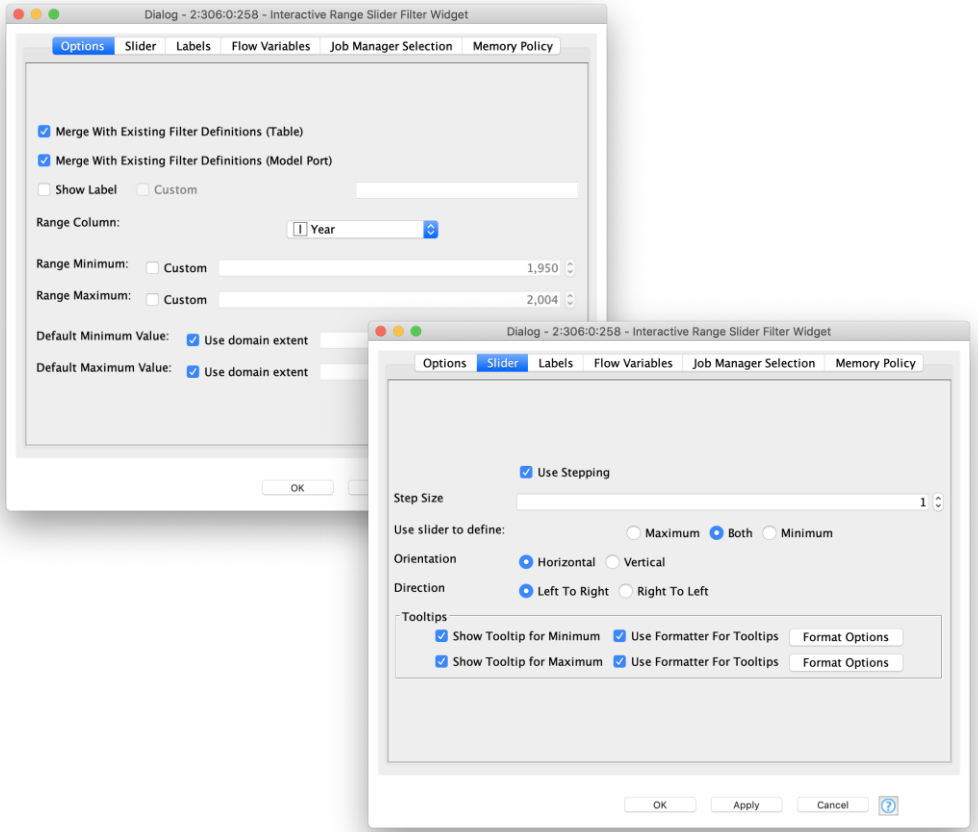
Interactivity across Charts: Selection and Filter Events



Interactive Range Slider Filter Widget

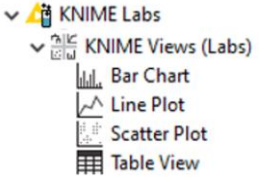
- Slider which can be used to trigger interactive filter events in the view of a component

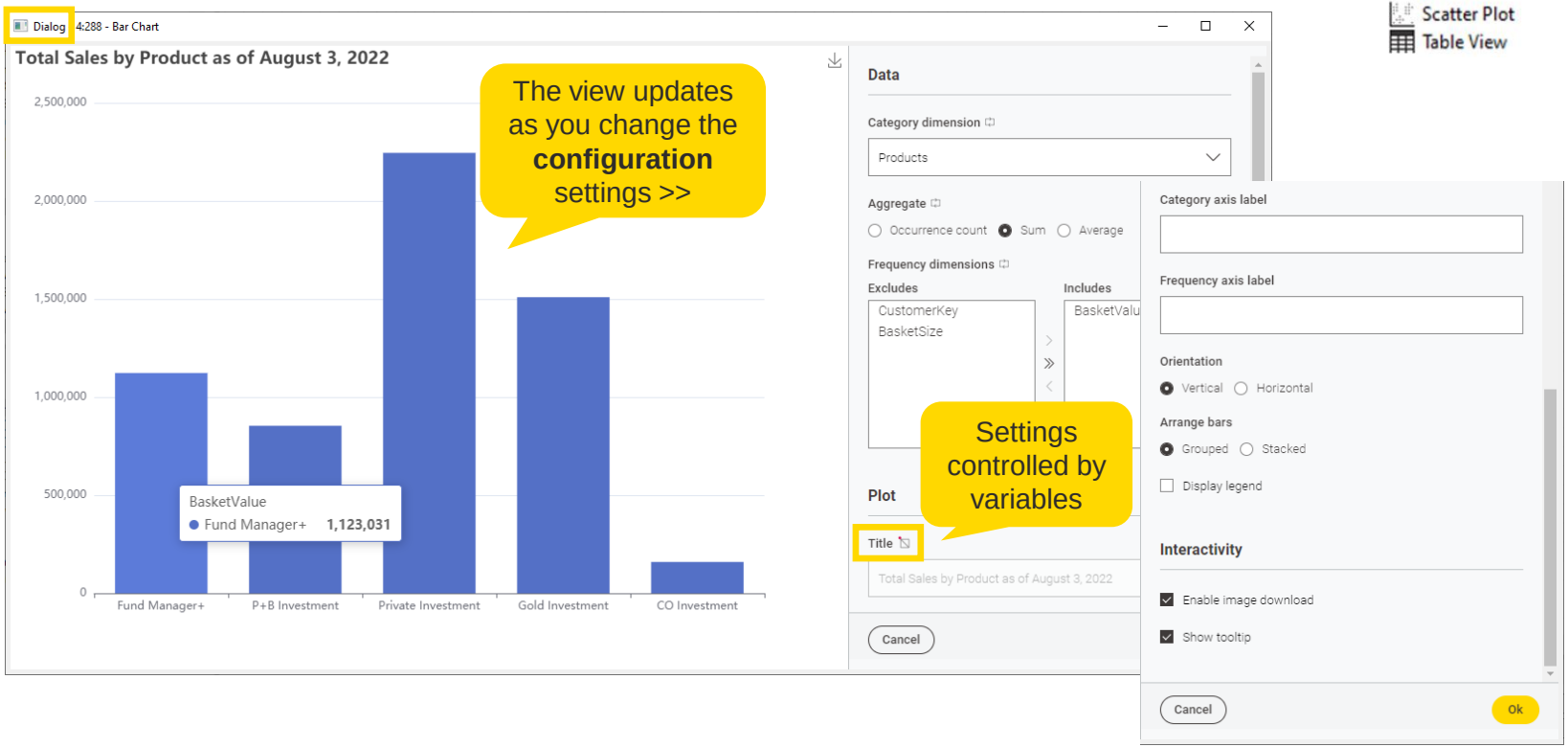




New Visualization Nodes in KNIME (Labs)

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

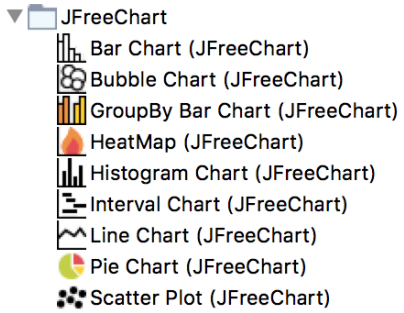




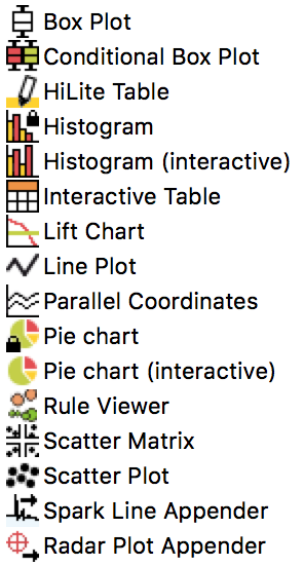
Legacy View Nodes: JFreeChart & KNIME Views

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

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



JFreeChart



KNIME Views

Confirmation of Attendance and Survey

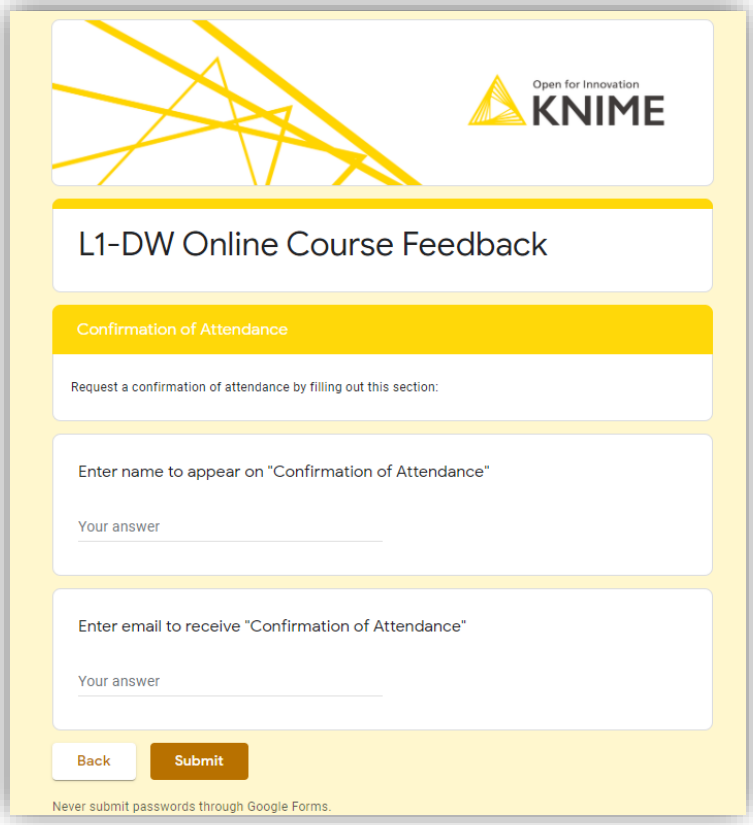
- If you would like to get a “Confirmation of Attendance” please click on the link below*

Confirmation of Attendance and Survey

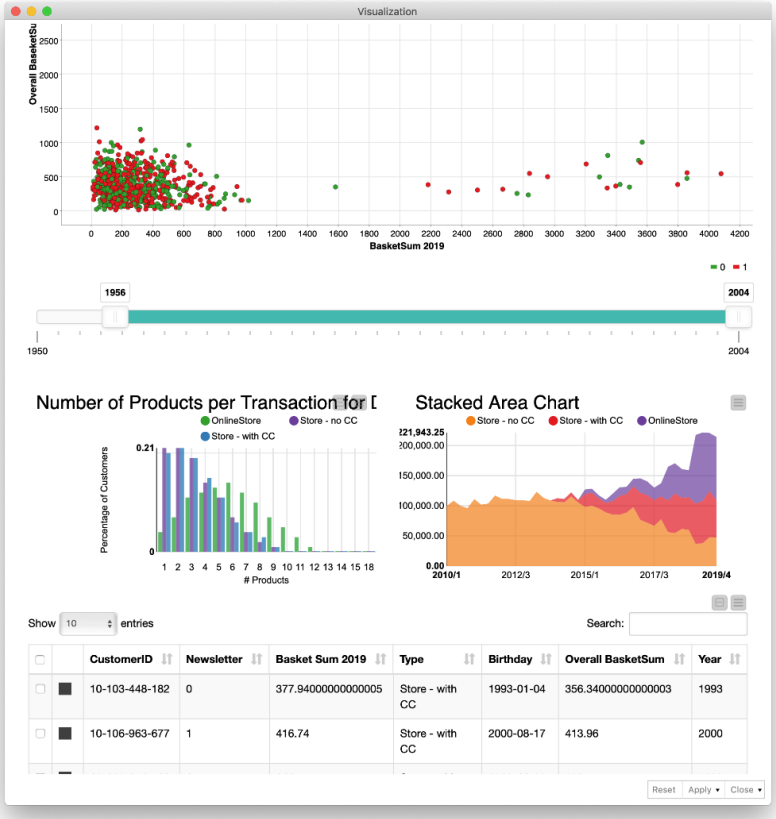
- The link also takes you to our course feedback survey. Filling it in is optional but highly appreciated!

Thank you!

*Please send your request within the next 3 days



Exercise: 07_Visualization – Goal

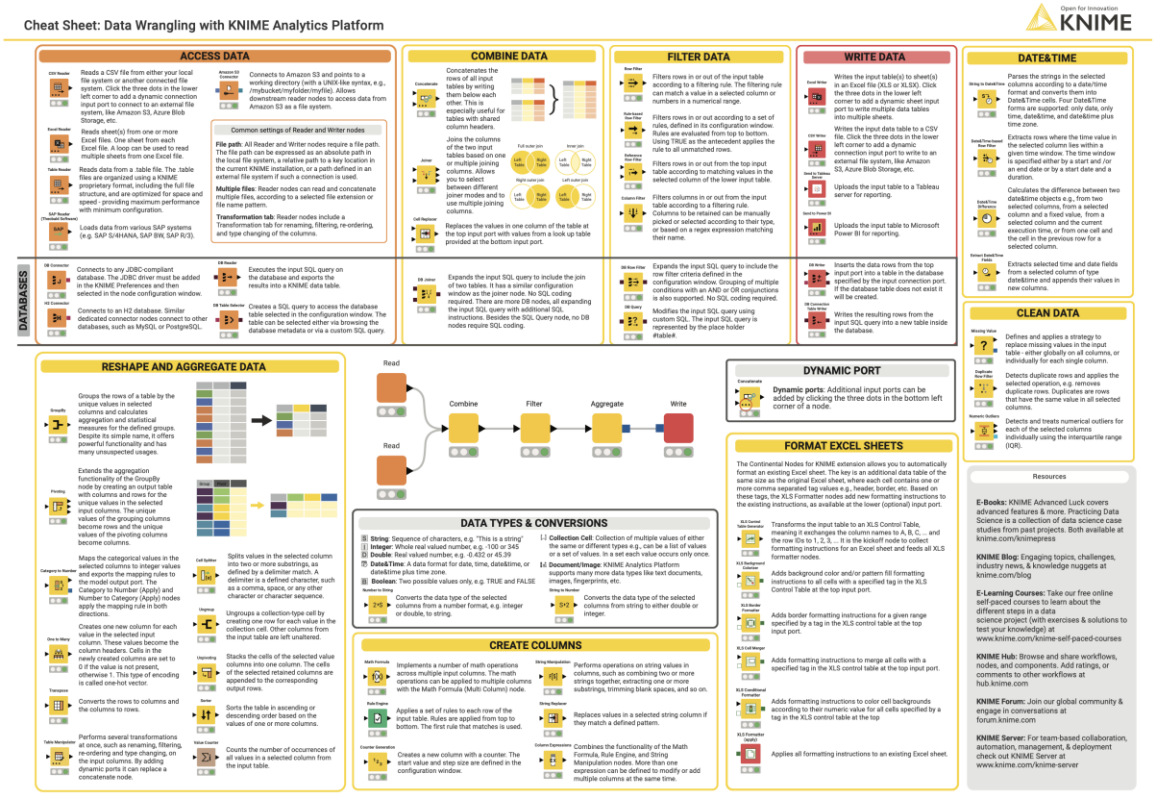


Exercise: 07_Visualization

- 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



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

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

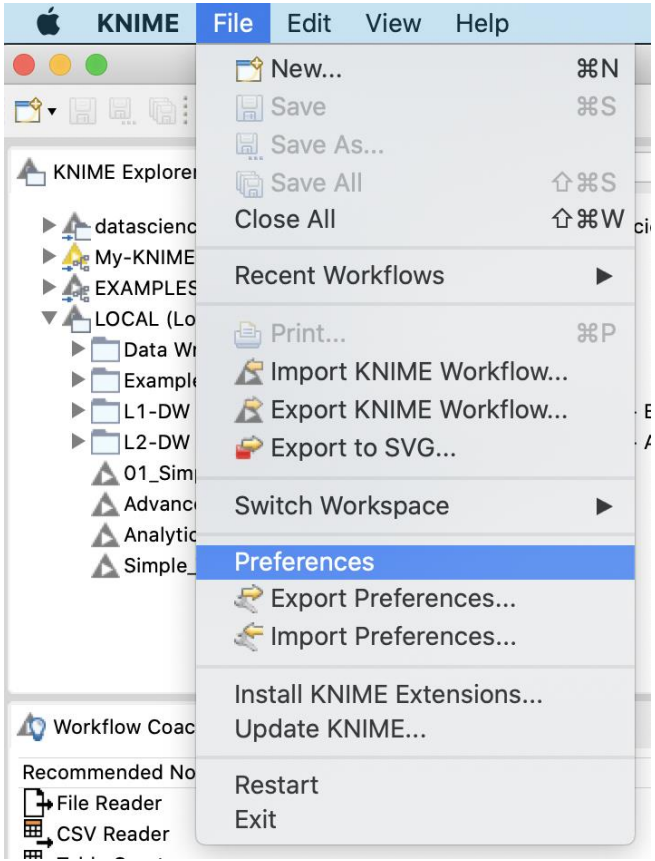
Attachment: How to use a local update site to install extensions

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

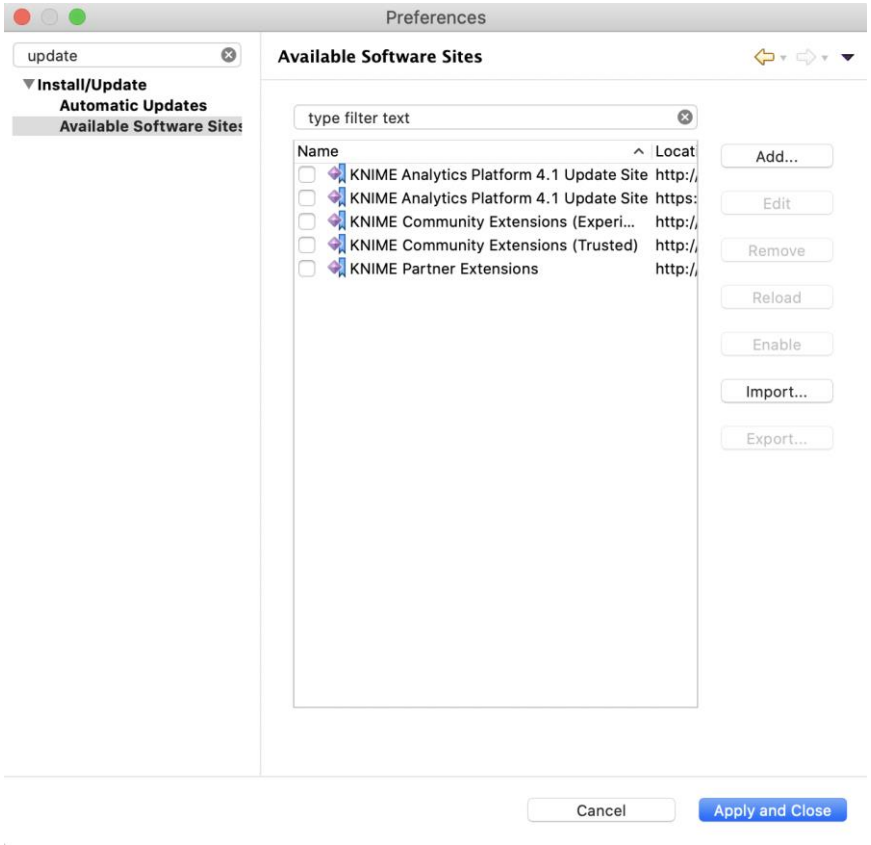
Adding a Local Update Site

- 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