



University of Minho  
School of Engineering



# Machine Learning and Decision-Making

ADI @ LEI/3º, MiEI/4º - 2º Semestre  
Filipe Gonçalves, Inês Alves, Cesar Analide

Part IV – March 2022

# Contents

2

Data Exploration

Data Preparation

Hands On

- Data Exploration
- Data Preparation
  - Join, Concatenation, Sorter, Filter and Aggregations
  - Feature Scaling, Outlier Detection, Feature Selection, Missing Values Treatment, Nominal Value Discretization, Binning and Feature Engineering
- Hands On

# Covariance

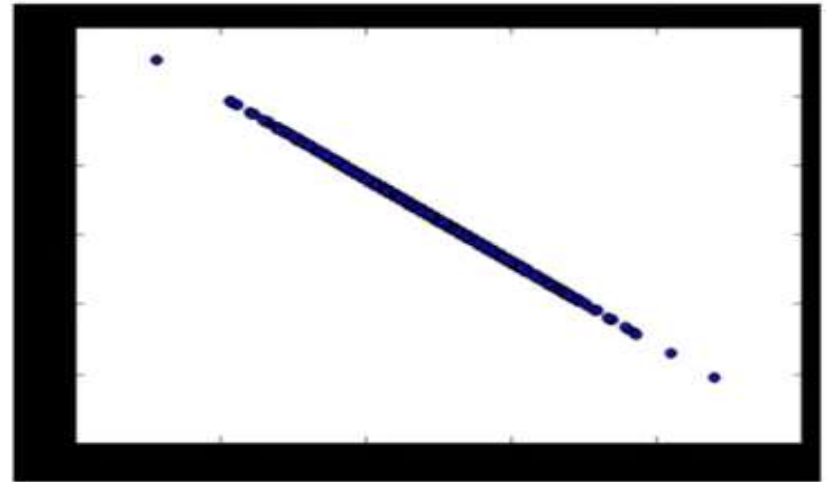
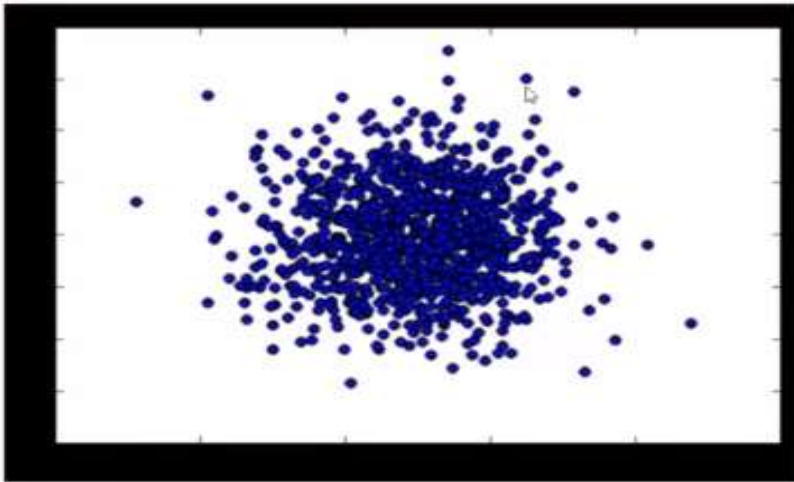
3

DATA EXPLORATION

Data Preparation

Hands On

Measures how **two variables vary in tandem from their means**, i.e., how 2 attributes depend on each other (left plot – low covariance / right plot – high covariance).



# Covariance

4

DATA EXPLORATION

Data Preparation

Hands On

Measuring covariance:

- Think of the datasets for the two variables as high-dimensional vectors
- Convert these to vectors of variances from the mean
- Take the dot product (cosine of the angle between them) of the two vectors
- Divide by the population size

Population Covariance Formula

$$\text{Cov}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N}$$

Sample Covariance

$$\text{Cov}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

# Covariance & Correlation

5

DATA EXPLORATION

Data Preparation

Hands On

Interpreting **covariance** is hard:

- Low covariance (close to 0) means there isn't much correlation between the two variables
- High covariance (far from 0 – can be negative for inverse relationships) means that there is a correlation

Interpreting **correlation** is easier:

- Normalization value of covariance divided by the standard deviations of both variables
  - Correlation of -1: perfect inverse correlation
  - Correlation of 0: no correlation
  - Correlation of 1: perfect correlation

# Covariance & Correlation

6

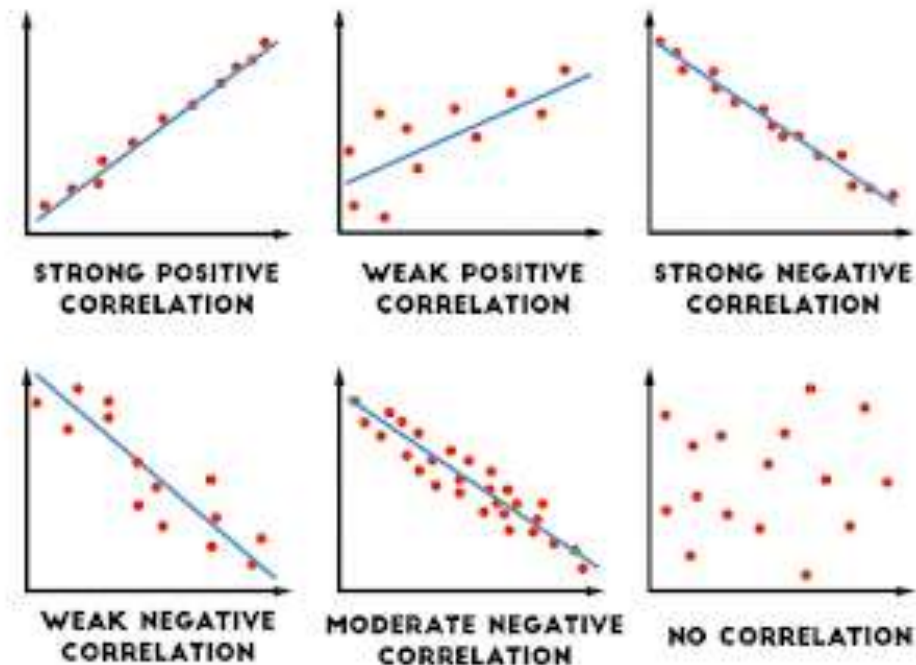
DATA EXPLORATION

Data Preparation

Hands On

But... **Correlation does not imply causation!!**

- Only a controlled, randomized experiment can give you insights on causation;
- Use correlation to decide what experiments to conduct.



# Correlation Matrix

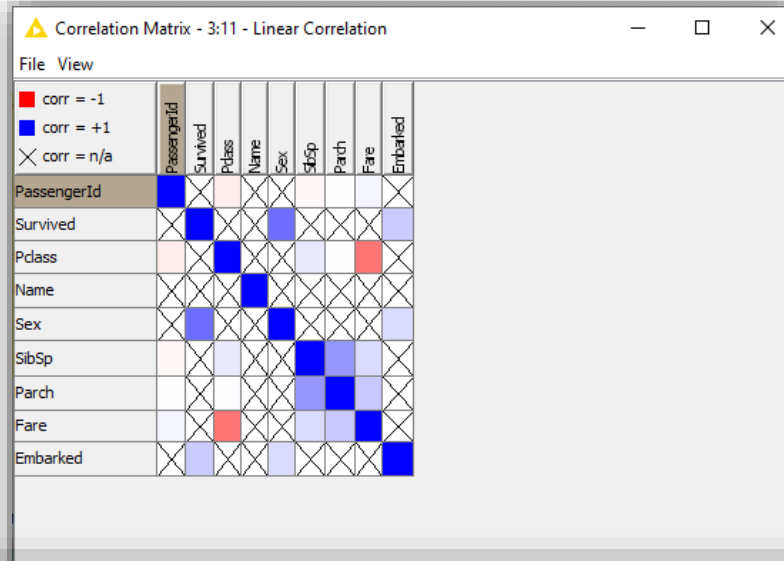
7

DATA EXPLORATION

Data Preparation

Hands On

Linear Correlation



Correlation measure - 3:11 - Linear Correlation

File Hilite Navigation View

Table "Correlation values" - Rows: 9 Spec - Columns: 9 Properties Flow Variables

Row ID	D Passen...	D Survived	D Pclass	D Name	D Sex	D SibSp	D Parch	D Fare	D Embarked
PassengerId	1.0	?	-0.074684...	?	?	-0.02572993...	0.0026940469...	0.04019030952...	?
Survived	?	1.0	?	?	0.5737...	?	?	?	0.20559893...
Pclass	-0.074684...	?	1.0	?	?	0.084898459...	0.0060928325...	-0.5393077410...	?
Name	?	?	?	1.0	?	?	?	?	?
Sex	?	0.57374697...	?	?	1.0	?	?	?	0.14153725...
SibSp	-0.0257299...	?	0.0848984...	?	?	1.0	0.4102760703...	0.14270976638...	?
Parch	0.00269404...	?	0.0060928...	?	?	0.410276070...	1.0	0.21318809003...	?
Fare	0.04019030...	?	-0.539307...	?	?	0.142709766...	0.2131880900...	1.0	?
Embarked	?	0.20559893...	?	?	0.1415...	?	?	?	1.0

# Correlation Matrix

9

DATA EXPLORATION

Data Preparation

Hands On

- Do we want to keep **highly-correlated features**?
- Both **positive** and **negatively correlated** ones?
- What about the **correlation between** the **dependent** and the **independent** features?
- ...



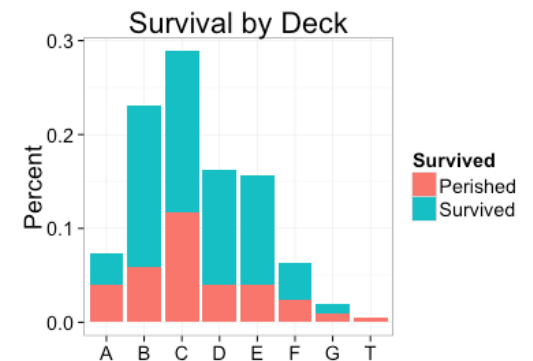
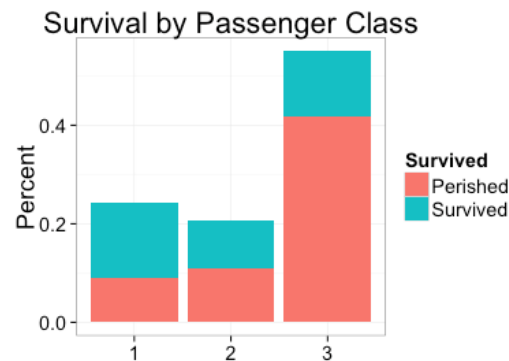
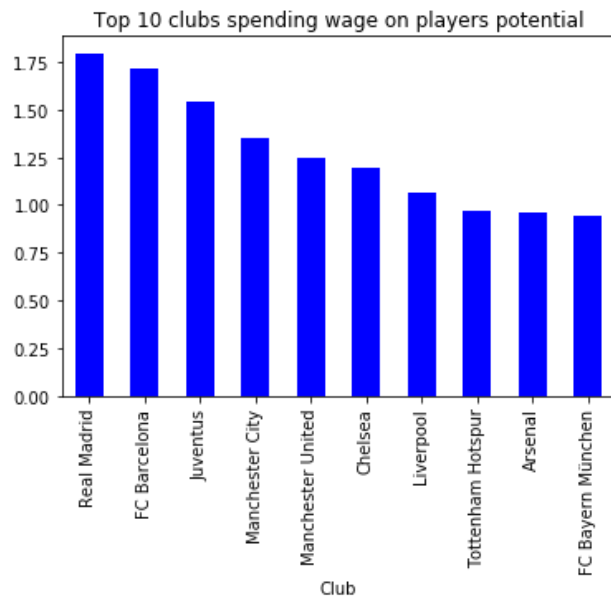
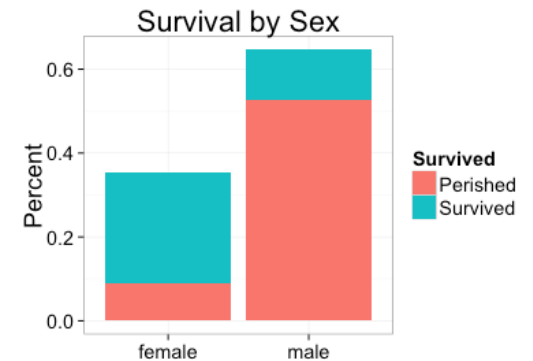
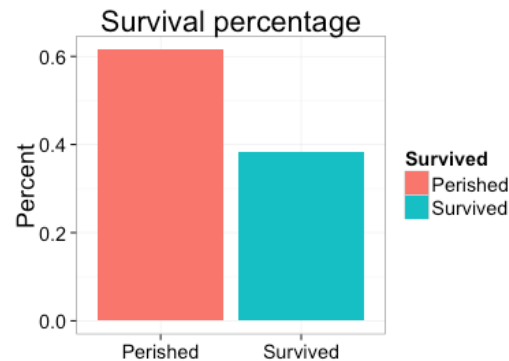
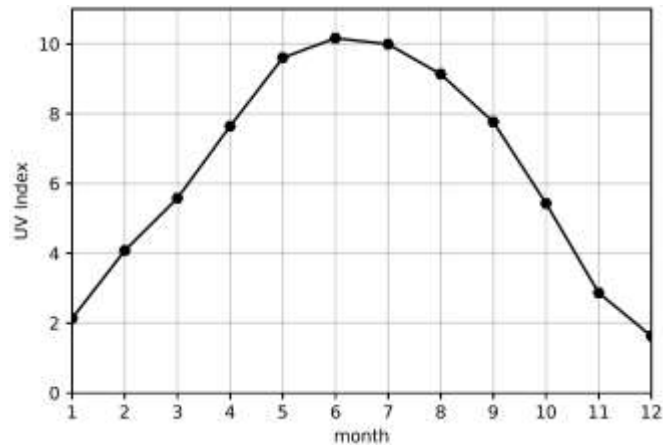
# Data Viz. <- Often Neglected

10

DATA EXPLORATION

Data Preparation

Hands On



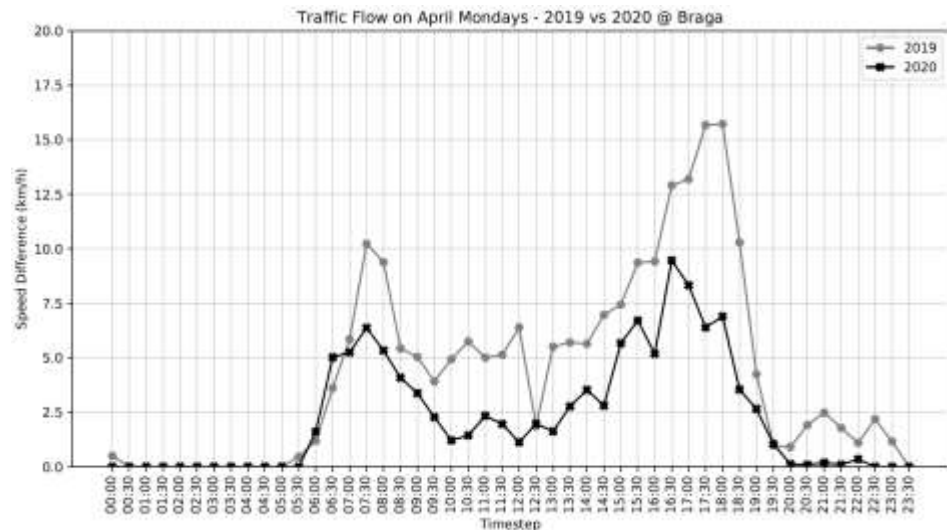
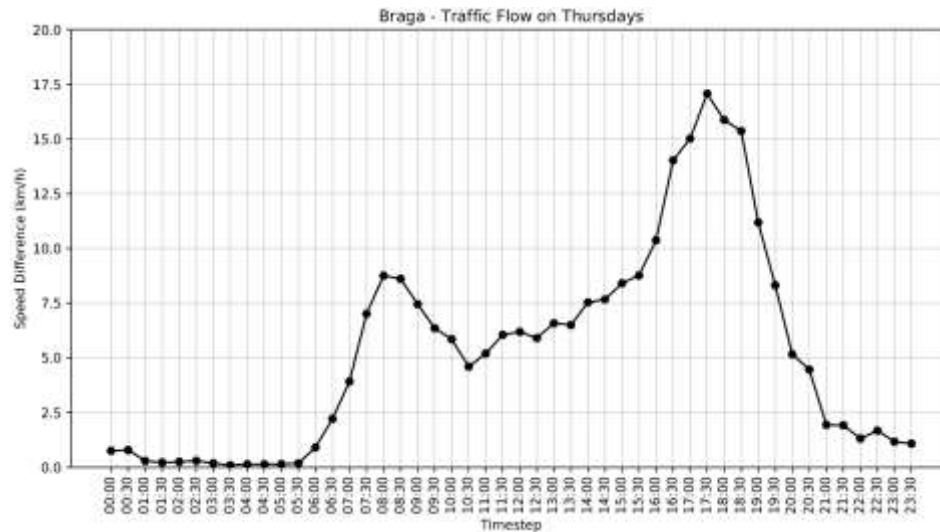
# Data Viz. <- Often Neglected

11

DATA EXPLORATION

Data Preparation

Hands On





# Join

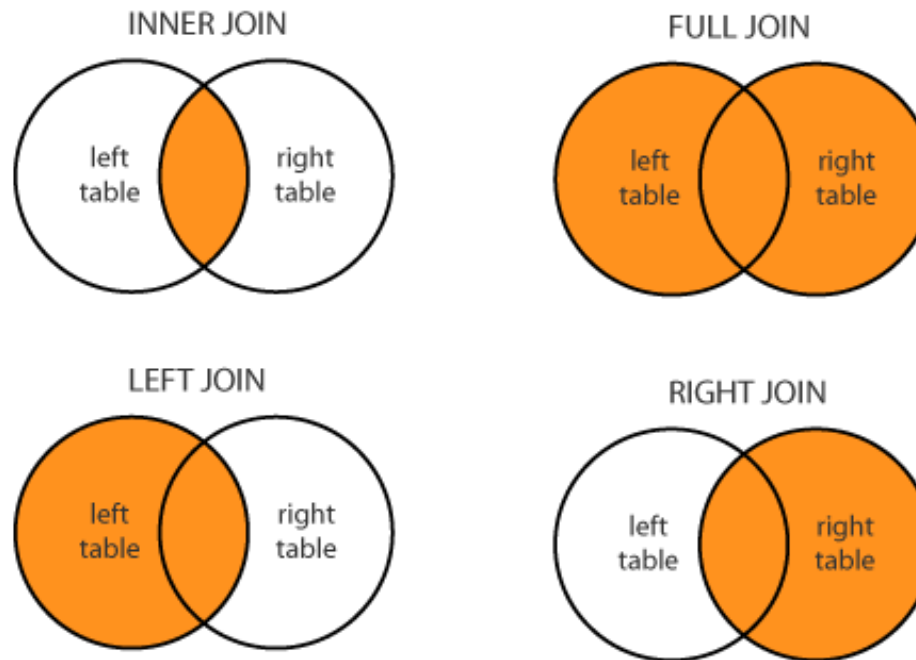
15

Data Exploration

**DATA PREPARATION**

Hands On

A **Join** is an operation that combines data from different tables



# Join

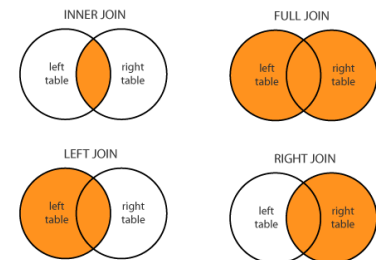
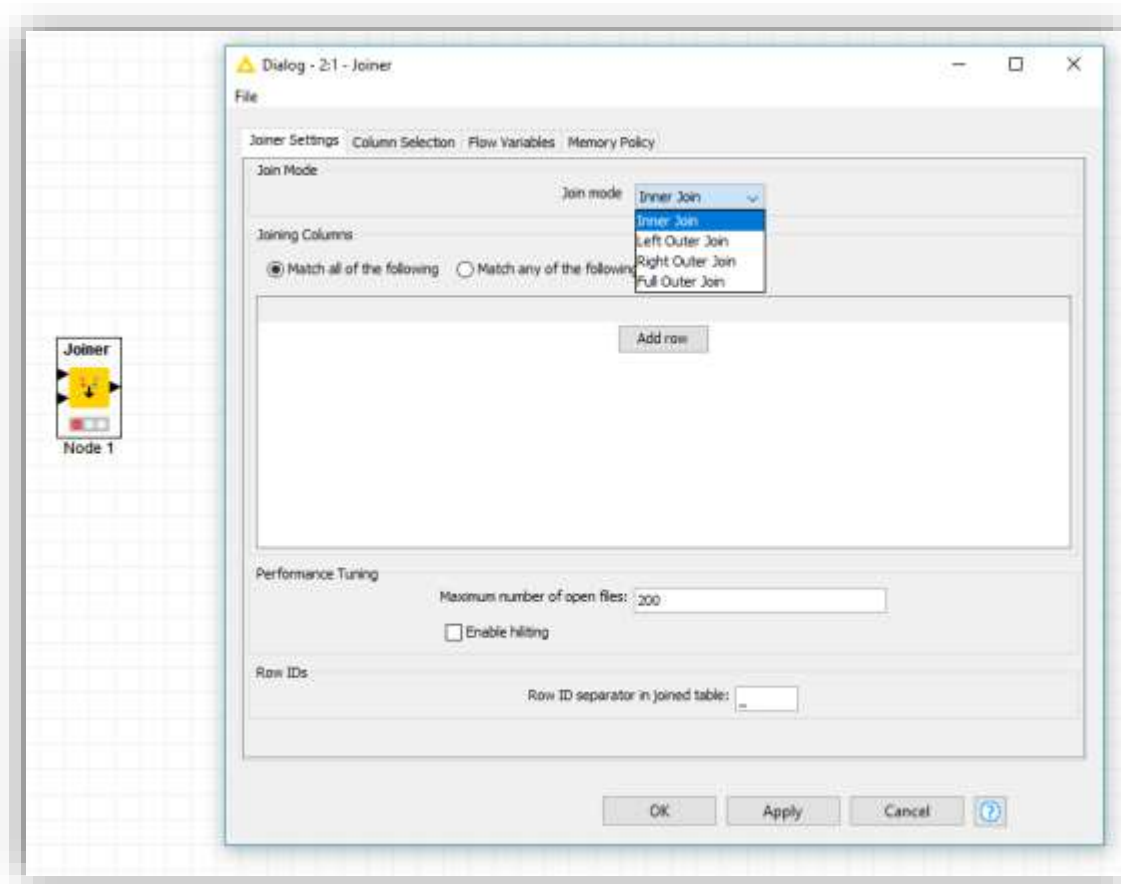
16

Data Exploration

DATA PREPARATION

Hands On

Knime offers **inner joins**, **right outer joins**, **left outer joins** and **full outer joins**



# Concatenation

17

Data Exploration

DATA PREPARATION

Hands On

## Union of columns

Manually created table - 2:2 - Table Creator

File Hilite Navigation View

Table "default" - Rows: 2 Spec - Columns: 3 Properties Flow Variables

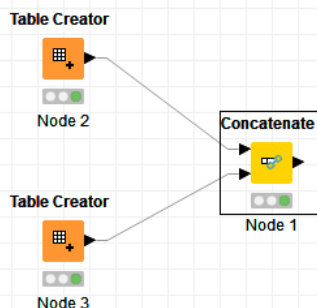
Row ID	S Id	S Name	S Age
Row0	1	Ze	19
Row1	2	Maria	22

Manually created table - 2:3 - Table Creator

File Hilite Navigation View

Table "default" - Rows: 3 Spec - Columns: 4 Properties Flow Variables

Row ID	S Id	S Address	S Name	S Phone
Row0	3	Braga	Rui	543
Row1	4	Porto	Ana	345
Row2	5	Braga	João	324



Dialog - 2:1 - Concatenate

File

Settings Flow Variables Memory Policy

Duplicate row ID handling

☐ Skip Rows

☒ Append Suffix:

☐ Fail Execution

Column handling

☐ Use intersection of columns

☒ Use union of columns

Hilting

☐ Enable hilting

OK Apply Cancel ?

Concatenated table - 2:1 - Concatenate

File Hilite Navigation View

Table "default" - Rows: 5 Spec - Columns: 5 Properties Flow Variables

Row ID	S Id	S Name	S Age	S Address	S Phone
Row0	1	Ze	19	?	?
Row1	2	Maria	22	?	?
Row0_dup	3	Rui	?	Braga	543
Row1_dup	4	Ana	?	Porto	345
Row2	5	João	?	Braga	324

# Concatenation

18

Data Exploration

DATA PREPARATION

Hands On

## Intersection of columns

Manually created table - 2:2 - Table Creator

File Hilite Navigation View

Table "default" - Rows: 2 Spec - Columns: 3 Properties Flow Variables

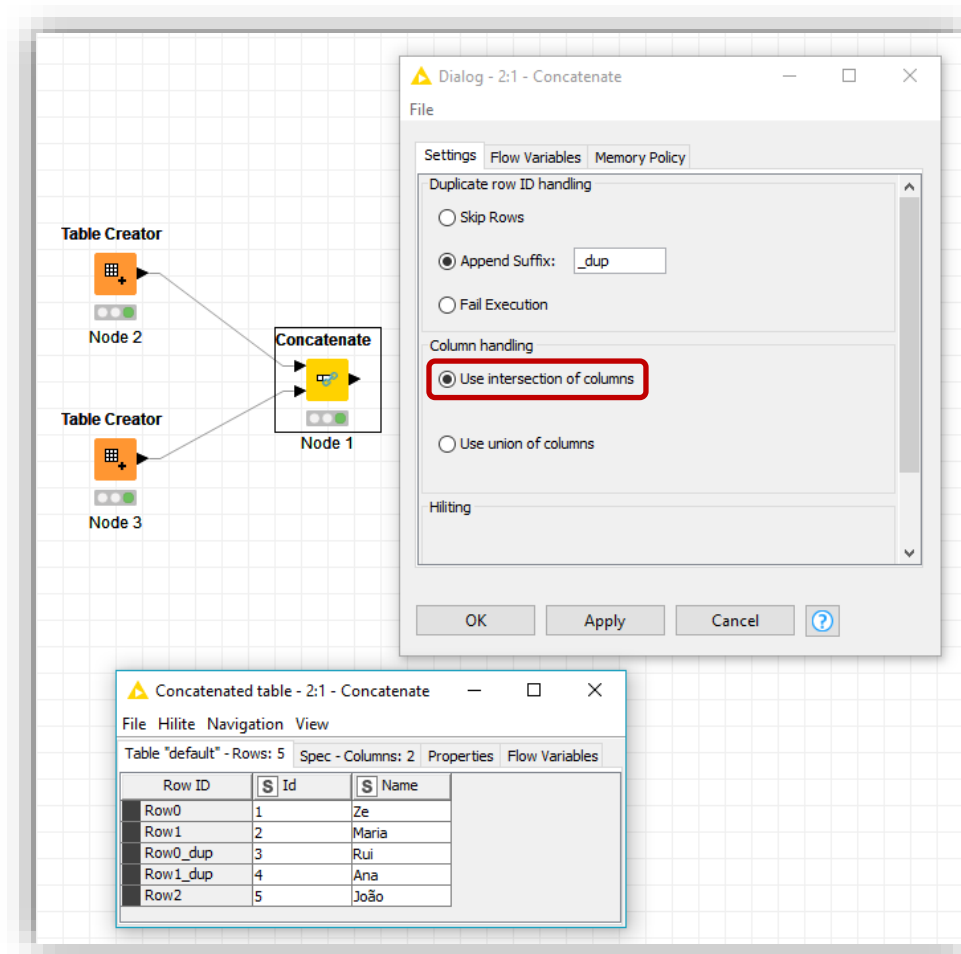
Row ID	S Id	S Name	S Age
Row0	1	Ze	19
Row1	2	Maria	22

Manually created table - 2:3 - Table Creator

File Hilite Navigation View

Table "default" - Rows: 3 Spec - Columns: 4 Properties Flow Variables

Row ID	S Id	S Address	S Name	S Phone
Row0	3	Braga	Rui	543
Row1	4	Porto	Ana	345
Row2	5	Braga	João	324



# Concatenation

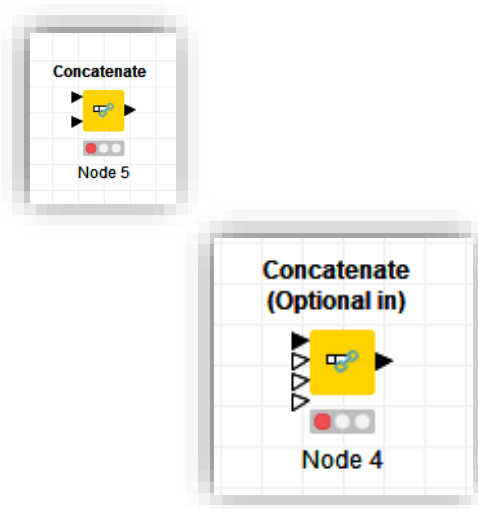
19

Data Exploration

**DATA PREPARATION**

Hands On

**Concatenate (Optional in)** works exactly like Concatenate but accepts up to 4 inputs!





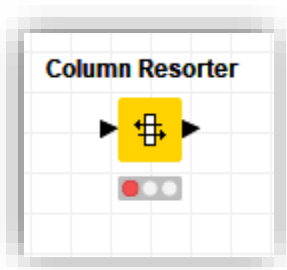
# Sorter

20

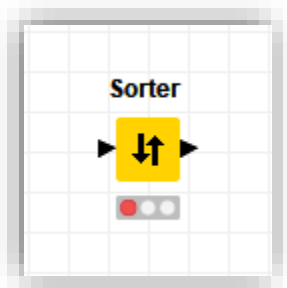
Data Exploration

DATA PREPARATION

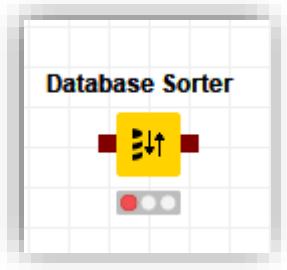
Hands On



Changes the order of the input **columns**, based on user defined settings



Sorts **rows** according to user-defined criteria



Allows **rows** to be sorted from the input database table (SQL ORDER BY clause)

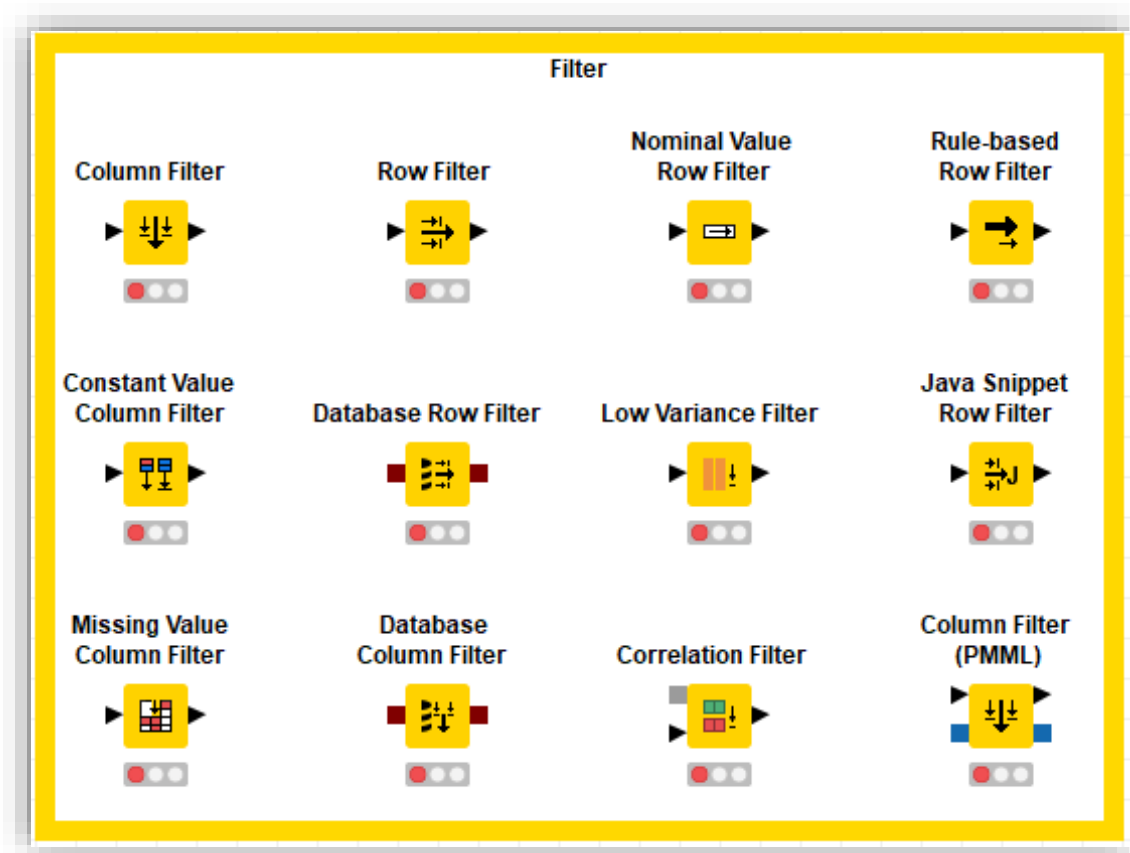
# Filter

21

Data Exploration

DATA PREPARATION

Hands On



Node Description ⓘ

### Column Filter

This node allows columns to be filtered from the input table while only the remaining columns are passed to the output table. Within the dialog, columns can be moved between the Include and Exclude list.

#### Dialog Options

**Include**

This list contains the column names that are included in the output table.

**Enforce Inclusion**

Select this option to enforce the current inclusion list to stay the same even if the input table specification changes. If some of the included columns are not available anymore, a warning is displayed. (New columns will automatically be added to the exclusion list.)

**Select**

Use these buttons to move columns between the Include and Exclude list.

**Search**

Use one of these fields to search either within the Include or Exclude list for certain column names or name substrings. Repeated clicking of the search button marks the next column that matches the search text. The checkbox 'Mark all search hits' causes all matching columns to be selected making them movable between the two lists.

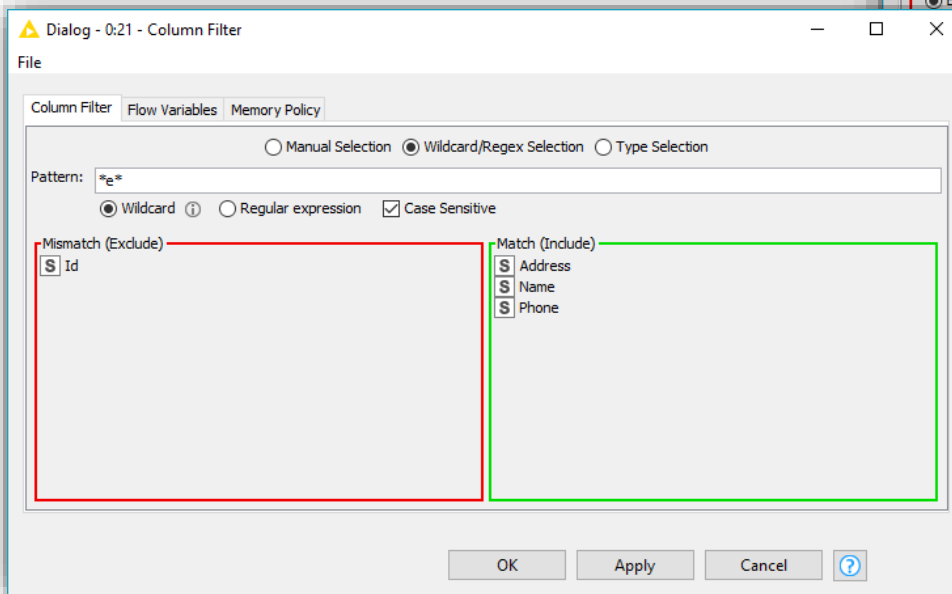
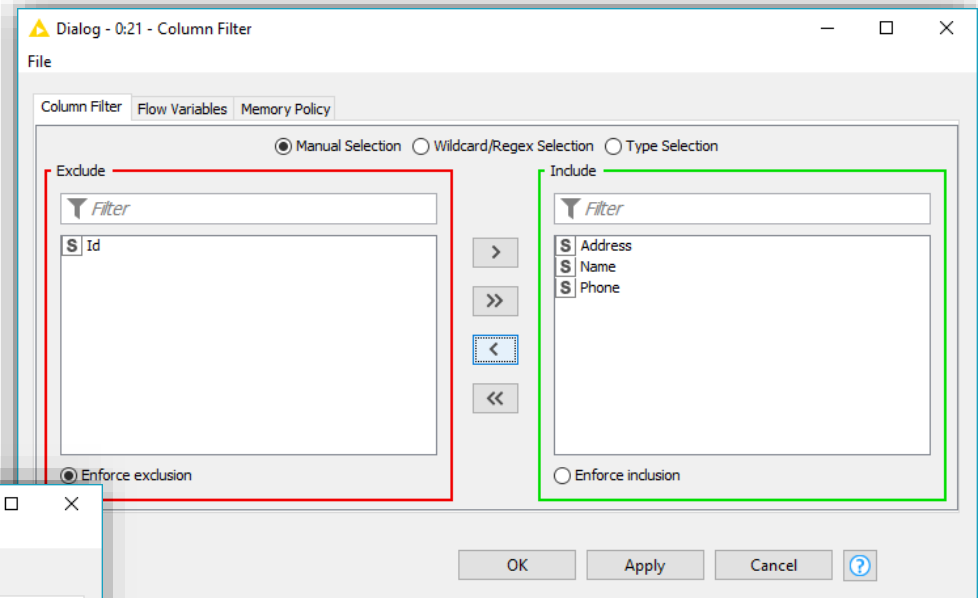
# Column Filter

22

Data Exploration

DATA PREPARATION

Hands On



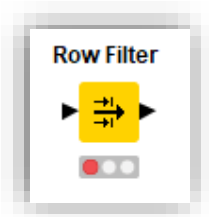
# Row Filter

23

Data Exploration

DATA PREPARATION

Hands On



Dialog - 0:30 - Row Filter

File

Filter Criteria | Flow Variables | Memory Policy

☒ Include rows by attribute value  
☐ Exclude rows by attribute value  
☐ Include rows by number  
☐ Exclude rows by number  
☐ Include rows by row ID  
☐ Exclude rows by row ID

Column value matching

Column to test: **S** Name

☐ filter based on collection elements

Matching criteria

☒ use pattern matching

\*ana\*

☐ case sensitive match ☒ contains wild cards  
☐ regular expression

☐ use range checking

lower bound:   
upper bound:

☐ only missing values match

OK Apply Cancel ?

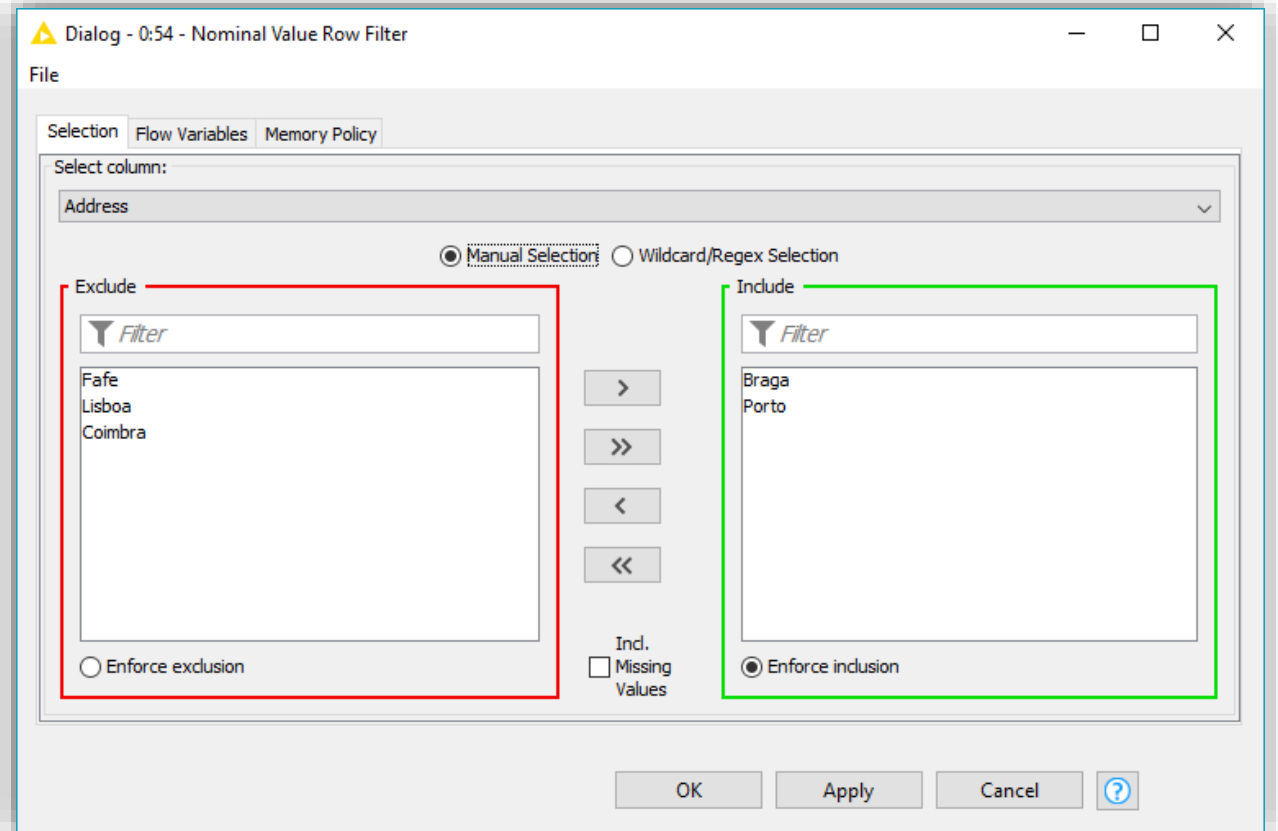
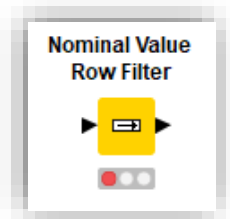
# Nominal Value Row Filter

24

Data Exploration

DATA PREPARATION

Hands On



# Rule-based Row Filter

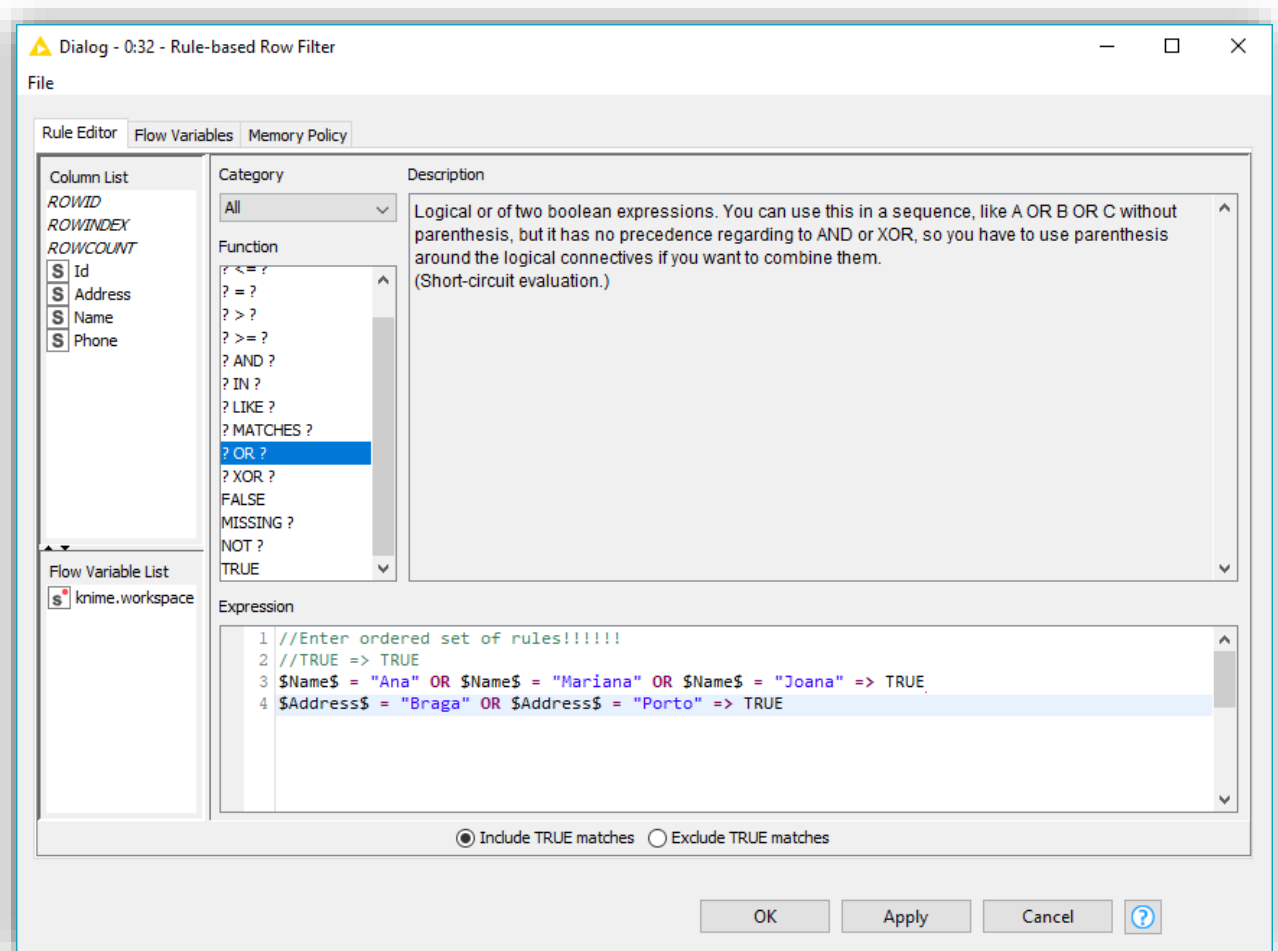
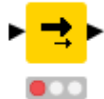
25

Data Exploration

DATA PREPARATION

Hands On

Rule-based  
Row Filter



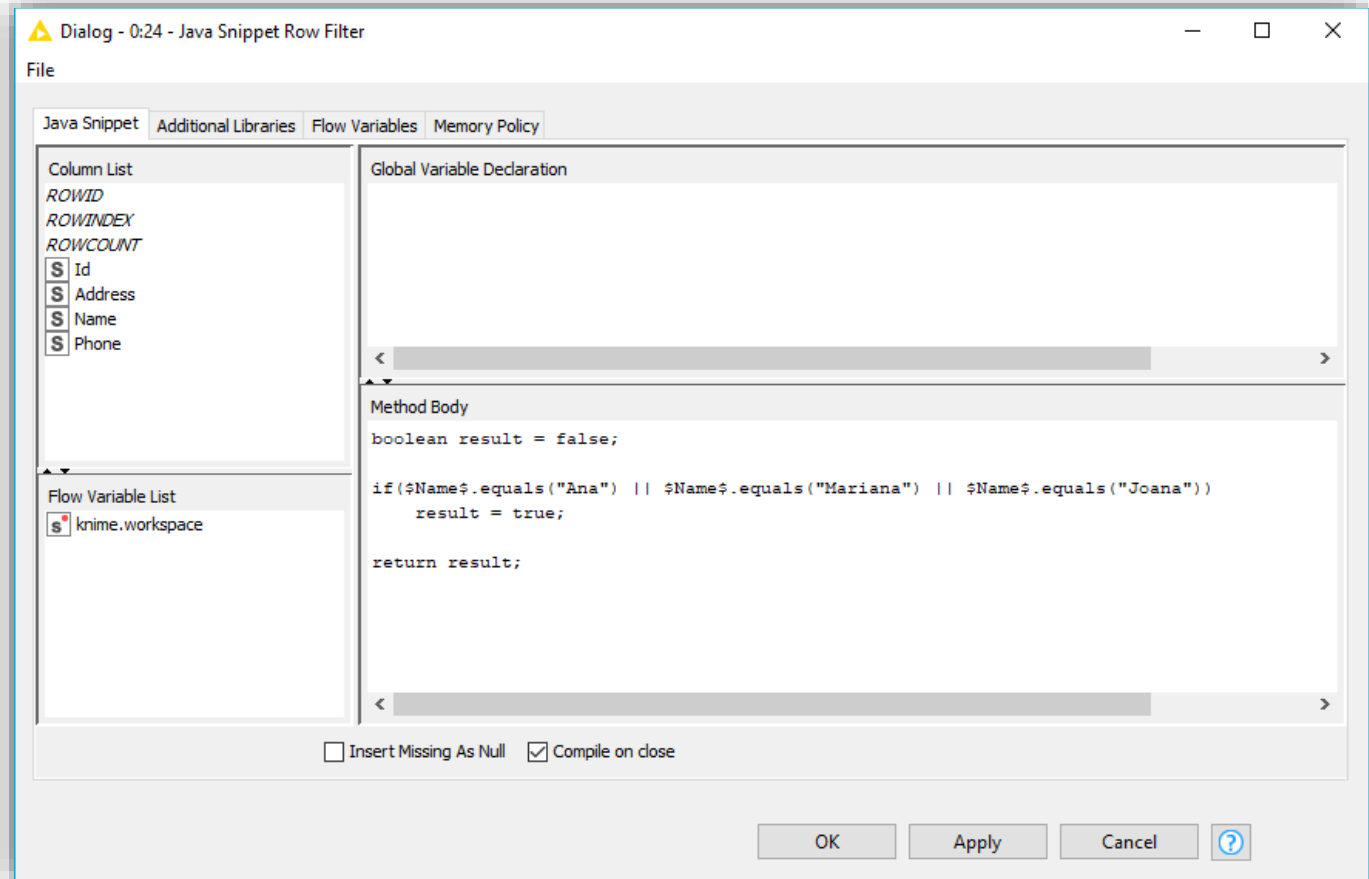
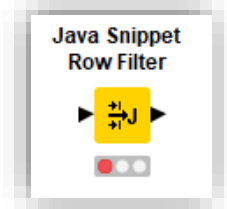
# Java Snippet Row Filter

26

Data Exploration

DATA PREPARATION

Hands On



# Basic Aggregations

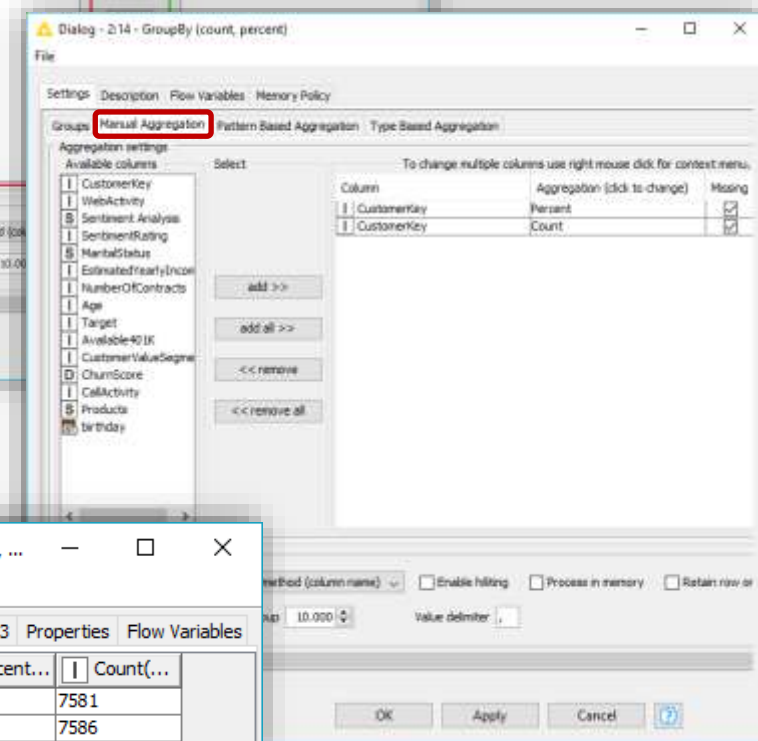
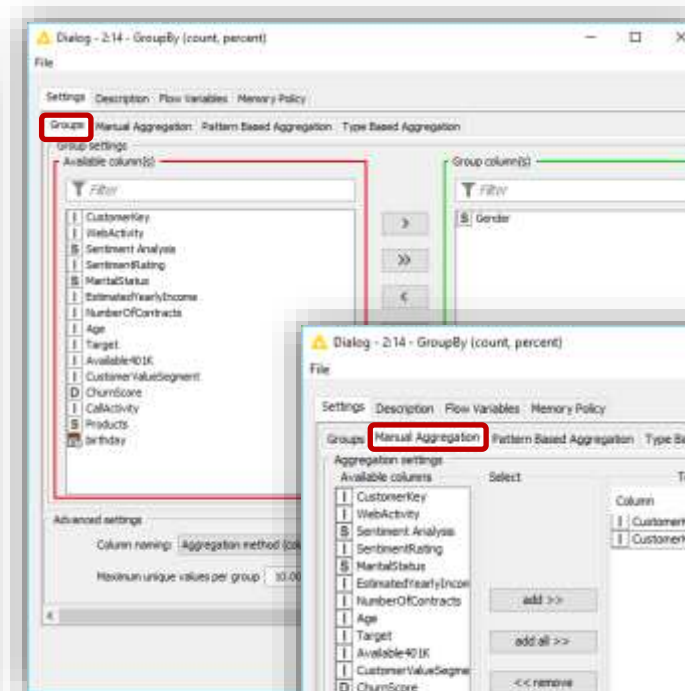
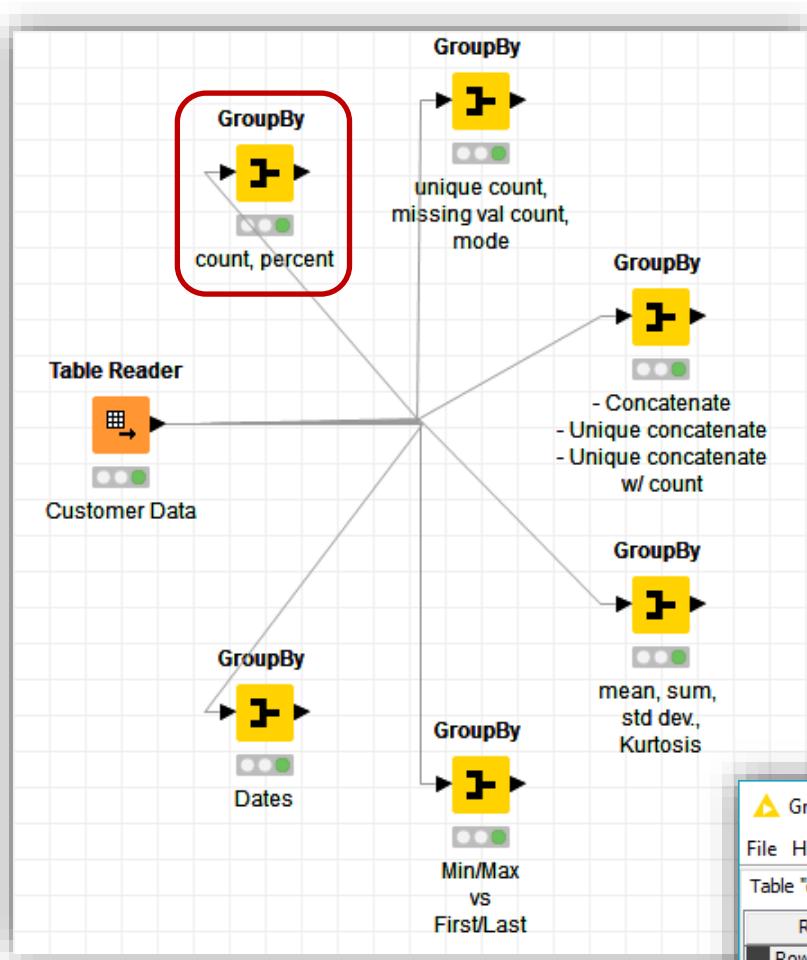
## Count and Percent

27

Data Exploration

DATA PREPARATION

Hands On



This screenshot shows the 'Group table - 2:14 - GroupBy (count, ...)' window, displaying the resulting data table with 2 rows and 3 columns.

Row ID	Gender	Percent...	Count(...)
Row0	F	49.984	7581
Row1	M	50.016	7586



# Basic Aggregations

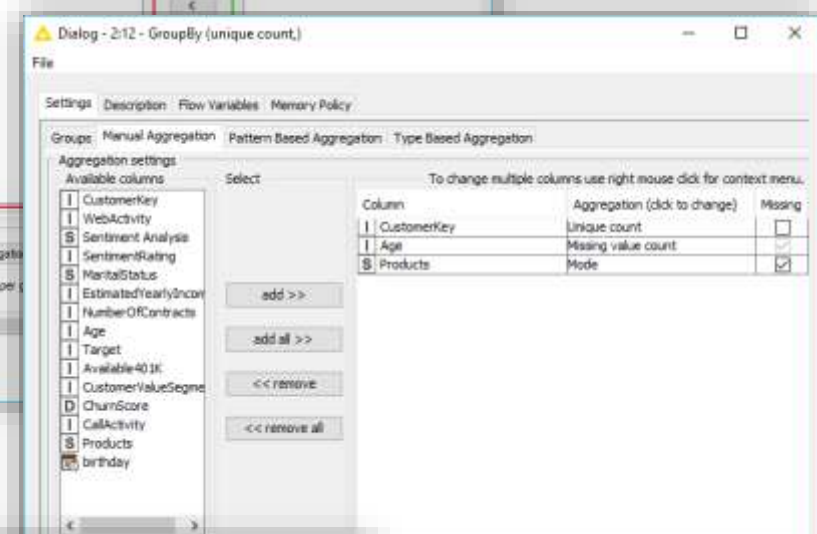
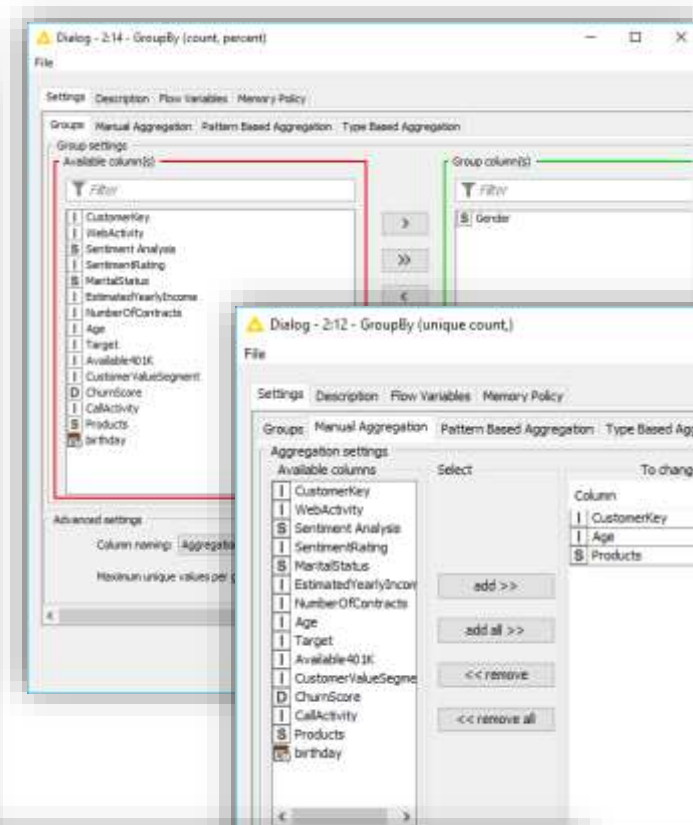
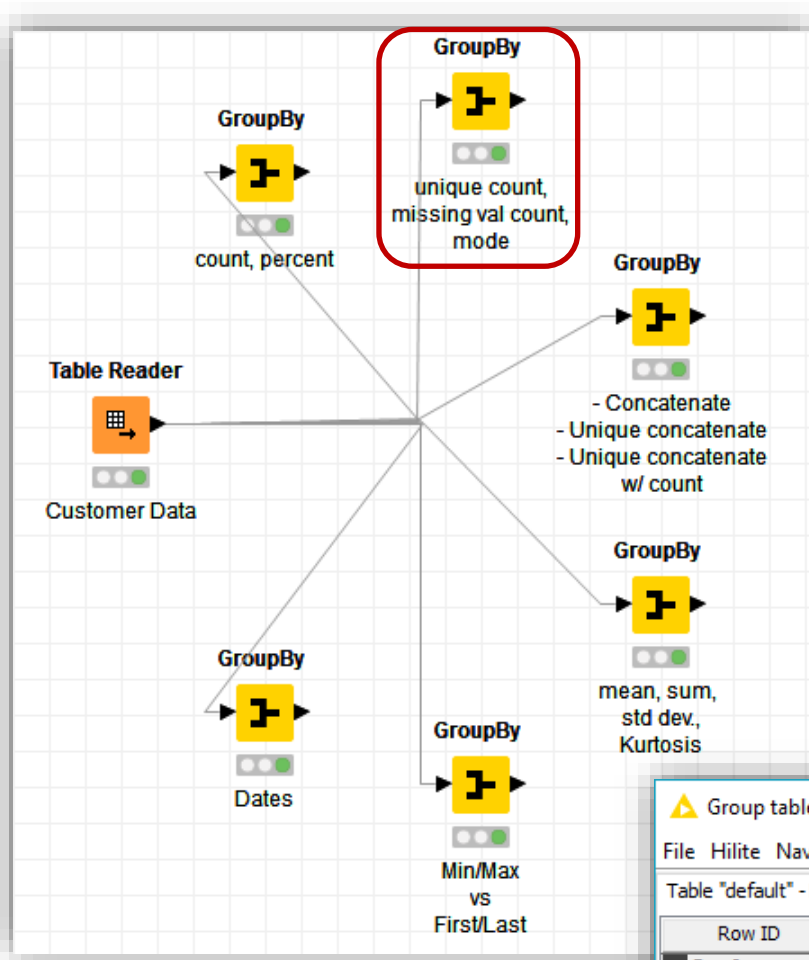
## Unique Count, Missing Values Count and Mode

28

Data Exploration

DATA PREPARATION

Hands On



Group table - 2:12 - GroupBy (unique count,)

File Hilite Navigation View

Table "default" - Rows: 2 Spec - Columns: 4 Properties Flow Variables

Row ID	S Gender	I Unique ...	I Missing ...	S Mode(Produ...
Row0	F	5763	0	private investment
Row1	M	5788	0	private investment

# Basic Aggregations

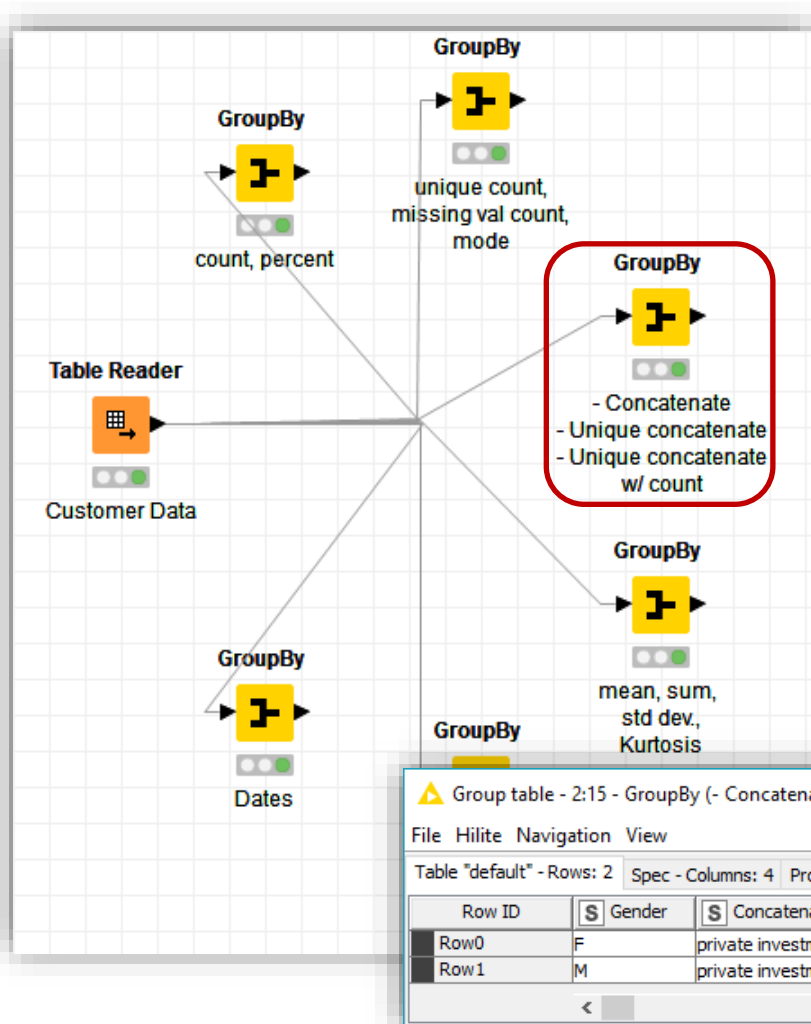
## Concatenate

29

Data Exploration

DATA PREPARATION

Hands On



Dialog - 2:14 - GroupBy (count, percent)

Settings | Description | Flow Variables | Memory Policy

Group settings: Manual Aggregation | Pattern Based Aggregation | Type Based Aggregation

Available column(s): CustomerKey, WebActivity, Sentiment Analysis, SentimentRating, MaritalStatus, EstimatedYearlyIncome, NumberOfContracts, Age, Target, Available401K, CustomerValueSegment, ChurnScore, CallActivity, Products, birthday

Group column(s): Filter, Gender

Dialog - 2:15 - GroupBy (- Concatenate)

Settings | Description | Flow Variables | Memory Policy

Group settings: Manual Aggregation | Pattern Based Aggregation | Type Based Aggregation

Aggregation settings: Available columns: CustomerKey, WebActivity, Sentiment Analysis, SentimentRating, MaritalStatus, EstimatedYearlyIncome, NumberOfContracts, Age, Target, Available401K, CustomerValueSegment, ChurnScore, CallActivity, Products, birthday

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

Column	Aggregation (click to change)	Missing
Products	Concatenate	<input checked="" type="checkbox"/>
Products	Unique concatenate	<input checked="" type="checkbox"/>
Products	Unique concatenate with count	<input checked="" type="checkbox"/>

Group table - 2:15 - GroupBy (- Concatenate)

File | Hilite | Navigation | View

Table "default" - Rows: 2 | Spec - Columns: 4 | Properties | Flow Variables

Row ID	Gender	Concatenate	Unique concatenate(Products)	Unique concatenate with count(Products)
Row0	F	private investme	private investment, p+b investment, gold...	private investment(2212), p+b investment(2139), gold inve...
Row1	M	private investme	private investment, p+b investment, gold...	private investment(2308), p+b investment(2009), gold inve...

# Basic Aggregations

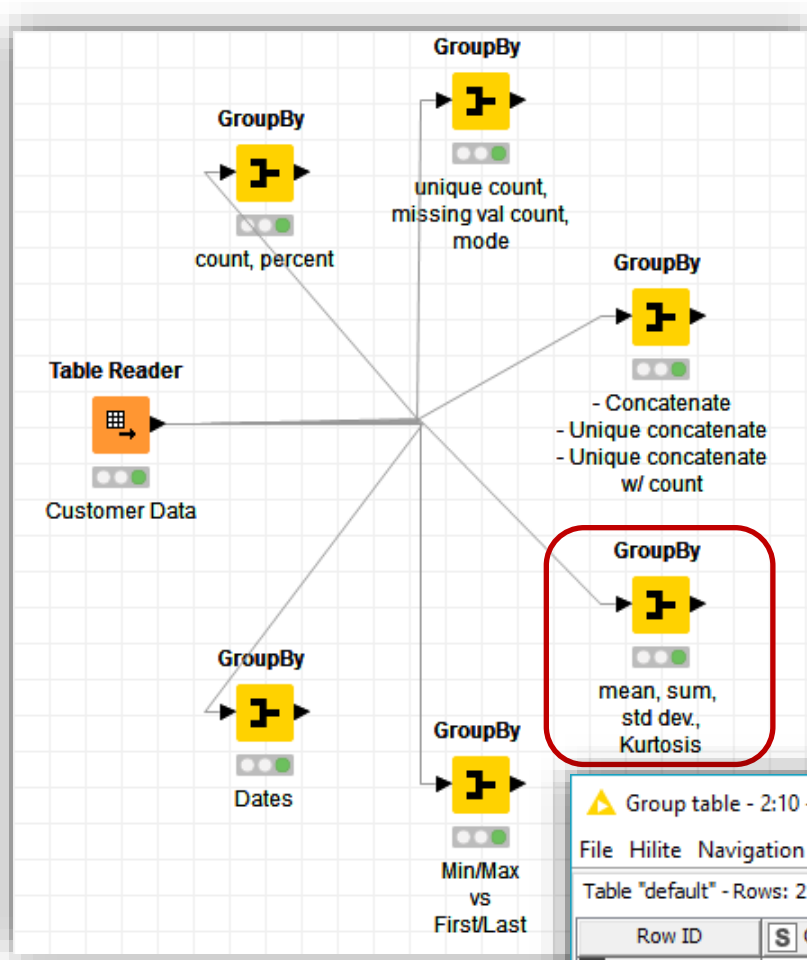
## Mean, Sum, Standard Deviation and Kurtosis

30

Data Exploration

DATA PREPARATION

Hands On



Dialog - 2:14 - GroupBy (count, percent)

Settings | Description | Flow Variables | Memory Policy

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

Group settings:

- Available column(s):
- Group column(s):

Advanced settings:

- Column naming: Aggregation
- Maximum unique values per group:

Dialog - 2:10 - GroupBy (mean, sum,)

Settings | Description | Flow Variables | Memory Policy

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

Aggregation settings:

- Available columns:
- Select:
- To change multiple columns use right mouse click for context menu.

Column	Aggregation (click to change)	Missing
EstimatedYearlyIncome	Mean	
NumberOfContracts	Sum	
EstimatedYearlyIncome	Standard deviation	
ChurnScore	Kurtosis	

Group table - 2:10 - GroupBy (mean, sum,)

File | Hilite | Navigation | View

Table "default" - Rows: 2 | Spec - Columns: 5 | Properties | Flow Variables

Row ID	Gender	Mean(E...	Sum(Nu...	Standa...	Kurtosi...
Row0	F	57,849.888	11110	31,609.743	0.243
Row1	M	57,586.343	11117	32,568.18	0.351

# Basic Aggregations

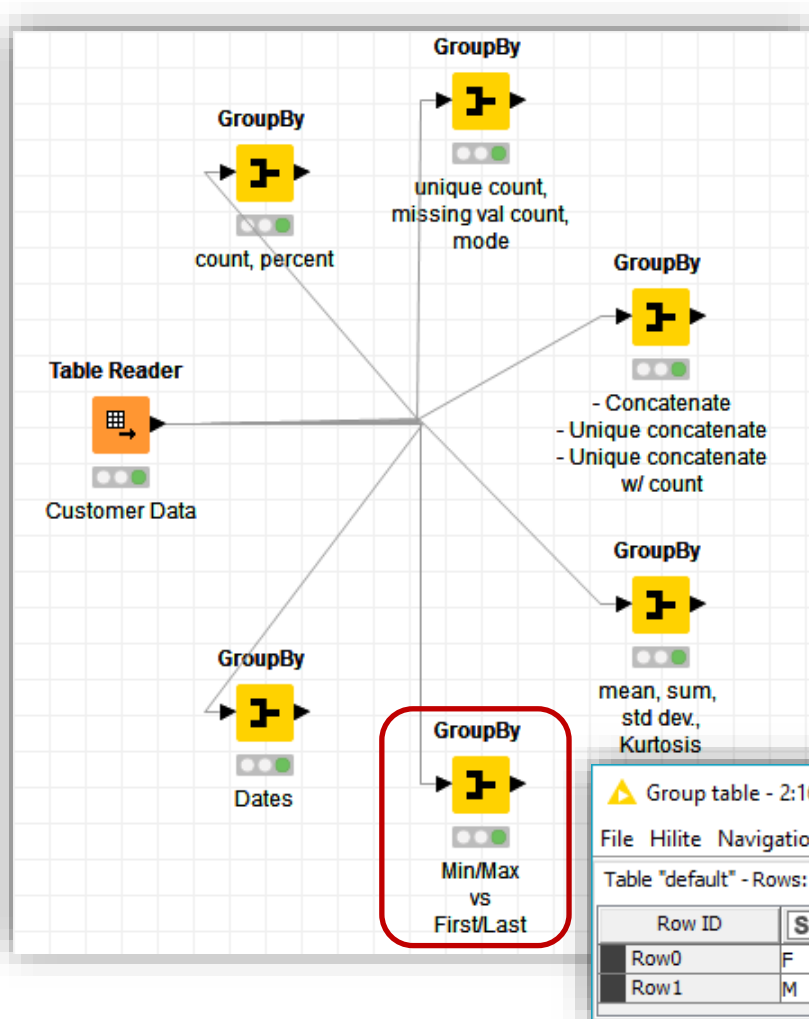
## Min/Max vs First/Last

31

Data Exploration

DATA PREPARATION

Hands On



The image shows two screenshots of the Alteryx GroupBy dialog box and a resulting data table.

**Dialog - 2:14 - GroupBy (count, percent)**

Available column(s): CustomerKey, WebActivity, Sentiment Analysis, Sentiment Rating, MaritalStatus, EstimatedYearlyIncome, NumberOfContracts, Age, Target, Available401K, CustomerValueSegment, ChurnScore, CallActivity, Products, birthday

Group column(s): Filter, Gender

**Dialog - 2:16 - GroupBy (Min/Max)**

Aggregation settings:

Column	Aggregation (click to change)	Missing
Age	Maximum	<input type="checkbox"/>
Age	Last	<input type="checkbox"/>
CustomerKey	Minimum	<input type="checkbox"/>
CustomerKey	First	<input type="checkbox"/>

**Group table - 2:16 - GroupBy (Min/Max)**

Table "default" - Rows: 2 Spec - Columns: 5 Properties Flow Variables

Row ID	Gender	Min*(Age)	Max*(Age)	First*(Age)	Last*(Age)
Row0	F	29	100	42	61
Row1	M	29	98	44	45

# Basic Aggregations

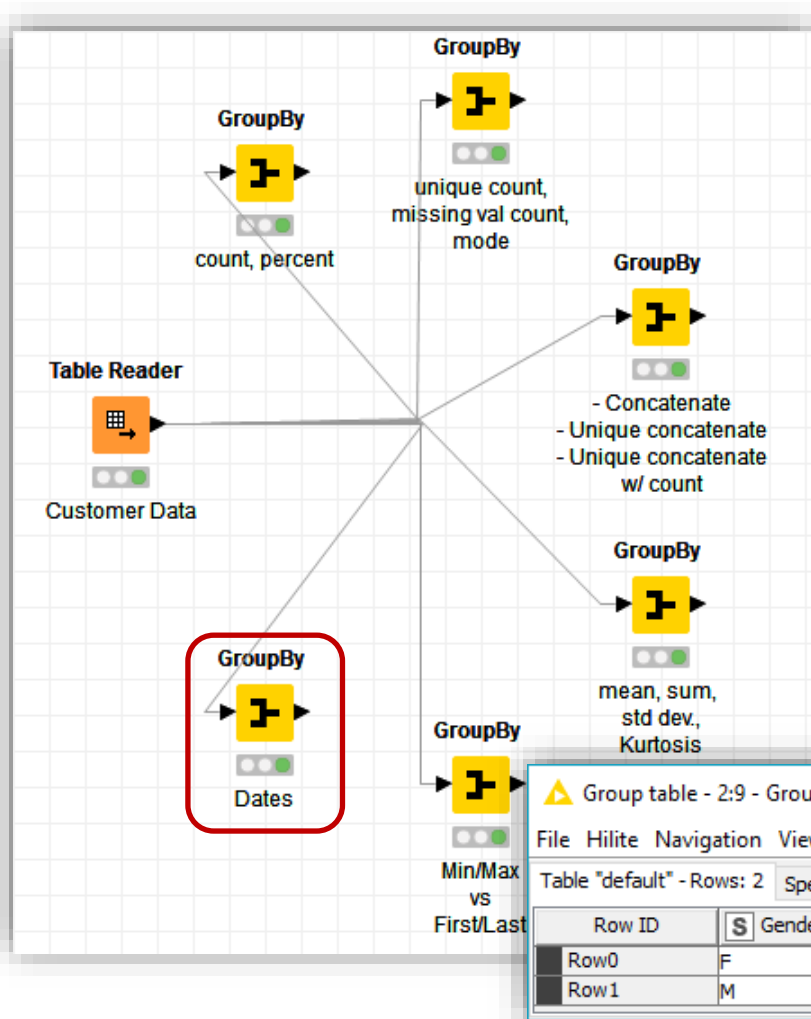
## Dates

32

Data Exploration

DATA PREPARATION

Hands On



Dialog - 2:14 - GroupBy (count, percent)

File Settings Description Flow Variables Memory Policy

Group: Manual Aggregation Pattern Based Aggregation Type Based Aggregation

Group settings

Available column(s)

Filter

Group column(s)

Filter

Advanced settings

Column naming: Aggregation

Maximum unique values per group

Dialog - 2:9 - GroupBy (Dates)

File Settings Description Flow Variables Memory Policy

Group: Manual Aggregation Pattern Based Aggregation Type Based Aggregation

Aggregation settings

Available columns

Select

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

Column	Aggregation (click to change)	Missing
birthday	Date range(day)	<input type="checkbox"/>
birthday	Mean date	<input type="checkbox"/>
birthday	Minimum	<input type="checkbox"/>
birthday	Maximum	<input type="checkbox"/>

add >>

add all >>

<< remove

<< remove all

Group table - 2:9 - GroupBy (Dates)

File Hilite Navigation View

Table "default" - Rows: 2 Spec - Columns: 5 Properties Flow Variables

Row ID	S Gender	D Date range(...)	Mean date...	Min*(birthday)	Max*(birthday)
Row0	F	26,115	29.ago.1967	27.set.1915	28.mar.1987
Row1	M	25,542	07.ago.1967	20.mai.1917	25.abr.1987

Cancel

# Advanced Aggregations

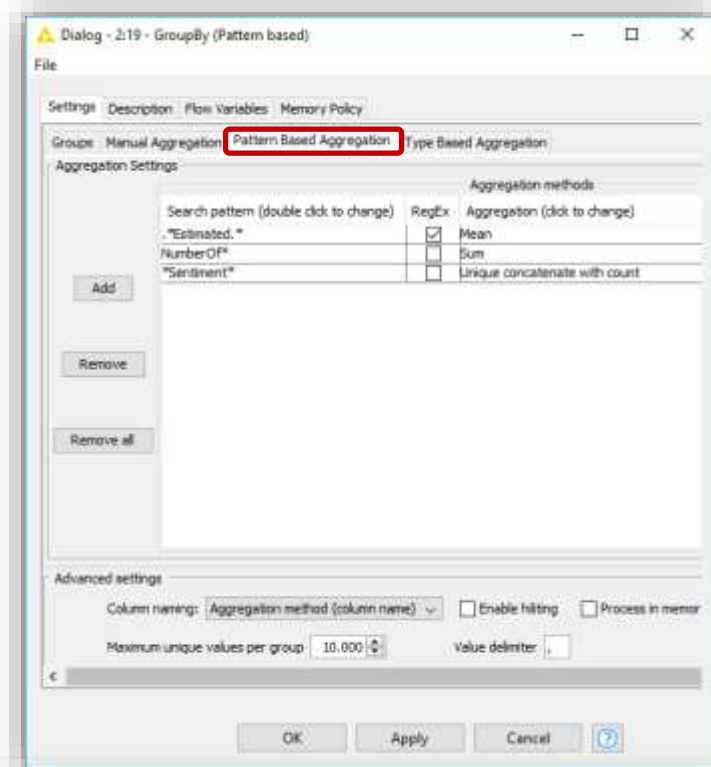
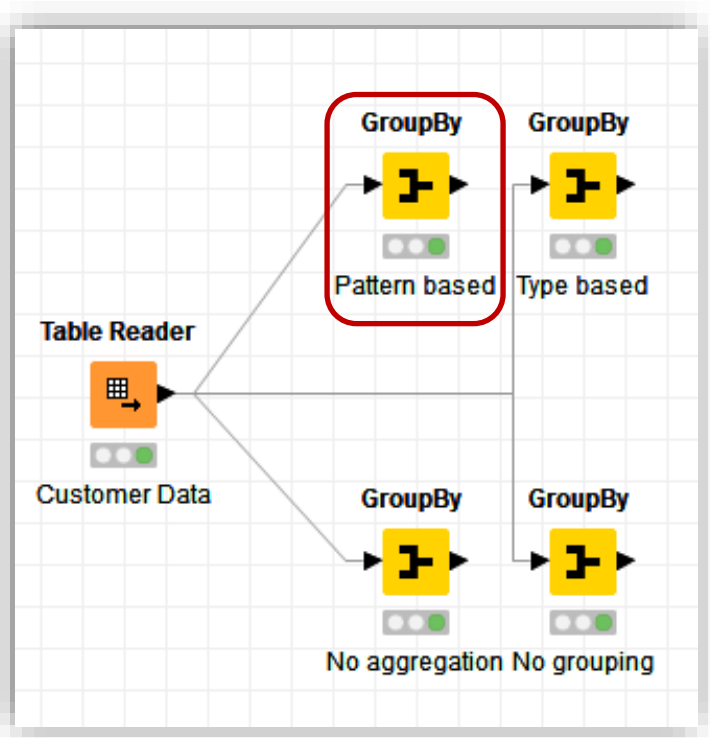
## Pattern Based

33

Data Exploration

DATA PREPARATION

Hands On



The screenshot shows the 'Group table - 2:19 - GroupBy (Pattern based)' window. It displays the resulting table structure with 5 columns and 2 rows.

Columns: 5	Column Type	Column Index	Color Handler
Gender	String	0	
Unique concatenate with count(Sentiment Analysis)	String	1	
Unique concatenate with count(SentimentRating)	String	2	
Mean(EstimatedYearlyIncome)	Number (do...	3	
Sum(NumberOfContracts)	Number (int...	4	



# Advanced Aggregations

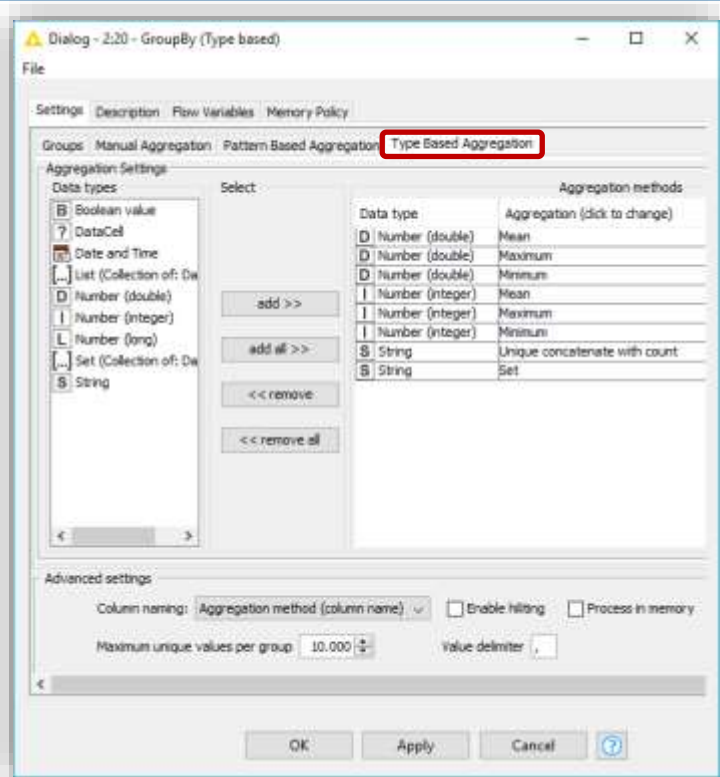
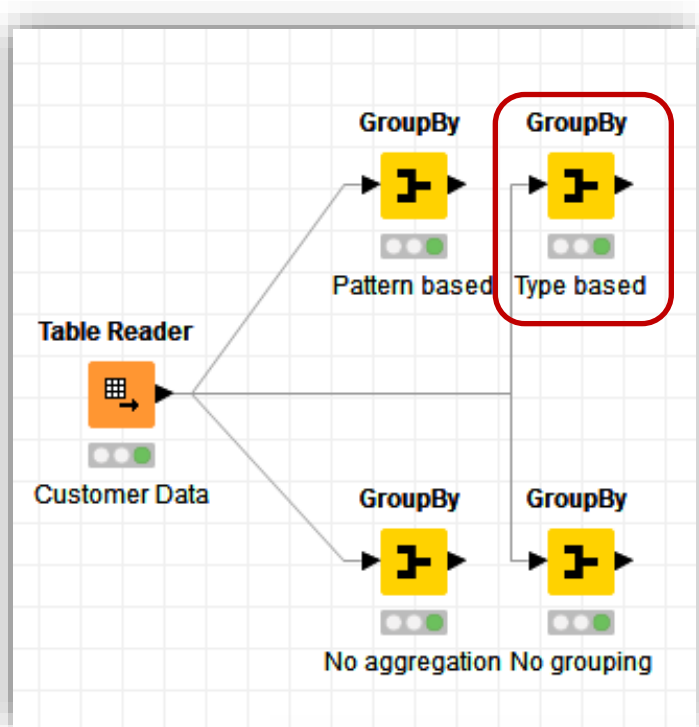
## Data Type Based

34

Data Exploration

DATA PREPARATION

Hands On



This screenshot shows the 'Group table - 2:20 - GroupBy (Type based)' window. The table displays the results of the aggregation, with columns for Row ID, Gender, and various statistical measures (Mean, Max, Min) for different data types (D for double, I for integer, S for string).

Row ID	S Gender	D Mean(C...	I Max*(C...	I Min*(C...	D Mean(C...	I Max*(C...	I Min*(C...	D Mean(...	I Max*(...	I Min*(W...	D Mean(...	I Max*(...	I Min*(W...	S
Row0	F	17,518.356	27333	11003	17,518.356	27333	11003	1.018	5	0	1.018	5	0	Ve
Row1	M	17,601.311	27336	11000	17,601.311	27336	11000	0.981	5	0	0.981	5	0	Sl

# Advanced Aggregations

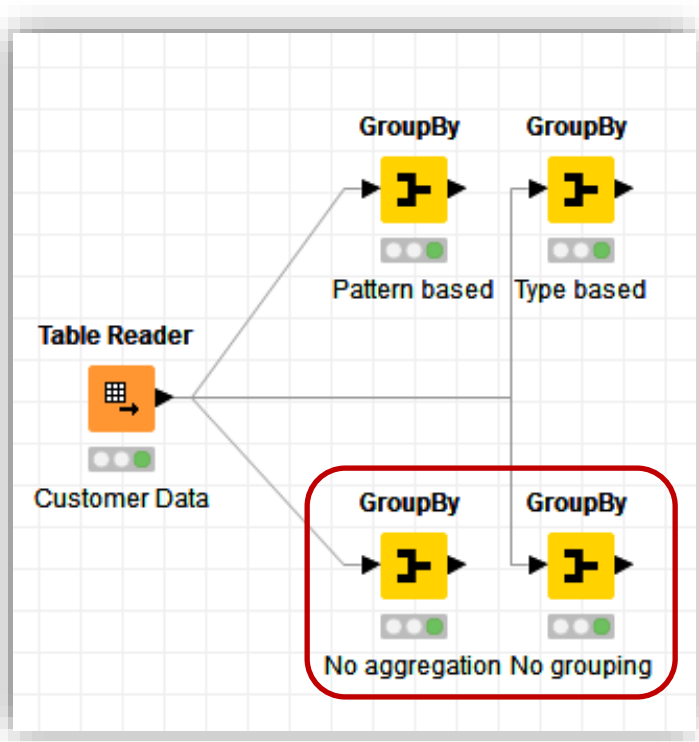
## No Aggregation vs No Grouping

35

Data Exploration

DATA PREPARATION

Hands On



Group table - 2:...

File Hilite Navigation View

Properties Flow Variables  
Table "default" - Rows: 12 Spec - Columns: 2

Row ID	S	Gen...	S	Sentim...
Row1	M			Negative
Row3	M			Positive
Row5	M			Slightly Neg...
Row7	M			Slightly Posit...
Row9	M			Very Negative
Row11	M			Very Positive
Row0	F			Negative
Row2	F			Positive
Row4	F			Slightly Neg...
Row6	F			Slightly Posit...
Row8	F			Very Negative
Row10	F			Very Positive

Group table - 2:18 - GroupBy (No grouping)

File Hilite Navigation View

Table "default" - Rows: 1 Spec - Columns: 3 Properties Flow Variables

Row ID	D	Mean(A...	I	Sum(Nu...	S	Unique concatenate with count(Sentiment Analysis)
Row0		48.203		22227		Slightly Negative(3023), Slightly Positive(1690), Very Negative(4173), Very Positive(1199), Positive(1960), Negative(3122)



# Advanced Data Preparation

36

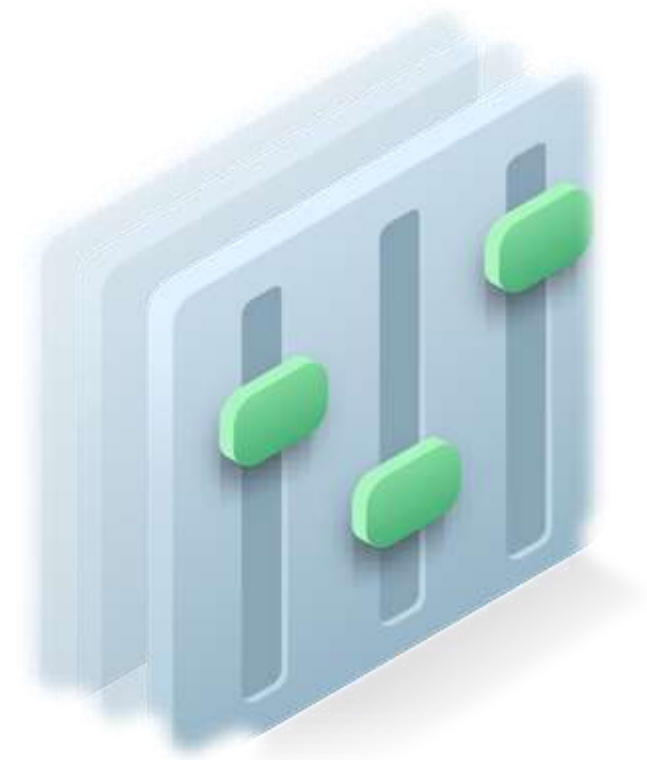
Data Exploration

**DATA PREPARATION**

Hands On

How?

1. Feature scaling
2. Outlier detection
3. Feature selection
4. Missing Values treatment
5. Nominal value discretization
6. Binning
7. Feature Engineering



# Feature scaling

37

Data Exploration

**DATA PREPARATION**

Hands On

## 1. Normalizing the range of the independent features

Rationale:

Many classifiers use **distance metrics** (ex.: Euclidean distance) and, if one feature has a broad range of values, the distance will be governed by this particular feature. Hence, the range should be normalized so that each feature may contribute proportionately to the final distance.



# Feature scaling

38

Data Exploration

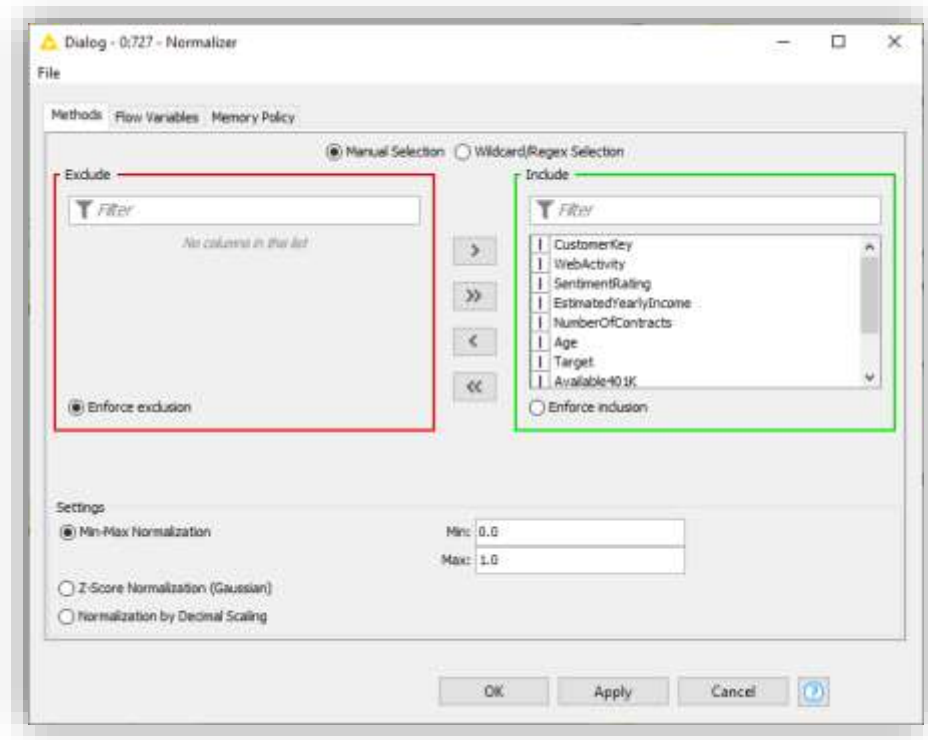
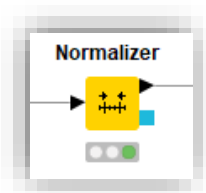
DATA PREPARATION

Hands On

1. Normalize the range of the independent features:

i. **Normalization**

Rescaling data so that all values fall within the range of 0 and 1, for example.



$$z = (b - a) \frac{x - \min(x)}{\max(x) - \min(x)} + a$$

# Feature scaling

39

Data Exploration

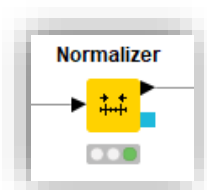
DATA PREPARATION

Hands On

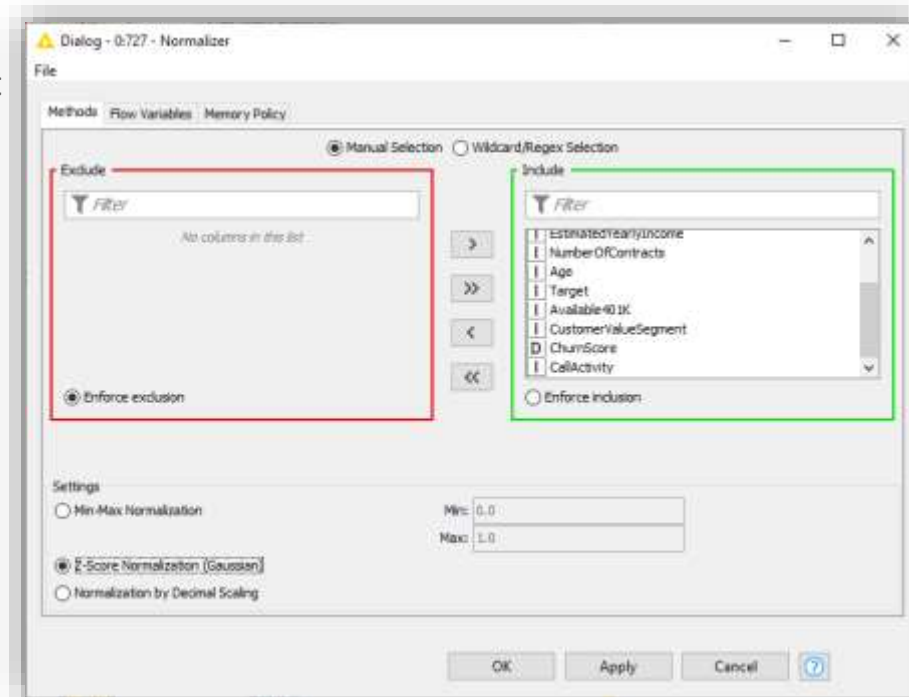
1. Normalize the range of the independent features:

ii. **Standardization** (or **Z-score Normalization**)

Rescaling the distribution of values so that the mean of observed values is 0 and the standard deviation is 1. Assumes observations fit a Gaussian distribution with a well-behaved mean and standard deviation, which may not always be the case.



$$z = \frac{x_i - \mu}{\sigma}$$



# Outlier Detection

40

Data Exploration

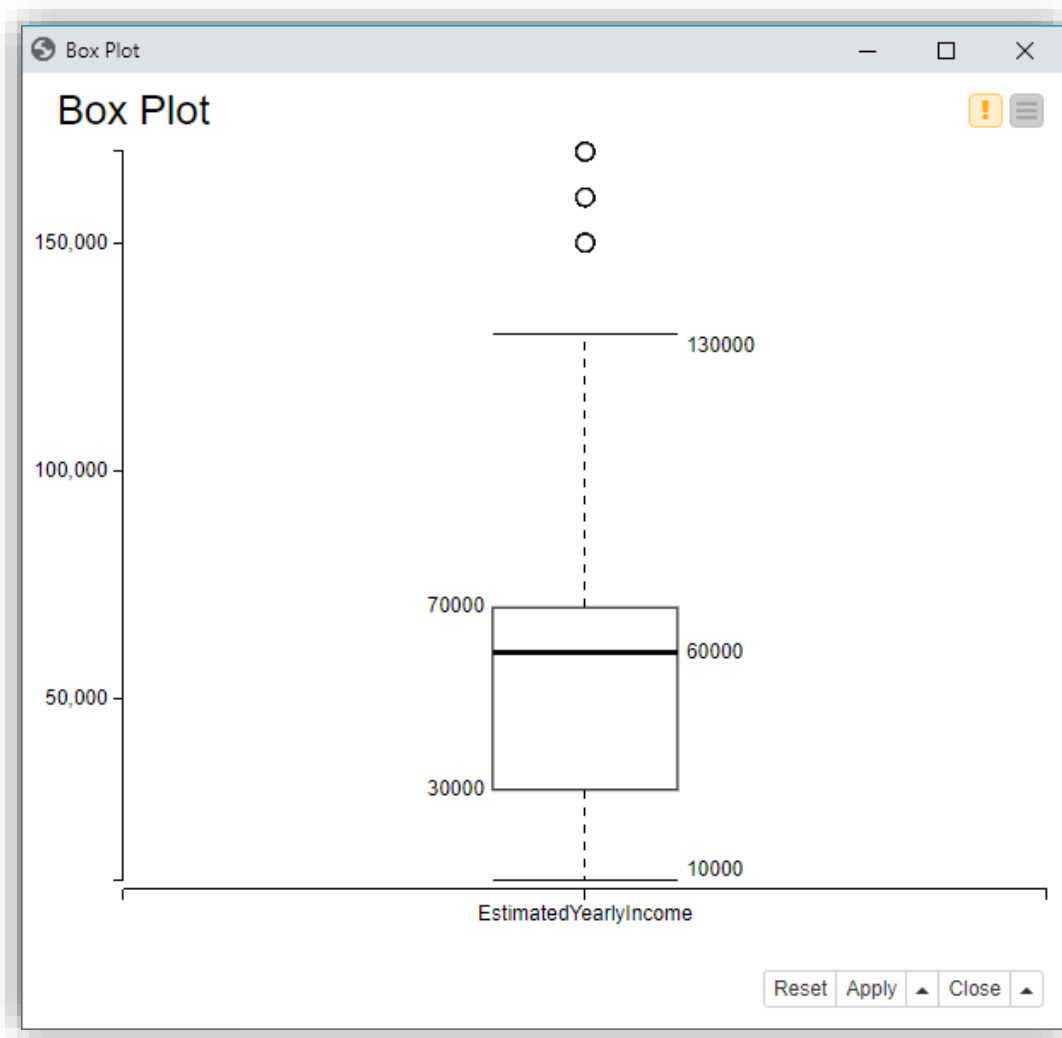
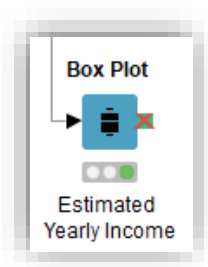
DATA PREPARATION

Hands On

## 2. Outlier detection:

### i. Statistical-based strategy

- Box Plots
- Z-Score (std. dev)



# Outlier Detection

41

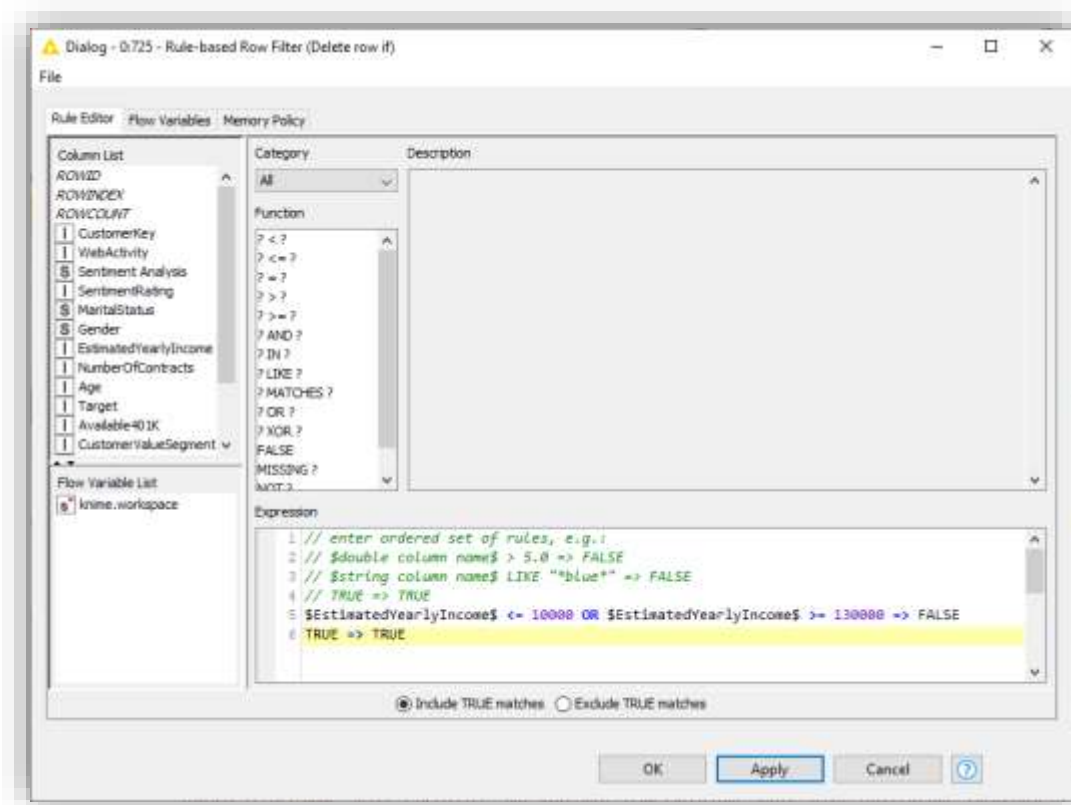
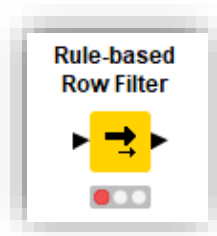
Data Exploration

DATA PREPARATION

Hands On

## 2. Outlier detection:

### ii. Knowledge-based strategy



# Outlier Detection

42

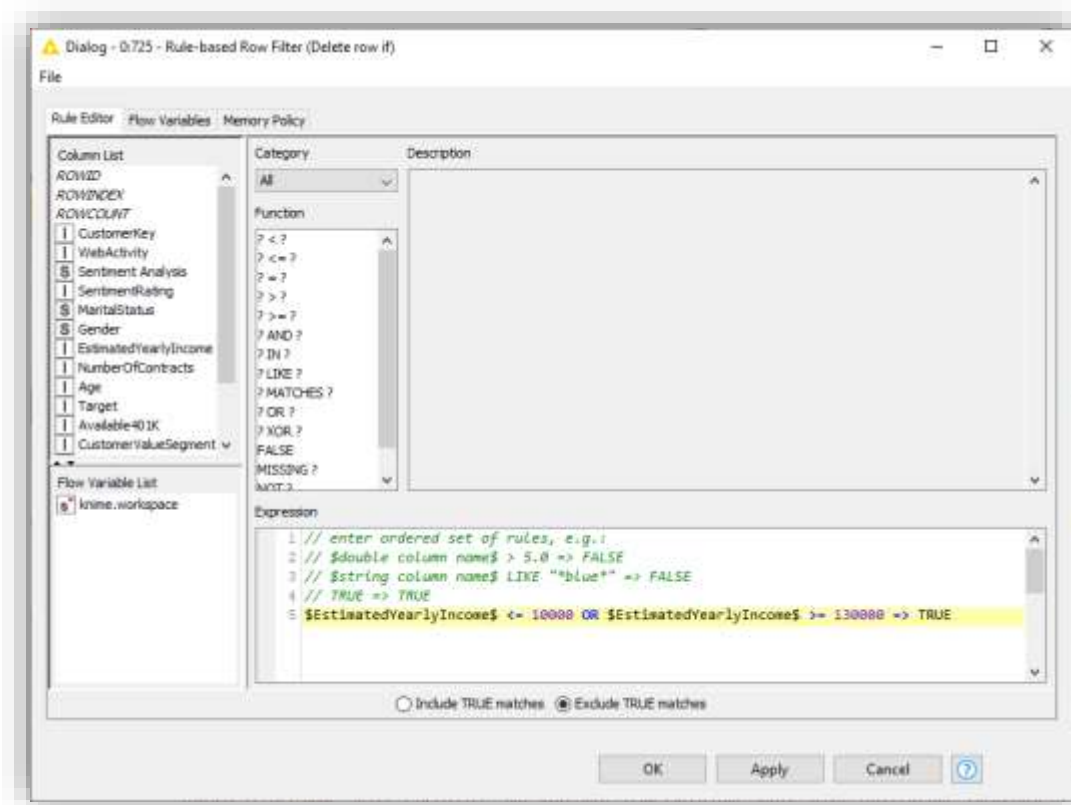
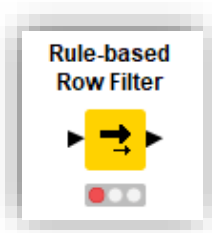
Data Exploration

DATA PREPARATION

Hands On

## 2. Outlier detection:

### ii. Knowledge-based strategy



# Outlier Detection

43

Data Exploration

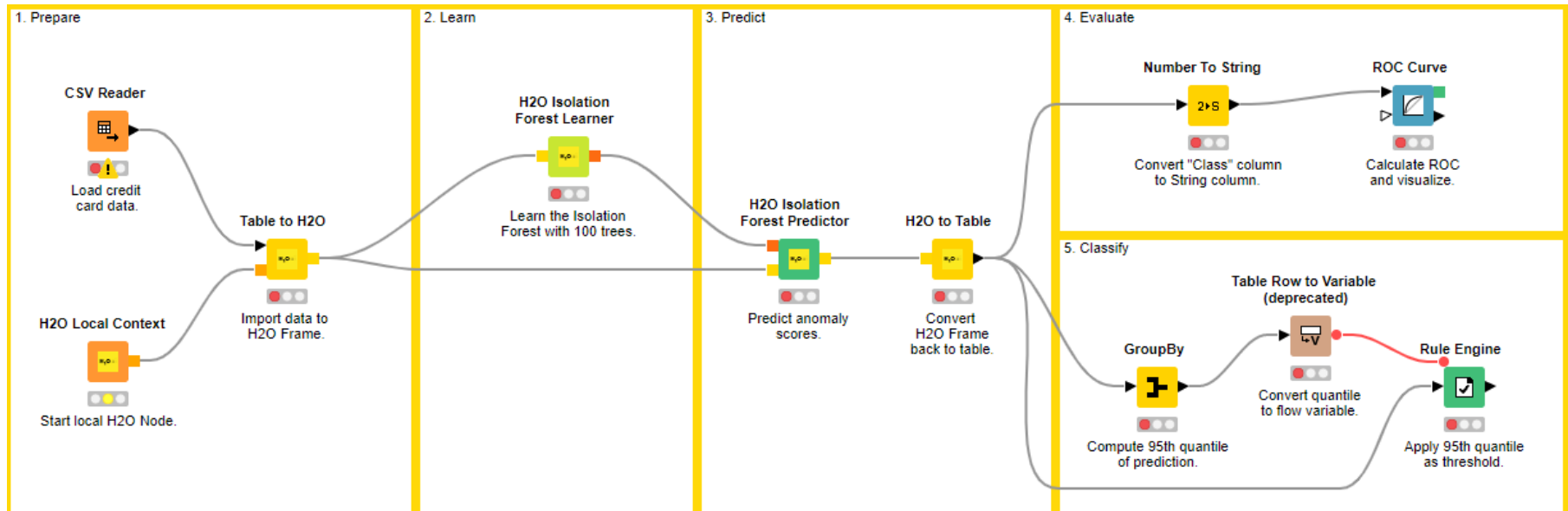
DATA PREPARATION

Hands On

## 2. Outlier detection:

### iii. Model-based strategy

- Isolation Forest
- One-Class SVM
- Minimum Covariance Determinant
- ...





# Outlier Detection

44

Data Exploration

**DATA PREPARATION**

Hands On

## 2. Outlier detection:

- i. Statistical-based strategy
- ii. Knowledge-based strategy
- iii. Model-based strategy

The Outlier Dilemma: Drop or Cap?

To keep the dataset size we may want to cap outliers instead of dropping them. However, it can affect the distribution of data!

# Feature Selection

45

Data Exploration

**DATA PREPARATION**

Hands On

## 3. Feature Selection (or dimensionality reduction)

Rationale:

Which features should we use to create a predictive model? Select a sub-set of the most important features to reduce dimensionality.

The removal of unimportant features:

- May **affect significantly the performance of a model**
- **Reduces overfitting** (less opportunity to make decisions based on noise)
- **Improves accuracy**
- Helps **reducing the complexity** of a model (reduces training time)

# Feature Selection

46

Data Exploration

**DATA PREPARATION**

Hands On

## 3. Feature Selection (or dimensionality reduction)

Rationale:

Which features should we use to create a predictive model? Select a sub-set of the most important features to reduce dimensionality.

The removal of unimportant features:

- May **affect significantly the performance of a model**
- **Reduces overfitting** (less opportunity to make decisions based on noise)
- **Improves accuracy**
- Helps **reducing the complexity** of a model (reduces training time)

What can we remove:

- **Redundant features** (duplicate)
- **Irrelevant and unneeded features** (non-useful)

Feature Selection Methods:

- **Filter methods**
- **Wrapper methods**
- **Embedded methods**

# Feature Selection

47

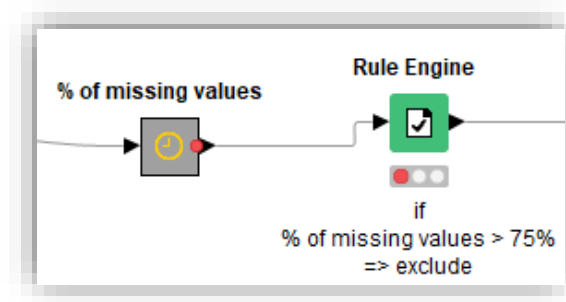
Data Exploration

DATA PREPARATION

Hands On

## 3. Filter Methods:

- i. Remove a feature if the **percentage of missing values** is **higher than** a threshold



- ii. Use the chi-square test to measure the **degree of dependency between a feature** and the **target class**
  - For each feature calculate  $X^2$
  - Normalize  $X^2$  and sort in descending order
  - Select  $n$  features with the highest importance (or those that are above the threshold)

# Feature Selection

48

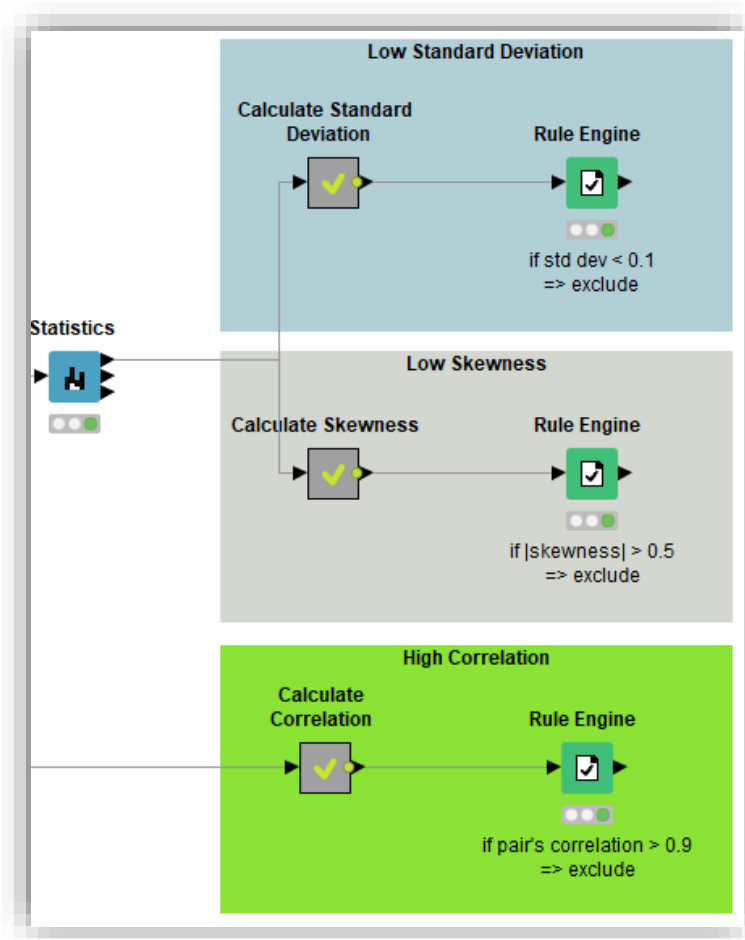
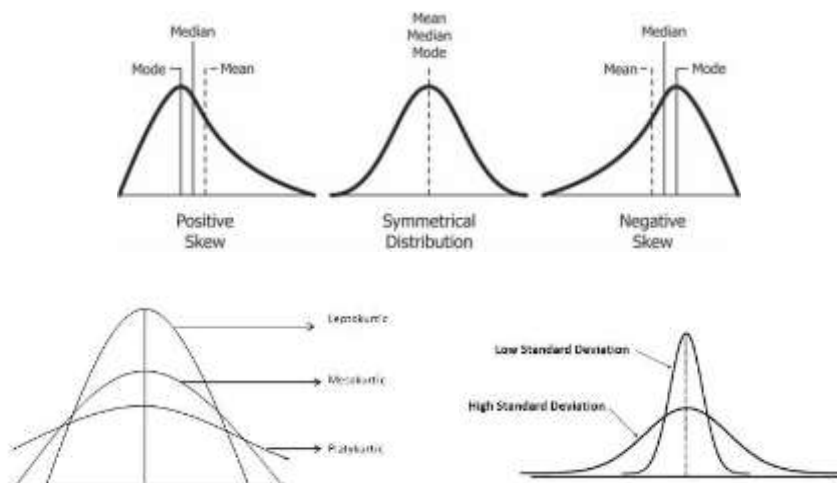
Data Exploration

DATA PREPARATION

Hands On

## 3. Filter Methods:

- iii. Remove feature if **low standard deviation**
- iv. Remove feature if data are **highly skewed**
- v. Remove features that are **highly correlated** between each other



# Feature Selection

49

Data Exploration

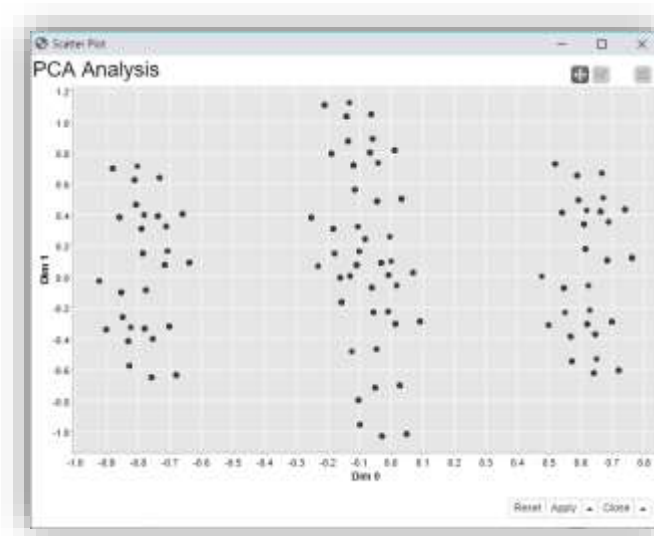
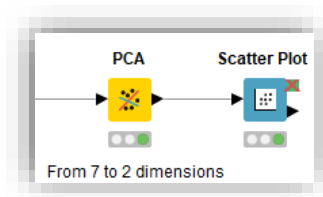
DATA PREPARATION

Hands On

## 3. Filter Methods:

### vi. Principal Component Analysis (PCA)

A technique to reduce the dimension of the feature space. The goal is to **reduce the number of features without losing too much information**. A popular application of PCA is for **visualizing higher dimensional data**.



# Feature Selection

50

# Data Exploration

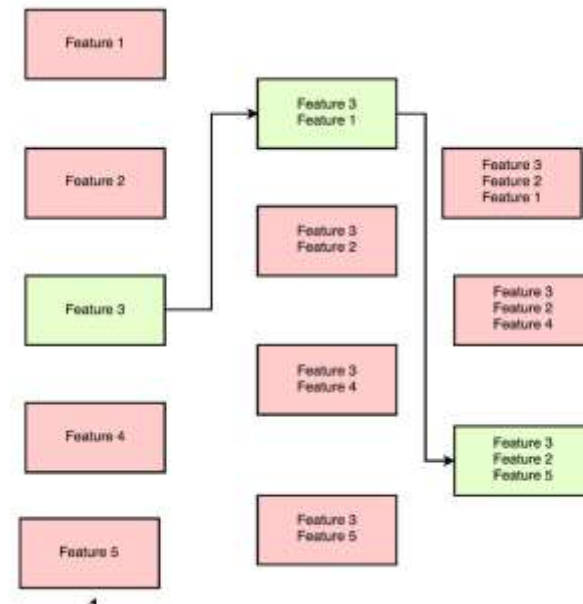
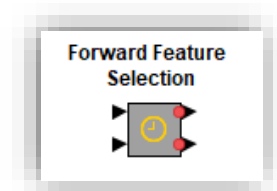
## DATA PREPARATION

## Hands On

### 3. Wrapper Methods:

Use a **ML algorithm** to select the most important features! Select a set of features as a search problem, prepare different combinations, evaluate and compare them!  
Measure the “usefulness” of features based on the classifier performance.

## vii. Sequential Forward Selection



# Feature Selection

51

Data Exploration

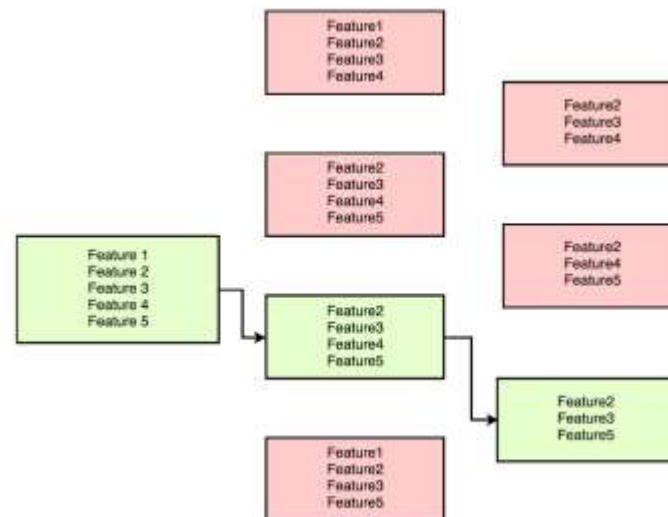
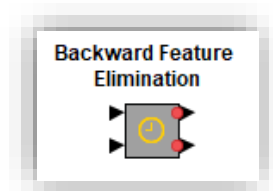
**DATA PREPARATION**

Hands On

## 3. Wrapper Methods:

Use a **ML algorithm** to select the most important features! Select a set of features as a search problem, prepare different combinations, evaluate and compare them! Measure the “usefulness” of features based on the classifier performance.

### vii. **Backward Feature Elimination**





# Feature Selection

52

Data Exploration

**DATA PREPARATION**

Hands On

## 3. Embedded Methods:

Algorithms that already have built-in feature selection methods.

Lasso, for example, has their own feature selection methods.

For example, if a feature's weight is zero then it has no importance!

Regularization - constrain/regularize or shrink the coefficient estimates towards zero!

# Missing Values

53

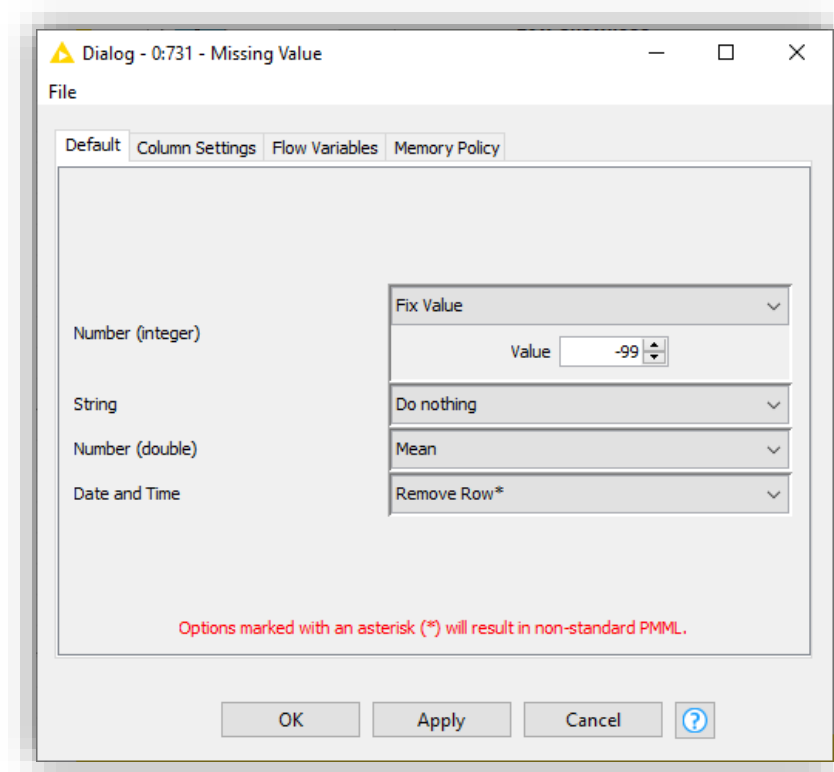
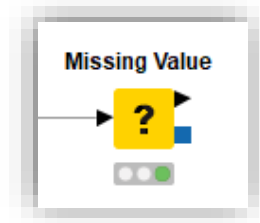
Data Exploration

DATA PREPARATION

Hands On

## 4. Missing Values treatment:

- i. First **analyse each feature** in regard to the **number and percentage of missing values**
- ii. Then decide what to do:
  - Remove
  - Mean
  - (Linear/...) Interpolation
  - Mask
  - ...



# Nominal Value Discretization/Encoding

54

Data Exploration

**DATA PREPARATION**

Hands On

## 5. Nominal value discretization:

Rationale:

**Categorical data** often called nominal data, are variables that **contain label values rather than numeric ones**. Several methods may be applied:

- One-Hot Encoding
- Label Encoding
- Binary Encoding

# Nominal Value Discretization/Encoding

55

Data Exploration

**DATA PREPARATION**

Hands On

Movie	Genre
Jumanji	Adventure
American Pie	Comedy
Braveheart	Drama
...	...

# Nominal Value Discretization/Encoding

56

Data Exploration

**DATA PREPARATION**

Hands On

Movie	Genre
Jumanji	Adventure
American Pie	Comedy
Braveheart	Drama
...	...

**Label Encoded**

Movie	Genre	Category
Jumanji	Adventure	0
American Pie	Comedy	1
Braveheart	Drama	2
...	...	

# Nominal Value Discretization/Encoding

57

Data Exploration

**DATA PREPARATION**

Hands On

Movie	Genre
Jumanji	Adventure
American Pie	Comedy
Braveheart	Drama
...	...

One-Hot Encoded

Movie	Adventure	Comedy	Drama
Jumanji	1	0	0
American Pie	0	1	0
Braveheart	0	0	1
...	...		

# Nominal Value Discretization/Encoding

58

Data Exploration

DATA PREPARATION

Hands On

Movie	Genre
Jumanji	Adventure
American Pie	Comedy
Braveheart	Drama
...	...

Label Encoded

Movie	Genre	Category
Jumanji	Adventure	0
American Pie	Comedy	1
Braveheart	Drama	2
...	...	

Integer values have a natural ordered relationship between each other. ML models may be able to understand such relationships.

One-Hot Encoded

Movie	Adventure	Comedy	Drama
Jumanji	1	0	0
American Pie	0	1	0
Braveheart	0	0	1
...	...		

Categorical features where no such ordinal relationship exists. However, for a huge number of categories...

# Nominal Value Discretization/Encoding

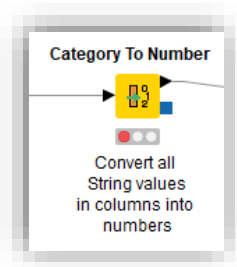
59

Data Exploration

**DATA PREPARATION**

Hands On

## 5. Nominal value discretization:



Dialog - 3:145:101 - Category To Number

File

Columns to transform | Flow Variables | Memory Policy

☒ Manual Selection ☐ Wildcard/Regex Selection

**Exclude**

Filter

No columns in this list

☒ Enforce exclusion

**Include**

Filter

S title  
S genres

☐ Enforce inclusion

☒ Append columns

Column suffix: (to number)

Start value: 0

Increment: 1

Max. categories: 100

Default value:

Map missing to:

OK Apply Cancel ?



# Nominal Value Discretization/Encoding

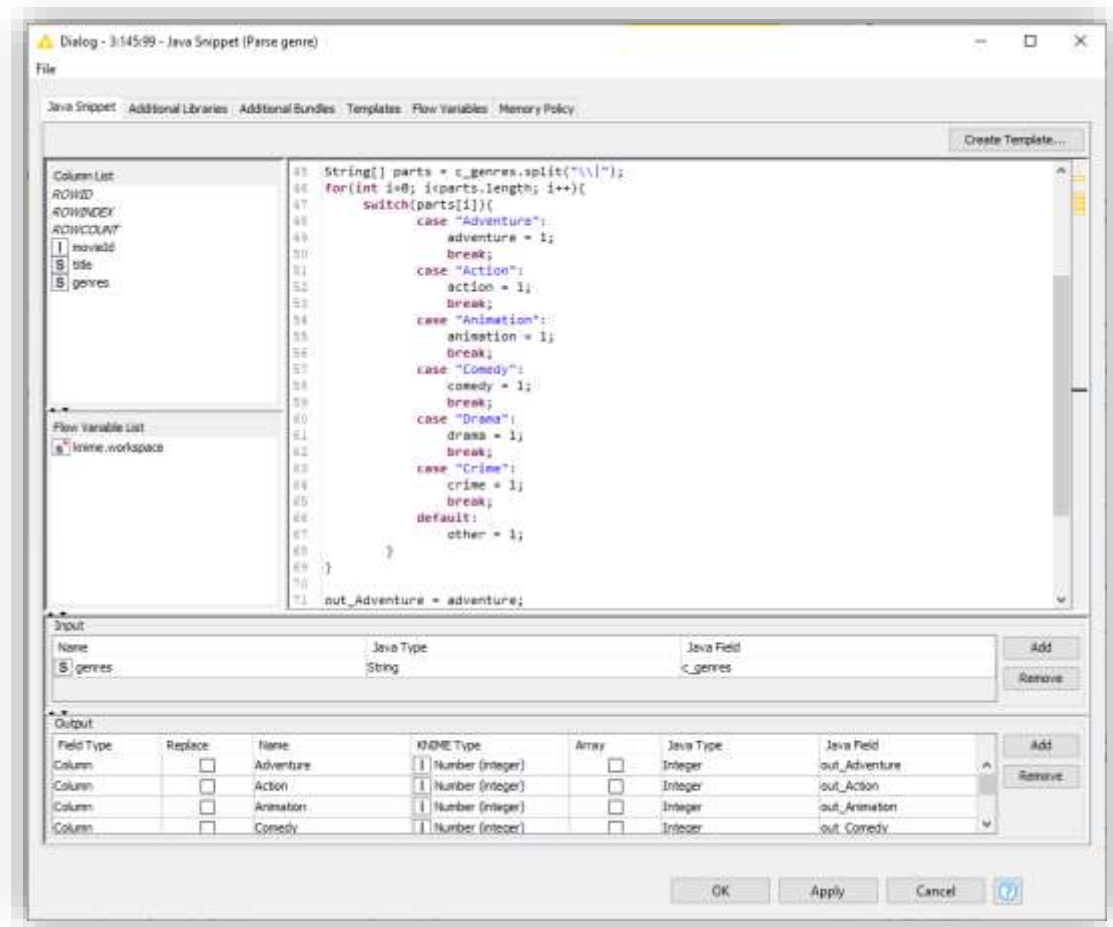
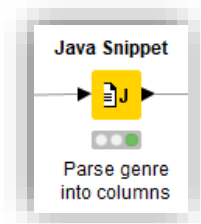
60

Data Exploration

DATA PREPARATION

Hands On

## 5. Nominal value discretization:



# Binning

61

Data Exploration

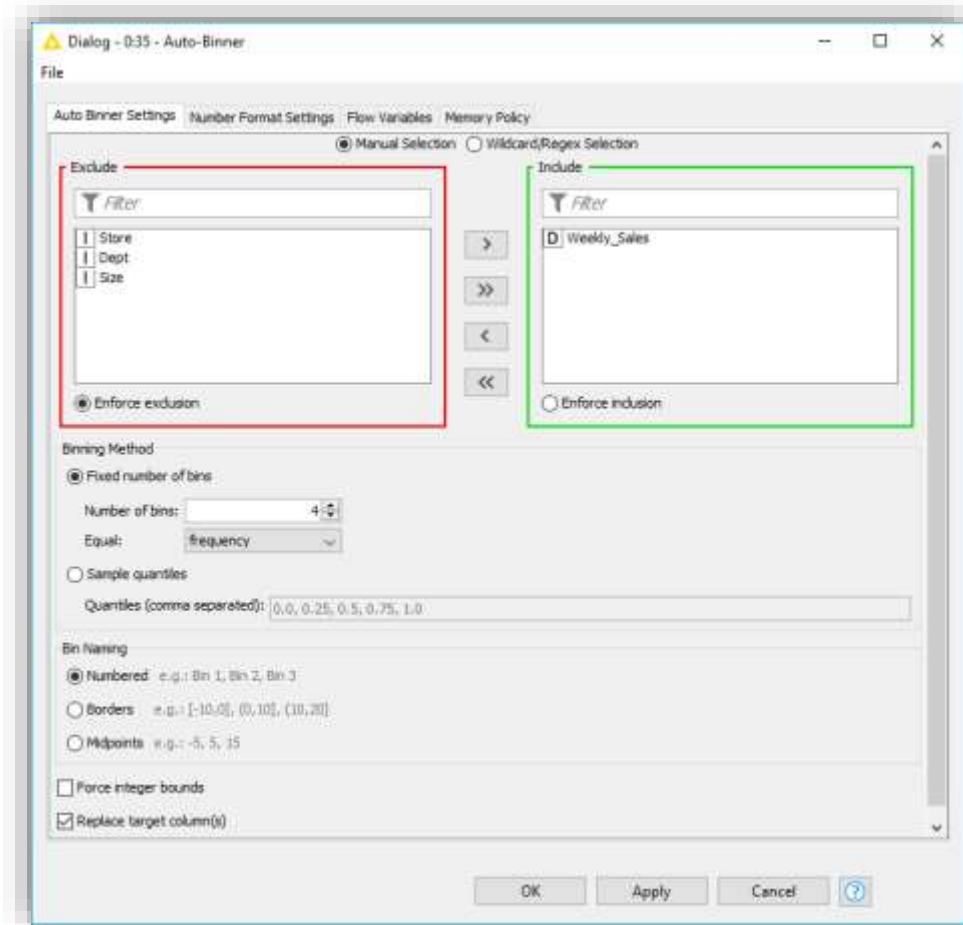
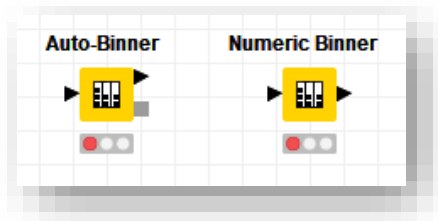
DATA PREPARATION

Hands On

## 6. Binning, i.e., group numeric data into intervals - called bins:

Rationale:

Make the model **more robust** and **prevent overfitting**. However, it **penalizes the model's performance** since every time you bin something, you sacrifice information.



# Binning

62

Data Exploration

DATA PREPARATION

Hands On

6. Binning, i.e., **group numeric data into intervals** - called **bins**:

Binned Data - 0:733 - Auto-Binner (Age into 4 bins)

File Hilite Navigation View

Table "default" - Rows: 15167 Spec - Columns: 16 Properties Flow Variables

Row ID	Custom...	WebAc...	Sentiment...	Sentim...	Marital...	Gender	Estim...	Number...	Age	Target
Row0_Row0_...	11000	0	Slightly Negative	2	M	M	90000	0	(39,46]	1
Row0_Row86...	11000	0	Slightly Negative	2	M	M	90000	0	(39,46]	1
Row1_Row1_...	11001	3	Slightly Positive	3	S	M	60000	1	(39,46]	1
Row1_Row86...	11001	3	Slightly Positive	3	S	M	60000	1	(39,46]	1
Row2_Row2_...	11002	3	Slightly Positive	3	S	M	60000	1	(39,46]	1
Row2_Row86...	11002	3	Slightly Positive	3	S	M	60000	1	(39,46]	1
Row3_Row3_...	11003	0	Very Negative	0	S	M	60000	1	(39,46]	1
Row3_Row86...	11003	0	Very Negative	0	S	M	60000	1	(39,46]	1
Row4_Row4_...	11004	5	Very Positive	5	S	M	60000	1	(39,46]	1
Row4_Row86...	11004	5	Very Positive	5	S	M	60000	1	(39,46]	1
Row5_Row5_...	11005	0	Very Negative	0	S	M	60000	1	(39,46]	1
Row5_Row86...	11005	0	Very Negative	0	S	M	60000	1	(39,46]	1
Row6_Row6_...	11006	0	Very Negative	0	S	M	60000	1	(39,46]	1
Row6_Row86...	11006	0	Very Negative	0	S	M	60000	1	(39,46]	1
Row7_Row7_...	11007	3	Slightly Positive	3	M	M	60000	2	(39,46]	1
Row7_Row87...	11007	3	Slightly Positive	3	M	M	60000	2	(39,46]	1
Row8_Row8_...	11008	4	Positive	4	S	F	60000	3	(39,46]	1
Row8_Row87...	11008	4	Positive	4	S	F	60000	3	(39,46]	1
Row9_Row9_...	11009	0	Very Negative	0	S	M	70000	1	(39,46]	1
Row9_Row87...	11009	0	Very Negative	0	S	M	70000	1	(39,46]	1
Row10_Row1...	11010	0	Very Negative	0	S	F	70000	1	(39,46]	1
Row10_Row8...	11010	0	Very Negative	0	S	F	70000	1	(39,46]	1
Row11_Row1...	11011	4	Positive	4	M	M	60000	4	(39,46]	1
Row11_Row8...	11011	4	Positive	4	M	M	60000	4	(39,46]	1

Possible Values

- [29,39]
- (39,46]
- (46,55]
- (55,100]

OK

# Feature Engineering

63

Data Exploration

**DATA PREPARATION**

Hands On

## 7. Feature Engineering:

Rationale:

The process of creating new features! The goal is to improve the performance of ML models.

Example: from the creation date of an observation what can we extract?

**2020-10-29 16h30**

# Feature Engineering

64

Data Exploration

**DATA PREPARATION**

Hands On

## 7. Feature Engineering:

Rationale:

The process of creating new features! The goal is to improve the performance of ML models.

Example: from the creation date of an observation what can we extract?

**2020-10-29 16h30**

We may extract new features such as:

- Year, month and day
- Hour and minutes
- Day of week (Thursday)
- Is Weekend? (No)
- Is Holiday? (No)
- ...

# Feature Engineering

65

Data Exploration

**DATA PREPARATION**

Hands On

## 7. Feature Engineering:

Rationale:

The process of creating new features! The goal is to improve the performance of ML models.

Example: from the geographic coordinates of a road

**(41.561859, -8.397455)**

# Feature Engineering

66

Data Exploration

**DATA PREPARATION**

Hands On

## 7. Feature Engineering:

Rationale:

The process of creating new features! The goal is to improve the performance of ML models.

Example: from the geographic coordinates of a road

**(41.561859, -8.397455)**

We may extract new features such as:

- Number of roads in the vicinity
- Are there schools nearby?
- ...

# Feature Engineering

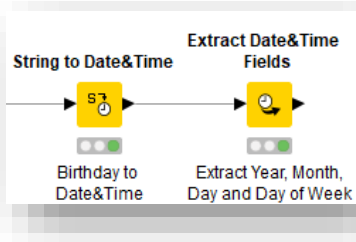
67

Data Exploration

DATA PREPARATION

Hands On

## 7. Feature Engineering:



Output table - 0:734 - Extract Date&Time Fields (Extract Year, Month,)

File Hilite Navigation View

Table "default" - Rows: 15167 Spec - Columns: 20 Properties Flow Variables

Row ID	..	D ChurnS...	I CallActi...	S Products	31 birthday	I Year	I Month (number)	I Day of year	S Day of week (name)
Row0_Row0_...	0.1	4	private investment	1972-01-14	1972	1	14	Sexta-feira	
Row0_Row86...	0.1	4	private investment	1971-08-28	1971	8	240	Sábado	
Row1_Row1_...	0	4	private investment	1970-06-26	1970	6	177	Sexta-feira	
Row1_Row86...	0	4	private investment	1971-02-11	1971	2	42	Quinta-feira	
Row2_Row2_...	0.2	4	private investment	1971-01-27	1971	1	27	Quarta-feira	
Row2_Row86...	0.2	4	private investment	1971-02-17	1971	2	48	Quarta-feira	
Row3_Row3_...	0.5	4	private investment	1973-11-07	1973	11	311	Quarta-feira	
Row3_Row86...	0.5	4	private investment	1974-02-14	1974	2	45	Quinta-feira	
Row4_Row4_...	0.1	4	private investment	1973-09-21	1973	9	264	Sexta-feira	
Row4_Row86...	0.1	4	private investment	1974-01-02	1974	1	2	Quarta-feira	
Row5_Row5_...	0.5	4	private investment	1970-06-05	1970	6	156	Sexta-feira	
Row5_Row86...	0.5	4	private investment	1970-05-06	1970	5	126	Quarta-feira	
Row6_Row6_...	0.5	4	private investment	1971-07-29	1971	7	210	Quinta-feira	
Row6_Row86...	0.5	4	private investment	1972-01-19	1972	1	19	Quarta-feira	
Row7_Row7_...	0	4	private investment	1970-01-03	1970	1	3	Sábado	
Row7_Row87...	0	4	private investment	1970-02-07	1970	2	38	Sábado	
Row8_Row8_...	1	4	private investment	1970-03-12	1970	3	71	Quinta-feira	
Row8_Row87...	1	4	private investment	1969-08-04	1969	8	216	Segunda-feira	
Row9_Row9_...	0.5	4	private investment	1969-07-21	1969	7	202	Segunda-feira	
Row9_Row87...	0.5	4	private investment	1969-10-28	1969	10	301	Terça-feira	
Row10_Row1...	0.5	4	private investment	1969-11-06	1969	11	310	Quinta-feira	



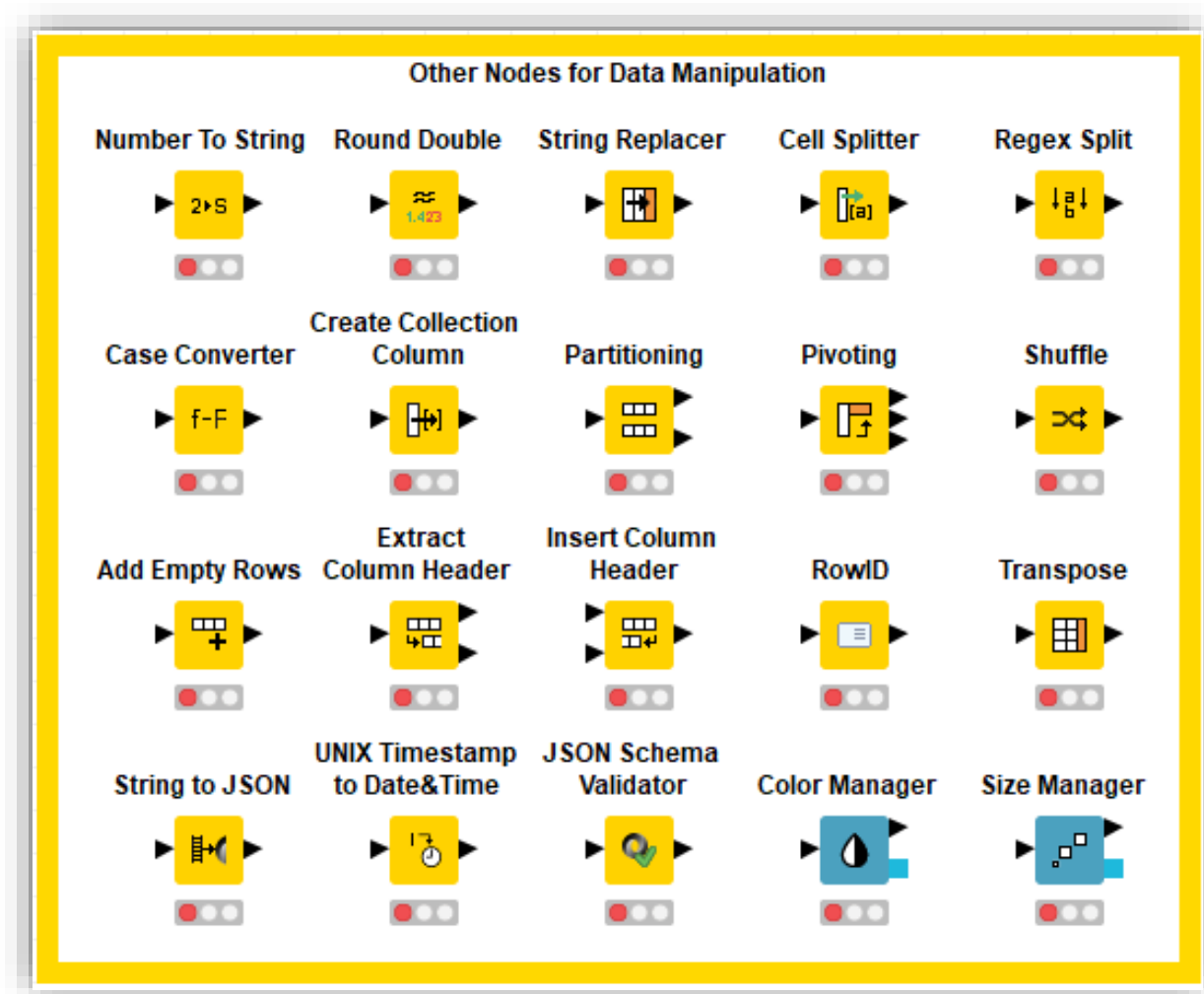
# More Nodes

68

Data Exploration

DATA PREPARATION

Hands On



# Hands On

69

Data Exploration

Data Preparation

**HANDS ON**

**HANDS ON**

