

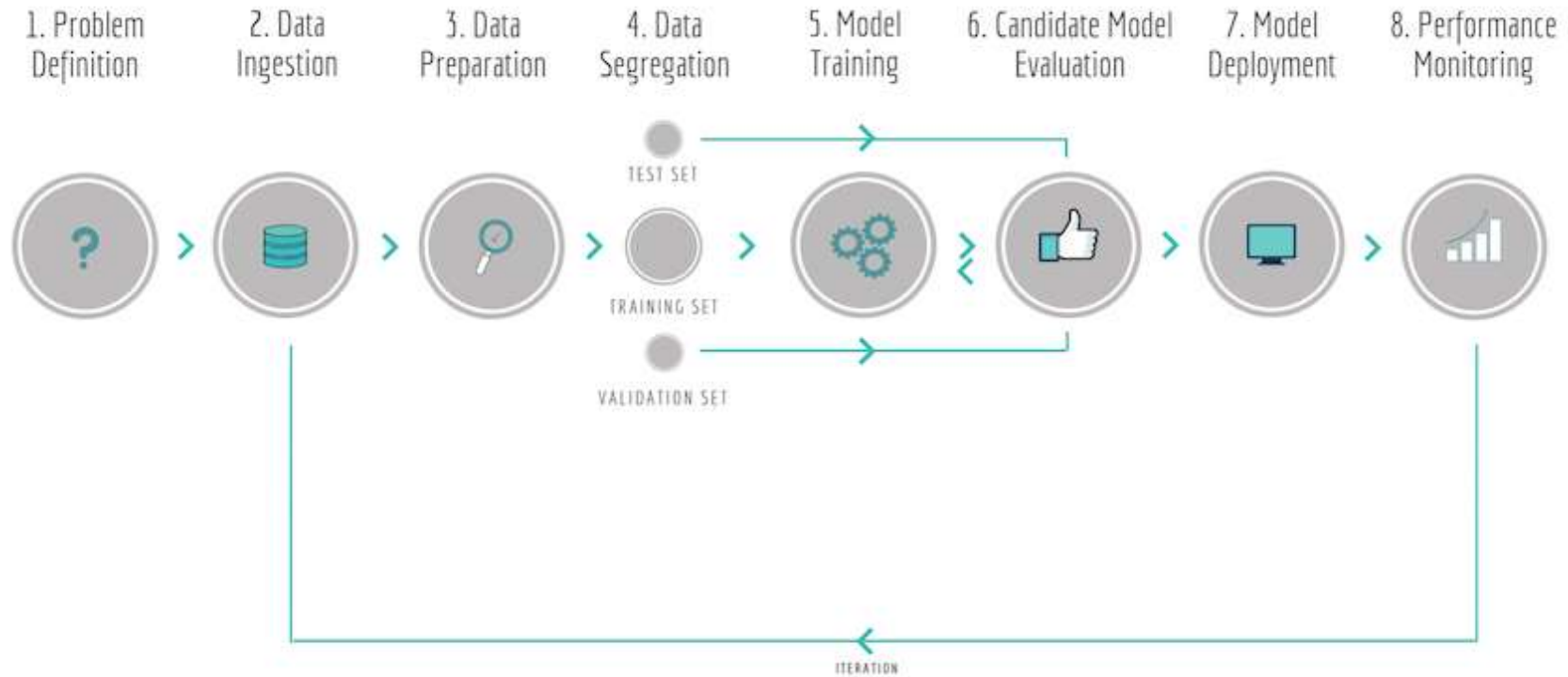
Preparação e Exploração avançada de dados com KNIME

LEI/MiEI @ 2022/2023, 2º sem
[ADI^3]

- Preparação de Dados
 - Join, Concatenation, Sorter, Filter and Aggregations
- Preparação e Exploração Avançada de Dados
 - Missing Values Treatment, Binning, Feature Scaling, Outlier Detection
 - Feature Selection, Nominal Value Discretization, Feature Engineering
- Experimentação
(*hands on*)

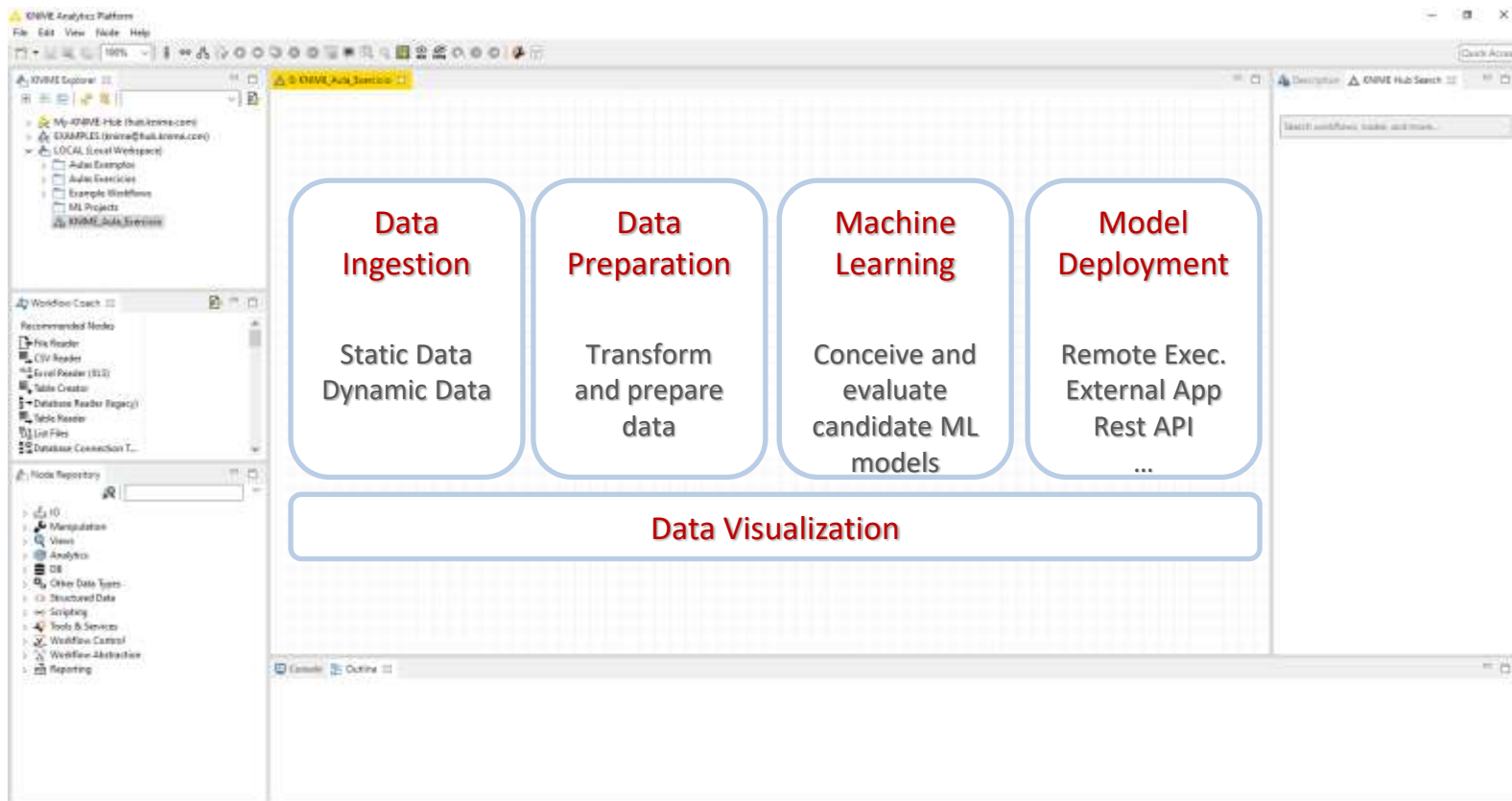


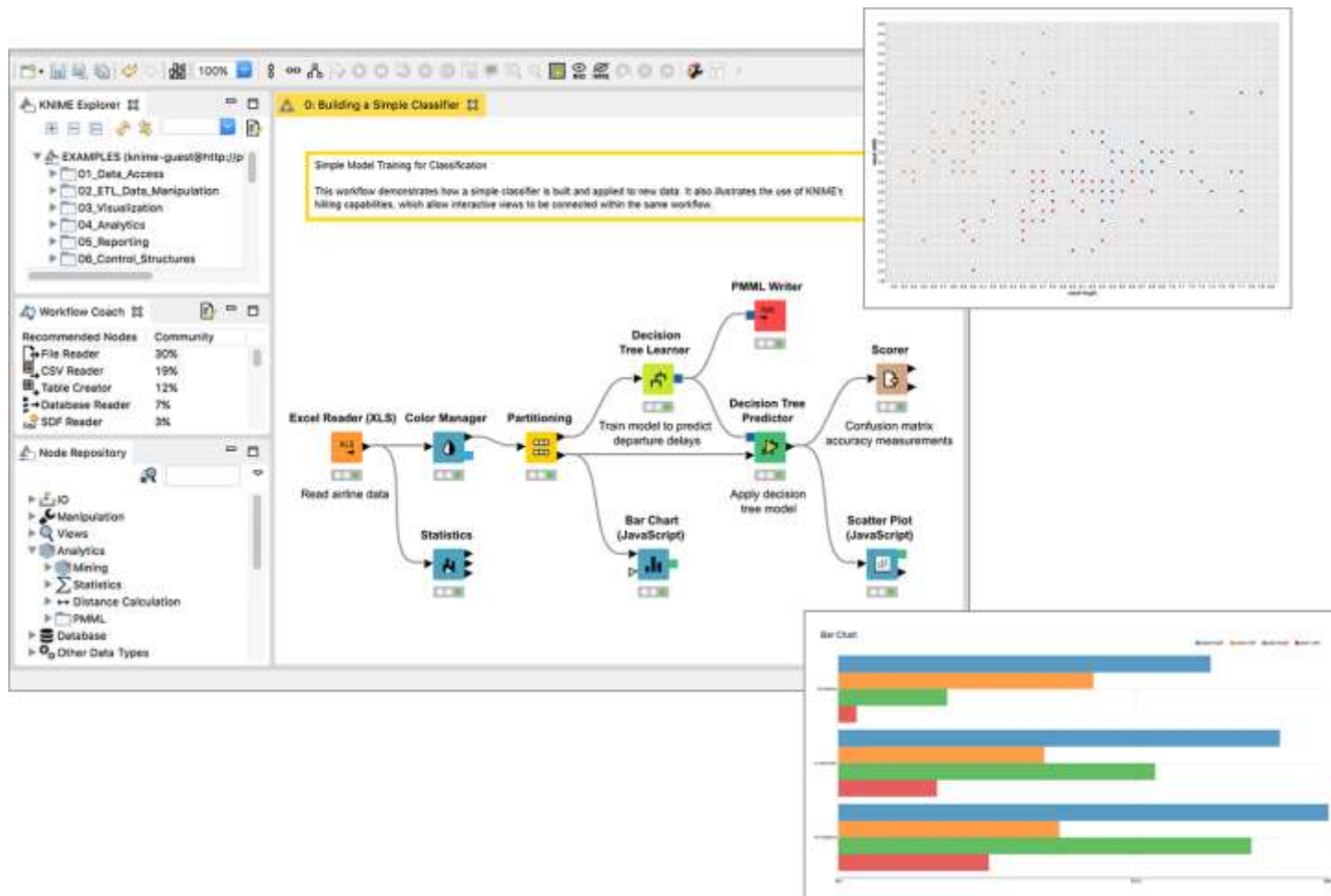
A Machine Learning Pipeline



(<https://towardsdatascience.com/architecting-a-machine-learning-pipeline-a847f094d1c7>)

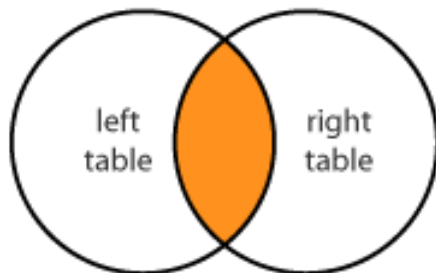
Fluxo de Trabalho Típico @ Knime



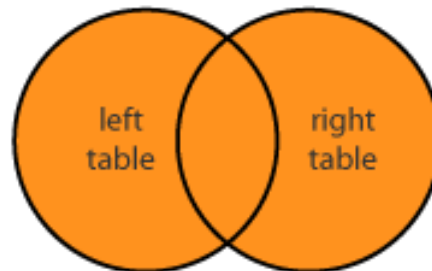


- Uma operação JOIN (junção) combina dados de diferentes fontes

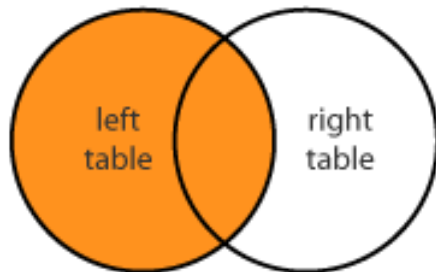
INNER JOIN



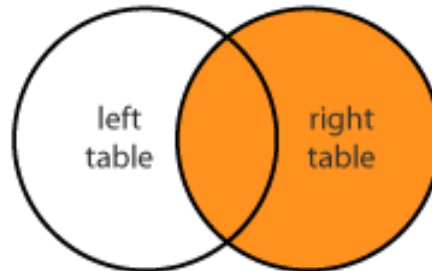
FULL JOIN



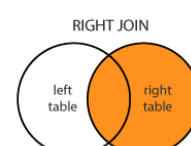
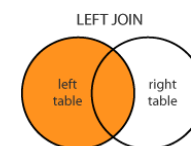
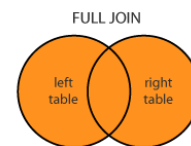
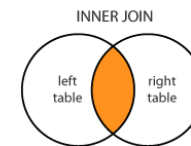
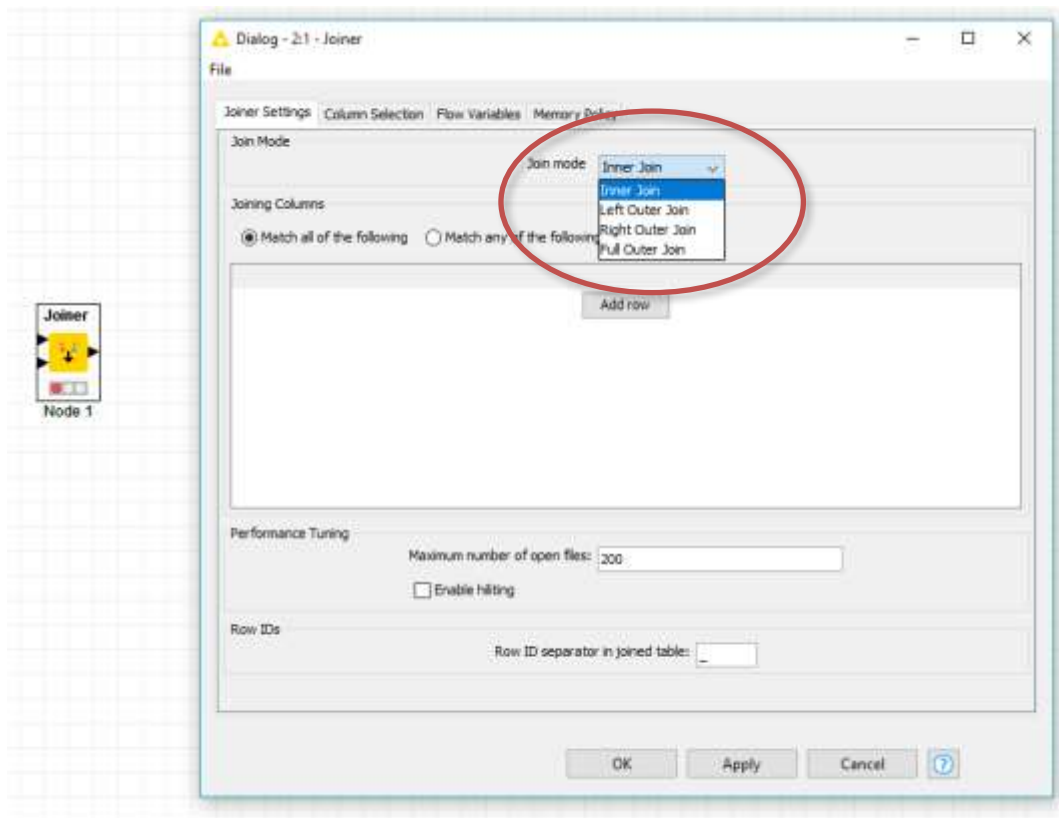
LEFT JOIN



RIGHT JOIN



- No KNIME estão disponíveis 4 tipos de operações JOIN



■ União de colunas

Manually created table - 2:2 - Table Creator

File Hilite Navigation View

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

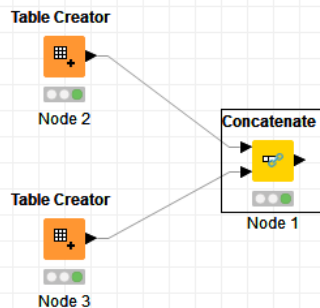
Row ID	S Id	S Name	S Age
Row0	1	Ze	19
Row1	2	Maria	22

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 hiling

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

Interseção de colunas

Manually created table - 2:2 - Table Creator

File Hilite Navigation View

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

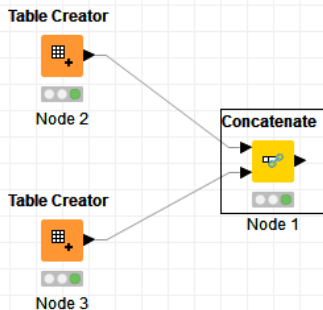
Row ID	S Id	S Name	S Age
Row0	1	Ze	19
Row1	2	Maria	22

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

OK Apply Cancel ?

Concatenated table - 2:1 - Concatenate

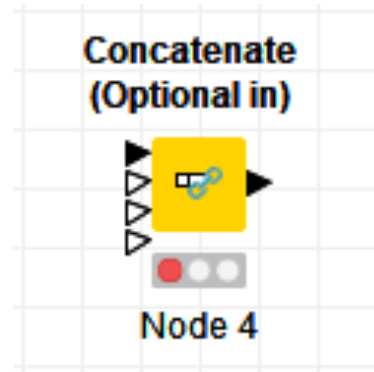
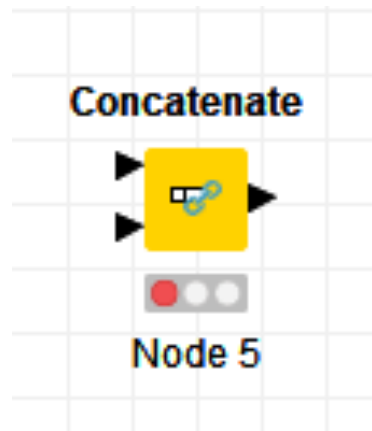
File Hilite Navigation View

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

Row ID	S Id	S Name
Row0	1	Ze
Row1	2	Maria
Row0_dup	3	Rui
Row1_dup	4	Ana
Row2	5	João

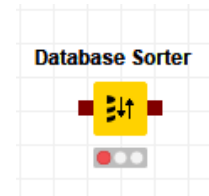
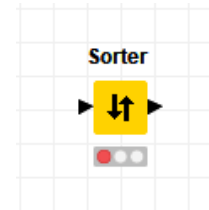
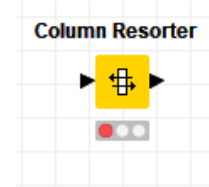
Concatenação *Concatenation*

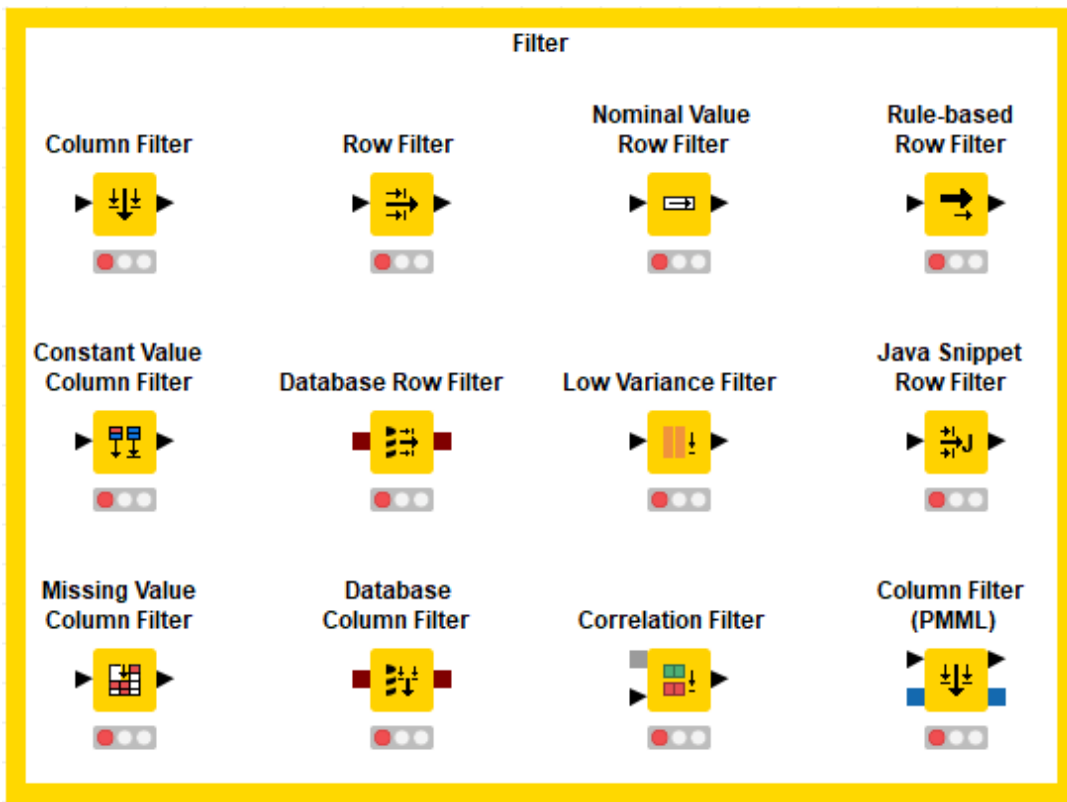
- Concatenate (Optional in) funciona como a concatenação (Concatenate), mas aceita até 4 *inputs*



Ordenação *Sorter*

- Altera a ordem das colunas de *input*, com base na definição dos parâmetros
- Ordena as linhas com base na definição dos parâmetros
- Permite que as linhas sejam classificadas a partir da tabela da base de dados de entrada
(cláusula SQL ORDER BY)





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.

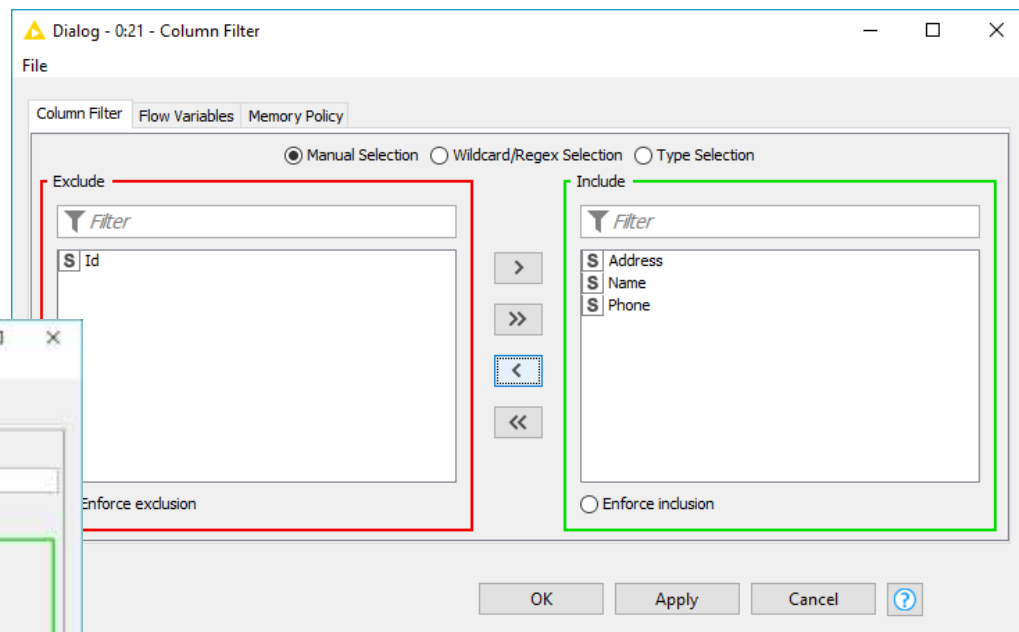
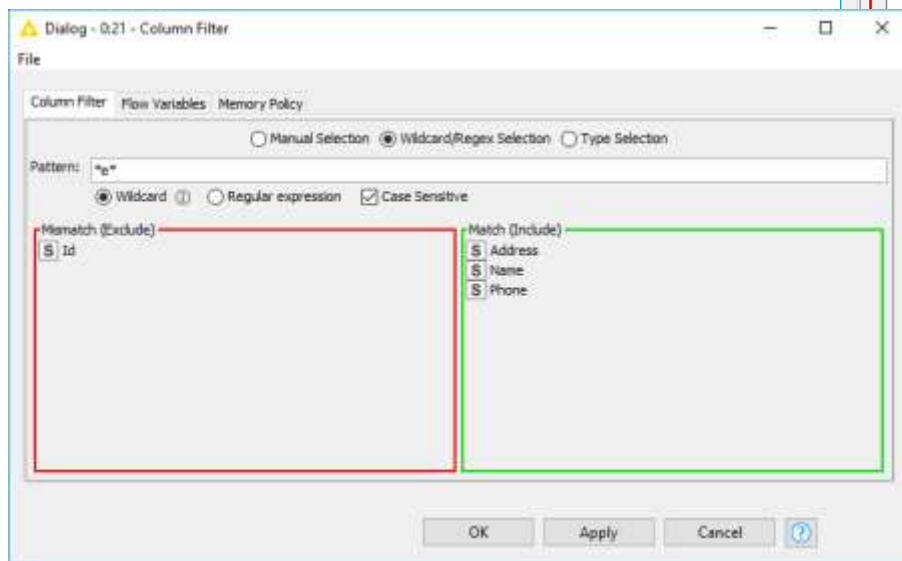
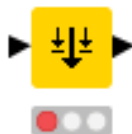
Search

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 check box 'Mark all search hits' causes all matching columns to be selected making them movable between the two lists.

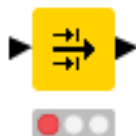
Filtros *Filter*

Filtro de Colunas *Column Filter*

Column Filter



Row Filter



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:

☐ filter based on collection elements

Matching criteria

☒ use pattern matching

☐ case sensitive match ☒ contains wild cards
☐ regular expression

☐ use range checking

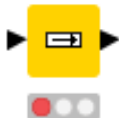
lower bound:
upper bound:

☐ only missing values match

OK Apply Cancel ?

Filtro de Linhas de Valores Nominais *Nominal Value Row Filter*

Nominal Value
Row Filter



Dialog - 0:54 - Nominal Value Row Filter

File

Selection Flow Variables Memory Policy

Select column:
Address

☒ Manual Selection ☐ Wildcard/Regex Selection

Exclude

Filter

Fafe
Lisboa
Coimbra

☐ Enforce exclusion

Include

Filter

Braga
Porto

☒ Enforce inclusion

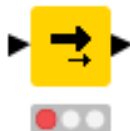
Ind. Missing Values

OK Apply Cancel ?

Filtro de Linhas de Baseado em Regras

Rule-based Row Filter

Rule-based
Row Filter



Dialog - 0:32 - Rule-based Row Filter

File

Rule Editor Flow Variables Memory Policy

Column List

- ROWID
- ROWINDEX
- ROWCOUNT
- [S] Id
- [S] Address
- [S] Name
- [S] Phone

Flow Variable List

- [S] knime.workspace

Category

All

Function

- ? <= ?
- ? = ?
- ? > ?
- ? >= ?
- ? AND ?
- ? IN ?
- ? LIKE ?
- ? MATCHES ?
- ? OR ?
- ? XOR ?
- FALSE
- MISSING ?
- NOT ?
- TRUE

Description

Logical or of two boolean expressions. You can use this in a sequence, like A OR B OR C without parenthesis, but it has no precedence regarding to AND or XOR, so you have to use parenthesis around the logical connectives if you want to combine them.
(Short-circuit evaluation.)

Expression

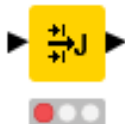
```
1 //Enter ordered set of rules!!!!!!
2 //TRUE => TRUE
3 $Name$ = "Ana" OR $Name$ = "Mariana" OR $Name$ = "Joana" => TRUE
4 $Address$ = "Braga" OR $Address$ = "Porto" => TRUE
```

☒ Include TRUE matches ☐ Exclude TRUE matches

OK Apply Cancel ?

Filtro de Linhas em JAVA Snippet JAVA Snippet Row Filter

Java Snippet
Row Filter



Dialog - 0:24 - Java Snippet Row Filter

File

Java Snippet Additional Libraries Flow Variables Memory Policy

Column List

- ROWID
- ROWINDEX
- ROWCOUNT
- [S] Id
- [S] Address
- [S] Name
- [S] Phone

Flow Variable List

- [S] knime.workspace

Global Variable Declaration

Method Body

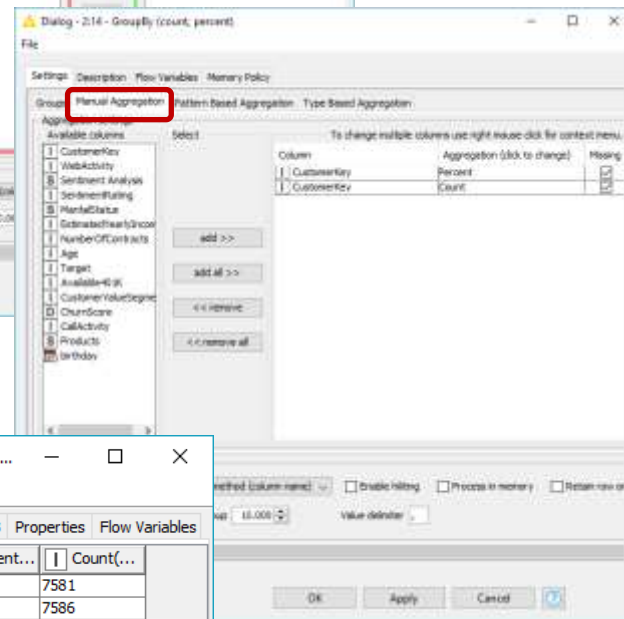
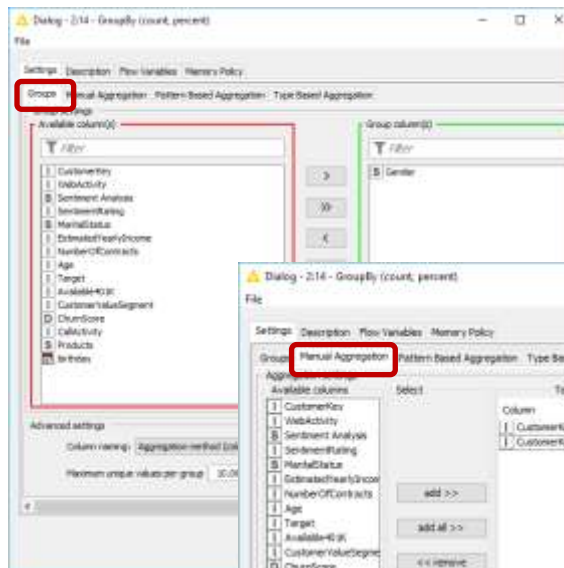
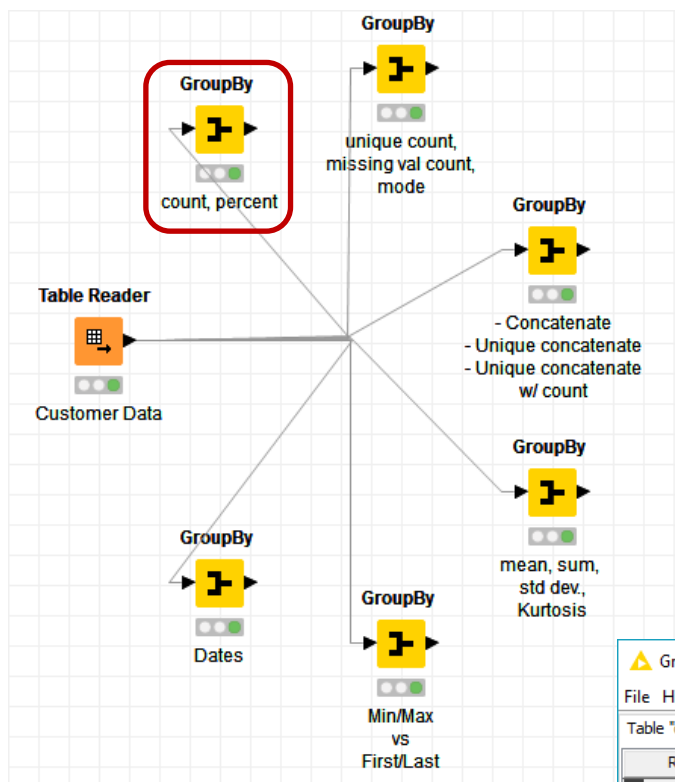
```
boolean result = false;  
  
if($Name$.equals("Ana") || $Name$.equals("Mariana") || $Name$.equals("Joana"))  
    result = true;  
  
return result;
```

☐ Insert Missing As Null ☒ Compile on close

OK Apply Cancel ?

Operações de Agregação

Count and Percent



Group table - 2:14 - GroupBy (count, ...

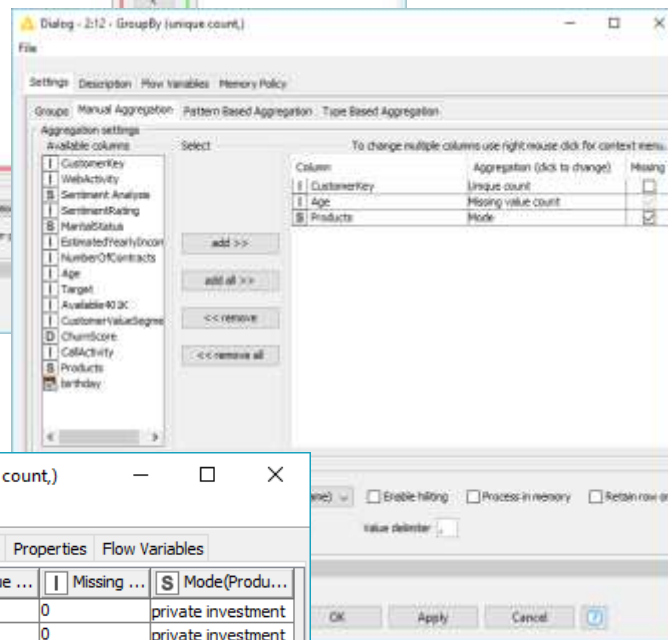
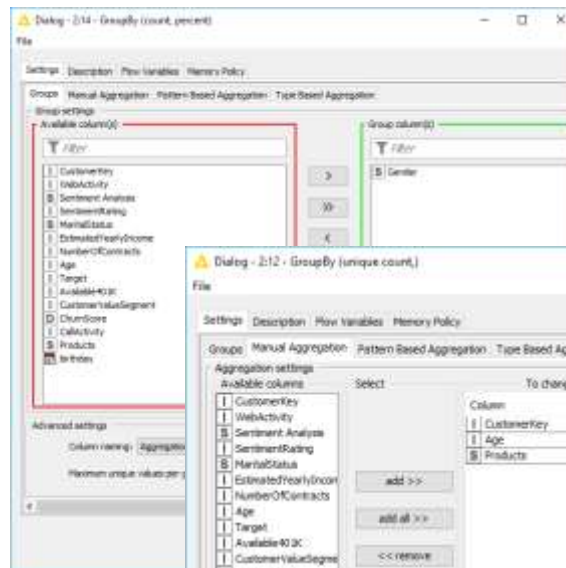
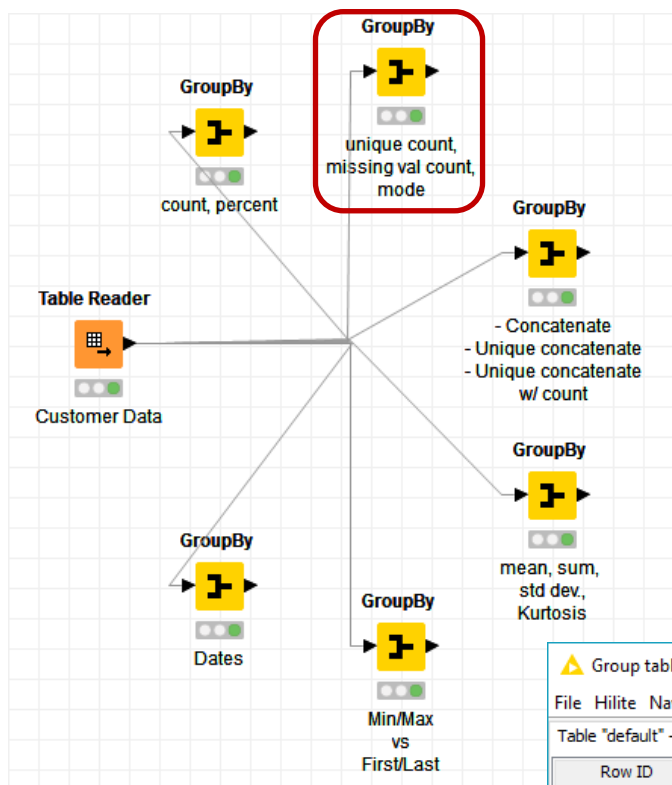
File Hilite Navigation View

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

Row ID	S Gender	D Percent...	I Count(...
Row0	F	49.984	7581
Row1	M	50.016	7586

Operações de Agregação

Unique Count, Missing Values Count and Mode



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

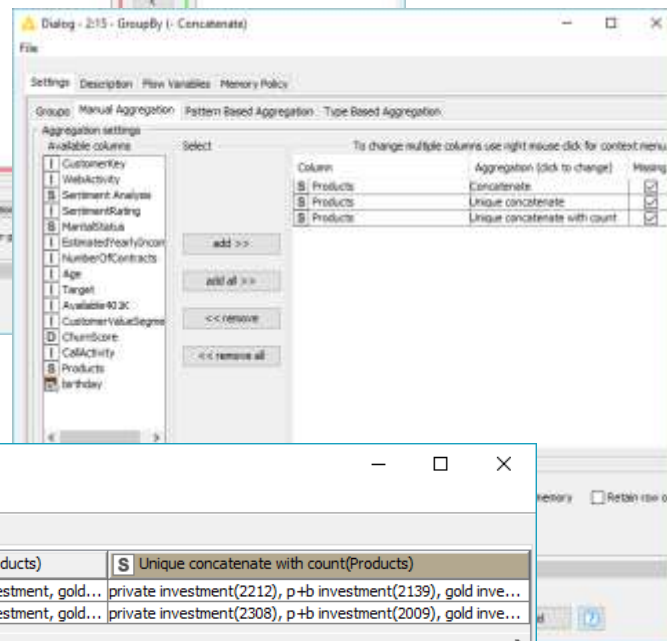
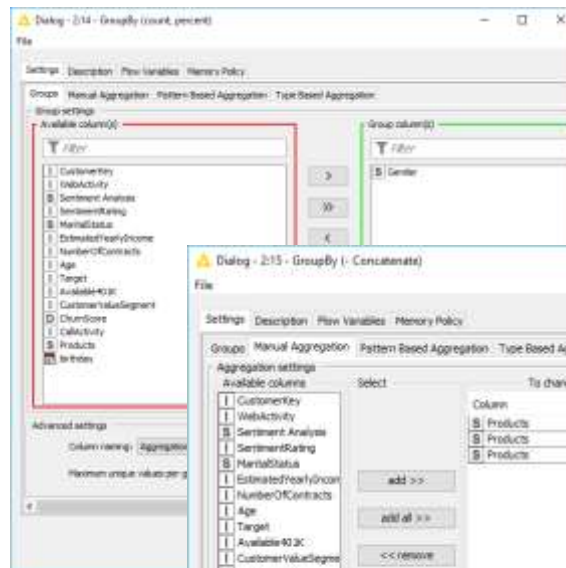
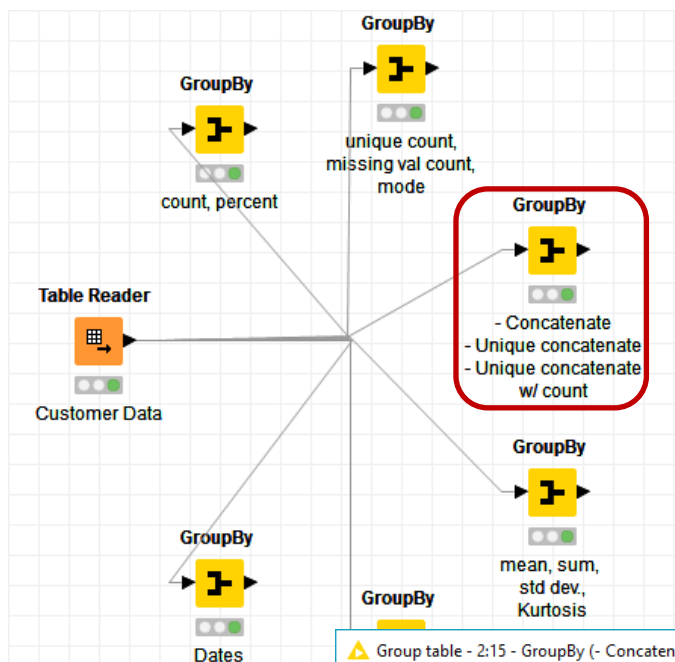
File Hilite Navigation View

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

Operações de Agregação

Concatenate



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

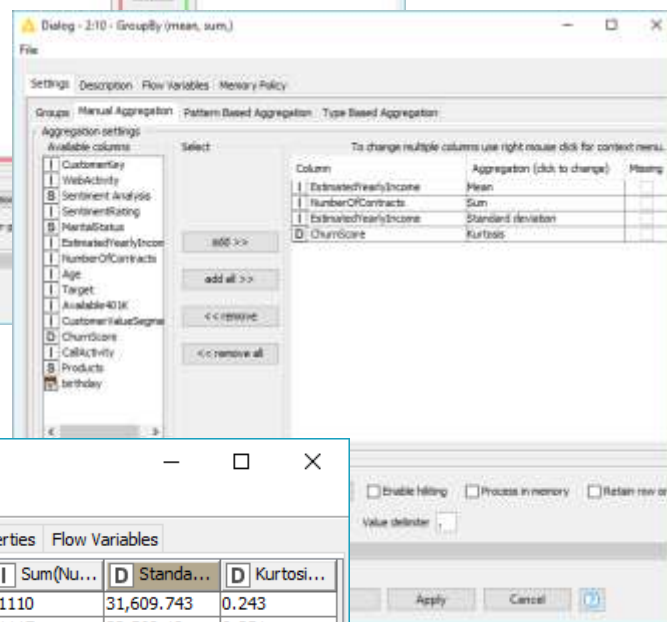
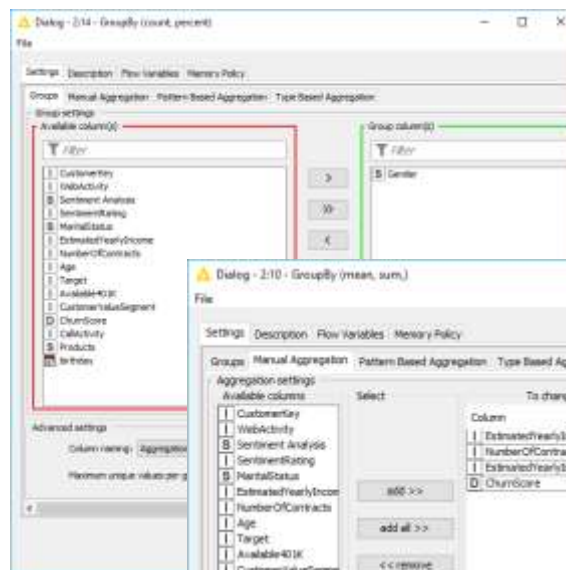
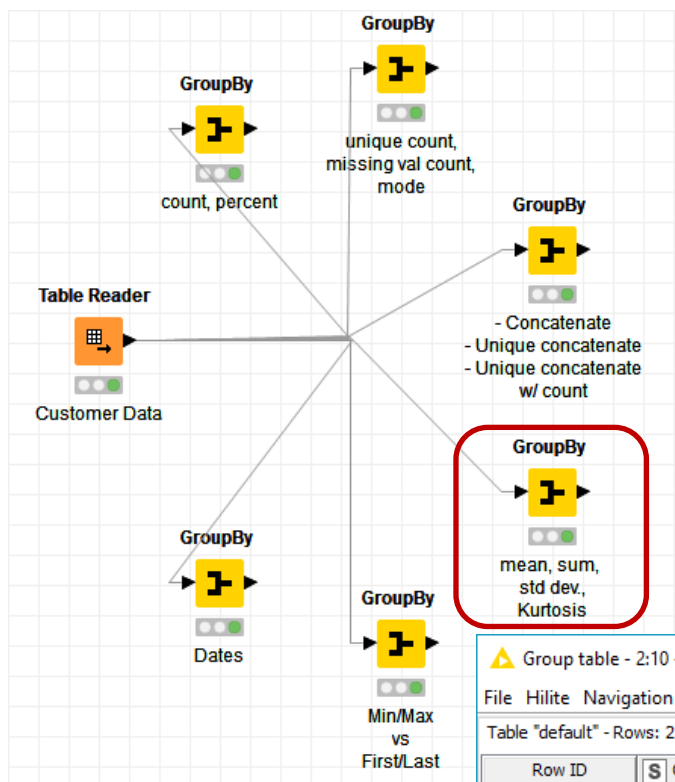
File Hilite Navigation View

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...

Operações de Agregação

Mean, Sum, Standard Deviation and Kurtosis



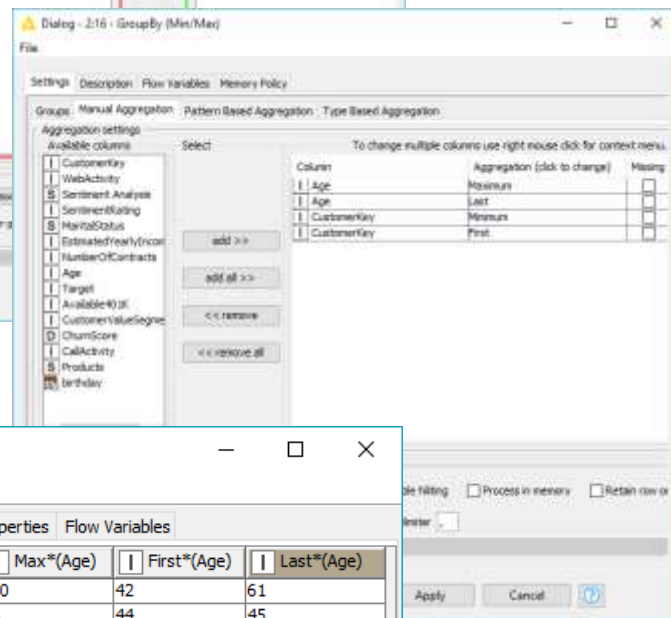
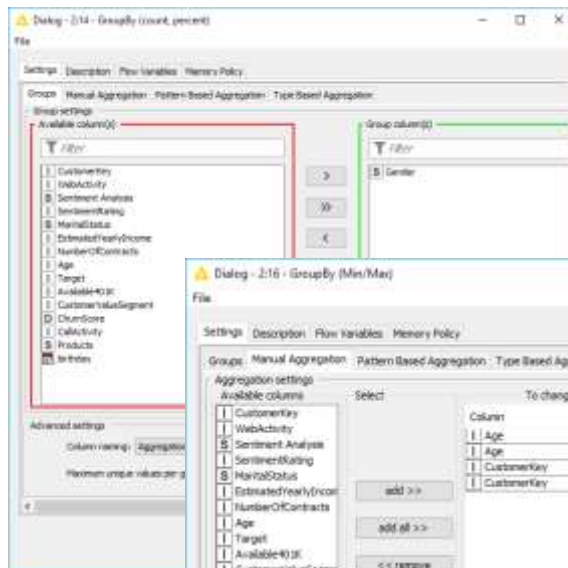
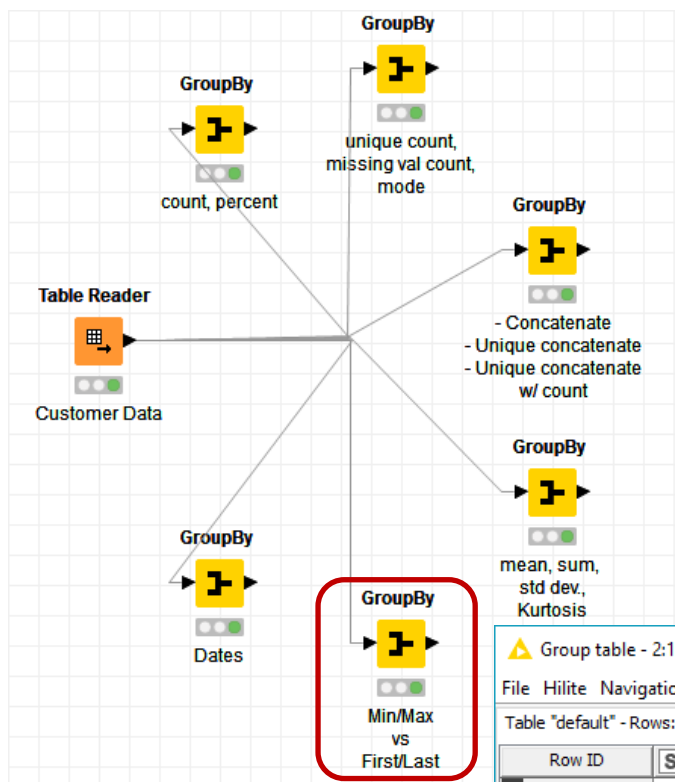
Group table - 2:10 - GroupBy (mean, sum,)

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

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

Operações de Agregação

Min/Max vs First/Last



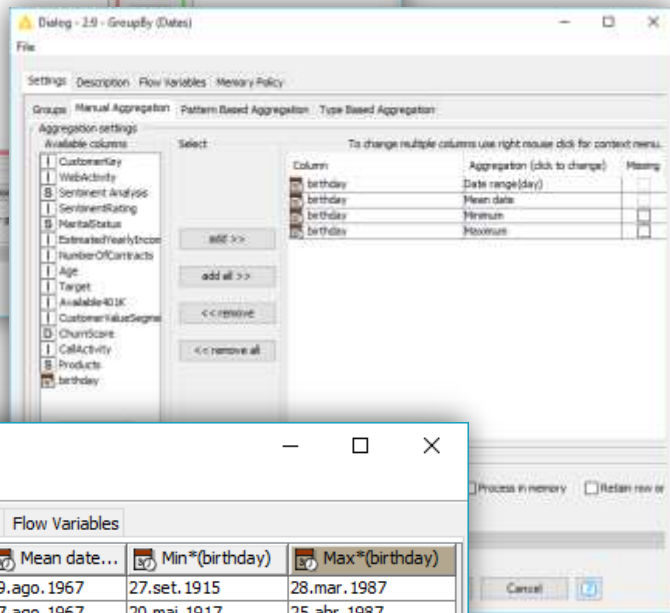
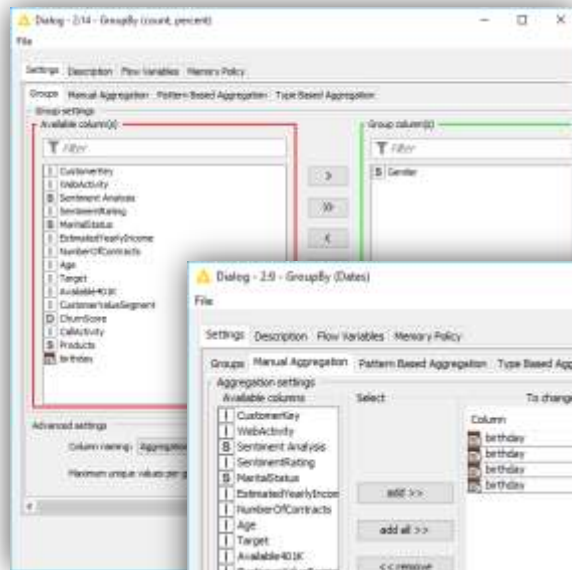
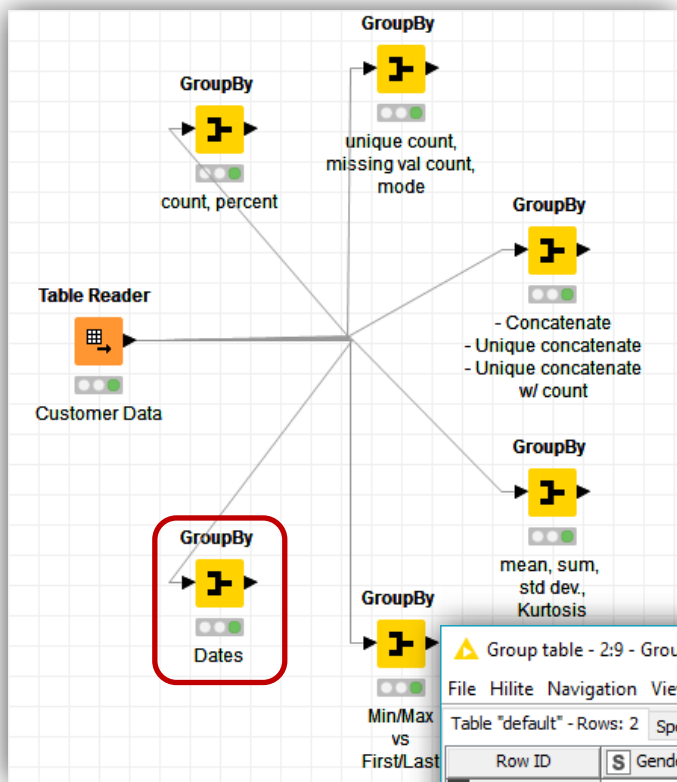
Group table - 2:16 - GroupBy (Min/Max)

File Hilite Navigation View

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

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

Operações de Agregação *Dates*



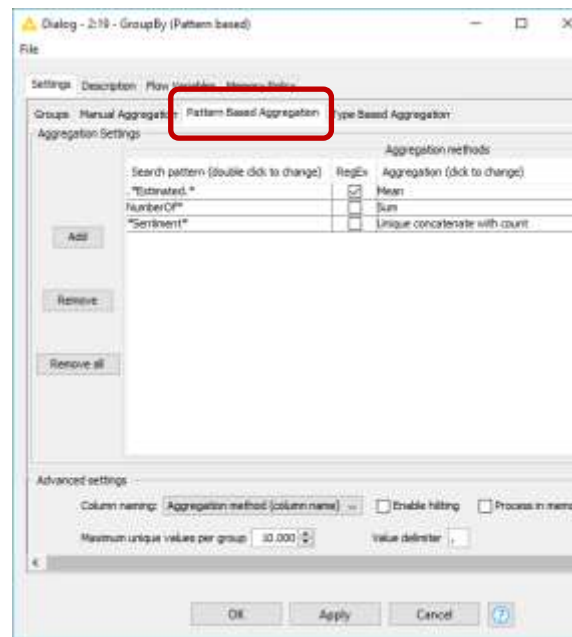
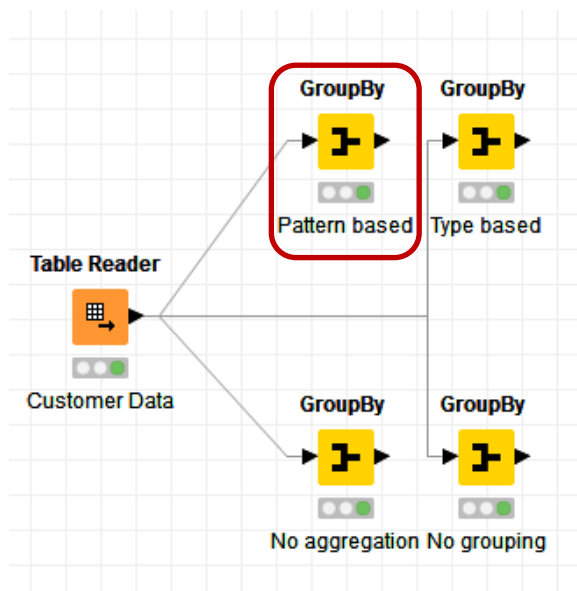
Group table - 2:9 - GroupBy (Dates)

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

Operações Avançadas de Agregação *Pattern Based*



Group table - 2:19 - GroupBy (Pattern based)

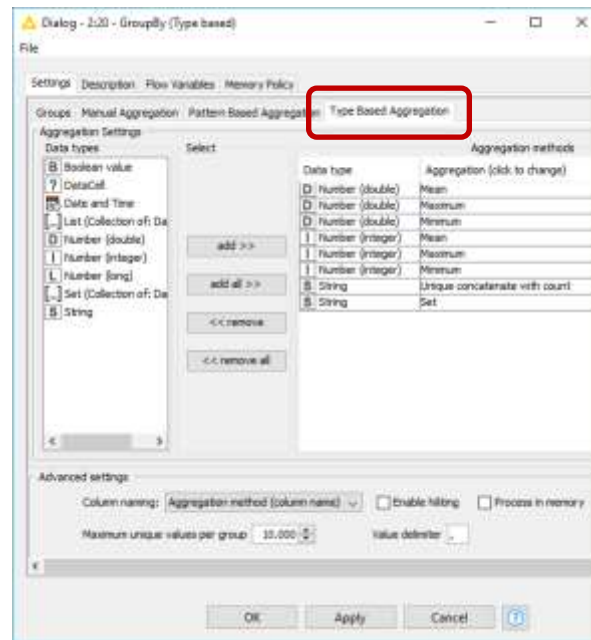
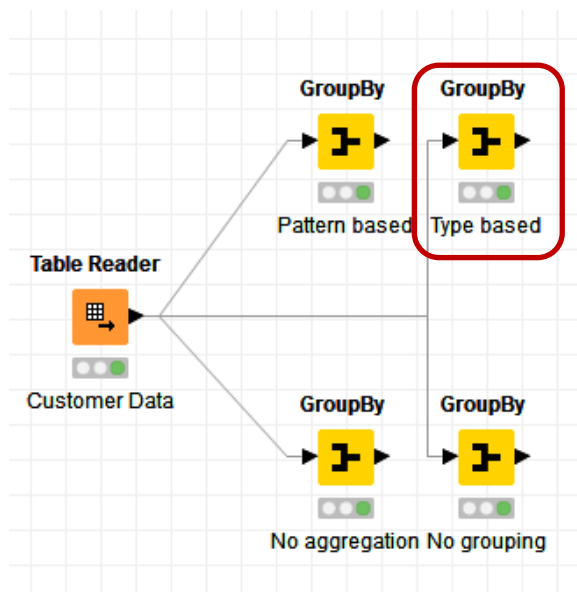
File

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

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	

Operações Avançadas de Agregação

Data Type Based



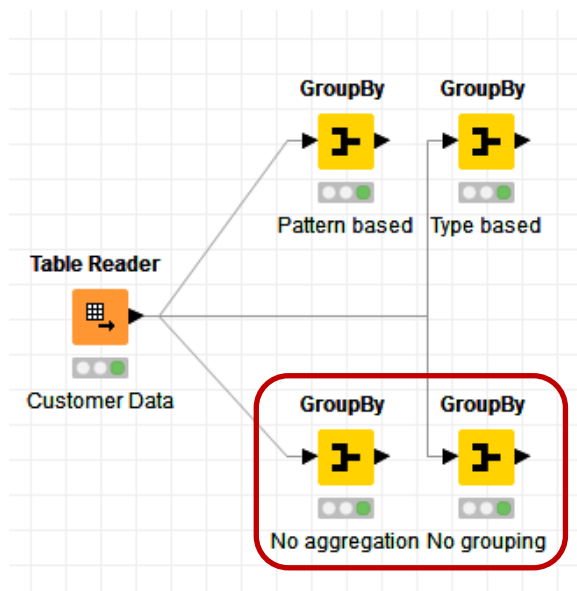
Group table - 2:20 - GroupBy (Type based)

File Hilite Navigation View

Table "default" - Rows: 2 Spec - Columns: 70 Properties Flow Variables

Row ID	S Gender	D Mean(C...	I Max*(C...	I Min*(C...	D Mean(C...	I Max*(C...	I Min*(C...	D Mean(C...	I Max*(C...	I Min*(C...	D Mean(C...	I Max*(C...	I Min*(C...	D Mean(C...	I Max*(C...	I Min*(C...
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	SM		

Operações Avançadas de Agregação *No Aggregation vs No Grouping*



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 - 2:16 - 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(3890), Very Negative(4173), Very Positive(1199), Positive(1960), Negative(3122)

Universidade do Minho

Escola de Engenharia

Departamento de Informática

Preparação e Exploração Avançada de Dados em KNIME

Preparação e Exploração Avançada de Dados

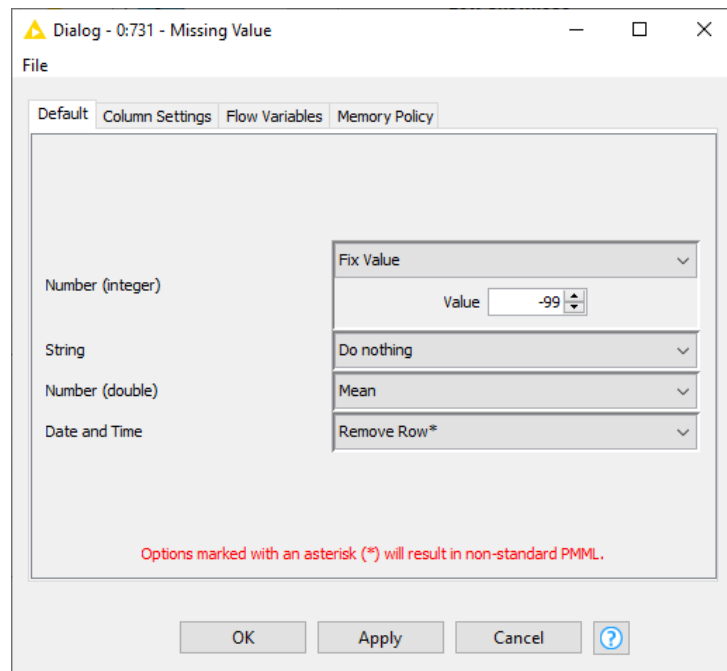
- Tratamento de Valores em Falta/ *Missing Values treatment*
- Discretização em Intervalos/ *Binning*
- Redimensionamento de atributos/ *Feature scaling*
- Detecção de valores fora de limites/ *Outlier detection*



Tratamento de Valores em Falta *Missing Values*

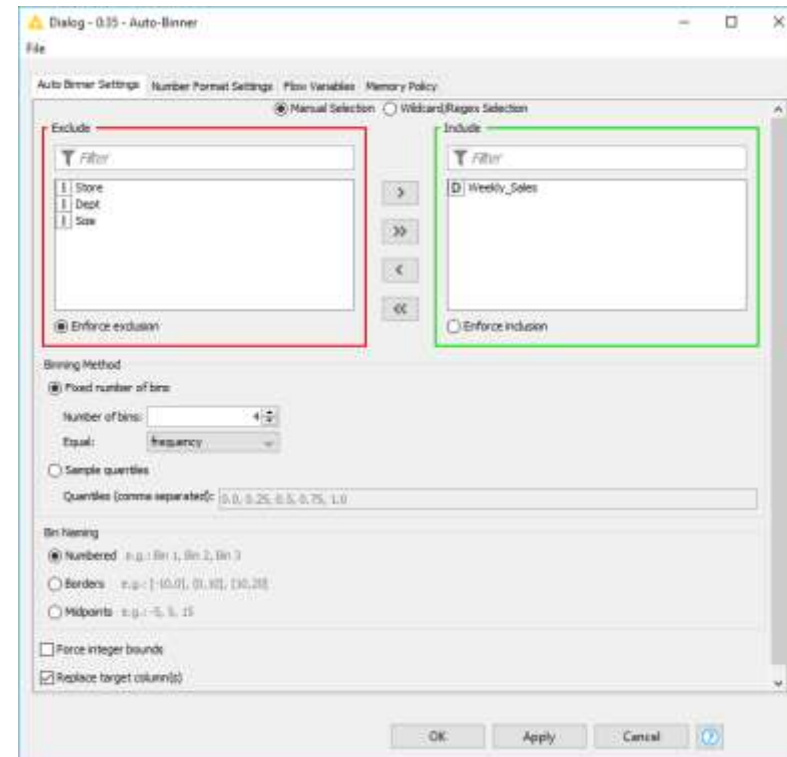
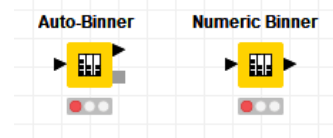
■ Tratamento de Valores em Falta

- Analisar cada atributo em relação à quantidade e proporção de valores em falta
- Decidir o que fazer:
 - Remover
 - Calcular a média
 - Interpolação linear
 - Criar máscaras
 - ...



Discretização em Intervalos *Binning*

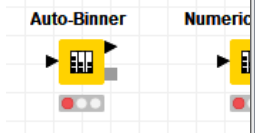
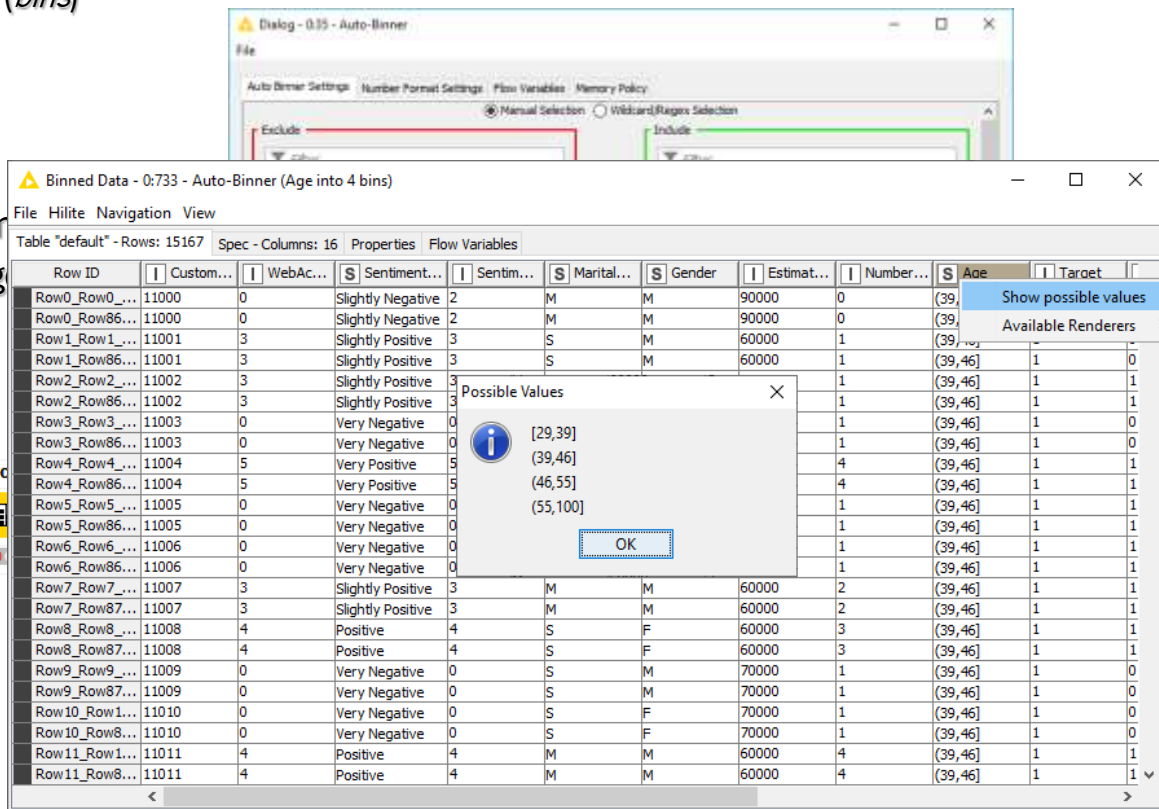
- Discretização em Intervalos ou ***Binning***
 - Agrupar valores numéricos em intervalos (*bins*)
- Tornar o modelo mais robusto e evitar o sobreajustamento.
- No entanto, penaliza o desempenho do modelo, uma vez que cada vez que se encerra algo, se sacrifica informação.



Discretização em Intervalos

Binning

- Discretização em Intervalos ou ***Binning***
 - Agrupar valores numéricos em intervalos (*bins*)
- Tornar o modelo mais robusto e evitar o sobreajustamento.
- No entanto, penaliza o desempenho do m uma vez que cada vez que se encerra algo se sacrifica informação.

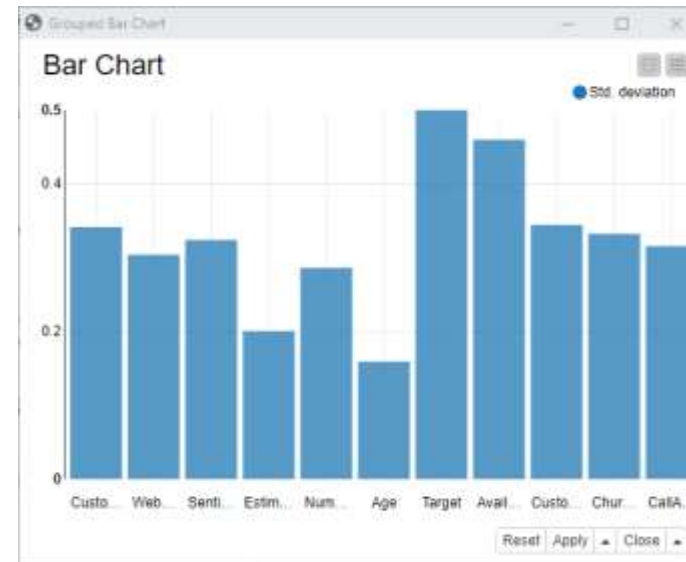
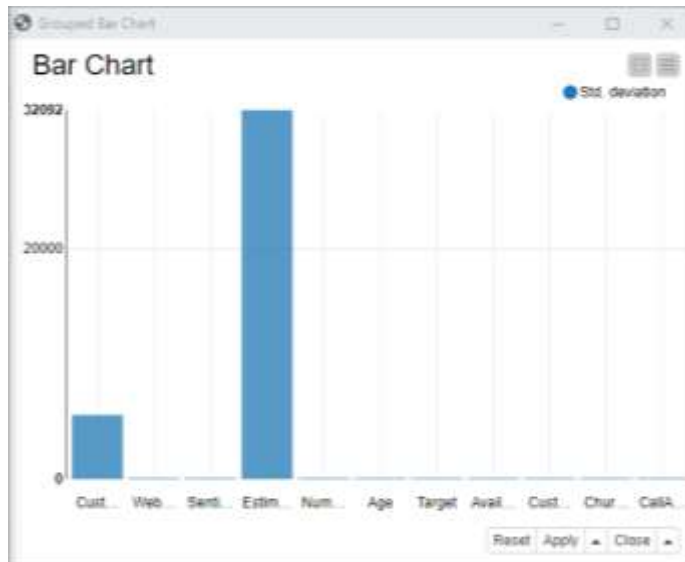
The screenshot shows the 'Dialog - 0.35 - Auto-Binner' window with the 'Auto Binner Settings' tab selected. Below it, the 'Binned Data - 0.733 - Auto-Binner (Age into 4 bins)' window displays a table with 16 columns and 15167 rows. A 'Possible Values' dialog box is open, showing the range [29,39] and [39,46] for the 'Age' variable.

Row ID	Custom...	WebAc...	Sentiment...	Sentim...	Marital...	Gender	Estimat...	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	4	(39,46]	1
Row4_Row86...	11004	5	Very Positive	5	S	M	60000	4	(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

Redimensionamento de atributos

Feature scaling

- Normalizar a gama de valores de atributos
- Muitos classificadores usam métricas de distância (ex.: distância euclidiana)
 - Se um atributo tiver uma gama ampla de valores, a distância será muito influenciada por esse atributo.
 - Assim, a gama de valores deve ser normalizada para que cada atributo possa contribuir proporcionalmente para a distância final.

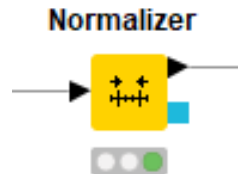


Redimensionamento de atributos *Feature scaling*

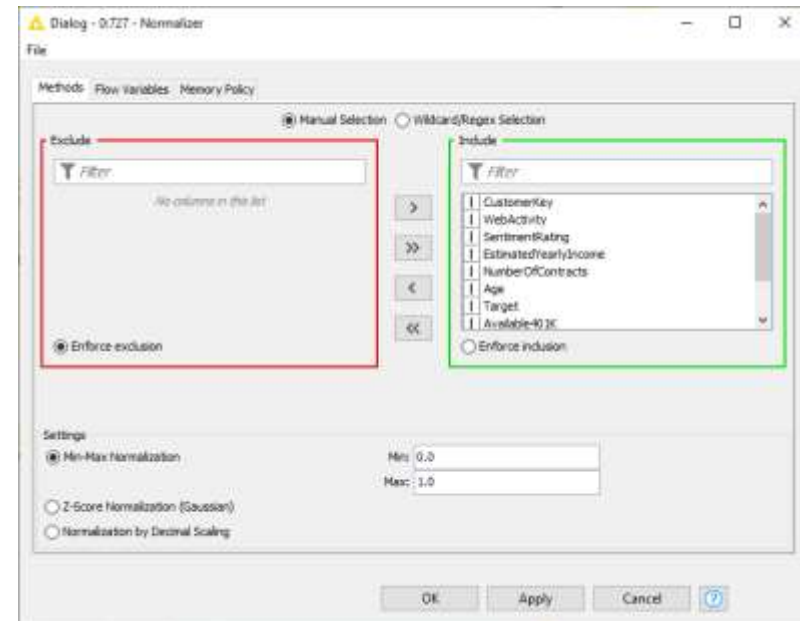
- Normalizar a gama de valores de atributos

- Normalização

Redimensionar os dados para que todos os valores caiam no intervalo de 0 e 1, por exemplo.



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



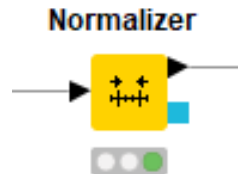
Redimensionamento de atributos *Feature scaling*

- Normalizar a gama de valores de atributos

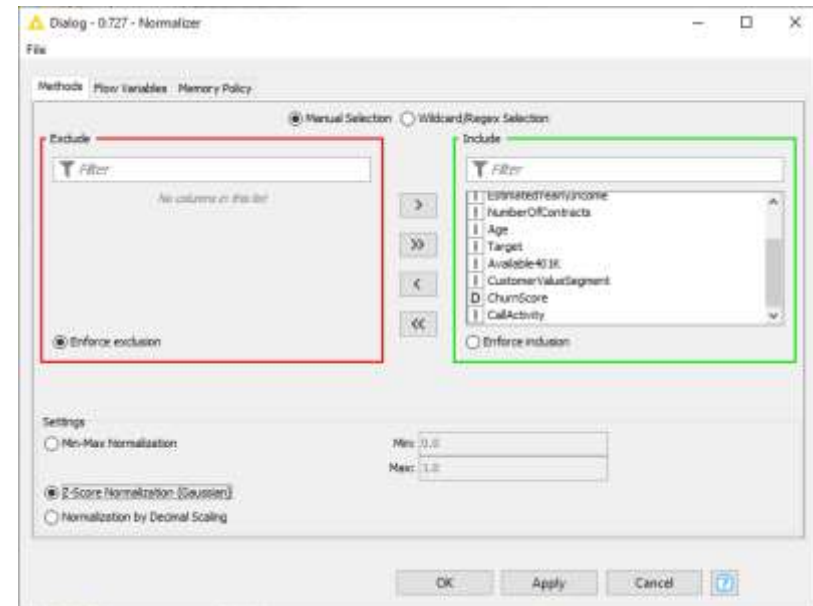
- *Standardization* (ou *Z-score Normalization*)

Redimensionar a distribuição de valores para que a média dos valores observados seja 0 e o desvio padrão seja 1.

Assume que os dados se ajustam a uma distribuição gaussiana com média e desvio padrão bem comportados, o que nem sempre é o caso.

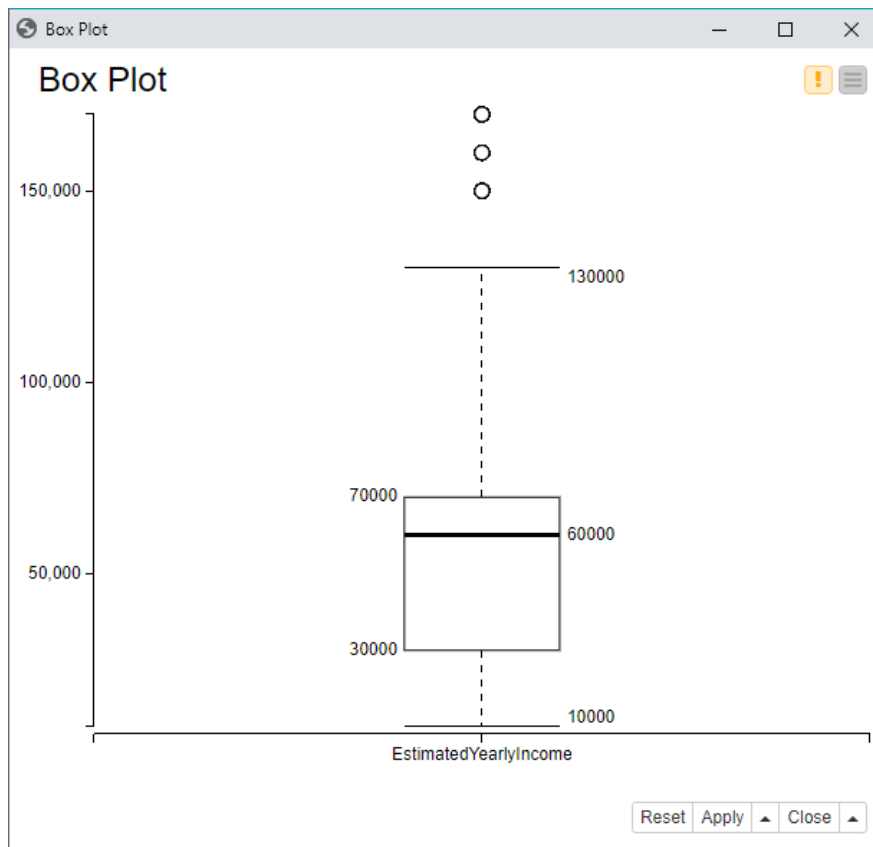
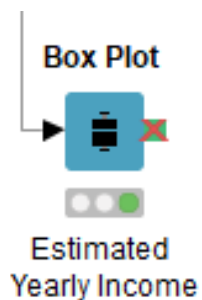


$$Z = \frac{x_i - \mu}{\sigma}$$



Deteção de valores fora de limites *Outlier Detection*

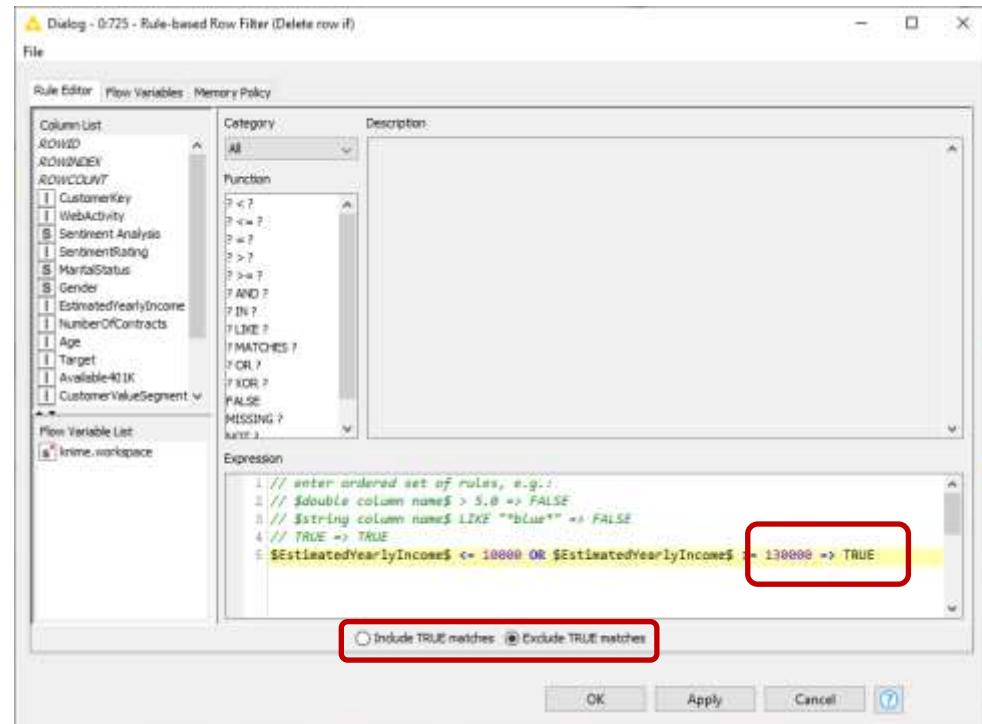
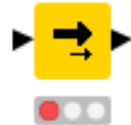
- Deteção de valores fora de limites
 - Estratégias baseadas em estatística
 - Box Plots
 - Z-Score (std. dev)



Deteção de valores fora de limites *Outlier Detection*

- Deteção de valores fora de limites
 - Estratégias baseadas em conhecimento

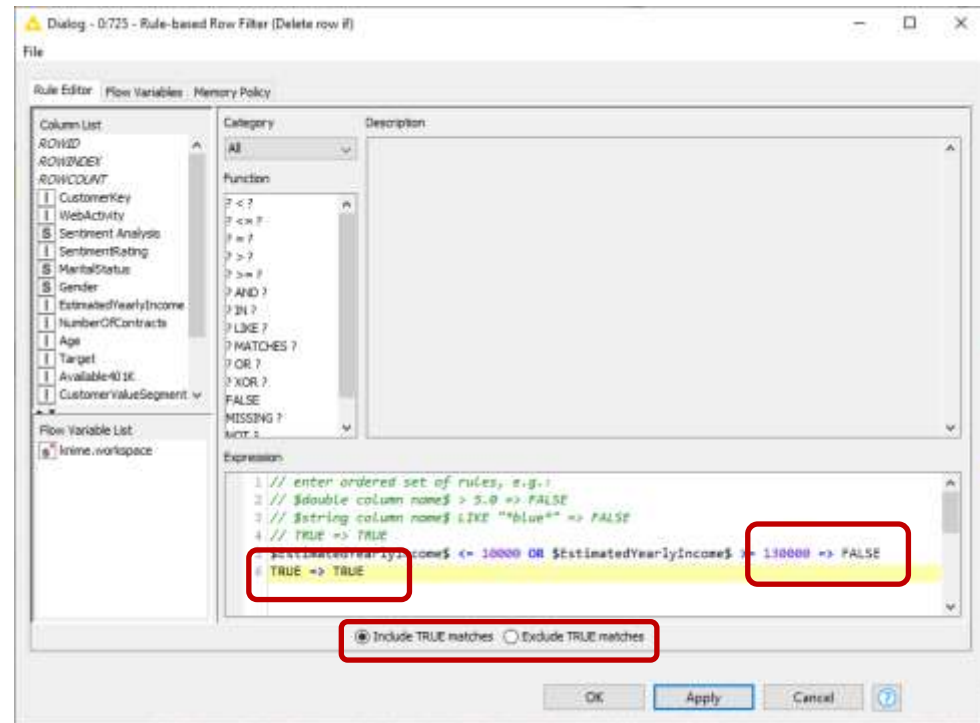
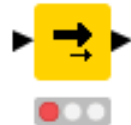
Rule-based
Row Filter



Deteção de valores fora de limites *Outlier Detection*

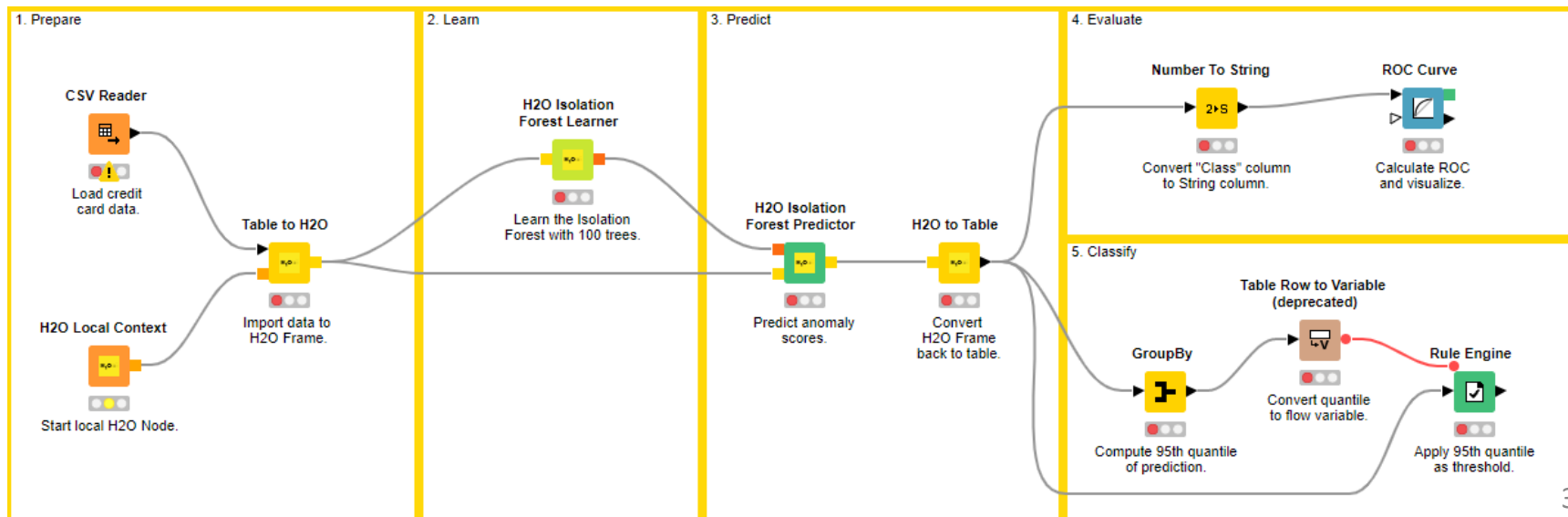
- Deteção de valores fora de limites
 - Estratégias baseadas em conhecimento

Rule-based Row Filter



Deteção de valores fora de limites *Outlier Detection*

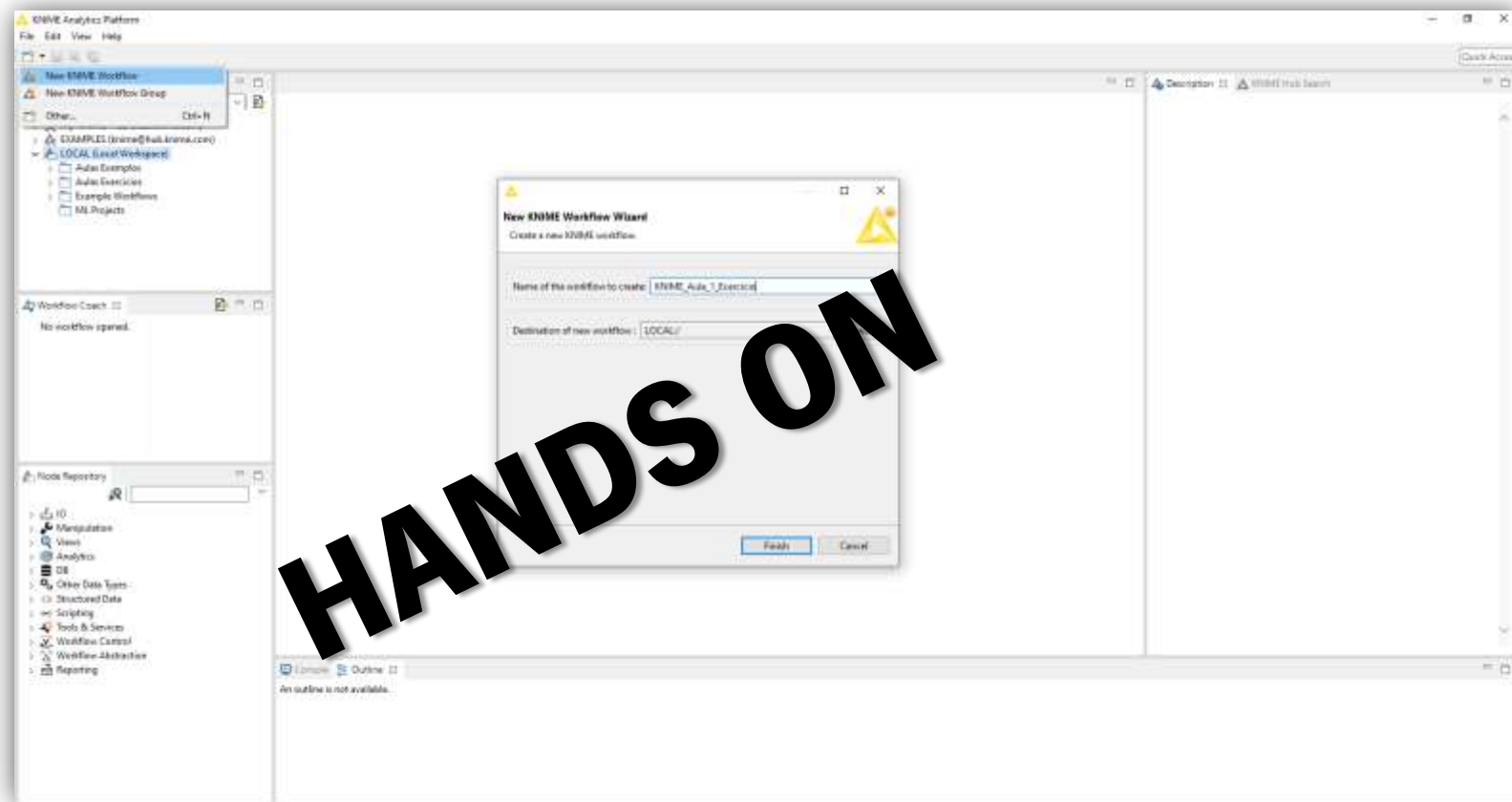
- Deteção de valores fora de limites
 - Estratégias baseadas em modelos
 - Isolation Forest
 - One-Class SVM
 - Minimum Covariance Determinant
 - ...



Deteção de valores fora de limites ***Outlier Detection***

- Deteção de valores fora de limites
 - Estratégias baseadas em estatística
 - Estratégias baseadas em conhecimento
 - Estratégias baseadas em modelos

- ***Outlier Dilemma: Drop or Cap?***
 - Para manter o tamanho do conjunto de dados, podemos querer limitar os outliers em vez de os apagar.
 - No entanto, isso pode afetar a distribuição dos dados!



Preparação e Exploração Avançada de Dados

- Tratamento de Valores em Falta/ *Missing Values treatment*
 - Discretização em Intervalos/ *Binning*
 - Redimensionamento de atributos/ *Feature scaling*
 - Detecção de valores fora de limites/ *Outlier detection*
-
- Seleção de Atributos/ *Feature selection*
 - Discretização de Valores Nominiais/ *Nominal value discretization*
 - Engenharia de Atributos/ *Feature Engineering*



