



Universidade do Minho
Departamento de Informática

Preparação e Exploração avançada de dados com KNIME

LEI/MiEI @ 2024/2025, 2º sem

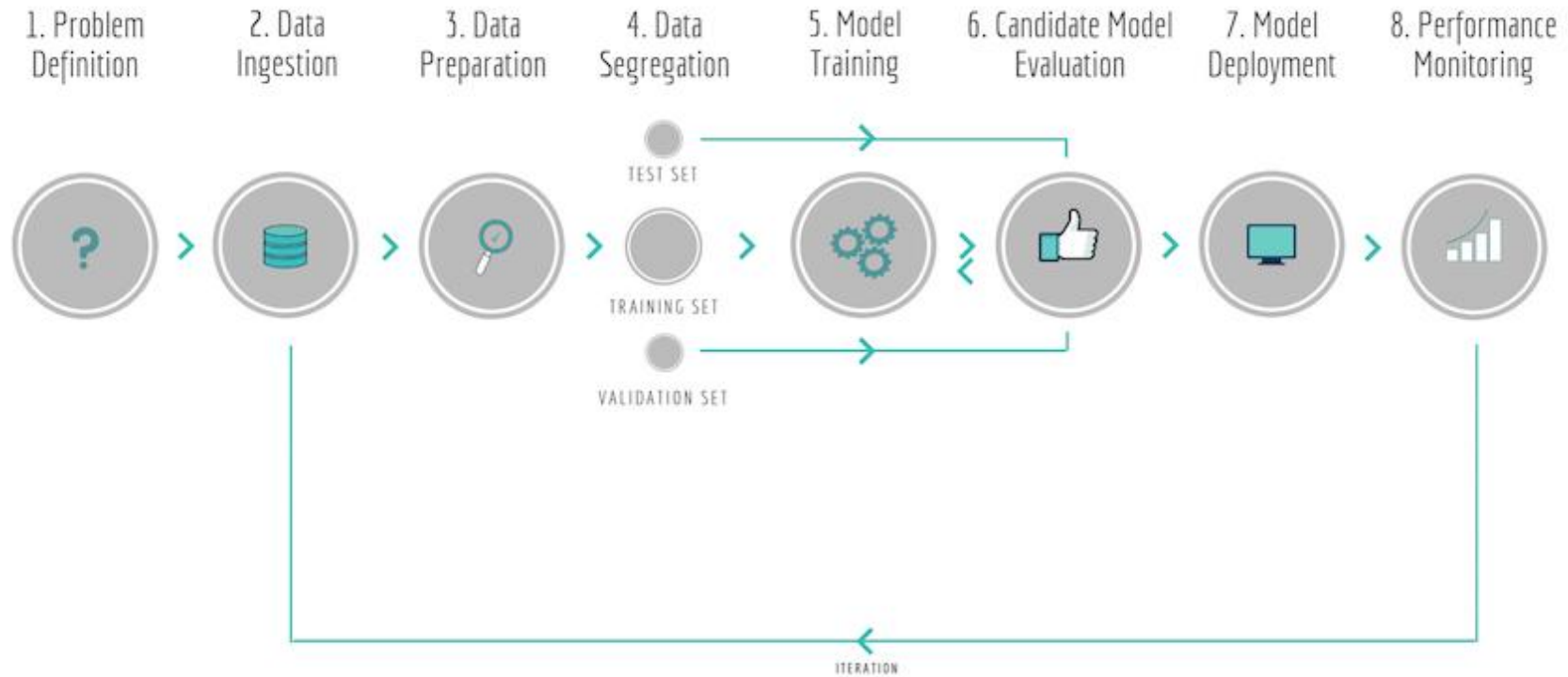


- Preparação de Dados
 - Join, Concatenation, Sorter, Filter and Aggregations
- Preparação e Exploração Avançada de Dados
 - Missing Values Treatment, Binning, Feature Scaling, Outlier Detection
 - Feature Selection, Nominal Value Discretization, Feature Engineering
- Experimentação
(*hands on*)





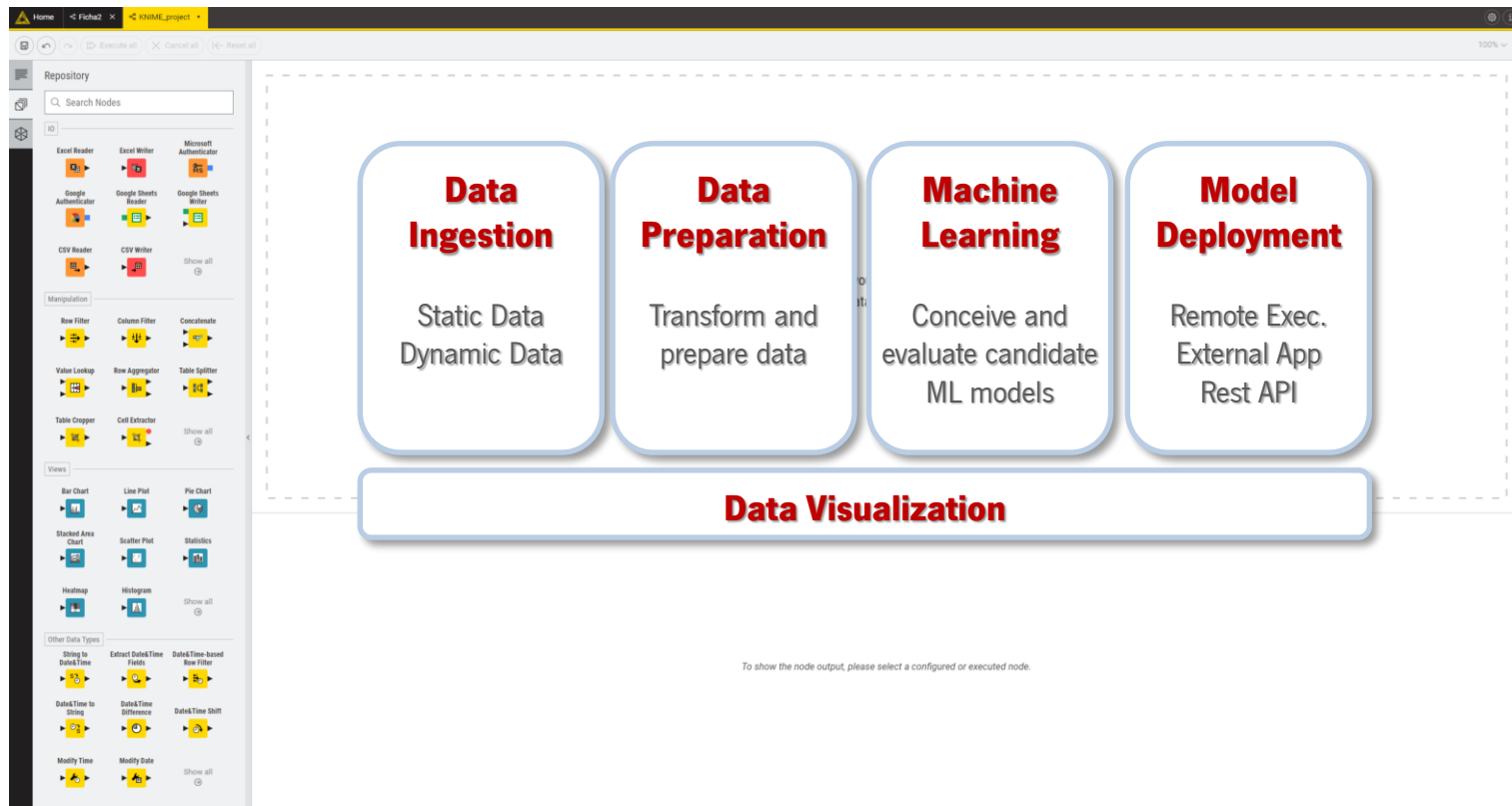
A Machine Learning Pipeline



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



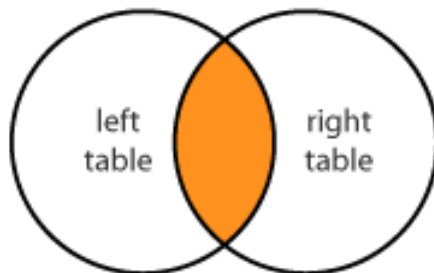
Fluxo de Trabalho Típico @ Knime



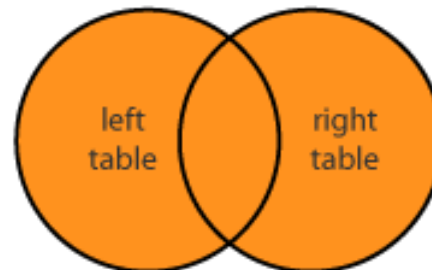


- Uma operação JOIN combina dados de diferentes fontes:

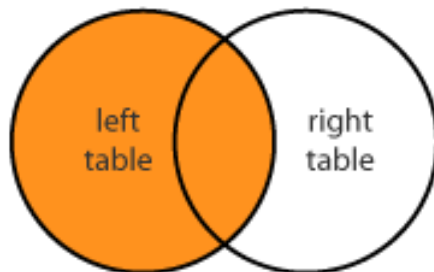
INNER JOIN



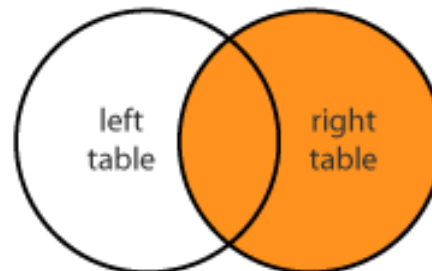
FULL JOIN



LEFT JOIN

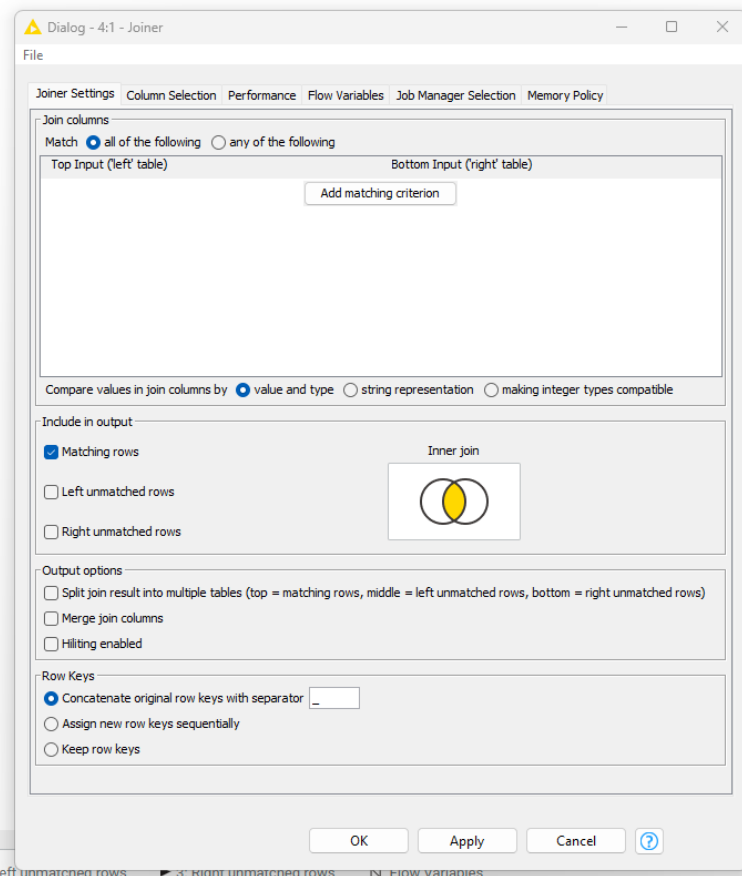
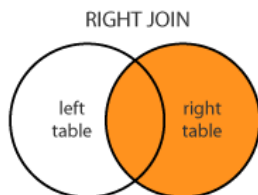
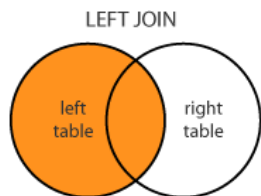
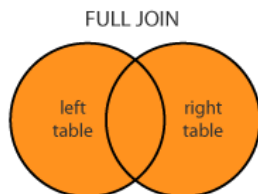
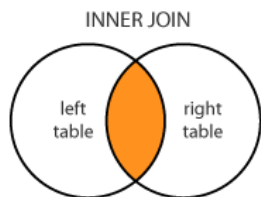


RIGHT JOIN





- No KNIME estão disponíveis diversas combinações no nodo JOINER:





Concatenação

Concatenation

União de colunas: CONCATENATE

Diagram illustrating the concatenation process in a data workflow:

The workflow consists of two **Table Creator** components feeding into a **Concatenate** component. The **Concatenate** component is configured with the following settings:

- How to combine input columns:** **Union** (selected).
- If there are duplicate RowIDs:** **Append suffix** (selected).
- Suffix:** **_dup**.
- Enable hilling:** ☐ (unchecked).
- Hide advanced settings:** [Hide advanced settings](#).

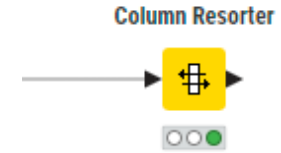
The resulting data table is shown below:

	name	age	city	column4
	String	Number (double)	String	Number (long)
0	Zé	19		
1	Maria	21		
2	Rui		Guimarães	919
3	Bela		Porto	926
4	Joana		Guimarães	938



- 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;

Ordenação *Sorter*





Filtros

Filter

Repository > Results

filter

Manipulation Filter Row Column +29

Row Filter Column Filter Nominal Value Row Filter

Reference Row Filter Top k Row Filter Duplicate Row Filter

Rule-based Row Filter Reference Column Filter Date&Time-based Row Filter

Missing Value Column Filter Sorter CSV Writer

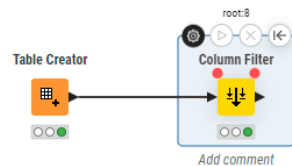
More advanced nodes

Filter Apply DB Row Filter DB Column Filter



Filtro de Colunas

Column Filter



Dialog - 4:8 - Column Filter

Column filter

Manual Wildcard Regex Type

Search Aa

Excludes

id
city

Includes

name
phone

> >> < <<

Any unknown columns

Cancel Ok

► 1: Filtered table Flow Variables

Rows: 3 | Columns: 2

#	Row...	name <small>String</small>	phone <small>Number (long)</small>
1	Row0	Rui	919
2	Row1	Bela	926
3	Row2	Joana	938



Filtro de Linhas

Row Filter

Diagram showing a workflow: Table Creator → Row Filter (root:7) → Add comment.

Dialog - 4:7 - Row Filter

File

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

Column value matching

Column to test: city

☐ filter based on collection elements

Matching criteria

☒ use pattern matching

Porto

☒ case sensitive match ☐ contains wild cards

☐ regular expression

☐ use range checking

lower bound:

upper bound:

☐ only missing values match

☒ 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

OK Apply Cancel ?

► 1: Filtered Flow Variables

Rows: 1 | Columns: 4

#	Row...	id	city	name
		Number (long)	String	String
1	Row1	12	Porto	Bela



Filtro de Linhas de Valores Nominais

Nominal Value Row Filter

Diagram illustrating the Nominal Value Row Filter configuration:

The workflow shows a **Table Creator** connected to a **Nominal Value Row Filter** (root:9). The filter is configured to filter rows based on the **city** column.

The configuration dialog, titled "Dialog - 4:9 - Nominal Value Row Filter", shows the following settings:

- Select column:** city
- Manual Selection** (selected) / Wildcard/Regex Selection
- Exclude** (red box):
 - Filter: Porto
 - ☐ Enforce exclusion
- Include** (green box):
 - Filter: Guimarães
 - ☒ Enforce inclusion

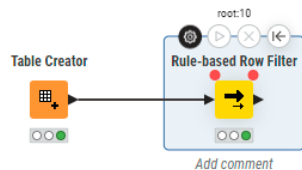
The resulting data table shows 2 rows and 4 columns:

#	Row...	id Number (long)	city String	name String	phone Number (long)
1	Row0	11	Guimarães	Rui	919
2	Row2	13	Guimarães	Joana	938



Filtro de Linhas Baseado em Regras

Rule-based Row Filter



Dialog - 4:10 - Rule-based Row Filter

File

Rule Editor | Flow Variables | Job Manager Selection | Memory Policy

Column List

- ROWID
- ROWINDEX
- ROWCOUNT
- L id
- S city
- S name
- L phone

Category

All

Function

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

Expression

```
1 // enter ordered set of rules, e.g.:
2 // TRUE => TRUE
3 $name$ = "Joana" AND $city$ = "Guimarães" => TRUE
4 $phone$ > 920 => TRUE
```

☒ Include TRUE matches ☐ Exclude TRUE matches

OK Apply Cancel ?

► 1: Filtered Flow Variables

Rows: 2 | Columns: 4

#	Row...	id Number (long)	city String	name String	phone Number
1	Row1	12	Porto	Bela	926
2	Row2	13	Guimarães	Joana	938



Filtro de Linhas em JAVA Snippet

JAVA Snippet Row Filter

Diagram showing a workflow with a **Table Creator** tool connected to a **Java Snippet Row Filter** tool. The filter tool has a comment "Add comment" and a label "root:11".

The **Dialog - 4:11 - Java Snippet Row Filter** is open, showing the following configuration:

- File** tab selected.
- Column List**: ROWID, ROWINDEX, ROWCOUNT, L id, S city, S name, I phone.
- Flow Variable List**: \$kname.workspace.
- Global Variable Declaration**: (Empty)
- Method Body**:

```
boolean result = false;

if( ($name$.equals("Joana") && $city$.equals("Guimarães")) || ($phone$ > 920) )
    result = true;

return result;
```
- Options**: ☐ Insert Missing As Null, ☒ Compile on close.
- Buttons**: OK, Apply, Cancel, ?

Below the dialog, the workflow results are displayed:

► 1: True match Flow Variables

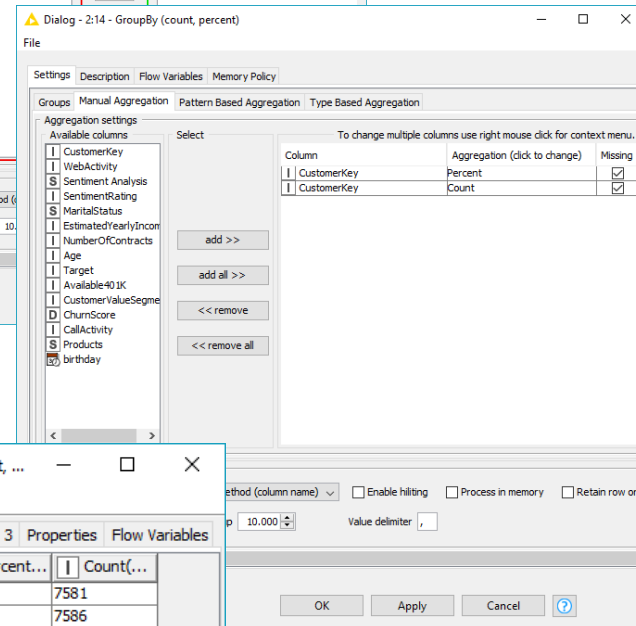
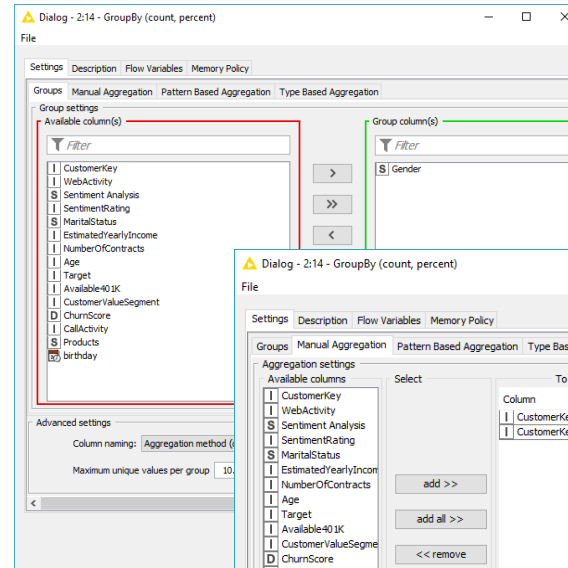
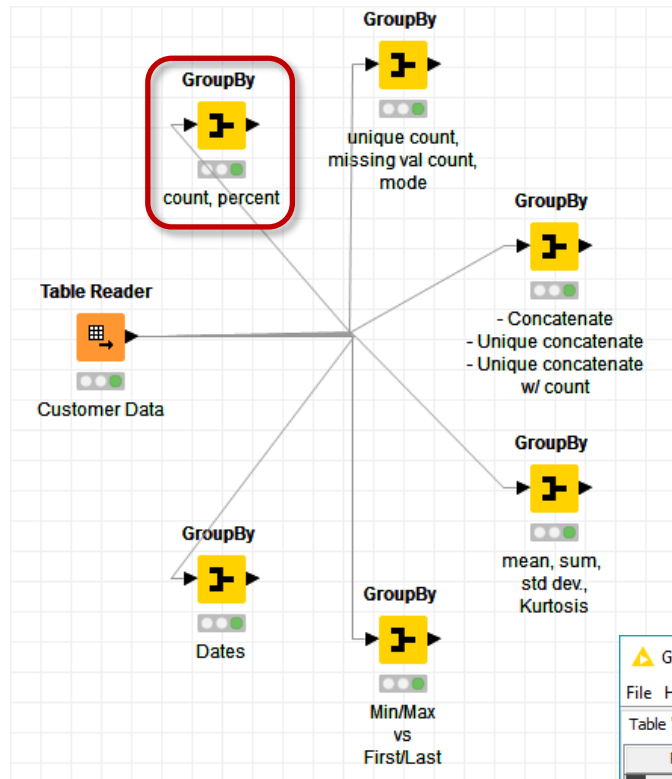
Rows: 2 | Columns: 4

#	Row...	id Number (long)	city String	name String	phone Number (integer)
1	Row1	12	Porto	Bela	926
2	Row2	13	Guimarães	Joana	938



Operações de Agregação

Count and Percent



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

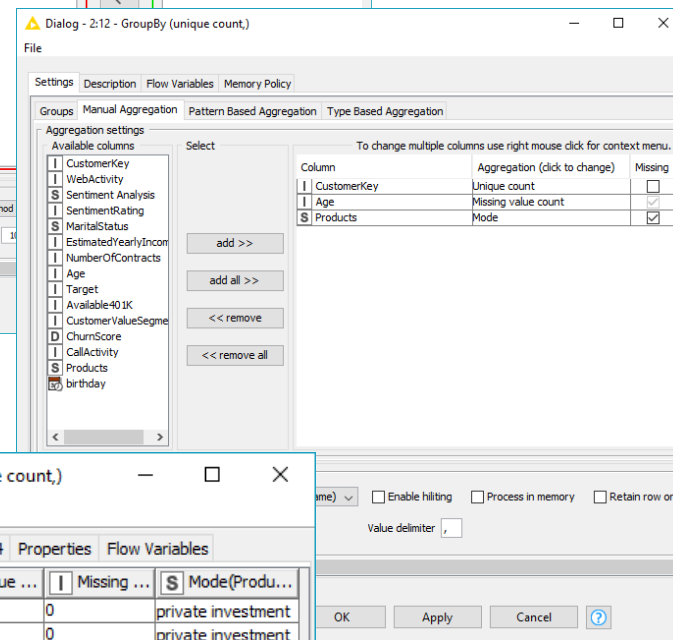
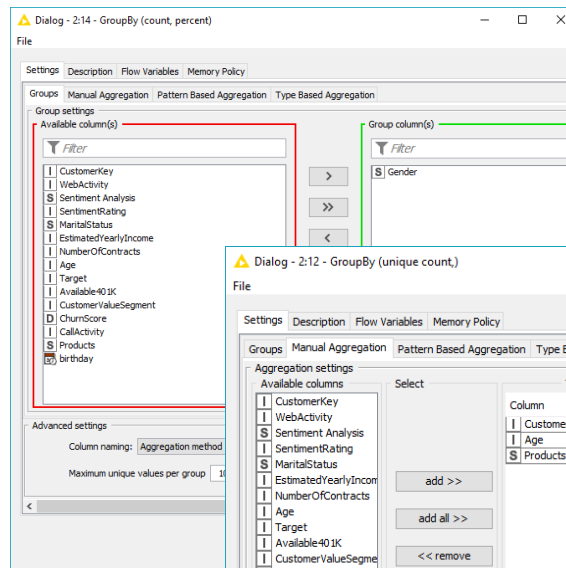
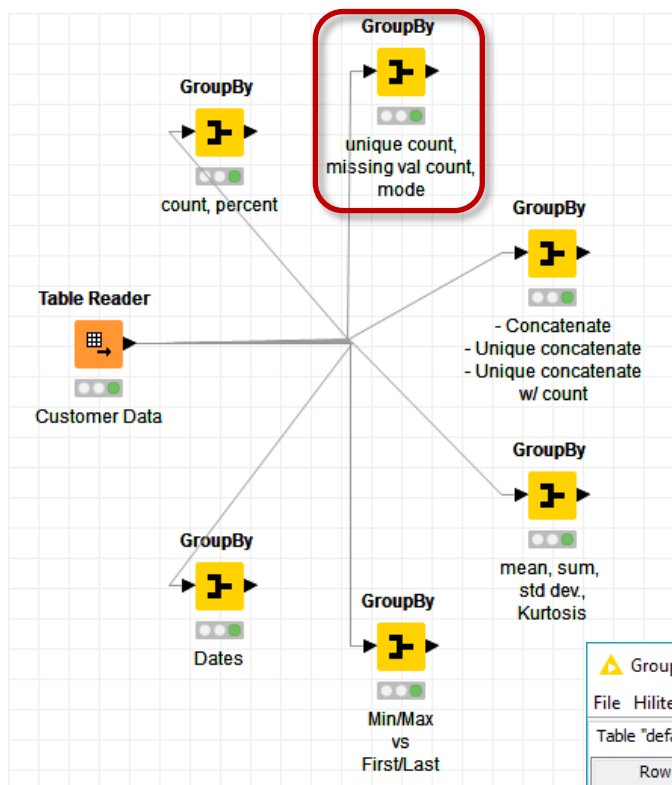
Table "default" - Rows: 2 Spec - Columns: 3 Properties Flow Variables

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

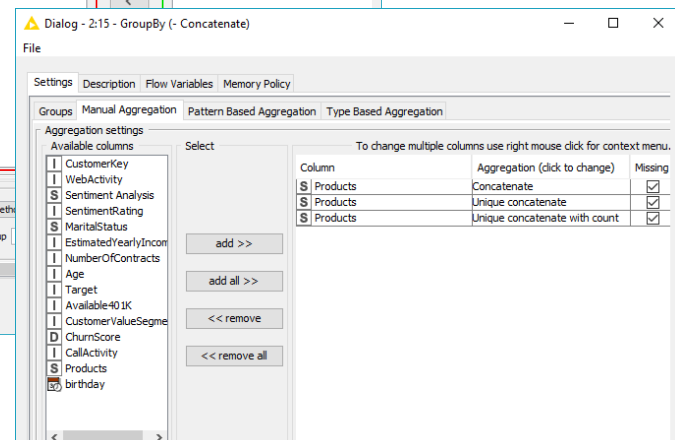
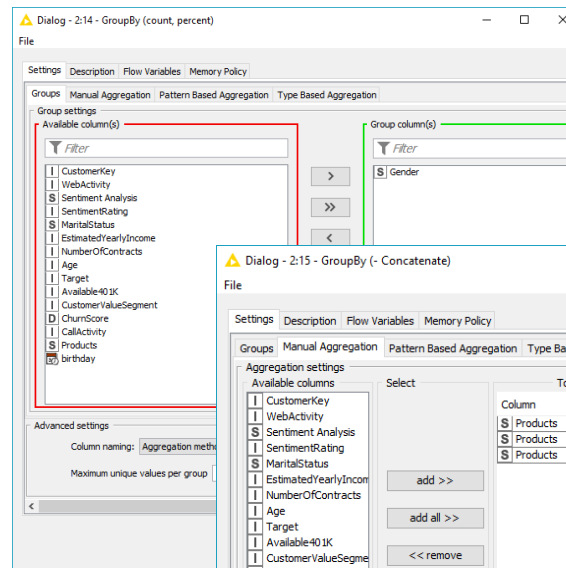
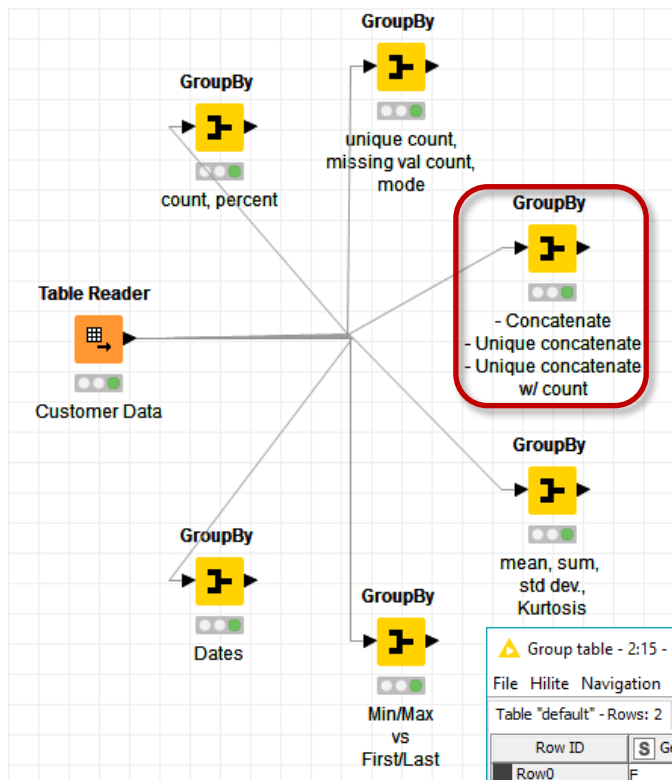
Table "default" - Rows: 2 Spec - Columns: 4 Properties Flow Variables

Row ID	Gender	Unique ...	Missing ...	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

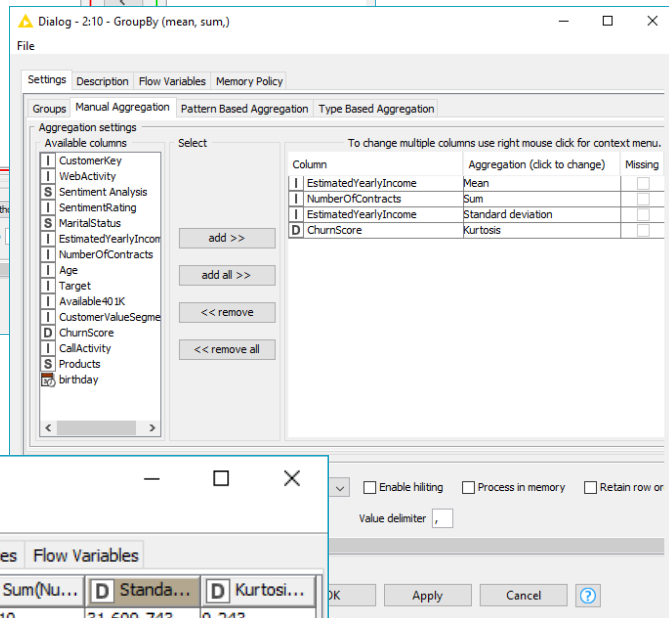
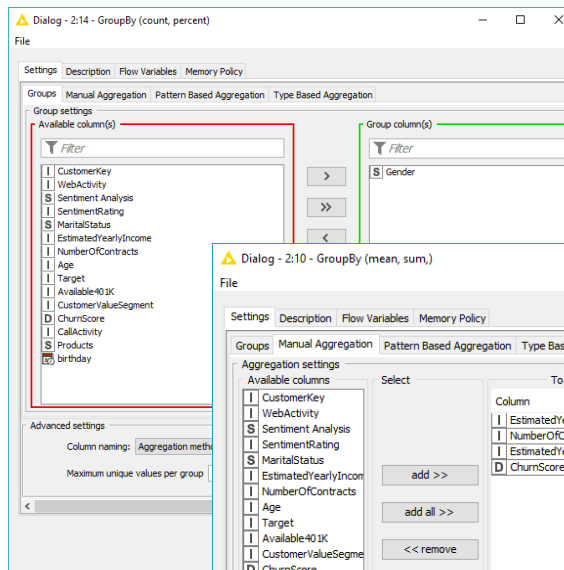
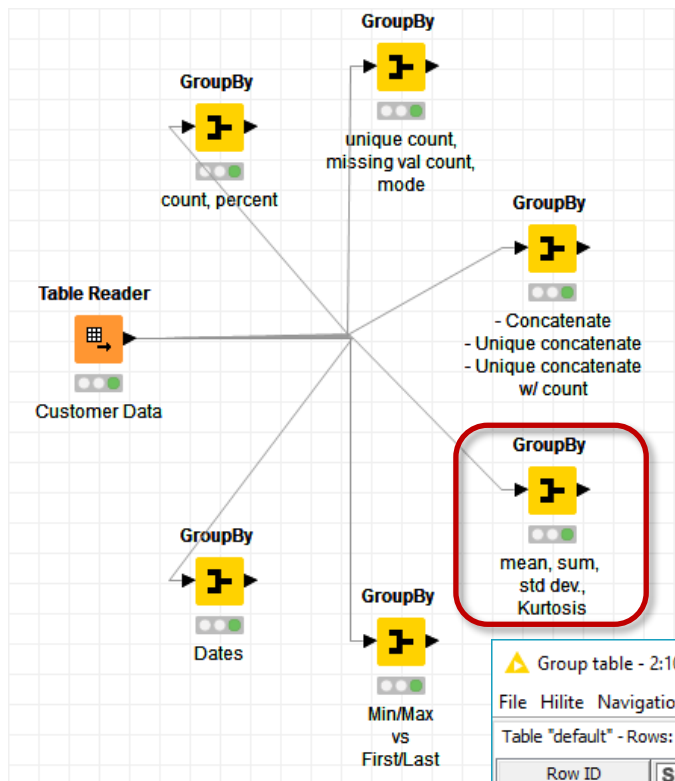
Table "default" - Rows: 2 Spec - Columns: 4 Properties Flow Variables

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,)

File Hilite Navigation View

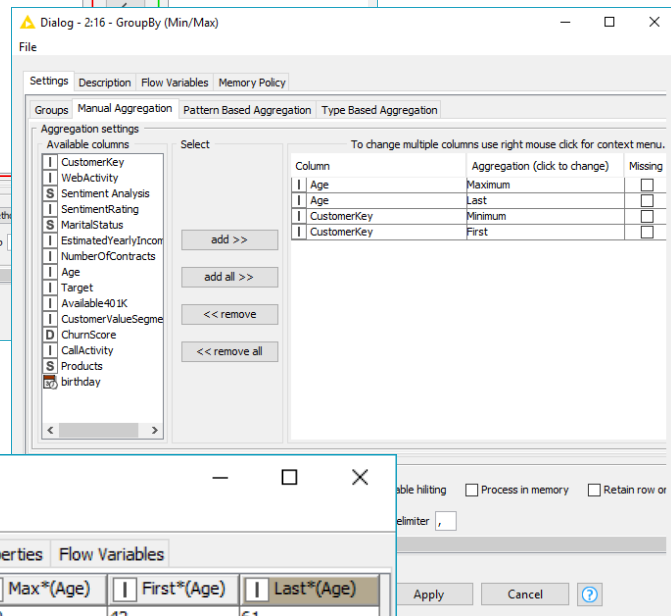
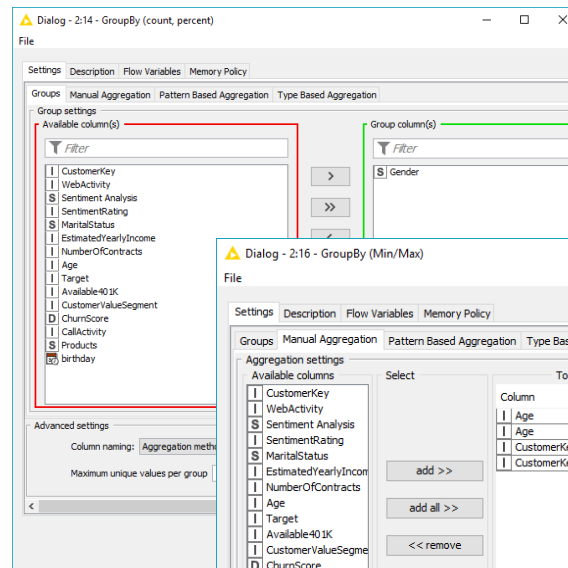
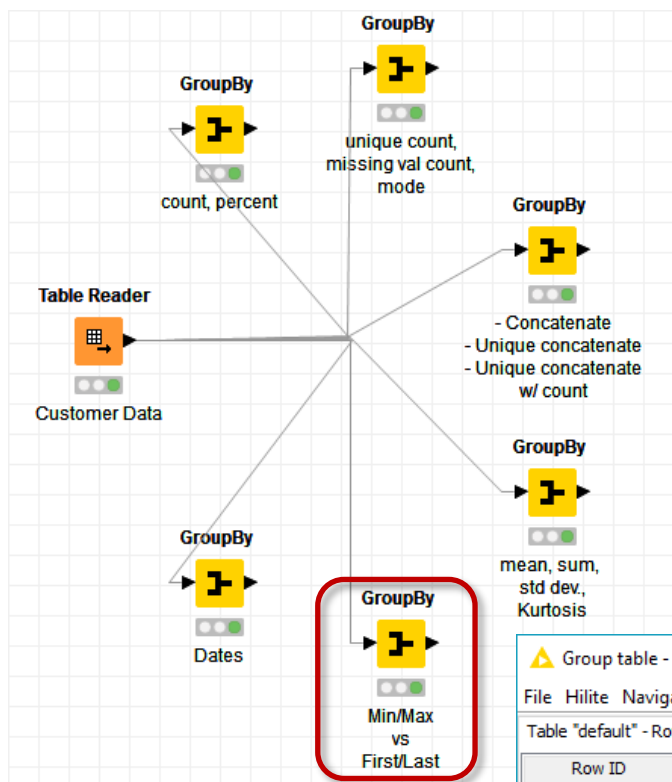
Table "default" - Rows: 2 Spec - Columns: 5 Properties Flow Variables

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)

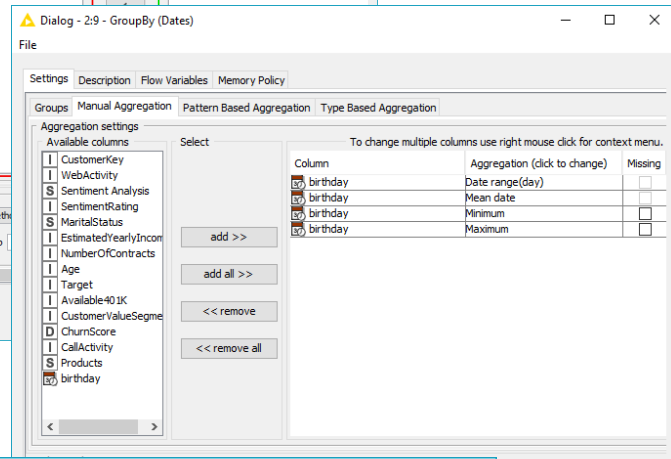
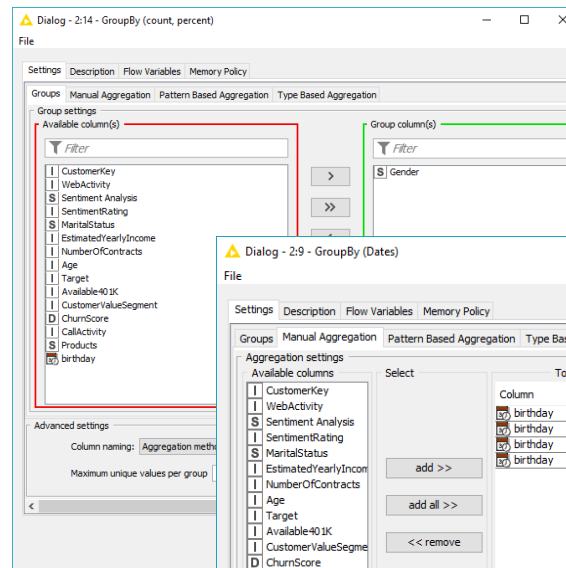
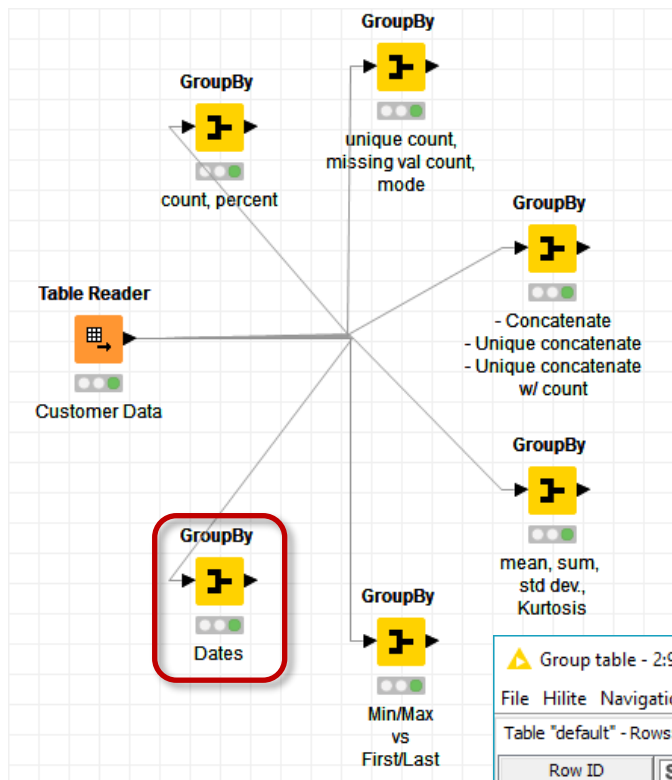
Table "default" - Rows: 2 Spec - Columns: 5 Properties Flow Variables

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



Operações de Agregação

Dates



Group table - 2:9 - GroupBy (Dates)

File Hilite Navigation View

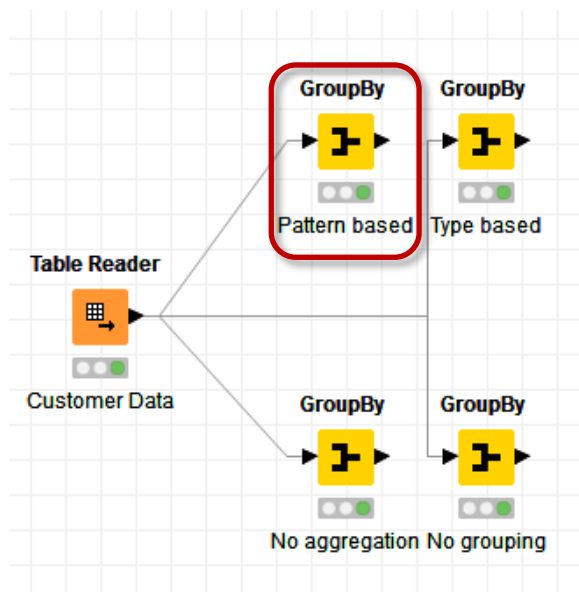
Table "default" - Rows: 2 Spec - Columns: 5 Properties Flow Variables

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



Dialog - 2:19 - GroupBy (Pattern based)

File

Settings | Description | Flow Variables | Memory Policy

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

Aggregation Settings

Search pattern (double click to change)	Regex	Aggregation methods
.*Estimated.*	<input checked="" type="checkbox"/>	Mean
NumberOf*	<input type="checkbox"/>	Sum
Sentiment	<input type="checkbox"/>	Unique concatenate with count

Add

Remove

Remove all

Advanced settings

Column naming: Aggregation method (column name) ☐ Enable hilling ☐ Process in memory

Maximum unique values per group: 10,000 Value delimiter: ,

< [OK] [Apply] [Cancel] [?]

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

File

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

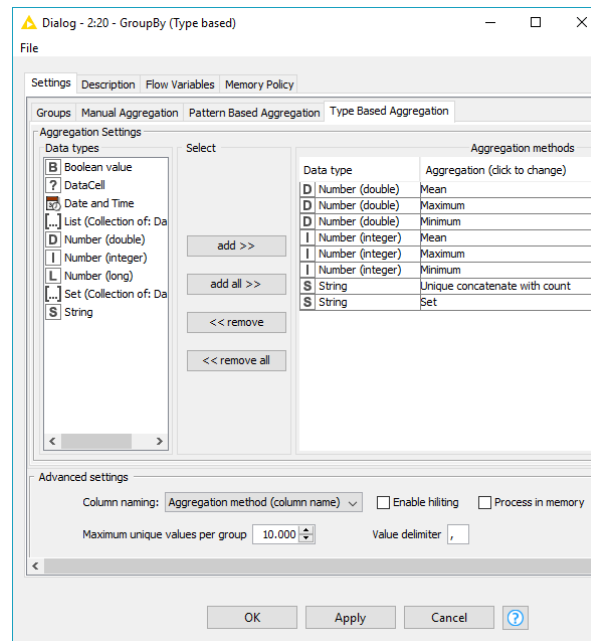
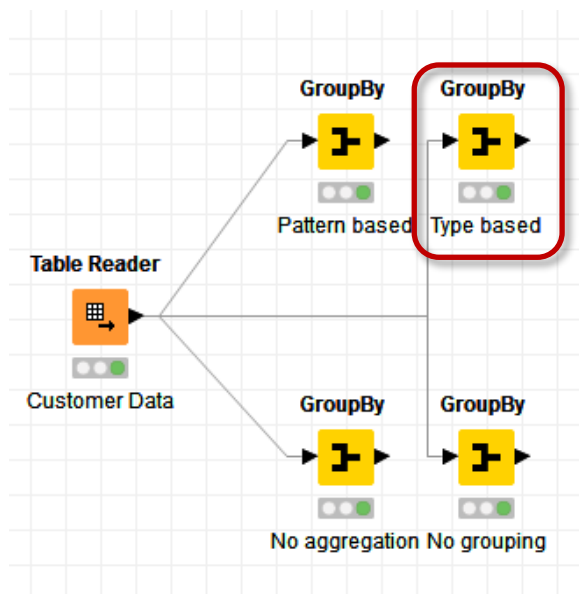
Columns: 5	Column Type	Column Index	Color Handler
Gender	String	0	
Unique concatenate with count(Sentiment Analysis)	String	1	
Unique concatenate with count(SentimentRating)	String	2	
Mean(EstimatedYearlyIncome)	Number (do...	3	
Sum(NumberOfContracts)	Number (int...	4	

< [] >



Operações Avançadas de Agregação

Data Type Based



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

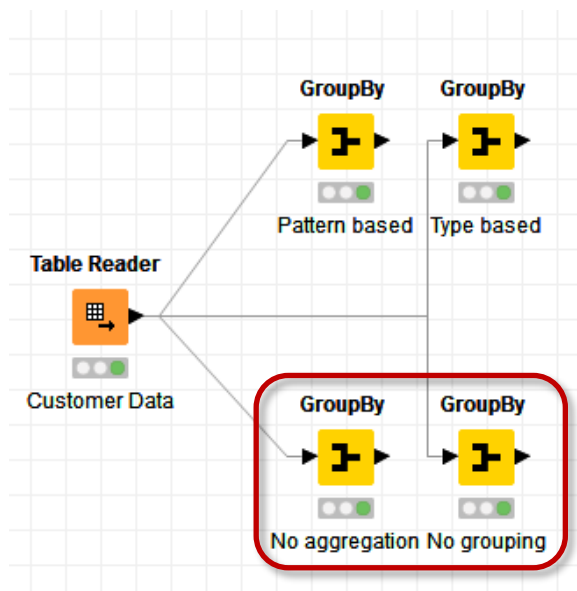
Table "default" - Rows: 2 | Spec - Columns: 70 | Properties | Flow Variables

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



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 - 2:18 - 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)



Preparação Avançada de Dados

- Redimensionamento de atributos/ *Feature scaling*
- Detecção de valores atípicos/ *Outlier detection*
- Seleção de atributos/ *Feature selection*
- Tratamento de valores em falta/ *Missing values treatment*
- Enumeração de valores nominais/ *Nominal value discretization*
- Divisão em intervalos/ *Binning*
- Engenharia de atributos/ *Feature engineering*

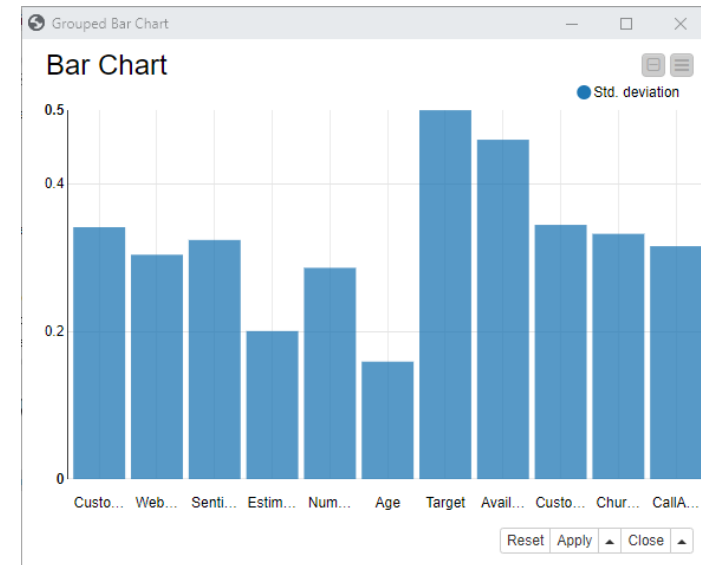
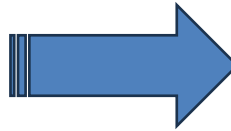
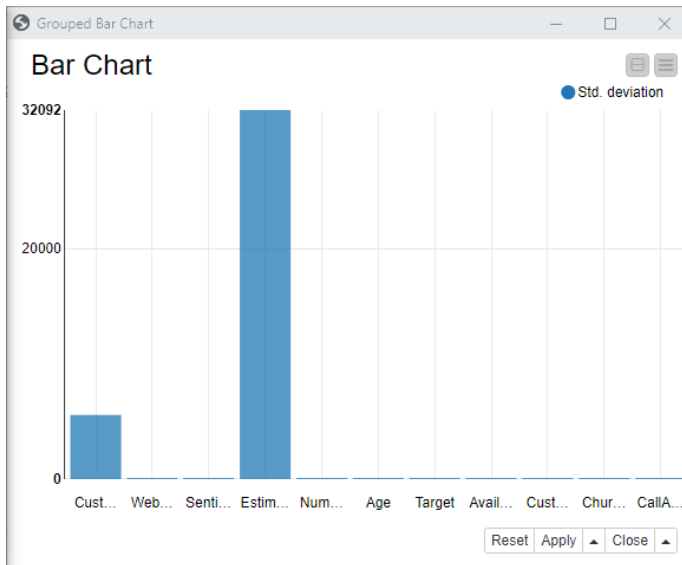




Redimensionamento de atributos

Feature scaling

- Normalizar a gama de valores de atributos;
- Muitos classificadores usam métricas de distância (ex.: distância euclidiana) e, se um atributo tiver uma gama alargada de valores, a distância será definida por esse atributo em particular. Por isso, a gama de valores deve ser normalizada para que cada atributo possa contribuir proporcionalmente para a distância final.

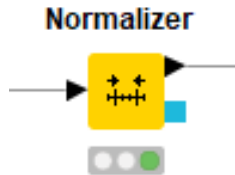




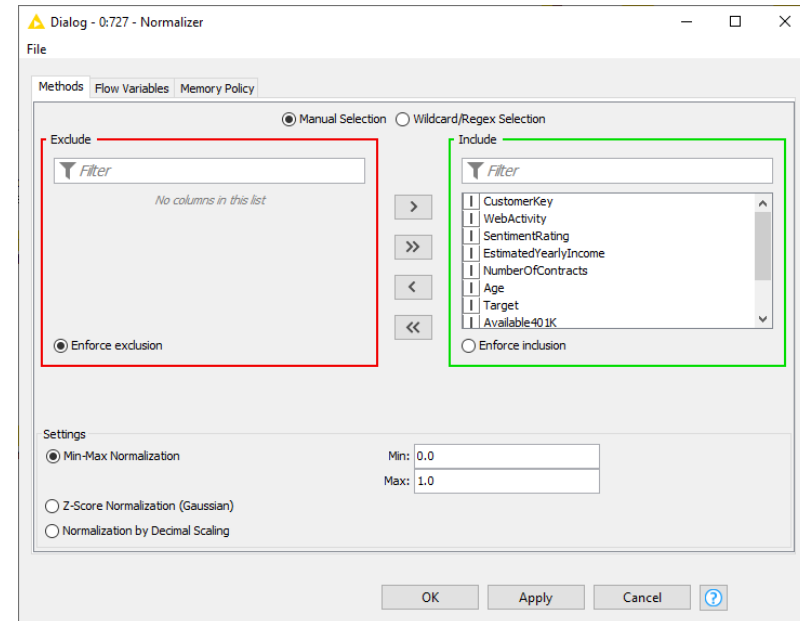
Redimensionamento de atributos

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

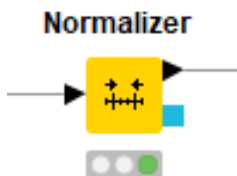
Feature scaling

- Normalizar a gama de valores de atributos:

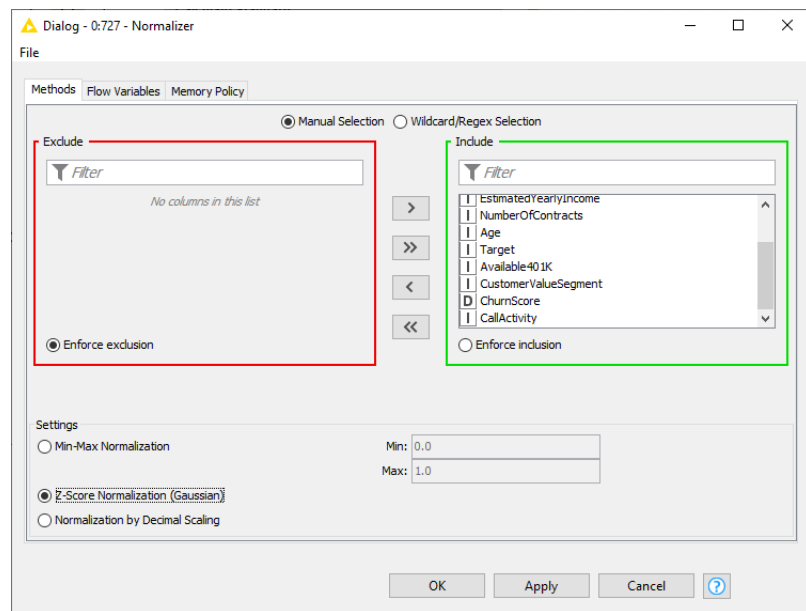
- Standardization* (ou *Z-score Normalization*):

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.



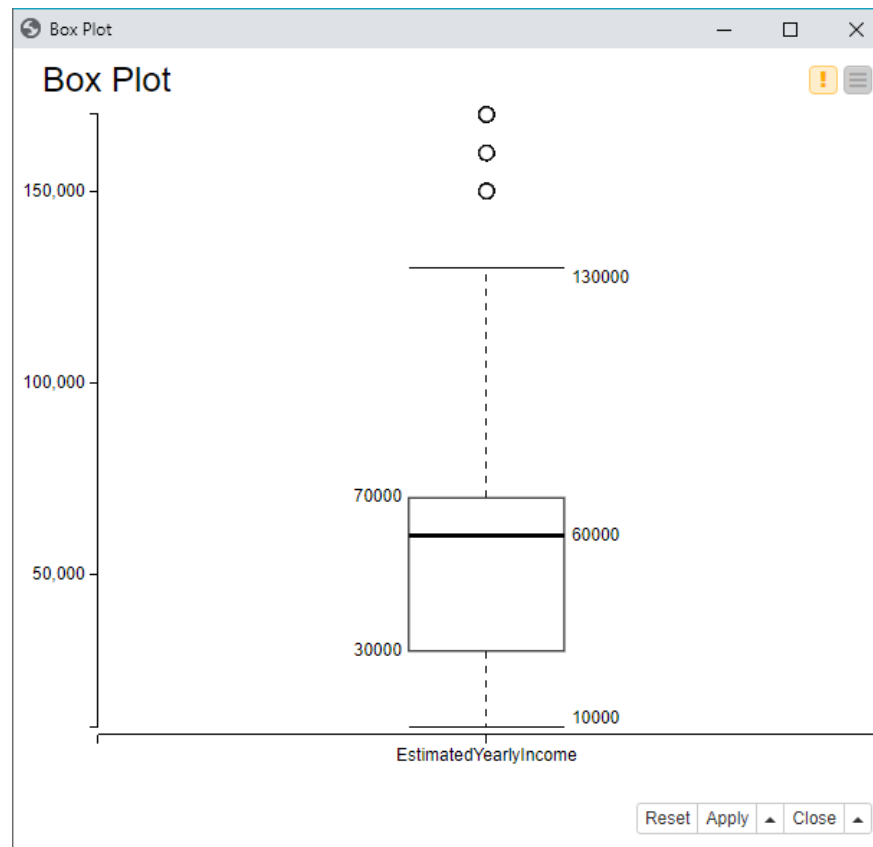
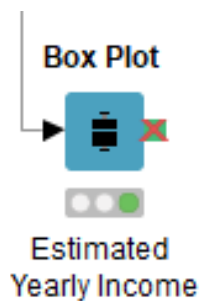
$$z = \frac{x_i - \mu}{\sigma}$$





Deteção de valores atípicos *Outlier Detection*

- Deteção de valores atípicos (*outliers*):
 - Estratégias baseadas em estatística:
 - Box Plots
 - Z-Score (std. dev)



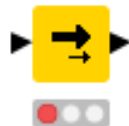


Deteção de valores atípicos

Outlier Detection

- Deteção de valores atípicos (*outliers*):
 - Estratégias baseadas em conhecimento

Rule-based
Row Filter



Dialog - 0:725 - Rule-based Row Filter (Delete row if)

File

Rule Editor | Flow Variables | Memory Policy

Column List

- ROWID
- ROWINDEX
- ROWCOUNT
- CustomerKey
- WebActivity
- Sentiment Analysis
- SentimentRating
- MaritalStatus
- Gender
- EstimatedYearlyIncome
- NumberOfContracts
- Age
- Target
- Available401K
- CustomerValueSegment

Flow Variable List

- knime.workspace

Category

All

Function

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

Description

Expression

```
1 // enter ordered set of rules, e.g.:
2 // $double column name$ > 5.0 => FALSE
3 // $string column name$ LIKE "*blue*" => FALSE
4 // TRUE => TRUE
5 $EstimatedYearlyIncome$ <= 10000 OR $EstimatedYearlyIncome$ >= 130000 => FALSE
6 TRUE => TRUE
```

☒ Include TRUE matches ☐ Exclude TRUE matches

OK Apply Cancel ?

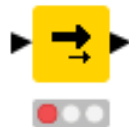


Deteção de valores atípicos

Outlier Detection

- Deteção de valores atípicos (*outliers*):
 - Estratégias baseadas em conhecimento

Rule-based
Row Filter



Dialog - 0:725 - Rule-based Row Filter (Delete row if)

File

Rule Editor Flow Variables Memory Policy

Column List

- ROWID
- ROWINDEX
- ROWCOUNT
- CustomerKey
- WebActivity
- Sentiment Analysis
- SentimentRating
- MaritalStatus
- Gender
- EstimatedYearlyIncome
- NumberOfContracts
- Age
- Target
- Available401K
- CustomerValueSegment

Flow Variable List

- knime.workspace

Category

All

Function

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

Description

Expression

```
1 // enter ordered set of rules, e.g.:
2 // $double column name$ > 5.0 => FALSE
3 // $string column name$ LIKE "*blue*" => FALSE
4 // TRUE => TRUE
5 $EstimatedYearlyIncome$ <= 10000 OR $EstimatedYearlyIncome$ >= 130000 => TRUE
```

☐ Include TRUE matches ☒ Exclude TRUE matches

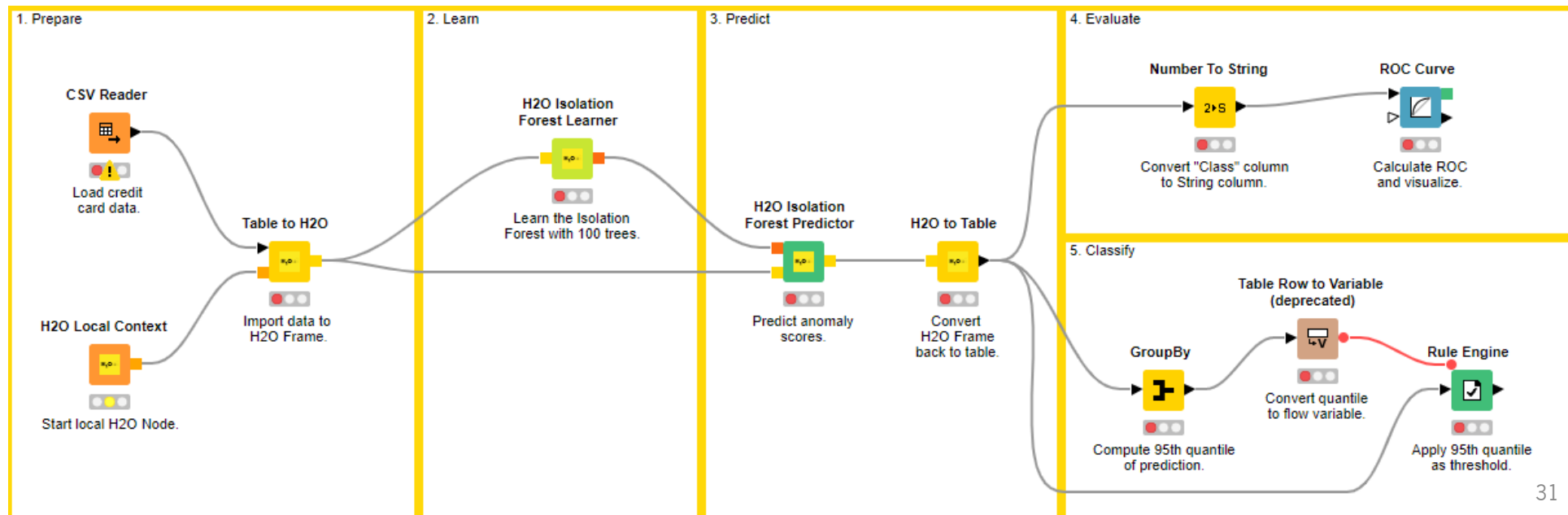
OK Apply Cancel ?



Deteção de valores atípicos

Outlier Detection

- Deteção de valores atípicos (*outliers*):
 - Estratégias baseadas em modelos:
 - Isolation Forest
 - One-Class SVM
 - Minimum Covariance Determinant
 - ...





Seleção de Atributos

Feature Selection

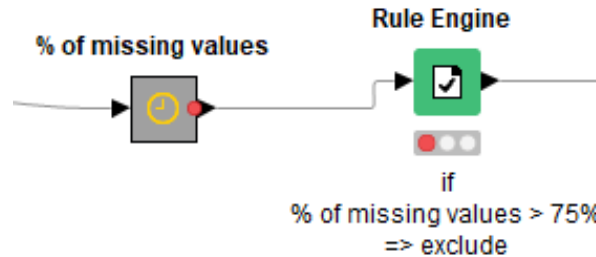
- Seleção de atributos:
 - Que atributos devem ser usados para criar um modelo de previsão?
 - Selecionar um subconjunto de atributos mais importantes para reduzir a dimensionalidade;
 - A remoção de atributos pouco importantes:
 - Pode afetar significativamente o desempenho de um modelo;
 - Reduz o *overfitting* (menor probabilidade de tomar decisões com base em ruído);
 - Melhora a precisão;
 - Ajuda a reduzir a complexidade de um modelo (reduz o tempo de treino);
 - O que podemos remover:
 - Atributos redundantes (duplicados);
 - Atributos irrelevantes e desnecessários (não úteis);
 - Métodos de seleção de atributos:
 - Métodos de filtro;
 - Métodos *wrapper*;
 - Métodos embebidos;



Seleção de Atributos *Feature Selection*

■ Métodos de Filtro:

- Remover uma *feature* se a percentagem de *missing values* for superior a um determinado valor estabelecido;



- Usar o teste “chi-square” para medir o grau de dependência entre uma *feature* e o *target*.
 - Para cada *feature* calcular X^2 ;
 - Normalizar X^2 e ordenar de forma decrescente;
 - Selecionar 'n' *features* com maior importância;
(ou as que estão acima de um determinado limite)

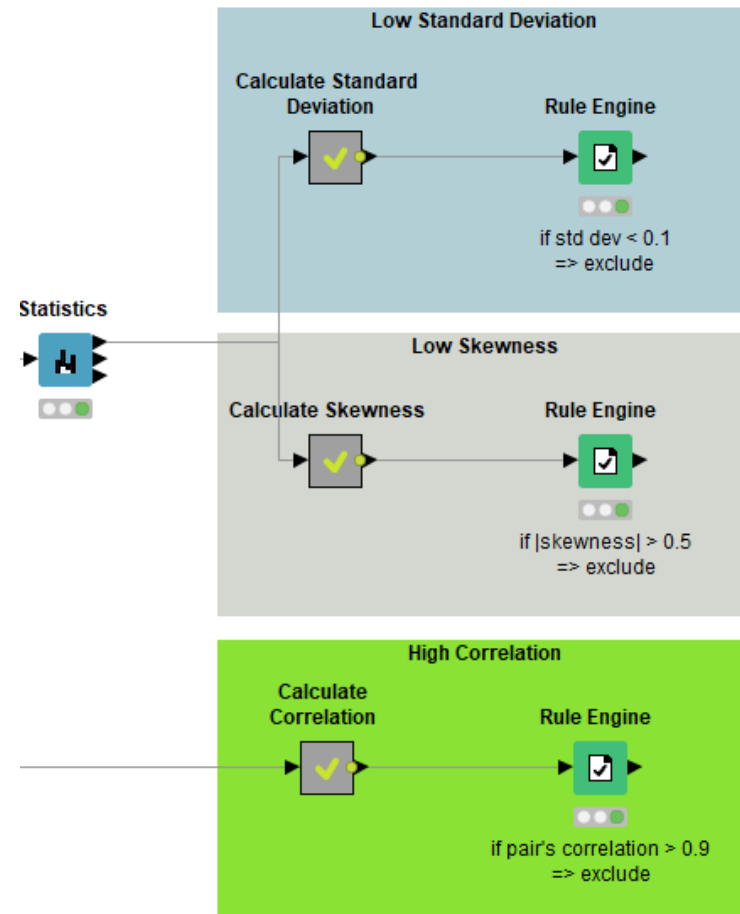
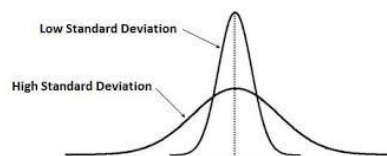
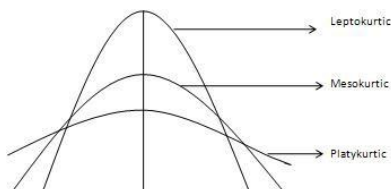
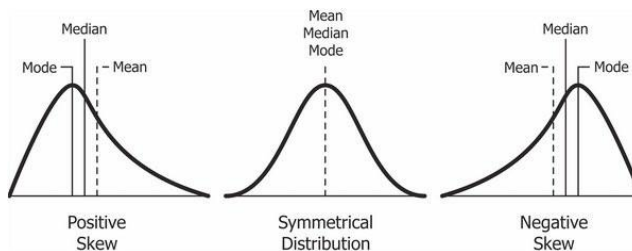


Seleção de Atributos

Feature Selection

■ Métodos de Filtro:

- ...
- Remover uma *feature* se o valor do desvio padrão for baixo;
- Remover uma *feature* se o valor de *skew* for elevado;
- Remover *features* com alta correlação;



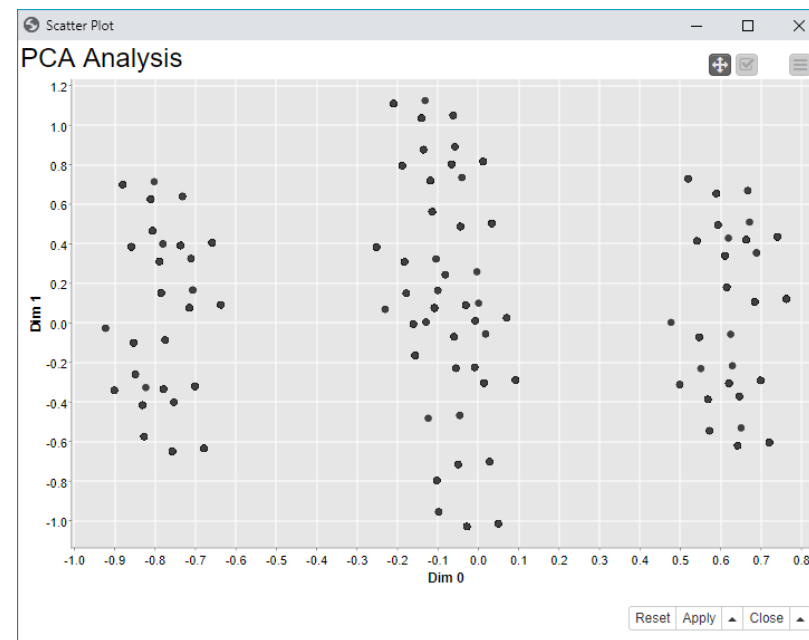
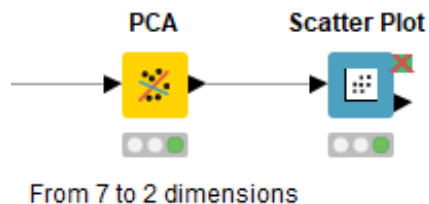


Seleção de Atributos

Feature Selection

■ Métodos de Filtro:

- ...
- Principal Component Analysis (PCA):
 - Técnica usada para reduzir a dimensão do espaço de *features*;
 - O objetivo é reduzir o número de *features* sem perder (demasiado) conhecimento;
 - Uma aplicação comum de PCA é para visualização de dados de grande dimensão;





Seleção de Atributos

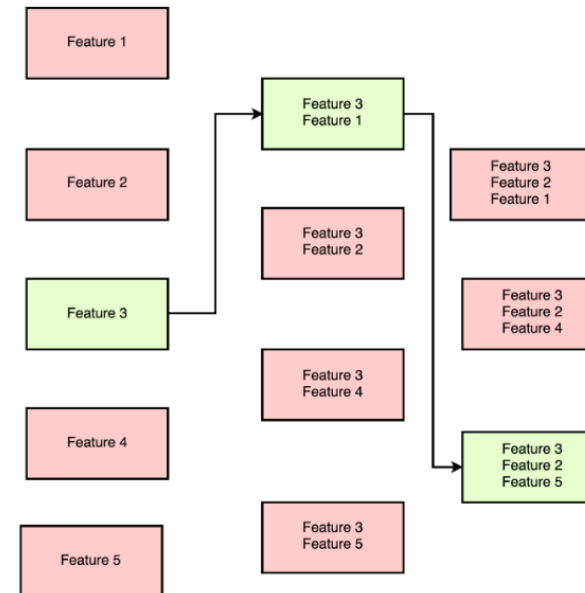
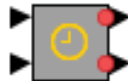
Feature Selection

■ Métodos *wrapper*.

- Utilizar técnicas de *machine learning* para selecionar as *features* mais importantes;
 - Selecionar um conjunto de *features* como um problema de pesquisa;
 - Preparar diferentes combinações;
 - Avaliar e comparar as diferentes combinações;
 - Medir a “utilidade” das *features* com base no desempenho do classificador;

○ *Sequential Forward Selection*

Forward Feature Selection





Seleção de Atributos

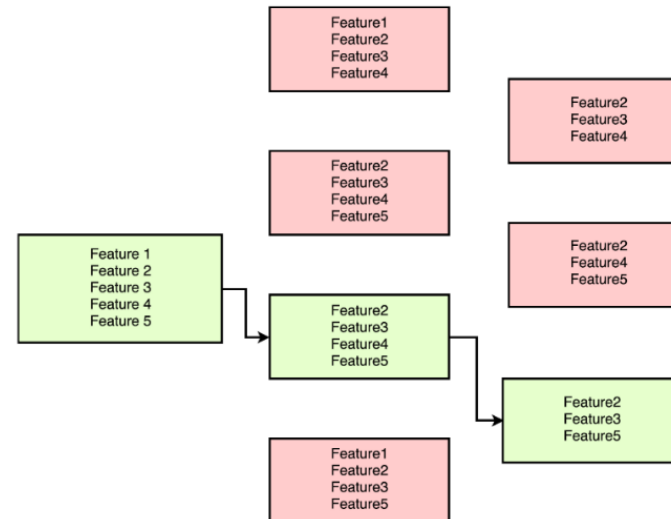
Feature Selection

■ Métodos *wrapper*.

- Utilizar técnicas de *machine learning* para selecionar as *features* mais importantes;
 - Selecionar um conjunto de *features* como um problema de pesquisa;
 - Preparar diferentes combinações;
 - Avaliar e comparar as diferentes combinações;
 - Medir a “utilidade” das *features* com base no desempenho do classificador;

○ *Backward Feature Elimination*

Backward Feature Elimination





Tratamento de Valores em Falta

Missing Values

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



Dialog - 0:731 - Missing Value

File

Default Column Settings Flow Variables Memory Policy

Number (integer)	Fix Value Value -99
String	Do nothing
Number (double)	Mean
Date and Time	Remove Row*

Options marked with an asterisk (*) will result in non-standard PMML.

OK Apply Cancel ?



Enumeração de Valores Nominais ***Nominal Value Discretization/Encoding***

- Enumeração de valores nominais:
 - Os dados categóricos/nominais contêm valores de “etiquetas” em vez de valores numéricos;
 - Podem ser aplicados vários métodos:
 - *Label Encoding*,
 - *One-Hot Encoding*,
 - *Binary Encoding*,



Enumeração de Valores Nominais

Nominal Value Discretization/Encoding

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



Enumeração de Valores Nominais

Nominal Value Discretization/Encoding

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

Label Encoded

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



Enumeração de Valores Nominais

Nominal Value Discretization/Encoding

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
...	...		



Enumeração de Valores Nominais

Nominal Value Discretization/Encoding

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
...	...		

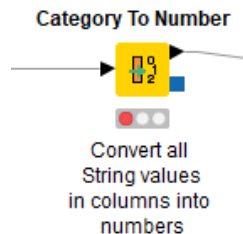
Categorical features where **no such ordinal relationship exists**. However, for a huge number of categories...



Enumeração de Valores Nominais

Nominal Value Discretization/Encoding

- Enumeração de valores nominais:



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:

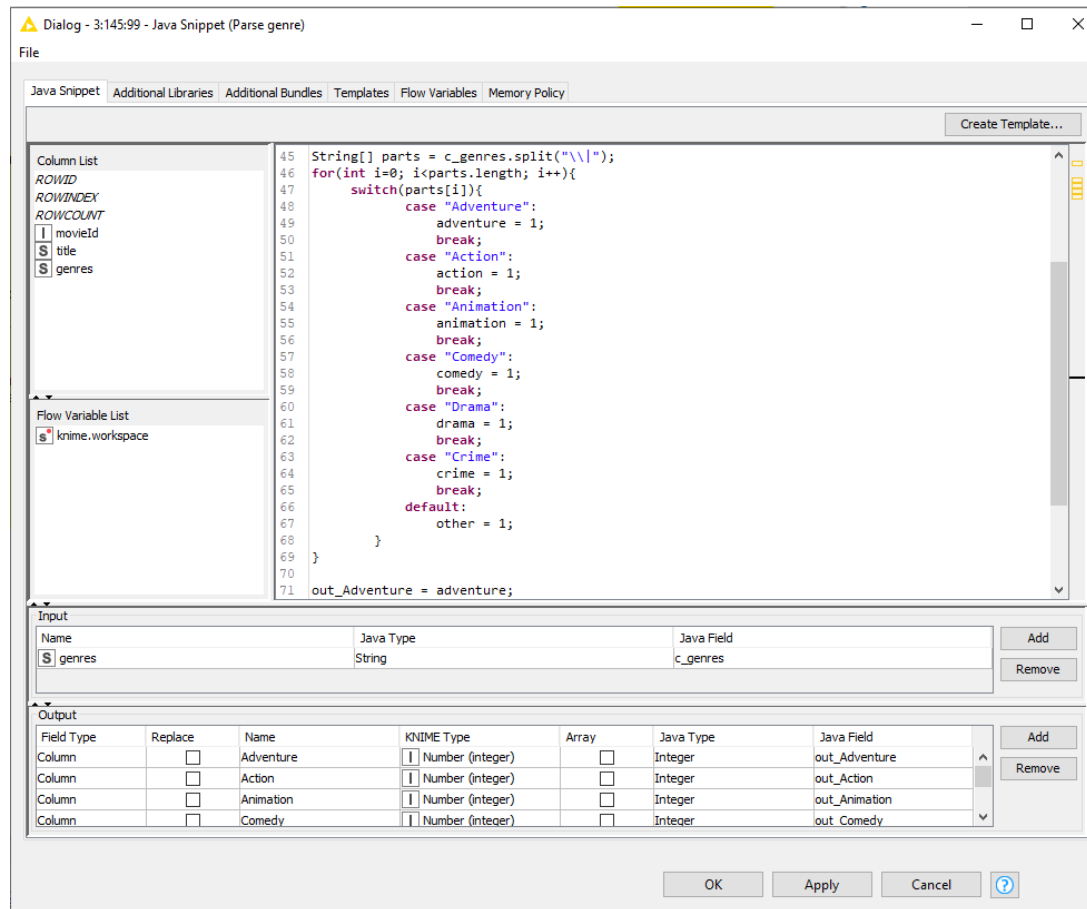
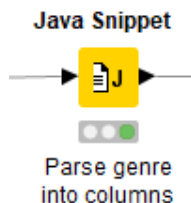
OK Apply Cancel ?



Enumeração de Valores Nominais

Nominal Value Discretization/Encoding

- Enumeração de valores nominais:

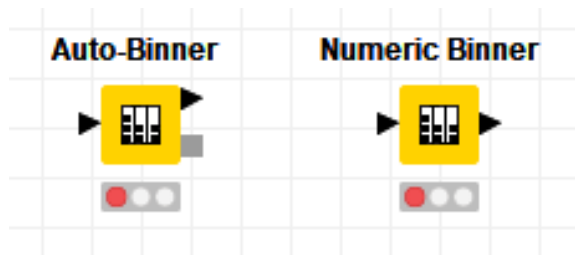




Divisão em Intervalos

Binning

- Divisão em intervalos:
 - Transformação de valores contínuos em discretos (*bins*):
 - Torna o modelo mais robusto e evita o *overfitting*;
 - Penaliza o desempenho do modelo, uma vez que, sempre que se descartam dados, perde-se conhecimento;



Dialog - 0:35 - Auto-Binner

File

Auto Binner Settings | Number Format Settings | Flow Variables | Memory Policy

☒ Manual Selection ☐ Wildcard/Regex Selection

Exclude

Filter

Store
Dept
Size

☒ Enforce exclusion

Include

Filter

D Weekly_Sales

☐ Enforce inclusion

Binning Method

☒ Fixed number of bins

Number of bins: 4

Equal: frequency

☐ Sample quantiles

Quantiles (comma separated): 0,0, 0,25, 0,5, 0,75, 1,0

Bin Naming

☒ Numbered e.g.: Bin 1, Bin 2, Bin 3

☐ Borders e.g.: [-10,0], (0,10], (10,20]

☐ Midpoints e.g.: -5, 5, 15

☐ Force integer bounds

☒ Replace target column(s)

OK Apply Cancel ?



Divisão em Intervalos

Binning

- Divisão em intervalos:
 - Transformação de valores contínuos em discretos (*bins*)
 - Torna o modelo mais robusto e evita o *overfitting*;
 - Penaliza o desempenho do modelo, uma vez que sempre que se descartam dados, perde-se conhecimento;

△ 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...	S Sentiment...	I Sentim...	S Marital...	S Gender	I Estim...	I Number...	S Age	I Target
Row0_Row0...	11000	0	Slightly Negative	2	M	M	90000	0	(39,	
Row0_Row86...	11000	0	Slightly Negative	2	M	M	90000	0	(39,	
Row1_Row1...	11001	3	Slightly Positive	3	S	M	60000	1	(39,	
Row1_Row86...	11001	3	Slightly Positive	3	S	M	60000	1	(39,46]	1
Row2_Row2...	11002	3	Slightly Positive	3				1	(39,46]	1
Row2_Row86...	11002	3	Slightly Positive	3				1	(39,46]	1
Row3_Row3...	11003	0	Very Negative	0				1	(39,46]	1
Row3_Row86...	11003	0	Very Negative	0				1	(39,46]	1
Row4_Row4...	11004	5	Very Positive	5				4	(39,46]	1
Row4_Row86...	11004	5	Very Positive	5				4	(39,46]	1
Row5_Row5...	11005	0	Very Negative	0				1	(39,46]	1
Row5_Row86...	11005	0	Very Negative	0				1	(39,46]	1
Row6_Row6...	11006	0	Very Negative	0				1	(39,46]	1
Row6_Row86...	11006	0	Very Negative	0				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

Show possible values
Available Renderers >



Engenharia de Atributos

Feature Engineering

- Engenharia de atributos:
 - Processo de criação de novos atributos (*features*);
 - Aumentar o conhecimento/aumentar o desempenho dos modelos;
- De um atributo do tipo “data”, que conhecimento se pode extrair?
 - 2020/10/29 16:30



Engenharia de Atributos

Feature Engineering

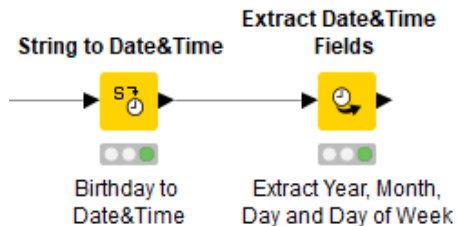
- Engenharia de atributos:
 - Processo de criação de novos atributos (*features*);
 - Aumentar o conhecimento/aumentar o desempenho dos modelos;
- De um atributo do tipo “data”, que conhecimento se pode extrair?
 - 2020/10/29 16:30
 - Ano, mês, dia
 - Horas, minutos, segundos
 - Dia da semana (quinta-feira)
 - Dia útil ou fim de semana?
 - Dia útil ou feriado?
 - ...



Engenharia de Atributos

Feature Engineering

- Engenharia de atributos:
 - Processo de criação de novos atributos (*features*);
 - Aumentar o conhecimento/aumentar o desempenho dos modelos;
- De um atributo do tipo “data”, que conhecimento se pode extrair?
 - 2020/10/29 16:30



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	ChurnS...	CallActi...	Products	birthday	Year	Month (number)	Day of year	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

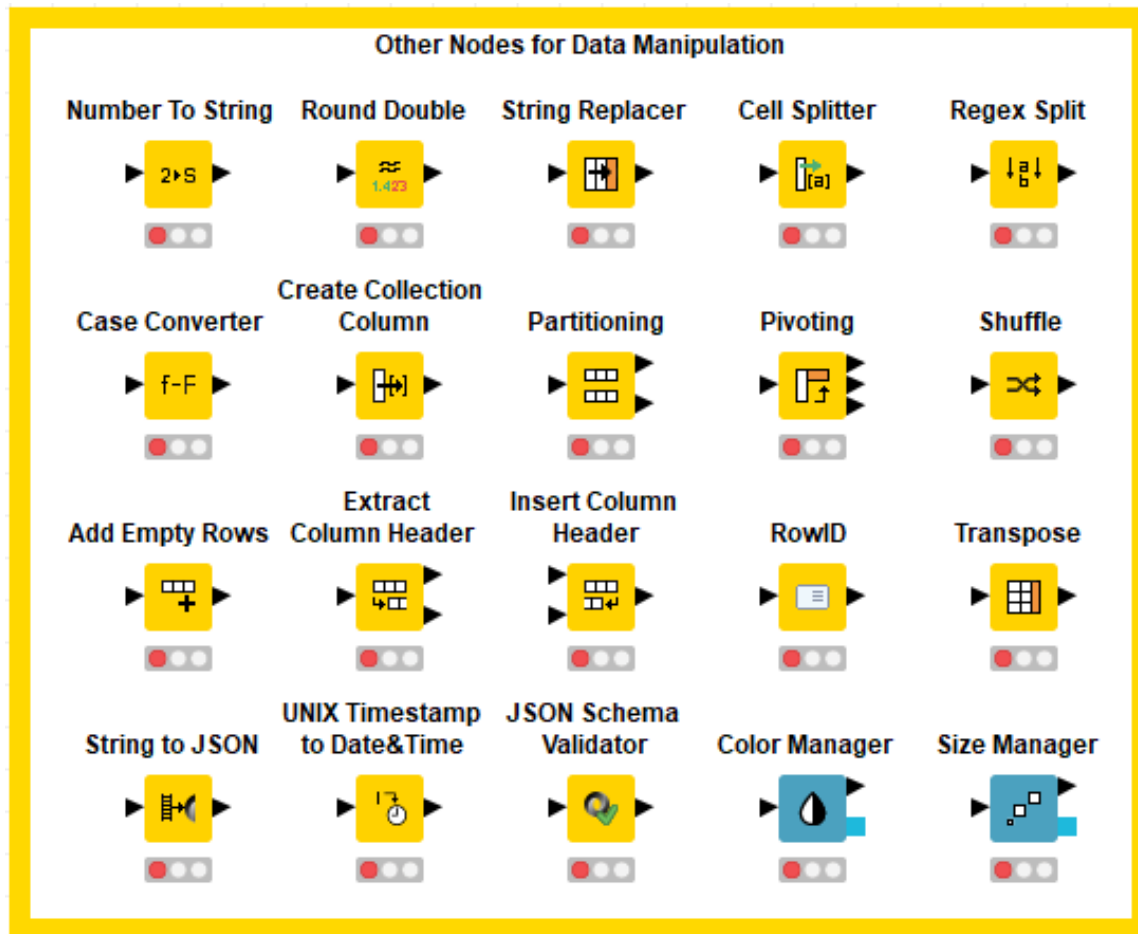


Engenharia de Atributos

Feature Engineering

- Engenharia de atributos:
 - Processo de criação de novos atributos (*features*);
 - Aumentar o conhecimento/aumentar o desempenho dos modelos;

- De um atributo com coordenadas geográficas?
 - 41.561859, -8.397455
 - Localização urbana ou rural?
 - Terra ou mar?
 - Quais as ruas na vizinhança deste ponto?
 - Há escolas/mercados/serviços nas imediações?
 - ...





Home

Building a Simple Classifier ×

Help

Preferences

Menu

Save

Undo

Redo

Execute all

Cancel all

Reset all

93%

Nodes

Search all nodes

IO

Excel Reader

Excel Writer

Microsoft Authenticator

CSV Reader

CSV Writer

Table Creator

SharePoint Online Connector

File Reader

Show all

Manipulation

Row Filter

Column Filter

Concatenate

Value Lookup

Row Aggregator

Table Splitter

String Cleaner

Table Cropper

Show all

Views

QUICK HANDS ON

To show the node output, please select a configured or executed node.