Universidade do Minho
Departamento de Informática

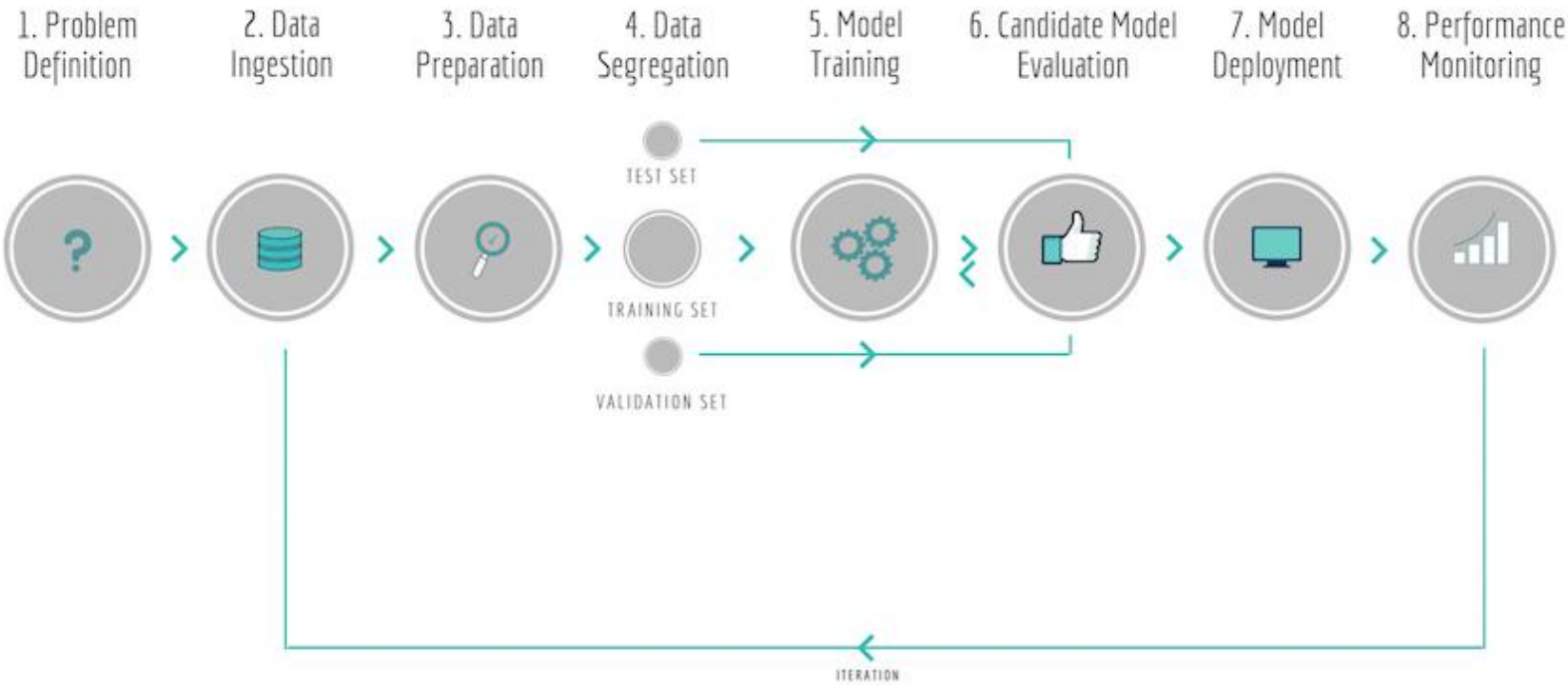
Preparação e Exploração avançada de dados com KNIME

LEI/MiEI @ 2024/2025, 2º sem

Agenda

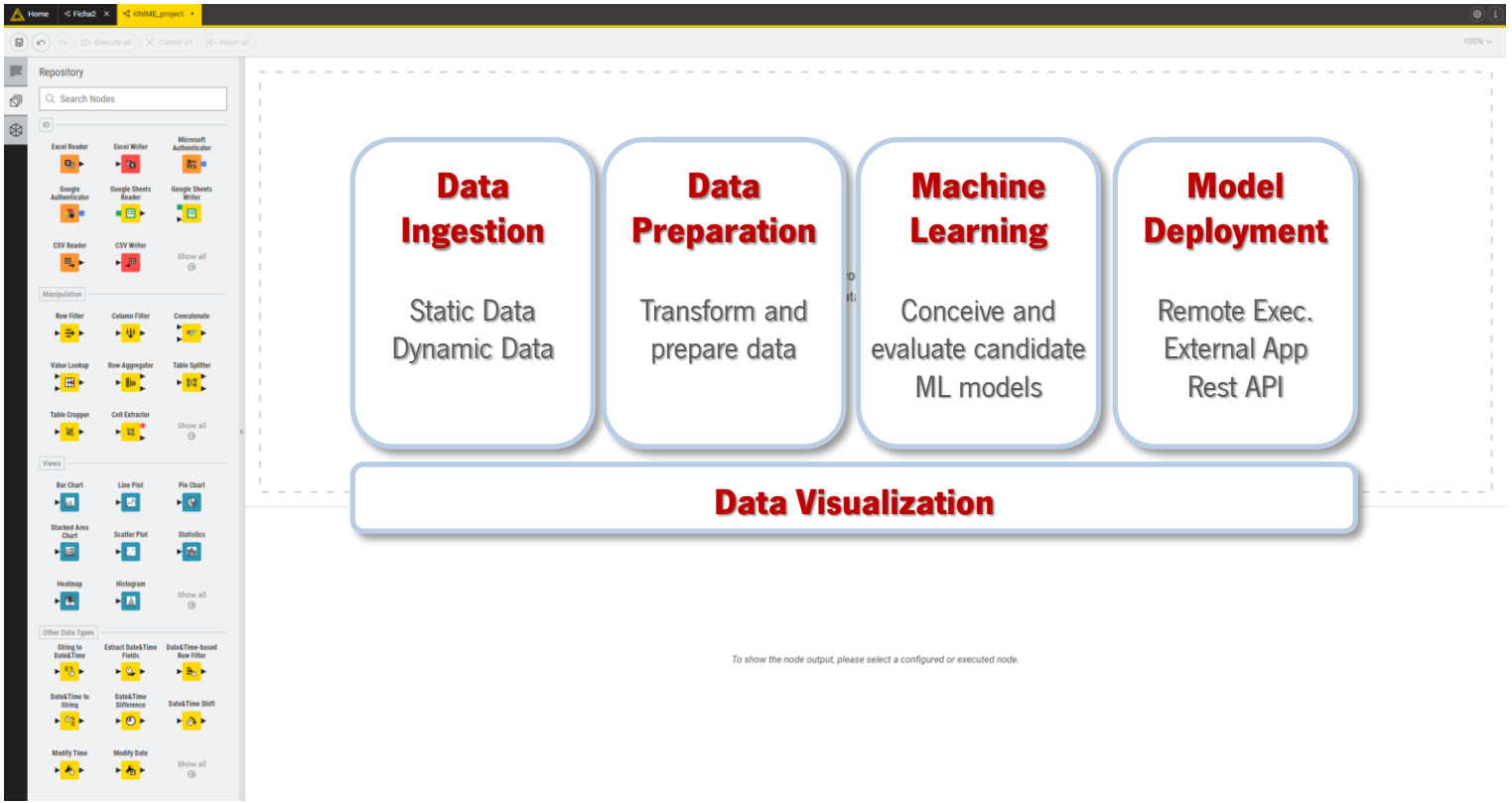
- Preparação de Dados
  - Join, Concatenation, Sorter, Filter and Aggregations
- Preparação e Exploração Avançada de Dados
  - Missing Values Treatment, Binning, Feature Scaling, Outlier Detection
  - Feature Selection, Nominal Value Discretization, Feature Engineering
- Experimentação (*hands on*)

A Machine Learning Pipeline



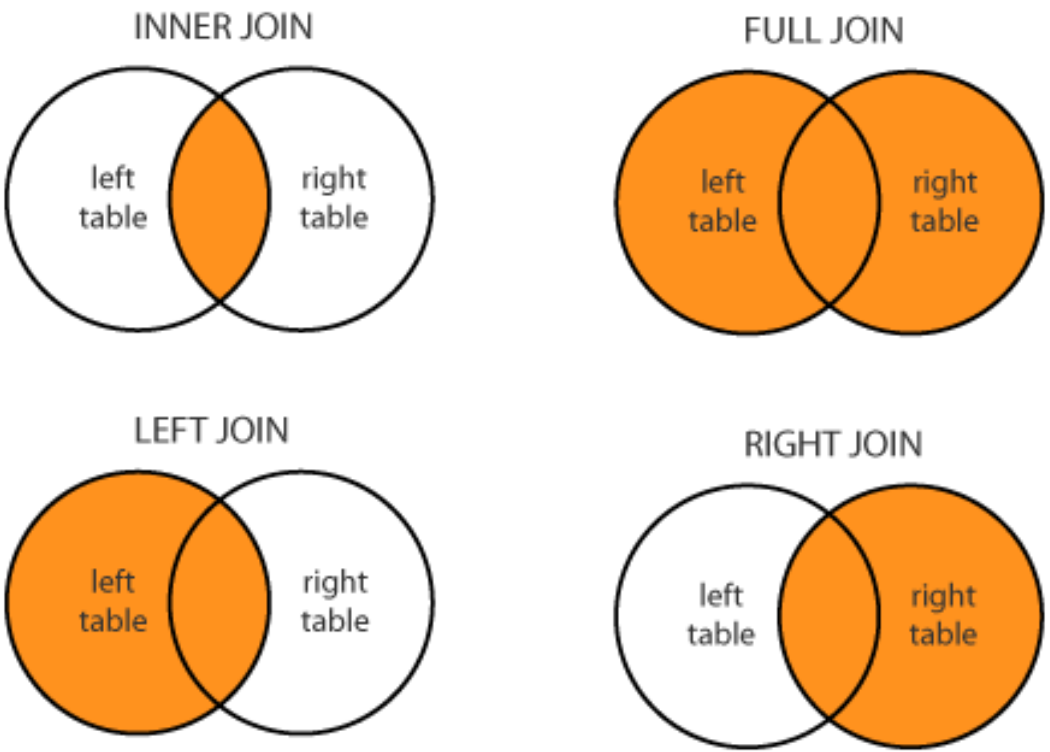
(https://towardsdatascience.com/architecting-a-machine-learning-pipeline-a847f094d1c7)

Fluxo de Trabalho Típico @ Knime



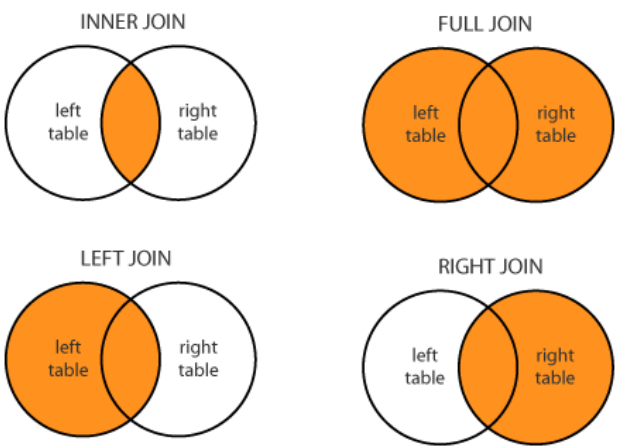
Junção
Join

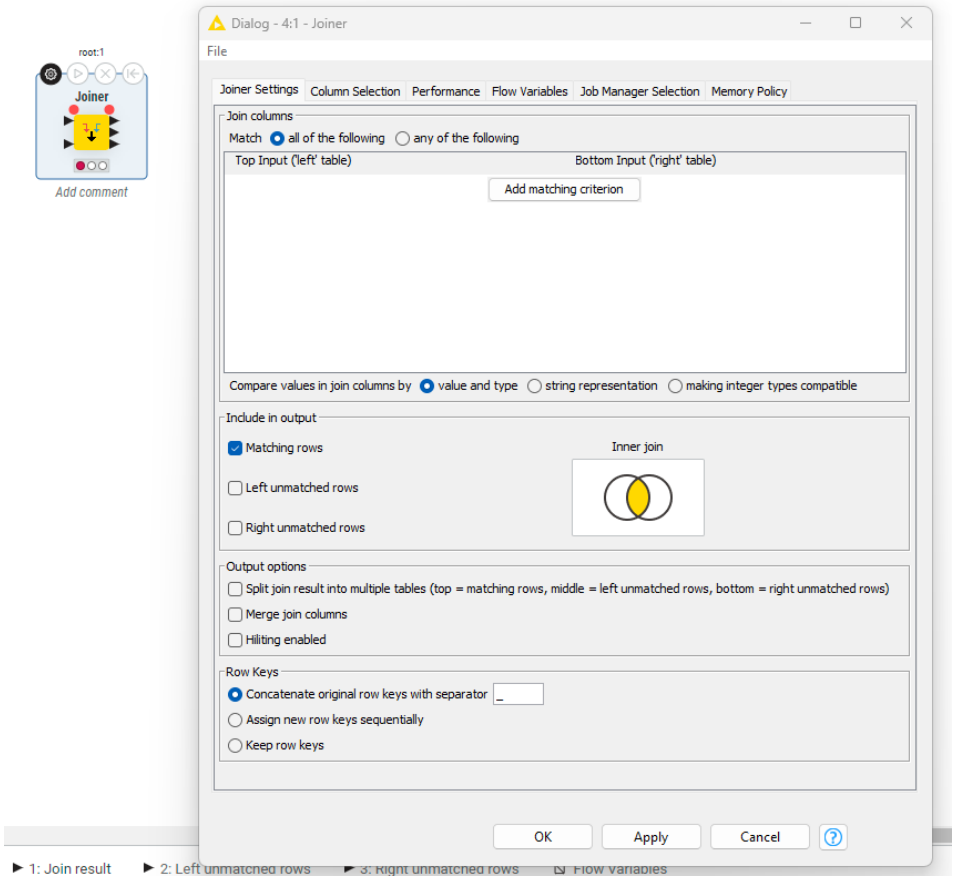
- Uma operação JOIN combina dados de diferentes fontes:



Junção
Join

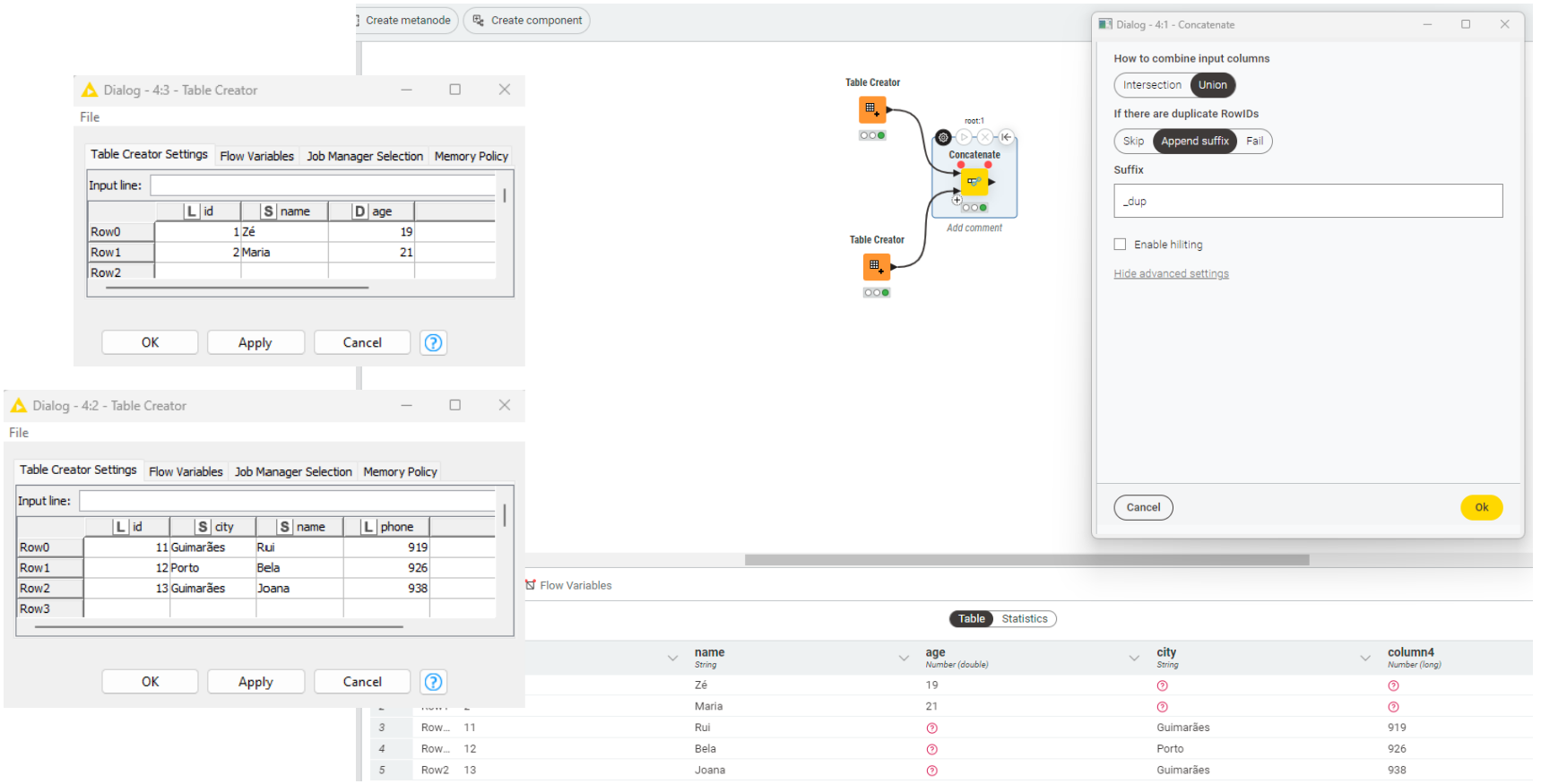
- No KNIME estão disponíveis diversas combinações no nodo JOINER:





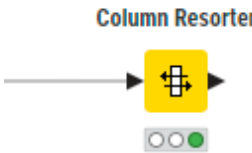
Concatenação
Concatenation

- União de colunas: CONCATENATE

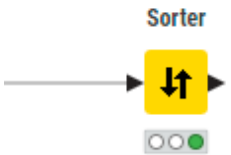


Ordenação
Sorter

- Altera a ordem das colunas de *input*, com base na definição dos parâmetros;

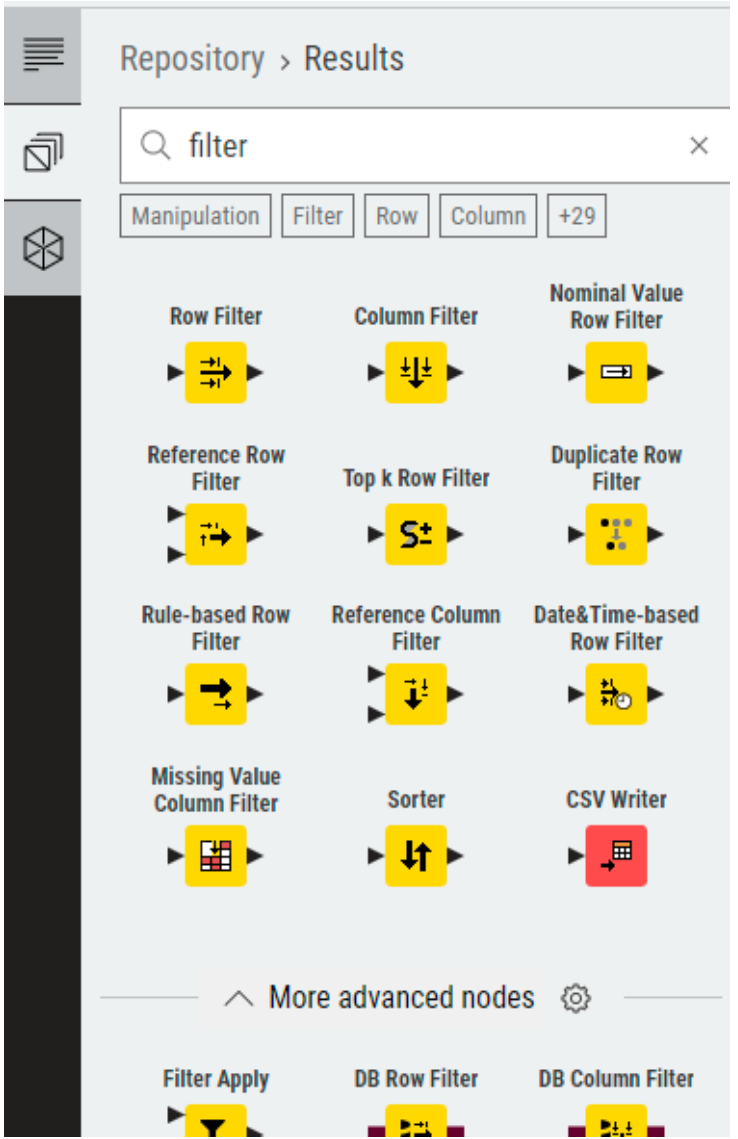


- Ordena as linhas com base na definição dos parâmetros;

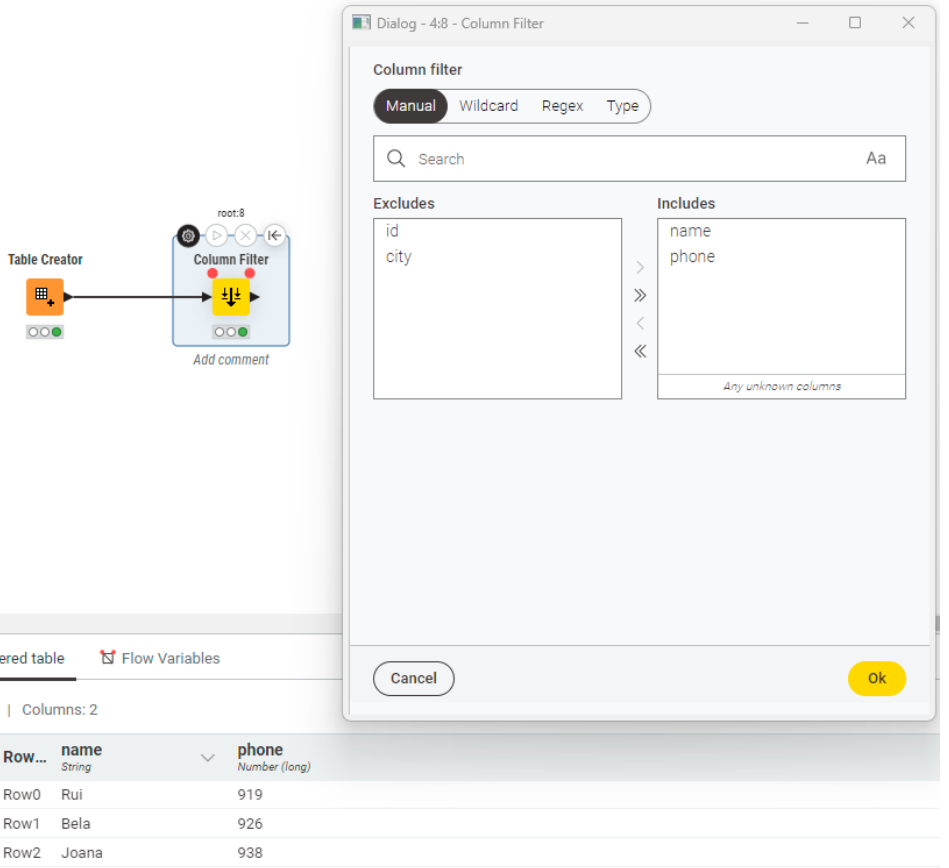


- Permite que as linhas sejam classificadas a partir da tabela da base de dados de entrada;

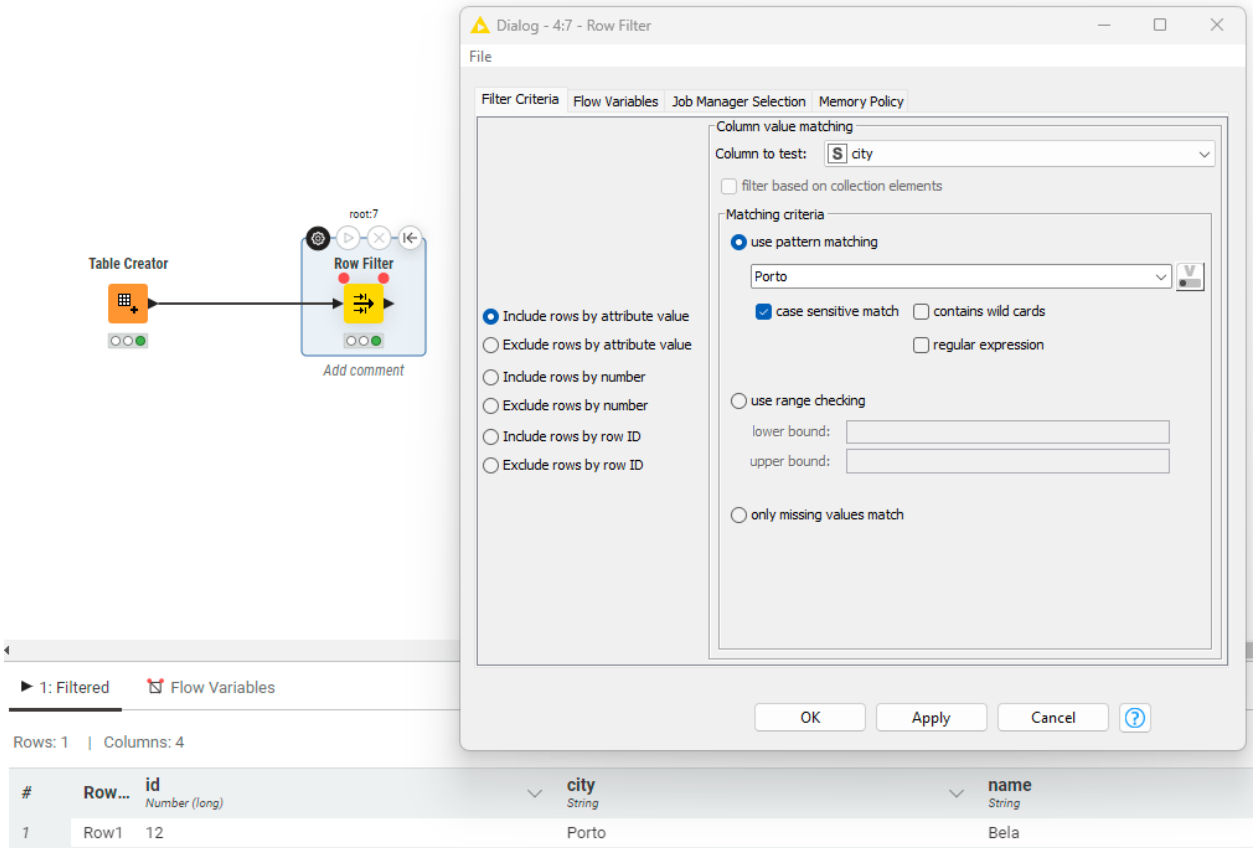
Filtros
Filter



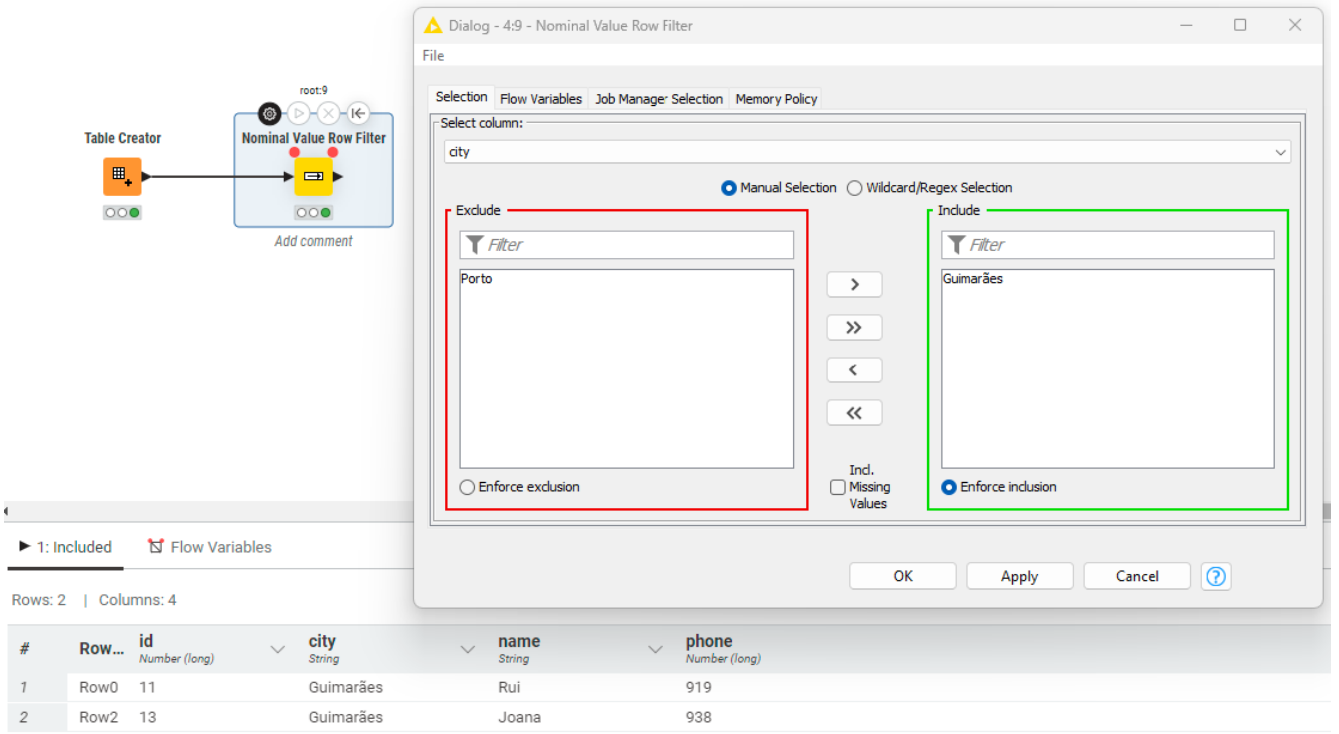
Filtro de Colunas
Column Filter



Filtro de Linhas
Row Filter

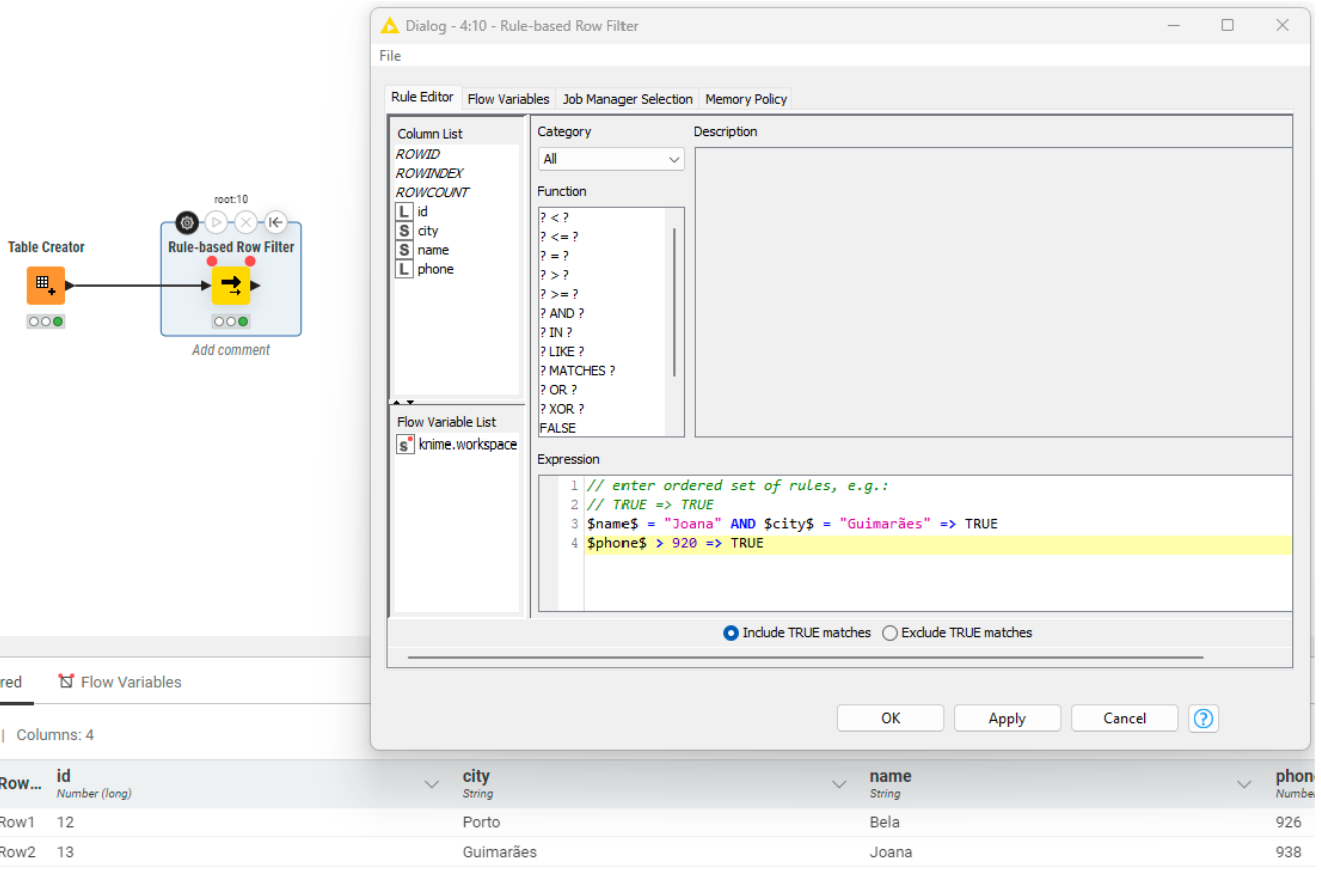


Filtro de Linhas de Valores Nominais
Nominal Value Row Filter

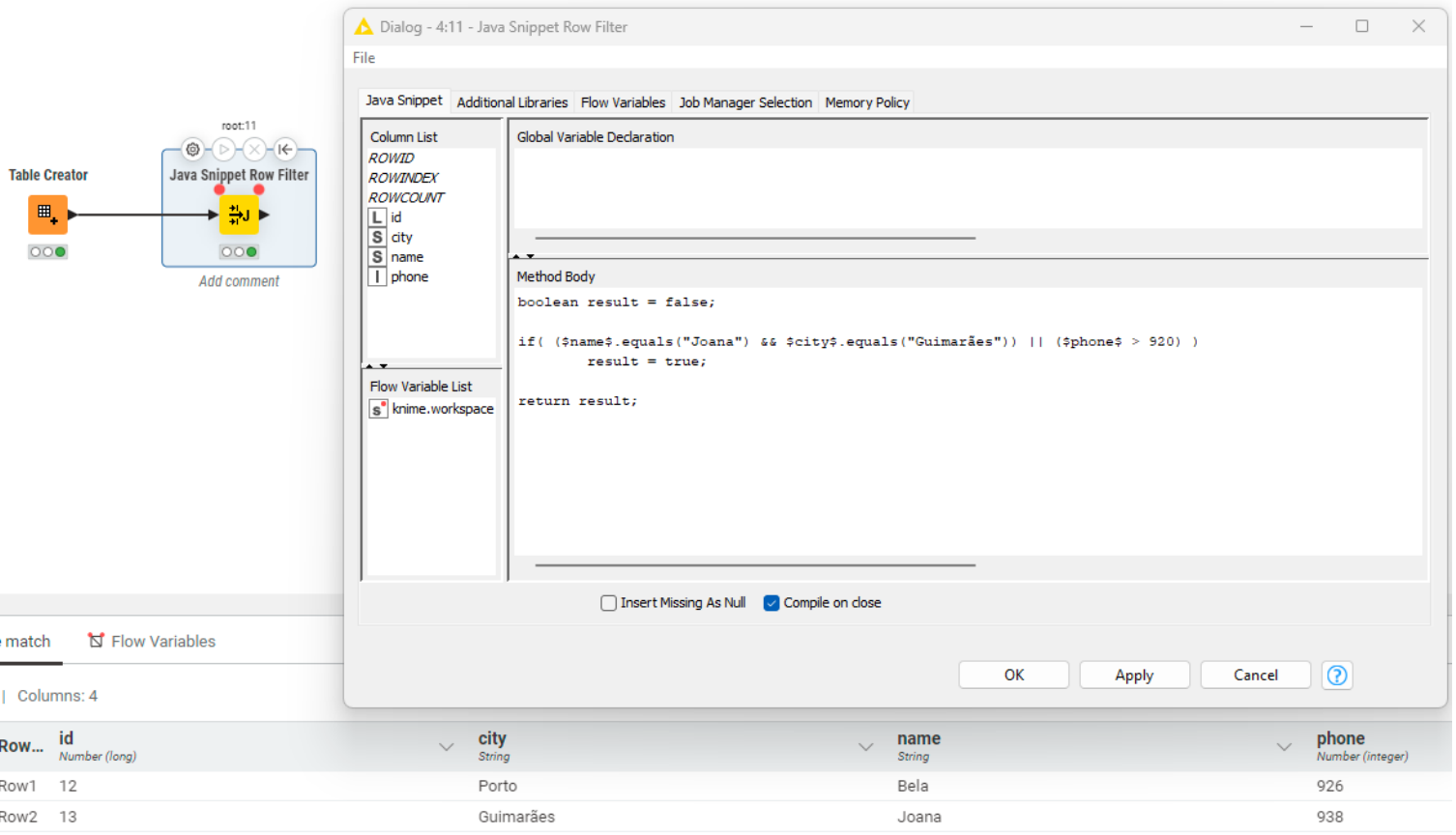


#	Row...	id *Number (long)*	city *String*	name *String*	phone *Number (long)*
1	Row0	11	Guimarães	Rui	919
2	Row2	13	Guimarães	Joana	938

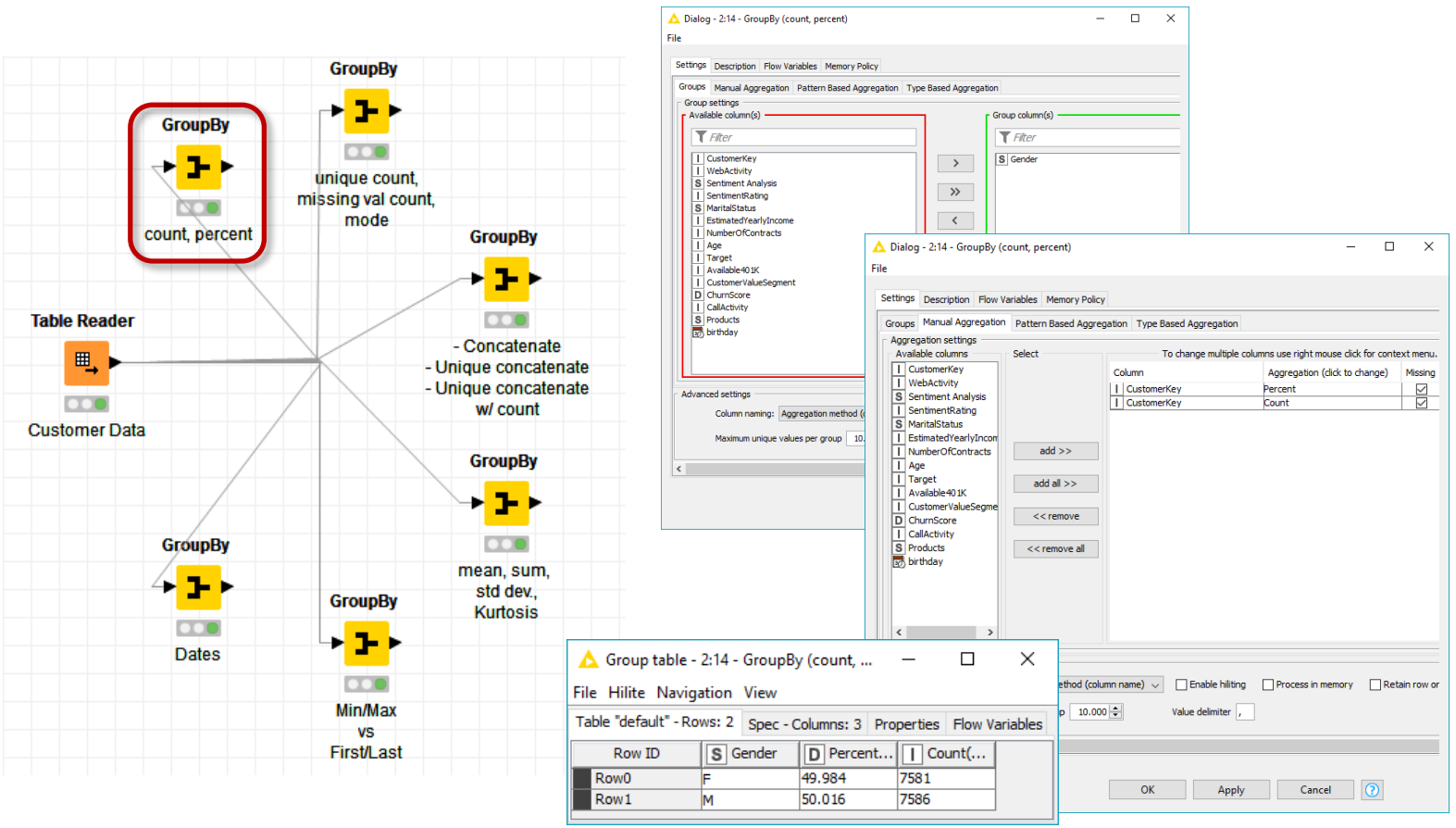
Filtro de Linhas Baseado em Regras
Rule-based Row Filter



Filtro de Linhas em JAVA *Snippet*
JAVA *Snippet Row Filter*

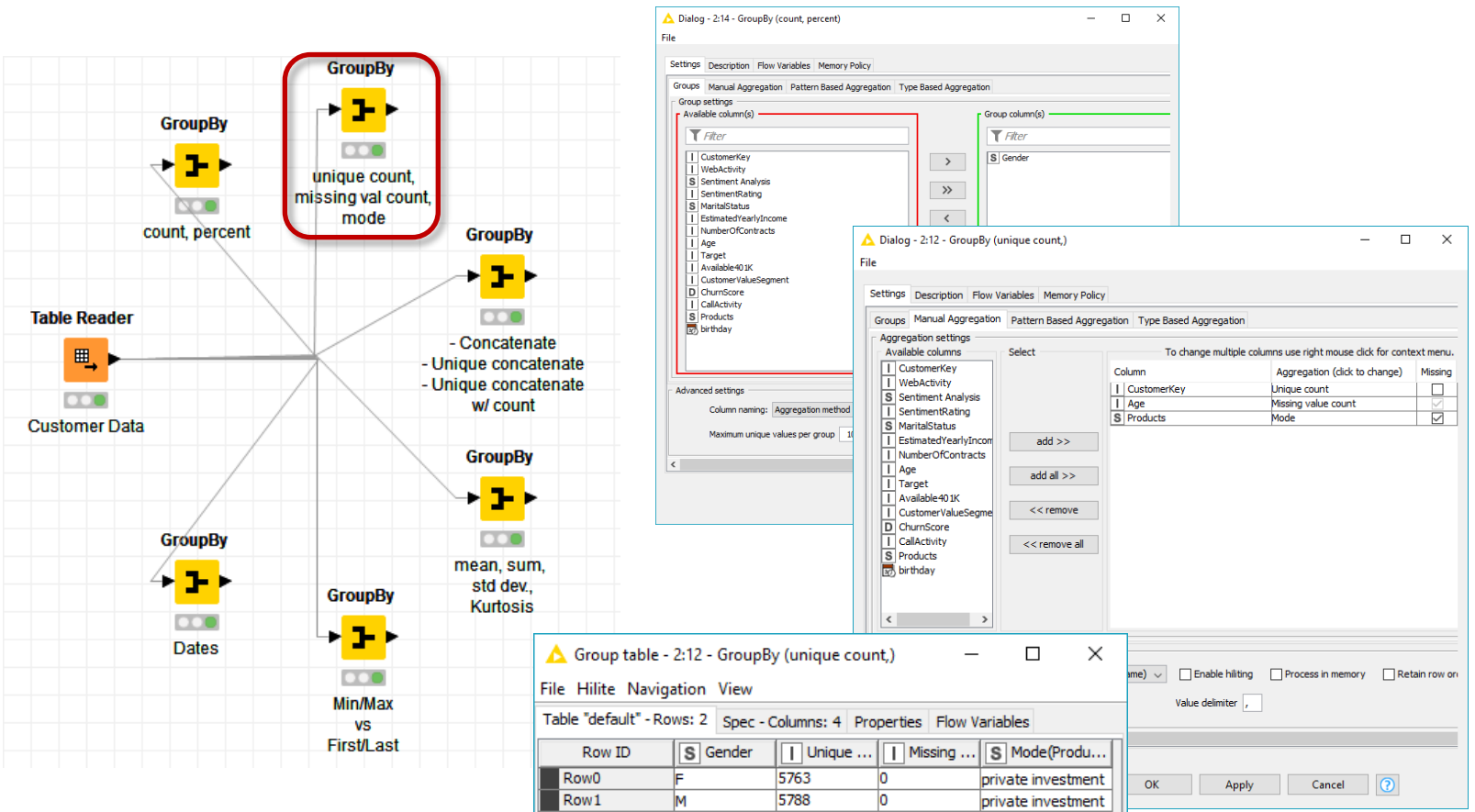


Operações de Agregação
Count and Percent

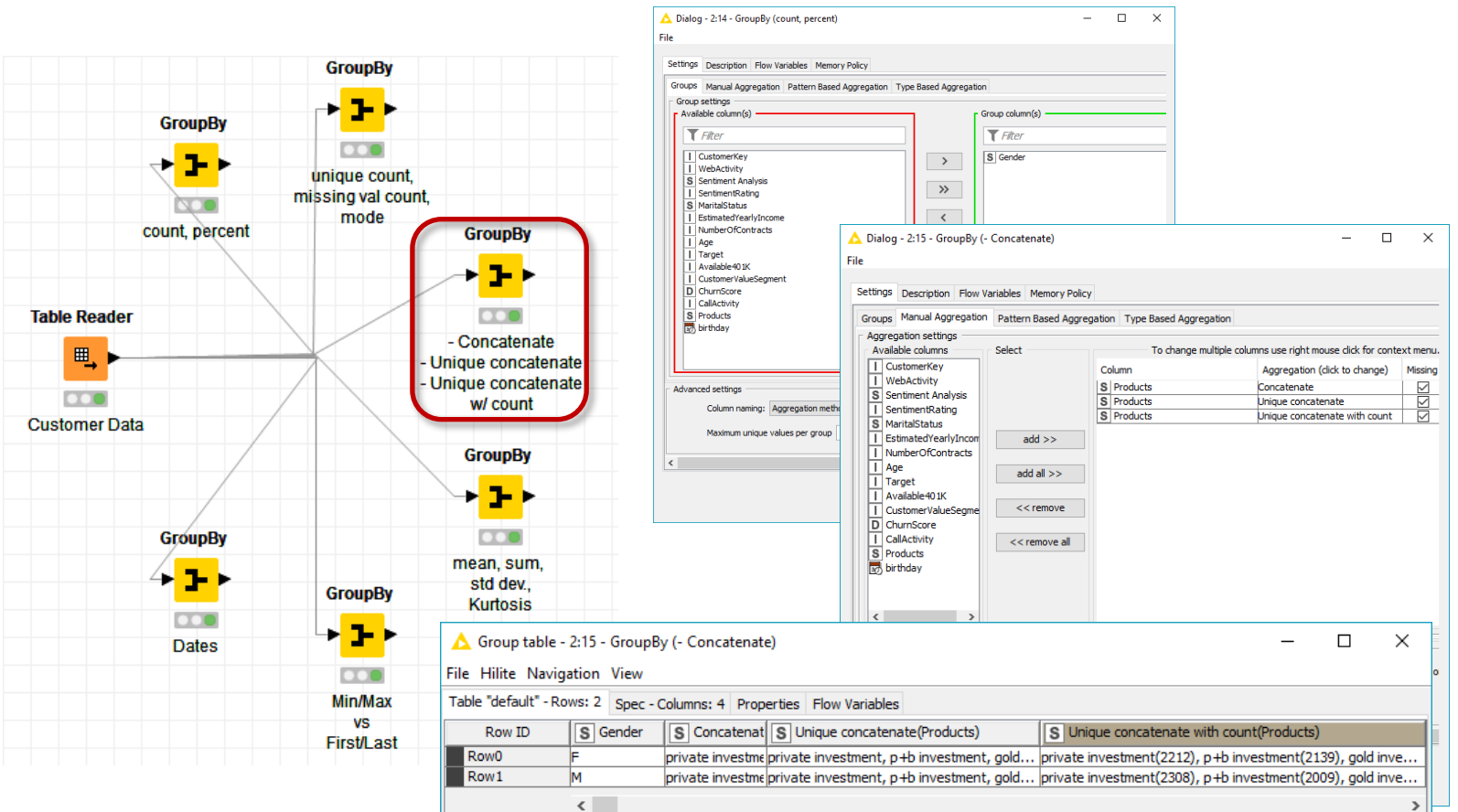


Operações de Agregação

Unique Count, Missing Values Count and Mode

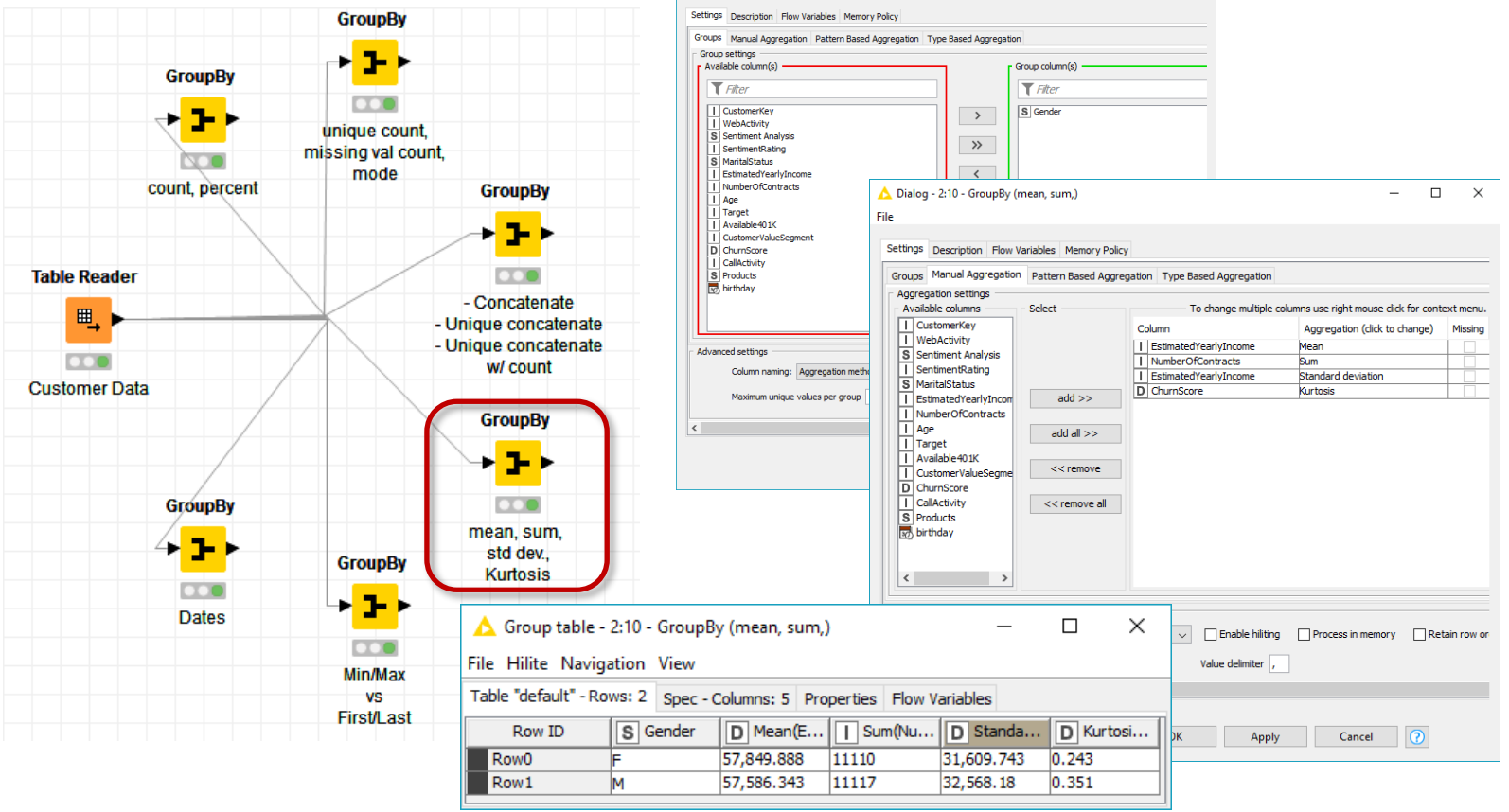


Operações de Agregação
Concatenate

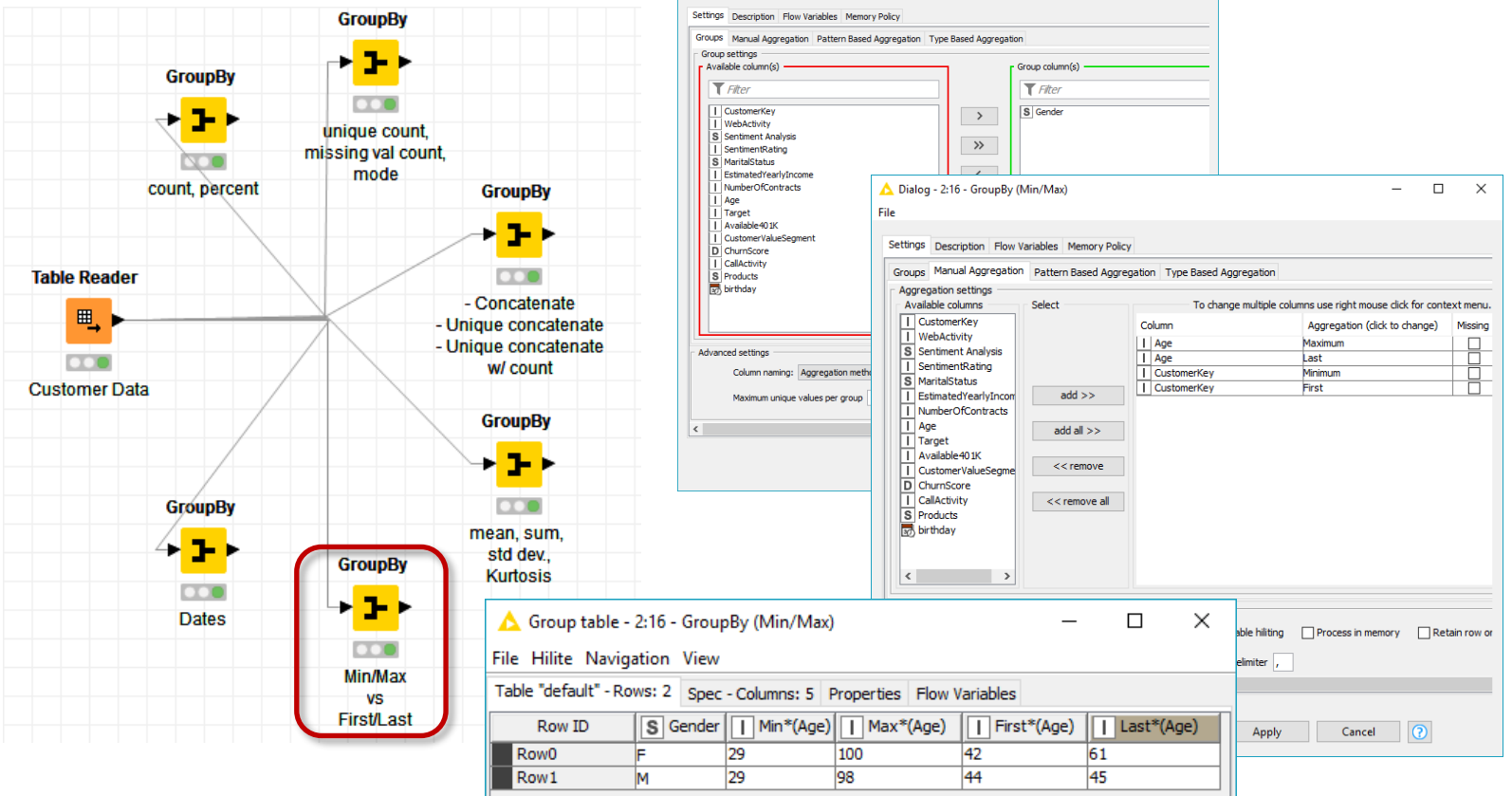


Operações de Agregação

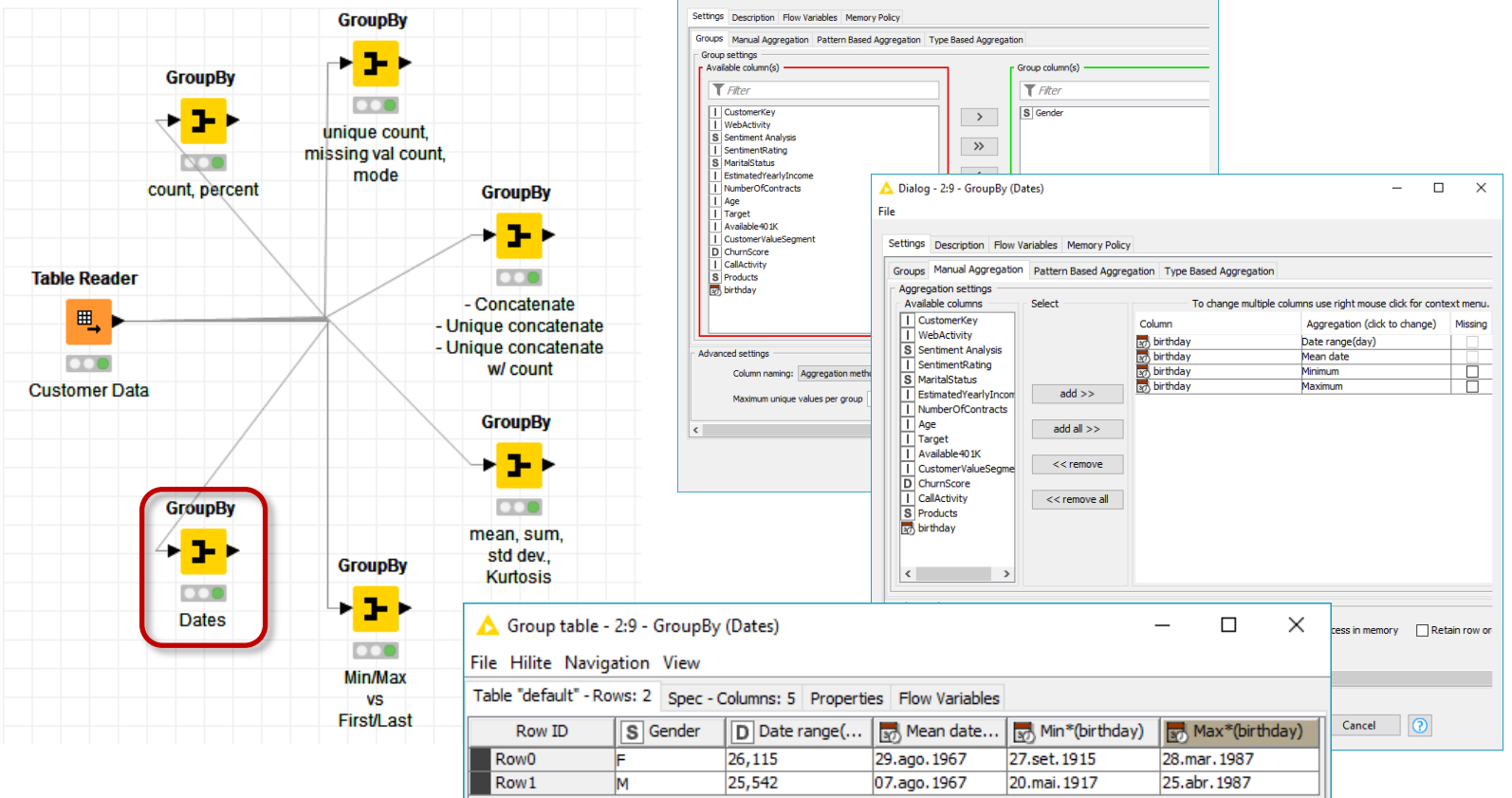
Mean, Sum, Standard Deviation and Kurtosis



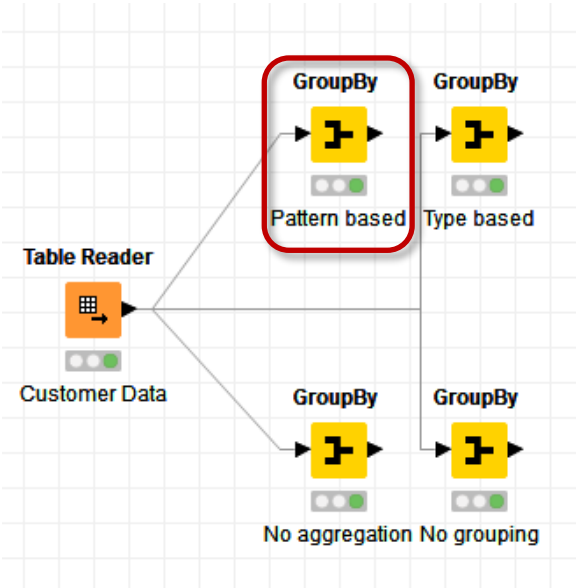
Operações de Agregação
Min/Max vs First/Last

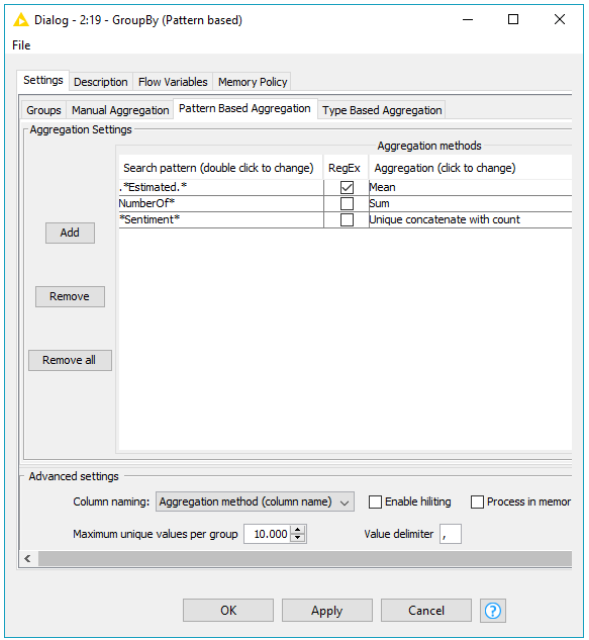


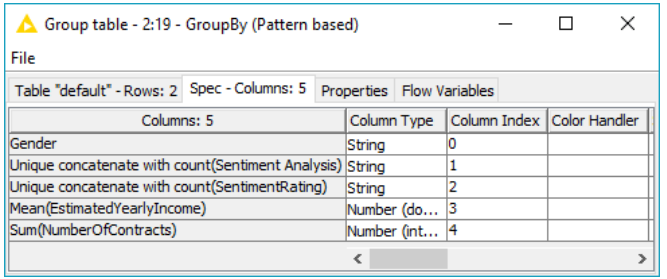
Operações de Agregação
Dates



Operações Avançadas de Agregação
Pattern Based

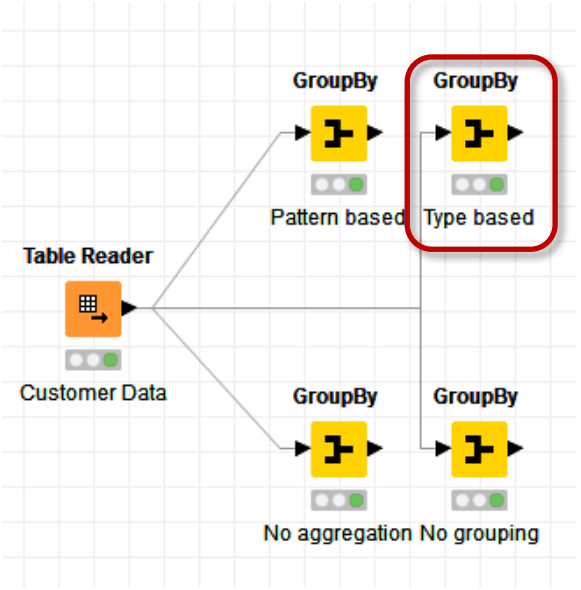


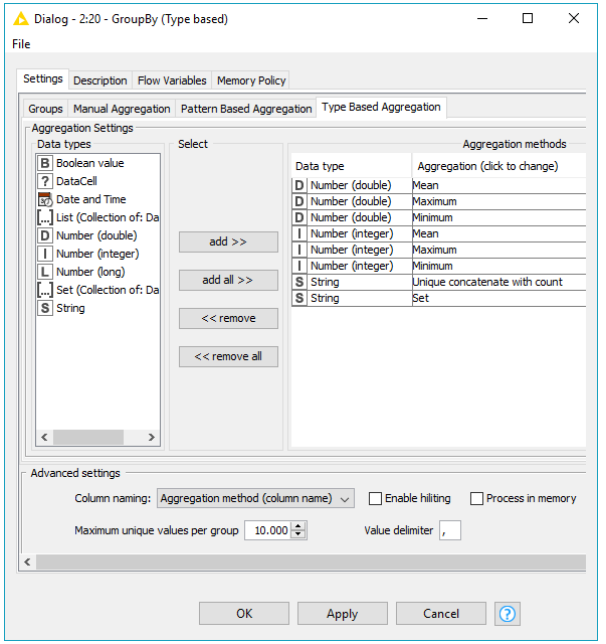


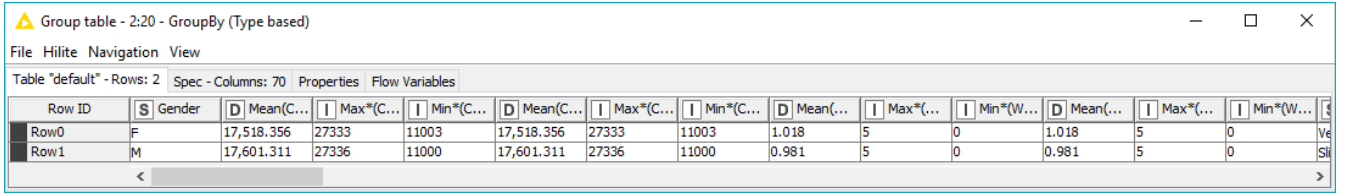


Columns: 5	Column Type	Column Index	Color Handler
Gender	String	0	
Unique concatenate with count(Sentiment Analysis)	String	1	
Unique concatenate with count(SentimentRating)	String	2	
Mean(EstimatedYearlyIncome)	Number (do…	3	
Sum(NumberOfContracts)	Number (int…	4	

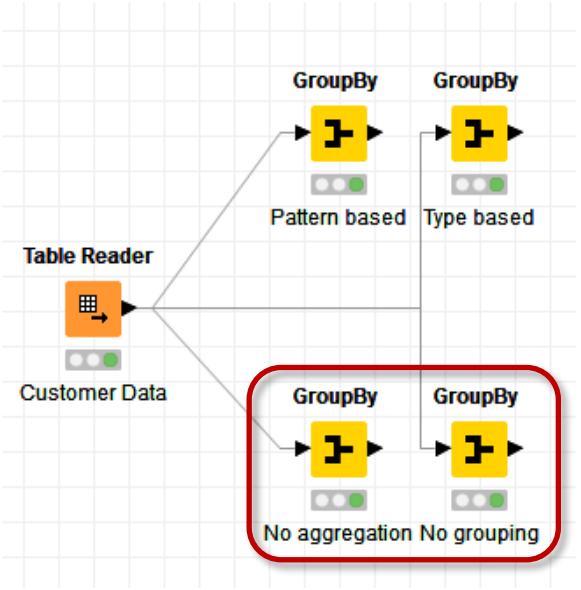
Operações Avançadas de Agregação
Data Type Based







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

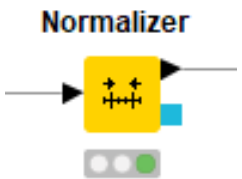
- Redimensionamento de atributos/ *Feature scaling*
- Deteção de valores atípicos/ *Outlier detection*
- Seleção de atributos/ *Feature selection*
- Tratamento de valores em falta/ *Missing values treatment*
- Enumeração de valores nominais/ *Nominal value discretization*
- Divisão em intervalos/ *Binning*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

Redimensionamento de atributos
Feature scaling

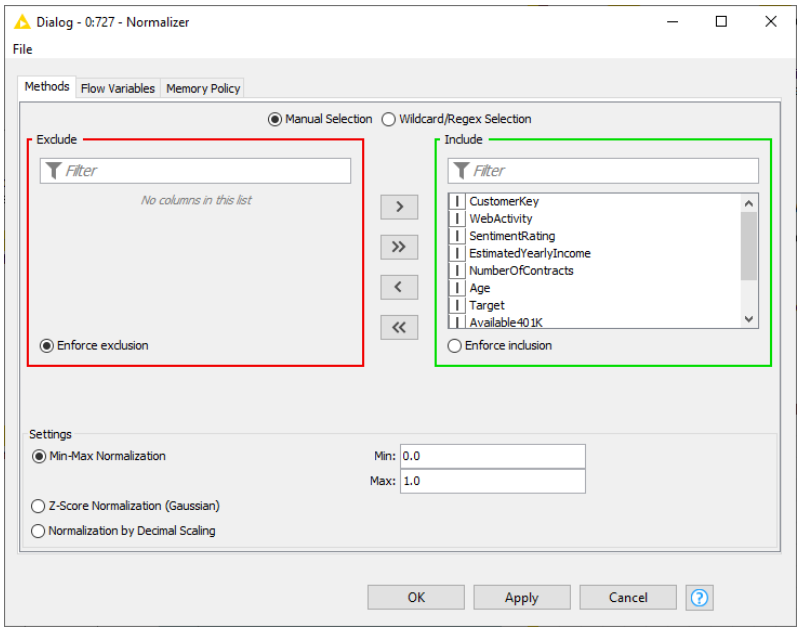
- Normalizar a gama de valores de atributos;
- Muitos classificadores usam métricas de distância (ex.: distância euclidiana) e, se um atributo tiver uma gama alargada de valores, a distância será definida por esse atributo em particular. Por isso, a gama de valores deve ser normalizada para que cada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

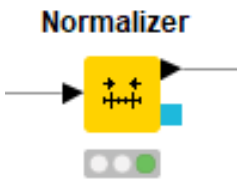


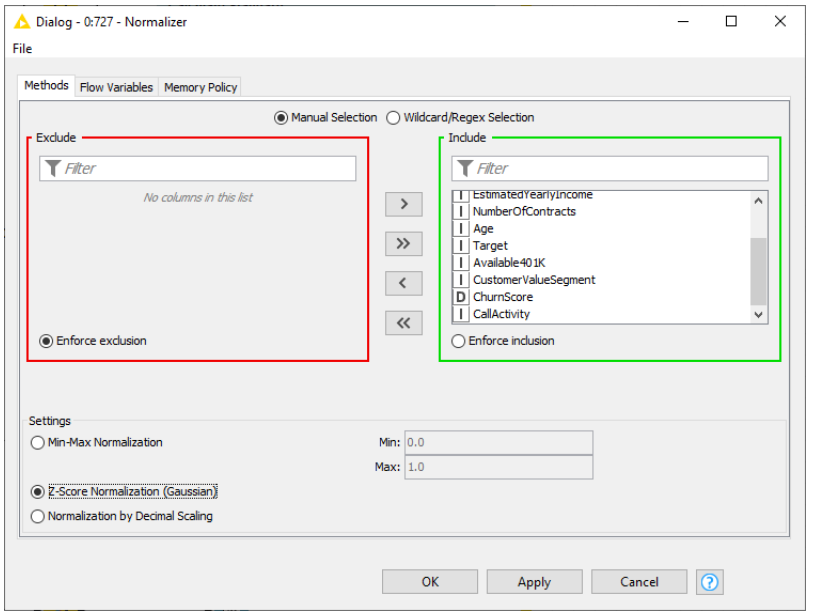
$$z = (b - a)\frac{x - \min(x)}{\max(x) - \min(x)} + a$$



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.

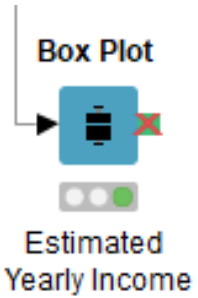


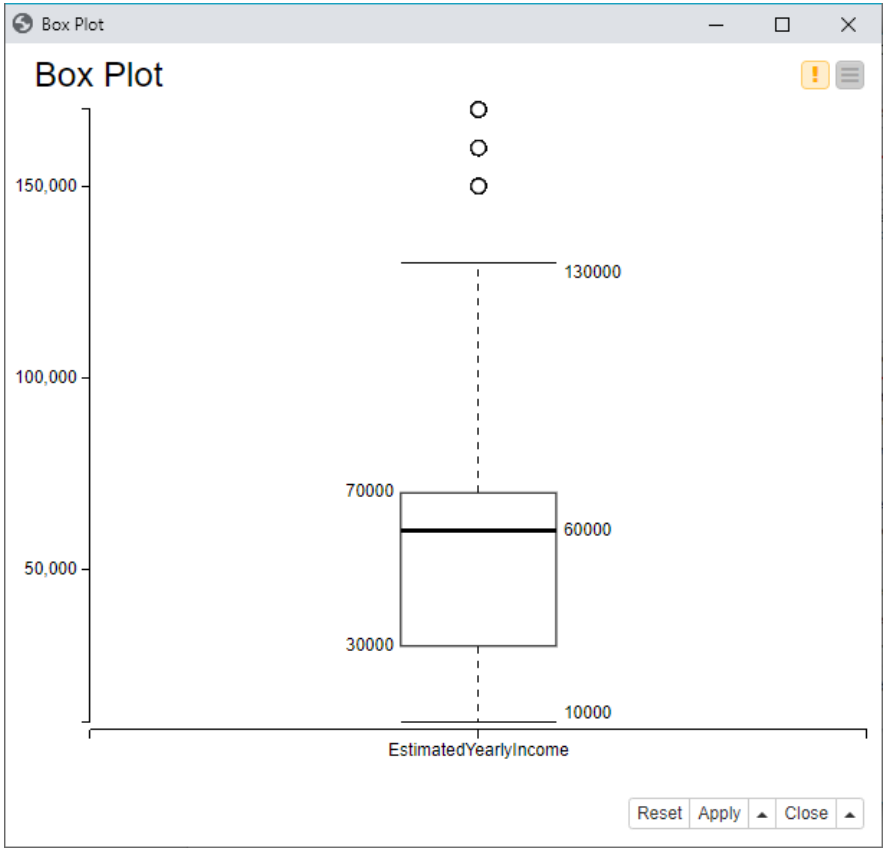


$$z = \frac{x_i - \mu}{\sigma}$$

Deteção de valores atípicos
Outlier Detection

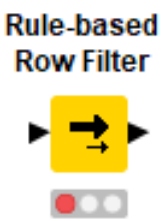
- Deteção de valores atípicos (*outliers*):
  - Estratégias baseadas em estatística:
    - Box Plots
    - Z-Score (std. dev)

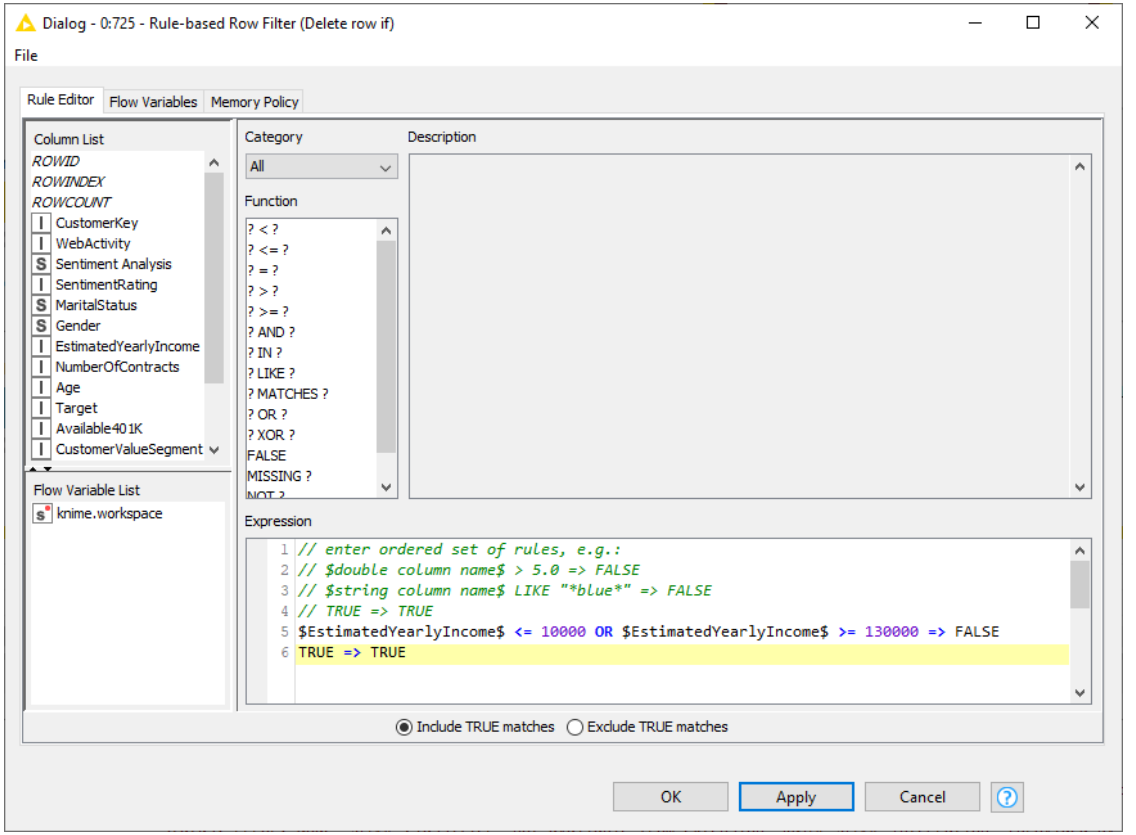




Deteção de valores atípicos
Outlier Detection

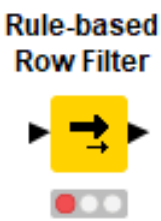
- Deteção de valores atípicos (*outliers*):
  - Estratégias baseadas em conhecimento

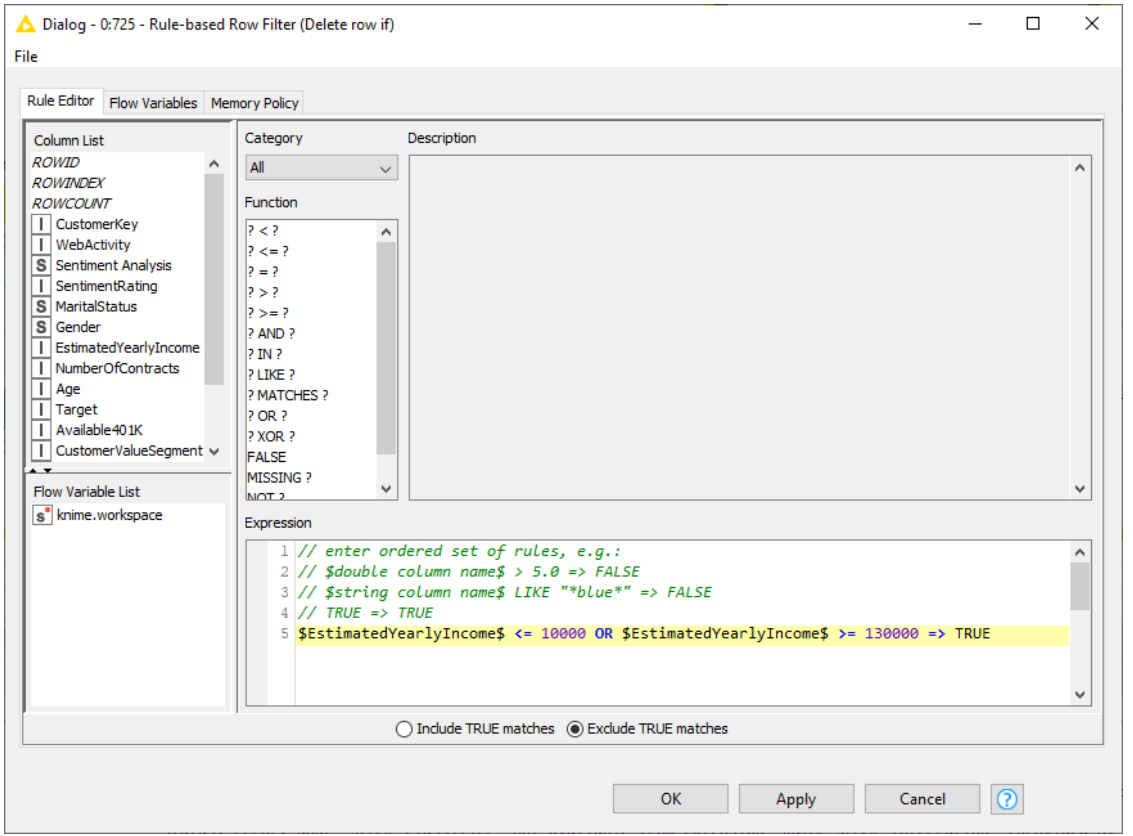




Deteção de valores atípicos
Outlier Detection

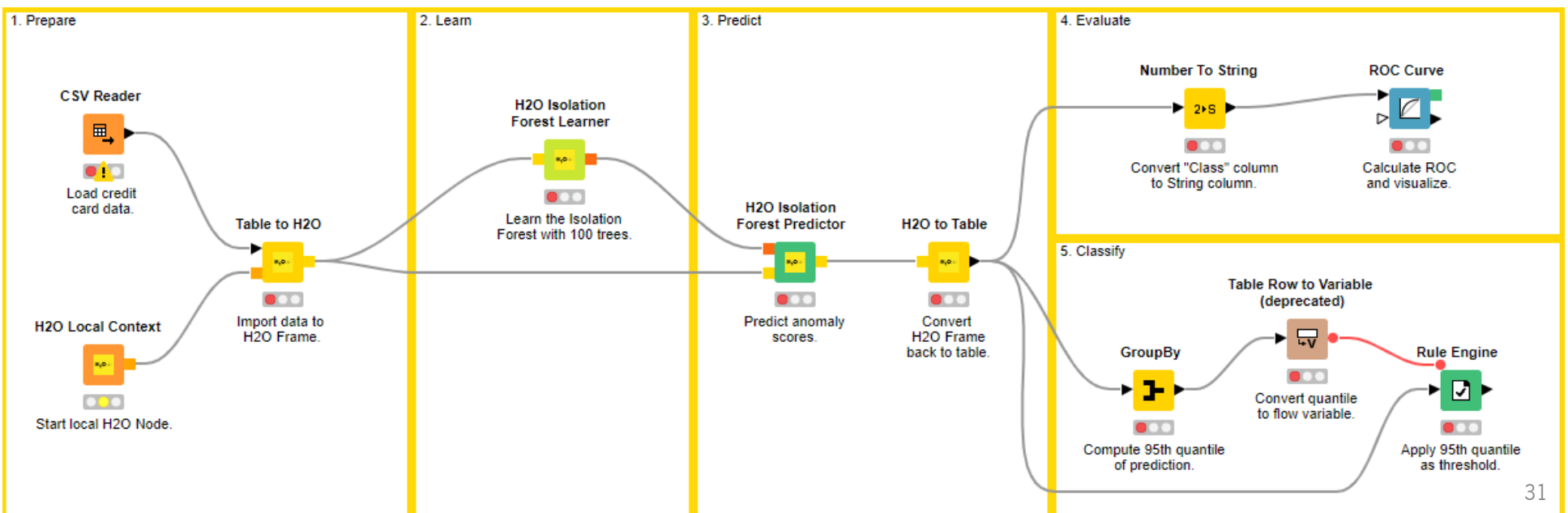
- Deteção de valores atípicos (*outliers*):
  - Estratégias baseadas em conhecimento





Deteção de valores atípicos
Outlier Detection

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

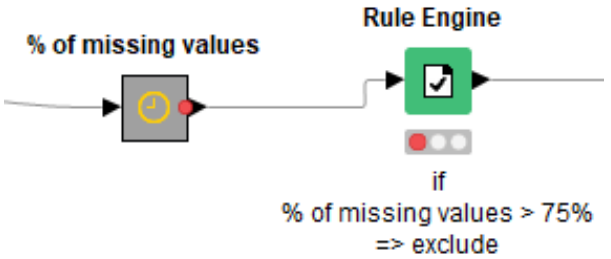


Seleção de Atributos
Feature Selection

- 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

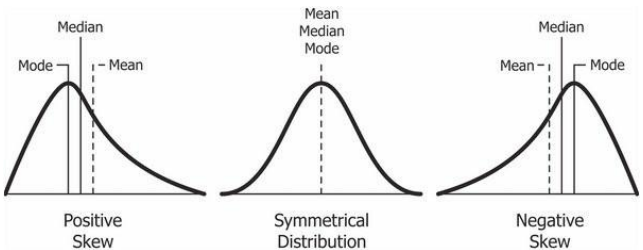
- Métodos de Filtro:
  - Remover uma *feature* se a percentagem de *missing values* for superior a um determinado valor estabelecido;

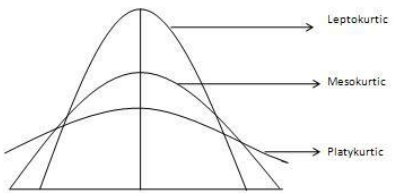


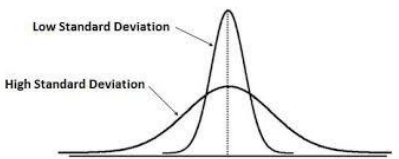
  - Usar o teste "chi-square" para medir o grau de dependência entre uma *feature* e o *target*:
    - Para cada *feature* calcular X^2;
    - Normalizar X^2 e ordenar de forma decrescente;
    - Selecionar 'n' *features* com maior importância;
      (ou as que estão acima de um determinado limite)

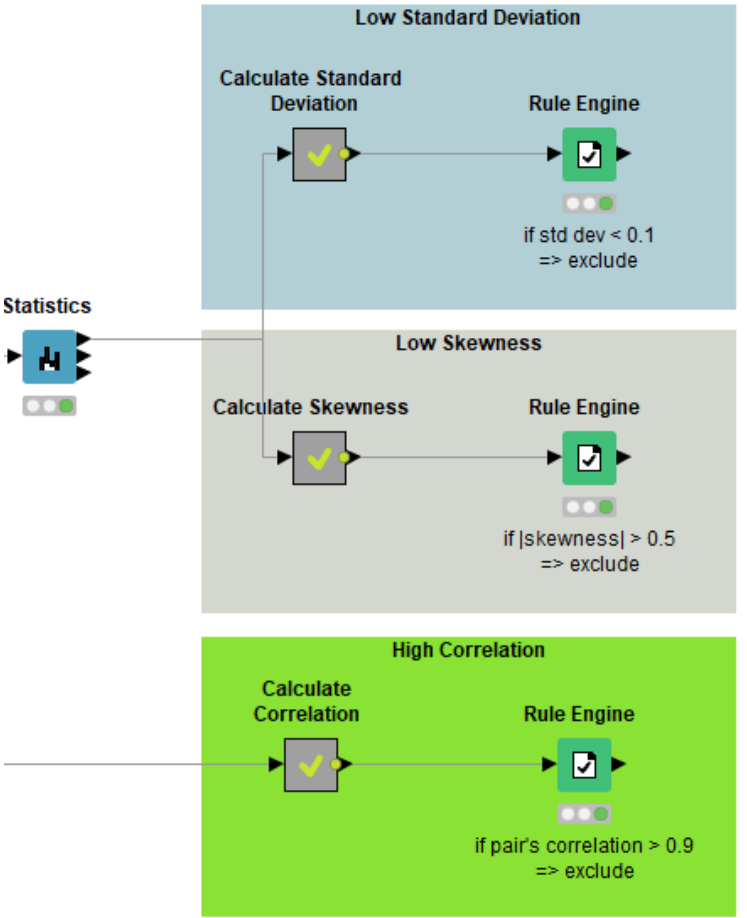
Seleção de Atributos
Feature Selection

- 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;



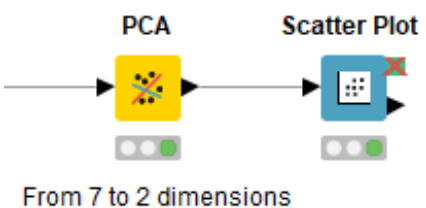


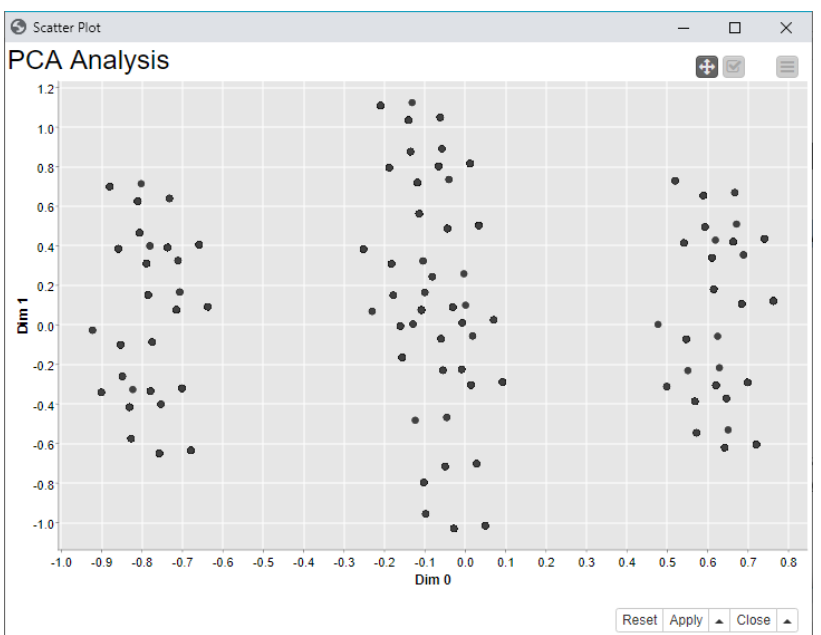




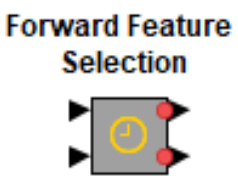
Seleção de Atributos
Feature Selection

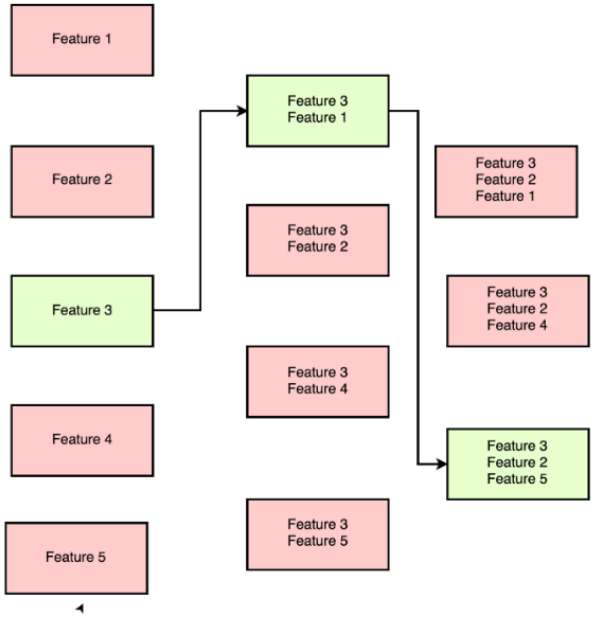
- Métodos de Filtro:
  - ...
  - Principal Component Analysis (PCA):
    - Técnica usada para reduzir a dimensão do espaço de *features*;
    - O objetivo é reduzir o número de *features* sem perder (demasiado) conhecimento;
    - Uma aplicação comum de PCA é para visualização de dados de grande dimensão;



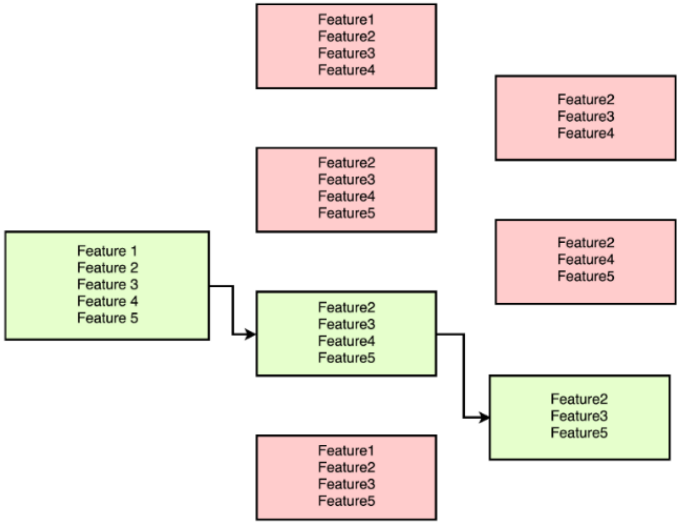


- Métodos *wrapper*:
  - Utilizar técnicas de *machine learning* para selecionar as *features* mais importantes;
    - Selecionar um conjunto de *features* como um problema de pesquisa;
    - Preparar diferentes combinações;
    - Avaliar e comparar as diferentes combinações;
    - Medir a "utilidade"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



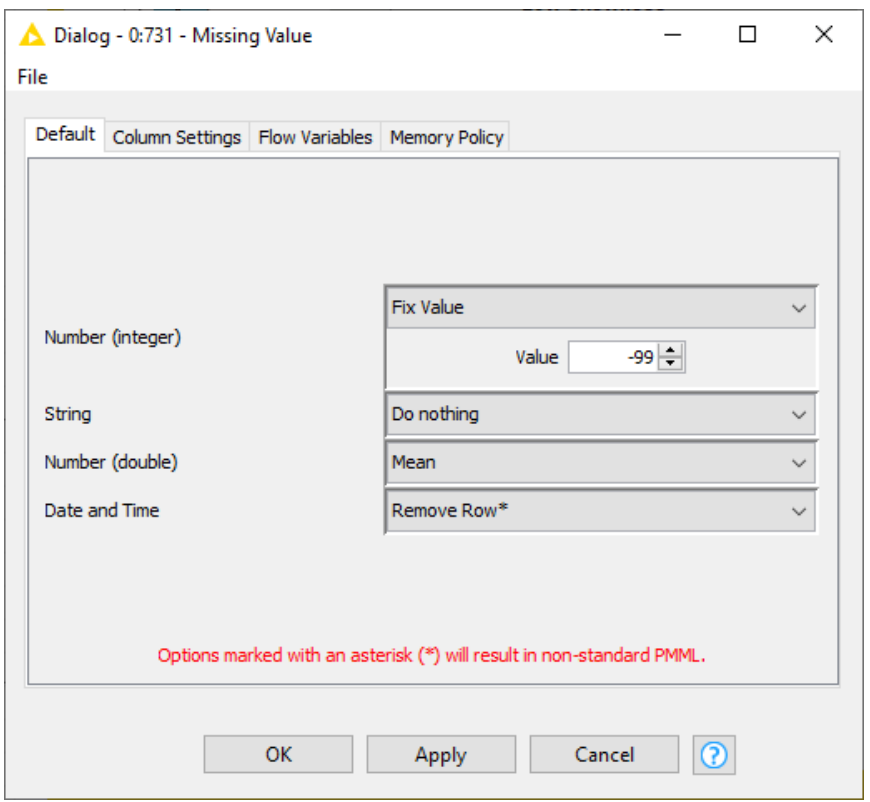


- Métodos *wrapper*:
  - Utilizar técnicas de *machine learning* para selecionar as *features* mais importantes;
    - Selecionar um conjunto de *features* como um problema de pesquisa;
    - Preparar diferentes combinações;
    - Avaliar e comparar as diferentes combinações;
    - Medir a "utilidade"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



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;
    - ...



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.

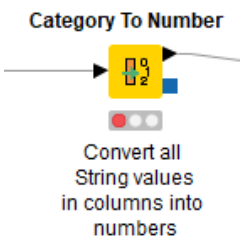
One-Hot Encoded

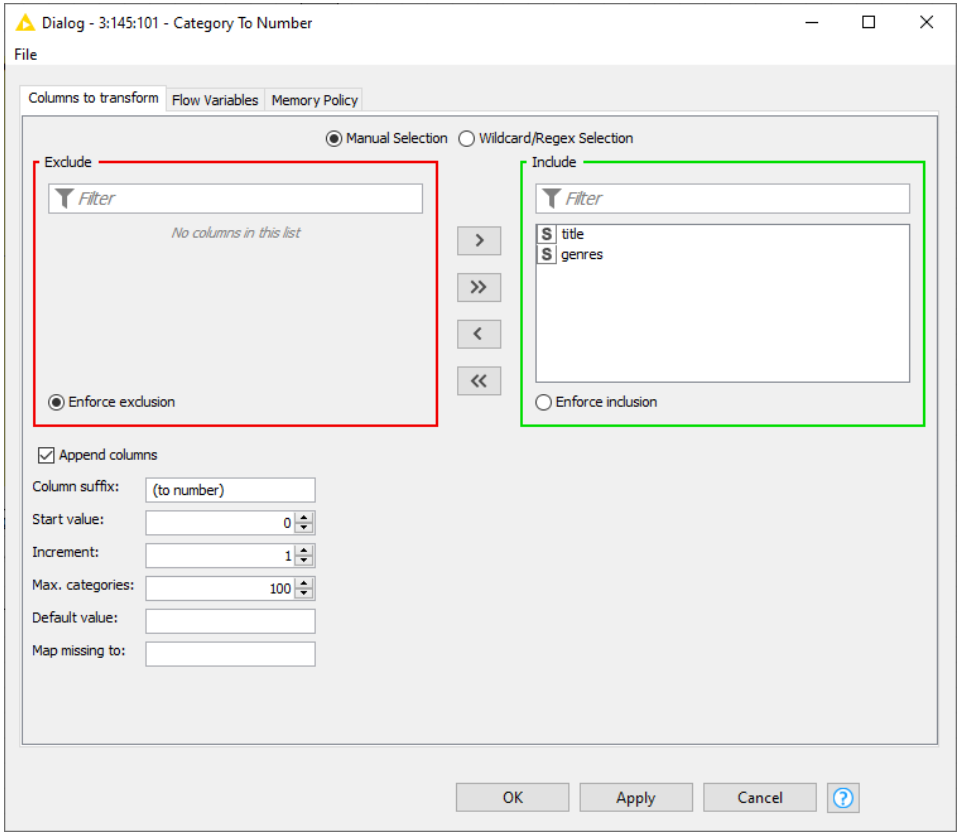
Movie	Adventure	Comedy	Drama
Jumanji	1	0	0
American Pie	0	1	0
Braveheart	0	0	1
…	…		

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

Enumeração de Valores Nominais
Nominal Value Discretization/Encoding

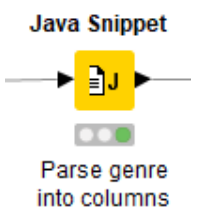
- Enumeração de valores nominais:

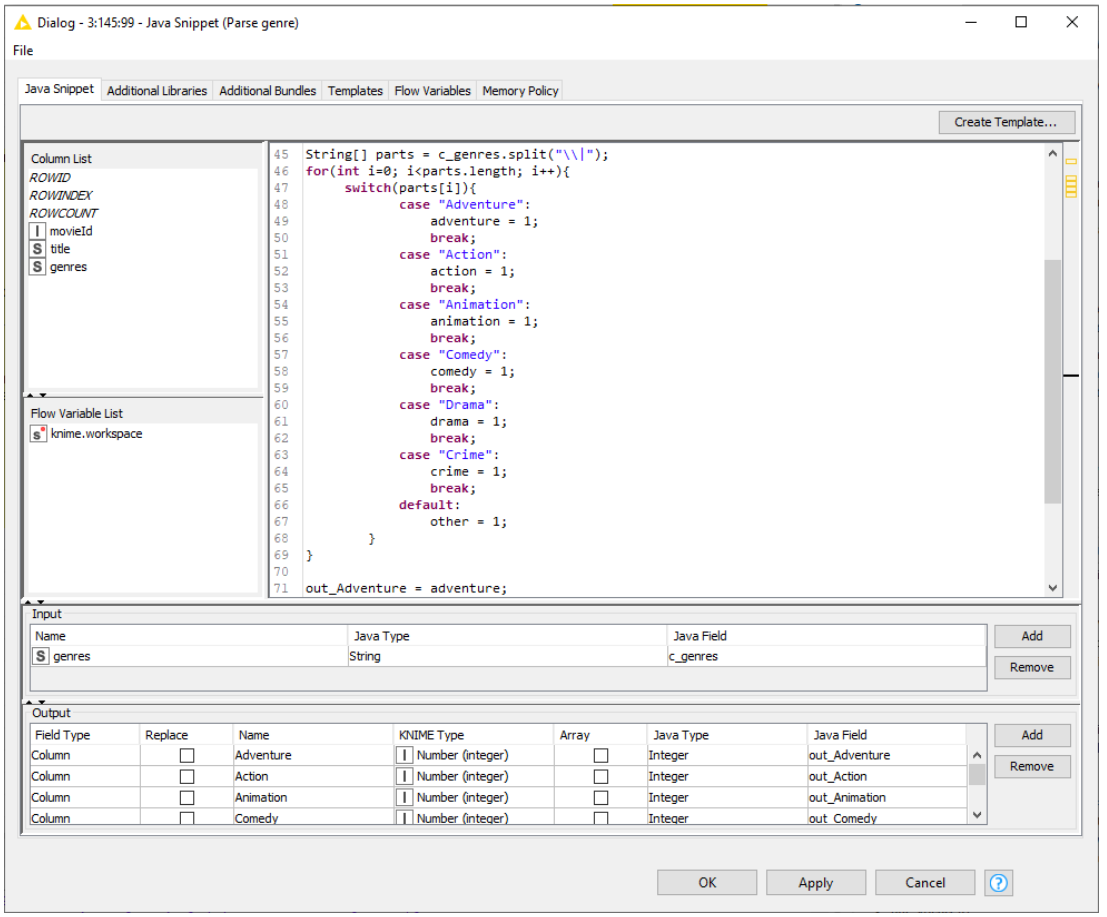




Enumeração de Valores Nominais
Nominal Value Discretization/Encoding

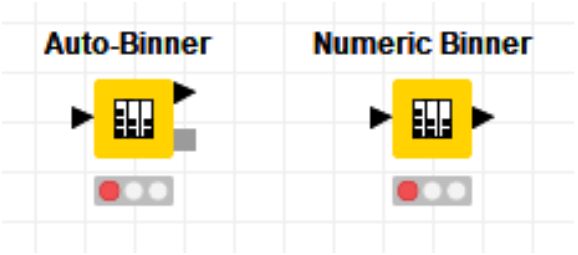
- Enumeração de valores nominais:

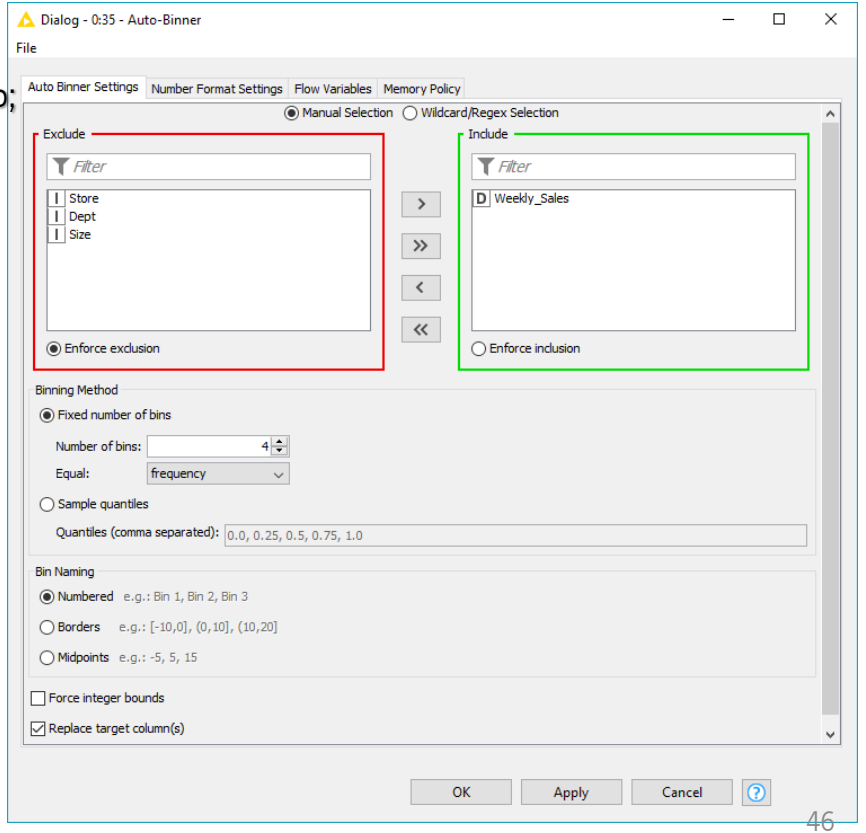




Divisão em Intervalos
Binning

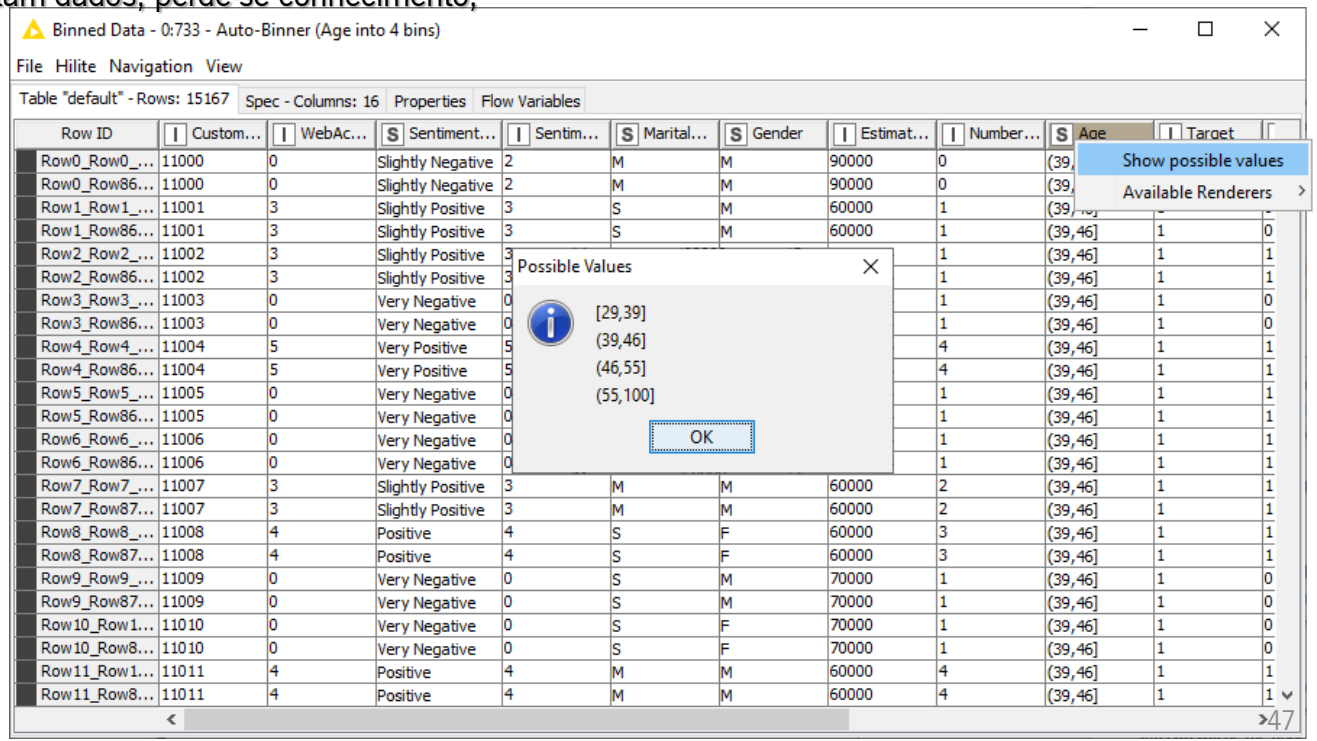
- Divisão em intervalos:
  - Transformação de valores contínuos em discretos (*bins*):
    - Torna o modelo mais robusto e evita o *overfitting*;
    - Penaliza o desempenho do modelo, uma vez que, sempre que se descartam dados, perde-se conhecimento;





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;



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

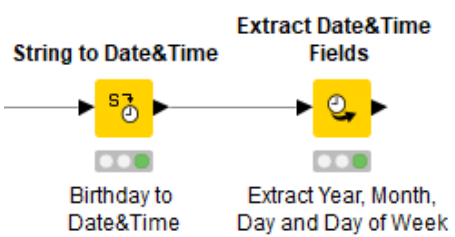
- Engenharia de atributos:
  - Processo de criação de novos atributos (*features*);
  - Aumentar o conhecimento/aumentar o desempenho dos modelos;

- 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:
  - 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

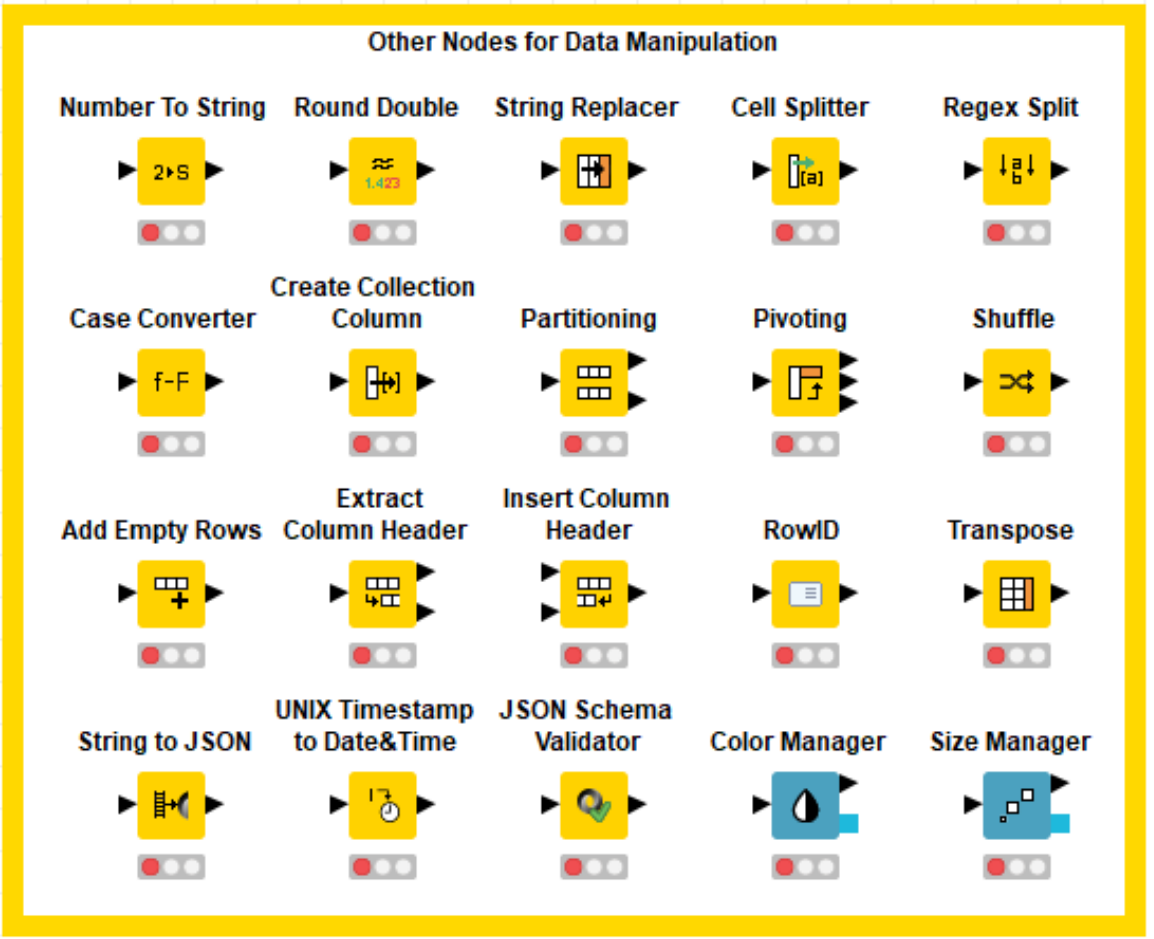
Engenharia de Atributos
Feature Engineering

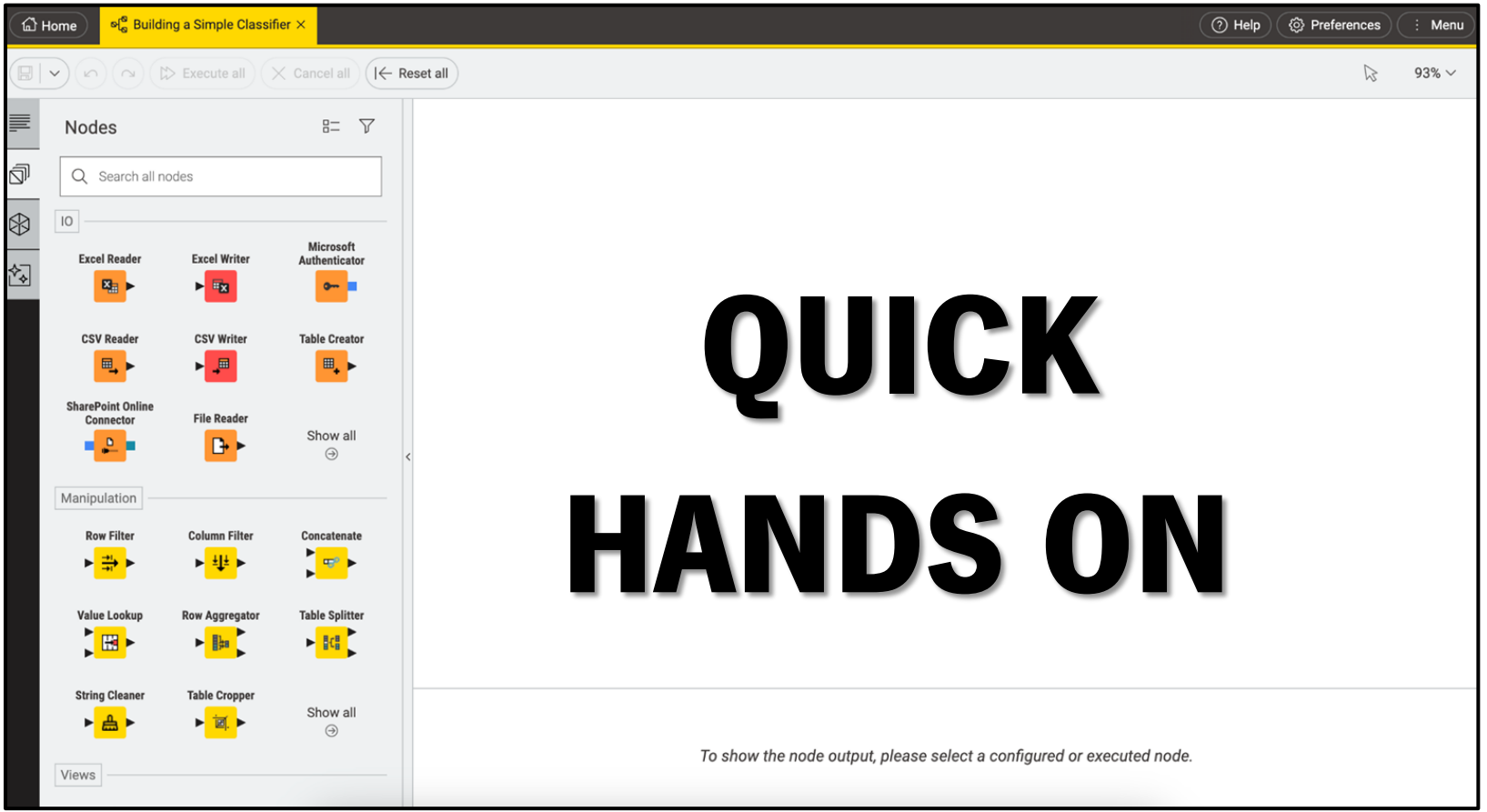
- Engenharia de atributos:
  - Processo de criação de novos atributos (*features*);
  - Aumentar o conhecimento/aumentar o desempenho dos modelos;

- 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?
    - ...

Outros Nodos




Home
Building a Simple Classifier
Help
Preferences
Menu
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.