



Universidade do Minho  
Departamento de Informática

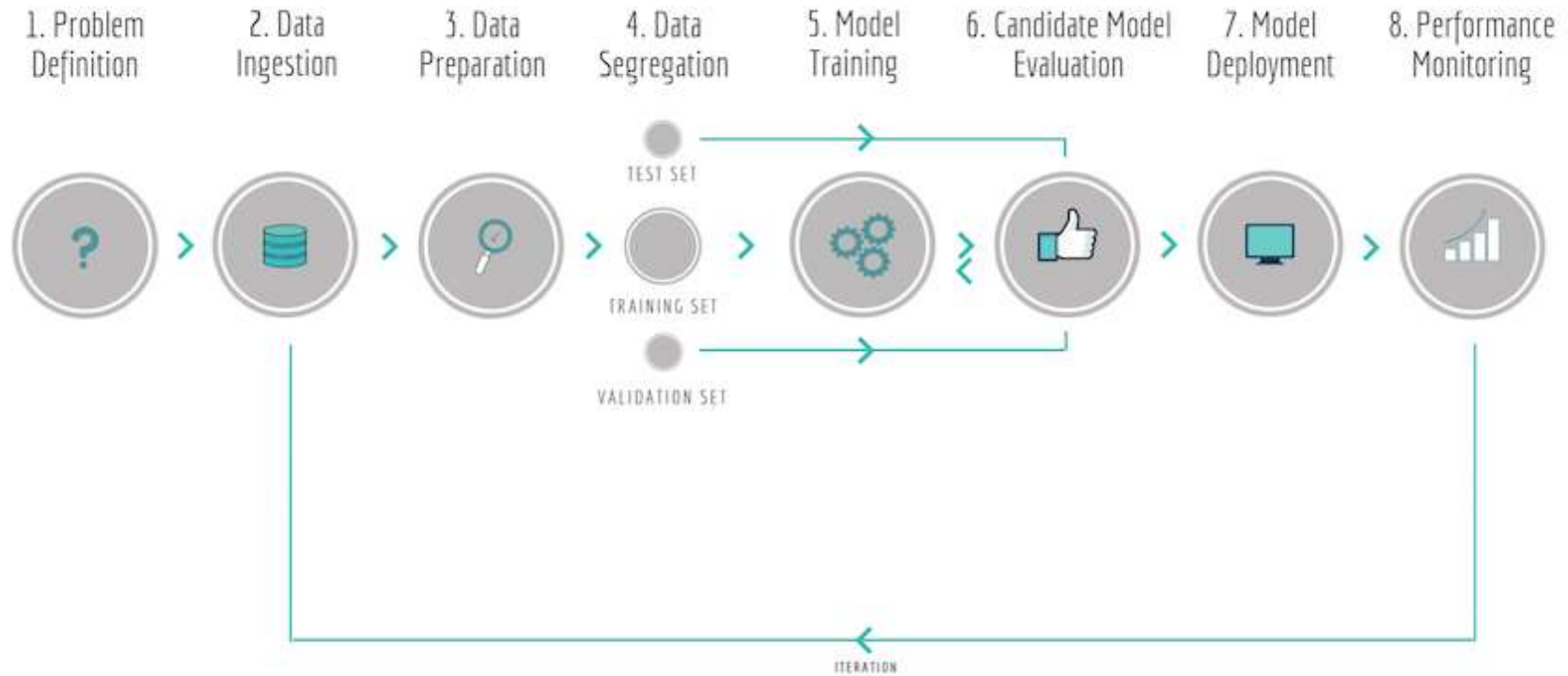
# **Exploração de dados com KNIME**

**LEI/MiEI @ 2022/2023, 2º sem**  
**[ADI<sup>3</sup>]**

- KNIME
  - Bons hábitos
  - Metanodos
  - Ingestão de dados
  - Partição de dados
- Qualidade dos dados
- Exploração de dados
- Experimentação  
(*hands on*)

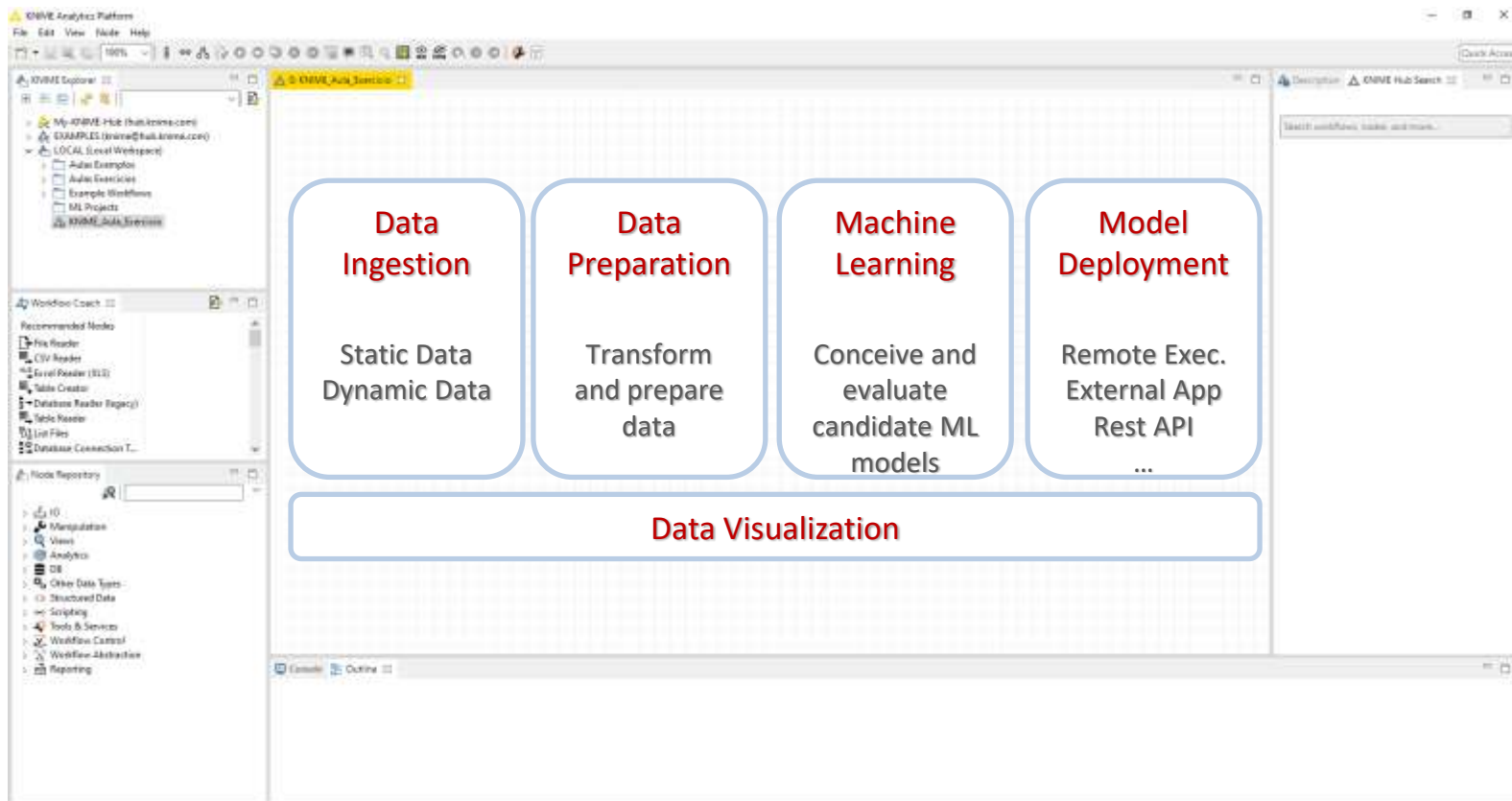


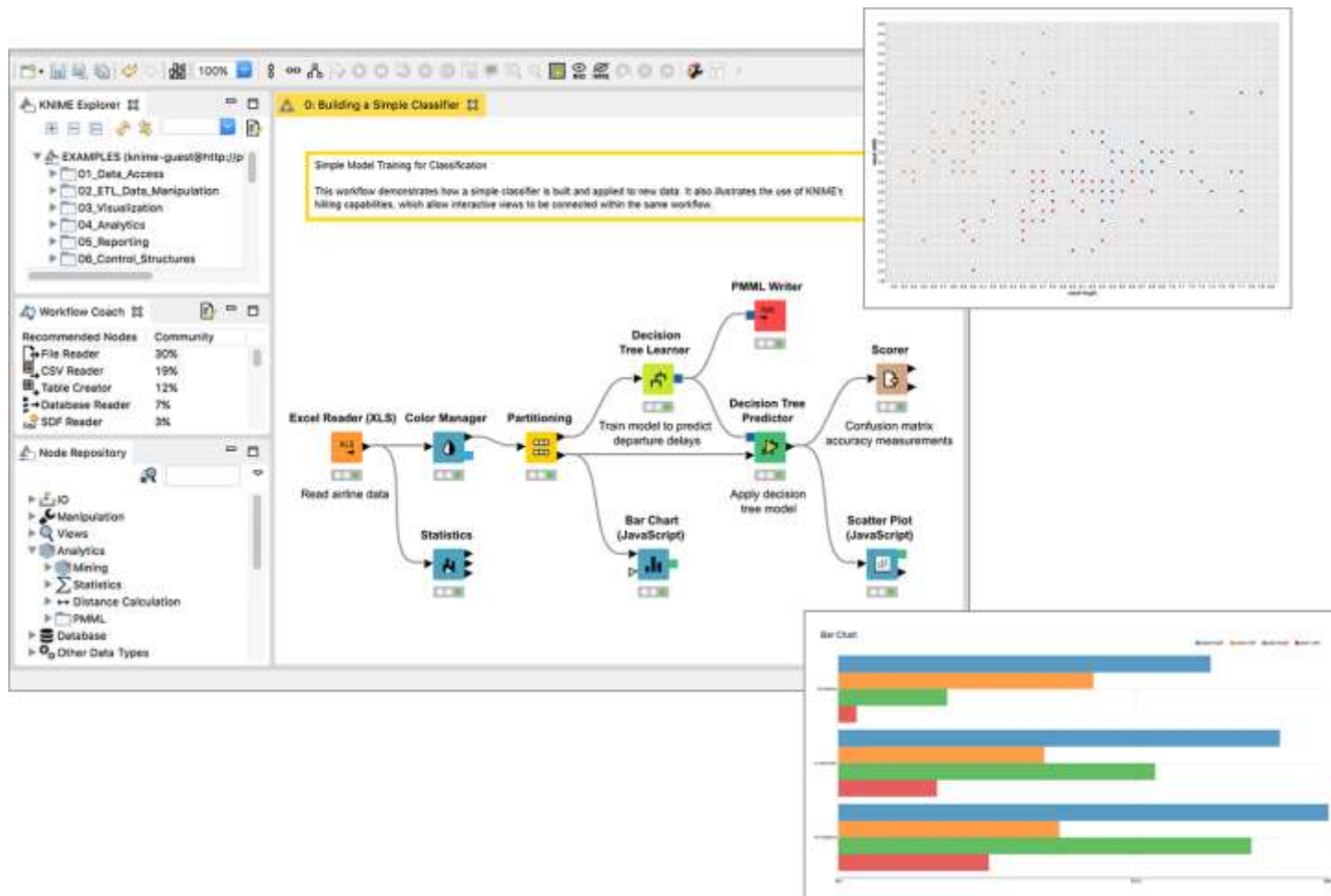
## A Machine Learning Pipeline



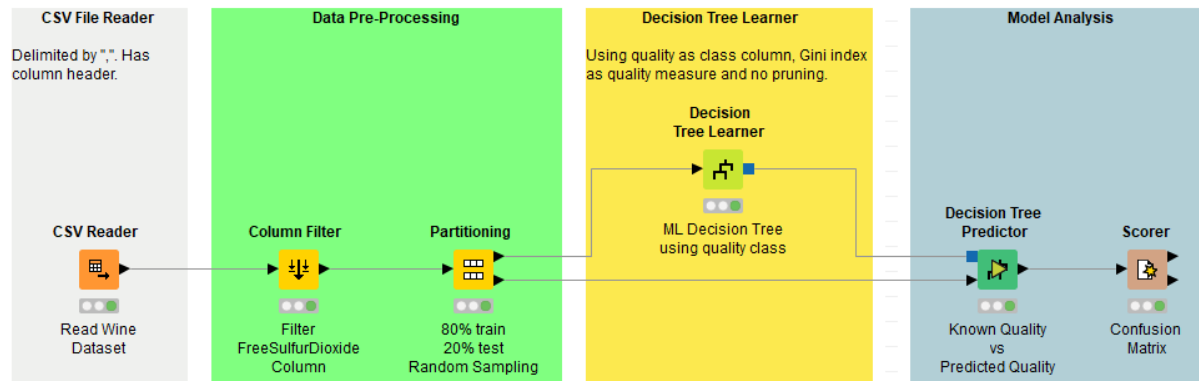
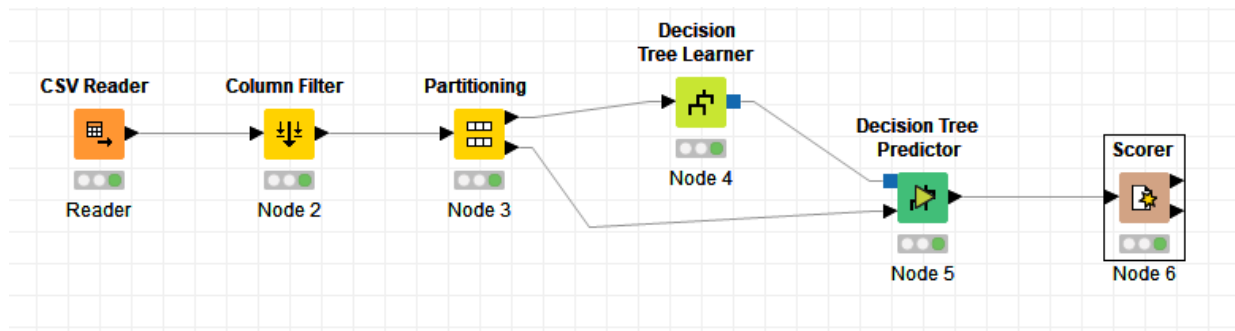
(<https://towardsdatascience.com/architecting-a-machine-learning-pipeline-a847f094d1c7>)

## Fluxo de Trabalho Típico @ Knime





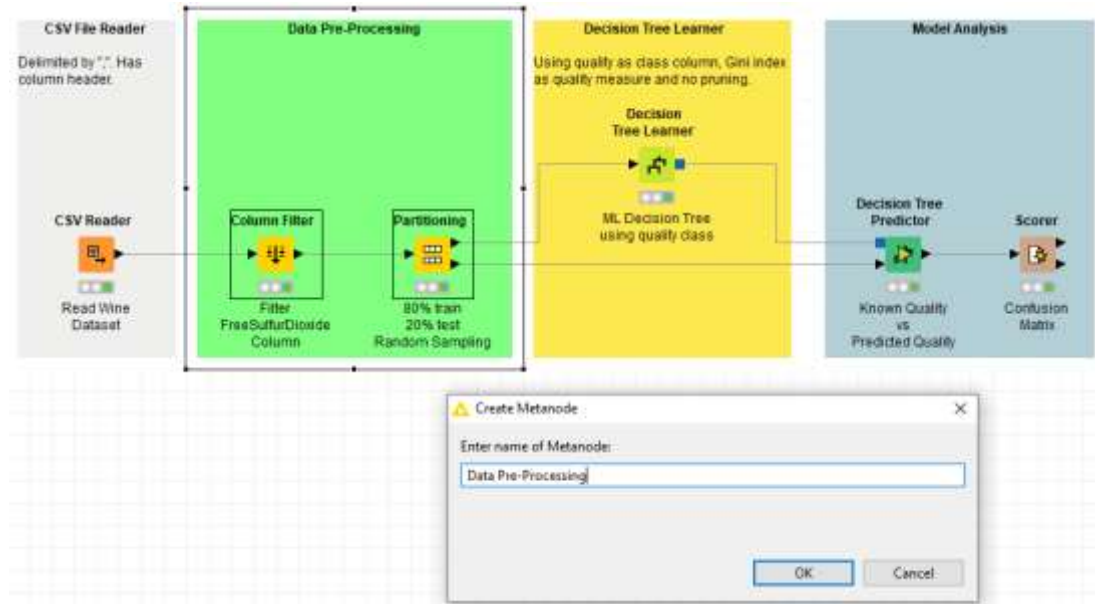
- Dar novos nomes aos nodos
- Adicionar anotações
- Utilizar metanodos



- Um metanodo é um nodo com outros nodos dentro!
- Usar metanodos para organizar o trabalho!

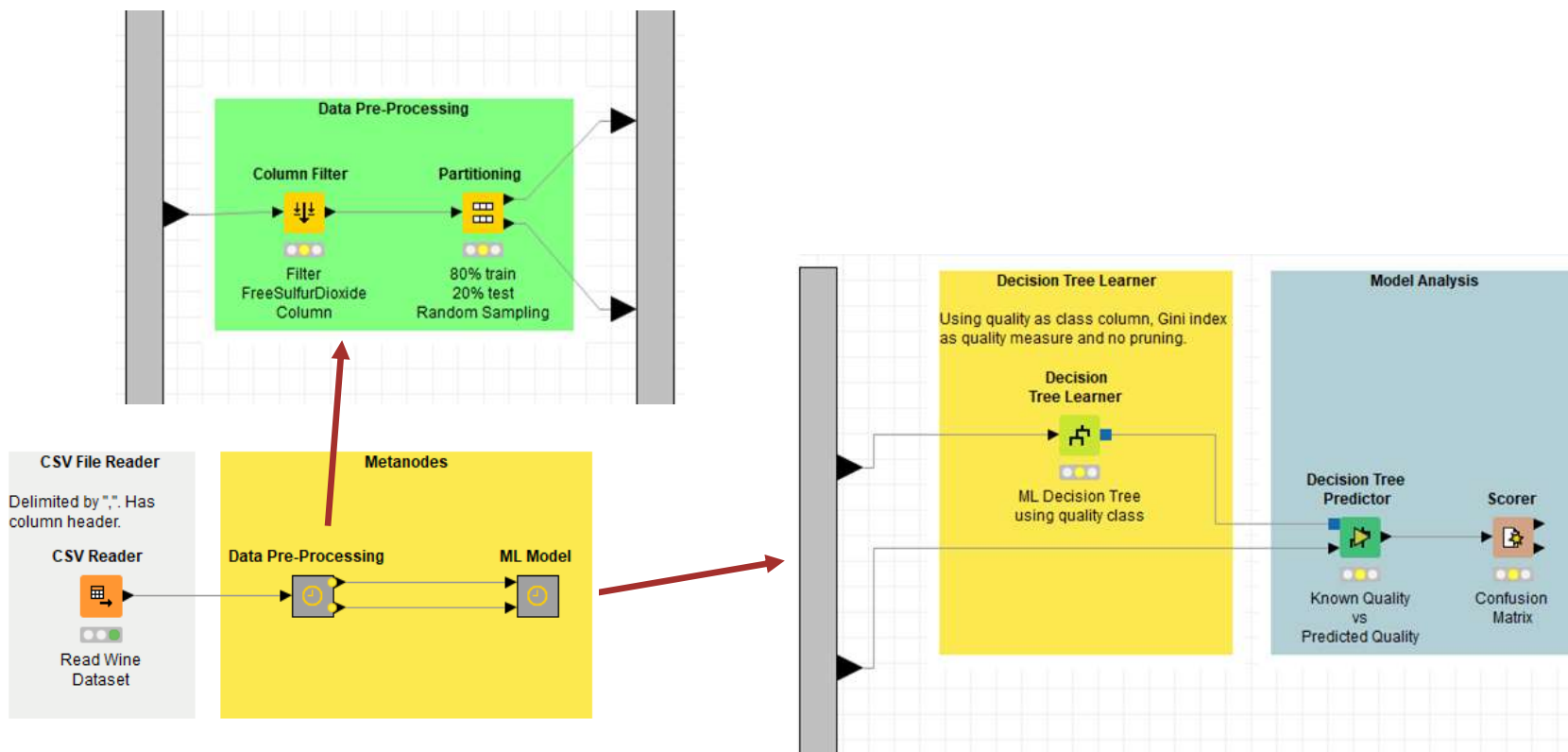


- Um metanodo é um nodo com outros nodos dentro!
- Usar metanodos para organizar o trabalho!

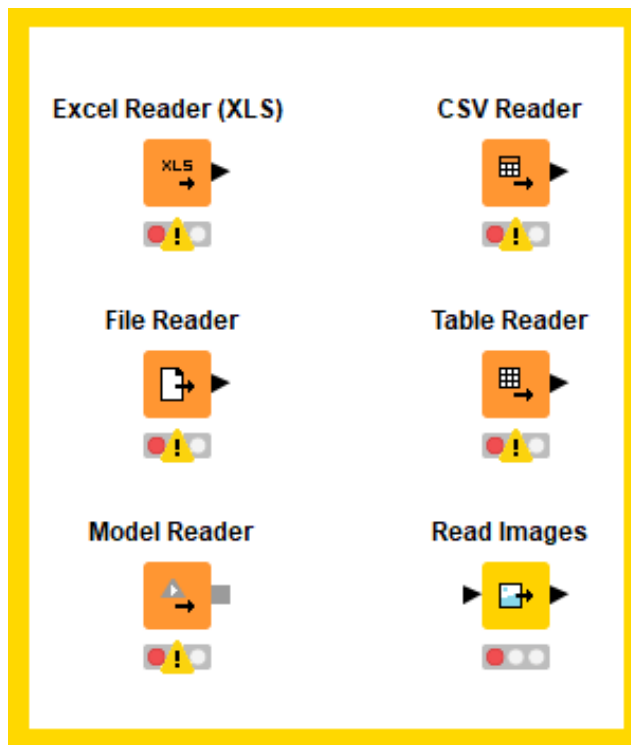








## Ingestão de dados ***KNIME Data Readers***



## Ingestão de dados *KNIME Data Readers*

CSV Reader



Dialog - 3:1 - CSV Reader (Read Wine)

File

Settings Limit Rows Encoding Flow Variables Memory Policy

Input location:

knime://knime.workflow/winequality-red.csv

☐ Custom connection timeout [s]:

Reader options:

Column Delimiter  Row Delimiter

Quote Char  Comment Char

☒ Has Column Header ☐ Has Row Header

☐ Support Short Lines

File Reader



Dialog - 3.9 - File Reader

File

Settings | Flow Variables | Memory Policy

Enter ASCII data file location: (press 'Enter' to update preview)

knime://knime.workflow/winequality-red.csv Browse...

☐ Preserve user settings for new location Rescan

Basic Settings

☐ read row IDs Column delimiter: , Advanced...

☒ read column headers ☒ ignore spaces and tabs

☐ Java-style comments Single line comment:

Preview

Click column header to change column properties (\* = name/type user settings)

Row ID	D fixed a...	D volatile ...	D citric acid	D residual...	D chlorides	D free sul...	D s
Row0	7.4	0.7	0	1.9	0.076	11	34
Row1	7.8	0.88	0	2.6	0.098	25	67
Row2	7.8	0.76	0.04	2.3	0.092	15	34
Row3	11.2	0.28	0.56	1.9	0.075	17	60
Row4	7.4	0.7	0	1.9	0.076	11	34
Row5	7.4	0.66	0	1.8	0.075	13	40
Row6	7.9	0.6	0.06	1.6	0.069	15	59
Row7	7.3	0.65	0	1.2	0.065	15	21
Row8	7.8	0.58	0.02	2	0.073	9	18
Row9	7.5	0.5	0.36	6.1	0.071	17	102
Row10	6.7	0.58	0.08	1.8	0.097	15	65
Row11	7.5	0.5	0.36	6.1	0.071	17	102
Row12	5.6	0.615	0	1.6	0.089	16	59
Row13	7.8	0.61	0.29	1.6	0.114	9	29
Row14	8.9	0.62	0.18	3.8	0.176	52	145
Row15	8.9	0.62	0.19	3.9	0.17	51	148
Row16	8.5	0.28	0.56	1.8	0.092	35	103
Row17	8.1	0.56	0.28	1.7	0.368	16	56
Row18	7.4	0.59	0.08	4.4	0.086	6	29
Row19	7.9	0.32	0.51	1.8	0.341	17	56
Row20	8.9	0.22	0.48	1.8	0.077	29	60
Row21	7.6	0.39	0.31	2.3	0.082	23	71
Row22	7.9	0.43	0.21	1.6	0.106	10	37
Row23	8.5	0.49	0.11	2.3	0.084	9	67
Row24	6.9	0.4	0.14	2.4	0.085	21	40

OK Apply Cancel ?

## Ingestão de dados *KNIME Data Readers*

## Excel Reader (XLS)



Dialog - 2.7 - Excel Reader (XLS) (Read Calls data)

File

XLS Reader Settings | Flow Variables | Memory Policy

Select file to read:

lreme://lreme.workflow/CallsData.xls Browse...

Adjust Settings:

Select the sheet to read: <first sheet with data> Connect timeout [s]: 5

Column Names:

☒ Table contains column names in row number: 1 (Row numbers start with 1. Mouse over header to see row number.)

Row IDs:

☒ Generate RowIDs (index incrementing, starting with Row0) ☐ Generate RowIDs (index as per sheet content, skipped rows will increment index)

☐ Table contains row IDs in column: A ☐ Make row IDs unique

Select the columns and rows to read:

☒ Read entire data sheet, or ...

read columns from: A to:

and read rows from: 1 to:

Tip: Mouse over the column and row headers in the "File Content" tab to identify cell coordinates

On evaluation error:

☒ Insert an error pattern: #XL\_EVAL\_ERROR#

☐ Insert a missing cell

More Options:

☐ Skip empty columns ☐ Reevaluate formulas (leave unchecked if uncertain; see node description for details)

☒ Skip hidden columns ☐ Disable Preview (does not compute the output table structure)

☒ Skip empty rows

Preview: File Content

Preview with current settings: CallsData.xls [ChurnDataset]

refresh

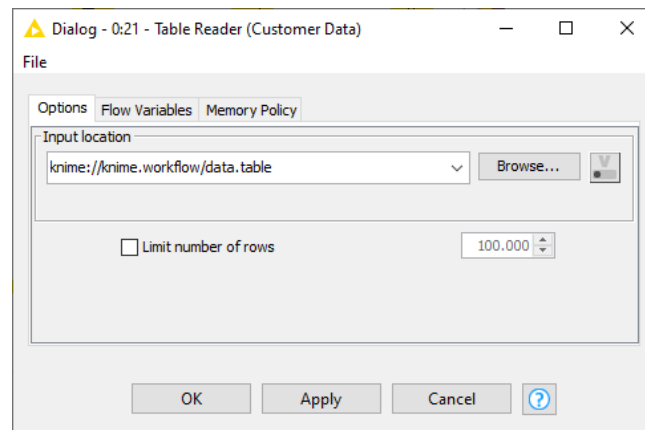
Row ID	I	Whal M...	D	Day Mins	D	Eve Mins	D	Night Mins	D	Snd Mins	I	CustSe...	I	Day Calls	D	Day Ch...	I	Eve C
Row0	25		265.1		197.4		244.7		10		1		110		45.07		99	
Row1	26		161.6		195.5		254.4		13.7		1		123		27.47		103	
Row2	0		245.4		121.2		162.6		12.2		0		114		41.38		110	
Row3	0		299.4		61.9		196.9		6.6		2		71		50.9		88	
Row4	0		166.7		148.3		186.9		10.1		3		113		28.34		122	
Row5	0		223.4		220.6		203.9		6.3		0		98		37.98		101	
Row6	24		218.2		348.5		212.6		7.5		3		88		37.09		108	
Row7	0		157		103.1		211.8		7.1		0		79		26.69		94	
Row8	0		184.5		351.6		215.8		8.7		1		97		31.37		80	
Row9	37		258.6		222		326.4		11.2		0		84		43.96		111	
Row10	0		129.1		228.5		208.8		12.7		4		137		21.95		83	
Row11	0		187.7		163.4		196		9.1		0		127		31.91		148	
Row12	0		128.8		104.9		141.1		11.2		1		96		21.9		71	

OK Apply Cancel ?

**gestão de dados**  
**IME Data Readers**

## Ingestão de dados *KNIME Data Readers*

Table Reader



Input Features/Input Vector

Target/Class/Label

File Table - 3:1 - CSV Reader (Read Wine)

File Hilite Navigation View

Table "winequality-red.csv" - Rows: 1599 Spec - Columns: 12 Properties Flow Variables

Row ID	D fixed a...	D volatile ...	D citric acid	D residual...	D chlorides	D free sul...	D total su...	D density	D pH	D sulphates	D alcohol	S quality
Row0	7.4	0.7	0	1.9	0.076	11	34	0.998	3.51	0.56	9.4	=5
Row1	7.8	0.88	0	2.6	0.098	25	67	0.997	3.2	0.68	9.8	=5
Row2	7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	=5
Row3	11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	=6
Row4	7.4	0.7	0	1.9	0.076	11	34	0.998	3.51	0.56	9.4	=5
Row5	7.4	0.66	0	1.8	0.075	13	40	0.998	3.51	0.56	9.4	=5
Row6	7.9	0.6	0.06	1.6	0.069	15	59	0.996	3.3	0.46	9.4	=5
Row7	7.3	0.65	0	1.2	0.065	15	21	0.995	3.39	0.47	10	=7
Row8	7.8	0.58	0.02	2	0.073	9	18	0.997	3.36	0.57	9.5	=7
Row9	7.5	0.5	0.36	6.1	0.071	17	102	0.998	3.35	0.8	10.5	=5
Row10	6.7	0.58	0.08	1.8	0.097	15	65	0.996	3.28	0.54	9.2	=5
Row11	7.5	0.5	0.36	6.1	0.071	17	102	0.998	3.35	0.8	10.5	=5
Row12	5.6	0.615	0	1.6	0.089	16	59	0.994	3.58	0.52	9.9	=5
Row13	7.8	0.61	0.29	1.6	0.114	9	29	0.997	3.26	1.56	9.1	=5
Row14	8.9	0.62	0.18	3.8	0.176	52	145	0.999	3.16	0.88	9.2	=5
Row15	8.9	0.62	0.19	3.9	0.17	51	148	0.999	3.17	0.93	9.2	=5
Row16	8.5	0.28	0.56	1.8	0.092	35	103	0.997	3.3	0.75	10.5	=7
Row17	8.1	0.56	0.28	1.7	0.368	16	56	0.997	3.11	1.28	9.3	=5
Row18	7.4	0.59	0.08	4.4	0.086	6	29	0.997	3.38	0.5	9	=4
Row19	7.9	0.32	0.51	1.8	0.341	17	56	0.997	3.04	1.08	9.2	=6
Row20	8.9	0.22	0.48	1.8	0.077	29	60	0.997	3.39	0.53	9.4	=6
Row21	7.6	0.39	0.31	2.3	0.082	23	71	0.998	3.52	0.65	9.7	=5
Row22	7.9	0.43	0.21	1.6	0.106	10	37	0.997	3.17	0.91	9.5	=5
Row23	8.5	0.49	0.11	2.3	0.084	9	67	0.997	3.17	0.53	9.4	=5
Row24	6.9	0.4	0.14	2.4	0.085	21	40	0.997	3.43	0.63	9.7	=6



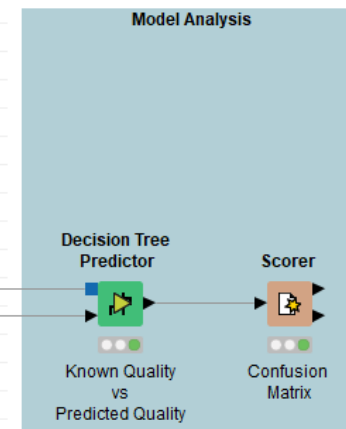
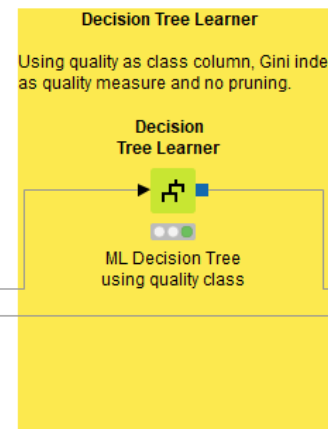
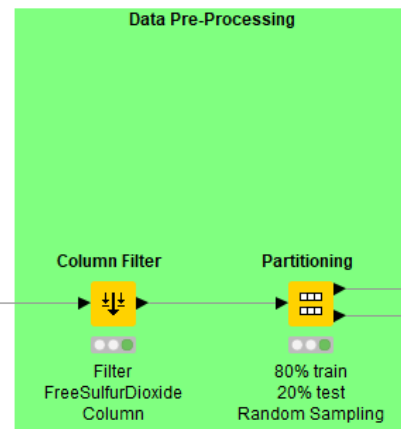
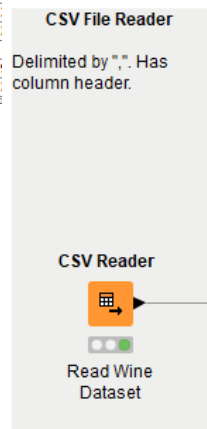
# Partição de dados

File Table - 3:1 - CSV Reader (Read Wine)

File Hilite Navigation View

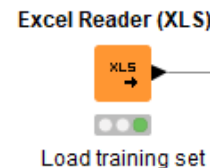
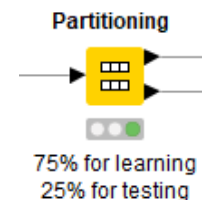
Table "winequality-red.csv" - Rows: 1599 Spec - Columns: 12 Properties Flow Variables

Row ID	[D] fixed a...	[D] volatile ...	[D] citric acid	[D] residual...	[D] chlorides	[D] free sul...	[D] total su...	[D] density	[D] pH	[D] sulphates	[D] alcohol	[S] quality
Row0	7.4	0.7	0	1.9	0.076	11	34	0.998	3.51	0.56	9.4	=5
Row1	7.8	0.88	0	2.6	0.098	25	67	0.997	3.2	0.68	9.8	=5
Row2	7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	=5
Row3	11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	=6
Row4	7.4	0.7	0	1.9	0.076	11	34	0.998	3.51	0.56	9.4	=5
Row5	7.4	0.66	0	1.8	0.075	13	40	0.998	3.51	0.56	9.4	=5
Row6	7.9	0.6	0.06	1.6	0.069	15	59	0.996	3.3	0.46	9.4	=5
Row7	7.3	0.65	0	1.2	0.065	15	21	0.995	3.39	0.47	10	=7
Row8	7.8	0.58	0.02	2	0.073	9	18	0.997	3.36	0.57	9.5	=7
Row9	7.5	0.5	0.36	6.1	0.071	17	102	0.998	3.35	0.8	10.5	=5
Row10	6.7	0.58	0.08	1.8	0.097	15	65	0.996	3.28	0.54	9.2	=5
Row11	7.5	0.5	0.36	6.1	0.071	17	102	0.998	3.35	0.8	10.5	=5
Row12	5.6	0.615	0	1.6	0.089	16	59	0.994	3.58	0.52	9.9	=5
Row13	7.8	0.61	0.29	1.6	0.114	9	29	0.997	3.26	1.56	9.1	=5
Row14	8.9	0.62	0.18	3.8	0.176	52	145	0.999	3.16	0.88	9.2	=5
Row15	8.9	0.62	0.19	3.9	0.17	51	148	0.999	3.17	0.93	9.2	=5
Row16	8.5	0.28	0.56	1.8	0.092	35	103	0.997	3.3	0.75	10.5	=7
Row17	8.1	0.56	0.28	1.7	0.368	16	56	0.997	3.11	1.28	9.3	=5
Row18	7.4	0.59	0.08	4.4	0.086	6	29	0.997	3.38	0.5	9	=4
Row19	7.9	0.32	0.51	1.8	0.341	17	56	0.997	3.04	1.08	9.2	=6
Row20	8.9	0.22	0.48	1.8	0.077							
Row21	7.6	0.39	0.31	2.3	0.082							
Row22	7.9	0.43	0.21	1.6	0.106							
Row23	8.5	0.49	0.11	2.3	0.084							
Row24	6.9	0.4	0.14	2.4	0.085							

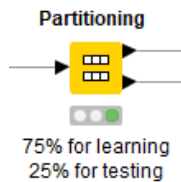


## Partição de dados

- Construção de um modelo supervisionado de ML
  - Usar um *dataset* de treino para treinar um modelo
    - Aprendizagem
  - Usar um dataset de teste para testar o modelo
    - Avaliar o modelo com dados “não vistos” durante o treino
  - Usar um *dataset* de validação (aconselhável)
    - Promove uma validação sem viés de um modelo sobre o *dataset* de treino enquanto se afinam os hiperparâmetros



## Partição de dados



Dialog - 2:14 - Partitioning (75% for learning)

File

First partition | Flow Variables | Memory Policy

Choose size of first partition

☐ Absolute

☒ Relative[%]

☐ Take from top

☐ Linear sampling

☐ Draw randomly

☒ Stratified sampling

☒ Use random seed

OK Apply Cancel ?

## **Exploração dos dados**

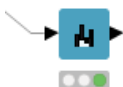
- Porquê?
  - Compreender os dados e as suas características
  - Analisar e avaliar a qualidade dos dados
  - Perceber padrões e informações relevantes

- Como?
  - Tendências centrais:
    - average, mode, median...
  - Dispersão estatística
    - variance, standard deviation, interquartile range...
  - Distribuição de probabilidades
    - Gaussian, Uniform, Exponential...
  - Correlação/Dependência
    - between pairs of features, with the dependent feature...
  - Visualização de dados
    - tables, charts, boxplots, scatter plots, histograms, ...

# Exploração de dados

## ***KNIME Data Explorer Node***

Data Explorer



Data Explorer View

Summary | Summary | Data Preview

Search:

Column	Exclude Column	Minimum	Maximum	Mean	Standard Deviation	Variance	Skewness	Kurtosis	Overall Sum	No. zeros	No. missing
CustomerKey		11000	27336	17559.647	5576.039	3102215.201	0.333	-1.566	266330291	0	0
WebActivity		0	5	0.999	1.520	2.310	1.395	0.687	15198	9199	0
SentimentRating		0	5	1.851	1.620	2.624	0.482	-0.958	28073	4175	0
EstimatedYearlyIncome		10000	170000	57718.072	32091.910	1029890707.928	0.796	0.617	875410080	0	0
NumberOfContracts		0	4	1.465	1.145	1.311	0.430	-0.457	22227	3711	0
Age		29	109	48.203	11.300	127.694	0.571	-0.182	731101	0	0
Target		0	1	0.487	0.500	0.250	0.053	-1.897	7303	7794	0
Available401K		0	1	0.666	0.480	0.211	-0.854	-1.279	10962	4605	0
CustomerValueSegment		1	3	2.597	0.689	0.475	-0.129	-0.898	31808	0	0
ChurnScore		0	1	0.269	0.332	0.110	1.254	0.296	4076.380	5299	0
CallActivity		1	5	3.237	1.262	1.594	-0.302	-0.915	49094	0	0

Showing 1 to 11 of 11 entries

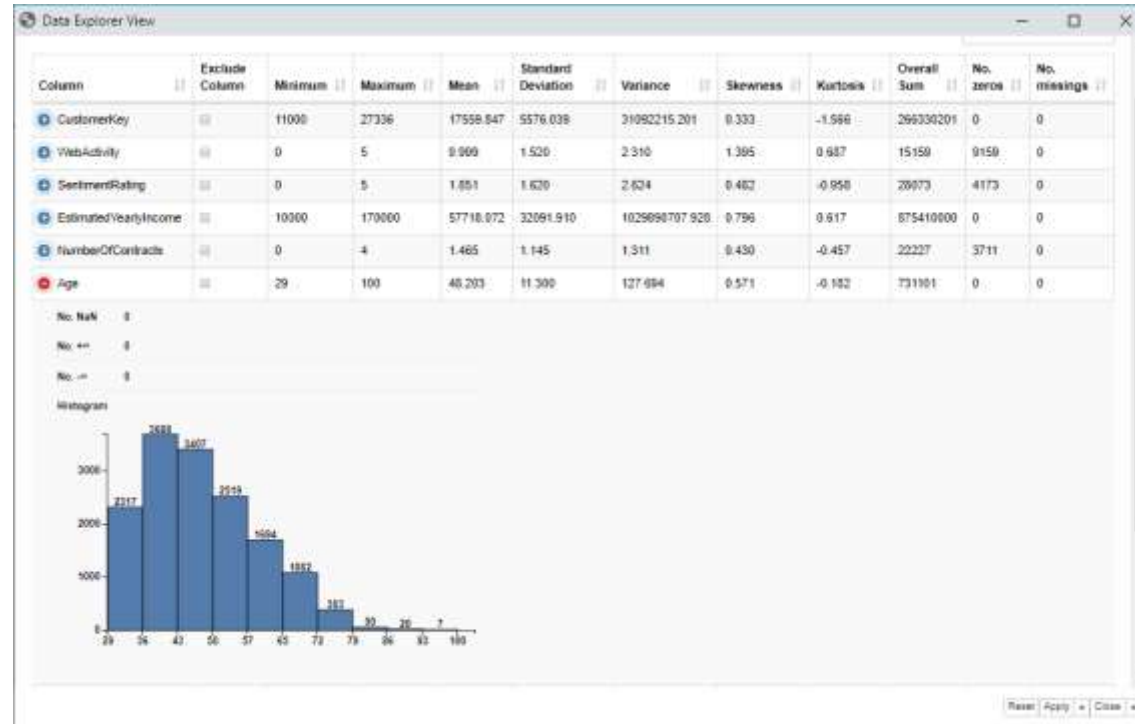
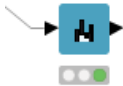
Reset Apply Close



Install **KNIME JavaScript Views (Labs)** extension

## Exploração de dados *KNIME Data Explorer Node*

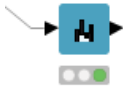
Data Explorer



Install **KNIME JavaScript Views (Labs)** extension

## Exploração de dados ***KNIME Data Explorer Node***





Data Explorer



Data Explorer View

Summary | Overview | Data Preview

Search:

Column	Exclude Column	No. missing	Unique values	All nominal values	Frequency Bar Chart
Sentiment Analysis	<input type="checkbox"/>	0	5	Very Negative, Negative, Slightly Negative, Positive, Slightly Positive, Very Positive	
MaritalStatus	<input type="checkbox"/>	0	2	M, S	
Gender	<input type="checkbox"/>	0	2	M, F	
Products	<input type="checkbox"/>	0	5	private investment, p-b investment, fund manager, gold investment, co investment	

Showing 1 to 4 of 4 entries

Reset | Apply | Close

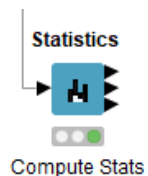


Install **KNIME JavaScript Views (Labs)** extension



# Exploração de dados

## ***KNIME Statistics Node***



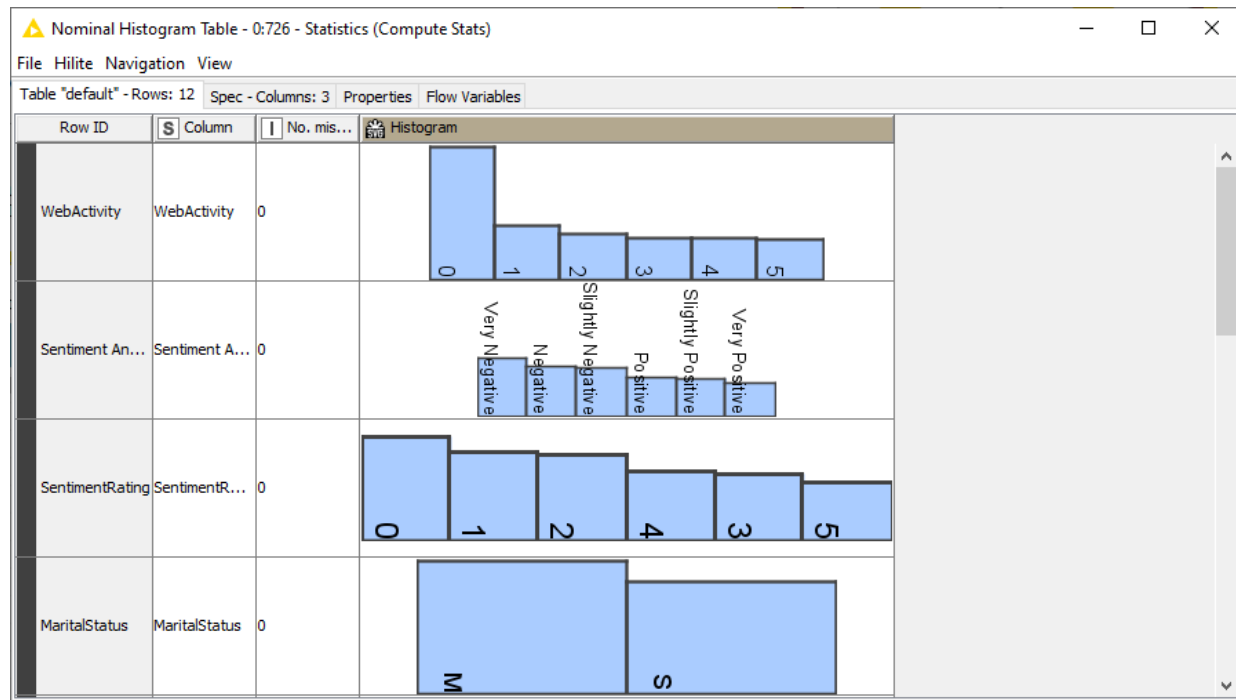
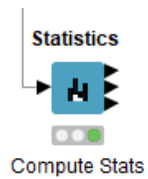
Statistics Table - 0:726 - Statistics (Compute Stats)

File Hilite Navigation View

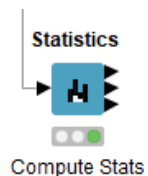
Table "default" - Rows: 11 Spec - Columns: 16 Properties Flow Variables

Row ID	S Column	D Min	D Max	D Mean	D Std. deviation	D Variance	D Skewness	D Kurtosis	D Overall sum	I No. missings	I Nc
EstimatedYea...	EstimatedYe...	10,000	170,000	57,718.072	32,091.91	1,029,890,7...	0.796	0.617	875,410,000	0	0
NumberOfCo...	NumberOfC...	0	4	1.465	1.145	1.311	0.43	-0.457	22,227	0	0
Age	Age	29	100	48.203	11.3	127.694	0.571	-0.182	731,101	0	0
Target	Target	0	1	0.487	0.5	0.25	0.053	-1.997	7,383	0	0
Available401K	Available401K	0	1	0.696	0.46	0.211	-0.854	-1.27	10,562	0	0

## Exploração de dados *KNIME Statistics Node*



# Exploração de dados *KNIME Statistics Node*



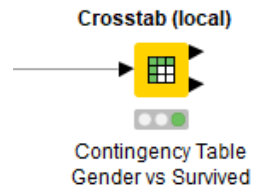
Occurrences Table - 0:726 - Statistics (Compute Stats)

File Hilite Navigation View

Table "default" - Rows: 68 Spec - Columns: 36 Properties Flow Variables

Row ID	WebActivity	Count (WebActivity)	Relative Frequency (WebActivity)	Sentiment Analysis	Count (Sentiment Analysis)	Relative Frequency (Sentiment Analysis)
Row0	0	9159	0.604	Very Negative	4173	0.275
Row1	1	1983	0.131	Negative	3122	0.206
Row2	2	1366	0.09	Slightly Negative	3023	0.199
Row3	3	963	0.063	Positive	1960	0.129
Row4	4	925	0.061	Slightly Positive	1690	0.111
Row5	5	771	0.051	Very Positive	1199	0.079
Row6	?	?	?	?	?	?
Row7	?	?	?	?	?	?
Row8	?	?	?	?	?	?
Row9	?	?	?	?	?	?
Row10	?	?	?	?	?	?
Row11	?	?	?	?	?	?
Row12	?	?	?	?	?	?
Row13	?	?	?	?	?	?
Row14	?	?	?	?	?	?
Row15	?	?	?	?	?	?
Row16	?	?	?	?	?	?
Row17	?	?	?	?	?	?
Row18	?	?	?	?	?	?
Row19	?	?	?	?	?	?
Row20	?	?	?	?	?	?
Row21	?	?	?	?	?	?

## Exploração de dados *KNIME Contingency Tables*



Cross Tabulation of Survived by Sex

Frequency	female	male	Total	<input checked="" type="checkbox"/> Frequency <input type="checkbox"/> Expected <input type="checkbox"/> Deviation <input type="checkbox"/> Percent <input type="checkbox"/> Row Percent <input type="checkbox"/> Column Percent <input type="checkbox"/> Cell Chi-Square
0	81	468	549	
1	233	109	342	
Total	314	577	891	

Max rows: 10  
Max columns: 10


Statistics for Table of Survived by Sex

Statistic	DF	Value	Prob
Chi-Square	1	263,0506	3,71E-39
Fisher's Exact Test (2-tail)			6,46E-60

# Exploração de dados

## ***KNIME Contingency Tables***

**Crosstab (local)**



Contingency Table  
Gender vs Survived

**Cross tabulation - 0:732 - Crosstab (local)**

File

Frequency Percent	F	M	Total
Negative	1.585	1.537	3.122
	10,4503%	10,1338%	20,5842%
Positive	941	1.019	1.960
	6,2043%	6,7185%	12,9228%
Slightly Negative	1.501	1.522	3.023
	9,8965%	10,0349%	19,9314%
Slightly Positive	861	829	1.690
	5,6768%	5,4658%	11,1426%
Very Negative	2.054	2.119	4.173
	13,5426%	13,9711%	27,5137%
Very Positive	639	560	1.199
	4,2131%	3,6922%	7,9053%
Total	7.581	7.586	15.167
	49,9835%	50,0165%	100%

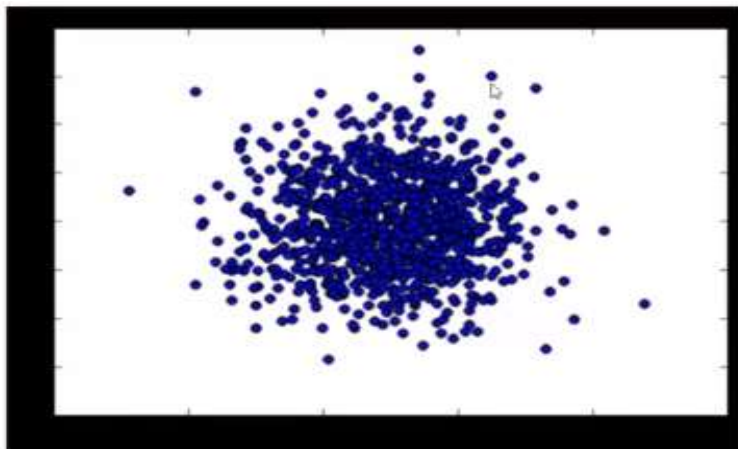
☒ Frequency  
☐ Expected  
☐ Deviation  
☒ Percent  
☐ Row Percent  
☐ Column Percent  
☐ Cell Chi-Square

Max rows: 10  
Max columns: 10

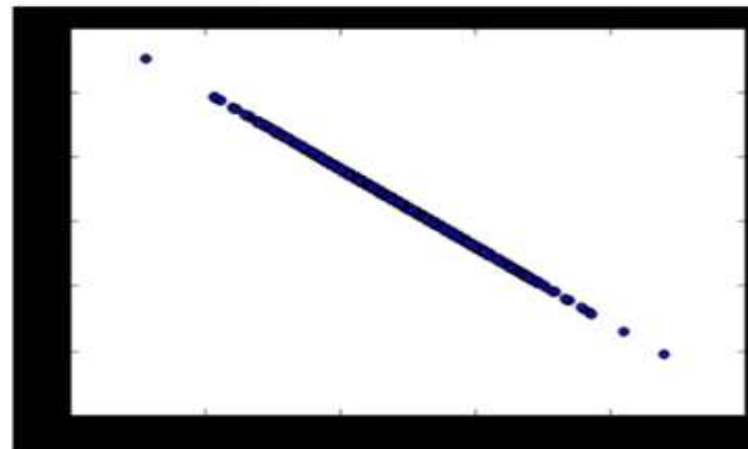
Statistics for Table of Sentiment Analysis by Gender

Statistic	DF	Value	Prob
Chi-Square	5	10,8099	0,0553

- A covariância mede quanto duas variáveis (atributos, colunas, *features*) dependem uma da outra



(baixa covariância)



(alta covariância)

- Calcular a covariância:
  - Considere os dados das duas variáveis como vetores de alta dimensão;
  - Converta-os em vetores de variâncias da média;
  - Calcule o produto escalar (cosseno do ângulo entre eles) dos dois vetores;
  - Divida pelo tamanho da população;

Population Covariance Formula

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

Sample Covariance

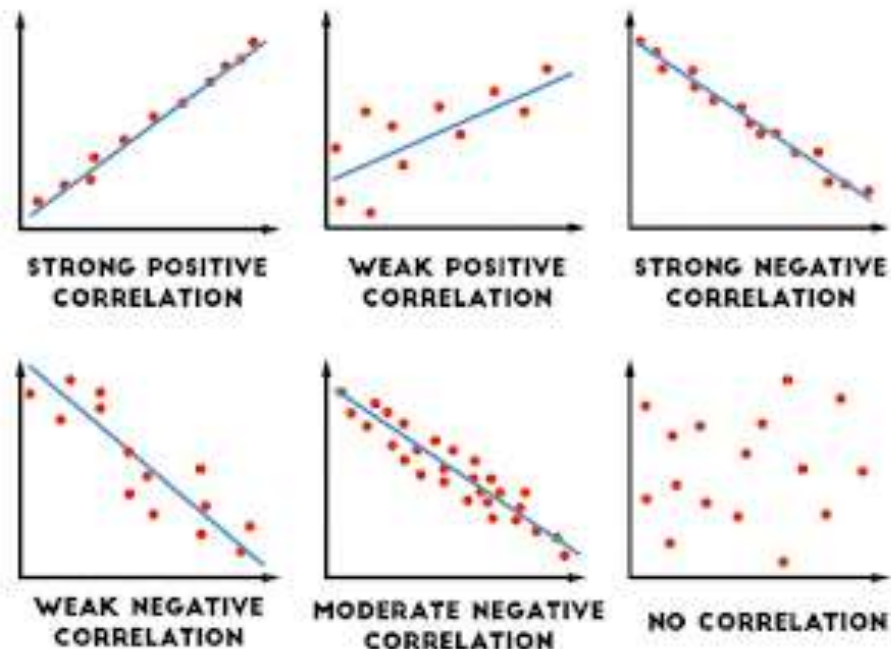
$$\text{Cov}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

## Covariância & Correlação

- Interpretar a **covariância** é difícil:
  - Baixa covariância (próxima de 0) significa que não há muita correlação entre as duas variáveis;
  - Alta covariância (muito superior a 0, ou negativa para relações inversas) significa que existe uma correlação entre as variáveis;
  
- Interpretar a **correlação** é mais fácil:
  - Valor de normalização da covariância dividido pelos desvios padrão de ambas as variáveis:
    - Correlação = -1: correlação inversa perfeita
    - Correlação = 0: sem correlação
    - Correlação = 1: correlação perfeita

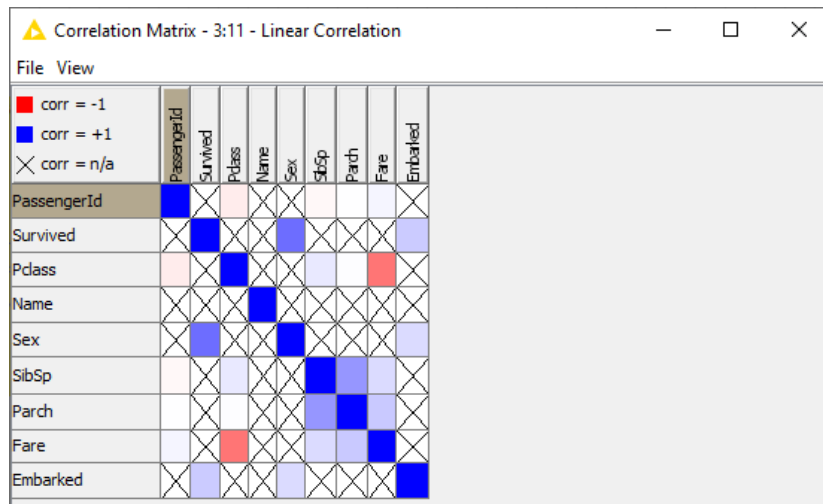
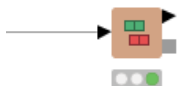


- Mas... a correlação não implica causalidade!!!
  - Apenas uma experiência controlada e aleatória pode fornecer informações sobre a causalidade;
  - Use a correlação para decidir quais as experiências a realizar.



## Matriz de Correlação

Linear Correlation



Correlation measure - 3:11 - Linear Correlation

File Hilite Navigation View

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

Row ID	D Passen...	D Survived	D Pclass	D Name	D Sex	D SibSp	D Parch	D Fare	D Embarked
PassengerId	1.0	?	-0.074684...	?	?	-0.02572993...	0.0026940469...	0.04019030952...	?
Survived	?	1.0	?	?	0.5737...	?	?	?	0.20559893...
Pclass	-0.0746846...	?	1.0	?	?	0.084898459...	0.0060928325...	-0.5393077410...	?
Name	?	?	?	1.0	?	?	?	?	?
Sex	?	0.57374697...	?	?	1.0	?	?	?	0.14153725...
SibSp	-0.0257299...	?	0.0848984...	?	?	1.0	0.4102760703...	0.14270976638...	?
Parch	0.00269404...	?	0.0060928...	?	?	0.410276070...	1.0	0.21318809003...	?
Fare	0.04019030...	?	-0.539307...	?	?	0.142709766...	0.2131880900...	1.0	?
Embarked	?	0.20559893...	?	?	0.1415...	?	?	?	1.0

## Matriz de Correlação

- Queremos manter atributos (*features*) altamente correlacionados?
- Ambos positivos e negativamente correlacionados?
- E quanto à correlação entre os atributos dependentes e independentes?
- ...

## Visualização de Dados

