

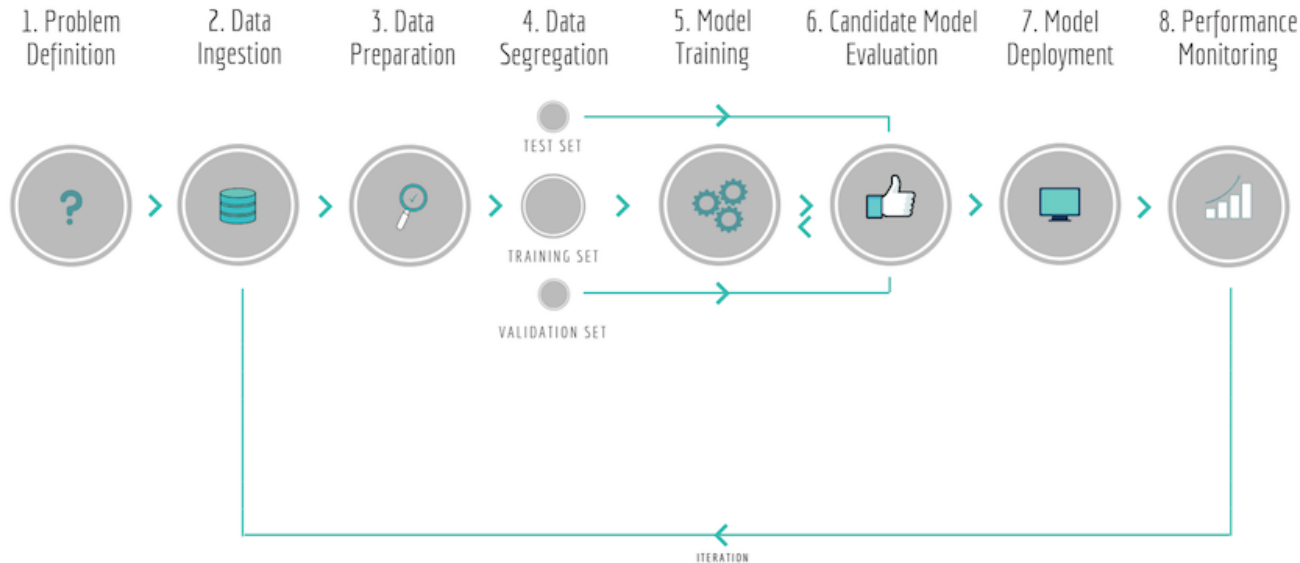
DADOS e APRENDIZAGEM AUTOMÁTICA

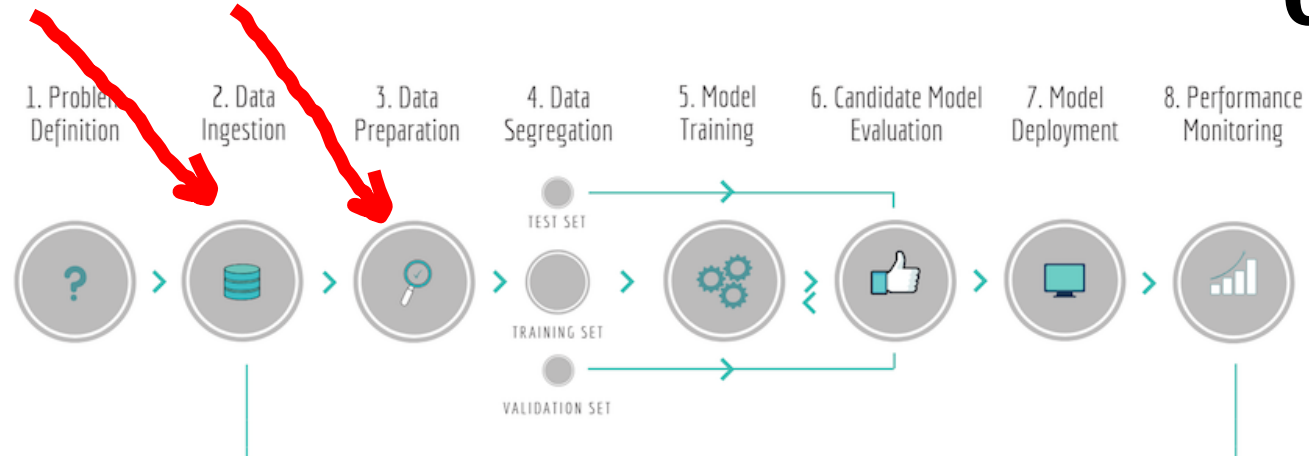
Data Exploration and Preparation

MESTRADO (integrado) EM ENGENHARIA INFORMÁTICA

Fluxo Aprendizagem Automatica

Fluxo de um processo de Aprendizagem Automática





- **Data Quality and Exploration**
- **Basic Data Preparation**
- **Advanced Data Preparation**
 - **Feature Scaling**
 - **Outlier Detection**
 - **Feature Selection**
 - **Missing Values Treatment**
 - **Nominal Value Discretization**
 - **Binning/Discretization**
 - **Feature Engineering**

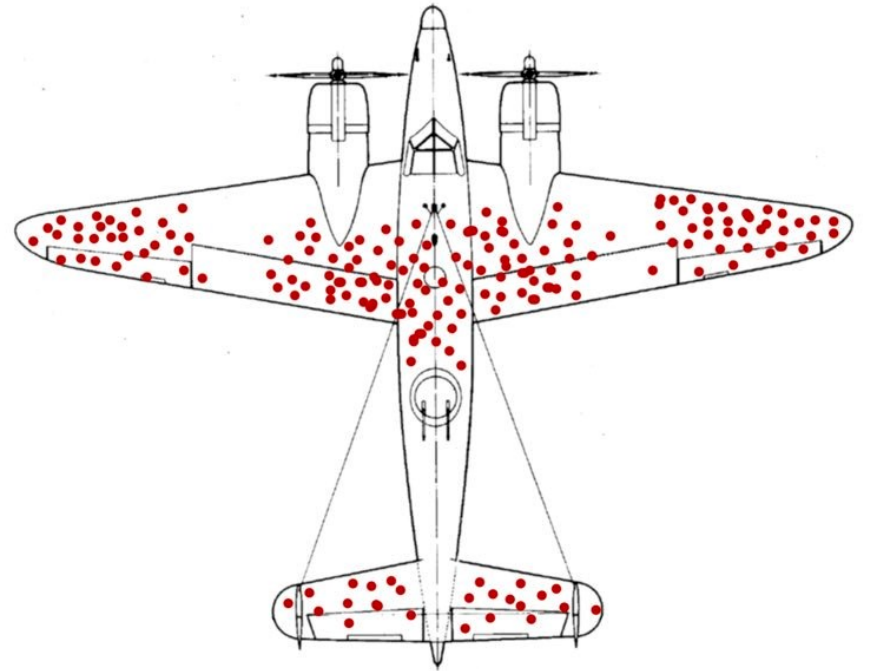
Think clearly...

During WWII, the US Navy tried to determine where they needed to armor their aircraft to ensure they came back home. They ran an analysis of where planes had been shot up.

Everybody told that, obviously, the places that needed to be up-armored are the wingtips, the central body, and the elevators. That's where the planes were all getting shot up!

Abraham Wald, a statistician, disagreed.

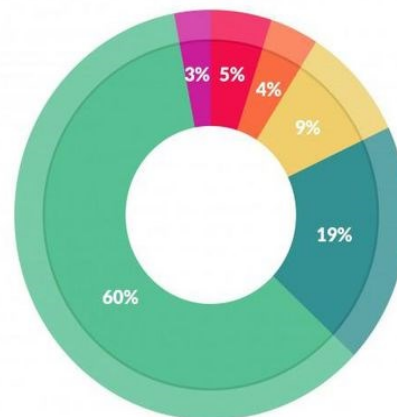
Why?



Data Quality

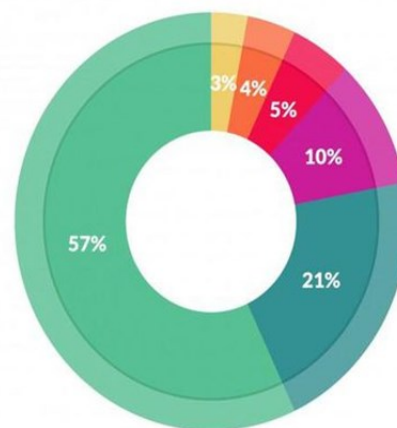
Indeed... Cleaning and manipulating data may be considered as the:

- **Most Time-Consuming task**
- **Least Enjoyable task** (by some!)



What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets: 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%



What's the least enjoyable part of data science?

- Building training sets: 10%
- Cleaning and organizing data: 57%
- Collecting data sets: 21%
- Mining data for patterns: 3%
- Refining algorithms: 4%
- Other: 5%

A few problems... How to solve them?

- **Missing values**

- Information that is not available because it wasn't collected or because it consisted of sensitive information
- Features that are not applicable in all cases

- **Duplicated Records**

- Same (or similar) data collected from different sources

File Table - 2:1 - File Reader (Reading adult.csv)

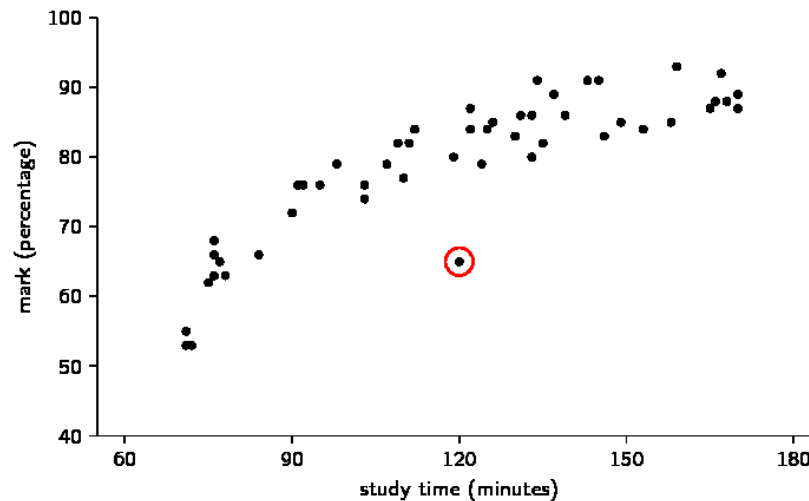
File Hilite Navigation View

Table "adult.csv" - Rows: 32561 Spec - Columns: 15 Properties Flow Variables

Row ID	age	workclass	fnlwgt	education	educati...	marital...	occupa...	relation...	race	sex	I
Row30711	18	?	157131	HS-grad	9	Never-married	?	Own-child	White	Female	0
Row30712	27	Local-gov	255237	Bachelors	13	Never-married	Prof-specialty	Not-in-family	White	Female	0
Row30713	56	?	192325	Some-college	10	Divorced	?	Not-in-family	White	Female	0
Row30714	40	Private	163342	HS-grad	9	Never-married	Adm-clerical	Not-in-family	White	Female	0
Row30715	31	Private	Missing Value	Bachelors	13	Married-civ...	Sales	Husband	White	Male	0
Row30716	18	Private	206008	Some-college	10	Never-married	Sales	Unmarried	White	Male	217
Row30717	25	Private	397317	Assoc-acdm	12	Never-married	Prof-specialty	Not-in-family	White	Female	0
Row30718	36	Private	745768	Some-college	10	Never-married	Protective-s...	Unmarried	Black	Female	0
Row30719	38	Private	141550	10th	6	Divorced	Craft-repair	Not-in-family	White	Male	0
Row30720	52	Private	35576	HS-grad	9	Widowed	Craft-repair	Not-in-family	White	Male	0
Row30721	23	Private	376383	HS-grad	9	Never-married	Other-service	Unmarried	White	Male	0
Row30722	48	Self-emp-no...	200825	Some-college	10	Married-civ...	Exec-manag...	Husband	White	Male	0
Row30723	34	?	362787	HS-grad	9	Never-married	?	Unmarried	Black	Female	0
Row30724	46	Private	116789	HS-grad	9	Married-civ...	Adm-clerical	Husband	White	Male	0
Row30725	26	Private	160300	HS-grad	9	Married-spo...	Protective-s...	Not-in-family	White	Male	0
Row30726	47	Private	362654	HS-grad	9	Married-civ...	Machine-op...	Husband	White	Male	0
Row30727	21	?	107801	Some-college	10	Never-married	?	Own-child	White	Female	0
Row30728	65	Private	170939	Bachelors	13	Divorced	Prof-specialty	Not-in-family	White	Male	672
Row30729	31	Local-gov	224234	HS-grad	9	Married-civ...	Transport-m...	Husband	Black	Male	0
Row30730	38	Private	478346	HS-grad	9	Married-civ...	Exec-manag...	Wife	White	Female	768
Row30731	68	Private	211162	HS-grad	9	Married-civ...	Exec-manag...	Husband	White	Male	0
Row30732	26	Private	147638	Bachelors	13	Never-married	Adm-clerical	Other-relative	Asian-Pac-Is...	Female	0
Row30733	42	Private	104647	HS-grad	9	Divorced	Other-service	Not-in-family	White	Male	0
Row30734	49	Private	67365	HS-grad	9	Married-civ...	Craft-repair	Husband	White	Male	0

A few problems... How to solve them?

- **Noise**
 - Modifications to the original records (data that is **corrupted** or **distorted**) due to technological limitations, sensor error or even human error
- **Outliers**
 - A data point that differs significantly from other observations



Data Exploration

Why?

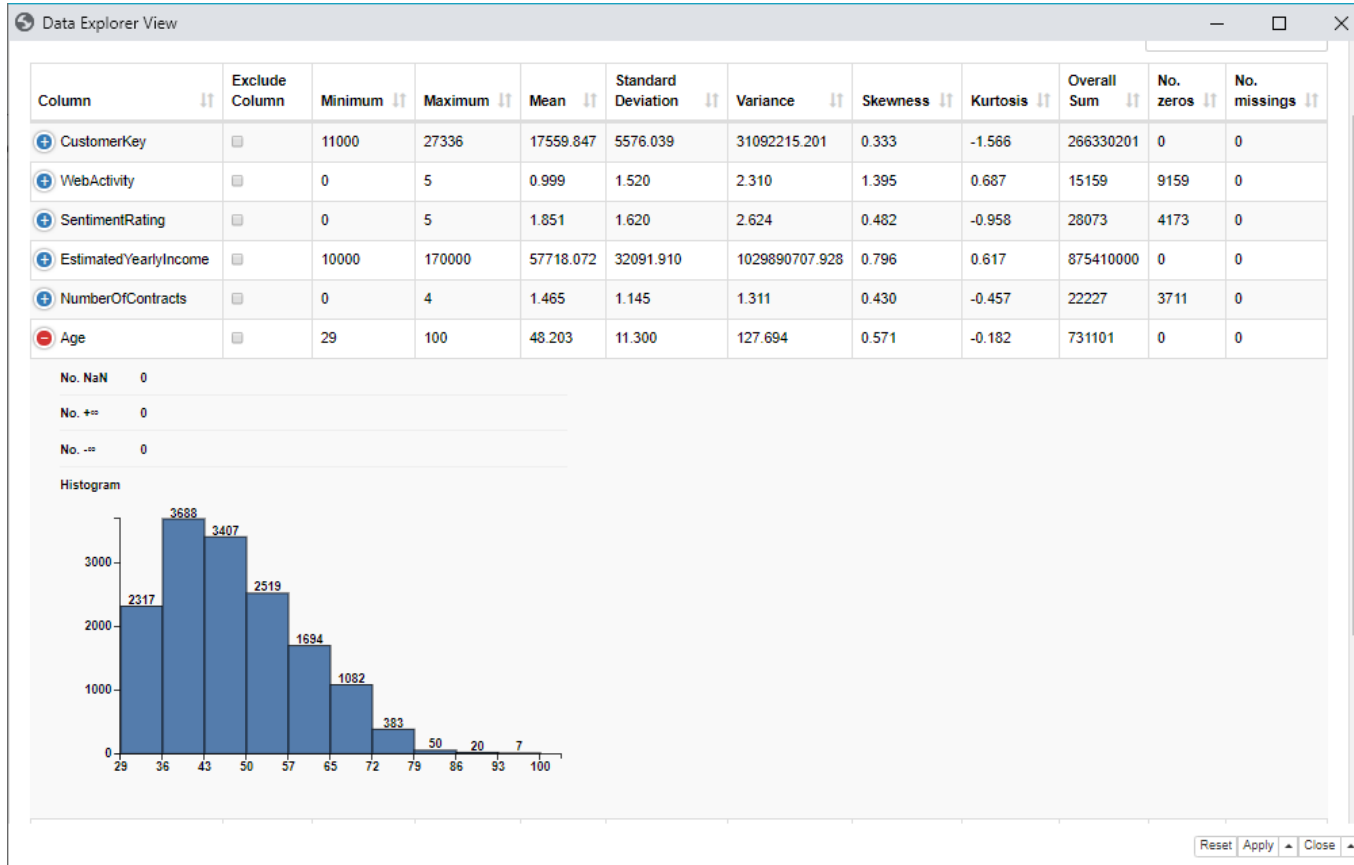
- Understand the **data** and its **characteristics**
- Evaluate its **quality**
- Find **patterns** and **relevant information**

Data Exploration

How?

- **Central Tendency**: average, mode, median...
- **Statistical dispersion**: variance, standard deviation, interquartile range...
- **Probability distribution**: Gaussian, Uniform, Exponential...
- **Correlation/Dependence**: between pairs of features, with the dependent feature...
- **Data viz**: tables, charts, boxplots, scatter plots, histograms, ...

Data Exploration



Data Exploration - Contingency Tables

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

Data Exploration - Correlation Matrix

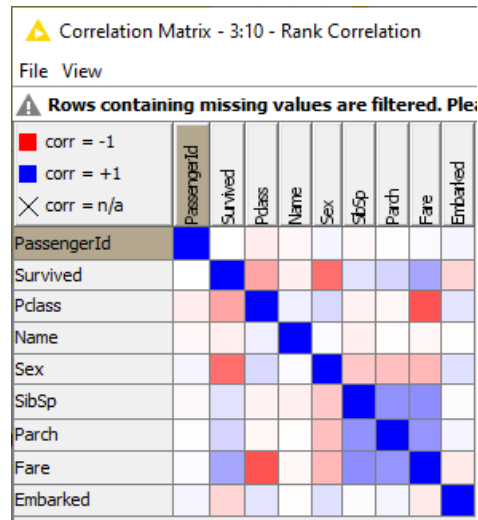
What doesn't make sense here?

Correlation measure - 3:10 - Rank Correlation

File Hilite Navigation View

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

Row ID	D Passeng...	D Survived	D Pclass	D Name	D Sex	D SibSp	D Parch	D Fare	D Embarked
Passeng...	1.0	-0.00174741...	-0.07274348...	-0.03142687...	0.0404574...	-0.02578774...	0.009305237...	0.019408214...	0.0380052514...
Survived	-0.00174741...	1.0	-0.35175083...	-0.06396644...	-0.571791...	0.113309900...	0.169963720...	0.350965386...	-0.1699515717...
Pclass	-0.07274348...	-0.35175083...	1.0	0.063095975...	0.1499334...	-0.05215363...	-0.035163461...	-0.67404431...	0.1039440255...
Name	-0.03142687...	-0.06396644...	0.06309597...	1.0	0.0211633...	-0.06164769...	-0.012894871...	-0.03634128...	-0.0135577924...
Sex	0.040457449...	-0.57179163...	0.14993345...	0.021163349...	1.0	-0.21708710...	-0.250155569...	-0.28053568...	0.1228843568...
SibSp	-0.02578774...	0.11330990...	-0.05215363...	-0.06164769...	-0.217087...	1.0	0.432451332...	0.447081227...	0.0122413820...
Parch	0.009305237...	0.16996372...	-0.03516346...	-0.01289487...	-0.250155...	0.432451332...	1.0	0.416985332...	0.0417986920...
Fare	0.019408214...	0.35096538...	-0.67404431...	-0.03634128...	-0.280535...	0.447081227...	0.416985332...	1.0	-0.082027478...
Embarked	0.038005251...	-0.16995157...	0.10394402...	-0.01355779...	0.1228843...	0.012241382...	0.041798692...	-0.08202747...	1.0



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

What are those?



Universidade do Minho
Departamento de Informática



Data Exploration - Features

Input Features/Input Vector
(independent variables)

Target/Class/Label
(dependent variable)

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

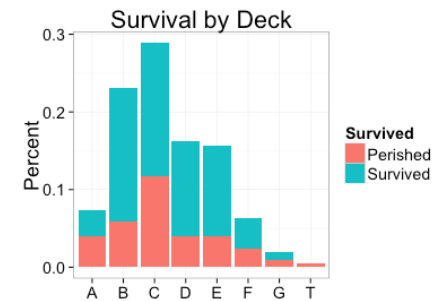
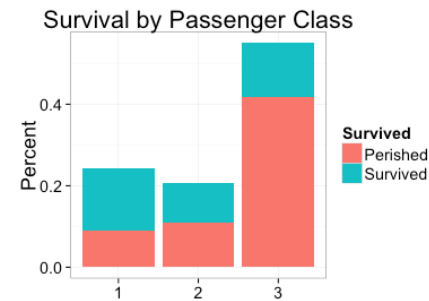
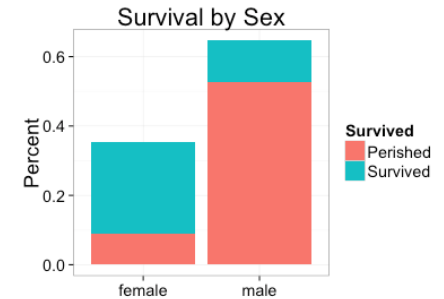
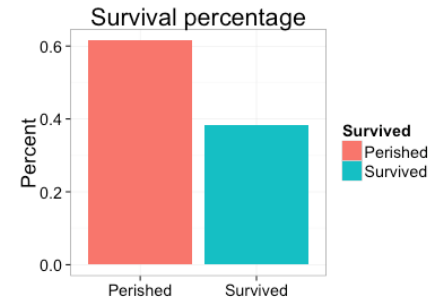
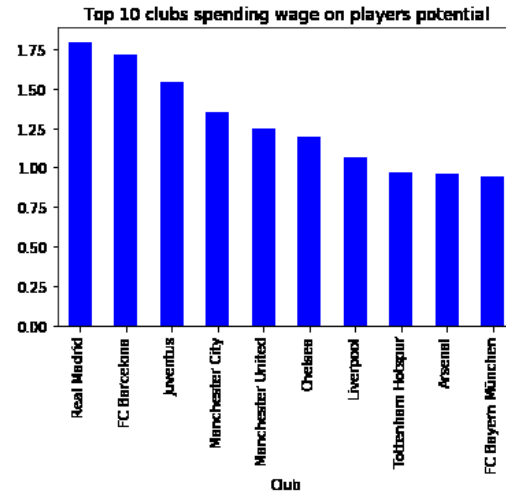
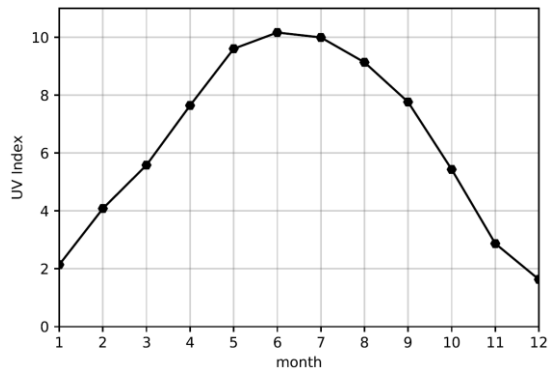
File Hilite Navigation View

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

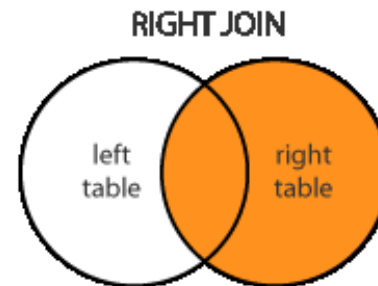
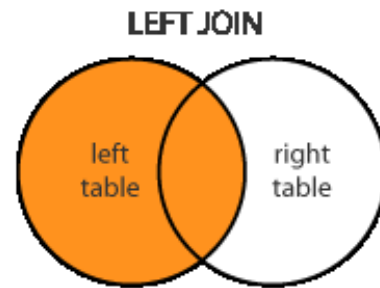
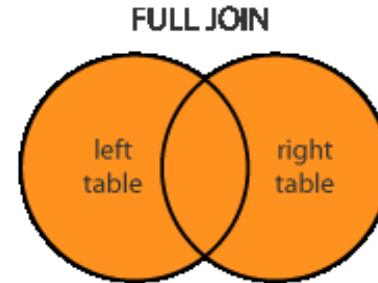
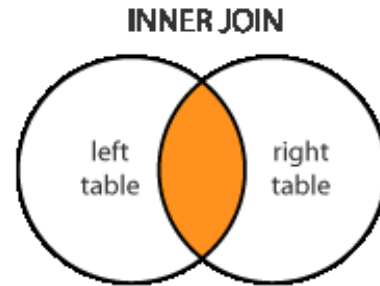
Data Viz.

← Often Neglected



Data Preparation - Basic Preparation

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



Data Preparation - Basic Preparation

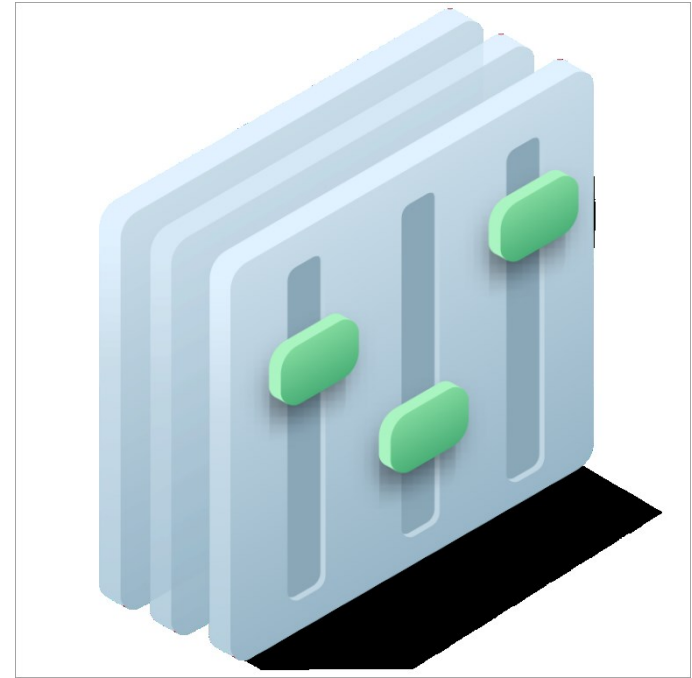
A set of basic data preparation techniques can be used:

- Union/intersection of columns;
- Concatenation
- Sorters
- Filters (column, row, nominal, rule-based, ...)
- Basic aggregations (counts, unique, mean/sum, ...)

Data Preparation - Advanced Preparation

How?

- Feature scaling
- Outlier detection
- Feature selection
- Missing Values treatment
- Nominal value discretization
- Binning
- Feature Engineering

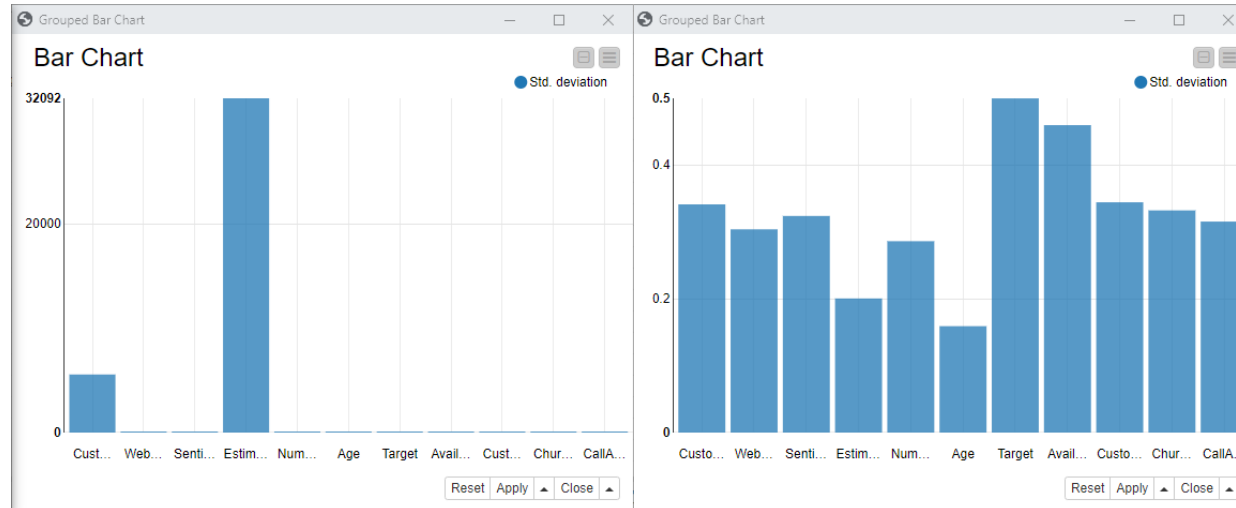


Data Preparation - Feature Scaling

1. Normalizing the range of the independent features

Rationale:

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



Data Preparation - Feature Scaling

1. Normalizing the range of the independent features

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

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

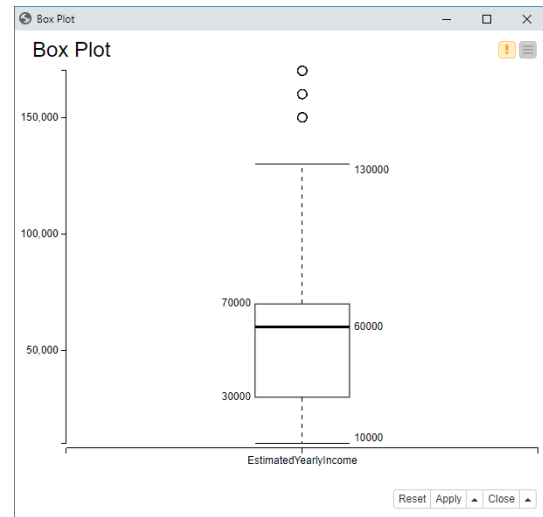
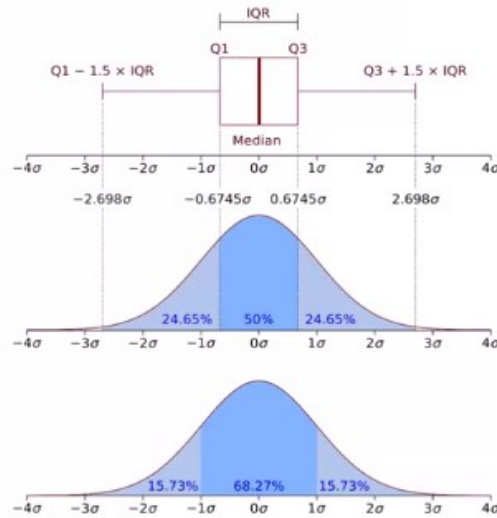
- **Standardization** (or **Z-score Normalization**): Rescaling the distribution of values so that the mean of observed values is 0 and the standard deviation is 1. Assumes observations fit a Gaussian distribution with a well-behaved mean and standard deviation, which may not always be the case.

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

Data Preparation - Outlier Detection

2. Outlier Detection:

- **Statistical-based strategy:** Z-Score, Box Plots, ...
- **Knowledge-based strategy:** Based on domain knowledge. For example, exclude everyone with a monthly salary higher than 1M € ...
- **Model-based strategy:** Using models such as one-class SVMs, isolation forests, clustering, ...



The Outlier Dilemma: **Drop** or **Cap**?

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

Data Preparation - Feature Selection

3. Feature Selection (or dimensionality reduction):

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

The removal of unimportant features:

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

What can we remove:

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

Feature Selection Methods:

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

Data Preparation - Feature Selection

3. Feature Selection (or dimensionality reduction):

- Remove a feature if the **percentage of missing values** is **higher than** a threshold;
- Use the **chi-square test** to measure the **degree of dependency** between a feature and the **target class**;
- Remove feature if **low standard deviation**;
- Remove feature if data are **highly skewed**;
- Remove features that are **highly correlated** between each other.

Data Preparation - Feature Selection

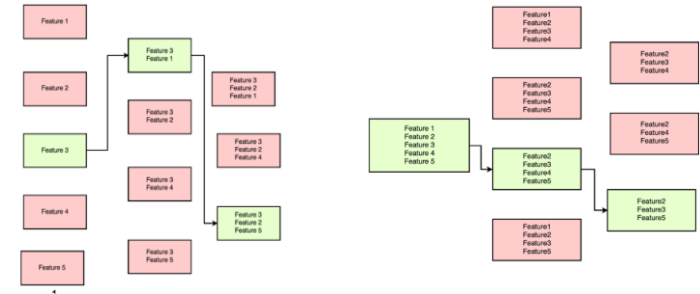
3. Feature Selection (or dimensionality reduction):

- **Principal Component Analysis (PCA)**: a technique to **reduce the dimension of the feature space**. The goal is to reduce the number of features without losing too much information. A popular application of PCA is for visualizing higher dimensional data.

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

- **Embedded Methods**: Algorithms that already have built-in feature selection methods. Lasso, for example, has their own feature selection methods. For example, if a feature’s weight is zero than it has no importance! Regularization - constrain/regularize or shrink the coefficient estimates towards zero!

- *Sequential Forward Selection* - *Sequential Backward Selection*



Data Preparation - Missing Values

4. Missing Values Treatment:

First analyze each feature in regard to the number and percentage of missing values. Then decide what to do:

- Remove
- Mean
- Interpolation
- Mask
- ...

Data Preparation - Nominal Value Discretization

5. Nominal value discretization:

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

- **One-Hot Encoding**
- **Label Encoding**
- **Binary Encoding**

Nominal value discretization:

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...

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

Row ID	Custom...	Webac...	Sentiment...	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,	Show possible values
Row0_Row86...	11000	0	Slightly Negative	2	M	M	90000	0	(39,	Available Renderers
Row1_Row1_...	11001	3	Slightly Positive	3	M	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]	0
Row3_Row86...	11003	0	Very Negative	0				1	(39,46]	0
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]	0
Row9_Row87...	11009	0	Very Negative	0	S	M	70000	1	(39,46]	0
Row10_Row1...	11010	0	Very Negative	0	S	F	70000	1	(39,46]	0
Row10_Row8...	11010	0	Very Negative	0	S	F	70000	1	(39,46]	0
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

Data Preparation - Feature Engineering

7. Feature Engineering:

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

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

2021-10-29 16h30

Data Preparation - Feature Engineering

7. Feature Engineering:

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

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

2021-10-29 16h30

We may extract new features such as:

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

DADOS e APRENDIZAGEM AUTOMÁTICA

Data Exploration and Preparation

MESTRADO (integrado) EM ENGENHARIA INFORMÁTICA