POLITECNICO
DI TORINO

Data Science Lab

Data Exploration

DataBase and Data Mining Group

Tania Cerquitelli and Elena Baralis

Data Exploration

- It includes a set of preliminary analyses that allows exploring and characterizing the data
- It is a very important step that aims to better design all the steps of the KDD pipeline
  - Quality of data has an impact on the quality of extracted knowledge
  - Understanding the input data allows the data scientist to make better decision on further and deeper analysis
  - Time saving

Dataset

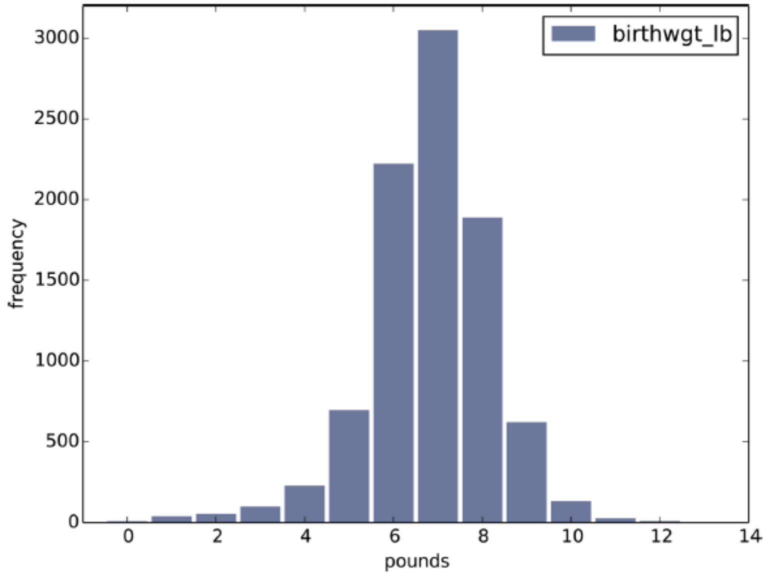
- A dataset is a collection of data
  - e.g., a tabular representation of data includes rows and columns
    - **Rows** correspond to objects, records, point, case, sample, entity, or instance
    - **columns** are the attributes
- The size of the dataset has an impact on the choice of the analyses
  - Some algorithms require considerable hardware resources when applied to large datasets, in some cases it is not possible to execute them at all.
  - There are solutions to reduce the size of the dataset preserving the completeness of the data
    - data sampling can reduce the dataset size in terms of number of rows
    - feature selection can reduce the number of attributes

Feature/attribute

- Each column of the dataset represents one attribute/feature
  - Data exploration can be performed in a univariate or multivariate fashions
- For further analysis please consider the following basic information for each attribute
  - Unit of measurement
  - Attribute Type
    - Categorical (not numerical or fixed number of possible values)
    - Conitinuous (numerical)
  - Attribute Domain
    - It is a good practice to verify if the attribute values satisfy the domain-driven constraints

Univariate analysis: Distribution

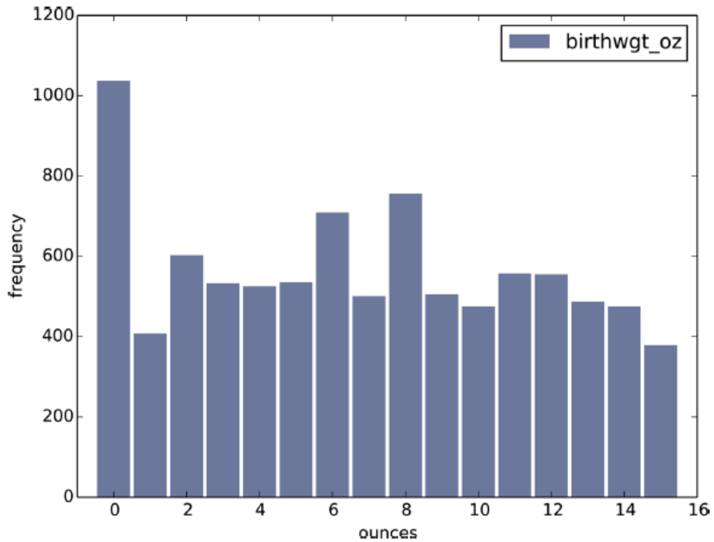
- Distribution of the attribute
  - Attribute description through the plot that shows for each attribute value how many times it appears in the dataset
  - The most common representation of a distribution is a **histogram**
    - A graph that shows the **frequency** of each value.



- Example of a histogram that shows the distribution of the pound part of birth weight
  - The distribution is approximately bell-shaped, which is the shape of the **normal** distribution

Univariate analysis: Distribution

- Distribution of the attribute
  - Attribute description through the plot that shows for each value how many times it appears in the dataset.
  - The most common representation of a distribution is a **histogram**
    - A graph that shows the **frequency** of each value.



- Example of a histogram that shows the distribution of the ounces part of birth weight
  - This distribution is not **uniform**
    - 0 is more common than the other values,
    - 1 and 15 are less common, probably because respondents round off birth weights that are close to an integer value.

Characterizing Distributions

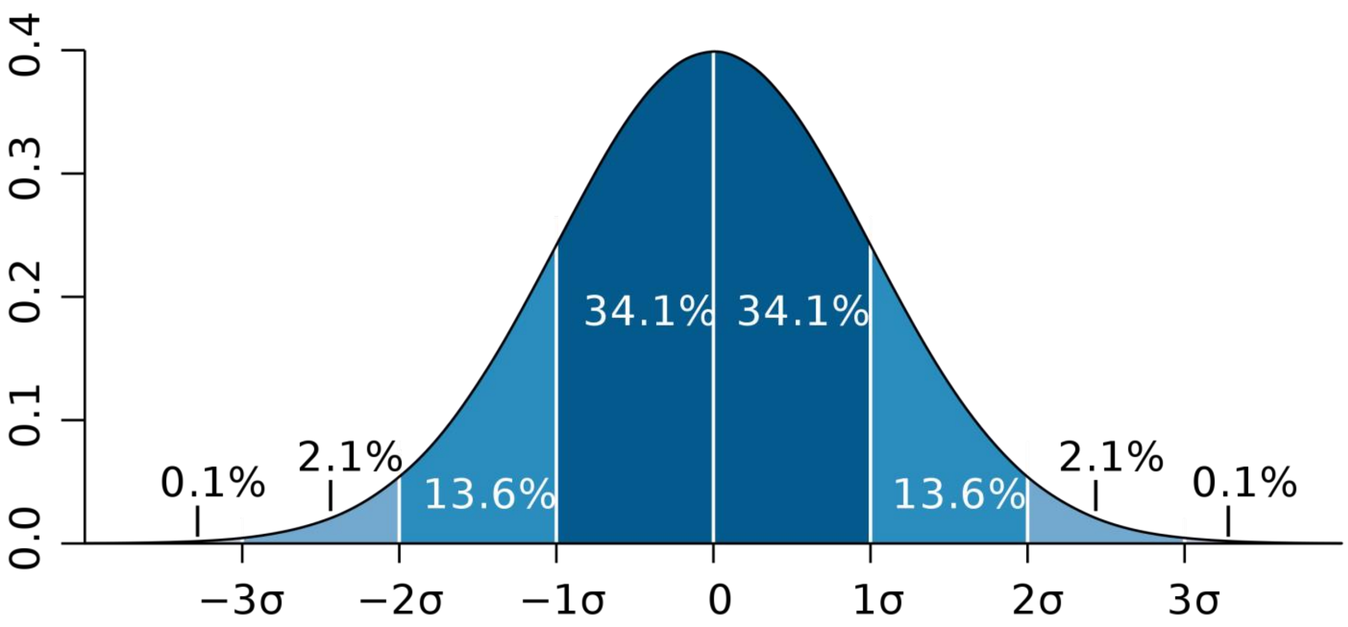
- Given a distribution
  - Minimum value
  - Maximum value
  - Mean value
  - Number of samples
  - Standard deviation
  - Mathematical functions
    - **A probability distribution** that provides the probabilities of occurrence of different possible values in a dataset
      - Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean
      - Kurtoisis is a measure of the "tailedness" of the probability distribution of a real-valued random variable
  - …

Univariate analysis: Distribution

- Probability distributions are usually classified as
  - **Continuous probability distribution**
    - The set of possible outcomes can assume values in a continuous range
    - It is typically described by a probability density functions (modelling)
  - **Discrete probability distribution**
    - Characterized by a discrete list of the probabilities of the outcomes
    - It is typically described by a probability mass function (description)

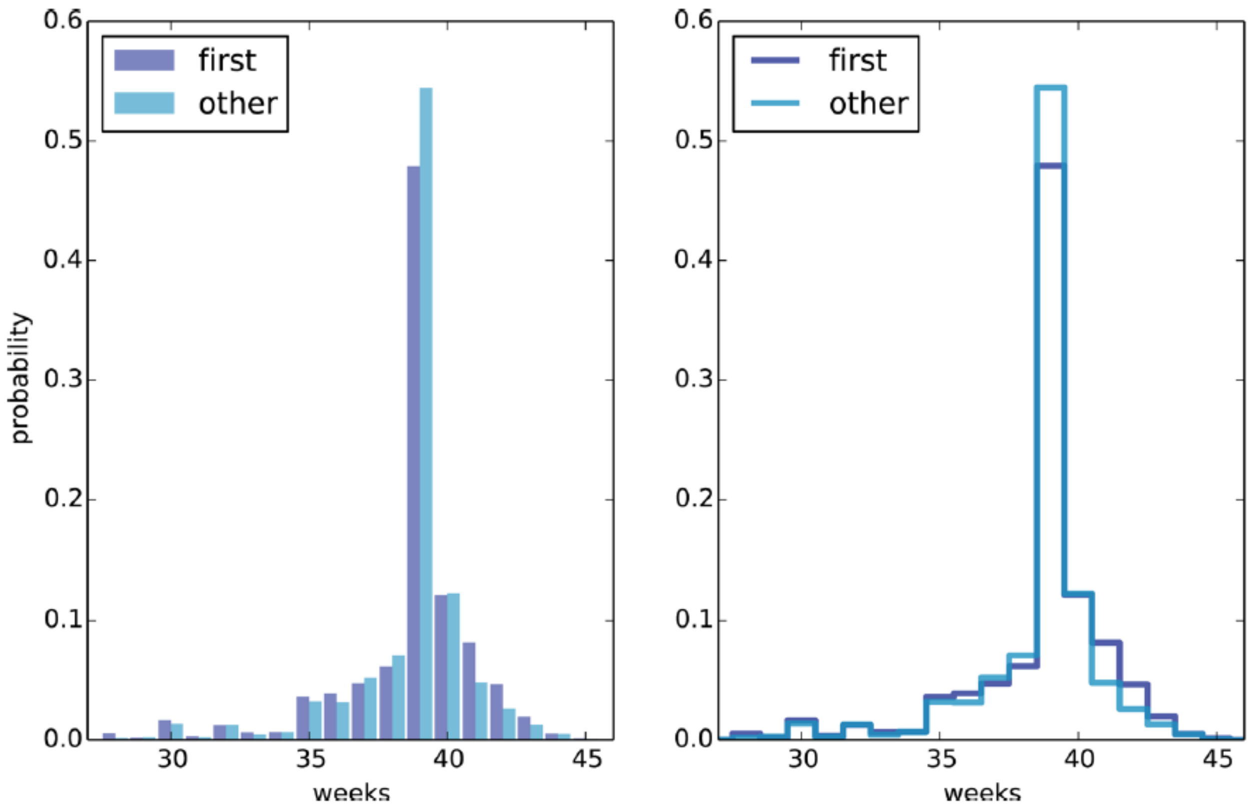
Continuous probability

- Example of the **probability density function** (PDF) of the normal distribution



Discrete probability

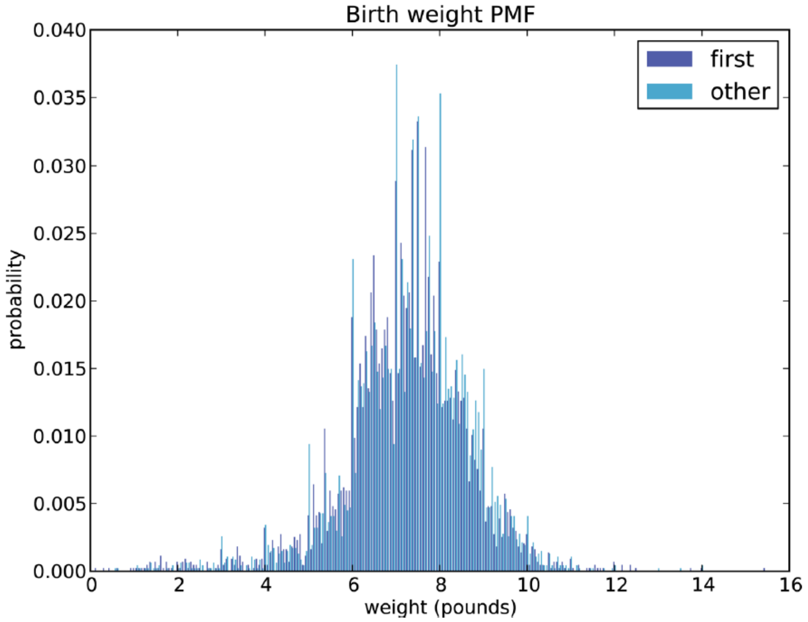
- Example of the **probability mass function (PMF)** of a **discrete** probability distribution



PMF of pregnancy lengths for first babies and others, using bar graphs (left) and step functions (right)

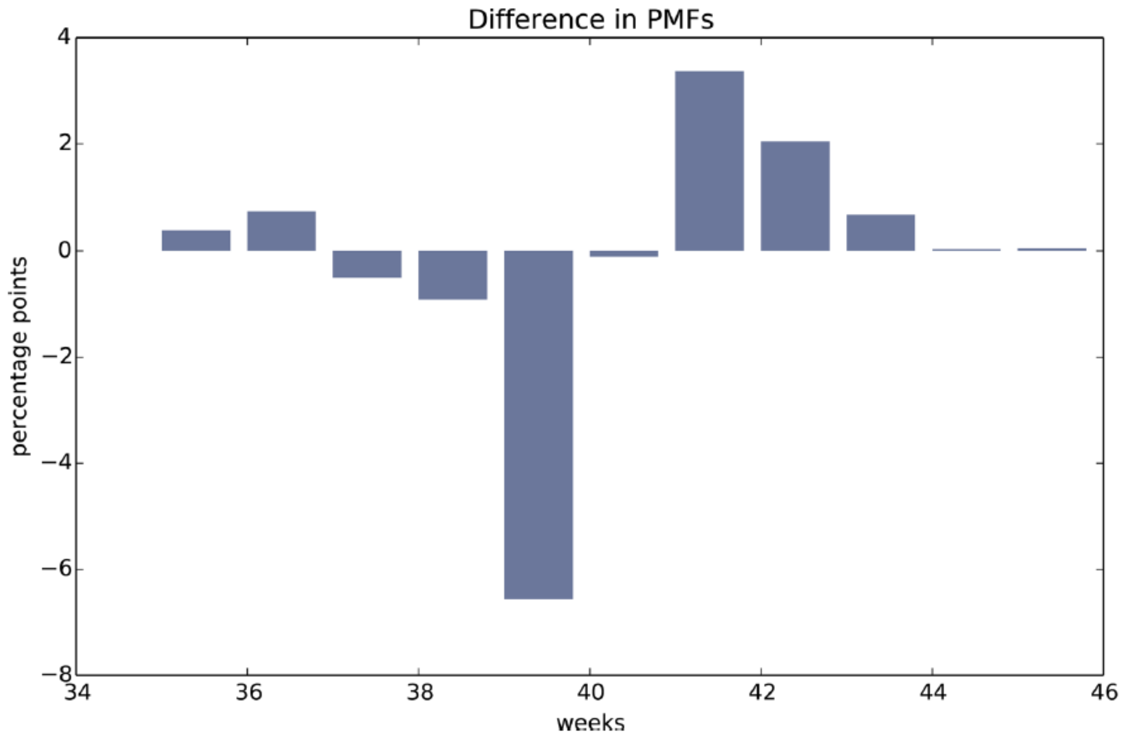
PMF limitations

- In cases of a lot of samples to show it is very hard to read information from a PMF plot
- Possible solutions include
  - calculating the cumulative distribution function (CDF)
  - Showing the difference of distributions



Difference of distributions

- Example of difference between two PMFs (probability mass functions)



Difference, in percentage points, by week

Cumulative Distribution Function (CDF)

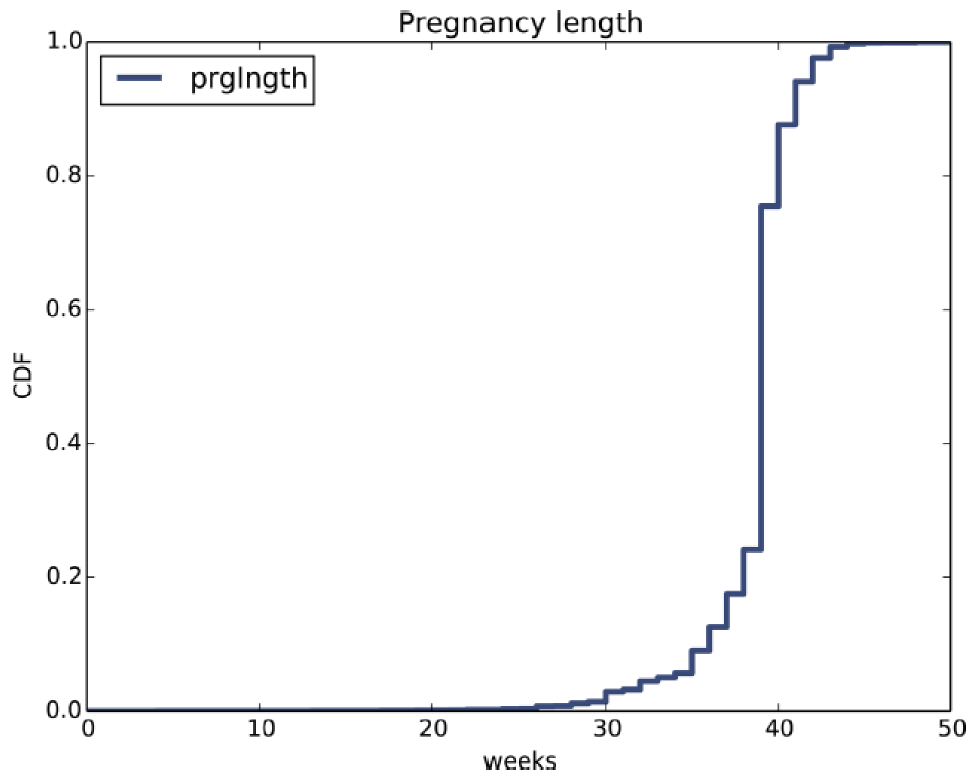
- The CDF is the function that maps a value to its percentile rank.
  - The CDF is a function of *x*, where *x* is any value that might appear in the distribution.
- The CDF also provides a visual representation of the shape of the distribution
  - Common values appear as steep or vertical sections of the CDF
- To evaluate CDF(x) for a particular value of x, we compute the fraction of values in the distribution less than or equal to x.

$$Fx(x) = P(X \leq x)$$

$$P(a < X \leq b) = Fx(b) - Fx(a)$$

Cumulative Distribution Function

- Example of CDF



E.g. about 10% of pregnancies are shorter than 36 weeks, and about 90% are shorter than 41 weeks.

The CDF also provides a visual representation of the shape of the distribution. Common values appear as steep or vertical sections of the CDF; in this example, the mode at 39 weeks is apparent.

There are few values below 30 weeks, so the CDF in this range is flat.

POLITECNICO
DI TORINO

Data Science Lab

Data Exploration: Outlier detection

DataBase and Data Mining Group

Tania Cerquitelli and Elena Baralis

Outliers

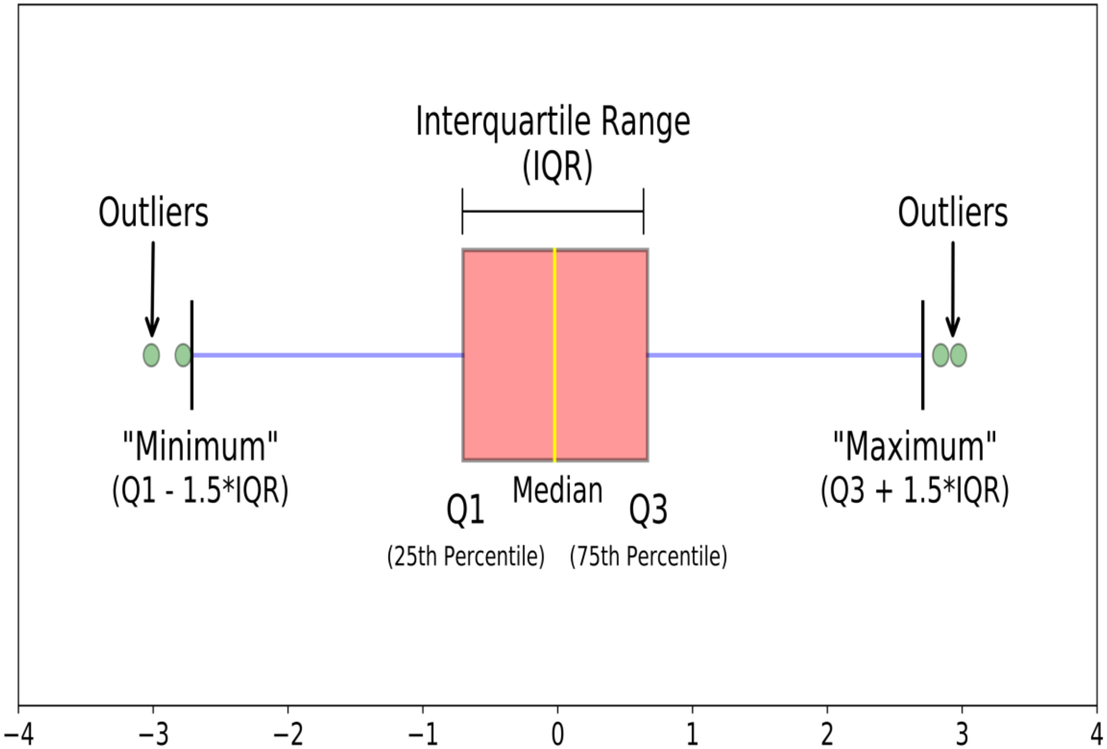
- **Outliers** are extreme values that might be
  - errors in measurement and recording
  - accurate reports of rare events
- The best way to handle outliers depends on “**domain knowledge**”
  - Information about where the data come from and what they mean
  - it depends on what analysis you are planning to address

Outliers

- Outliers can be detected through
  - Univariate analysis
    - Boxplot
    - Percentiles
    - Histograms
    - GESD
    - …
  - Multivariate analysis
    - DBSCAN
    - …
  - More specific techniques

Outliers Detection
Boxplot

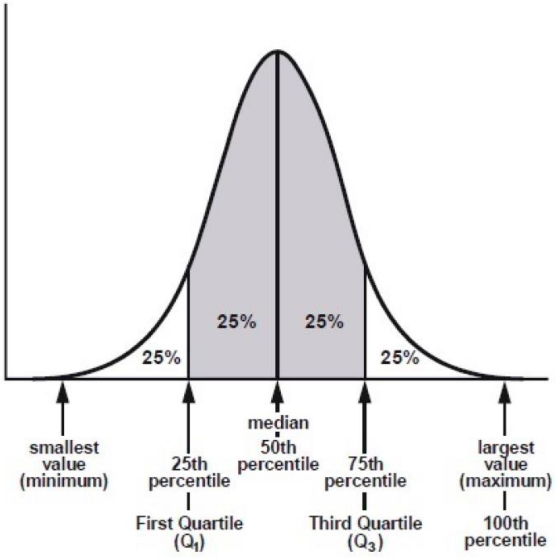
Boxplots are a standardized way of displaying the **distribution** of data based on a **five** number summary (“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”).



- **median** (Q2/50th Percentile): the middle value of the attribute.
- **first quartile** (Q1/25th Percentile): the middle number between the smallest number (not the “minimum”) and the median of the attribute.
- **third quartile** (Q3/75th Percentile): the middle value between the median and the highest value (not the “maximum”) of the attribute.
- **interquartile range** (IQR): 25th to the 75th percentile.
- **whiskers** (shown in blue)
- **outliers** (shown as green circles)

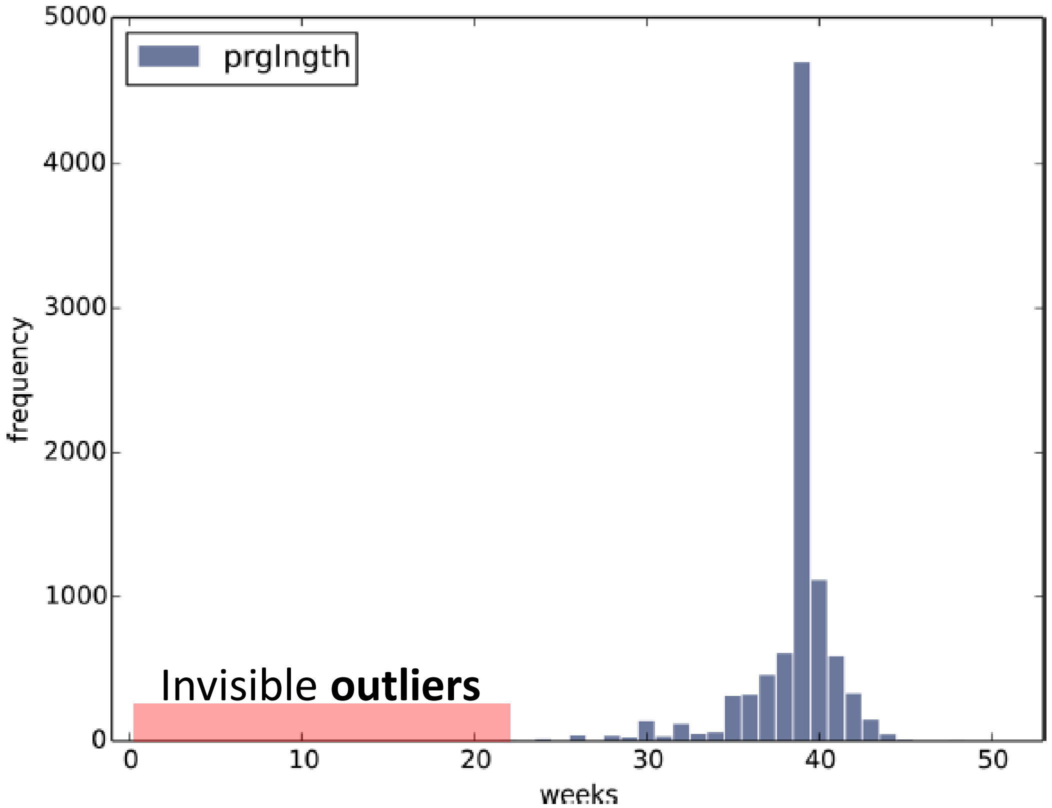
Percentiles

- **Percentile** indicates the value below which a given percentage of observations in a group of observations falls
- Representing a feature/attribute/variable through percentiles allow representing the entire distribution
  - Selecting the four percentiles
  - Selecting the ten percentiles
    - selected **10 percentiles**: 10, 20, 30, 40, 50, 60, 70, 80, 90, 99
- Outliers
  - e.g., values in the **first** and **last** percentile of the distribution



Histogram

- Example of a histogram that shows the distribution of the of pregnancy length in weeks



- The most common value of pregnancy length is 39 weeks. The left tail is longer than the right;
  - early babies are common, but need to establish the threshold where the value is a rare case or when it is an error
  - pregnancies seldom go past 43 weeks, and doctors often intervene if they do.

Outliers Detection
GESD

Generalized **E**xtreme **S**tudentized **D**eviate (GESD)

- It is used to detect one or more **outliers** in a **univariate** data set that follows an approximately normal distribution

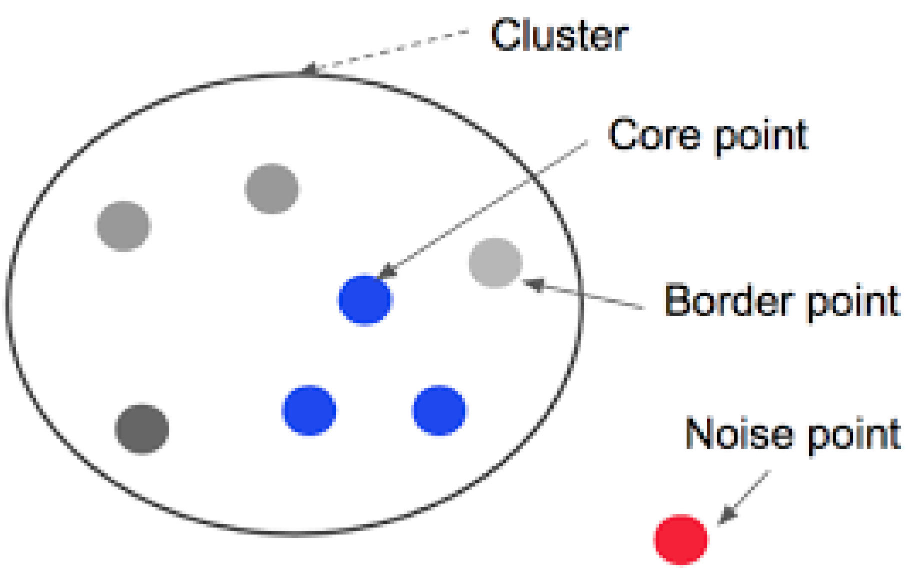
Given the upper bound, $\boldsymbol{r}$, the generalized ESD test essentially performs r separate tests: a test for one outlier, a test for two outlier, and so on up to r outliers.

$$R_i = \frac{\max_{\mathrm{i}} |x_i - \bar{x}|}{s}$$

With $\bar{x}$ and s denoting the sample mean and sample standard deviation, respectively

Outliers Detection
DBSCAN

DBSCAN is a **density-based clustering** non-parametric algorithm: given a set of points in some space, it **groups together** points that are closely packed together (points with many nearby neighbors), marking as **outliers** points that lie alone in low-density regions (whose nearest neighbors are too far away)



POLITECNICO
DI TORINO

Data Science Lab

Data Exploration: Correlation analysis

DataBase and Data Mining Group

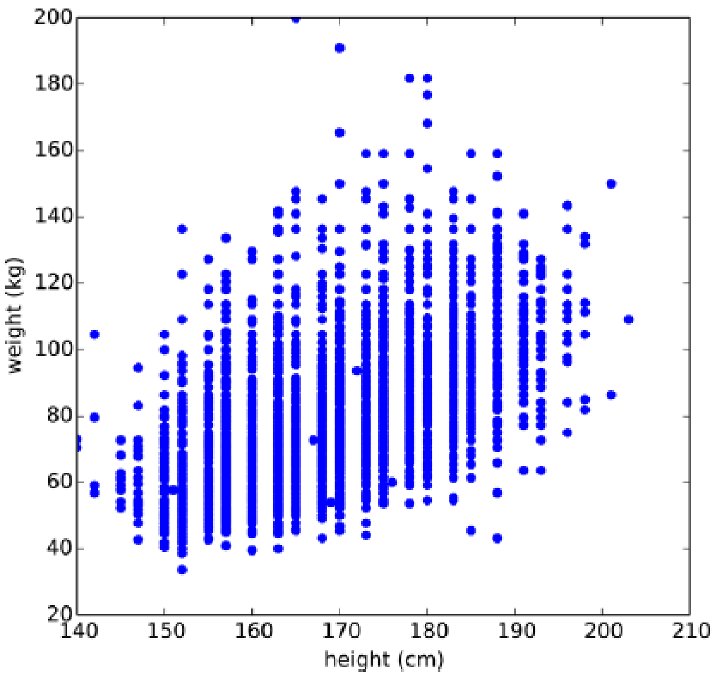
Tania Cerquitelli and Elena Baralis

Characterizing multivariate dataset

- A dataset usually includes different features/attributes
  - The description of the main relations between attributes assumes a key role
- Statistical descriptions includes
  - Scatter plot
  - Scatter plot percentiles
  - Correlation analysis
  - …

Scatter Plot

- The simplest way to check for a relationship between two variables is a **scatter plot**
  - e.g. plot the correlation between height and weight



In real case people who are taller tend to be heavier

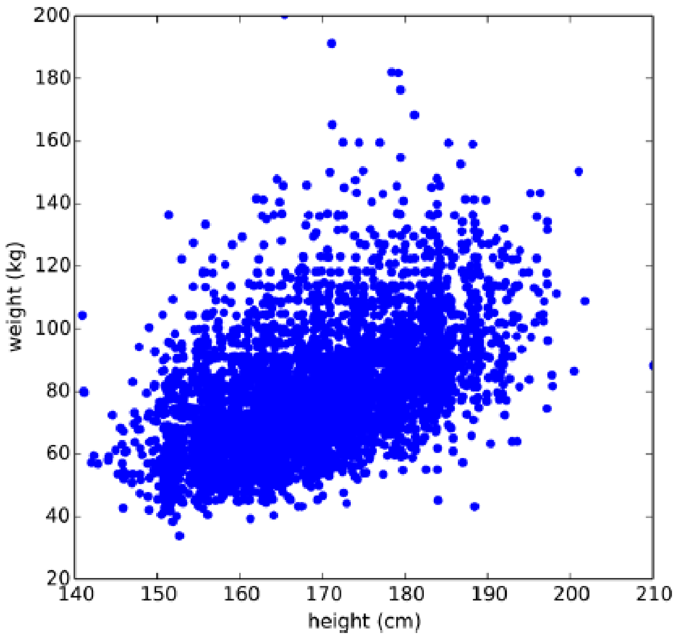
- the coordinates do not faithfully represent reality due to rounding and subsequent conversion done in this example dataset (inch to cm)

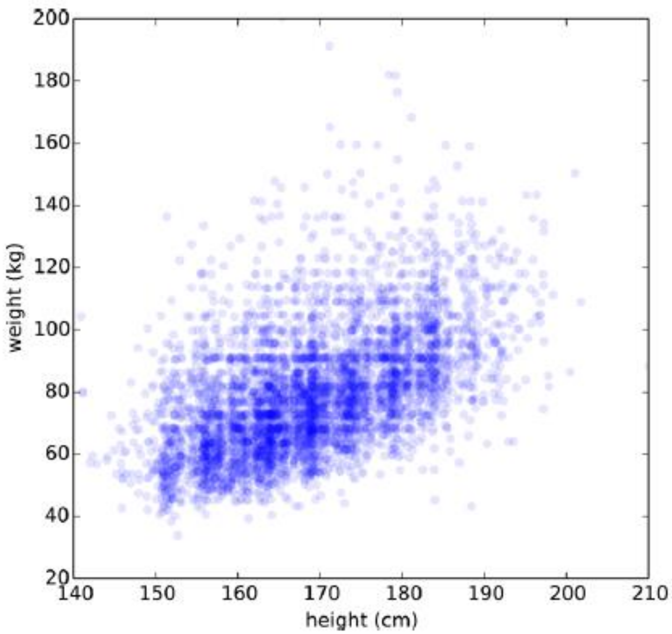
Scatter Plot

A possible solution is to **jittering** the data, which means adding random noise to reverse the effect of rounding off

Then adding **alpha** parameter to each point in order to retrieve density information

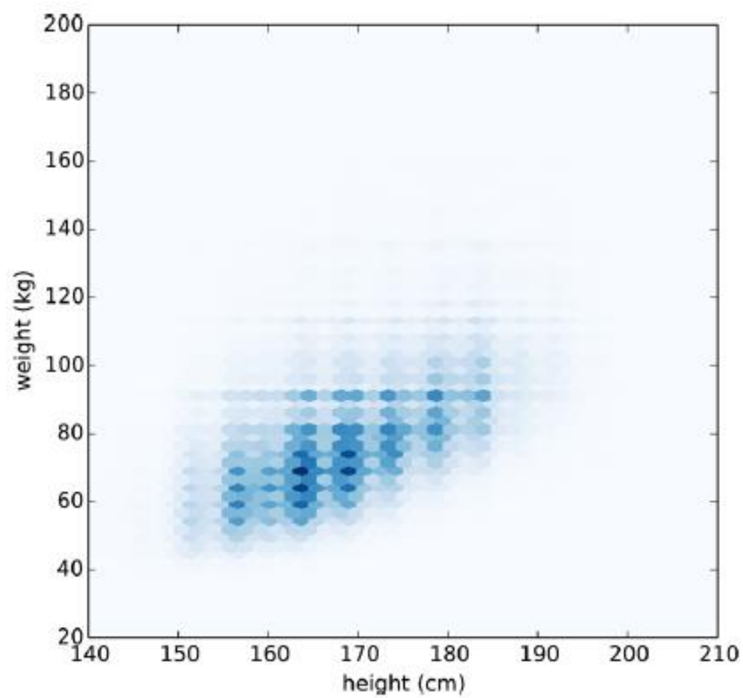
Darker zones corresponds to higher density zones





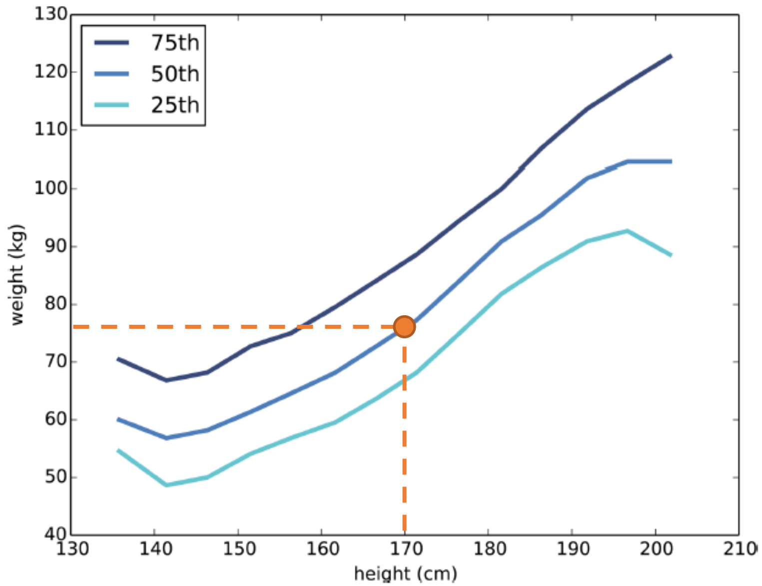
Scatter Plot

- Scatter plot is converted into **hexbin plot**
  - The hexbin plot uses **hexagonal bins** that are colored according to how many data points fall in it
- The main issue of the scatter plot is the limitation of representing huge quantity of points



Scatter Plot Percentiles

- This technique includes to bin one variable and plot percentiles of the other
  - It is an alternative to scatter plot



Line plot where the 25th, 50th and 75th percentiles are shown

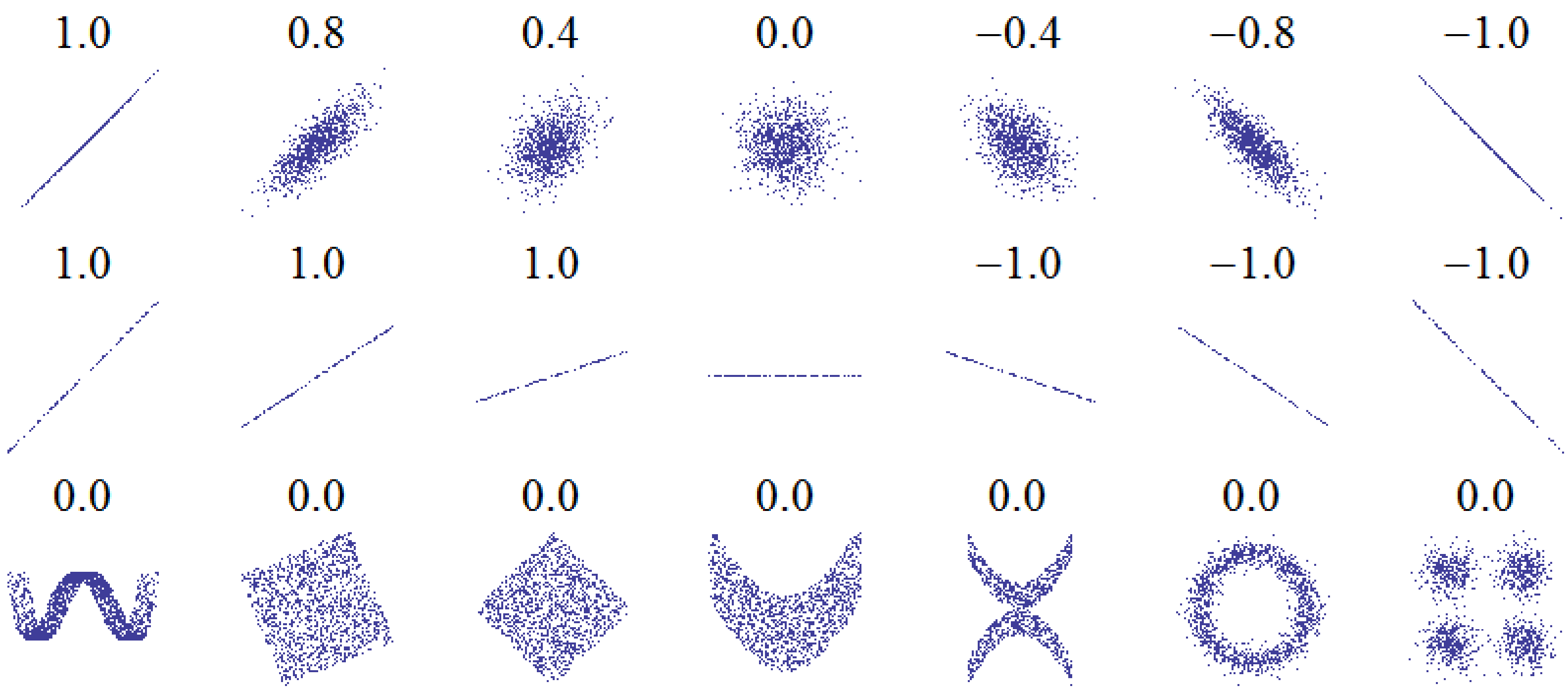
e.g. orange intersection means that 50% of people 170 cm tall weigh less than 75kg

Correlation

- A **correlation** is a statistic intended to quantify the strength of the relationship between two variables.
- Possible way to compute correlations are:
  - Covariance
    - In order to compare two variables, they must have the same unit of measurement
    - alternatively, they must be **normalised**
  - Pearson
    - Solves the problem of normalization
    - Only detect linear correlation
  - Spearman

Nonlinear Correlation

- There are some types of nonlinear correlations that Pearson's correlation can't detect. Here is an example

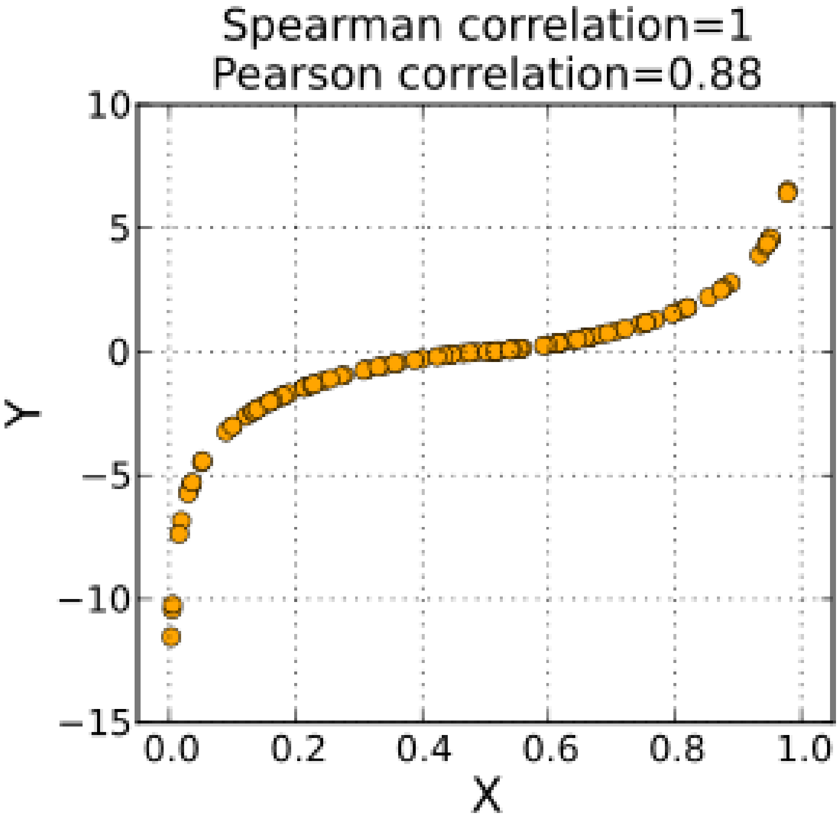


Spearman's Rank Correlation

- In some cases the Spearman index allows to find a correlation when the Pearson index returns a value close to 0
- The Spearman index or Spearman's rank uses the variables rank instead of Pearson that uses the variables themselves
- Can find monotonic function correlation between two variables

Spearman's Rank Correlation

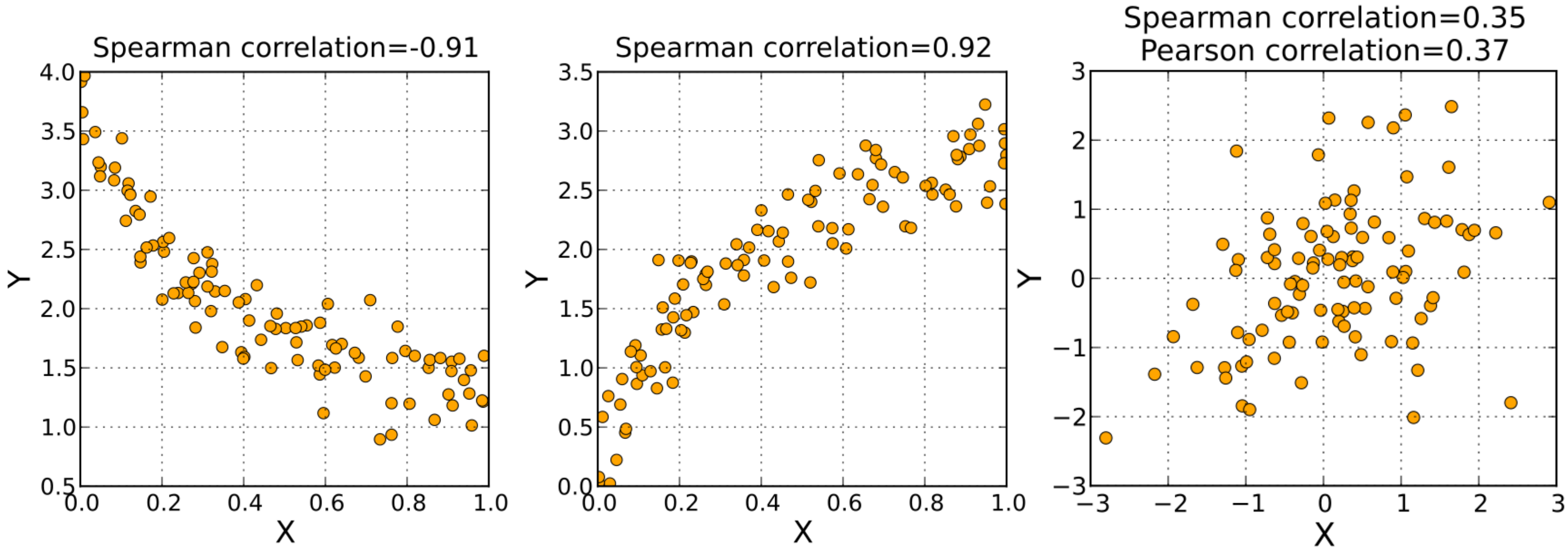
- Example of a monotonic correlation, Spearman index is 1 while Pearson is only 0.88 because the correlation in non linear



assesses monotonic relationships (whether linear or not)

Spearman's Rank Correlation

- Spearman's rank is
  - 1 when two variables are correlated by an increasing monotonic function
  - −1 if the function is decreasing monotonic
  - 0 if there isn't a monotonic function correlation



POLITECNICO
DI TORINO

Data Science Lab

Feature Engineering

DataBase and Data Mining Group

Tania Cerquitelli and Elena Baralis

Feature Engineering

- *Feature engineering* is the act of extracting features from raw data and transforming them into formats that are suitable for the machine learning model
- A *feature* is a numeric representation of an aspect of raw data

Feature Engineering

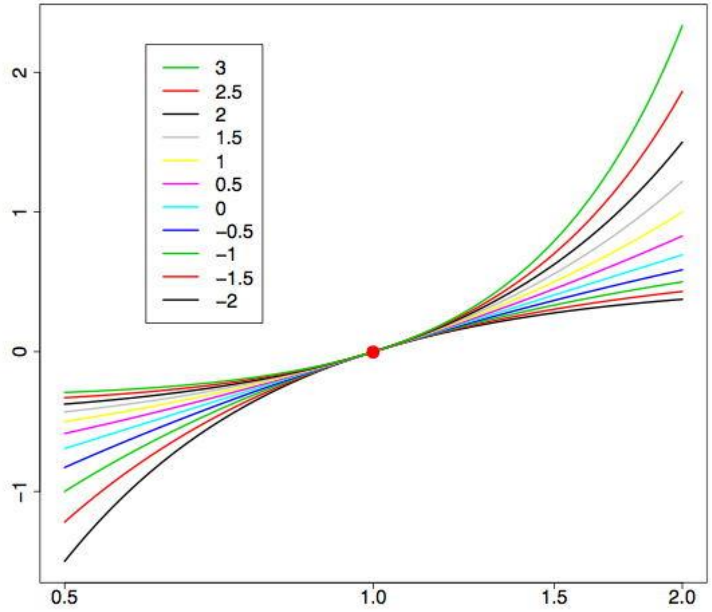
- Accordingly to the type of data under analysis different feature engineering techniques are needed
  - Structured
    - Numerical data, Categorical data
  - Unstructured
    - Text, Images, Signals
  - Mixed
- Basic types of feature engineering techniques include
  - Normalization
  - Discretization
  - Binarization
  - Data transformation

Data transformation

- Data transformation is the process of converting data from one format to another
- Why transforming data
  - Non numerical data is difficult to analyze if not transformed into numerical
  - To fit simpler models (e.g. normal distribution)
  - To better visualize the data (e.g. transform linear scale to logarithmic scale in audio context)
  - …

Power Transform: Box-Cox

- **Power transform** try to fit the distribution to the Normal distribution in order to achieve better results in further analysis
  - Some state-of-art techniques better perform with specific data distributions
- The Box-cox transformation changes the distribution of the variable so that the variance is no longer dependent on the mean



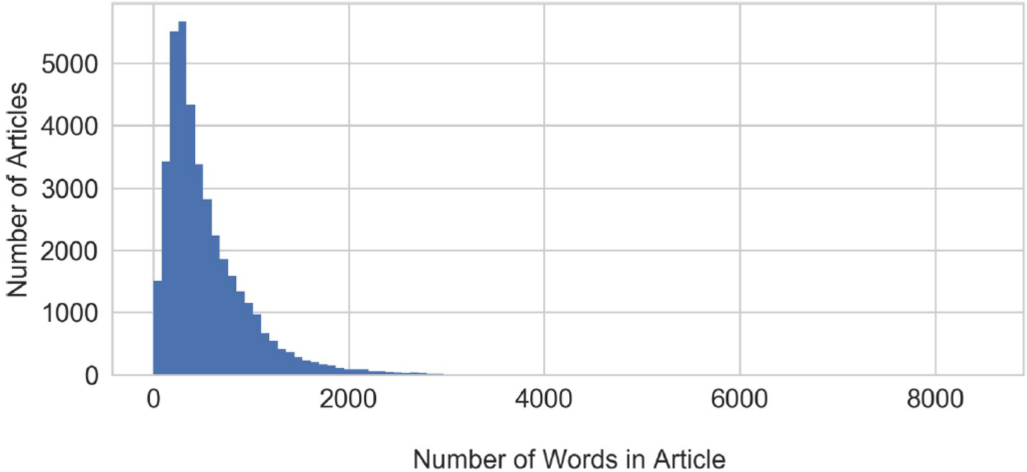
Box-cox Formula:

$$\tilde{x} = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln(x) & \text{if } \lambda = 0. \end{cases}$$

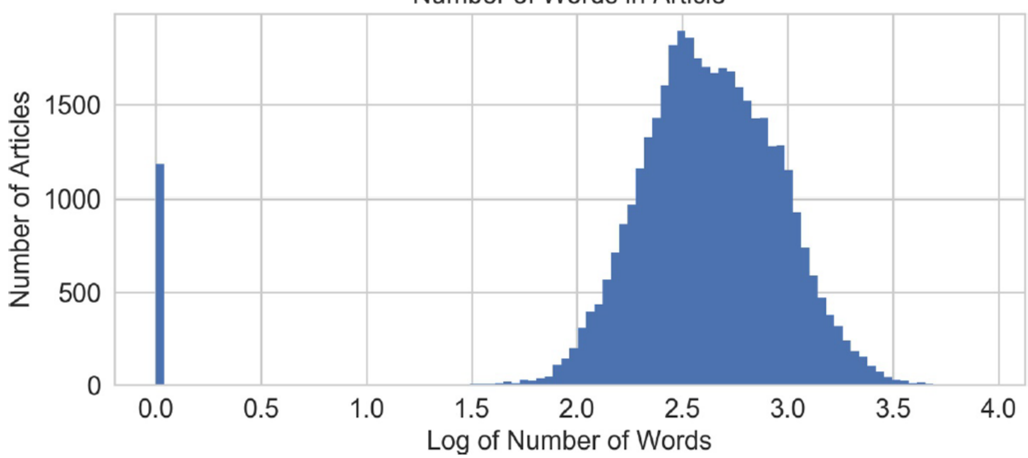
e.g.
λ = 0 corresponds to Log Transformation

Power Transform: Log Transformation

- Example of Log Transformation



Original data distribution. All values are close to 0



Data is transformed with Log Transformation in order to fit the Normal Distribution

Categorical Variables

- A *categorical variable*, as the name suggests, is used to represent categories or labels.
  - e.g., cities, season, etc…
- Some encoding methods are required to use categorial variables with some data analytics algorithms:
  - One-Hot Encoding
  - Dummy Coding
  - Effect Coding

One-Hot Encoding

- One-Hot Encoding use a group of bits.
- Each bit represents a possible category.
- If the variable cannot belong to multiple categories at once, then only one bit in the group can be "on."
  - Example: the attribute city assumes only 3 values
  - The one-hot encoding representation is reported below

	e1	e2	e3
San Francisco	1	0	0
New York	0	1	0
Seattle	0	0	1

Dummy Coding

- The problem with one-hot encoding is that it allows for k degrees of freedom, but the variable itself needs only $k-1$.
- Dummy Coding encodes the effect of each category relative to the reference category encoded with zeroes (Seattle)
- Example of a dummy coding

	e1	e2
San Francisco	1	0
New York	0	1
Seattle	0	0

Effect Coding

- It is similar to dummy coding, with the difference that the reference category is now represented by the vector of all −1's
- Example of an effect coding

	e1	e2
San Francisco	1	0
New York	0	1
Seattle	-1	-1

Pro−Cons

	PRO	CONS
One Hot	• each feature clearly corresponds to a category • missing data can be encoded as the all zeros Vector • output should be the overall mean of the target variable	• Redundant
Dummy	• Not Redundant	• cannot easily handle missing data, since the all-zeros vector is already mapped to the reference category.
Effect	• using a different code for the reference Category(-1)	• the vector of all –1's is a dense vector, which is expensive for both storage and computation

Feature engineering vs feature reduction

- Feature engineering also means feature reduction
  - Why reduce the number of features?
    - Reduce overfitting
    - Better performance on reduced data
    - Improve the generalization of models.
    - Gain a better understanding of the features and their relationship to the response variables.
- Feature reduction
  - **Dimensionality reduction using Singular Value Decomposition**
    - it can work with sparse matrices efficiently
  - **Principal component analysis (PCA)**
    - subtract the mean sample from each row of the data matrix
    - perform SVD on the resulting matrix.
  - **Linear Discriminant Analysis (LDA)**

Feature Selection

- Feature engineering might also require the feature selection step
- Traditional approaches can be used
  - Recursive feature elimination
  - Analysis of variance (ANOVA)
  - Exploiting feature importance of interpretable models
  - Automatic feature selection
  - …

Feature Selection: Recursive feature elimination

- **Recursive feature elimination**
  - assigns weights to features
    - e.g., the coefficients of a linear model
  - selects features by recursively considering smaller and smaller sets of features
  - First, the estimator is trained on the initial set of features and the importance of each feature is obtained
  - The least important features are pruned from current set of features
  - That procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached

- **Analysis of variance (ANOVA)**
  - It is a statistical technique that allows to compare two or more groups of data by comparing the *variance* within these groups with the *variance* between groups
    - e.g. **Fisher-Snedecor** analysis is used to compare the variance between two variable
  - Assigns ranks to the features that help to select the most important ones

Feature Selection

- Exploiting feature importance of interpretable models
  - Some models give the information about feature importance
  - e.g. Decision tree, Linear Regression
- Automatic feature selection
  - The components of this decomposition techniques allow to identify which are the most important features/components in the data
  - e.g. PCA, SVD

POLITECNICO
DI TORINO

Data Science Lab

Data Visualization

DataBase and Data Mining Group

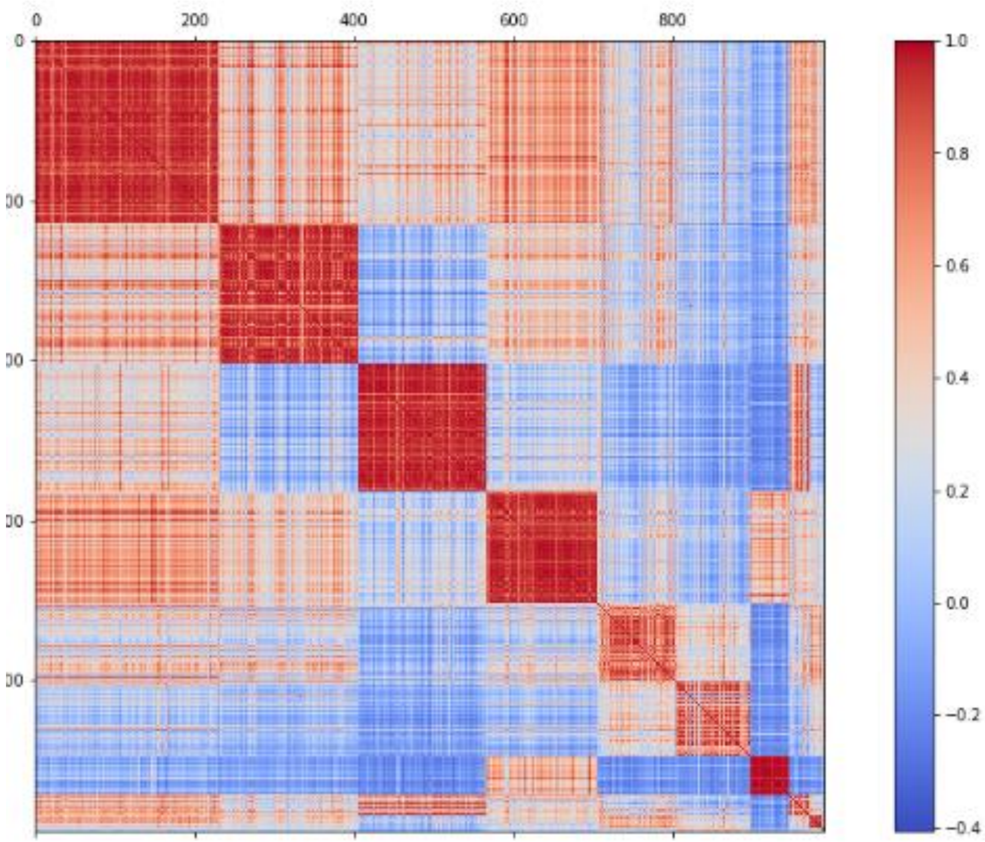
Tania Cerquitelli and Elena Baralis

Data Visualization

- It is important to visualize your data when possible
  - To explore the raw input data
  - To analyze your output
- Choosing the correct visualization method is not trivial
  - Different kind of analytics tasks require proper visualization techniques

Visualization: heatmap

- Correlation matrix can be visualized through **heatmaps** to represents the correlation between two variables



e.g. correlation between documents.

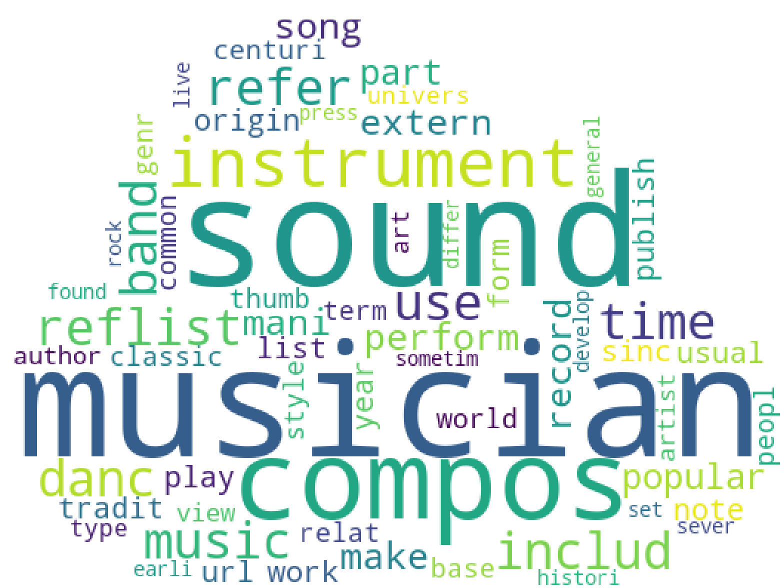
x and y axes represents the documents

documents were previously clustered by topics

Red dots represent high correlation, while blue dots low correlation.

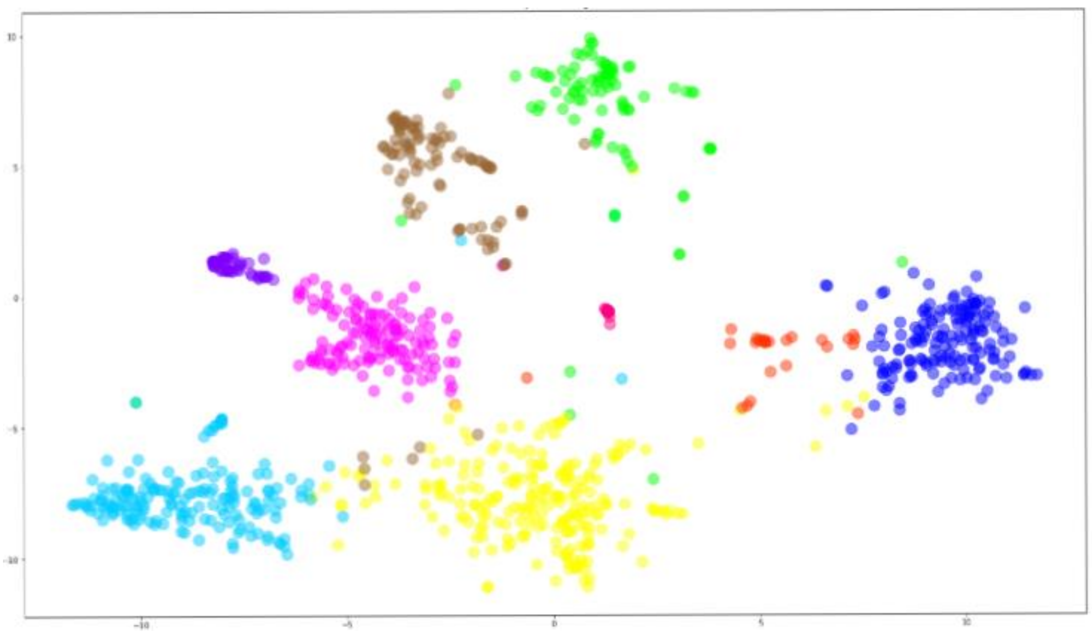
Visualization: word cloud

- In the context of text mining the **word cloud** can easily represents the topics of a group of similar documents
- Each word cloud contains the most important words characterizing a topic



Visualization: t-sne

- In some cases it is useful to reduce the size of attributes to show information in plots
- e.g **t-distributed stochastic neighbor embedding (t-SNE)**

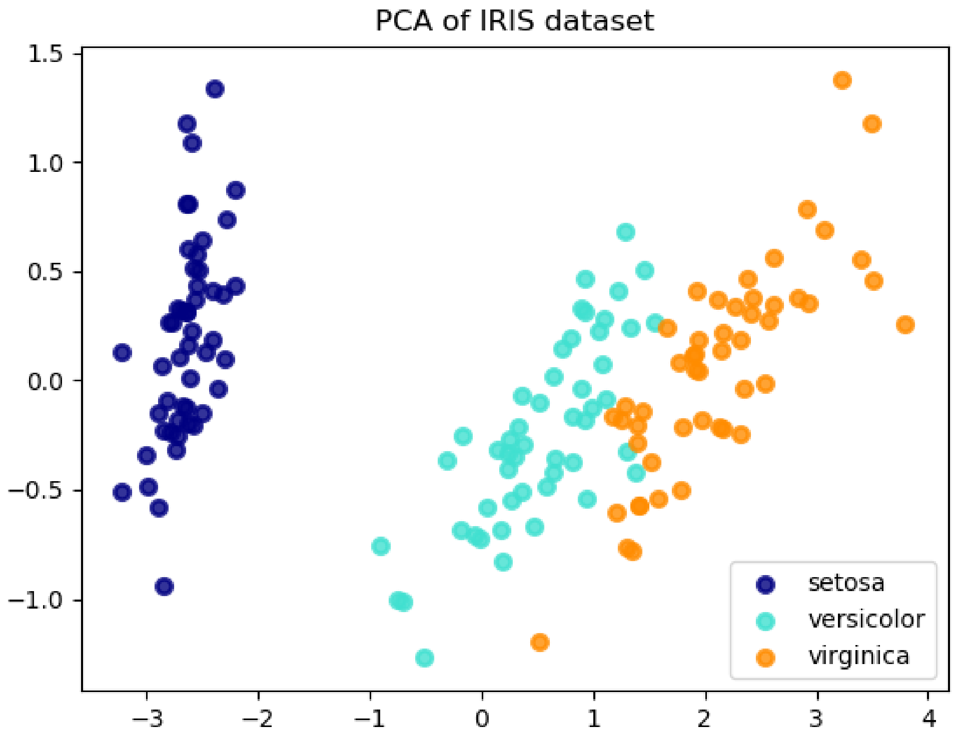


In the example T-sne shows text document in bidimensional space. Each color corresponds to a cluster label

WARNING: using the t-sne as dimensionality reduction technique in ML pipelines is not suggested since it does not preserve the information of your data.

Visualization: PCA

- In some cases it is useful to reduce the size of attributes to show information in plots
- e.g **Principal component analysis (PCA)**



In the example, the **Iris** dataset wad reduced with **PCA** in two features and represented in scatter plot. Each color corresponds to the original labels. See how the 3 category are separated in bidimensional space

Credits

- Think Stats, Allen B. Downey - Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists