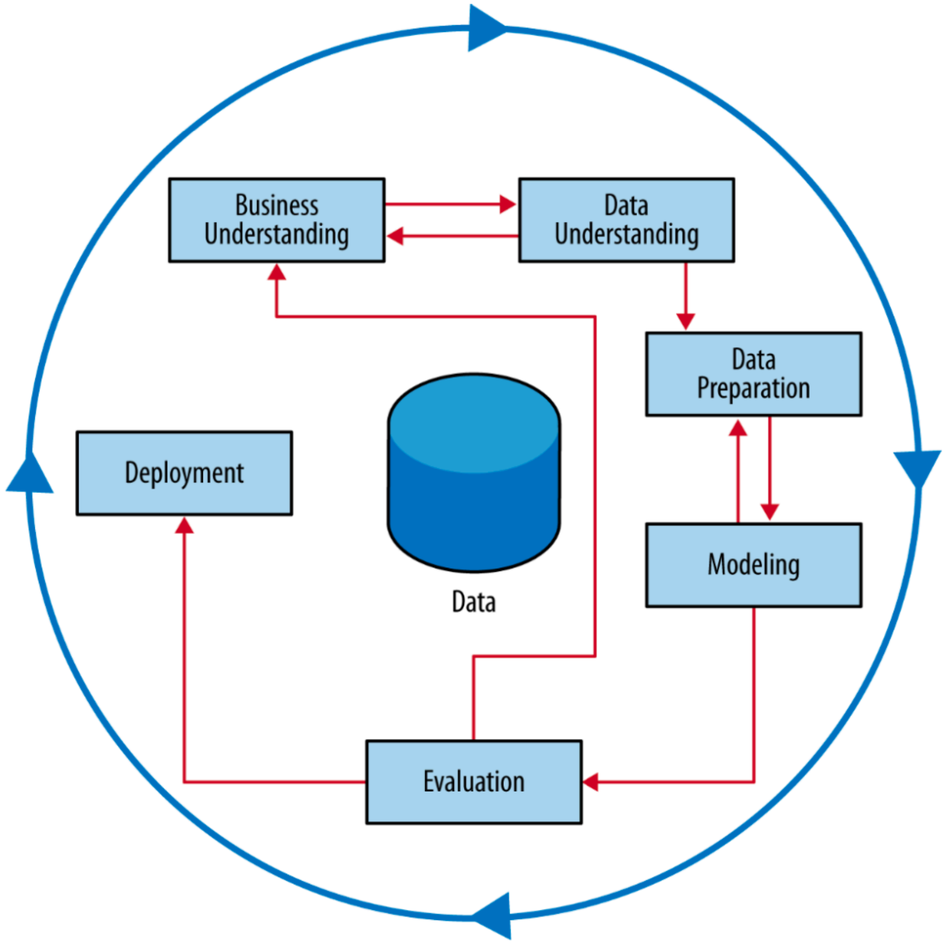
Introduction to Data Science

Lesson 3

Data Preparation and Exploratory Data Analysis (EDA).

Marija Stankova Medarovska, PhD
marija.s.medarovska@uacs.edu.mk

Data science process



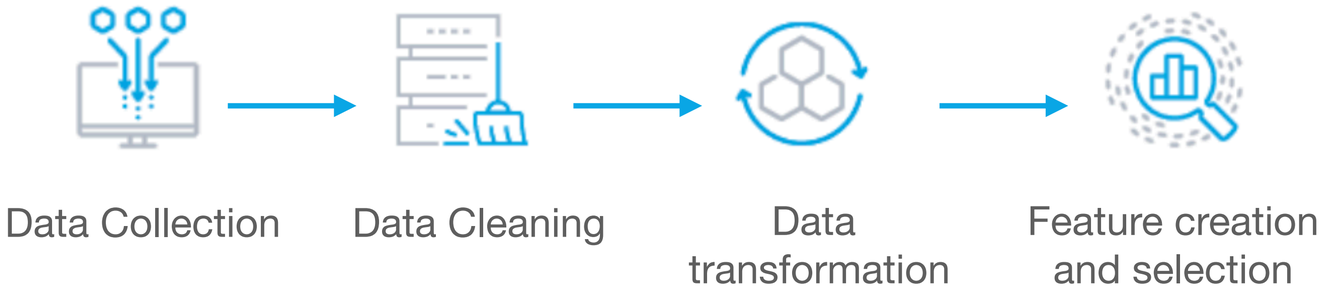
Data Preparation

Data preparation

- Data preparation is the process of cleaning, transforming and organizing raw data into a usable format
- Clean and well-prepared data leads to more accurate and reliable models
- Best practices in data preparation:
  - maintaining reproducibility
  - documentation of data preparation steps

Steps in data preparation

- Data collection: Collect data from various sources like databases, APIs, scraping the web, files, etc.
- Data cleaning: Handle missing values, outliers, correct errors, etc.
- Data transformation: Convert data into a appropriate format for modelling, such as scaling, encoding, normalizing, etc.
- Feature creation: Create new features or modify existing ones to improve model performance.



Data cleaning: Missing data

- Missing values are data points that are absent for a specific variable in a dataset.
- They can be represented in various ways, such as blank cells, null values, or special symbols like “NA”, “NaN” or “unknown.”

.

	School ID	Name	Address	City	Subject	Marks	Rank	Grade
0	101.0	Alice	123 Main St	Los Angeles	Math	85.0	2	B
1	102.0	Bob	456 Oak Ave	New York	English	92.0	1	A
2	103.0	Charlie	789 Pine Ln	Houston	Science	78.0	4	C
3	NaN	David	101 Elm St	Los Angeles	Math	89.0	3	B
4	105.0	Eva	NaN	Miami	History	NaN	8	D
5	106.0	Frank	222 Maple Rd	NaN	Math	95.0	1	A
6	107.0	Grace	444 Cedar Blvd	Houston	Science	80.0	5	C
7	108.0	Henry	555 Birch Dr	New York	English	88.0	3	B

Data cleaning: Missing data

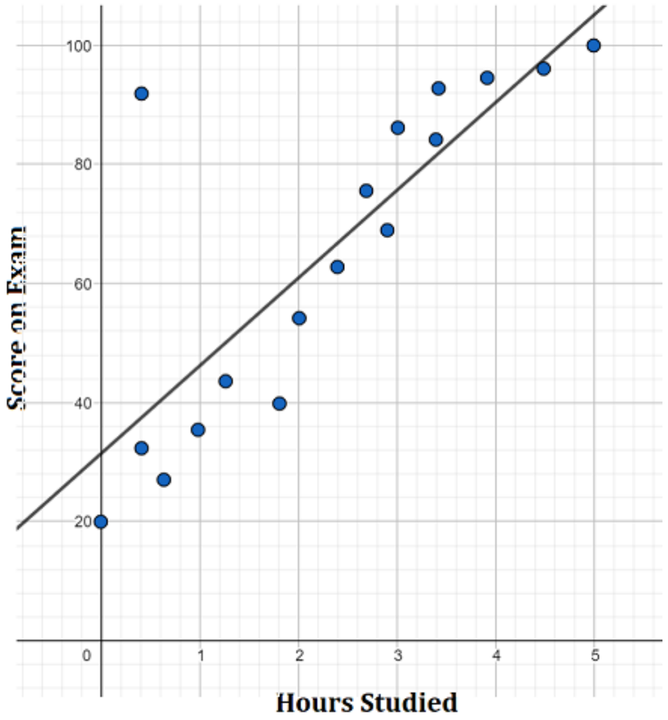
- Techniques for handling missing values:
  - Removing rows with missing data: removes instances with missing values altogether
  - Imputation methods: replacing missing values with estimated values (ex. mean, median and mode) or predictions (ex. k-nearest neighbors imputation that uses the values of the nearest neighbors to predict the missing data)

Data cleaning: Outliers

- An item of data that is significantly different from the rest of the data is called an outlier

- Outliers can be detected using statistical methods (z-score, interquartile range (IQR)), using visualisation techniques (box plot, scatterplot, histogram) etc.

- Handling techniques: removing outliers, imputation, transforming outliers



Data transformation: Feature scaling

Scale the features to bring them to a common scale, important for algorithms sensitive to feature magnitudes

- Normalization (also known as min-max scaling) is used to transform features to be on a similar scale, typically [0, 1].

- Standardization (also known as z-score normalization) transforms data to have a mean of 0 and a standard deviation of 1.

$$x^{(i)}_{std} = \frac{x^{(i)} - \mu_x}{\sigma_x}$$

$$x^{(i)}_{norm} = \frac{x^{(i)} - \mathbf{x}_{min}}{\mathbf{x}_{max} - \mathbf{x}_{min}}$$

standardization

min-max scaling ("normalization")

	input	standardized	normalized
0	0	-1.46385	0.0
1	1	-0.87831	0.2
2	2	-0.29277	0.4
3	3	0.29277	0.6
4	4	0.87831	0.8
5	5	1.46385	1.0

Data transformation: Encoding categorical variables

Data encoding is used to convert categorical variables into a numerical format

- in one-hot encoding each category is represented by a separate column with a 1 indicating its presence and 0s for all other categories.

- label encoding assigns a unique integer to each category

Color
Red
Red
Yellow
Green
Yellow

Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1
0	1	0

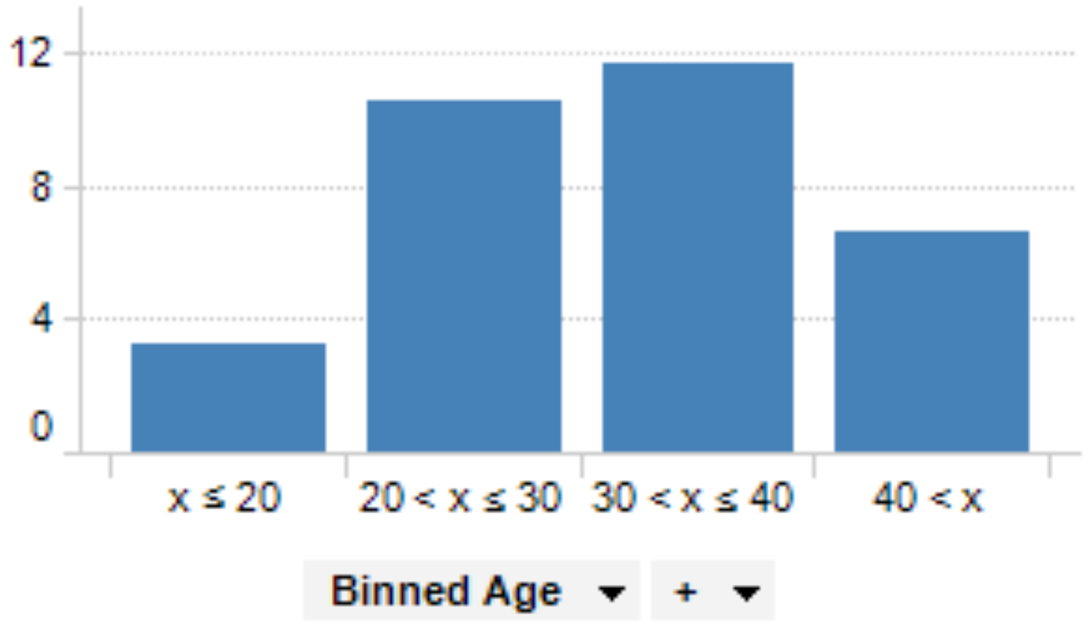
Example: One hot encoding

Data transformation: Discretization

Discretization (binning) is the process of converting continuous numerical data into discrete categories or intervals.

The range of a continuous feature is divided into bins (or intervals) and each data point is assigned to a corresponding bin

Types of binning: equal width binning, equal frequency binning, custom binning



Example: Discretization of a continuous variable

Feature creation

Feature creation is the process of generating new features based on domain knowledge or by observing patterns in the data.

- Domain specific: creating new features based on domain knowledge, such as creating features based on business rules or industry standards.
- Data driven: creating new features by observing patterns in the data, such as calculating aggregations or creating interaction features.
- Synthetic: generating new features by combining existing features or synthesizing new data points.

Feature selection

Feature selection involves selecting a subset of relevant features from the entire set of available features in a dataset.

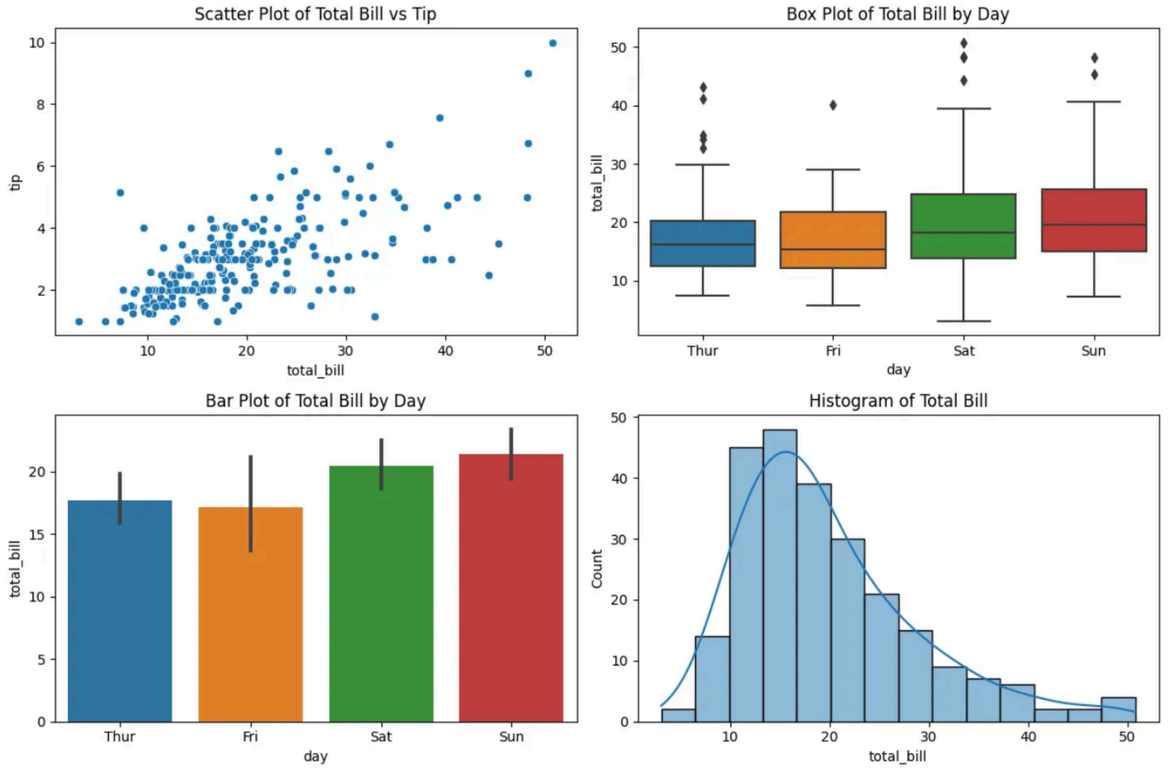
- Filter methods: evaluate the relevance of features by their statistical properties
- Wrapper methods: the feature subset that results in the best performance of the model is selected
- Embedded methods**:** incorporate feature selection as part of the model training process

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA)

John Tukey in 1977 coined the term exploratory data analysis (EDA) for the process of exploring data in order to gain an understanding of it, not to make predictions or test hypotheses.

This is done mostly with visualizations and summary statistics.

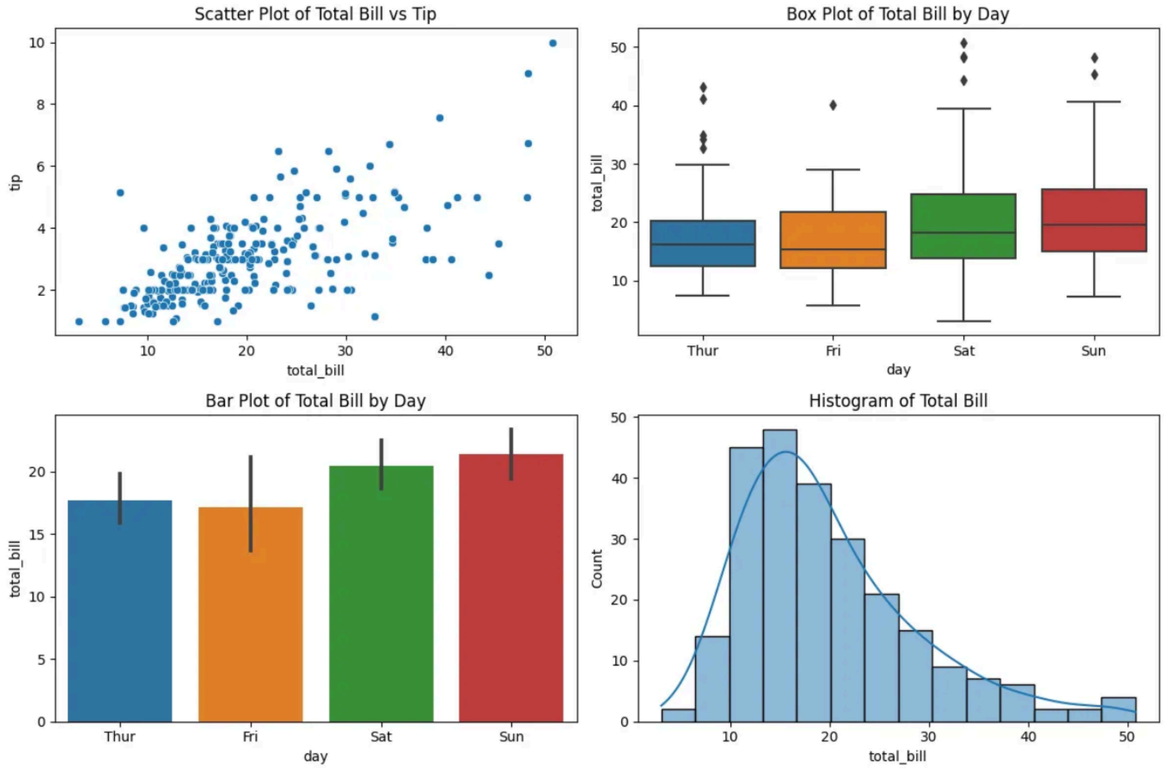


Example: Different types of plots for exploratory data analysis (EDA) and visualisation using the Python data visualization library Seaborn.

Exploratory Data Analysis (EDA)

EDA is important for:

- Understanding the data
- Identifying patterns and relationships
- Detecting outliers, missing values, errors in data, etc.
- Guiding the modelling process



Example: Different types of plots for exploratory data analysis (EDA) and visualisation using the Python data visualization library Seaborn.

Understanding the data structure

Simply by looking at the first several rows of the dataset we can understand a lot about the data

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NaN
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	NaN
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	NaN
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 metres	NaN

Example: Output using the head() function in Python (pandas)

[1] https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results/data

Descriptive statistics

Descriptive statistics provides summaries about the sample as an approximation of the population.

	ID	Age	Height	Weight	Year
count	271116.000000	261642.000000	210945.000000	208241.000000	271116.000000
mean	68248.954396	25.556898	175.338970	70.702393	1978.378480
std	39022.286345	6.393561	10.518462	14.348020	29.877632
min	1.000000	10.000000	127.000000	25.000000	1896.000000
25%	34643.000000	21.000000	168.000000	60.000000	1960.000000
50%	68205.000000	24.000000	175.000000	70.000000	1988.000000
75%	102097.250000	28.000000	183.000000	79.000000	2002.000000
max	135571.000000	97.000000	226.000000	214.000000	2016.000000

Example: Output using the describe() function in Python (pandas)

Descriptive statistics: Mean and Variance

If you have a sample of *n* values, x_i,the sample mean μ is the sum of the values divided by the number of values.

$$\mu = \frac{1}{n}\sum_i x_i$$

Variance σ^2 describes the spread of data. The square root of variance, is called the standard deviation

$$\sigma^2 = \frac{1}{n}\sum_i (x_i - \mu)^2$$

Descriptive statistics: Median and Mode

The median is the value that divides a dataset into two equal halves, meaning 50% of the values lie below the median and 50% lie above it.If the number of elements in the data is odd then the center element is the median and if it is even then the median would be the average of two central elements.

Mode is the value that has the highest frequency in the given data. The data set may have no mode if the frequency of all data points is the same. Also, we can have more than one mode if we encounter two or more data points having the same frequency.

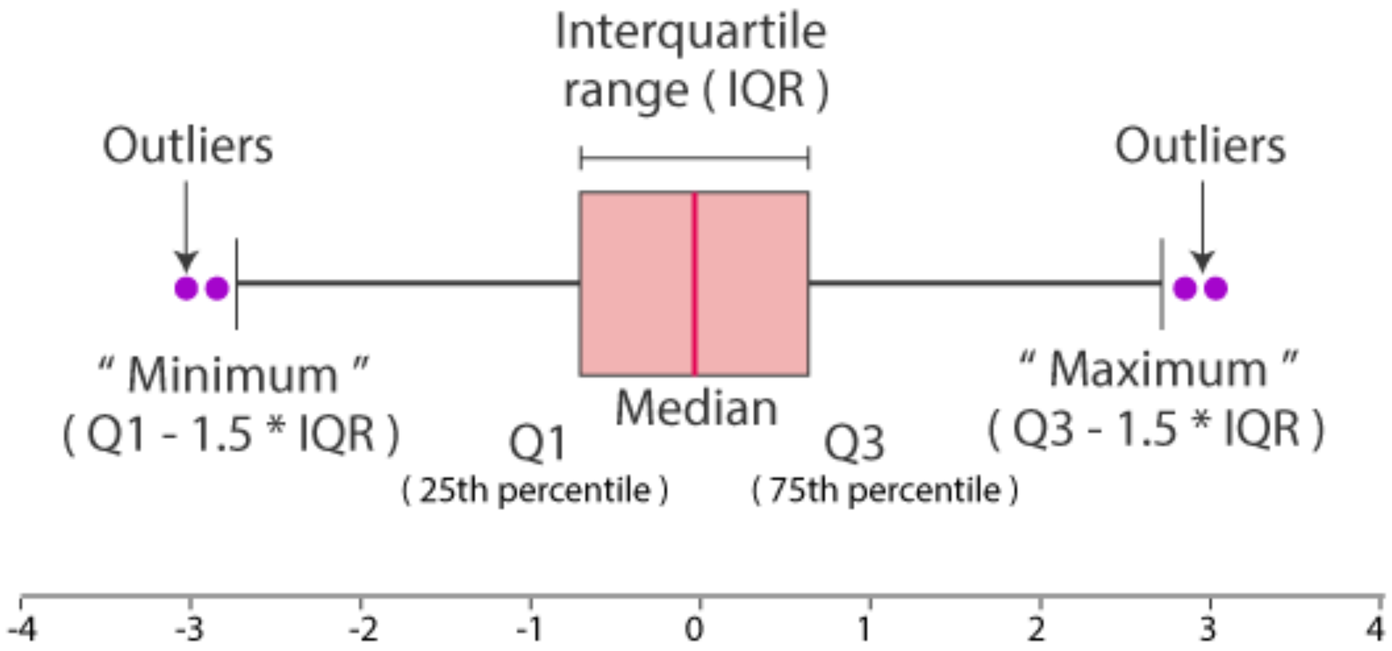
Descriptive statistics: Quantiles and Percentiles

Order the sample $\{x_i\}$, then find x_p so that it divides the data into two parts where:

- a fraction p of the data values is less than or equal to x_p and
- the remaining fraction $(1 - p)$ is greater than x_p.

That value, x_p, is the *p-th* quantile, or the $100 \times$ *p-th* percentile. For example, a 5-number summary is defined by the values $x_{min}, Q_1, Q_2, Q_3, x_{max}$, where Q_1 is the $25 \times$ *p-th* percentile, Q_2 is the $50 \times$ *p-th* percentile and Q_3 is the $75 \times$ *p-th* percentile.

Descriptive statistics: Boxplots and Outliers



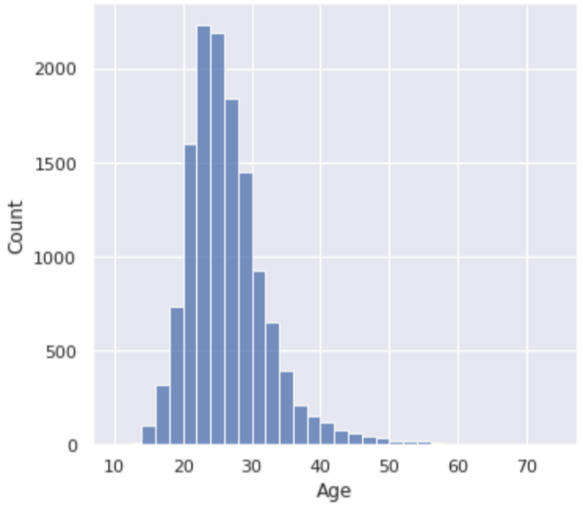
Example: Different parts of boxplot

Univariate Analysis

Summarizing data by just looking at their mean, median, and variance could be misleading: very different data can be described by the same statistics.

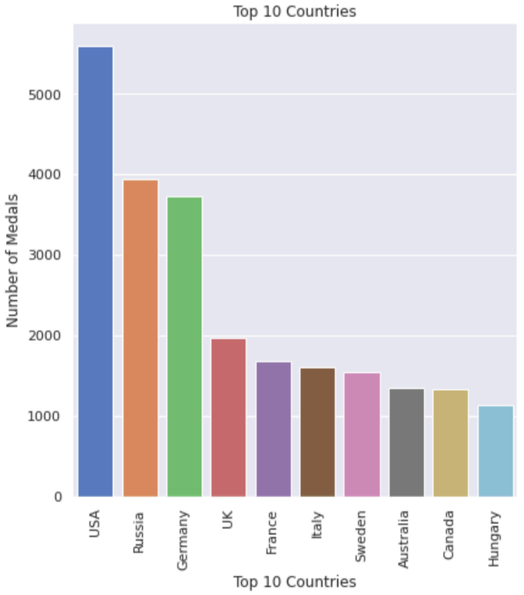
We can have a look at the data distribution, which describes how often each value appears (i.e., what is its frequency).

Commonly used visualisations for univariate analysis are histograms, box plots and bar charts



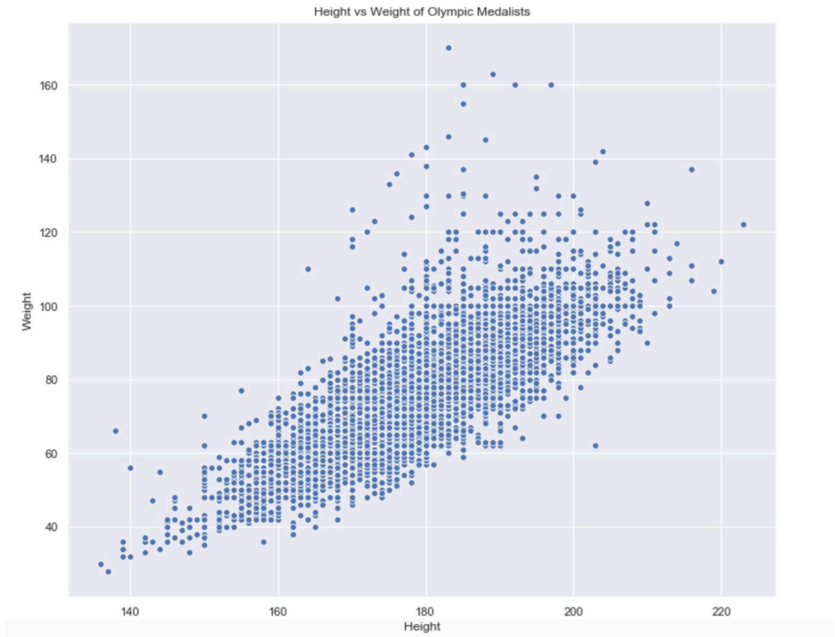
Example: Histogram graph that shows the frequency of each age value.

Example: Barplot graph that shows the count of medals for each country



Multivariate Analysis

With multivariate analysis we are seeking to understanding the relationships between two or more variables.



Example: Scatterplot that plots the height and weight of the athletes.

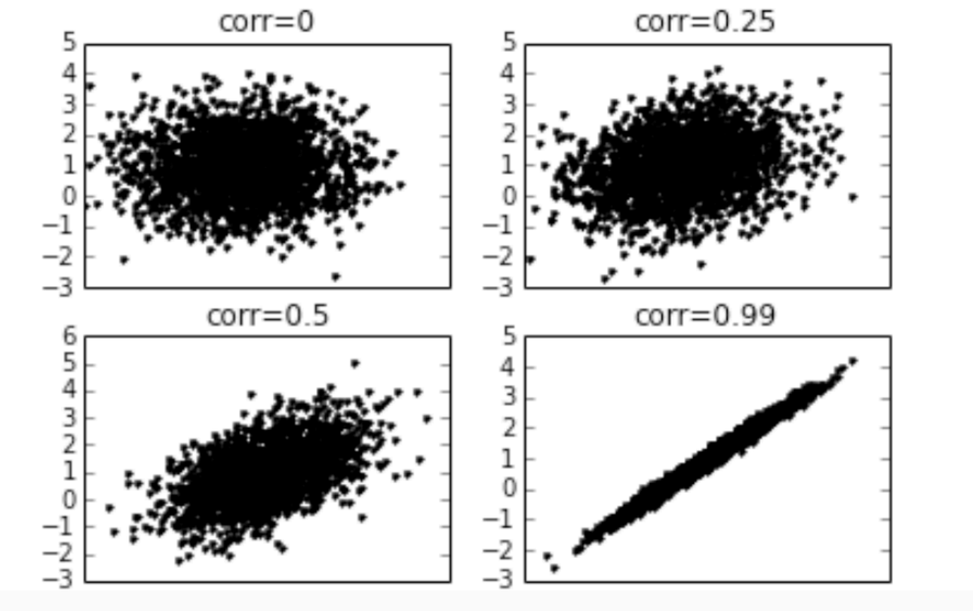
Example: A heatmap that shows the average age of medal winners in olympic games

Correlation

Correlation is a statistical measure that describes the strength and direction of a relationship between two variables. It shows whether and how strongly pairs of variables are related.

The most commonly used correlation coefficient is Pearson's correlation coefficient

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n} dx_i dy_i, \qquad \rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}.$$



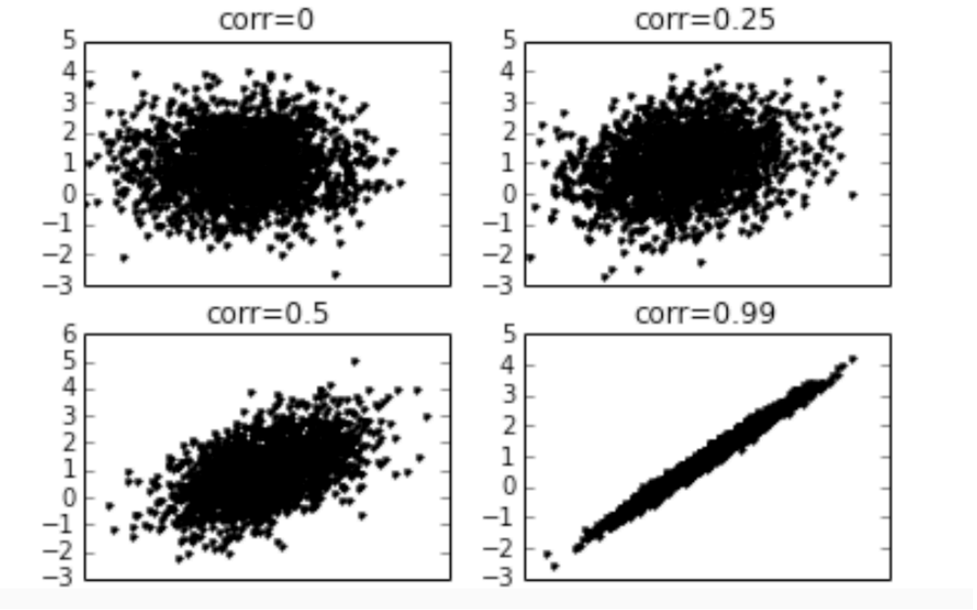
Example: Scatter plot of bivariate normally distributed variables with various correlations

Correlation

Pearson's correlation coefficient has a value between –1 and +1, where the magnitude depends on the degree of correlation. If the Pearson's correlation is 1, it means that the variables are perfectly positively correlated (for –1 negatively). This means that one variable can predict the other very well.

Pearson's correlation captures correlations of first order, but not nonlinear correlations.

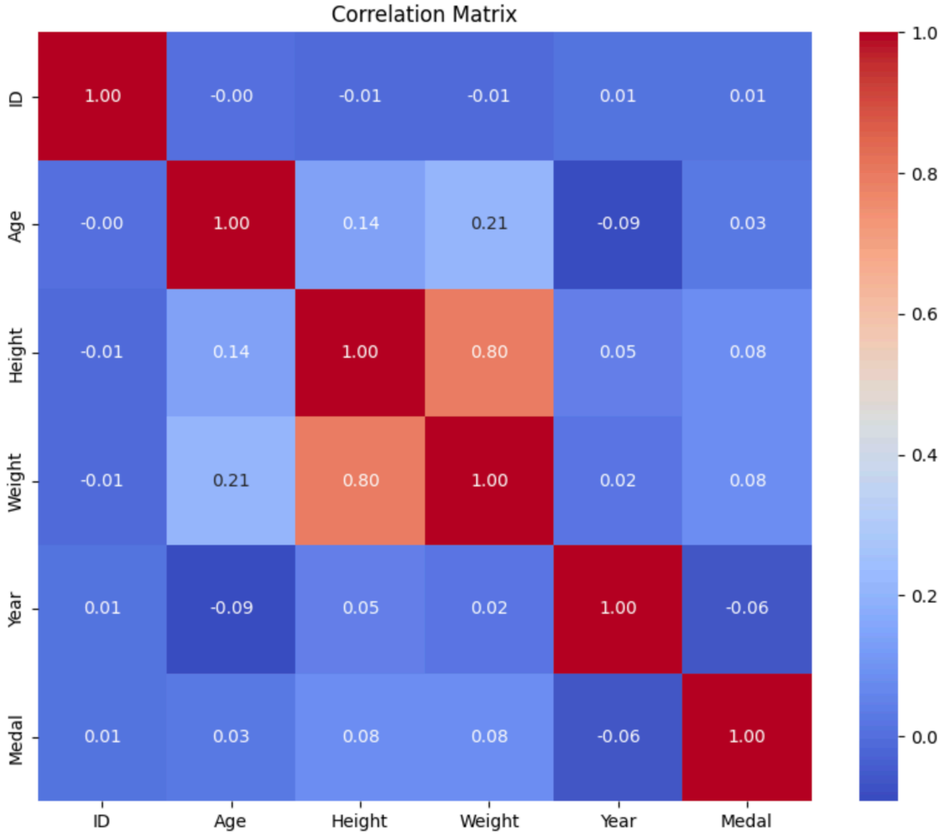
Spearman's rank correlation comes as a solution to the robustness problem of Pearson's correlation when the data contain outliers.



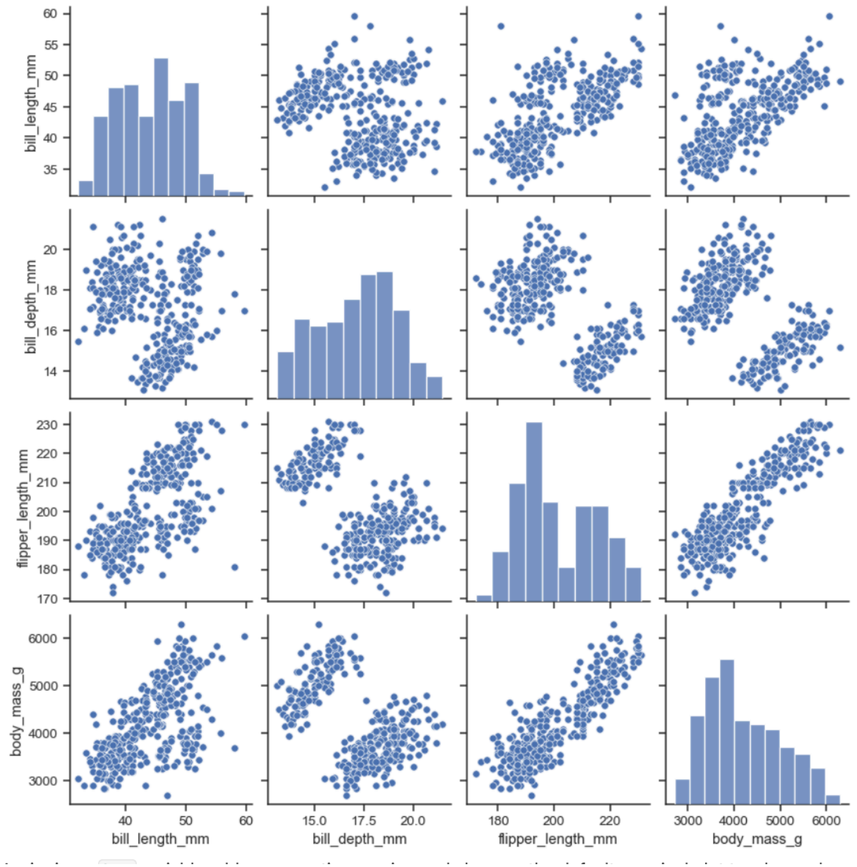
Example: Scatter plot of bivariate normally distributed variables with various correlations

Correlation matrix

A correlation matrix displays the correlation coefficients between multiple variables in a dataset.



Pairwise relationships

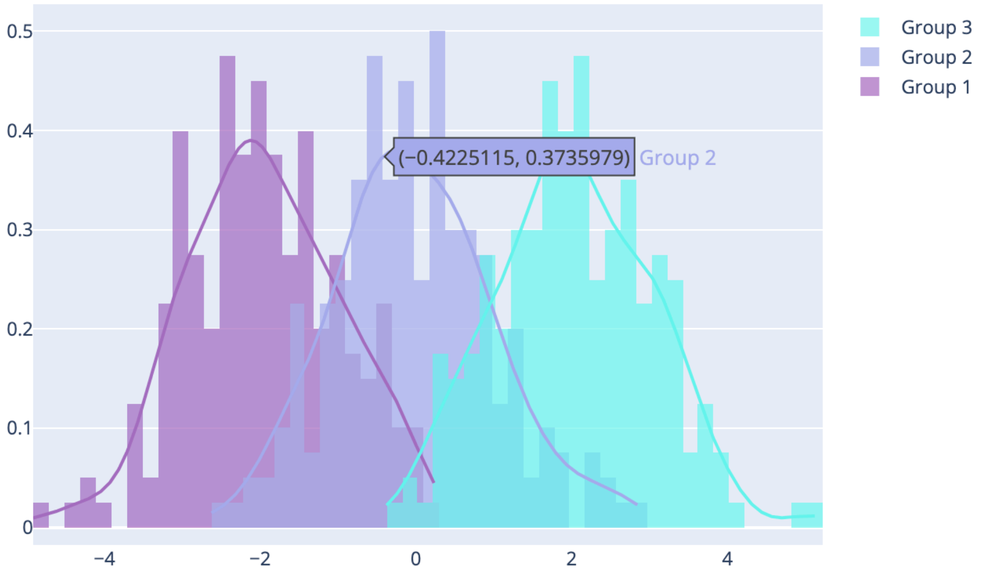


Example: Pairplot plots pairwise relationships in a dataset

Data Visualiation Tools

Python Libraries

- Several Python libraries can be used for visualising data such as Matplotlib, Seaborn, Plotly, etc.

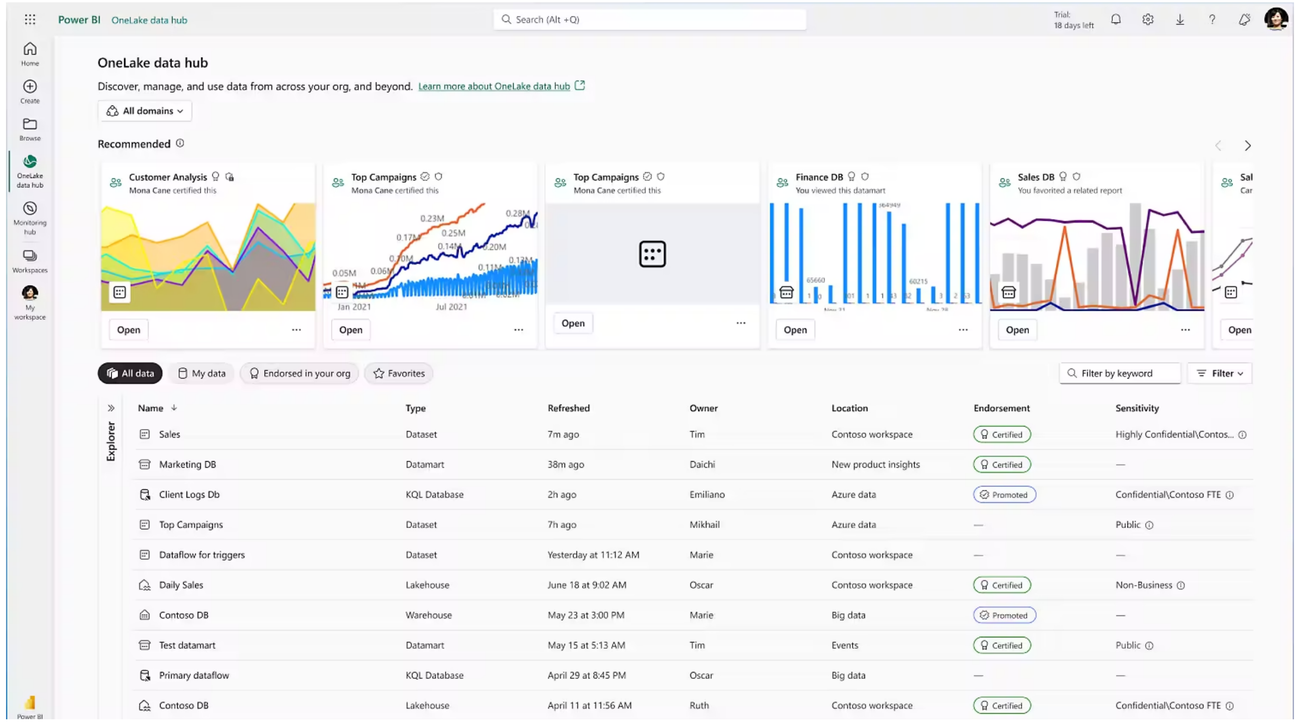


Example: Interactive histogram and curve plot în Plotly

[1] https://plotly.com/python/

Microsoft Power BI

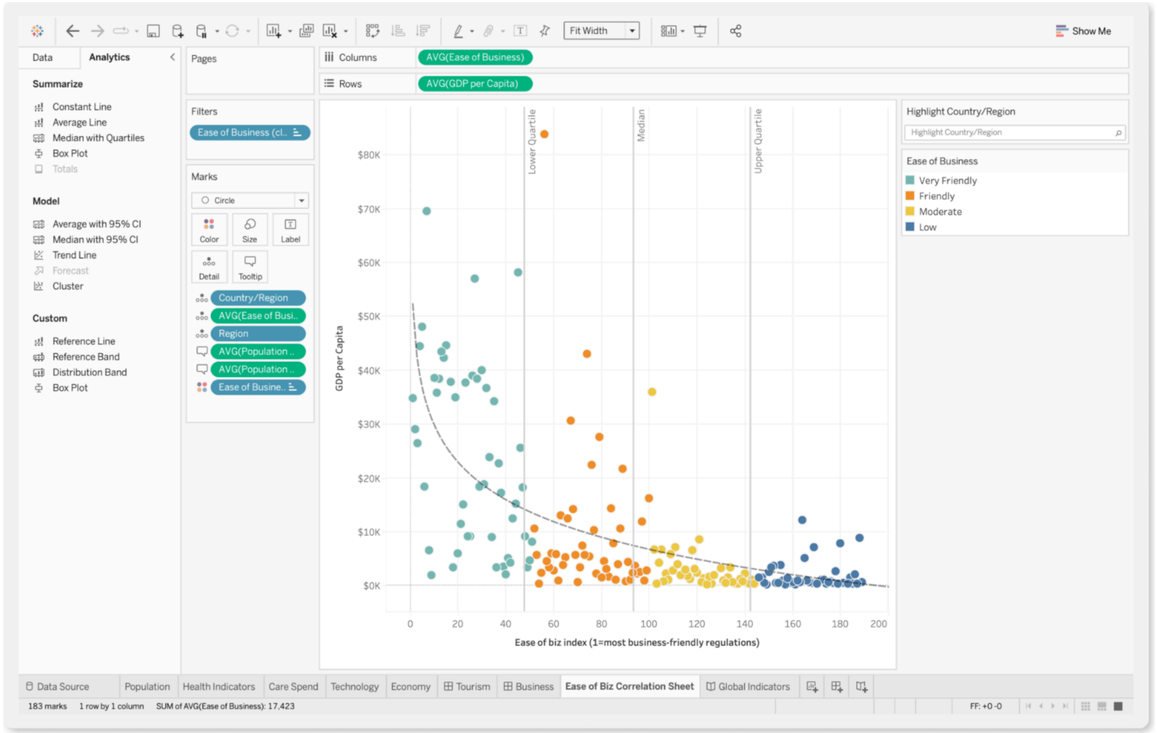
- Power BI is a data visualisation and business analytics tool by Microsoft that enables users to create interactive reports and dashboards



[1] https://learn.microsoft.com/en-us/power-bi/create-reports/sample-datasets

Tableau

- Tableau is a data visualisation tool that allows users to create interactive and shareable dashboards from a wide variety of data sources



[1] https://www.tableau.com/

Appendix