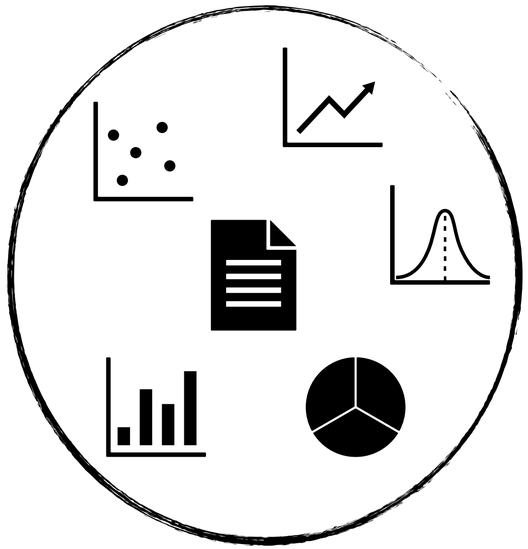
# Descriptive statistics

## Applied Data Analysis (ADA) - May 2025

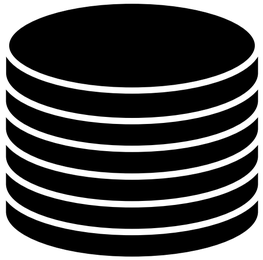
Nomades Advanced Technologies
Gaspard Villa

- Monday : Understand data structures
  - Population vs sampling
  - Central tendency measures
  - Dispersion measures
- Tuesday : Introduction to probability theory
- Wednesday : Central Limit Theorem confidence interval and test hypothesis
- Thursday : Feature selection and correlation matrix
- Friday : Statistics with scikit-learn

# How a project is built ?



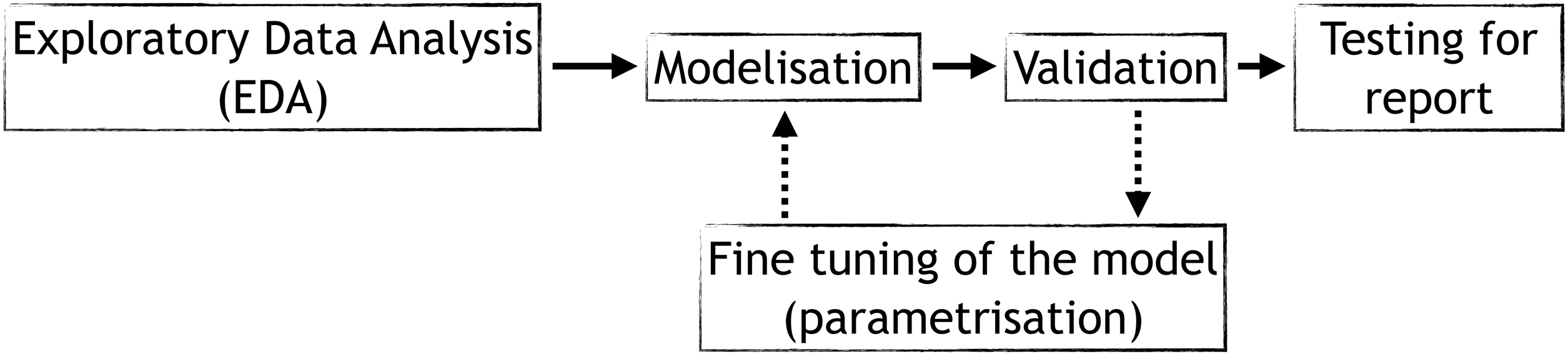
Raw data



Workable dataset

Modelisation, predictions, hidden informations, etc…

# How a project is built ?



# What’s descriptive statistics ?

**Definition :** descriptive statistics is about exploring and understanding a data set before going further into the modelisation.

Remark : Not the same as inferential statistics where we use a sample data set to make predictions on a larger population.

# Different types of data

## Unstructured

- Images
- Text
- Videos
- Time Series
- …

## Structured

- Numerical values
  - Continuous
  - Categorical

# Review on mean and median

1 - *Mean :* $\mu_X = \bar{X} = \frac{1}{n}\sum_{k=1}^{n} x_i$

`np.mean(x)`

2 - *Weighted mean :* $\bar{X} = \frac{1}{n}\sum_{k=1}^{n} w_i x_i$

`np.average(x, weights = w)`

3 - *Median :* $x_{\left[\frac{n}{2}\right]}$

`np.median(x)`

# Review on variability measures

1 - *Variance :* $\text{Var}[X] = \sigma_X^2 = \frac{1}{n}\sum_{k=1}^{n}\left(x_i - \mu_X\right)^2$

`np.var(x)`

2 - *Standard deviation :* $\sigma_X = \sqrt{\text{Var}[X]}$

`np.std(x)`

3 - *Covariance :* $\text{Cov}(X, Y) = \mathbb{E}\left[(X - \mu_X)(Y - \mu_Y)\right]$

`(Exercice)`

4 - *Correlation :* $\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$

`(Exercice)`

# Review on variability measures

1 - *Variance :* $\text{Var}[X] = \sigma_X^2 = \frac{1}{n}\sum_{k=1}^{n}\left(x_i - \mu_X\right)^2$

`np.var(x)`

2 - *Standard deviation :* $\sigma_X = \sqrt{\text{Var}[X]}$

`np.std(x)`

3 - *Covariance :* $\text{Cov}(X, Y) = \mathbb{E}\left[(X - \mu_X)(Y - \mu_Y)\right]$

`np.cov(X)`

4 - *Correlation :* $\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$

`np.corrcoef(X)`

# Key for analysis is Visualisation
# - See you data -

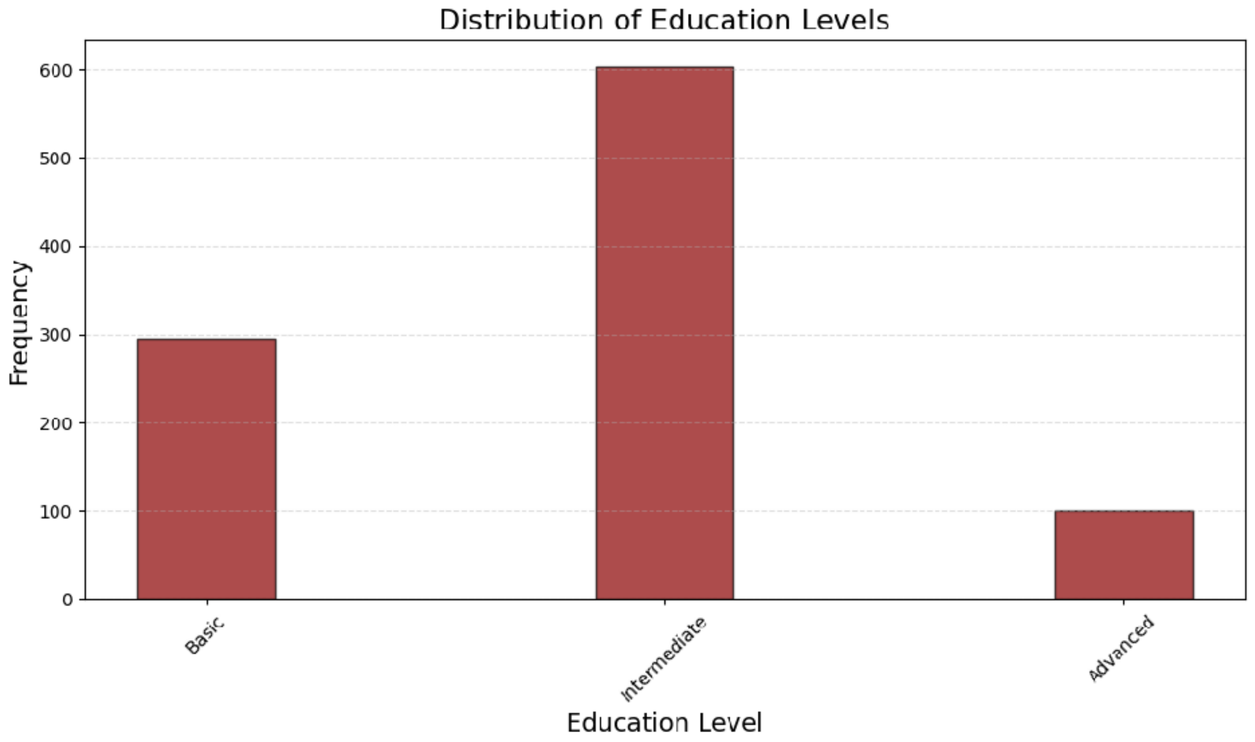
df.describe() is your friend
when you first see a data frame

| ID of the person | Age | Family size | Education level | Annual revenue [CHF] |
|---|---|---|---|---|
| **0** | 21 | “No family“ | Intermediate | 141 475 |
| **1** | 22 | “No family“ | Intermediate | 68 479 |
| **2** | 48 | Large | Basic | 129 630 |
| **3** | 52 | “No family“ | Intermediate | 159 280 |
| **4** | 62 | Small | Basic | 83 903 |
| **5** | 78 | “No family“ | Basic | 39 281 |
| **6** | 25 | “No family“ | Intermediate | 77 452 |
| **7** | 12 | Small | Intermediate | 358 865 |
| **8** | 53 | Medium | Advanced | 95 682 |

# Key for analysis is Visualisation
# - Univariate analysis -

1 - Univariate analysis for categorical variables

*Bar plots*

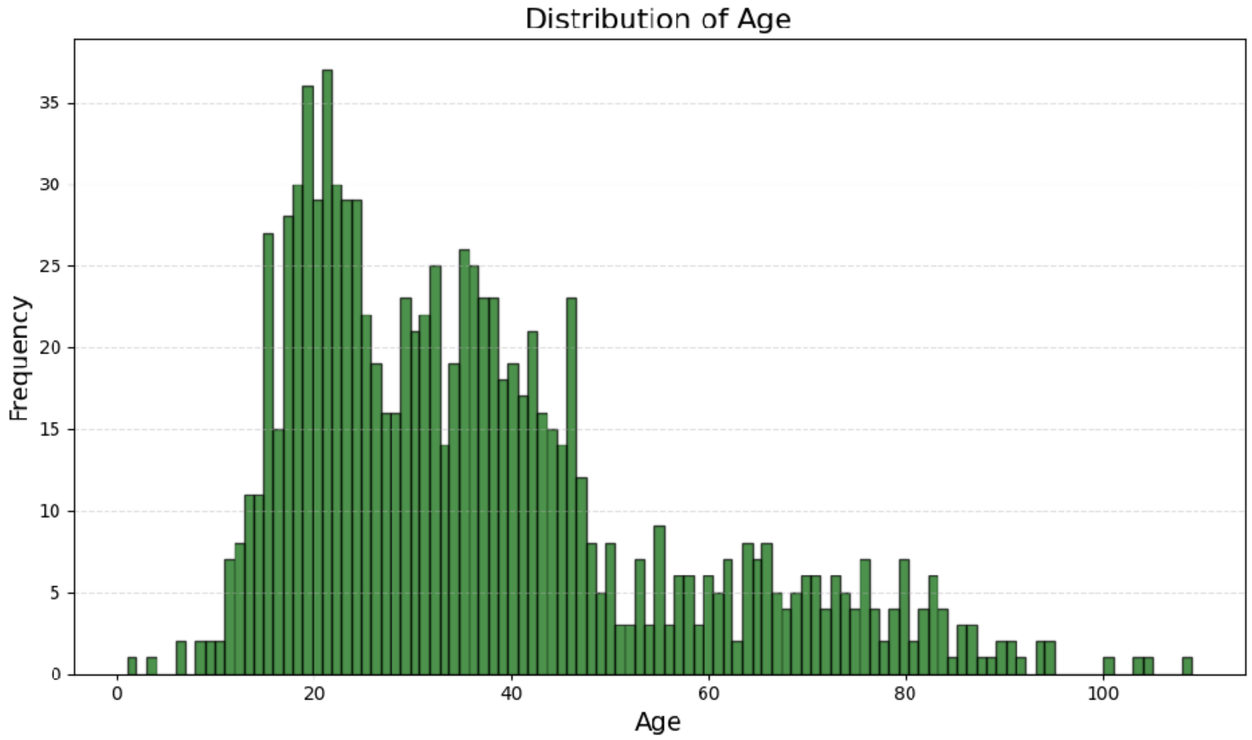



References

# Key for analysis is Visualisation
# - Univariate analysis -

2 - Univariate analysis for continuous variables

*Histograms*

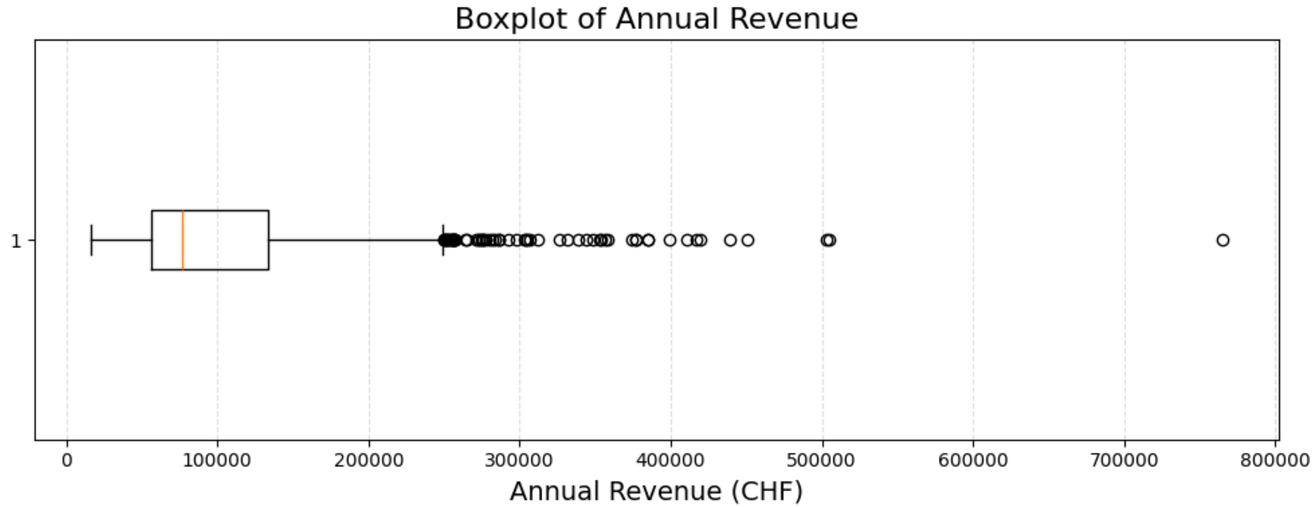



References

# Key for analysis is Visualisation
# - Univariate analysis -

3 - Univariate analysis for continuous variables

*Boxplots*

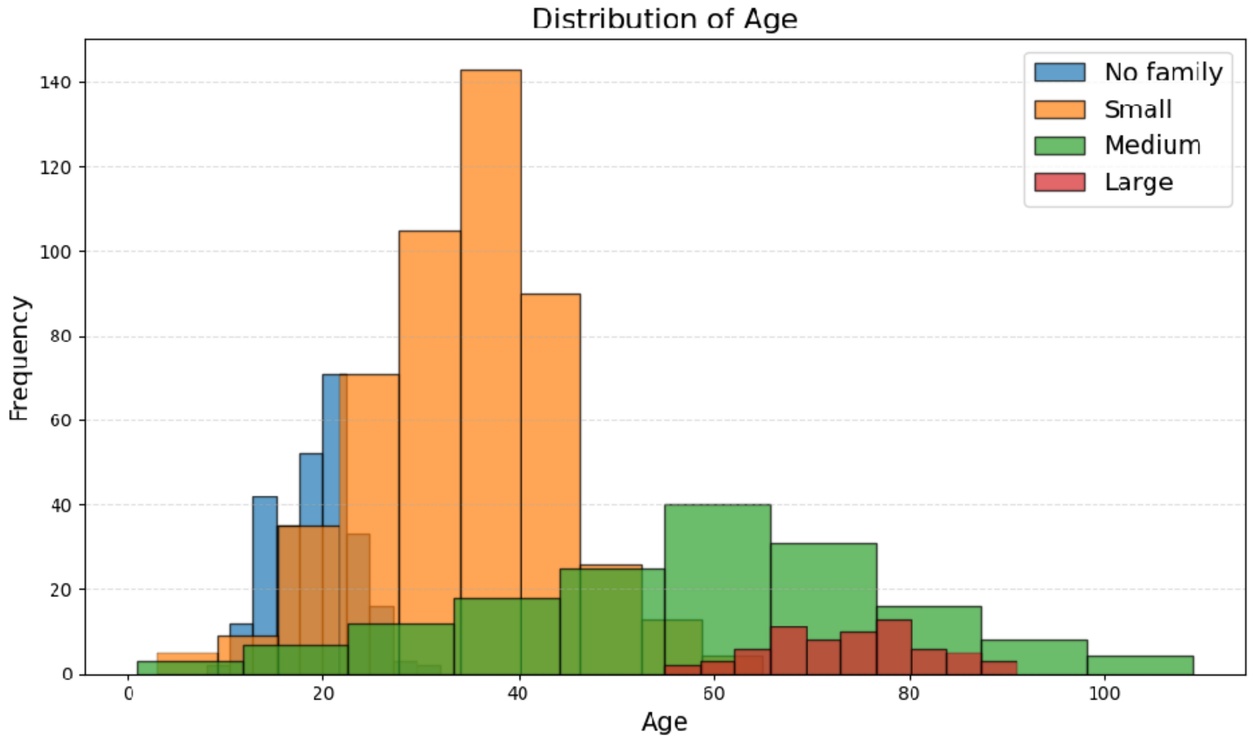



References

# Key for analysis is Visualisation
# - Multivariate analysis -

4 - Multivariate analysis for continuous and/or categorical variables

*Multivariate*

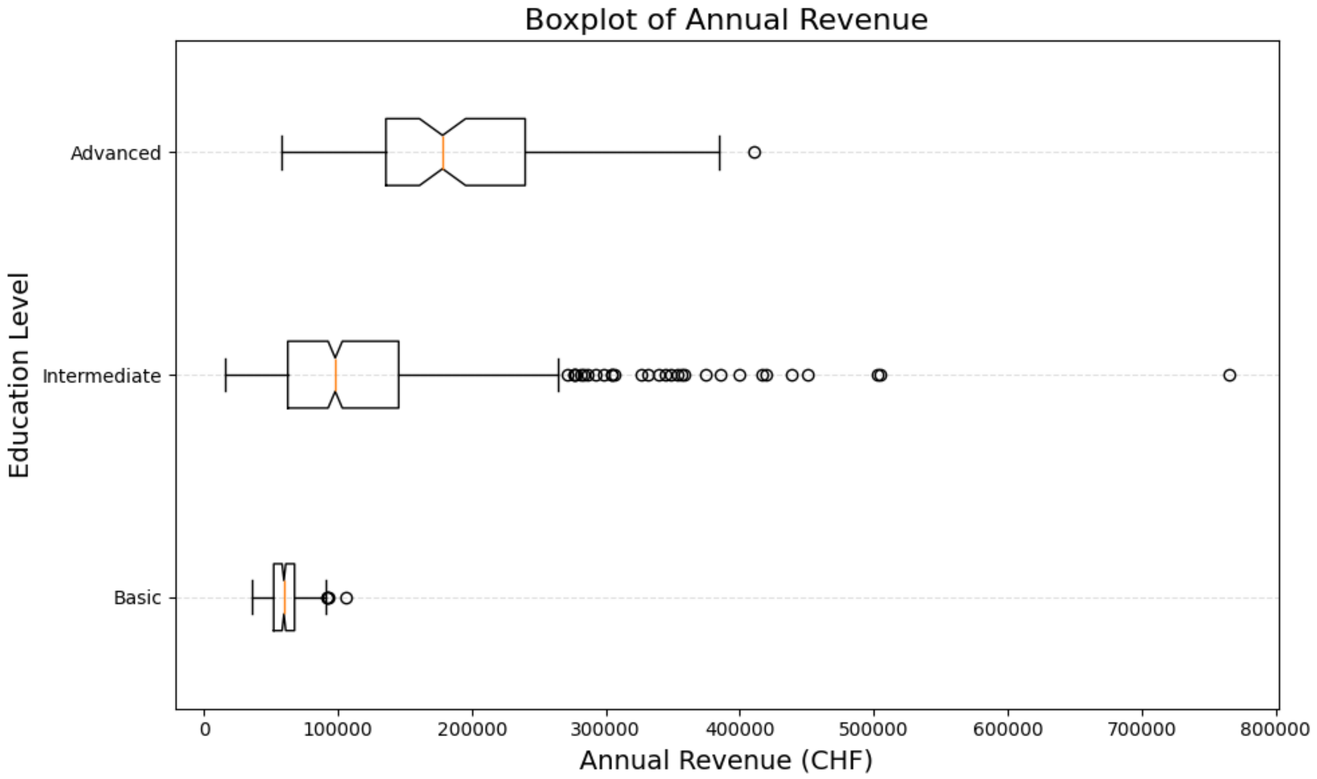



References

# Key for analysis is Visualisation
# - Multivariate analysis -

5 - Multivariate analysis for continuous and/or categorical variables

*Bivariate boxplots*

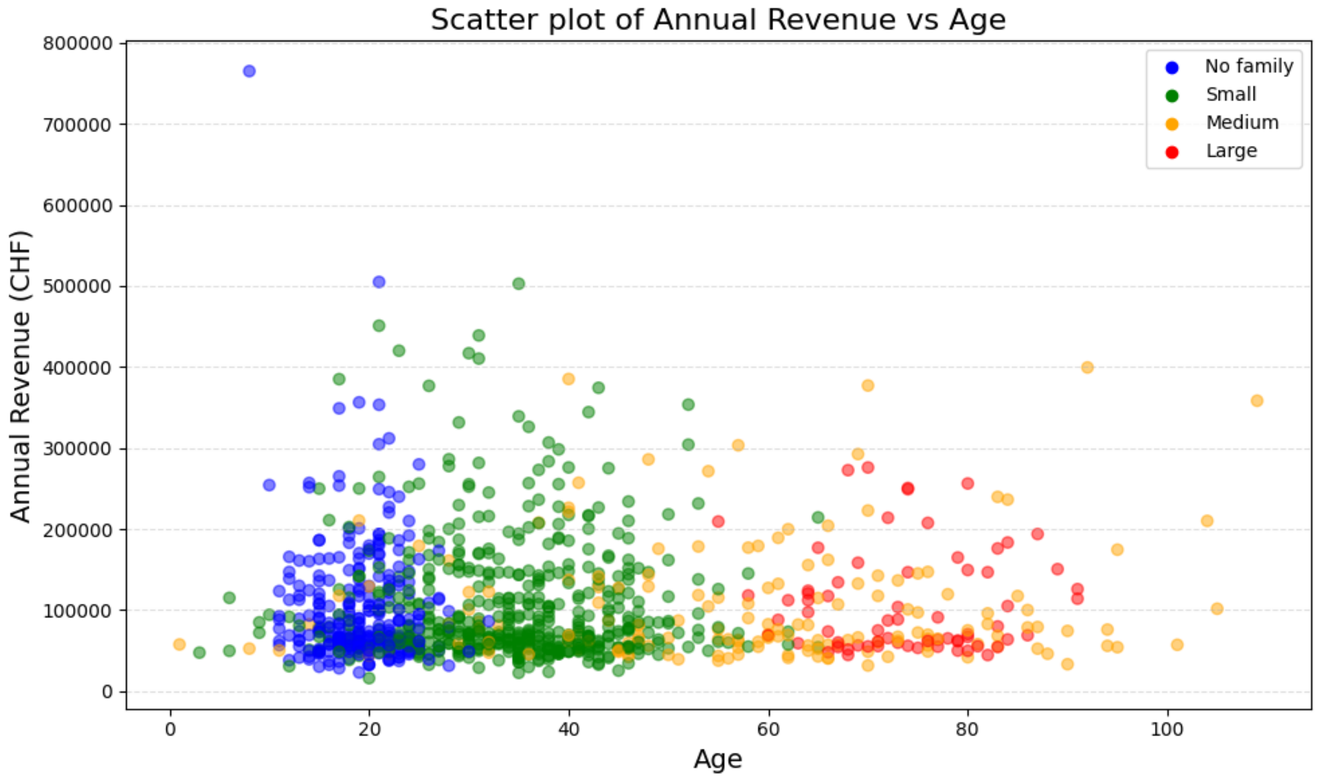



References

# Key for analysis is Visualisation
# - Multivariate analysis -

6 - Multivariate analysis for continuous and/or categorical variables

*Colored scatter plots*

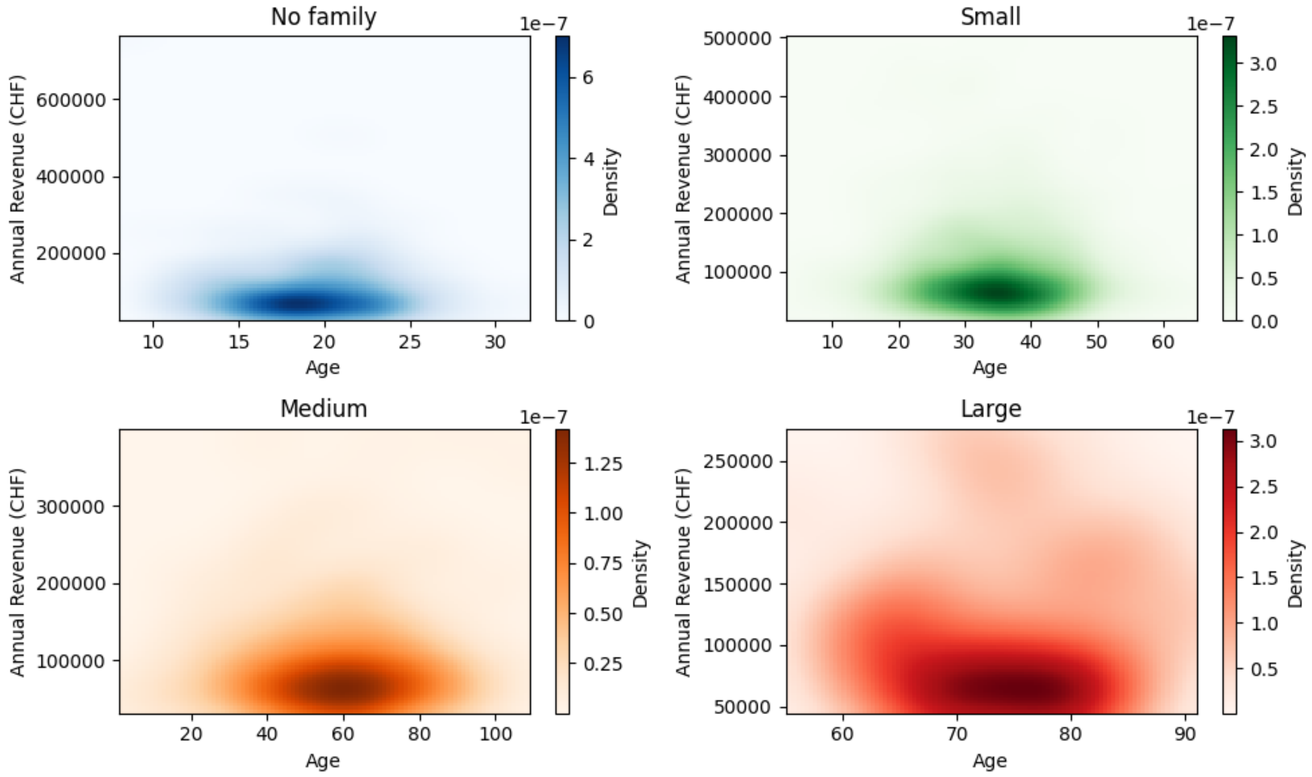



References

# Key for analysis is Visualisation
# - Multivariate analysis -

6 - Multivariate analysis for continuous and/or categorical variables

*Colored scatter plots*




References

# Data cleaning

Real-world datasets are generally never perfectly ready for use. Therefore, you often need to modify them to make them usable. Here are some steps:

- Remove duplicates / absurd data
- One-hot encoding
- Normalisation of continuous data
- Synchronisation of time series
- Manage outliers
- Missing data (be careful of bias)

# Remove duplicates

When working with multiple datasets from various sources, you might encounter duplicated data. Here are some key steps :

- Clearly identify each sample in the dataset
- Remove the samples that contain the least information or have the lowest quality assurance
- Warning : Samples can be very similar but still different! Be careful not to remove too many

# One-hot encoding

For the feature / attribute “Family size”, how would you tell your program whether you have no family or a large family ?

# One-hot encoding

For the feature / attribute “Family size”, how would you tell your program whether you have no family or a large family ?

*Idea 1 :* Assign a class value, e.g., No family = 0, Small = 1, Medium = 2, Large = 3.

*Idea 2 :* Use one-hot encoding! Add a column for each class and set the value to 1 if the sample belongs to that class, and 0 otherwise.

# One-hot encoding

| ID of the person | Age | Family size | Education level | Annual revenue [CHF] |
|---|---|---|---|---|
| **0** | 21 | “No family“ | Intermediate | 141 475 |
| **1** | 22 | “No family“ | Intermediate | 68 479 |
| **2** | 48 | Large | Basic | 129 630 |

| ID of the person | Age | Family No Family | Family Small | Family Medium | Family Large | Education Basic | Education Intermediate | Education Advanced | Annual revenue [CHF] |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 21 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 141 475 |
| **1** | 22 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 68 479 |
| **2** | 48 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 129 630 |

# Normalisation of continuous data set

Without normalisation, we may encounter the following issues:

- ✳ Large values of some features may have more importance during the modelisation
- ✳ Convergence and reliability of the models may be impacted during the process

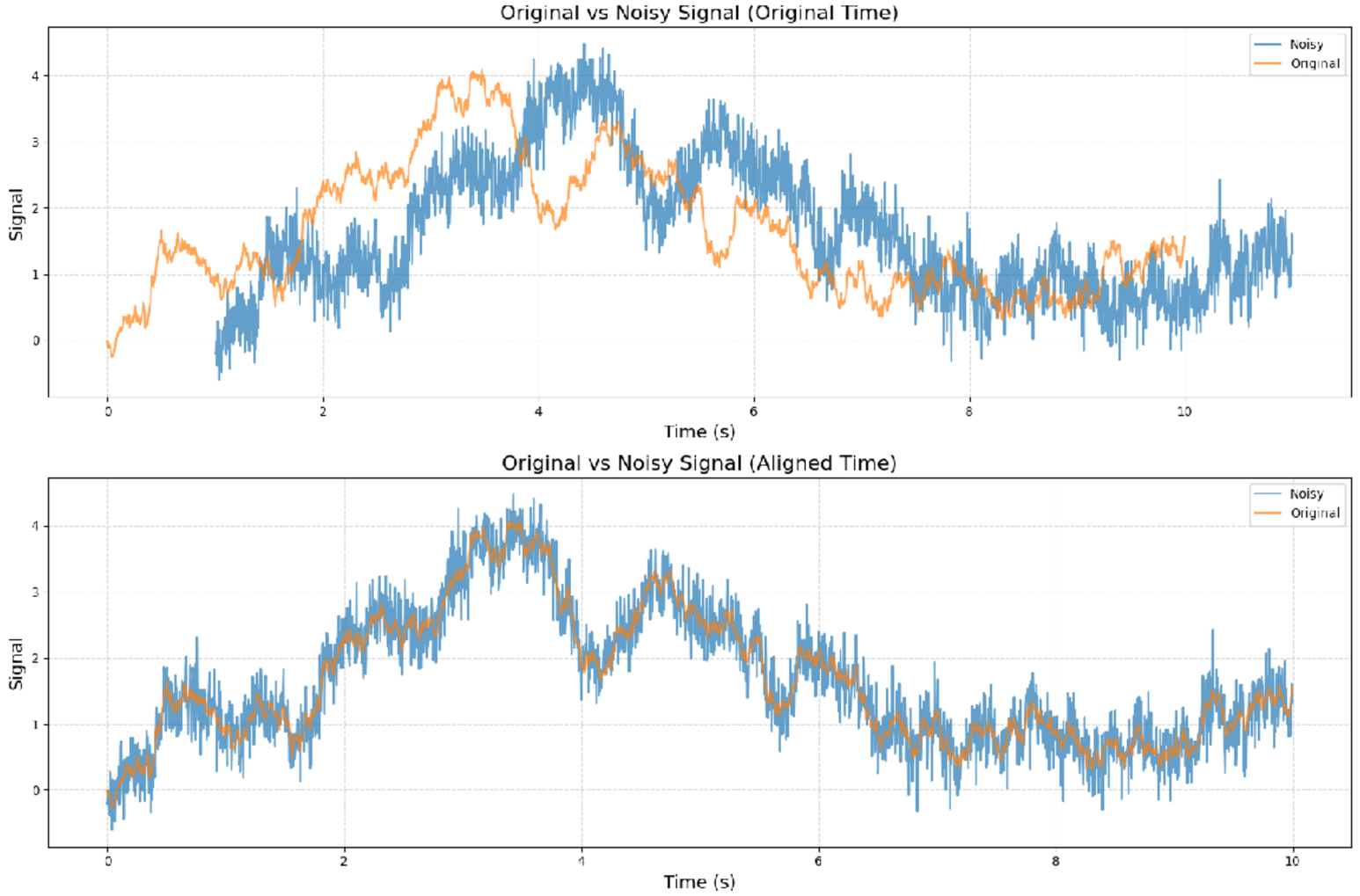
# Some normalisation formulas

1 - *Min-max normalisation :* $\tilde{x}_i = \dfrac{x_i - \min(X)}{\max(X) - \min(X)}$

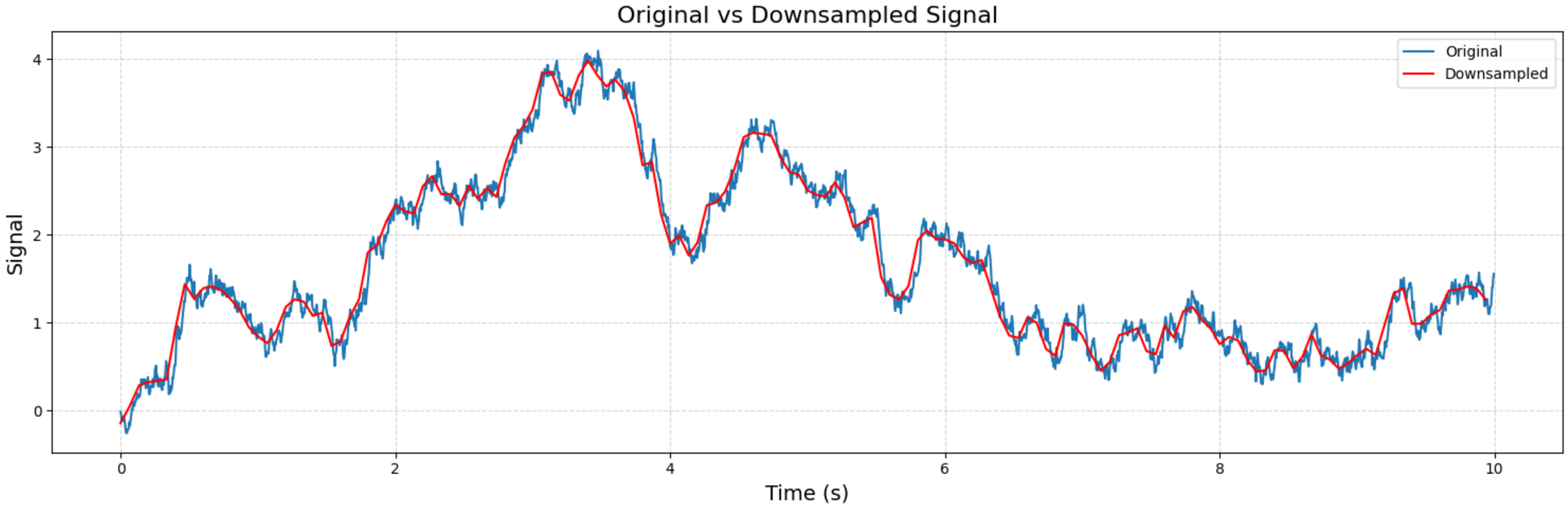
2 - *Standardisation :* $\tilde{x}_i = \dfrac{x_i - \mu}{\sigma}$

3 - *log-normalisation :* $\tilde{x}_i = \log(x_i + 1)$ or $\tilde{x}_i = \log(x_i)$

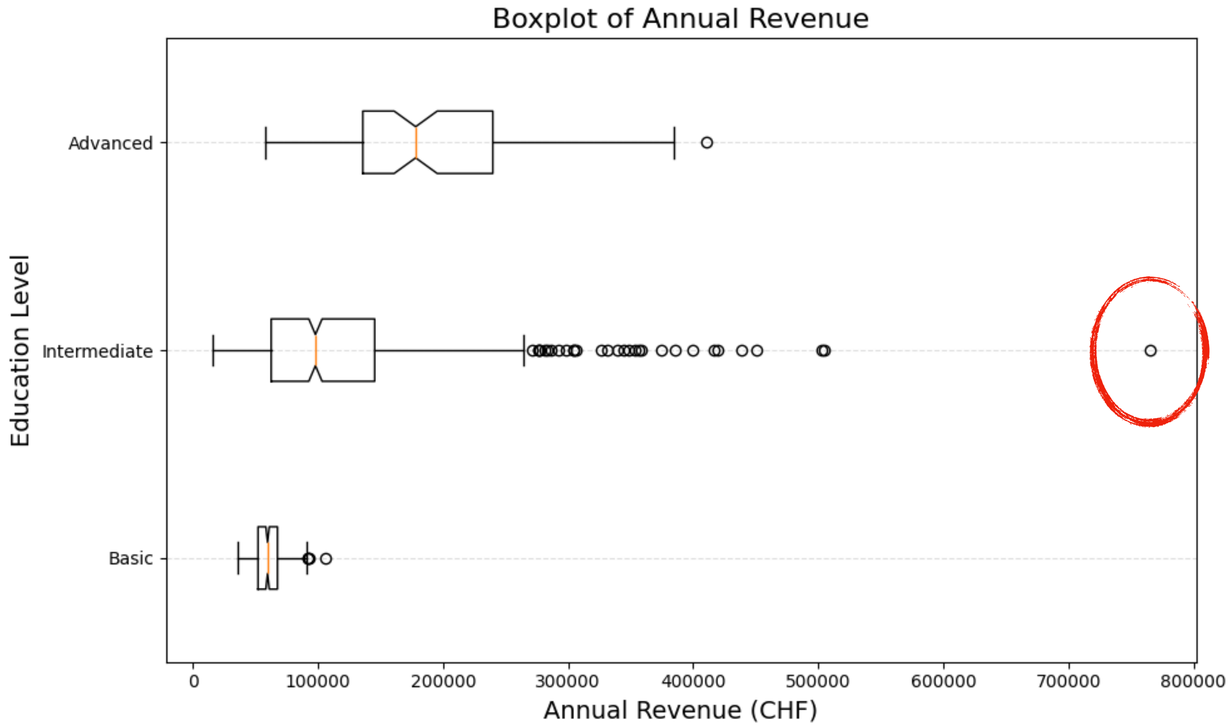
# Synchronisation of time series



# Re-sampling of time series



# Manage outliers




| ID of the person | Age | Family size | Education level | Annual revenue [CHF] |
|---|---|---|---|---|
| **0** | 21 | “No family“ | Intermediate | 141 475 |
| **1** | 22 | “No family“ | Intermediate | 68 479 |
| **2** | 48 | Large | Basic | 129 630 |
| **3** | 52 | “No family“ | Intermediate | 159 280 |
| **4** | 62 | Small | Basic | 83 903 |
| **5** | 78 | “No family“ | Basic | 39 281 |
| **6** | 25 | “No family“ | Intermediate | 77 452 |
| **7** | 12 | Small | Intermediate | 358 865 |
| **8** | 53 | Medium | Advanced | 95 682 |

# Missing data

| ID of the person | Age | Family size | Education level | Annual revenue [CHF] |
|---|---|---|---|---|
| **0** | 21 | “No family“ | Intermediate | 141 475 |
| **1** | 22 | “No family“ | Intermediate | 68 479 |
| **2** | 48 | Large | Basic | - None - |
| **3** | 52 | “No family“ | Intermediate | 159 280 |
| **4** | 62 | Small | Basic | 83 903 |
| **5** | 78 | “No family“ | Basic | 39 281 |
| **6** | 25 | - None - | Intermediate | - None - |
| **7** | 12 | Small | Intermediate | 358 865 |
| **8** | 53 | Medium | - None - | 95 682 |

# Exercice for tomorrow

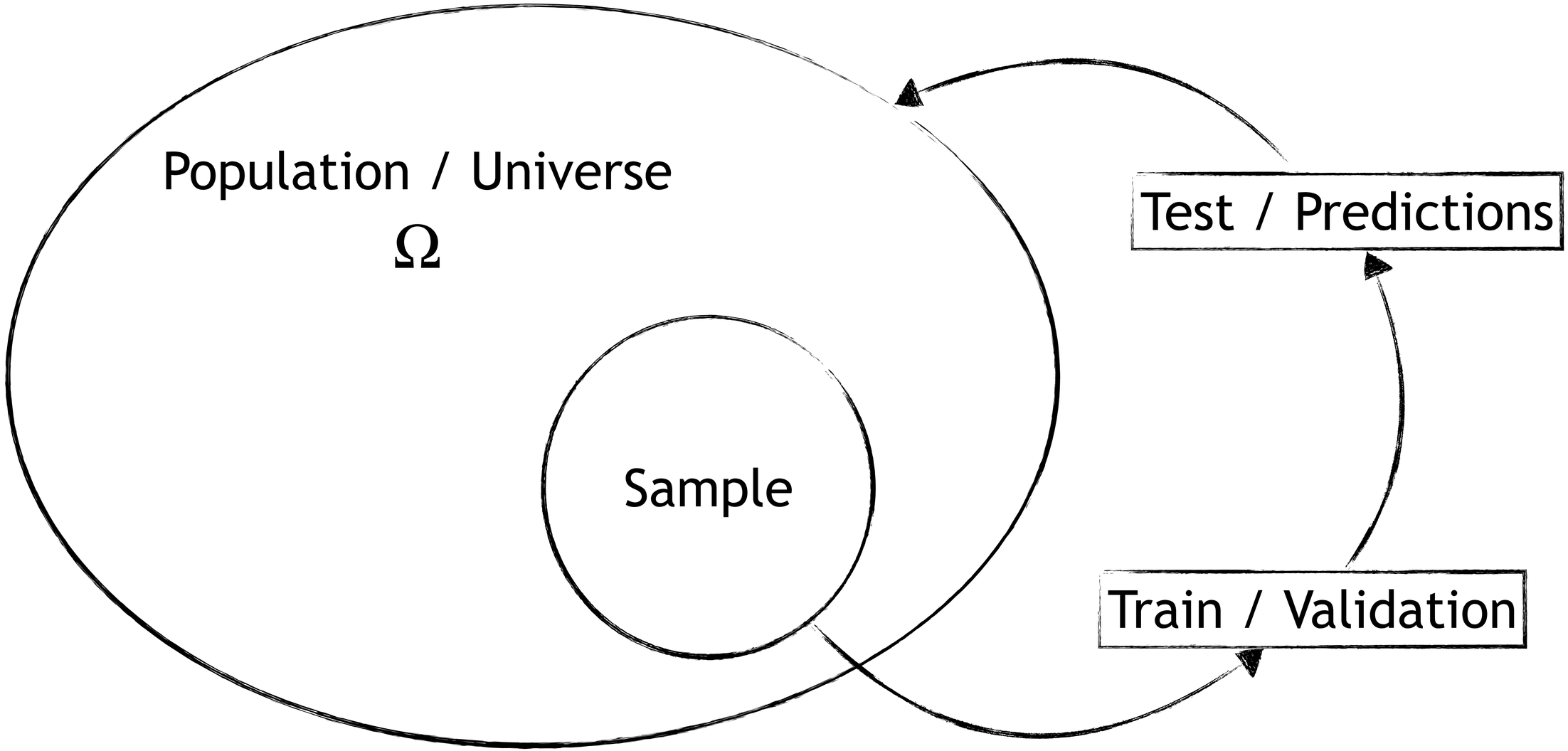
The idea is that you get a new data set of the houses there is in Boston. Multiple features are extracted in this data set such as:

- Garden size
- Distance from downtown
- Surface area
- Number of floors
- Type of walls [concrete, bricks, wood]
- Presence of a pool
- Estimated price

*Exercice :* Explore the features with the tools presented in the slides.

- Monday : Understand data structures
- Tuesday : Introduction to probability theory
  - Fundamental definitions
  - Probability law / distribution (discrete)
  - Discrete vs continuous probability
- Wednesday : Central Limit Theorem confidence interval and test hypothesis
- Thursday : Feature selection and correlation matrix
- Friday : Statistics with scikit-learn

# Introduction to probability theory



# Some definitions

- **Universe :** denoted as $\Omega$, it represents the complete set of all possible elements or outcomes you are studying.

- **Sample :** it represents a subset of the universe, selected for analysis.

- **Event :** it represents a set of outcomes from an experiment.

# Probability of an event

***Definition :*** Let $\Omega$ be a finite sample space (set of all possible outcomes) and $A \in \Omega$ an event, then the probability $\mathbb{P}[A]$ is defined as follow :

$$\mathbb{P}[A] = \frac{\text{Number of favorable outcomes for A}}{\text{Total number of outcomes in } \Omega} = \frac{|A|}{|\Omega|}$$

# Some practice

**Excercice :** You have a shuffled deck of 52 cards and you randomly pick one card. What is the probability of drawing a card with an even number (excluding face cards) ?

# Some practice

**Excercice :** You have a shuffled deck of 52 cards and you randomly pick one card. What is the probability of drawing a card with an even number (excluding face cards) ?

***Solution :*** You have to pick either 2, 4, 6, 8 and 10 (4 colours each), meaning you have 20 possible cards:

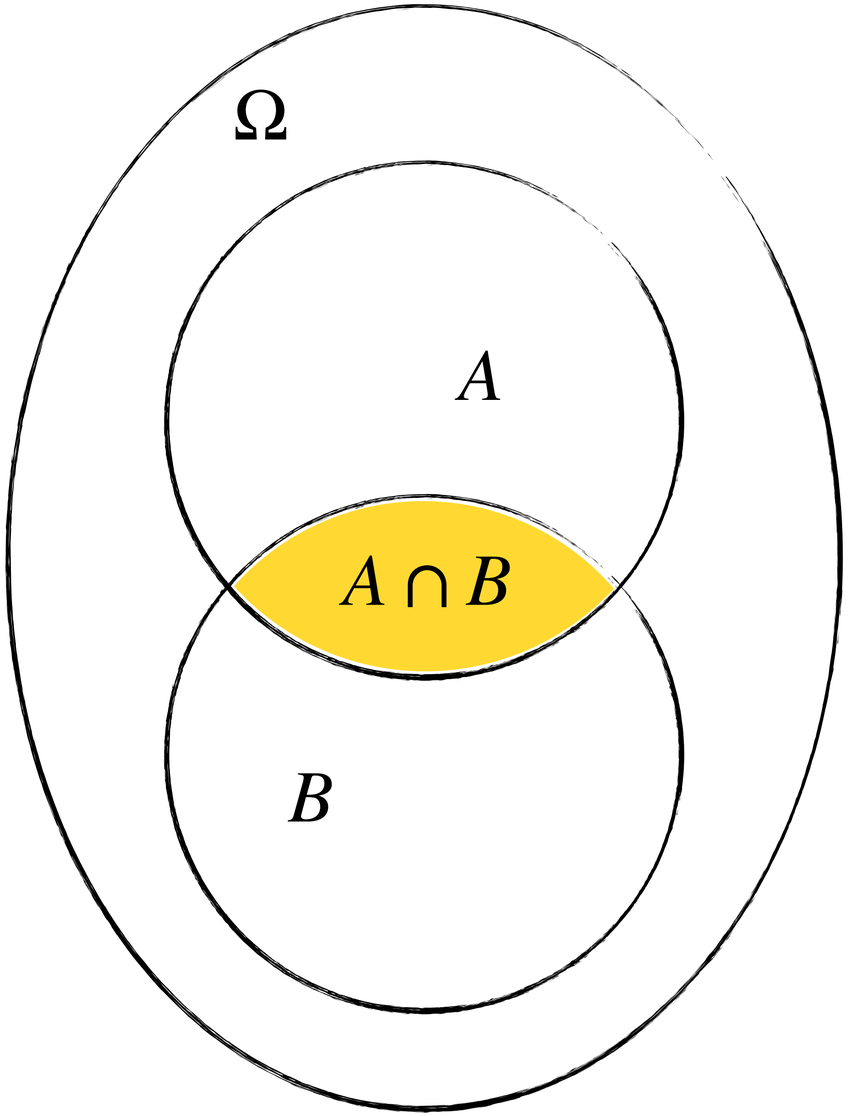
$$\mathbb{P}\left[\text{Pick even card}\right] = \frac{20}{52} \approx 0.38$$

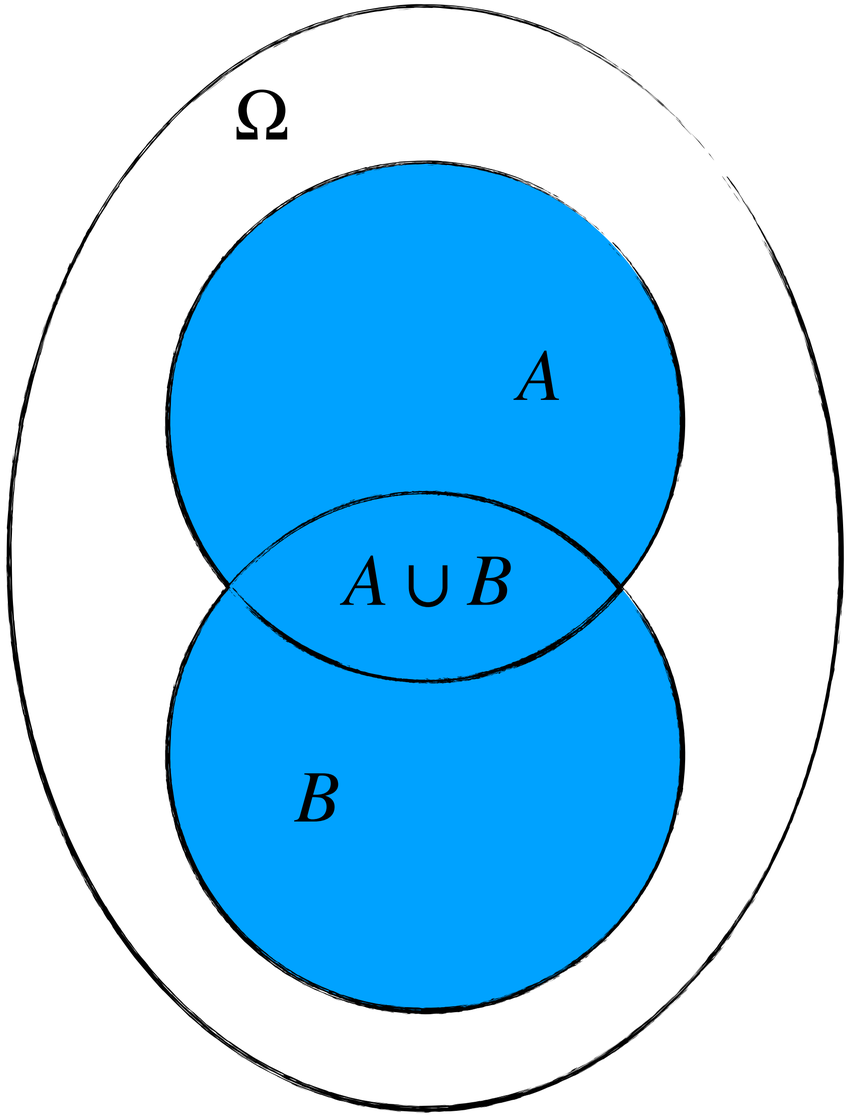
# Some properties

***Property :*** Let $\Omega$ be a finite sample space, then we have the following properties:

- $\mathbb{P}[\Omega] = 1$ and $\mathbb{P}[\varnothing] = 0$
- for all event $A \in \Omega$, $0 \leq \mathbb{P}[A] \leq 1$
- $\sum_{i=1}^{n} \mathbb{P}[A_i] = 1$ where $\Omega = \left\{ A_i \,|\, i = 1, \ldots, n \text{ and } A_i \cap A_j = \varnothing \text{ for } i \neq j \right\}$

# Visualisation of set theory



# And more properties

***Property :*** Let $\Omega$ be a finite sample space and $A, B \in \Omega$ two events such that $A \cap B = \varnothing$, then we have the following :

$$\mathbb{P}[A \cap B] = \mathbb{P}[A] + \mathbb{P}[B]$$

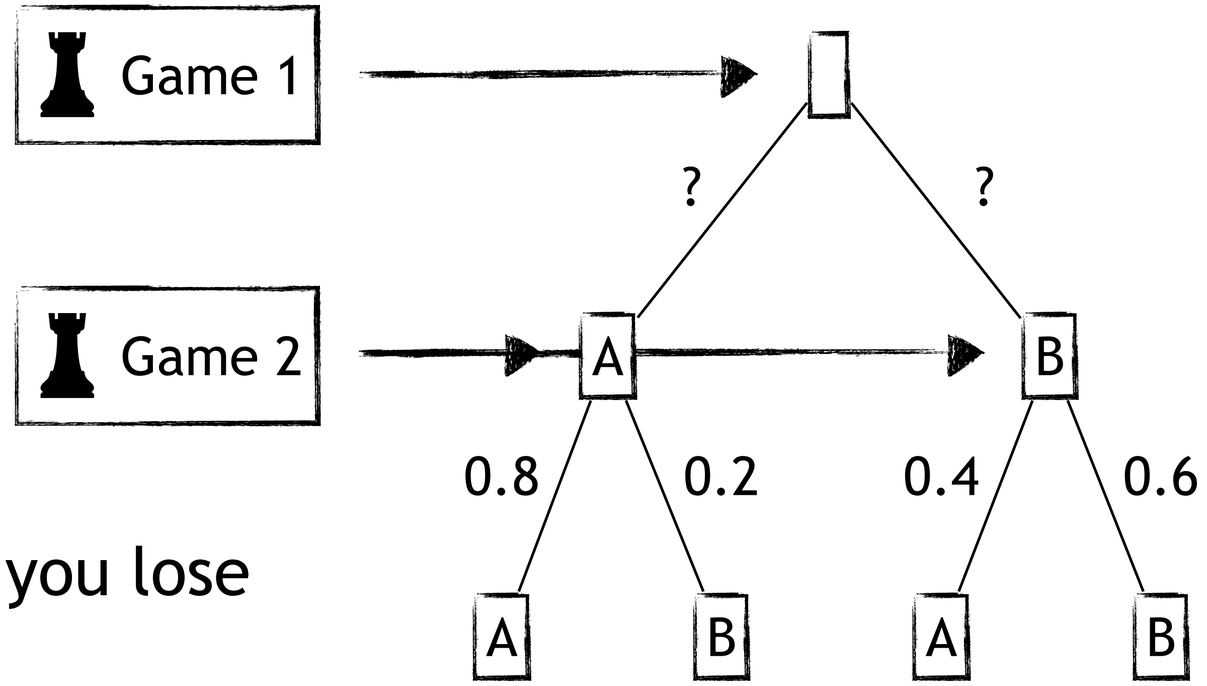
***Bayes' theorem :*** Let $\Omega$ be a finite sample space and $A, B \in \Omega$ two events such that $\mathbb{P}[B] > 0$, then we have the following :

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cup B]}{\mathbb{P}[B]}$$

# Let’s play chess !

You play a chess game against me and let’s define the following events :

- Event A : you win the game.
- Event B : I win the game.




Exo 1 - What is the probability for you to win a game after a lose ?

Exo 2 - Your probability to win both games is 0.56, what is the probability you lose the first game ?

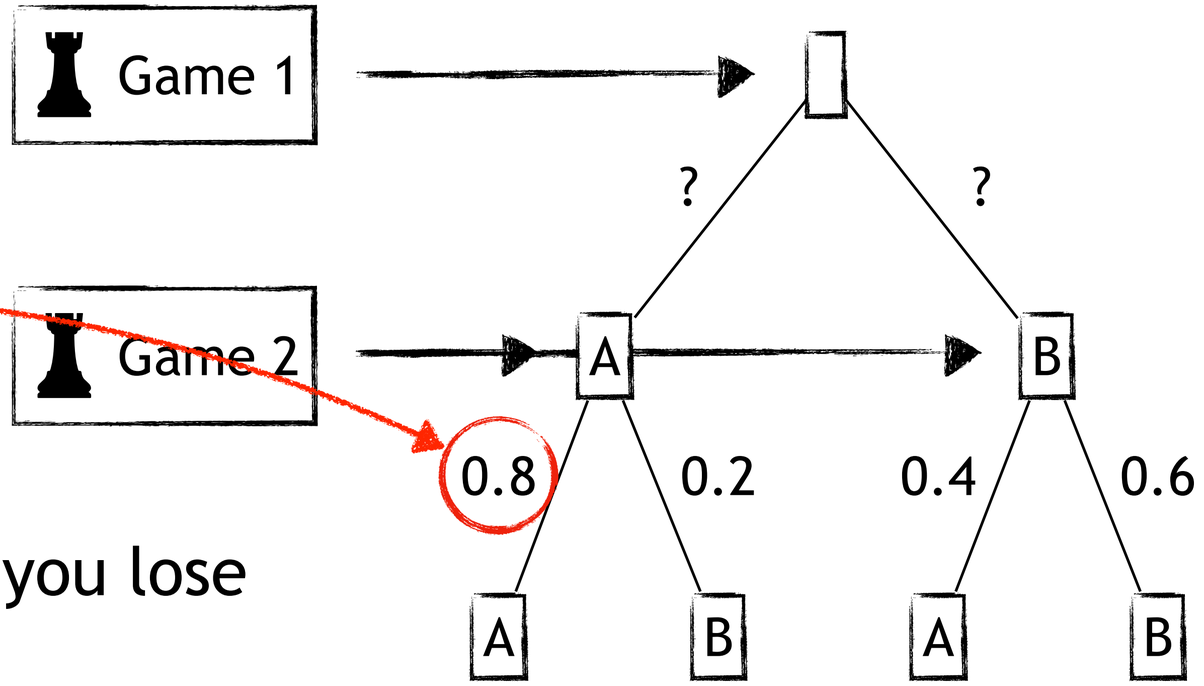
# Let’s play chess !

You play a chess game against me and let’s define the following events :

- Event A : you win the game.
- Event B : I win the game.

Exo 1 - What is the probability for you to win a game after a lose ?

Exo 2 - Your probability to win both games is 0.56, what is the probability you lose the first game ?



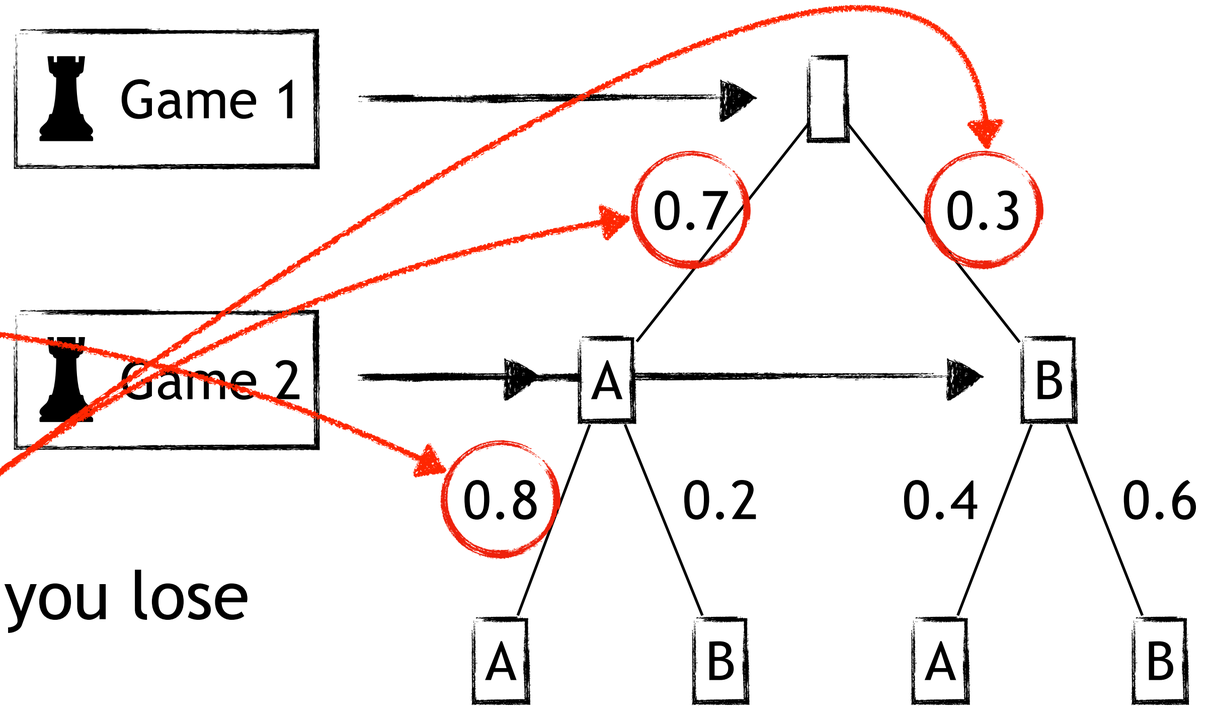
# Let’s play chess !

You play a chess game against me and let’s define the following events :

- Event A : you win the game.
- Event B : I win the game.

Exo 1 - What is the probability for you to win a game after a lose ?

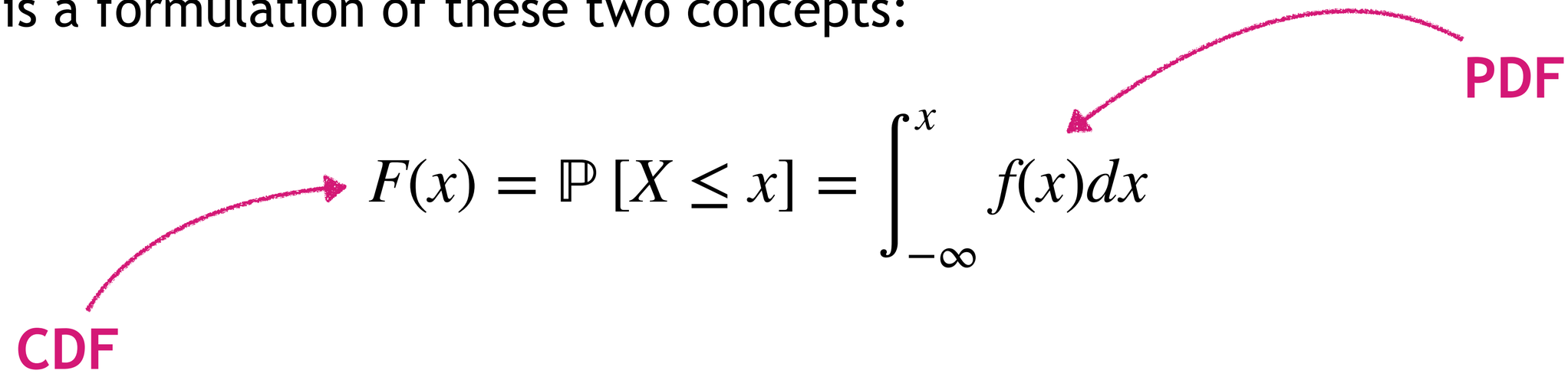
Exo 2 - Your probability to win both games is 0.56, what is the probability you lose the first game ?



# Last slide of theory !

In probability theory, there are two main concepts called the Cumulative Distribution Function (**CDF**) and the Probability Density Function (**PDF**), which describe how a random variable is theoretically defined.

Here is a formulation of these two concepts:



# Bernoulli distribution

Let’s take again the chess example:

- Event A : You win a game with p = 0.7
- Event B : You lose a game with q = 1 - p = 0.3

$$F(x) = \begin{cases} \mathbb{P}[A] = p \\ \mathbb{P}[B] = q = 1 - p \end{cases}$$

References

# Binomial distribution

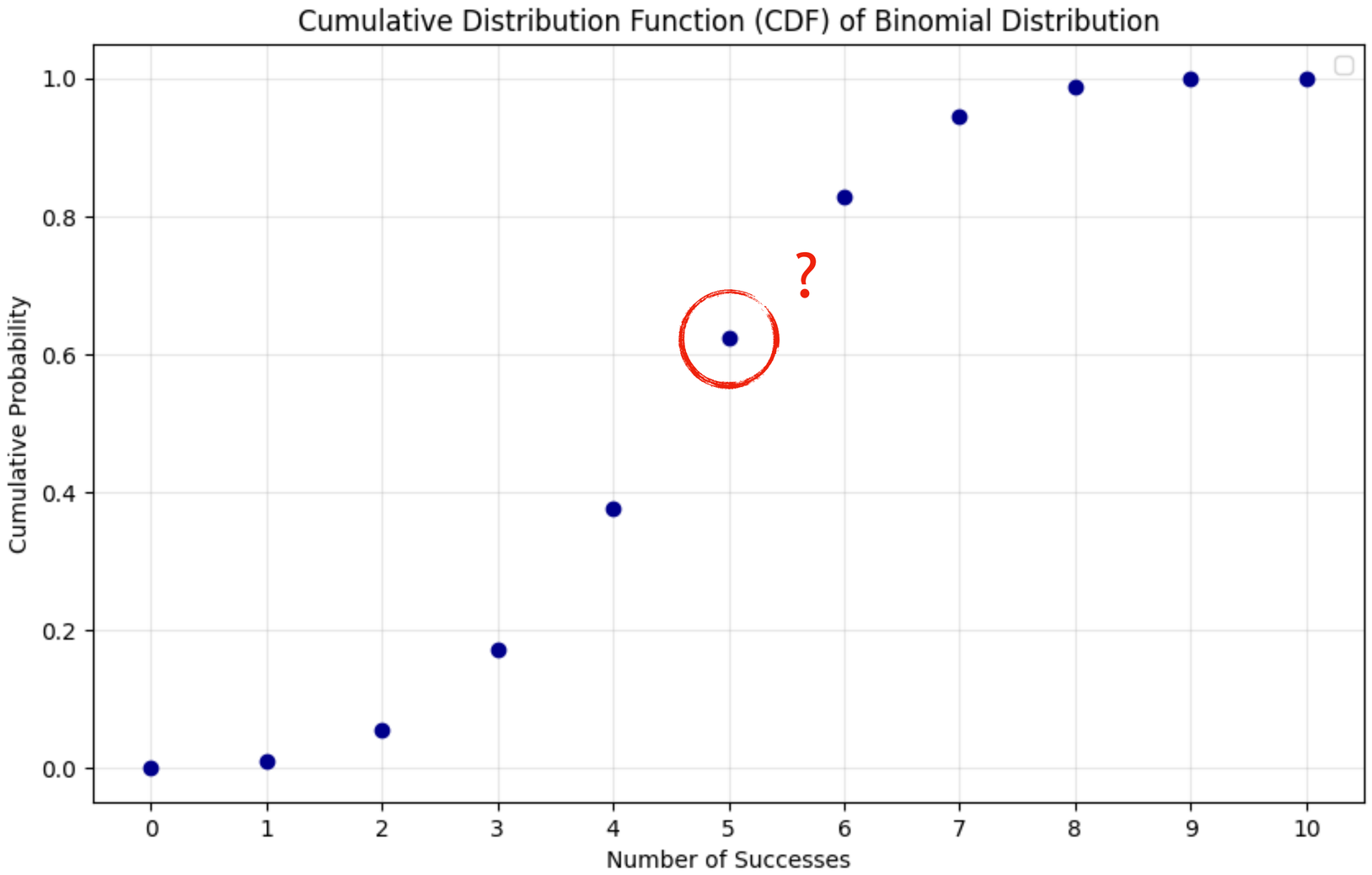
Let’s take again the chess example and you play $n$ games against me with the same probability after each game. What’s the probability you win k times ?

# Binomial distribution

Let’s take again the chess example and you play $n$ games against me with the same probability after each game. What’s the probability you win k times ?

$$\mathbb{P}\left[\text{Win } k \text{ times}\right] = \binom{n}{k} p^k (1-p)^{k-1}$$

# Binomial distribution



# Geometric distribution

Let’s suppose you don’t have the context: what kind of event $X$ could be represented by the following geometric distribution ?

$$\mathbb{P}\left[X = k\right] = (1-p)^{k-1}p$$

# Uniform distribution

Let’s consider the size of a randomly selected person in meters. For now, we consider a minimum at 1.60m and a maximum at 2.00m and there is an even chance to be in this range.

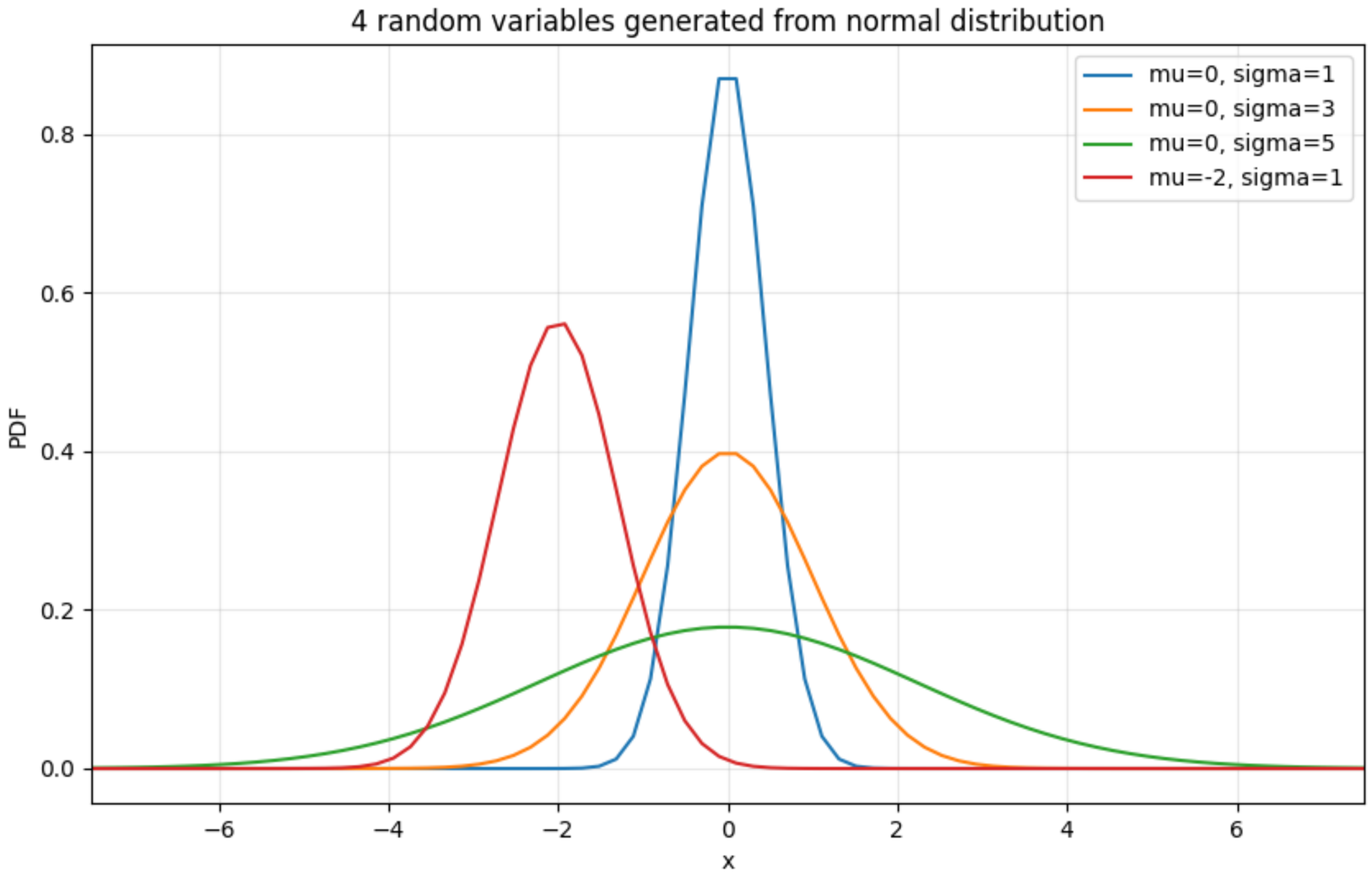
$$f(x) = \frac{1}{b-a}, \text{ for } x \in [a, b].$$

# Normal distribution

The distance of employees' houses from the company's main office can be represented by a normal distribution.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

# Normal distribution



# Exponential distribution

You are waiting for a bus to arrive and in average you know that a bus pass every 10 minutes.

- ✦ What would be the CDF and the PDF ?
- ✦ What is the probability you wait less than 2 minutes before a bus pass ?

# Exponential distribution

You are waiting for a bus to arrive and in average you know that a bus pass every 10 minutes.

- ✦ What would be the CDF and the PDF ?
- ✦ What is the probability you wait less than 2 minutes before a bus pass ?

$$\mathbb{E}[X] = \frac{1}{\lambda}$$

$$F(x) = \mathbb{P}[X \leq x] = \begin{cases} 1 - e^{-\lambda x} & \text{, if } x \geq 0 \\ 0 & \text{, otherwise} \end{cases}$$

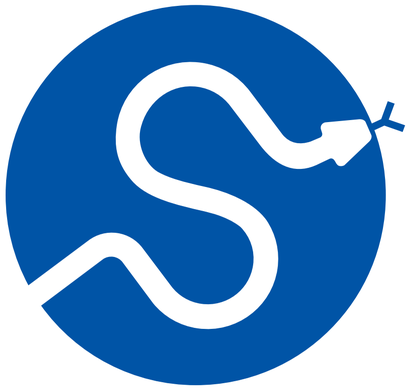
# Introduction to SciPy

*Name ?* - SciPy stands for Scientific Python

*What ?* - Open source library built-on NumPy

*Why ?* - Gives access to powerful tools such as:

- Statistics
- Signal processing -> for time series processing
- Linear algebra
- Test hypothesis
- and more…



# What are the main strengths of SciPy ?

- Easy-to-use scientific functions
- Optimised for performance and reliability
- Works perfectly with NumPy (and with pandas if you bridge your data through NumPy)
- Don’t waste time implementing functions yourself !
  - ✦ There’s a high probability that the function you want is already implemented in SciPy, NumPy or other libraries…

References

- Monday : Understand data structures
- Tuesday : Introduction to probability theory
- Wednesday : Central Limit Theorem confidence interval and test hypothesis
  - Central Limit Theorem
  - Confidence interval
  - Test hypothesis
- Thursday : Feature selection and correlation matrix
- Friday : Statistics with scikit-learn

- Monday : Understand data structures
- Tuesday : Numerical libraries (SciPy & Matplotlib)
- Wednesday : Introduction to probability theory
- Thursday : Confidence interval and test hypothesis
  - Confidence intervals
  - p-value
  - Tests hypothesis
- Friday : Statistics with scikit-learn

- Monday : Understand data structures
- Tuesday : Numerical libraries (SciPy & Matplotlib)
- Wednesday : Introduction to probability theory
- Thursday : Confidence intervals and test hypothesis
- Friday : Statistics with scikit-learn
  - Introduction to scikit-learn & methods
  - Model evaluation
  - Cross validation