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

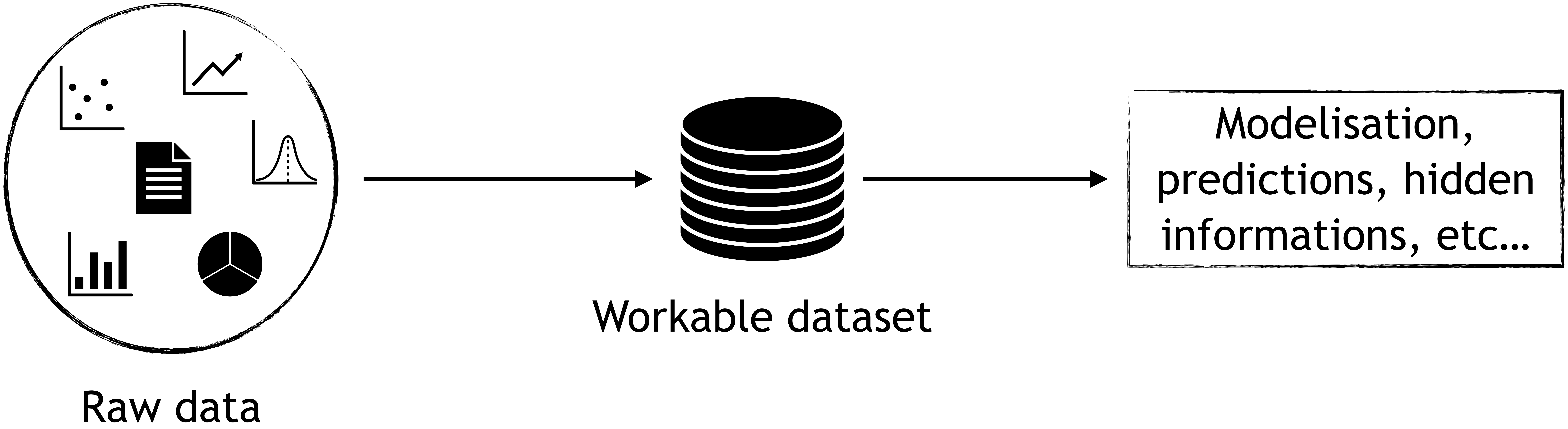
- Tuesday : Introduction to probability theory

- Wednesday : Central Limit Theorem confidence interval and test hypothesis

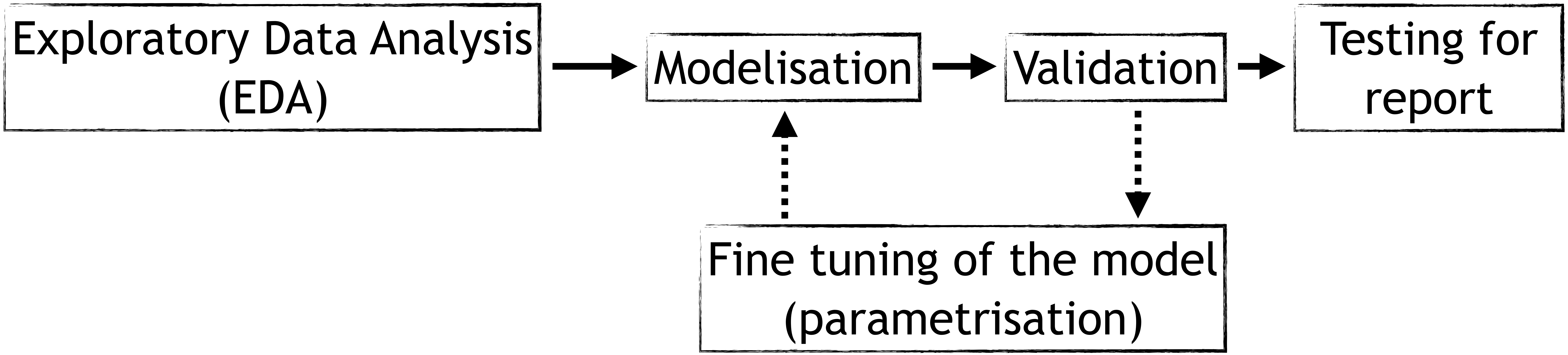
- Thursday : Feature selection and correlation matrix

- Friday : Statistics with scikit-learn

How a project is built ?



How a project is built ?



What’s descriptive statistics ?

<u>Definition :</u> descriptive statistics is about exploring and understanding a data set before going further into the modelisation.

<u>Remark :</u> 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 -

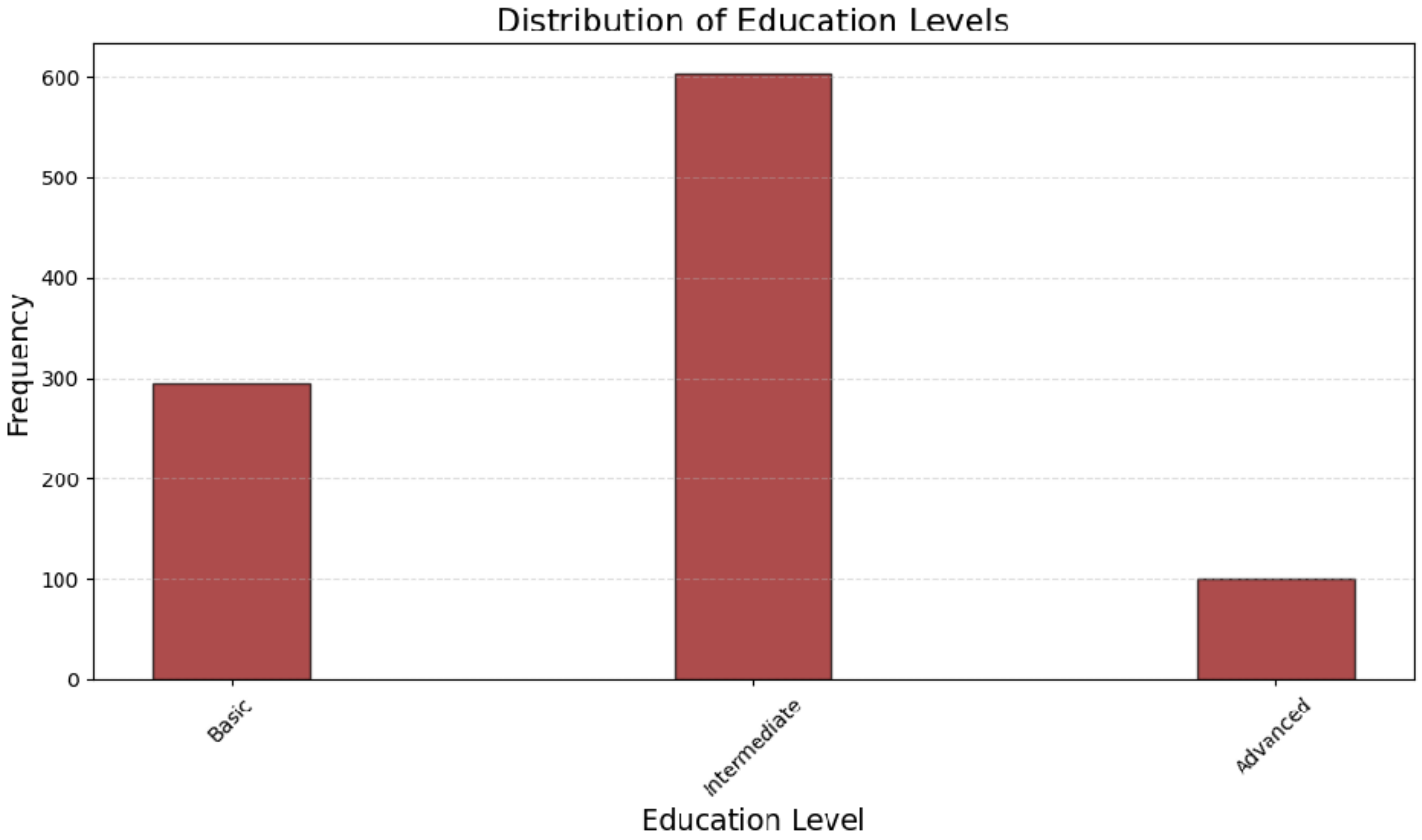
df.describe() is <u>your friend</u>
when you first see a data frame

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

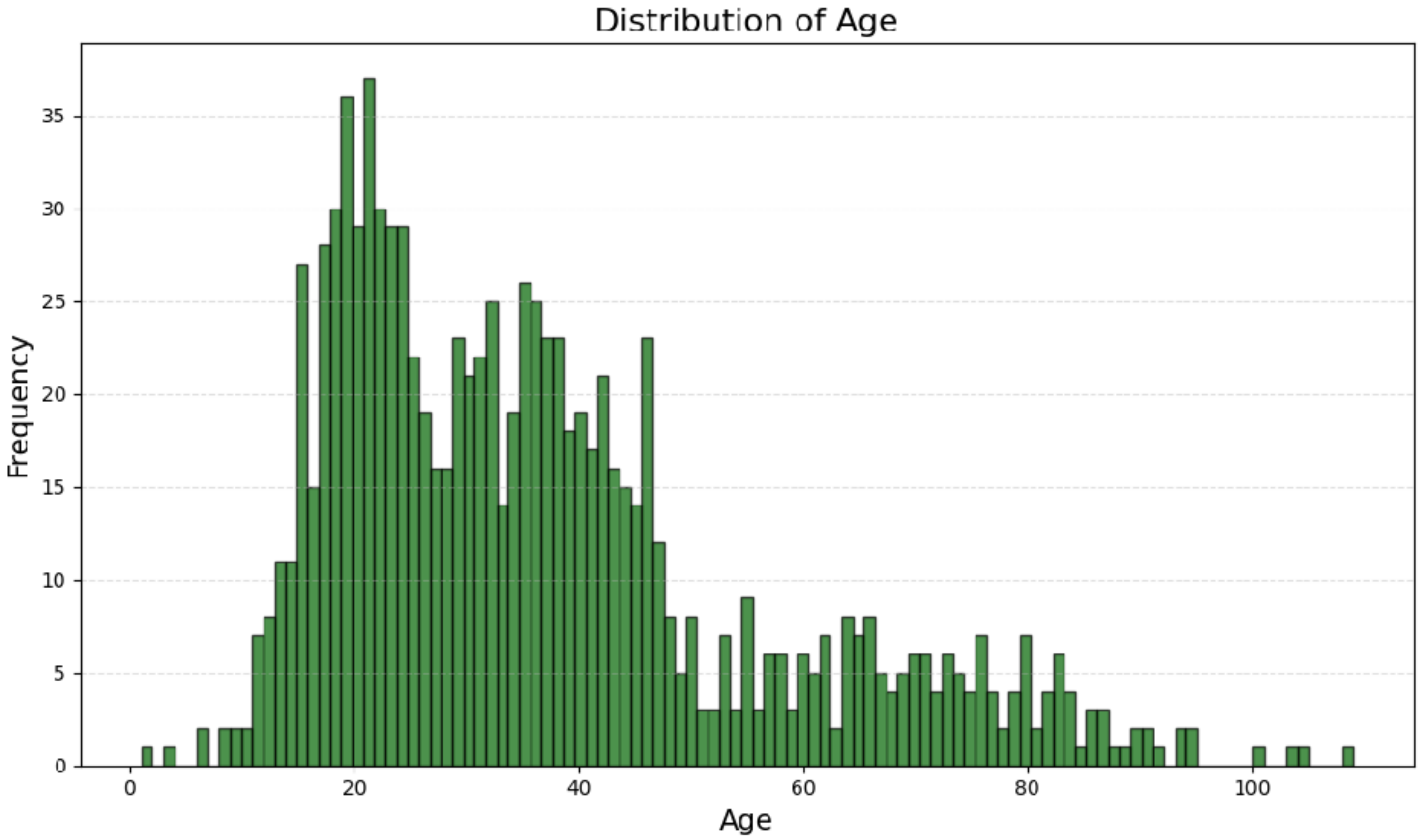


References

Key for analysis is Visualisation
- Univariate analysis -

2 - Univariate analysis for continuous variables

Histograms

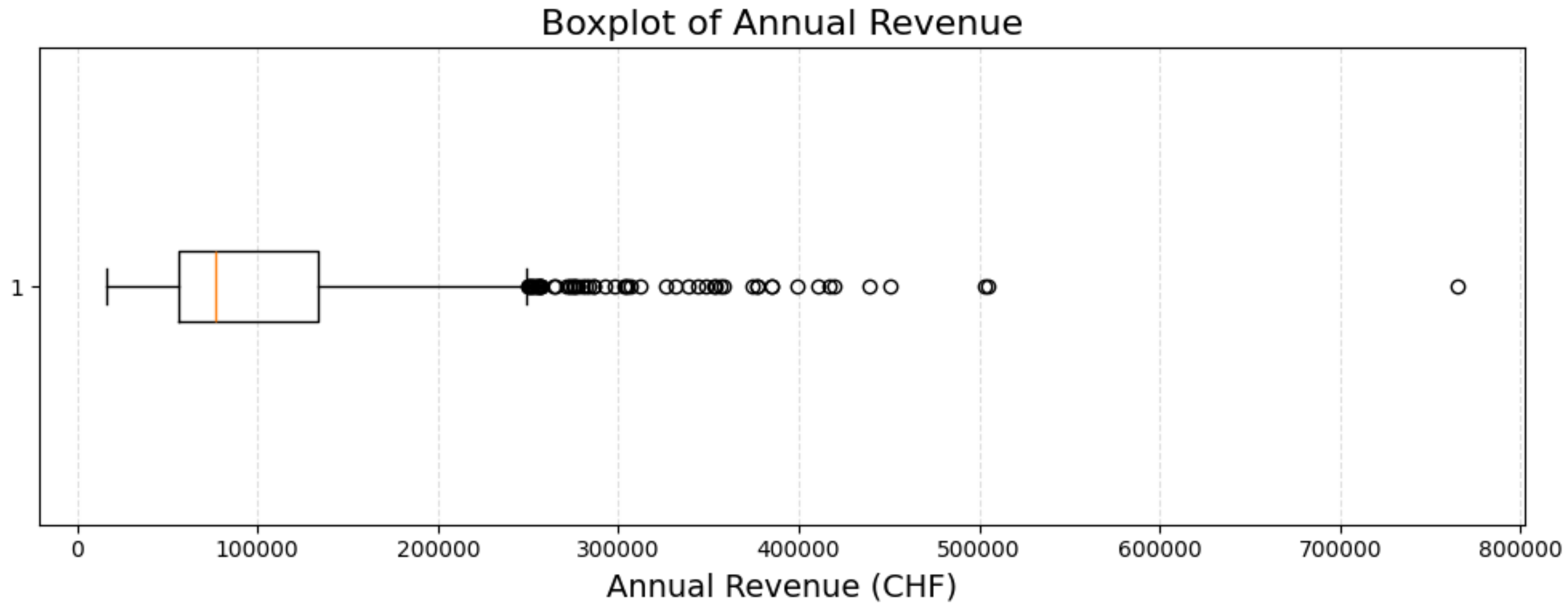


References

Key for analysis is Visualisation
- Univariate analysis -

3 - Univariate analysis for continuous variables

Boxplots

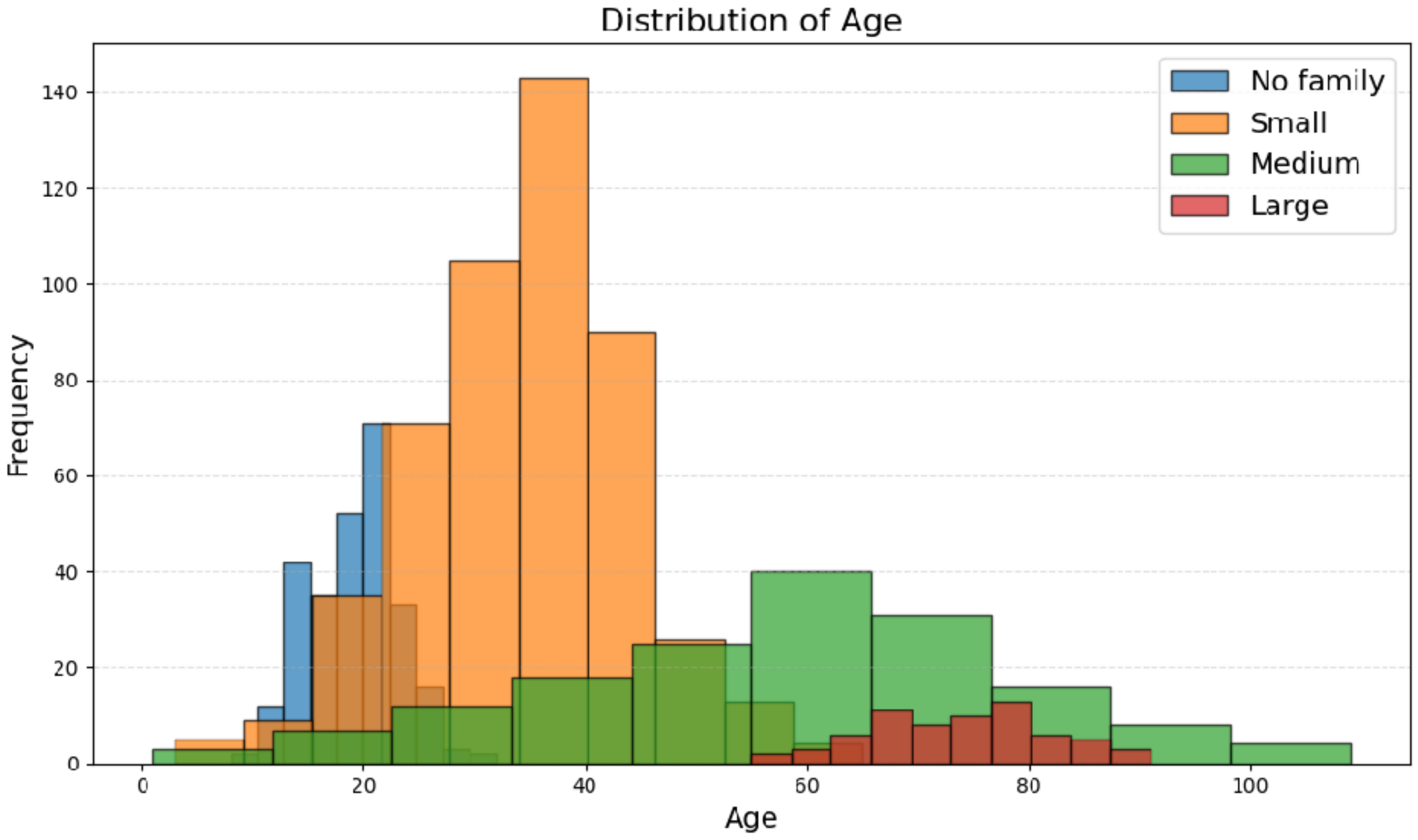


References

Key for analysis is Visualisation
- Multivariate analysis -

4 - Multivariate analysis for continuous and/or categorical variables

Multivariate

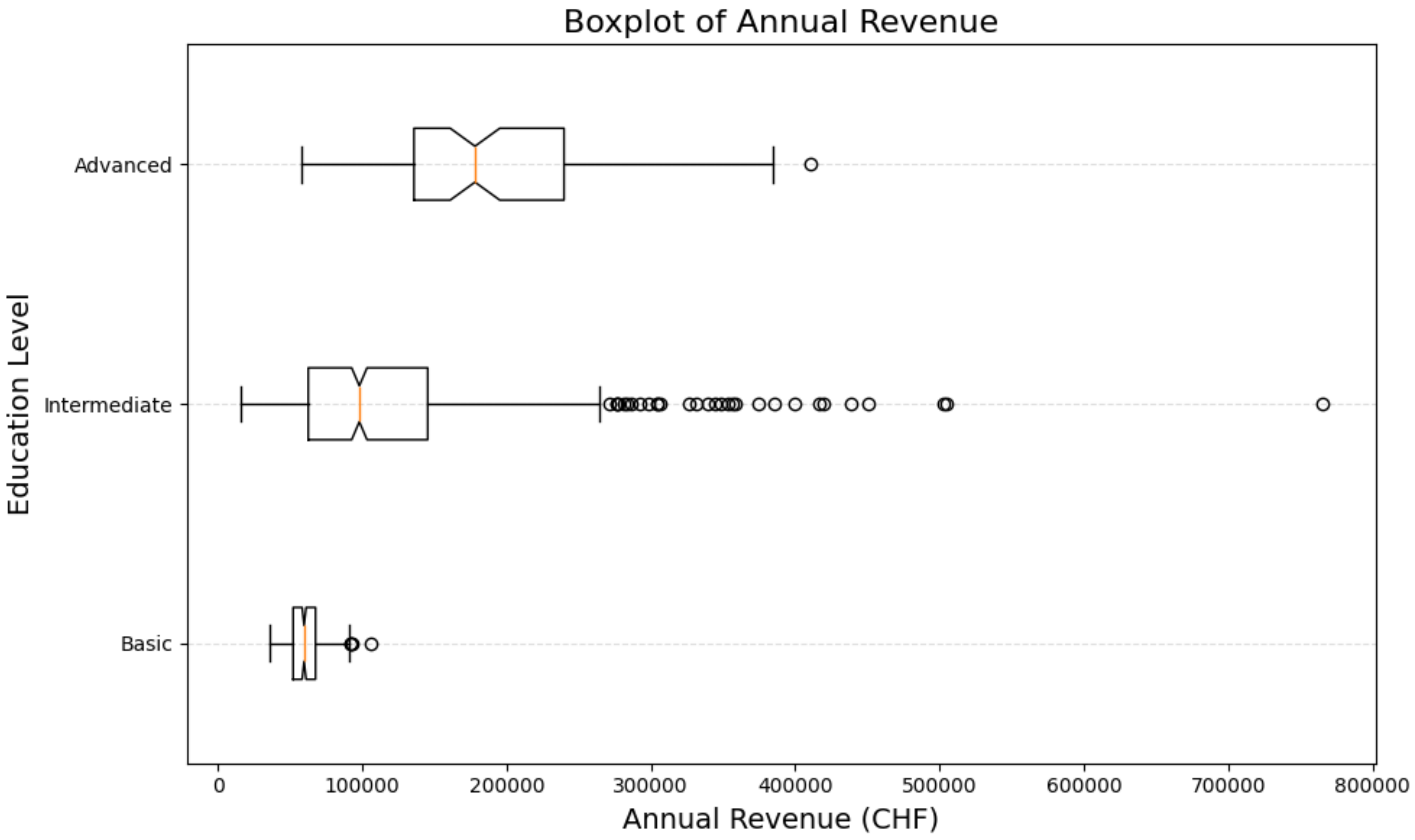


References

Key for analysis is Visualisation
- Multivariate analysis -

5 - Multivariate analysis for continuous and/or categorical variables

Bivariate boxplots

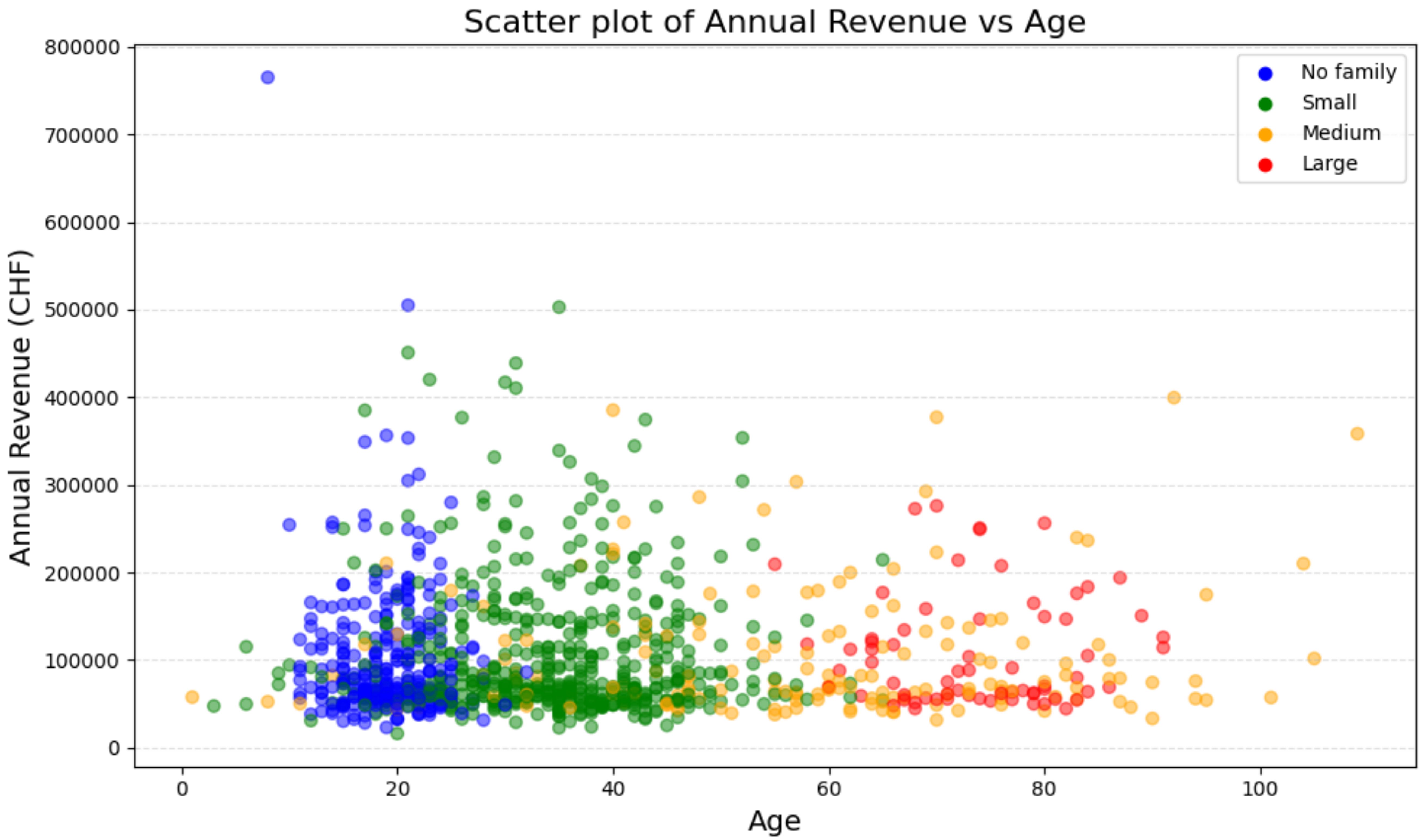


References

Key for analysis is Visualisation
- Multivariate analysis -

6 - Multivariate analysis for continuous and/or categorical variables

<u>Colored scatter plots</u>

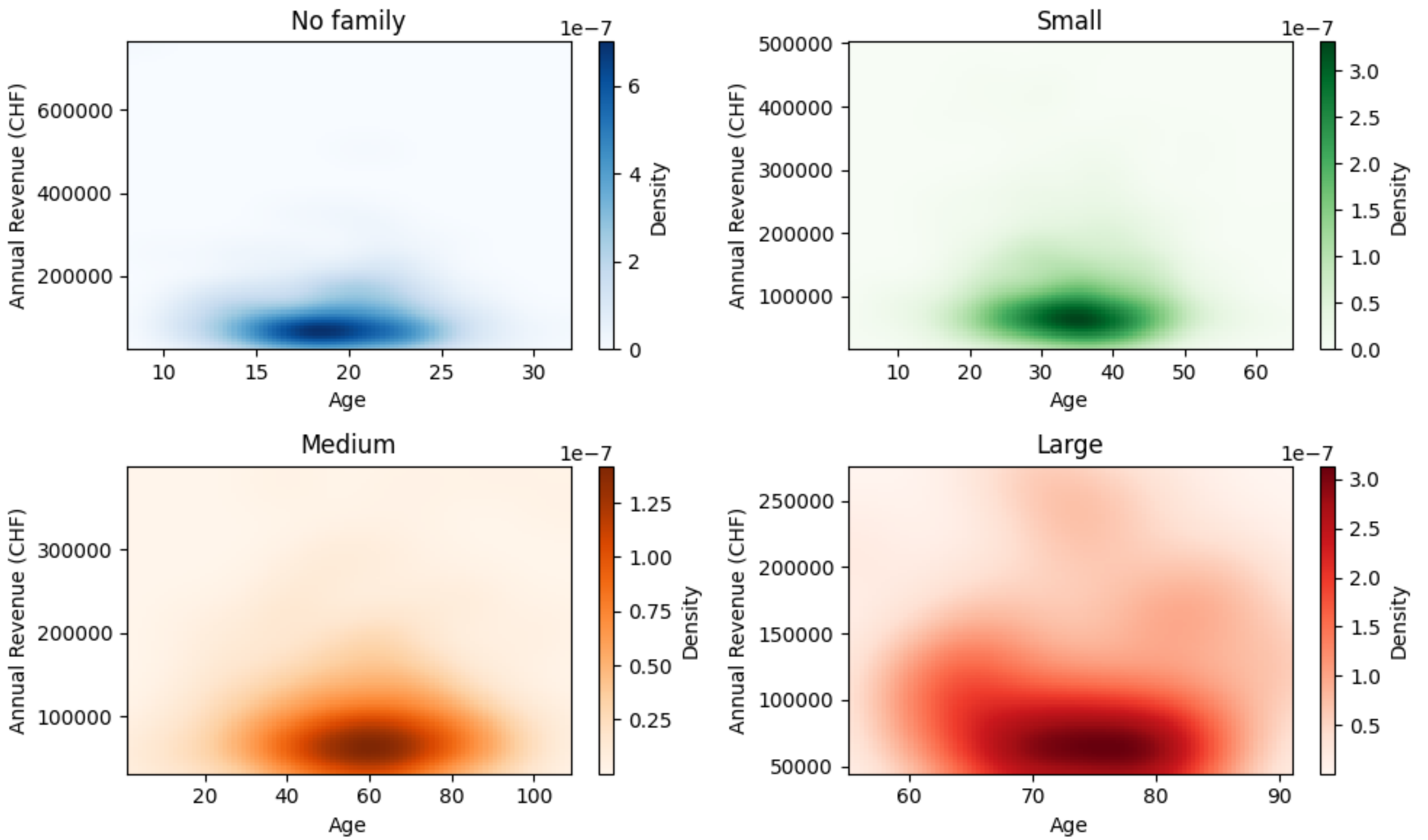


References

Key for analysis is Visualisation
- Multivariate analysis -

6 - Multivariate analysis for continuous and/or categorical variables

<u>Colored scatter plots</u>



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

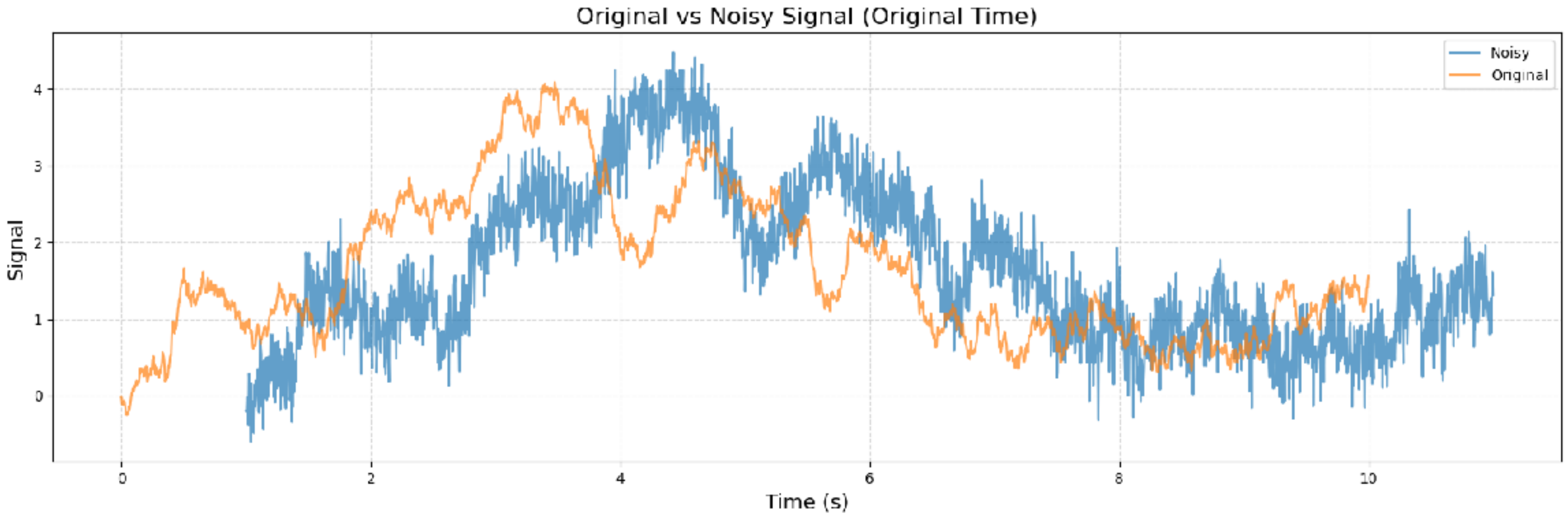
Some normalisation formulas

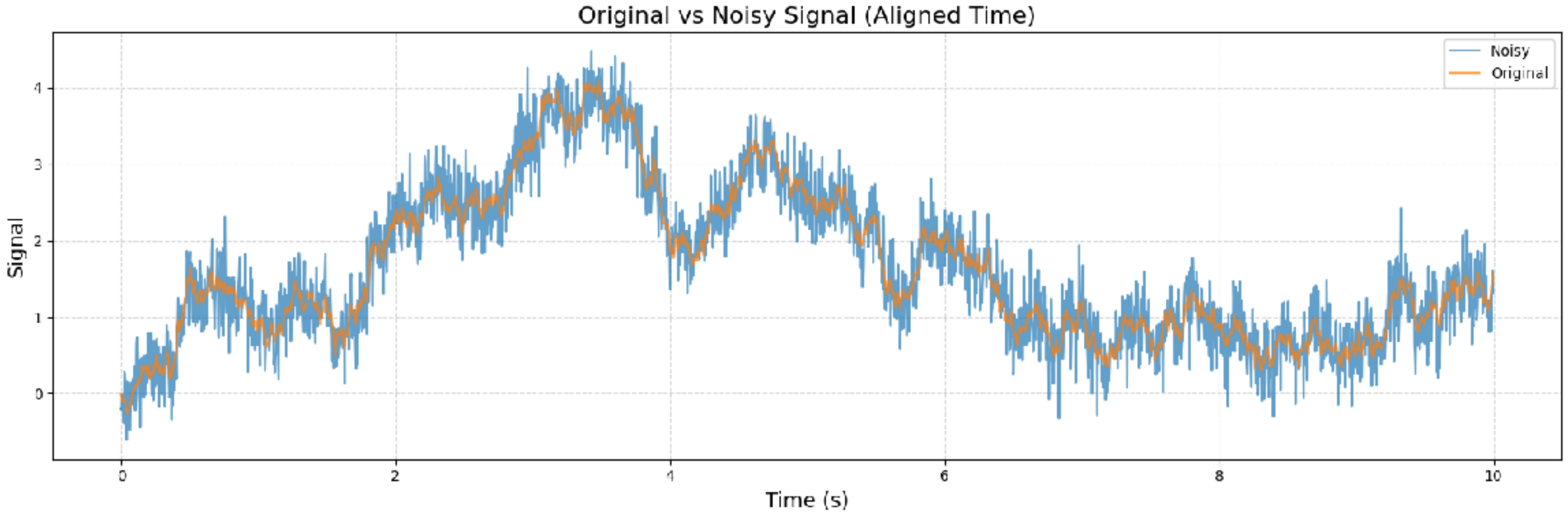
1 - <u>*Min-max normalisation :*</u> $\tilde{x}_i = \dfrac{x_i - \text{min}(X)}{\text{max}(X) - \text{min}(X)}$

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

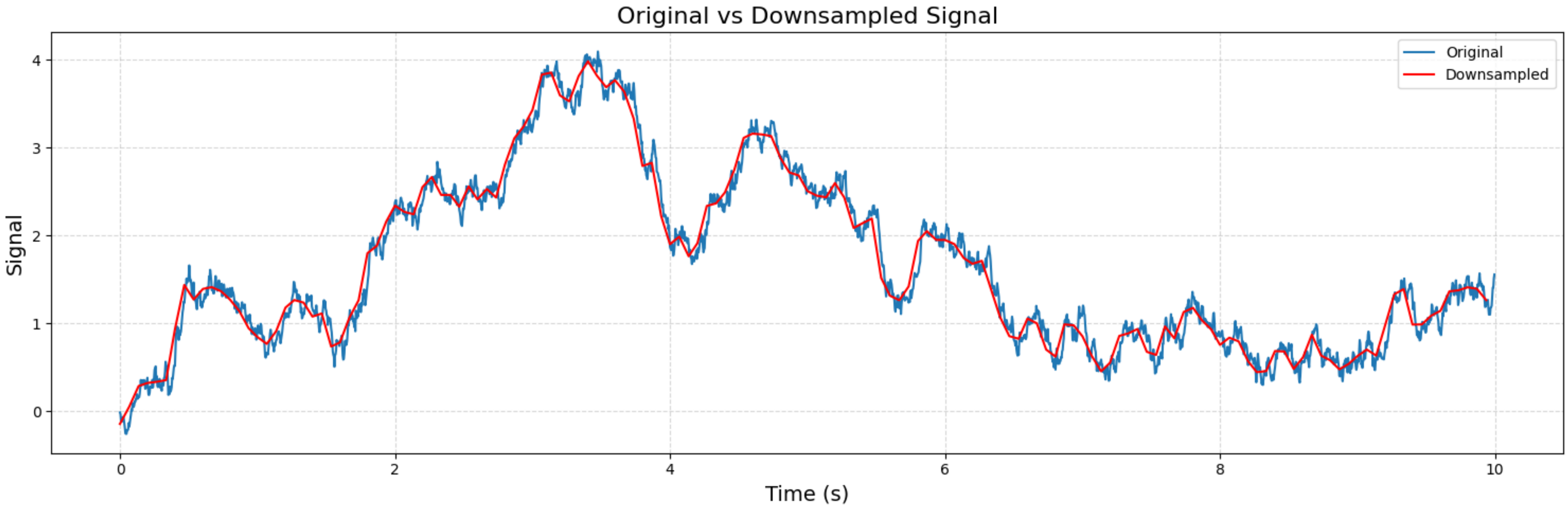
3 - <u>*log-normalisation :*</u> $\tilde{x}_i = \text{log}(x_i + 1)$ or $\tilde{x}_i = \text{log}(x_i)$

Synchronisation of time series

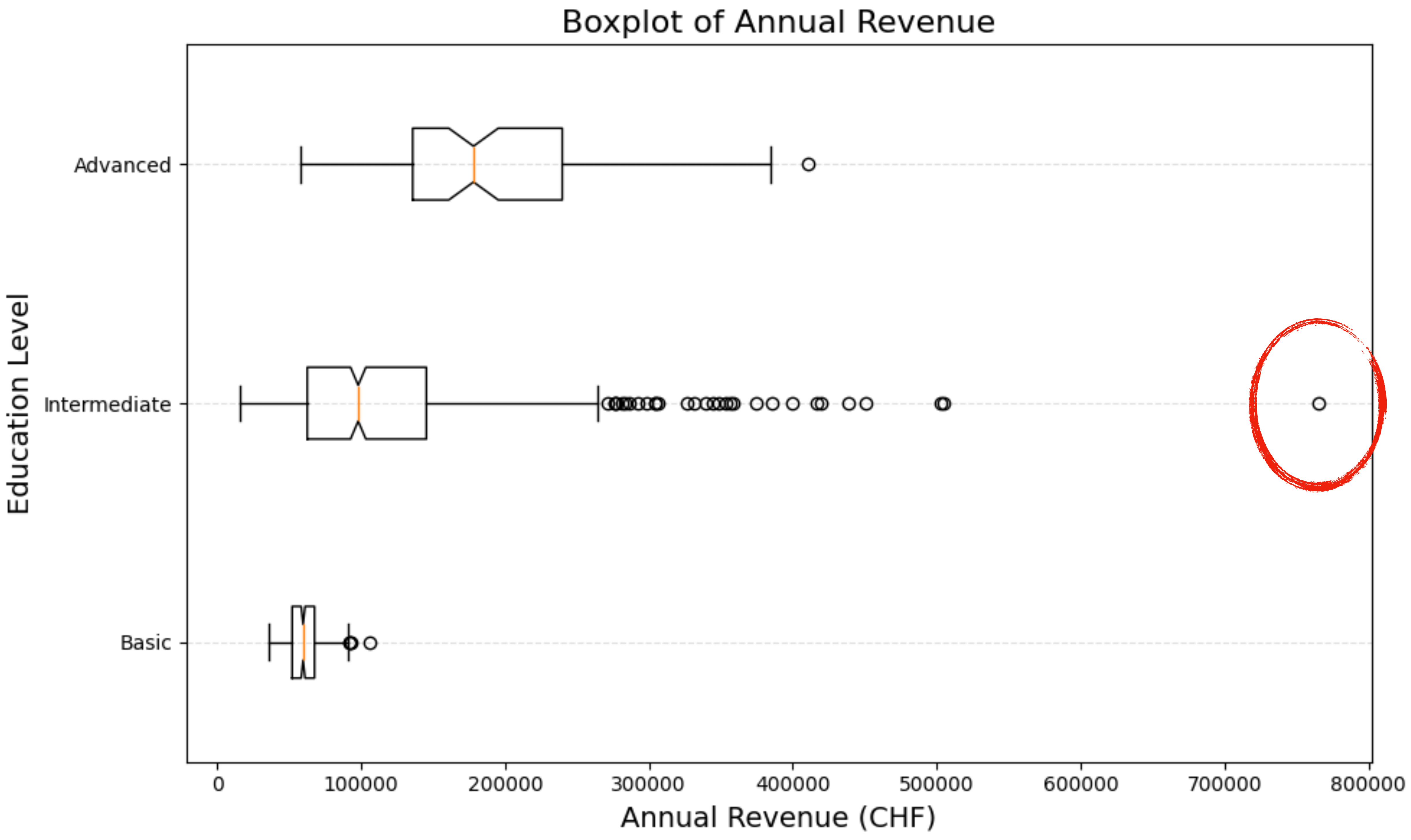




Re-sampling of time series



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

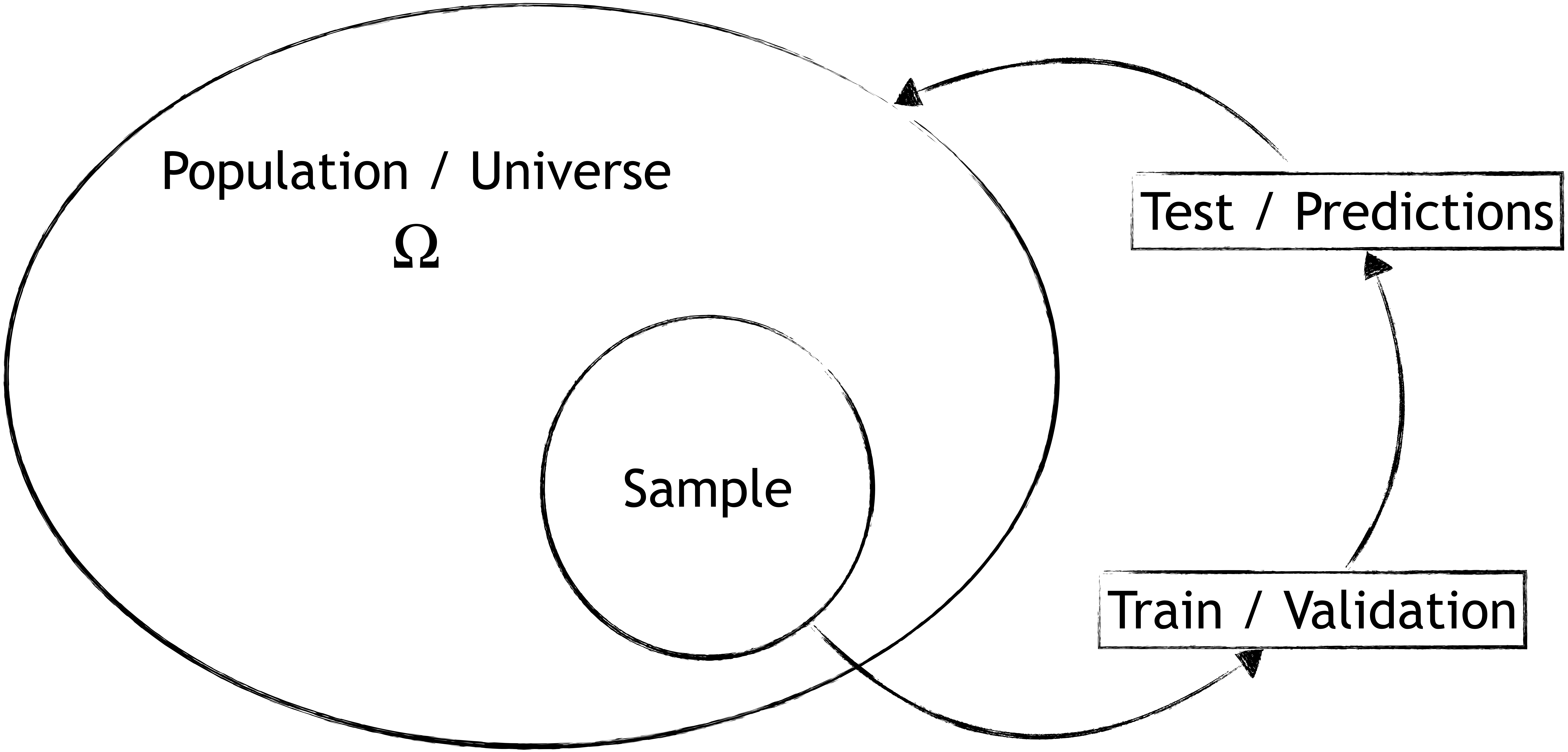
- Tuesday : Introduction to probability theory
  - Fundamental definitions
  - Probability law / distribution (discrete)
  - Discrete vs continuous probability

- 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

❖ **<u>Universe :</u>** denoted as Ω, it represents the <u>complete set</u> of all possible elements or outcomes you are studying.

❖ **<u>Sample :</u>** it represents a <u>subset of the universe</u>, selected for analysis.

❖ **<u>Event :</u>** it represents a <u>set of outcomes</u> from an experiment.

Probability of an event

<u>Definition :</u> Let Ω 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

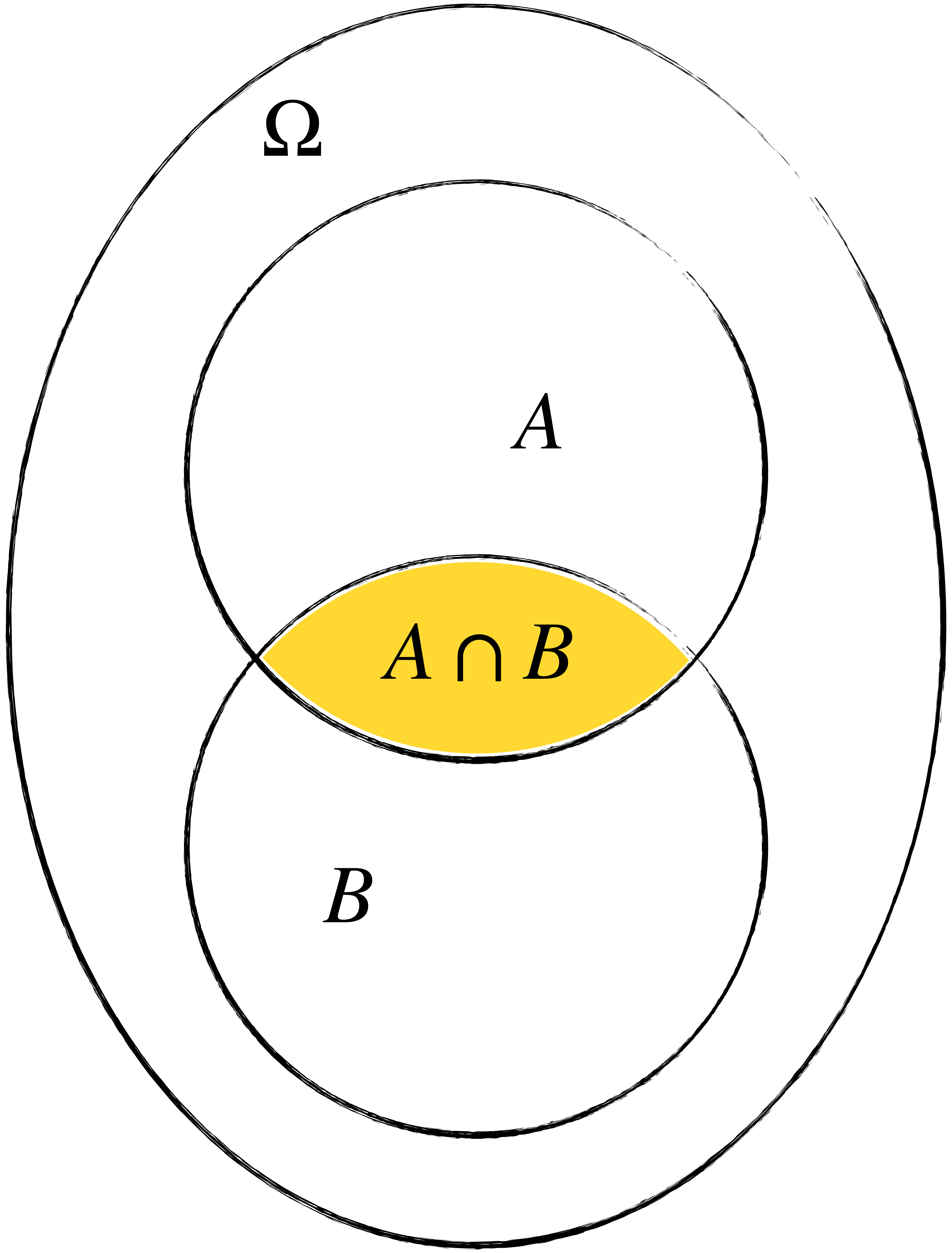
Property : Let Ω be a finite sample space, then we have the following properties:

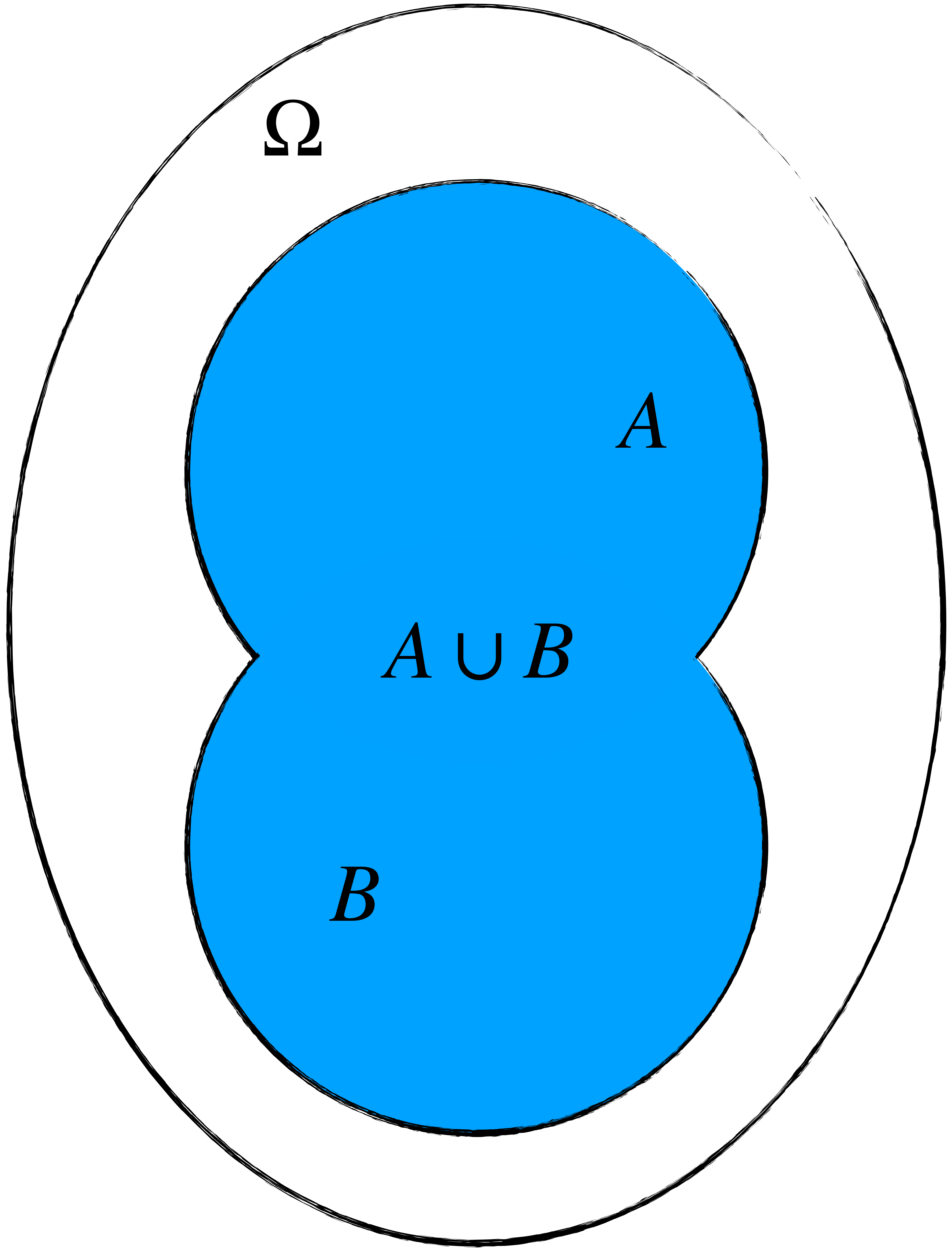
- $\mathbb{P}[\Omega] = 1$ and $\mathbb{P}[\varnothing] = 0$

- 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 Ω be a finite sample space and $A, B \in \Omega$ two events such that $A \cap B = \emptyset$, then we have the following :

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

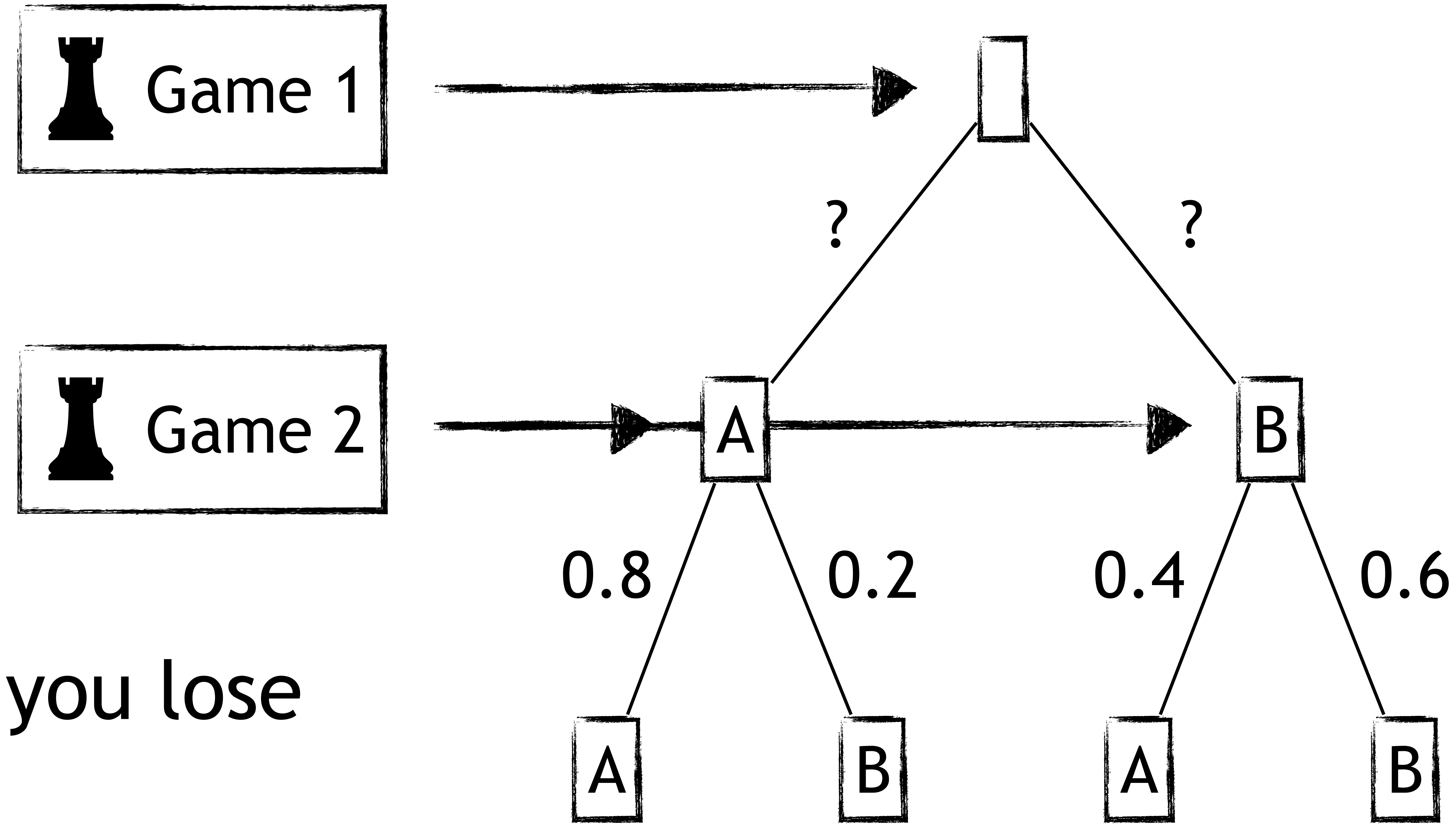
Bayes' theorem : Let Ω be a finite sample space and $A, B \in \Omega$ two events such that $\mathbb{P}[B] > 0$, then we have the following :

$$\mathbb{P}[A \,|\, B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$$

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.



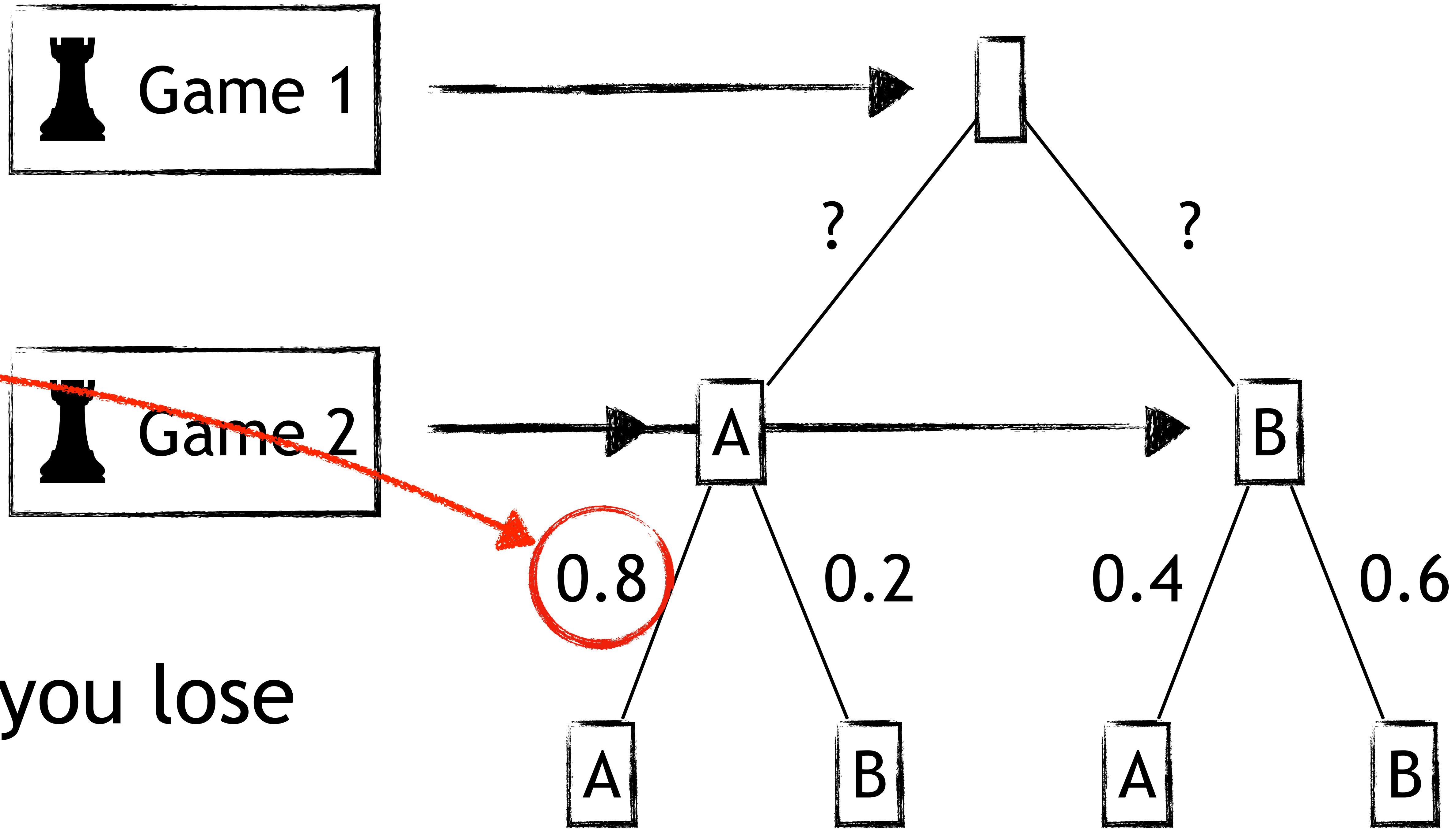
Exo 1 - What is the probability for you to win a game after a lose ?

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 ?

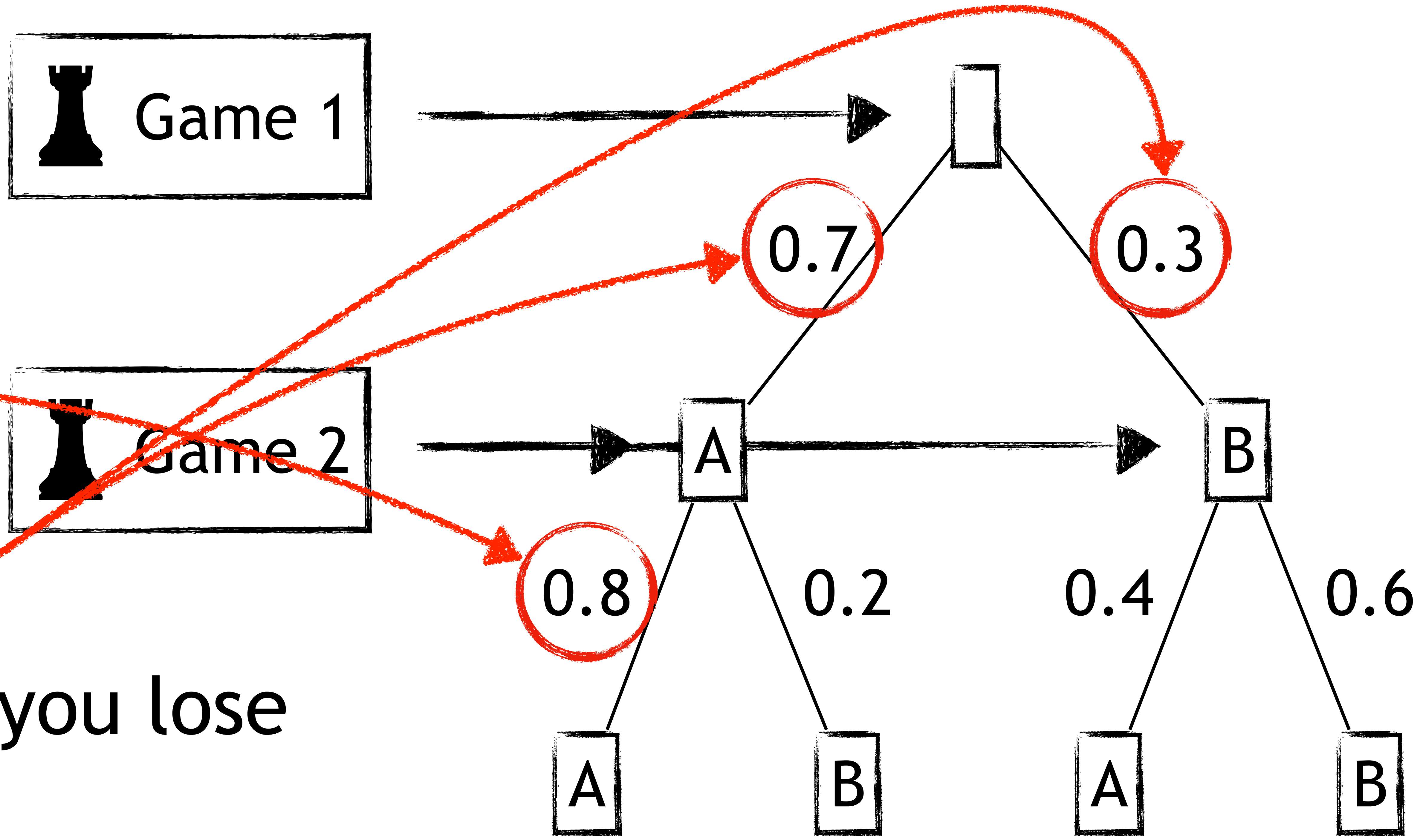
Let's play chess !

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:

PDF

$$F(x) = \mathbb{P}\left[X \leq x\right] = \int_{-\infty}^{x} f(x)dx$$

CDF

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

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

References

Binomial distribution

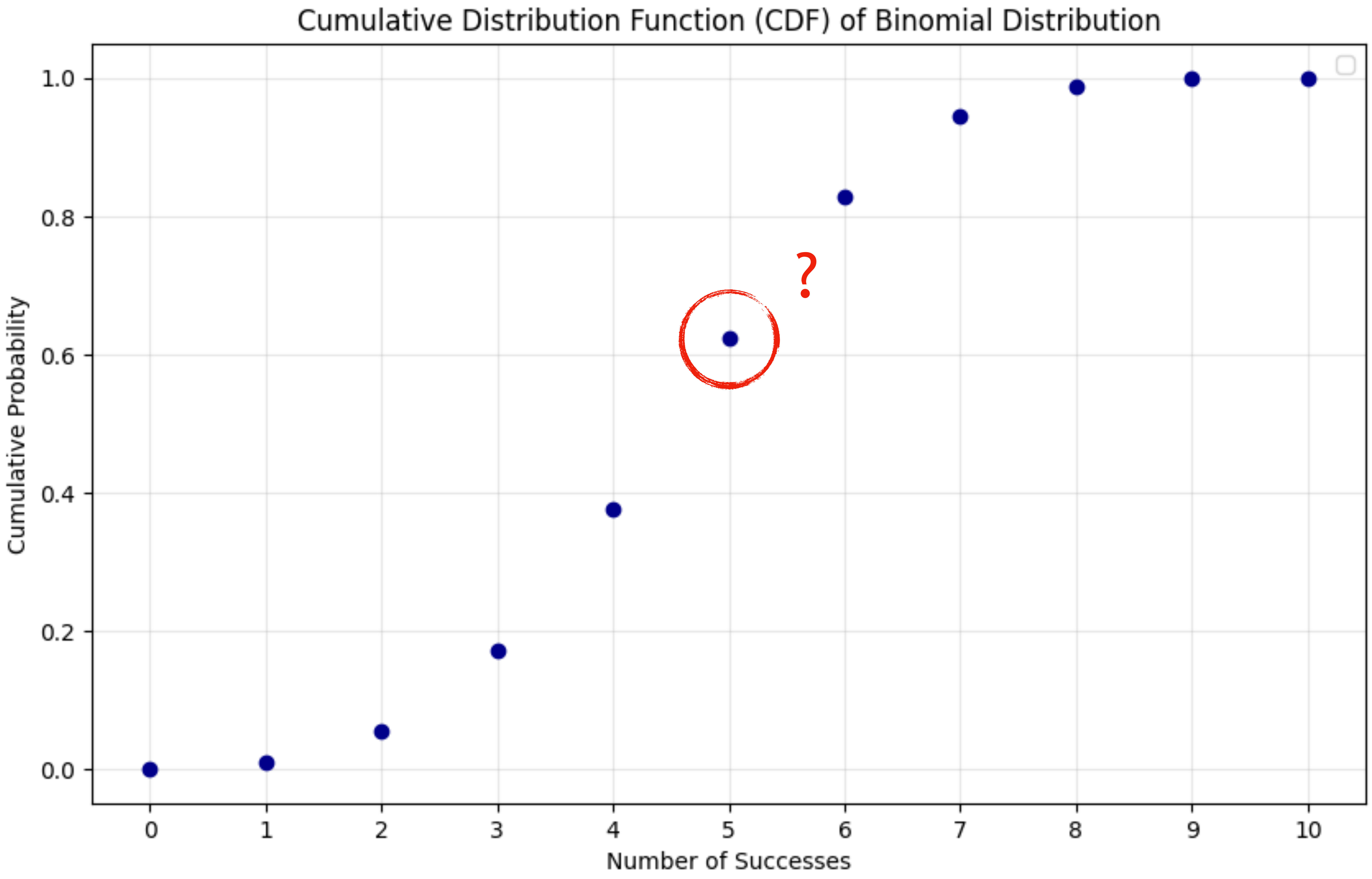
Let's take again the chess example and you play n games against me with the same probability after each game. What's the probability you win k times ?

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 ?

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

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 ?

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

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.

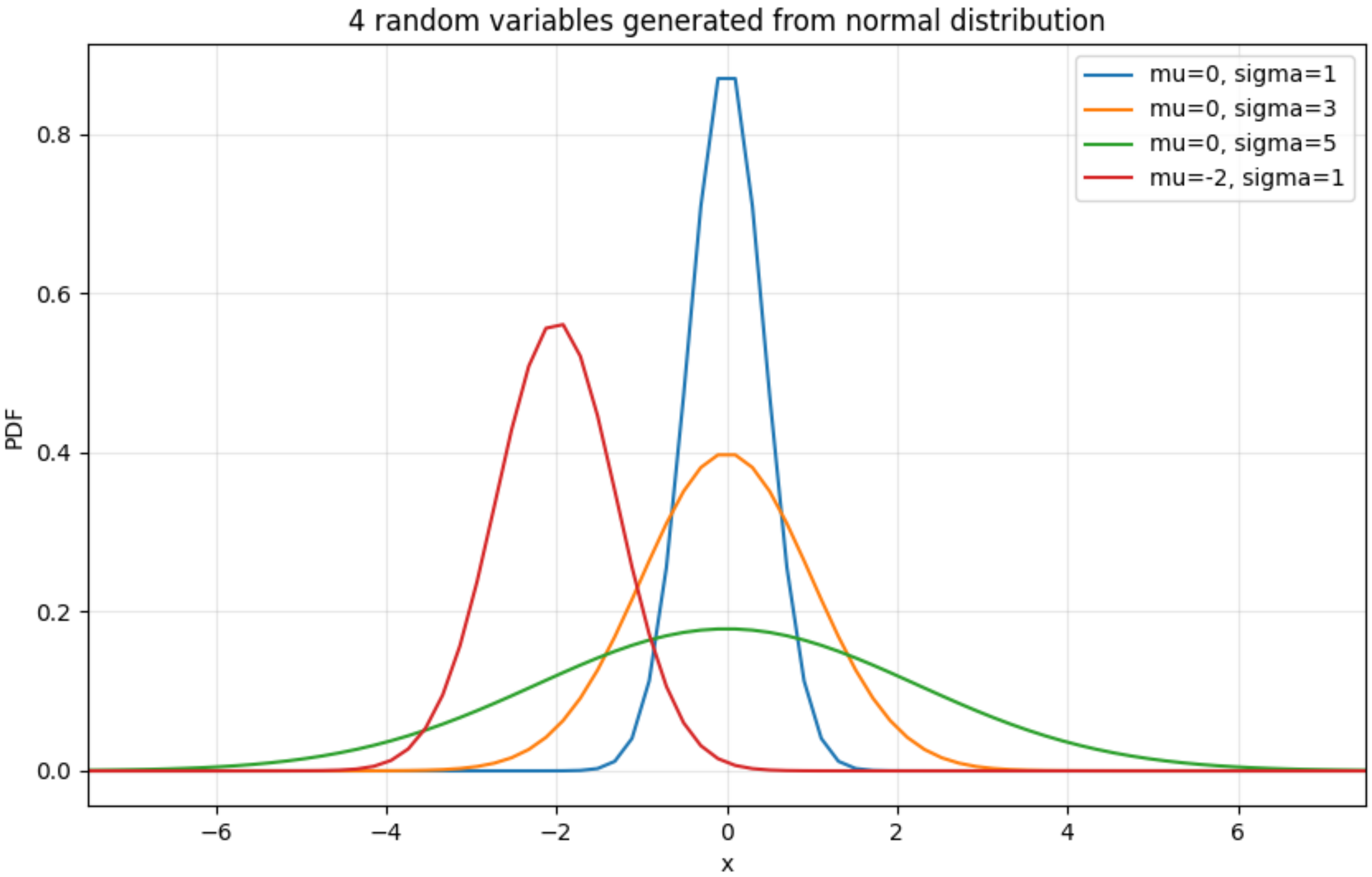
$$\boxed{f(x) = \frac{1}{b-a}, \text{ for } x \in [a, b].}$$

Normal distribution

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

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

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 ?

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

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

Introduction to SciPy

<u>Name ?</u> - SciPy stands for Scientific Python

<u>What ?</u> - Open source library built-on NumPy

<u>Why ?</u> - 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…

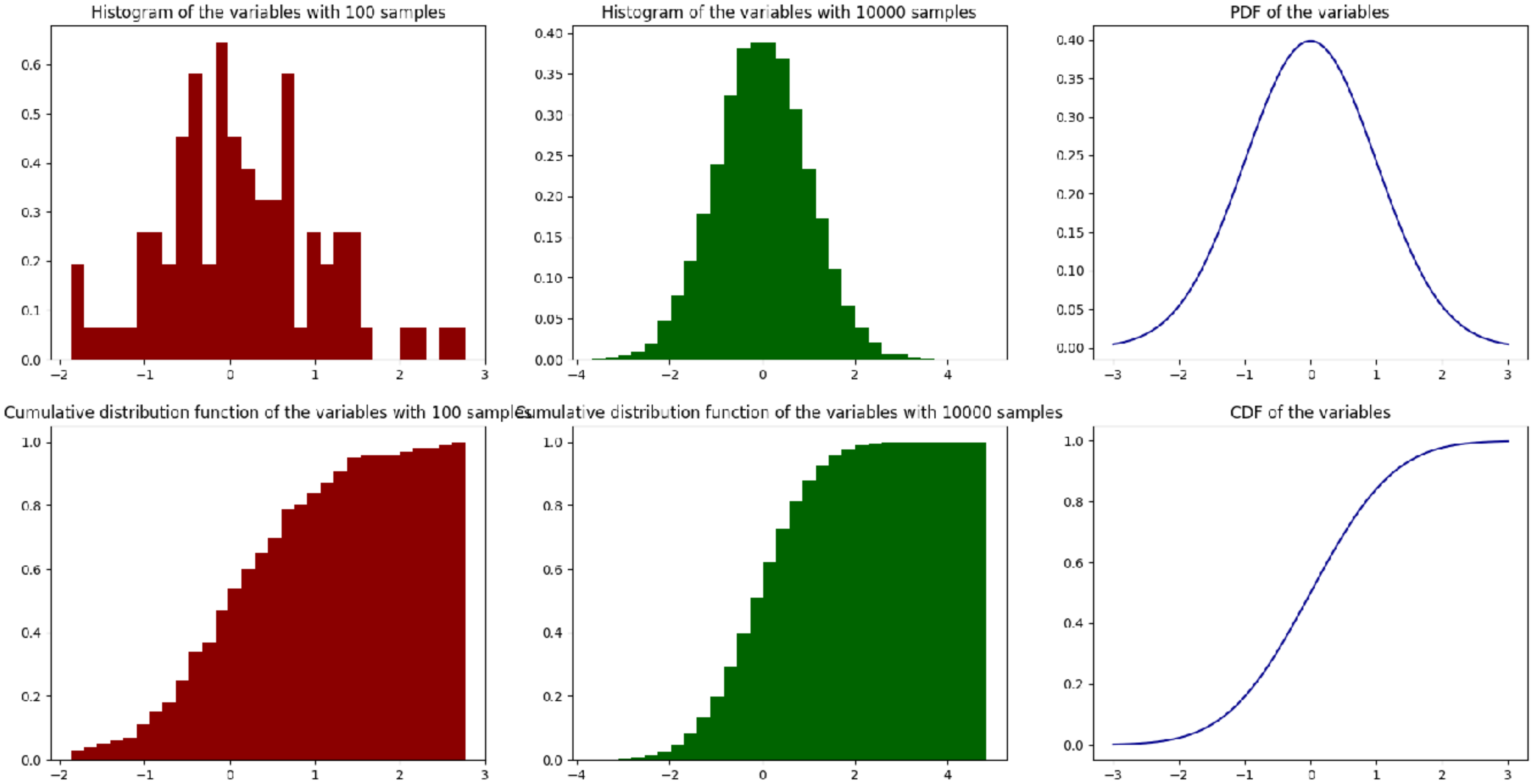
References

- 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

Small reminder on the difference between PDF and CDF

$$\boxed{F(x) = \mathbb{P}\left[X \leq x\right] = \int_{-\infty}^{x} f(x)dx}$$

Small reminder on the difference between PDF and CDF



Central Limit Theorem

<u>Question 1:</u> You have a fair 6-face dice and X represents the result after a roll of the dice as a random variable. What is the distribution of X ?

Central Limit Theorem

Question 1: You have a fair 6-face dice and X represents the result after a roll of the dice as a random variable. What is the distribution of X ?

Answer : a discrete uniform distribution ! There are two types of uniform distributions : continuous and discrete.

Central Limit Theorem

Still with the same fair dice, you roll it 5-times. So, you have 5 new random variables we note X_1, X_2, X_3, X_4 and X_5, which are all respectively discrete uniform distribution.

We note the average of the results as: $\bar{X} = \frac{1}{5}\sum_{i=1}^{5} X_i$.

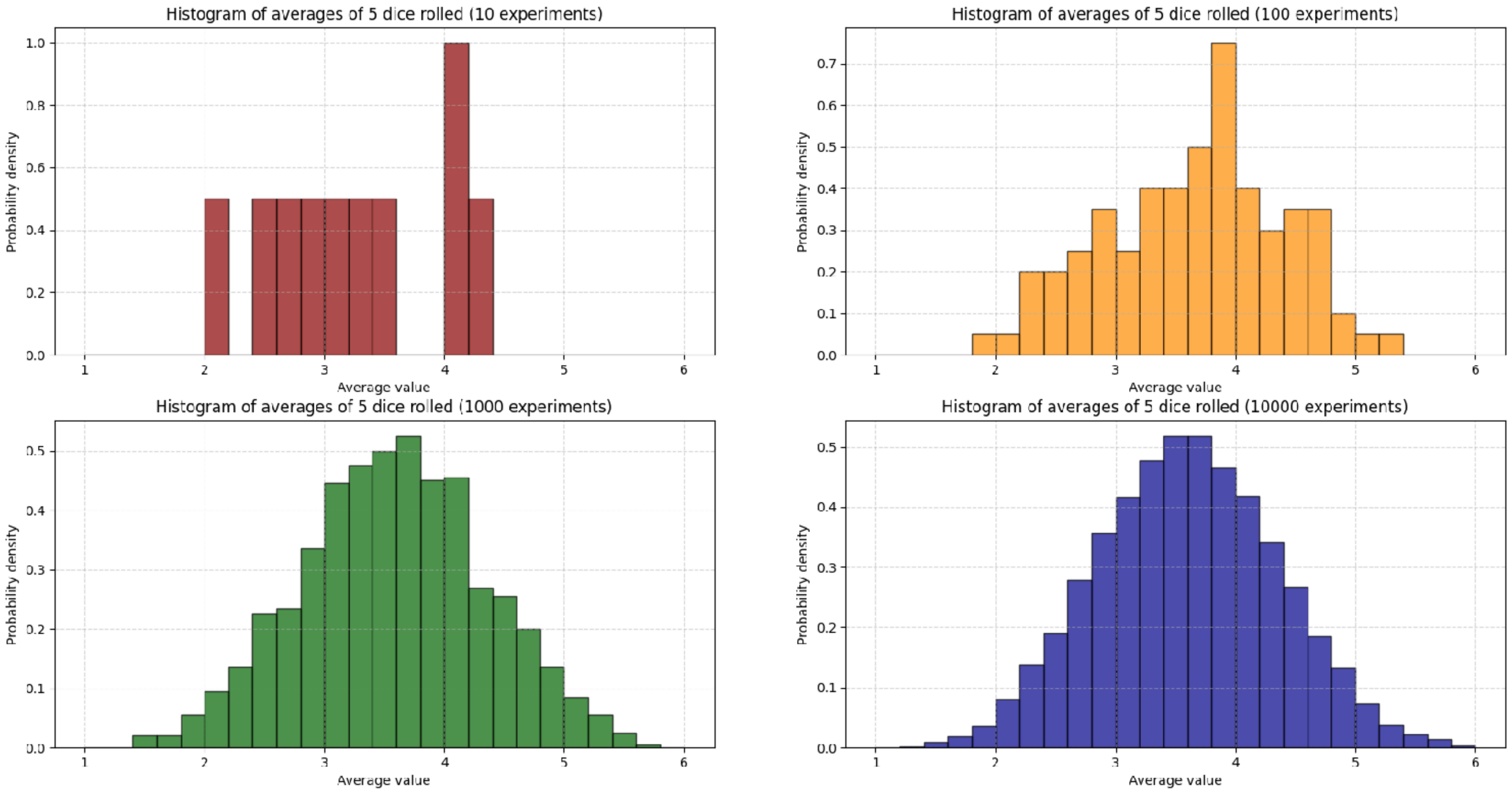
Question 2: Using the library SciPy, built a function that returns the average of the experiment describe above.

Central Limit Theorem

Now, you have a new random variable we note $\bar{X}$, corresponding to the average result of rolling a fair dice 5 time (uniform distribution).

<u>Question 3:</u> What happens if we repeat this operation 10 times? 100 times? 1,000 times? 10,000 times?

Central Limit Theorem



Central Limit Theorem

<u>Central Limit Theorem (CLT) :</u> Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables, each with mean μ and finite variance σ^2. Then, as n becomes large, the following normalized random variable:

$$Z_n = \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}$$

converges in distribution to a <u>standard normal distribution</u> (i.e. centrée et réduite), where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

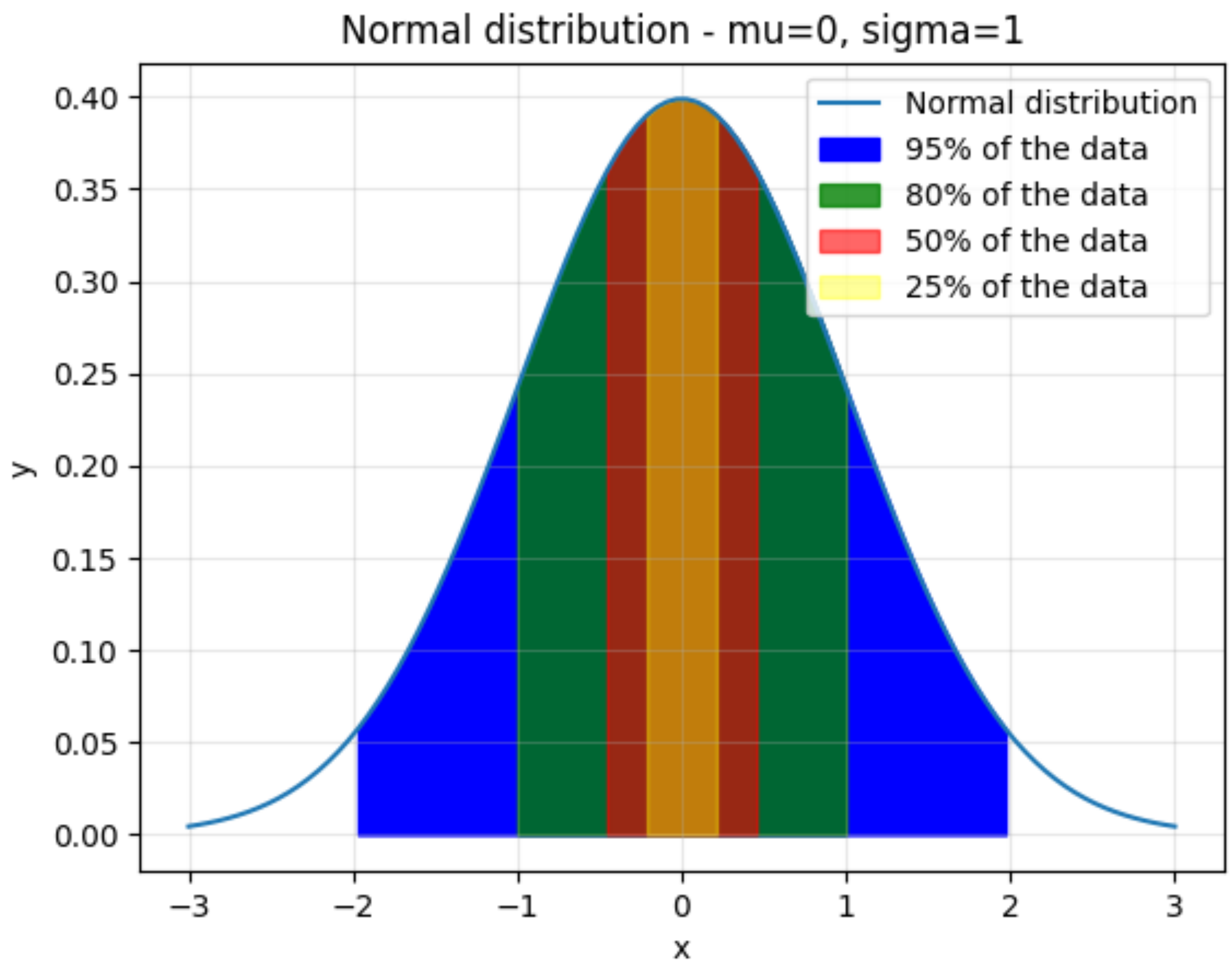
Confidence intervals

Let's recall the formula $Z_n = \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}$ and we know by the CLT that this distribution is standard normal.

If we note z_κ the value such that κ % of the possible values Z are less than z_κ, then we have:

$$\mathbb{P}\left[-z_{\alpha/2} \leq Z_n \leq z_{\alpha/2}\right] \approx 1 - \alpha$$

Confidence intervals



$$\mathbb{P}\left[-z_{\alpha/2} \leq Z_n \leq z_{\alpha/2}\right] \approx 1 - \alpha$$

Confidence intervals

$$\mathbb{P}\left[-z_{\alpha/2} \leq Z_n \leq z_{\alpha/2}\right] \approx 1 - \alpha$$

$$\mathbb{P}\left[-z_{\alpha/2} \leq \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \leq z_{\alpha/2}\right] \approx 1 - \alpha$$

$$\mathbb{P}\left[\bar{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \approx 1 - \alpha$$

Confidence intervals

Definition : Following the concepts introduced before, we define a $1 - \alpha\%$ confidence interval $C_{1-\alpha}$ for an estimator of μ such as:

$$C_{1-\alpha} = \left[\bar{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}; \bar{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right],$$

assuming σ is known and n is large enough to let the CLT apply.

Confidence intervals

Interpretation : « We are at $1 - \alpha\,\%$ confident that the true population mean μ lies within the interval $C_{1-\alpha}$. »

$$C_{1-\alpha} = \left[\bar{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} ; \bar{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Confidence intervals

Interpretation : « We are at $1 - \alpha\,\%$ confident that the true population mean μ lies within the interval $C_{1-\alpha}$. »

$$C_{1-\alpha} = \left[\bar{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} ; \bar{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Question : How do you compute the z-score ?

Confidence intervals

Interpretation : « We are at $1 - \alpha\,\%$ confident that the true population mean μ lies within the interval $C_{1-\alpha}$. »

$$C_{1-\alpha} = \left[\bar{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} ; \bar{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Question : How do you compute the z-score ?

Answer : Use SciPy ! (clue: norm.ppf)

Confidence intervals

Question : Now you know how to compute your z-score, use the formula below to compute the $90\,\%$, $95\,\%$ and $99\,\%$ confidence intervals of the distribution `normal_ci.csv` knowing that the standard deviation is equal to 6.

$$C_{1-\alpha} = \left[\bar{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} ; \bar{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Confidence intervals

The formula presented before suppose we know exactly the parameter of the standard deviation σ, which is not case in practice.

What is the solution ?

Confidence intervals

The formula presented before suppose we know exactly the parameter of the standard deviation σ and n is large enough, which is not always the case in practice.

What is the solution ?

=> Use an estimator of the standard deviation with the sample standard deviation s you can compute directly from the data you have.

Confidence intervals

By using an approximation of the standard deviation, you can no longer use the normal distribution to determine the confidence interval.

You have to use the Student's distribution.

Confidence intervals

You want the Student's distribution CDF and PDF ?

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

$$F(t) = 1 - \frac{1}{2} I_{\frac{\nu}{\nu+t^2}}\left(\frac{\nu}{2}, \frac{1}{2}\right) \quad \text{for } t \geq 0$$

Confidence intervals

Question : Now you know how to compute your z-score, use the formula below to compute the $90\,\%$, $95\,\%$ and $99\,\%$ confidence intervals of the distribution `exponential_ci.csv` using the student's distribution because you don't know the standard deviation σ.

$$C_{1-\alpha} = \left[\bar{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} ; \bar{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Test hypothesis

<u>Null hypothesis :</u> It's a baseline assumption, usually there is no effect or no difference.

$$H_0 : \mu = \mu_0$$

<u>Alternative hypothesis :</u> What we want to test for, typically that there is an effect or a difference.

One sided - $H_1 : \mu \neq \mu_0$

Tow sided - $H_1 : \mu > \mu_0$ *or* $\mu < \mu_0$

Test hypothesis

The test statistic presented before (normal and student's) measures how far the sample mean is from the hypothesized population mean under H_0, in units of standard error.

This test allows us to decide wether you reject or not the hypothesis H_0.

Now, how to compute the test statistic under the hypothesis H_0 ?
=> We need to talk about p-values.

p-values

<u>Definition :</u> The p-value is the probability, under the null hypothesis H_0, of observing a test statistic as extreme as or more extreme than the value actually obtained in the sample.

$$p\text{-value} = \mathbb{P}[\text{Test statistic} \geq \text{observed}] \qquad \text{(one-tailed)}$$

$$p\text{-value} = 2\mathbb{P}[\text{Test statistic} \geq |\text{observed}|] \qquad \text{(two-tailed)}$$

p-values

Example : In a z-test using the z-score presented before, the p-value is computed as follow:

$$p\text{-value} = 2\mathbb{P}\left[Z > |z_{\text{obs}}|\right]$$

We then compare this p-value to a significance level α we have chosen (typically 0.05).

p-values

For the comparison we have the following situations:

- p-value < α : reject H_0
- p-value > α : fails to reject H_0, not enough evidence to contradict

Interpretation : The smaller the p-value, the stronger the evidence against H_0.

p-values vs confidence intervals

- A **confidence level** $1-\alpha$ (e.g., 95%) defines the range within which we expect the true parameter to lie.
- If the **null value** μ_0 lies **outside** the confidence interval, the **p-value will be less than** α → reject H_0.
- Conversely, if μ_0 is **inside** the confidence interval, the p-value will be **greater than** α→ fail to reject H_0.

- Monday : Understand data structures

- Tuesday : Introduction to probability theory

- Wednesday : Central Limit Theorem confidence interval and test hypothesis

- Thursday : Feature selection and correlation matrix
    - Correlation matrix
    - Feature selection process

- Friday : Statistics with scikit-learn

Correction of the exercices

Quick review of all possible statistical tests

Feature type	Target type	Test name
Continuous	**Binary classification**	t-test
Continuous	**Categorical**	ANOVA
Categorical	**Categorical**	Chi Squared
Continuous	**Continuous**	Correlation

Correlation
- Some reminders -

The correlation coefficient is defined between two random variables (i.e. two different features) by using the covariance coefficient and normalise it as follow:

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

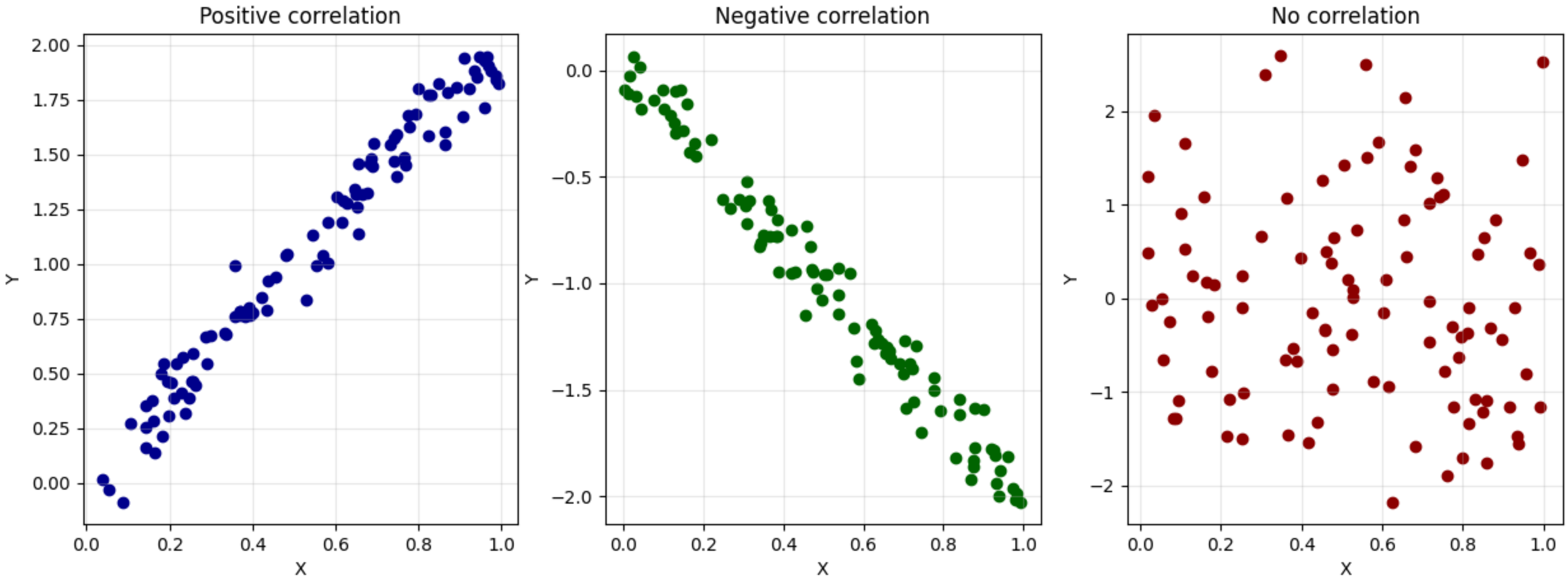
This values is located between -1 and 1.

Interpretation of the correlation coefficient

Correlation measures the relationship between variables. Here are the interpretation of some correlation values:

- -1 : perfect negative correlation (negative relationship)
- 0 : no linear correlation
- +1 : perfect positive correlation (positive relationship)

Interpretation of the correlation coefficient



Correlation matrix

The coefficients of a correlation matrix are defined as follow :

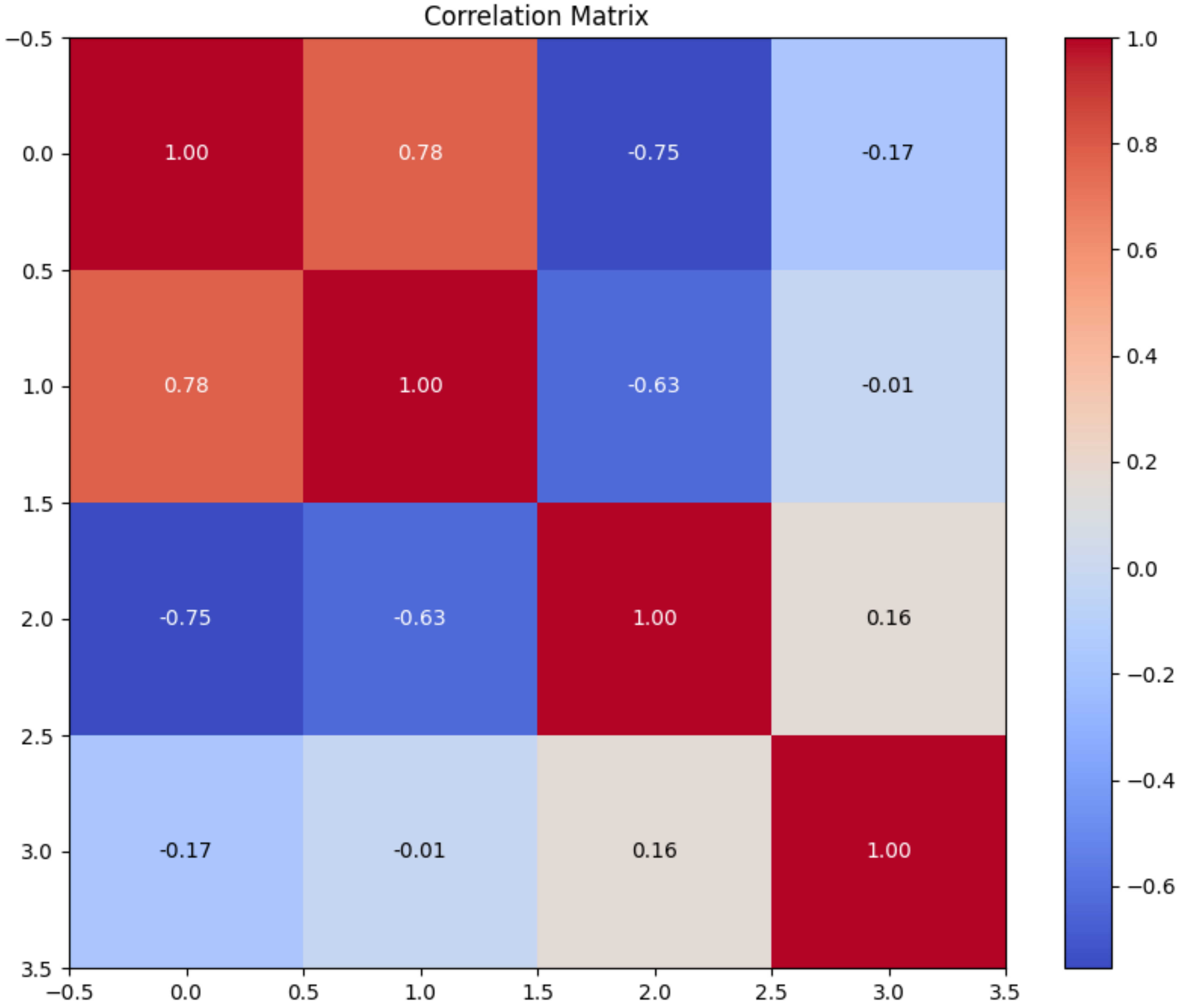
$$\rho_{ij} = \frac{\text{Cov}(X_i, X_j)}{\sigma_{X_i}\sigma_{X_j}},$$

where X_i, X_j are random variables with standard deviation $\sigma_{X_i}, \sigma_{X_j}$ respectively.

Correlation matrix

1.00	0.78	-0.75	-0.17
0.78	1.00	-0.63	-0.01
-0.75	-0.63	1.00	0.16
-0.17	-0.01	0.16	1.00

Correlation matrix



Correlation matrix

What are the situations when we want use a correlation matrix ?

- ✦ Feature selection
- ✦ Multicollinearity detection : very useful for regression models
- ✦ Exploratory Data Analysis (EDA) : Spot relationships

Correlation matrix
- Mistakes on the interpretation -

1 - Correlation shows association and ***not*** causality !

Example : Suppose we observe a strong positive correlation between coffee consumption and productivity.

=> If there is a strong correlation, it doesn’t mean drinking coffee causes higher productivity.

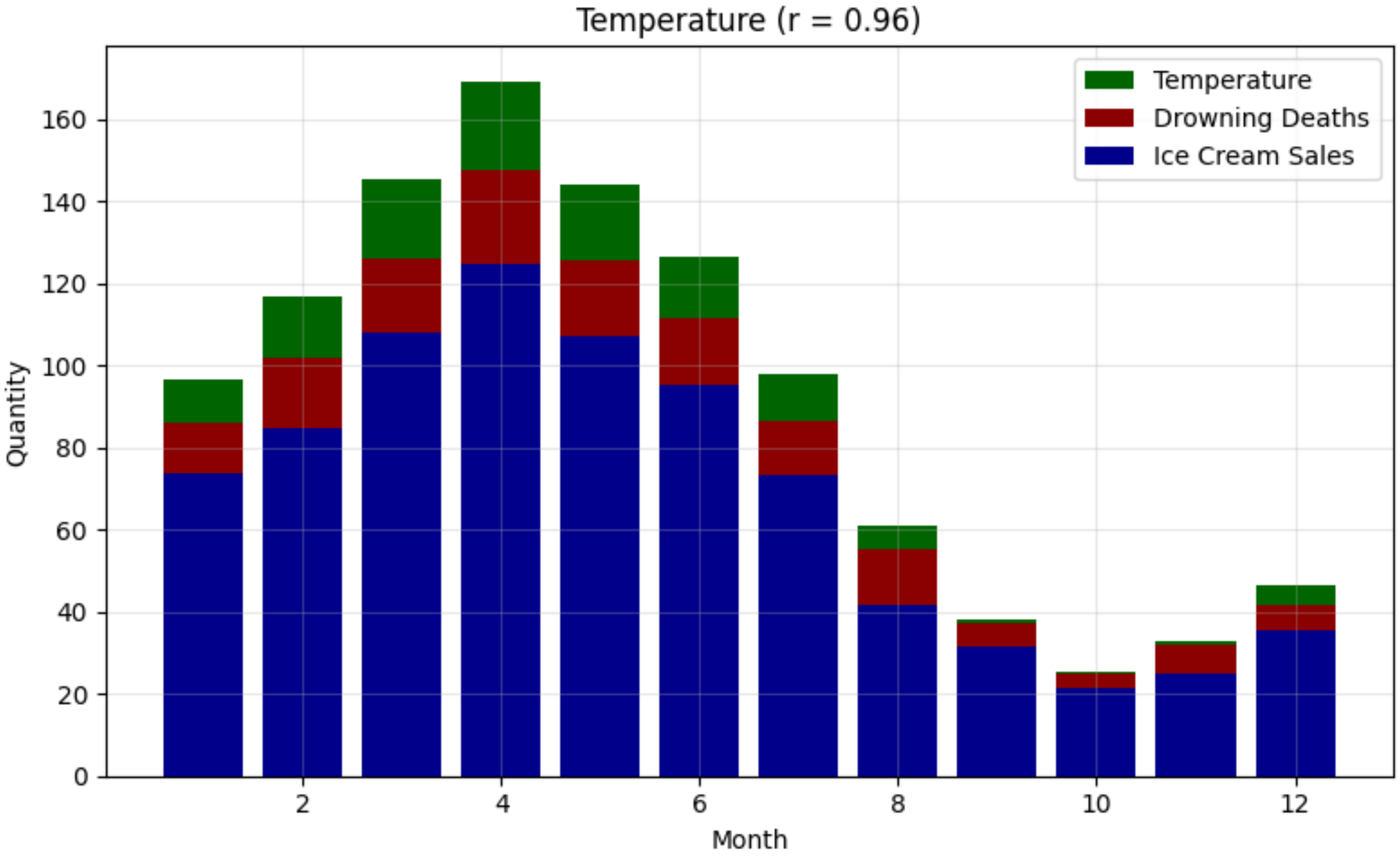
=> Maybe more productive people are just more likely to stay awake with coffee.

Correlation matrix
- Cofounding variables -

Question 1: Let's explore the ice_cream_sales_drowning_deaths.csv file, what is the correlation coefficient between Ice cream sales and drowing deaths ?

Question 2: and ice_cream_sales_drowning_deaths_temperature.csv ?

Correlation matrix
- Cofounding variables -



Spurious correlations

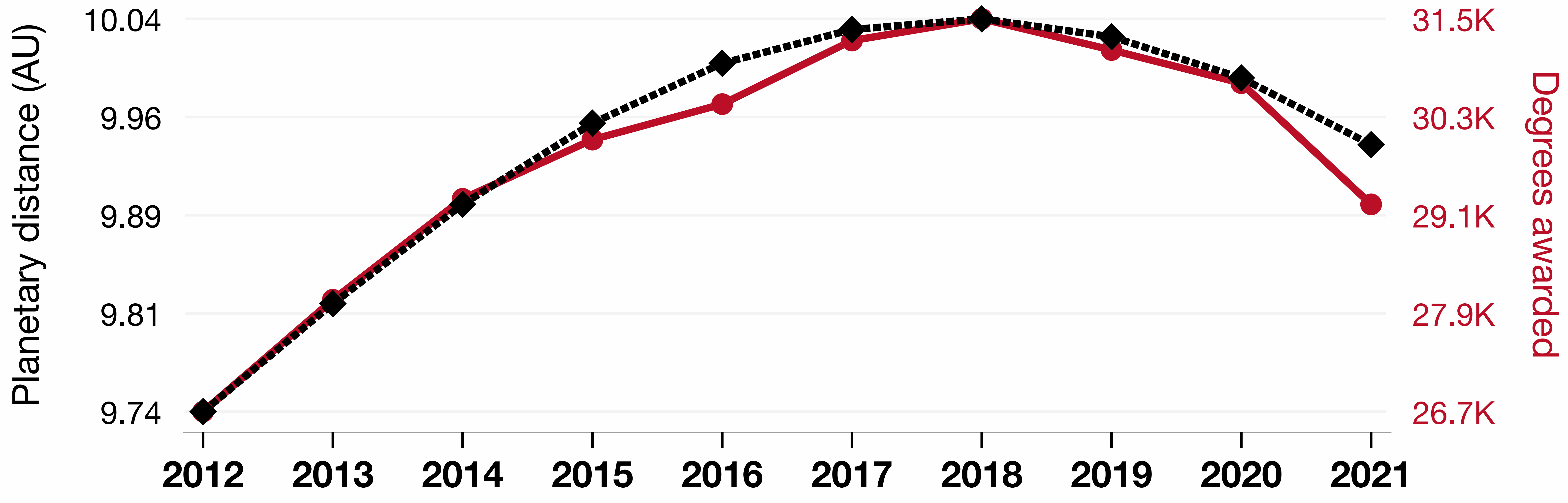
« Sometimes two things are completely unrelated, but randomly happen to correlate over a period of time »

Spurious correlations

The distance between Saturn and the moon

correlates with

Bachelor's degrees awarded in Physical sciences



The average distance between Saturn and the moon as measured on the first day of each month · Source: Caclculated using Astropy

Bachelor's degrees conferred by postsecondary institutions, in field of study: Physical sciences and science technologies · Source: National Center for Education Statistics

2012-2021, $r=0.987$, $r^2=0.974$, $p<0.01$ · tylervigen.com/spurious/correlation/2656

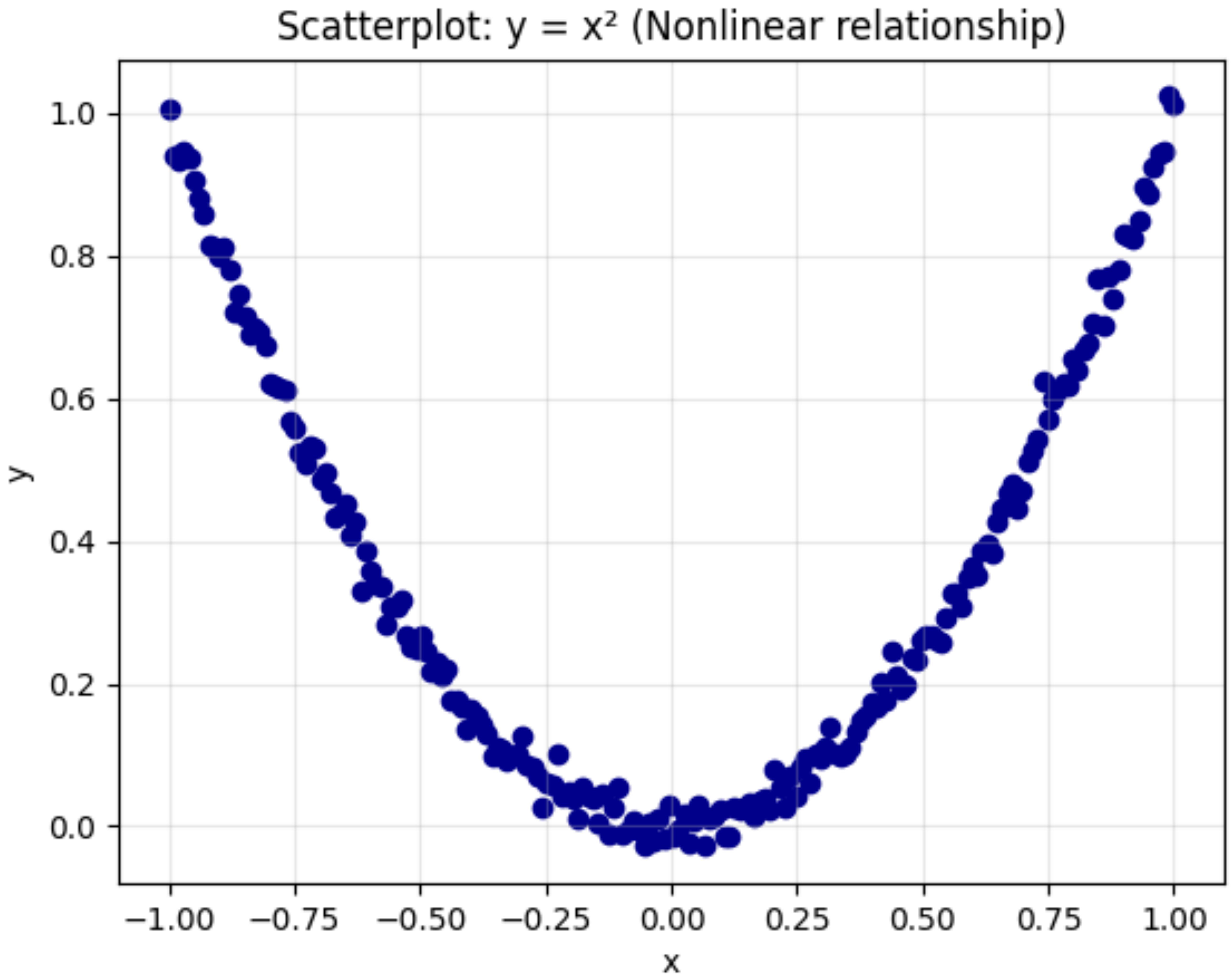
References

Non-linear relationships

<u>Question 1:</u> In the file non_linear_relationships.csv, compute the pearson correlation coefficient between variables x and y.

<u>Question 2:</u> Now make a scatter plot with these two features. What do you observe ?

Non-linear relationships



Feature selection
(finally)

Definition : The process of choosing a subset of relevant features (input variables). The idea is to reduce the complexity of the model, make it more accurate and speed up the computation.

Feature selection
- Why it does matter ? -

- Remove noisy and irrelevant variables
- Reduces the complexity of the model which reduces overfitting
- Speeds up training time
- Improves model interpretability

Different way of feature selection methods

Method	Description	Techniques
Filter	Based on statistical tests	Correlation, chi2, ANOVA
Wrapper	Uses model performances as a guide	Forward, backward, Recursive
Embedded	Feature selection built into the model	Lasso (L1), Tree based models

Feature selection
- Filter methods -

Idea : The idea is to rank all the features using their statistical relationship with the target variable you're trying to predict.

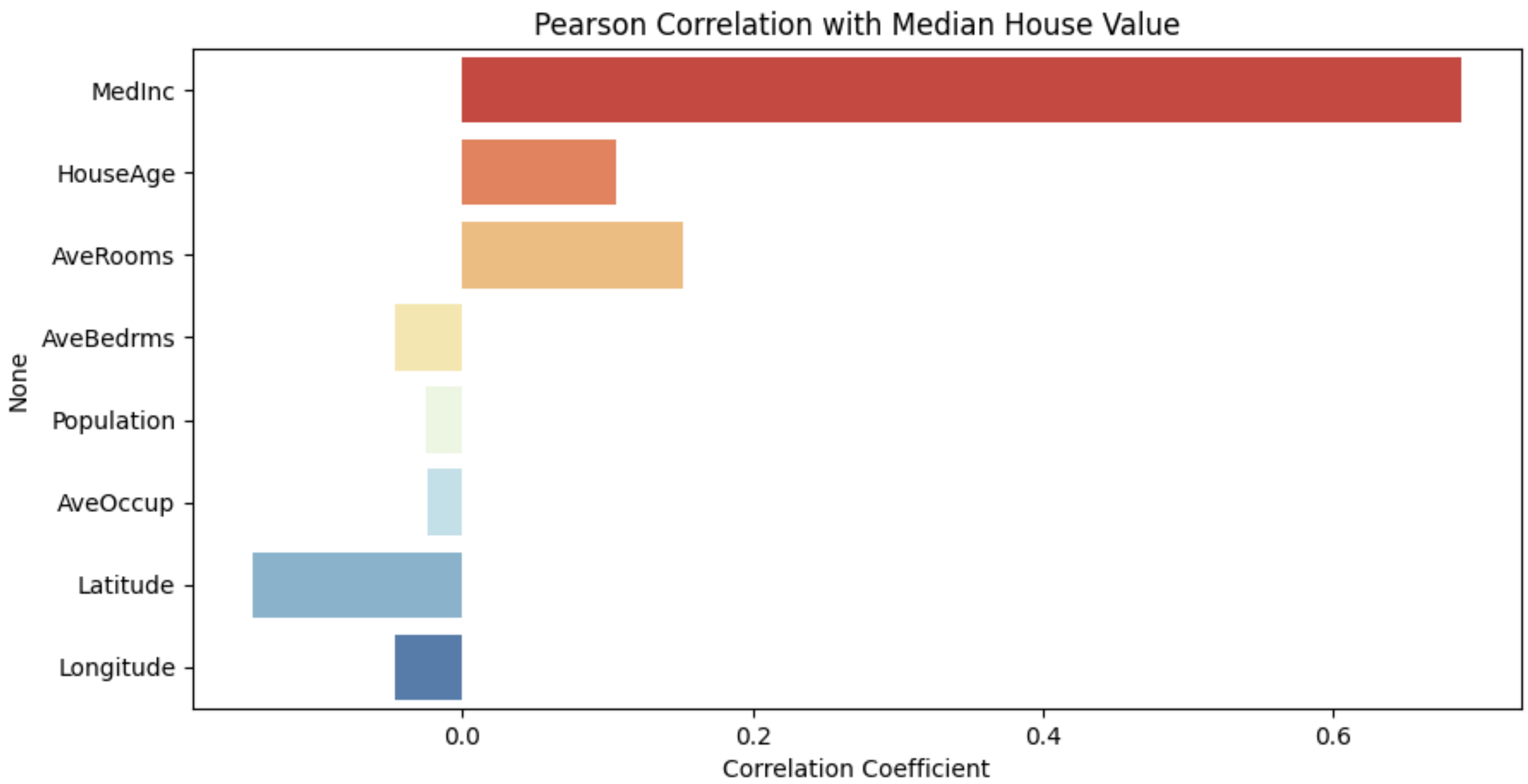
Their statistical relationships can be defined with Pearson correlation coefficient (for continuous variables) or with Chi-squared test (for categorical variables).

➡ Model agnostic
➡ Easy implementation and fast computation
➡ But it ignores interactions between features

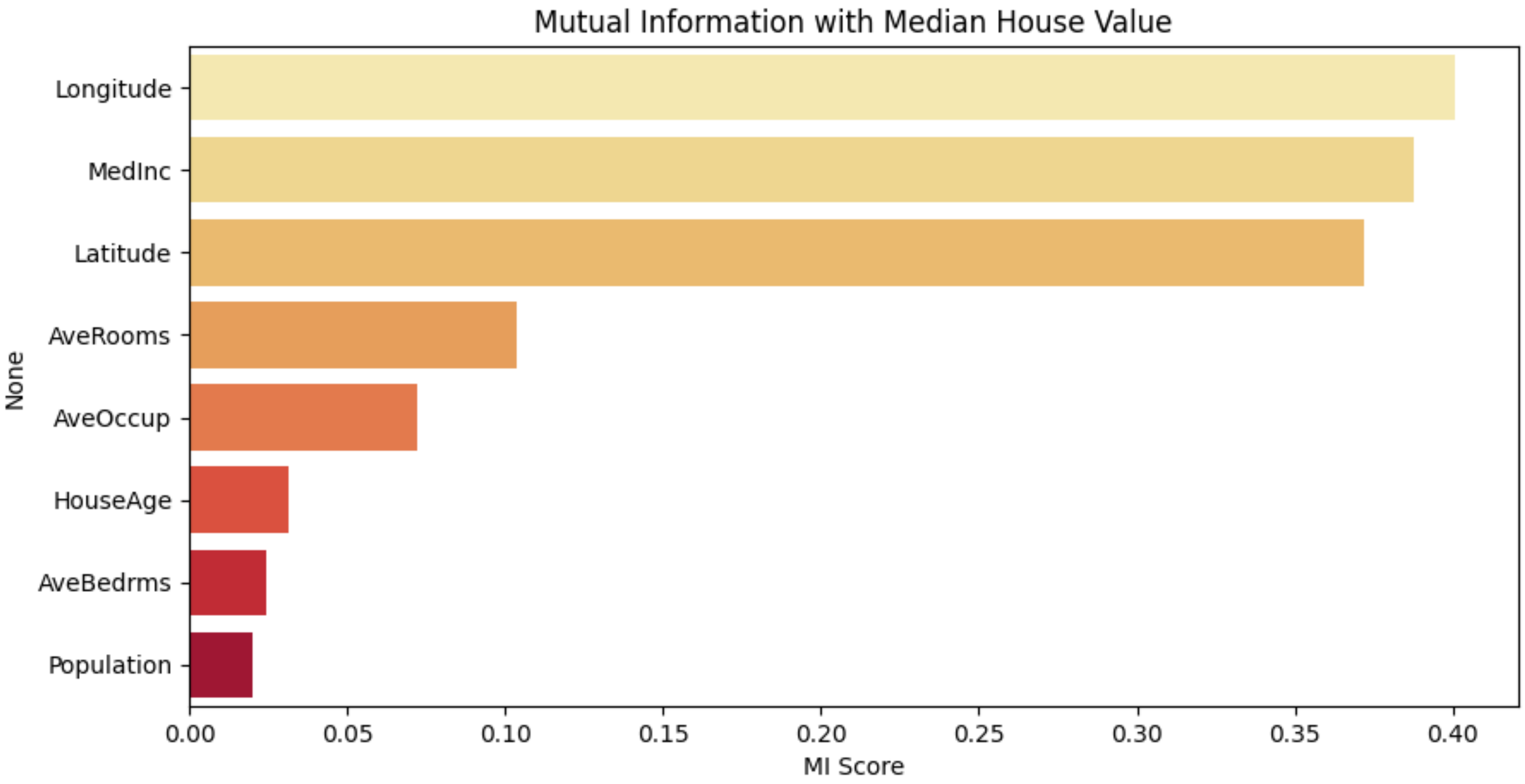
Feature selection
- Filter methods -

<u>Exercice :</u> Using the `fetch_california_housing` function from Python library `scikit-learn`, apply the filter methods using the Pearson’s correlation coefficients as a selection attributes.

Feature selection
- Filter methods -



Feature selection
- Filter methods -



Feature selection
- Wrapper methods -

Idea : You start with a model, usually easy to understand and train and you use the model performances to assess wether a feature needs to be selected.

- ✴ Forward selection
- ✴ Backward selection

Wrapper selection
- Forward selection -

Idea : Start with an empty set of features and recursively add one feature at a time and stop when there is no more improvements.

How to measure improvements ?

Wrapper selection
- Forward selection -

<u>Idea :</u> Start with an empty set of features and recursively add one feature at a time and stop when there is no more improvements.

<u>How to measure improvements ?</u>

=> Using model performances like R^2 or accuracy !

Wrapper selection
- Backward selection -

Idea : Start with all features and recursively remove feature with the least importance.

When do you stop the process ?

Feature selection
- Embedded methods -

Idea : Some models include natively a feature selection process. It is the case for regression trees, random forest models or XGboost models.

=> All these models will be seen next week 😉

Feature selection
- Risks -

✦ Removing too many features : underfitting risks !

✦ Selection is made only on the <u>train set</u>

✦ Correlated features

Feature selection
- Practice -

<u>Question :</u> Load the data set `load_breast_cancer` from scikit-learn and use the following feature selection methods:

- Filter methods: pearson’s correlation
- Filter methods: mutual information
- Wrapper methods : Forward
- Wrapper methods : Backward

- Monday : Understand data structures

- 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
  - Introduction to scikit-learn & methods
  - Model evaluation
  - Cross validation