Descriptive statistics

Applied Data Analysis (ADA) - October 2025

Nomades Advanced Technologies
Gaspard Villa

- Monday : Understand data structures
    - Population vs sampling
    - Central tendency measures
    - Dispersion measures

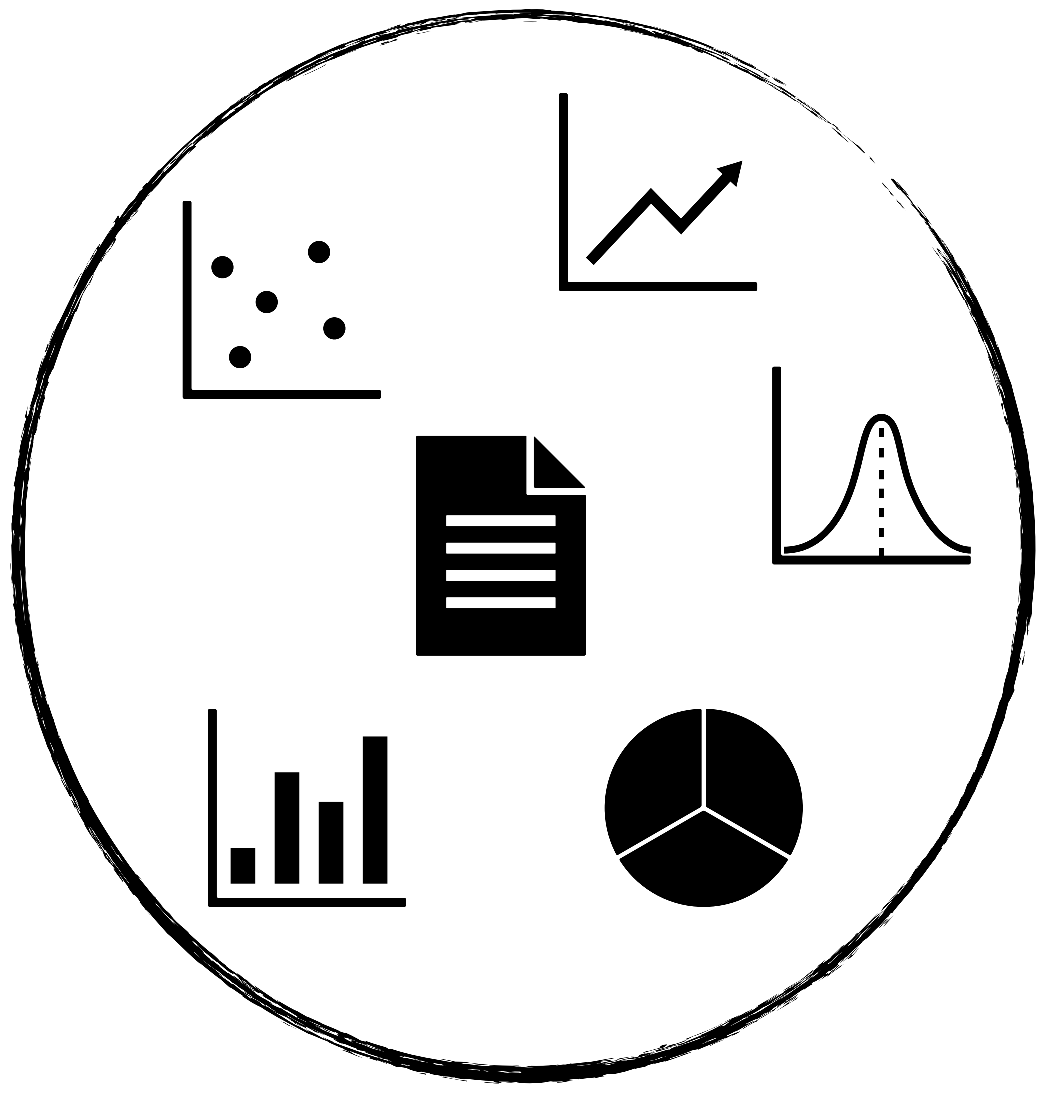
- Tuesday : Introduction to probability theory

- Wednesday : Central Limit Theorem, confidence intervals and test hypothesis

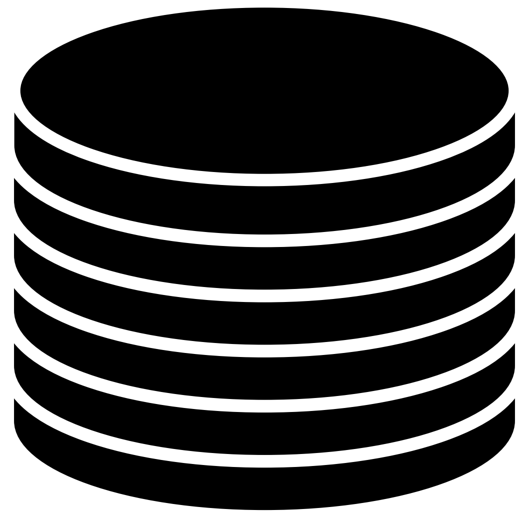
- Thursday : Data cleaning methods & pandas manipulations

- Friday : Statistics with scikit-learn

How a project is built ?



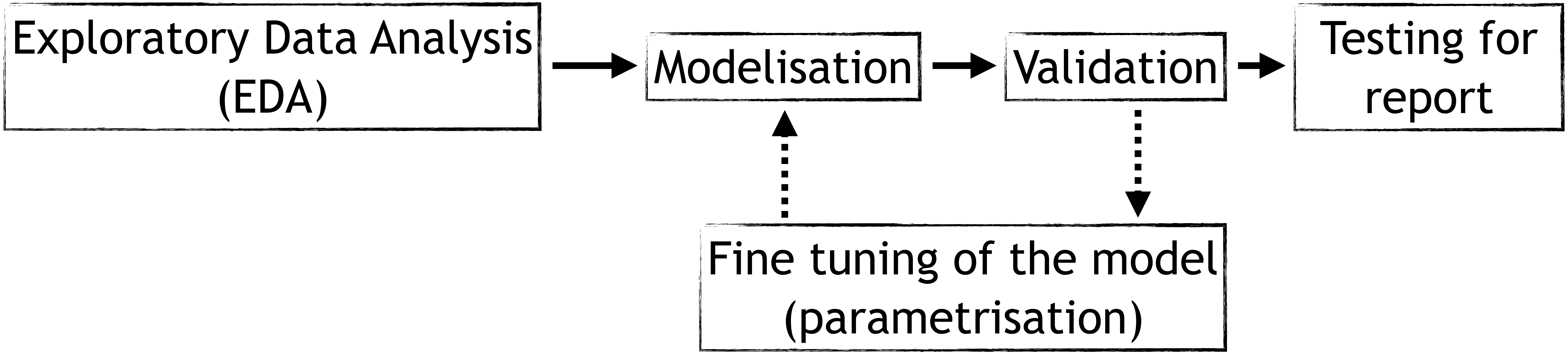
Raw data



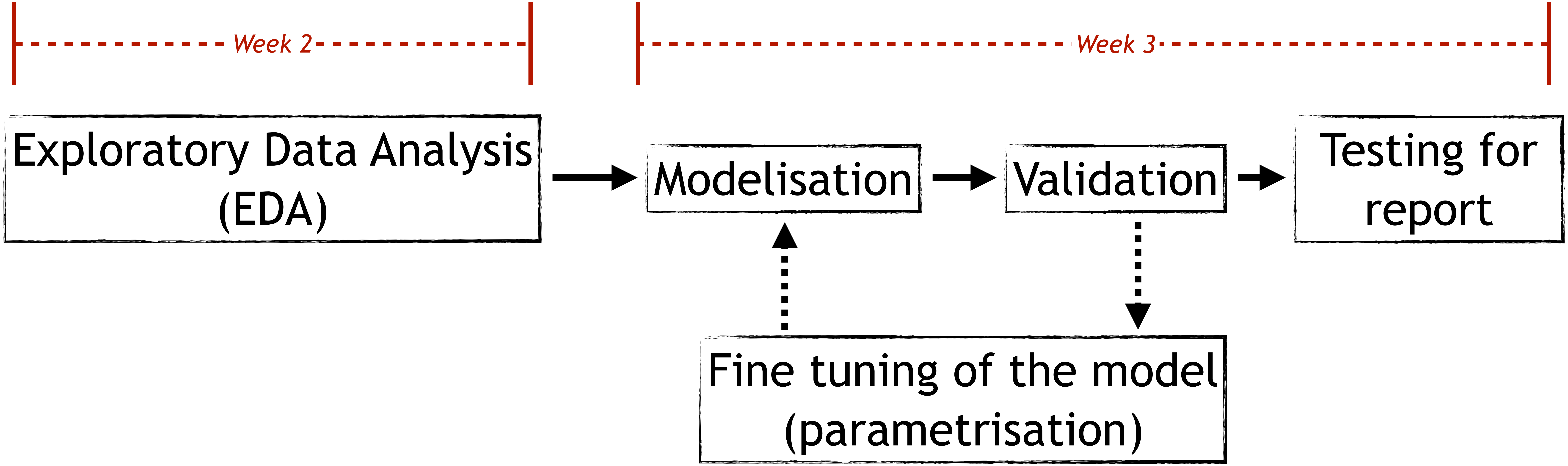
Workable dataset

Modelisation, predictions, hidden informations, etc…

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.

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

`np.cov(X)`

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

`np.corrcoef(X)`

Different types of data

Unstructured

- Images
- Text
- Videos
- Time Series
- …

Structured

- Numerical values
  - Continuous
  - Categorical

Let’s stick to structured data

For a structured dataset, it can be stored into a <u>table</u>.

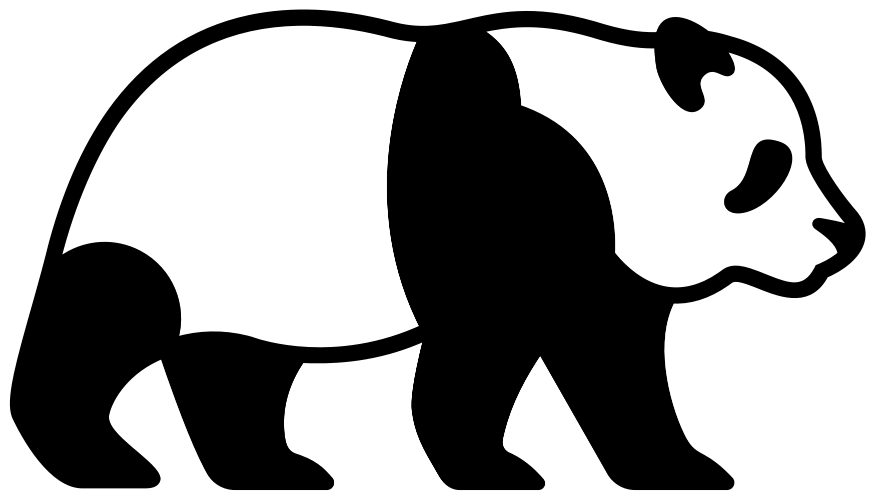
=> Do you know a nice tool/library in Python to manipulate tables ?

Let’s stick to structured data
- Pandas -

For a structured dataset, it can be stored into a table.

=> Do you know a nice tool/library in Python to manipulate tables ?

PANDAS



Pandas : basics (reminder)

What is it used for ?

You have a data set that is structured with one or more "features" that you can fit into a table.

For example, the grades of each of you given by each instructor during your final presentation.

How this table would be built ?

Pandas : basics (reminder)

What is it used for ?

You have a data set that is structured with one or more “features” that you can fit into a table.

For example, the grades of each of you given by each instructor during your final presentation.

How this table would be built ?

Name of the student	Grade [Antonio]	Grade [Haim]	Grade [Gabriel]	Grade [Gaspard]

Pandas : basics (reminder)

Generally, you have a csv or excel file that you need to import into your code base.

For that you use pandas, for example it can be done with the functions:

- pd.read_csv()
- pd.read_json()
- pd.read_excel()

Pandas : basics (reminder)

Now, you have full access to your data set. You want to check the general informations about it: how many data ? what kind / types of data ?

Here are some functions useful to answer these questions:

- .info()
- .shape
- .dtypes
- .columns

Pandas : basics (reminder)

It’s time now you look directly at your data set.

The idea is to check wether the data were well imported, check the look of different samples, etc… To help you have the following functions :

- .head()
- .tail()
- .sample()
- .value_counts()

Pandas : basics (reminder)

How do you catch a specific column ? A specific row ?

- df['Age']
- df.iloc[idx]
- df[['Age', 'Education level']]

How do you catch only the people older than 40 years old ?

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

Pandas : basics (reminder)

How do you catch a specific column ? A specific row ?

- df['Age']
- df.iloc[idx]
- df[['Age', 'Education level']]

How do you catch only the people older than 40 years old ?

=> df[df['Age'] > 40]

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

Pandas : basics (reminder)

Finally, some useful functions are already implemented for you so you don't have to recode them.

Here is an overview of some simple functions / operations you may apply to your data set:

- .mean()
- .std()
- .sum()
- .count()
- .describe()

Pandas : a bit of practice

Exercise 1 : Go to the notebook of today and do the first exercise. You will explore, understand and use the different functions presented before.

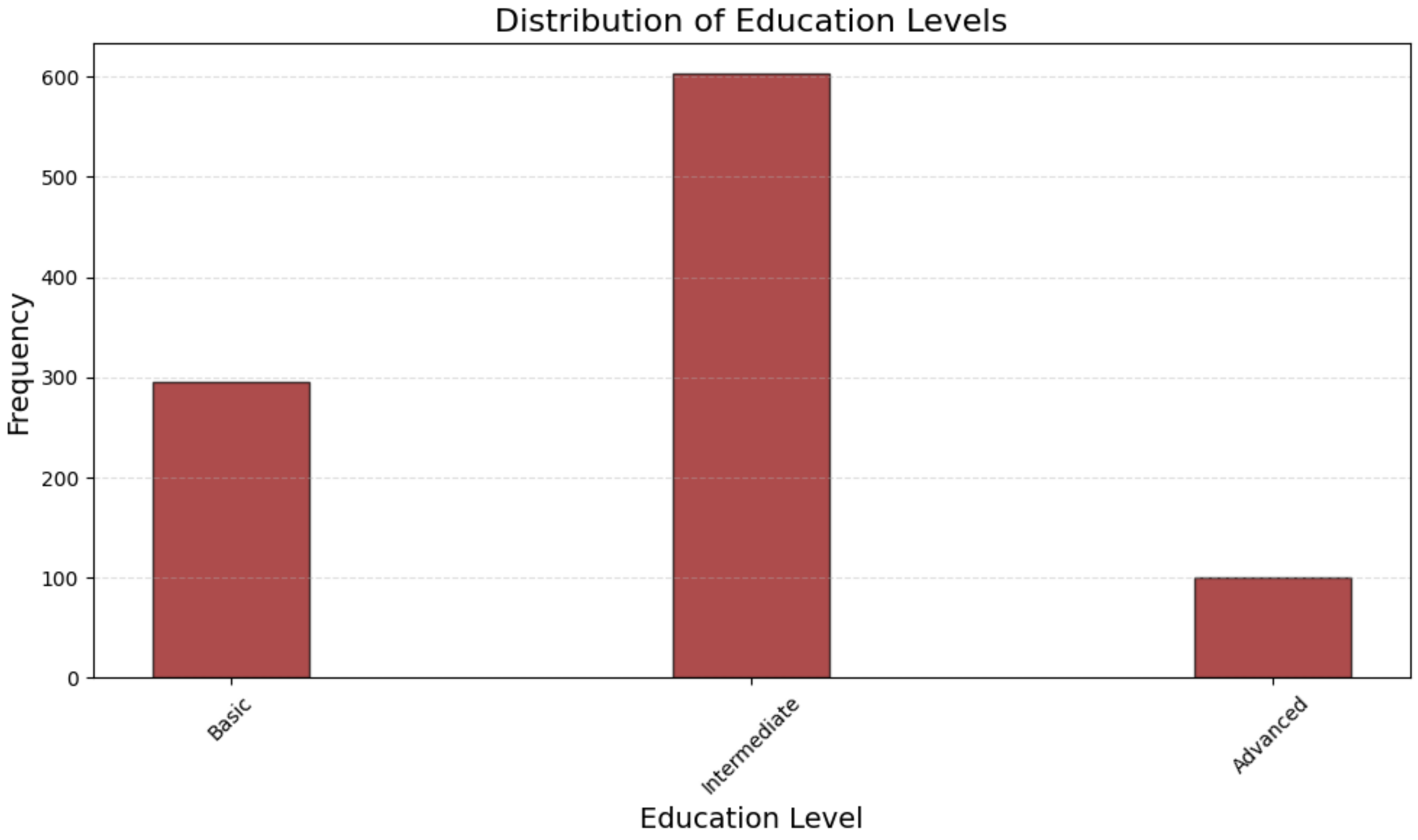
Key for analysis is Visualisation
- See your data -

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	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

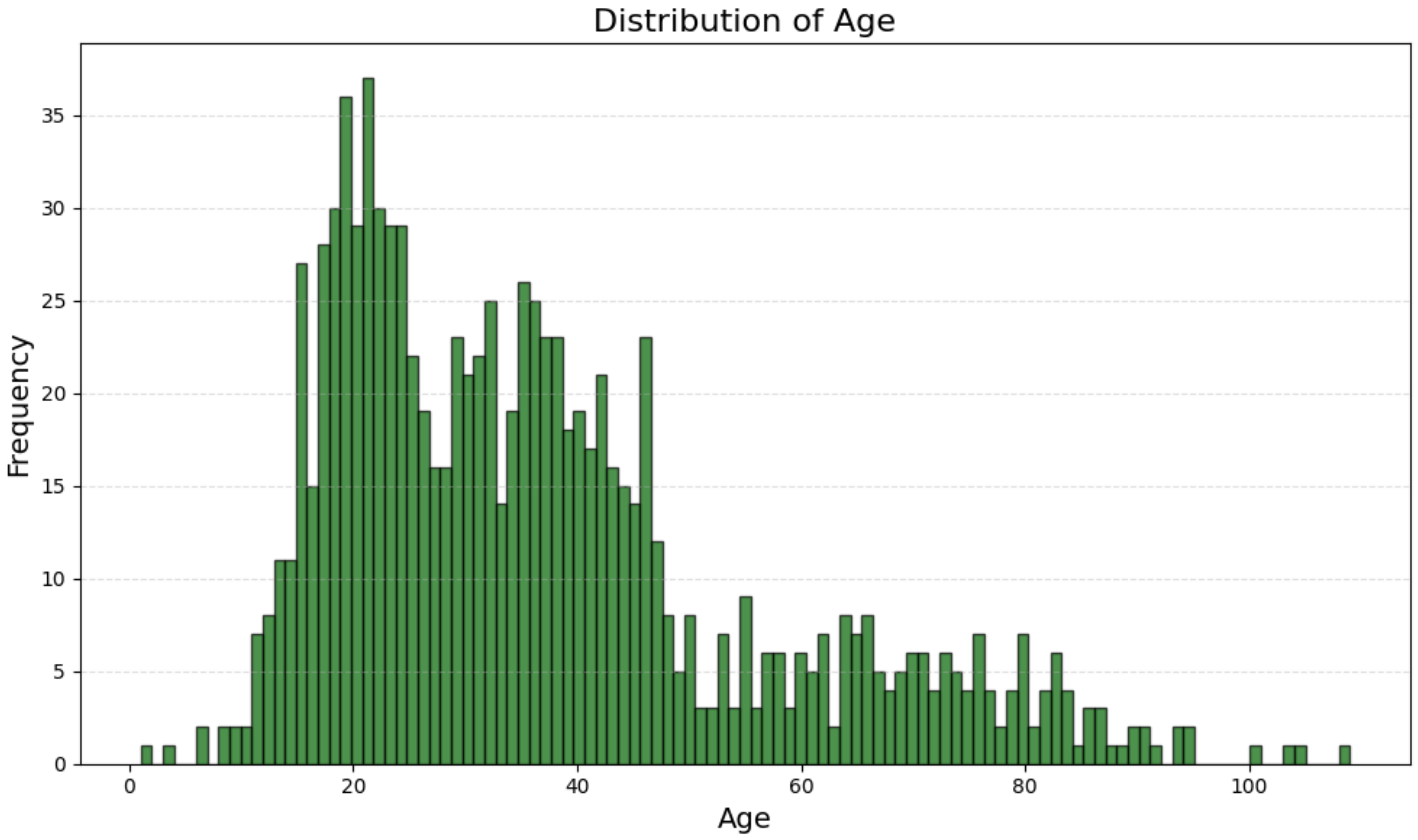


References

Key for analysis is Visualisation
- Univariate analysis -

2 - Univariate analysis for continuous variables

Histograms

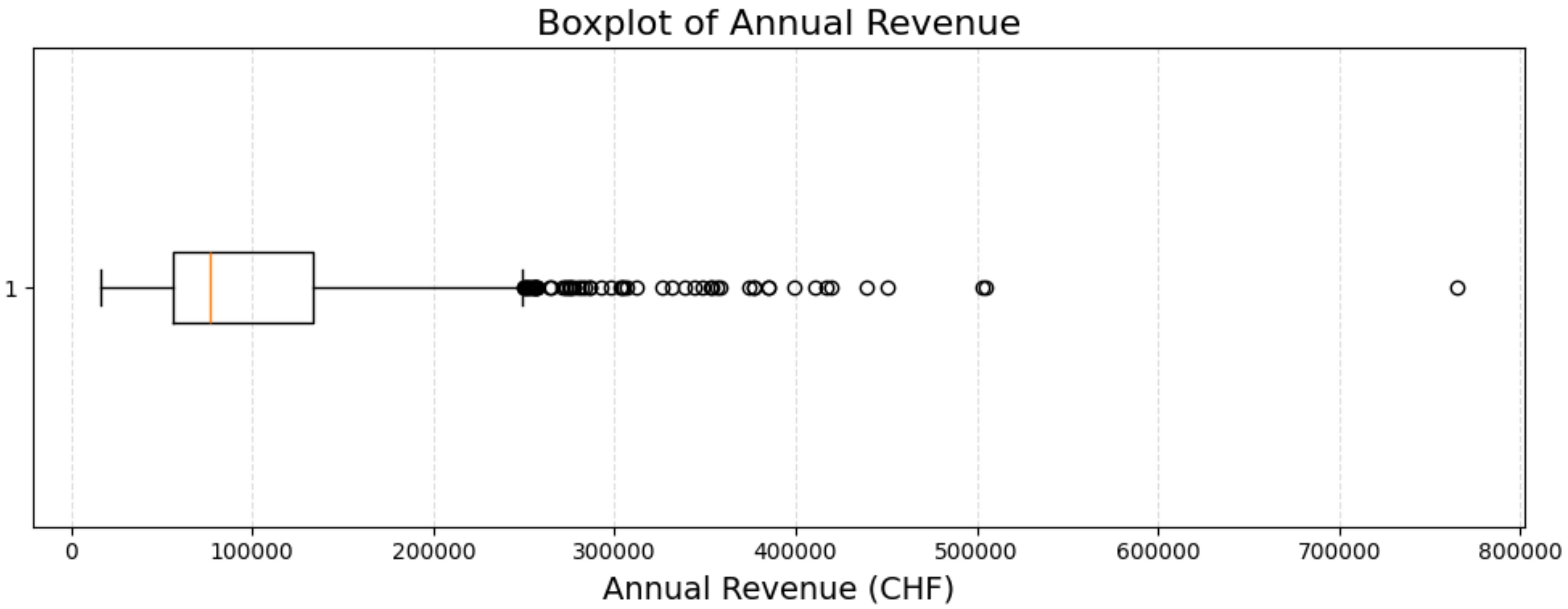


References

Key for analysis is Visualisation
- Univariate analysis -

3 - Univariate analysis for continuous variables

Boxplots

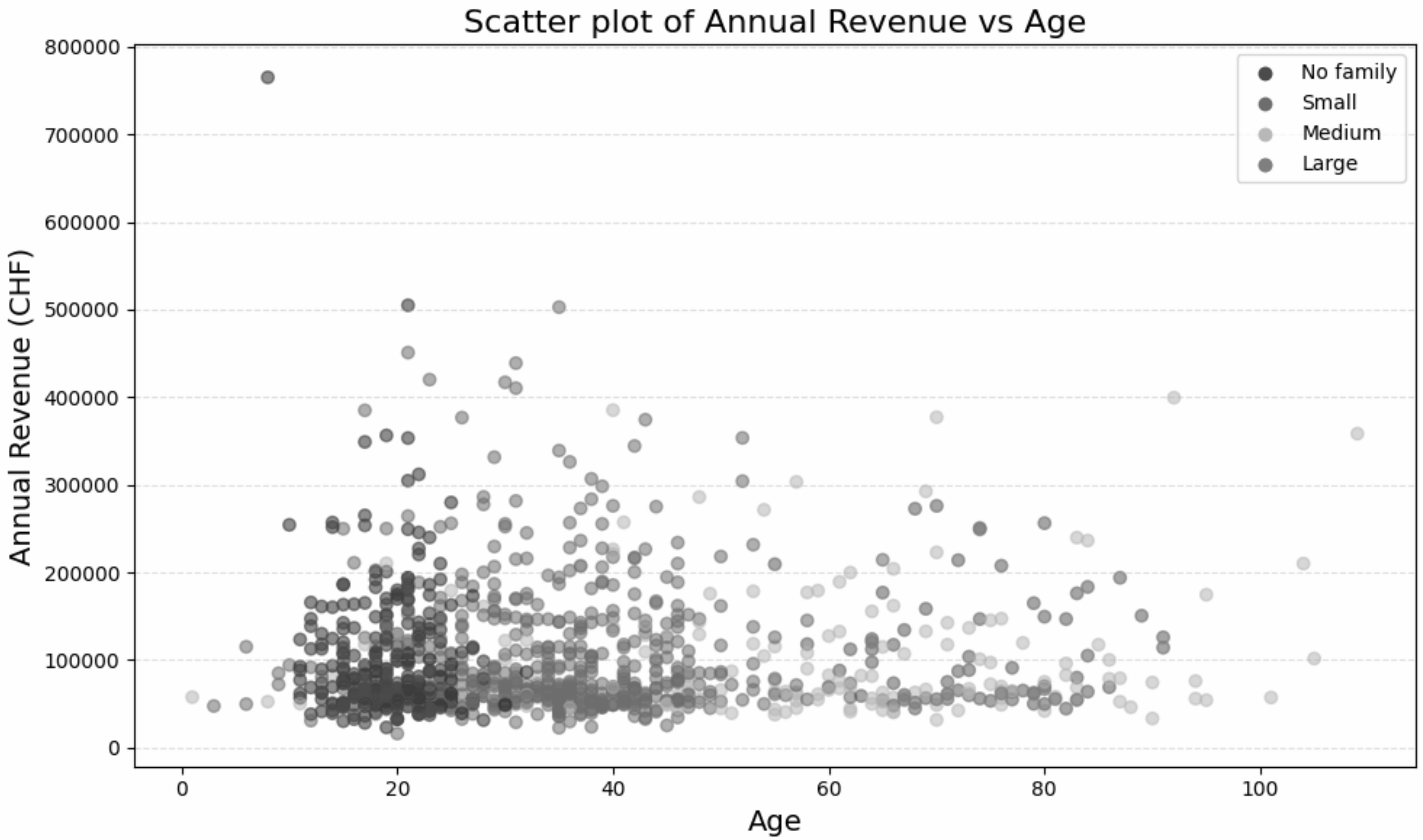


References

Key for analysis is Visualisation
- Multivariate analysis -

4 - Bivariate analysis for continuous variables

Bivariate scatter plot

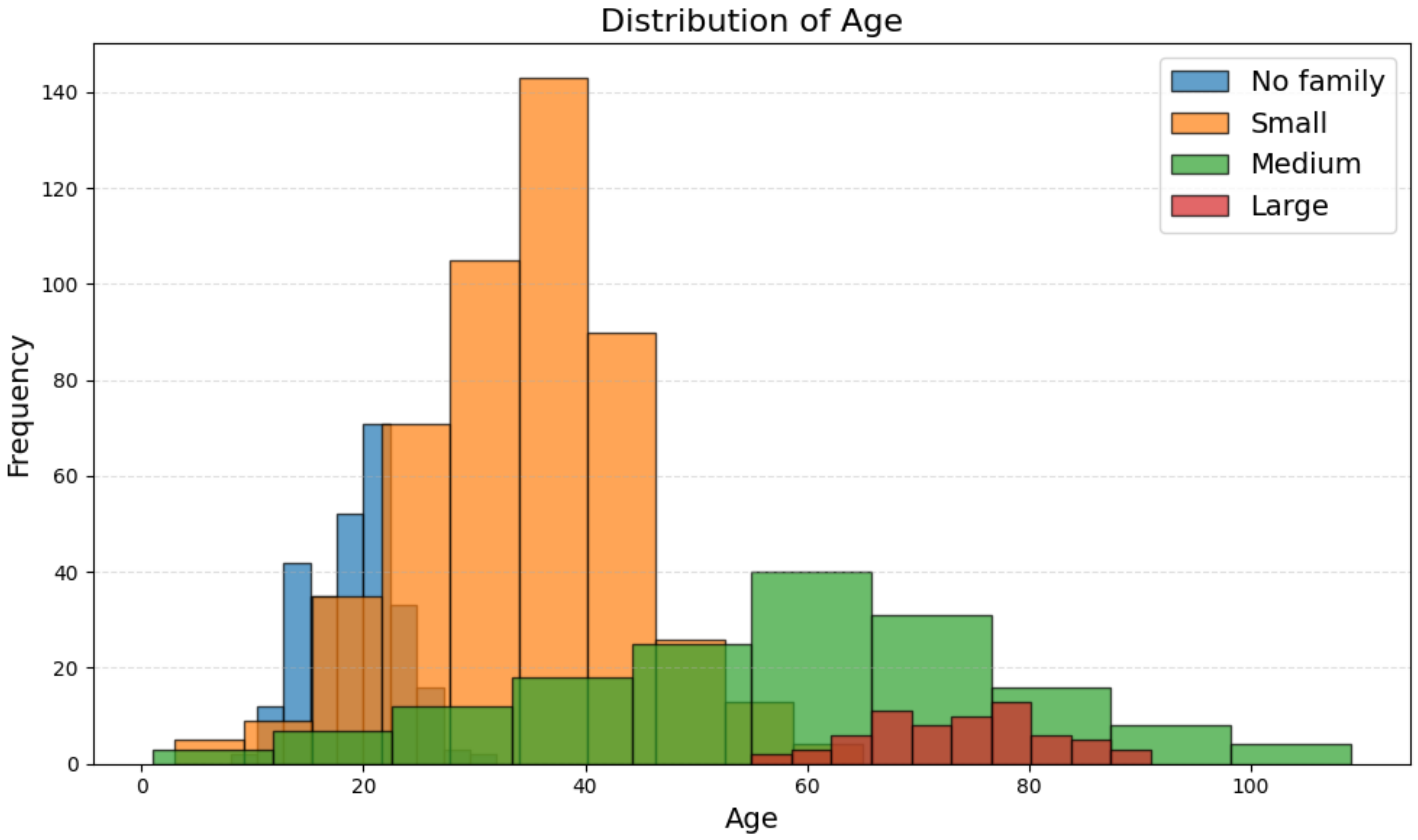


References

Key for analysis is Visualisation
- Multivariate analysis -

4 - Multivariate analysis for continuous and/or categorical variables

Multivariate

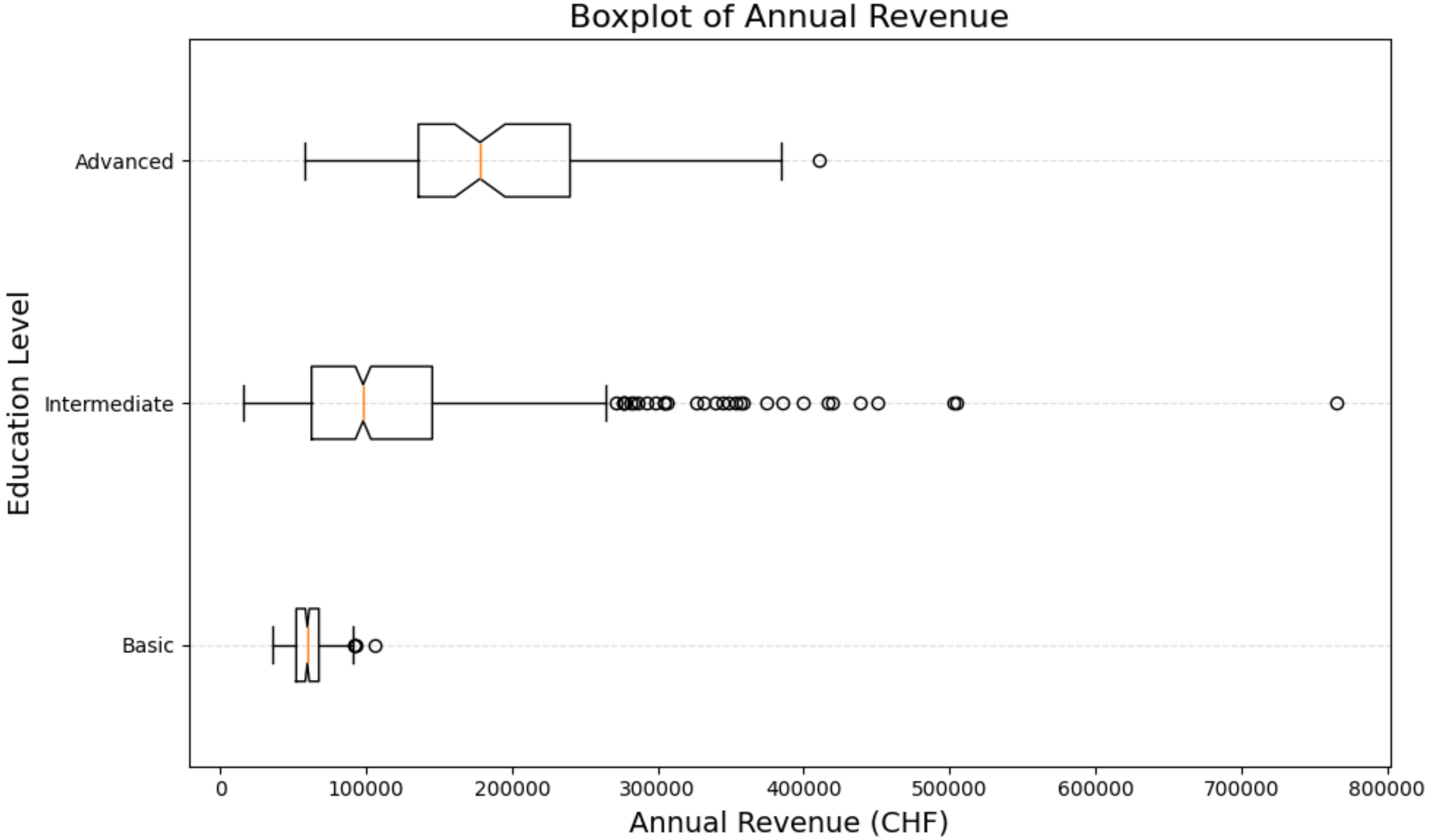


References

Key for analysis is Visualisation
- Multivariate analysis -

5 - Multivariate analysis for continuous and/or categorical variables

Bivariate boxplots

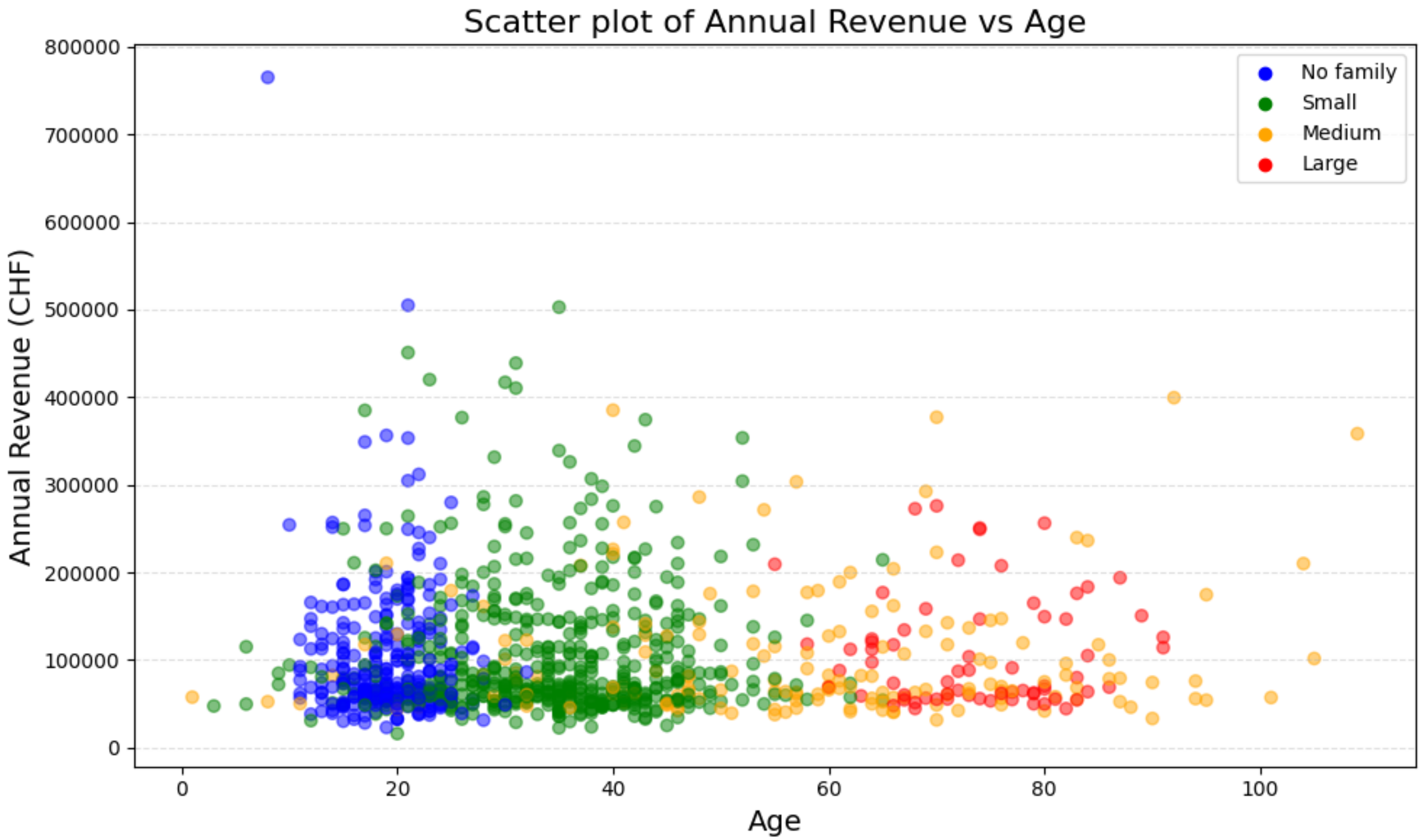


References

Key for analysis is Visualisation
- Multivariate analysis -

6 - Multivariate analysis for continuous and/or categorical variables

Colored scatter plots

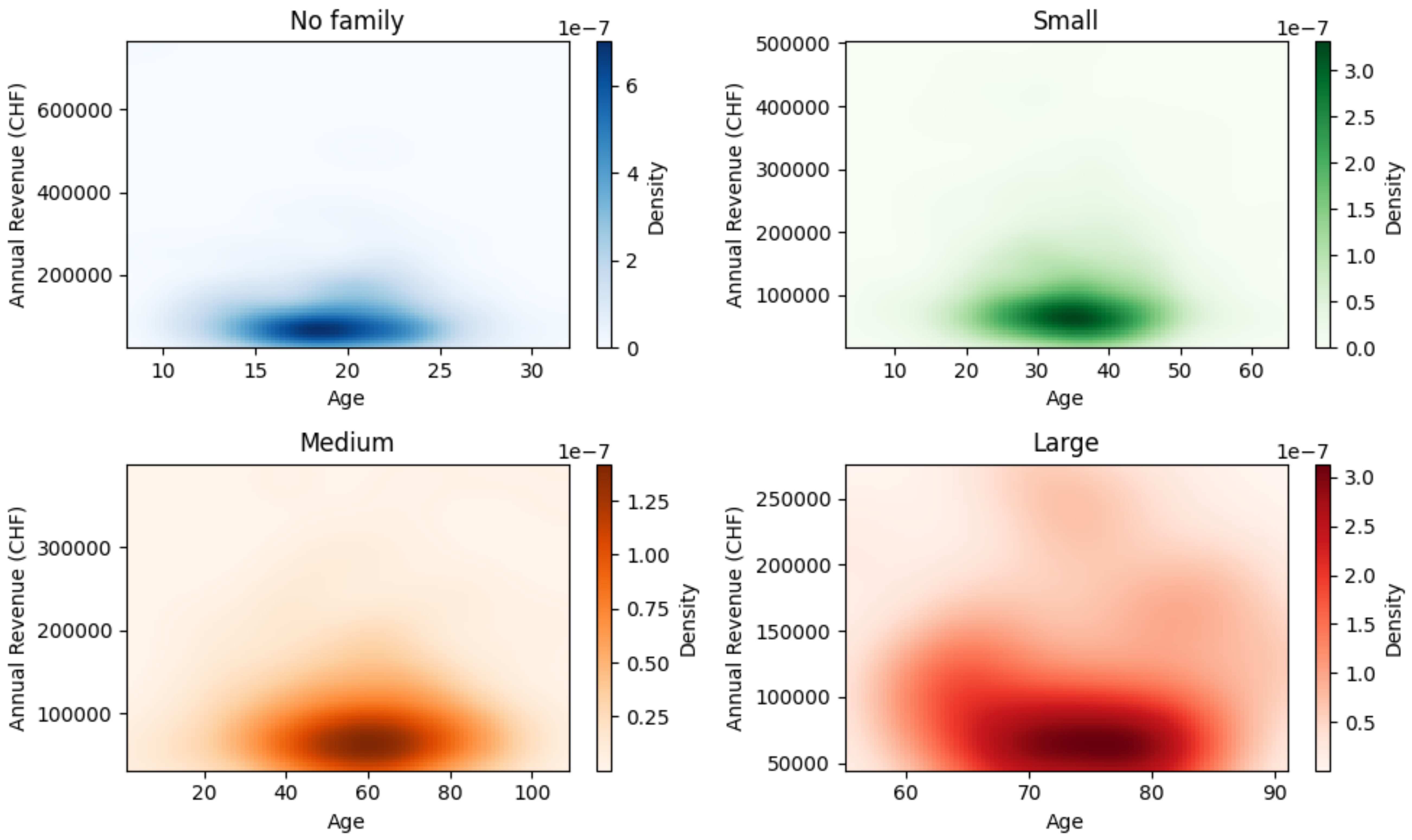


References

Key for analysis is Visualisation
- Multivariate analysis -

7 - Multivariate analysis for continuous and/or categorical variables

Density plots



References

Visualization : a bit of practice

Exercise 2 : Go to the notebook of today and do the second exercise. You will learn how to plot exactly the different features so that you may understand better the dataset.

Exercice for tomorrow

For tomorrow, you have the Exercise 3 from the notebook of today's exercises.

The idea for you is to use the "fake" dataset from waves.csv and extract the hidden informations you may find into it.

- Monday : Understand data structures

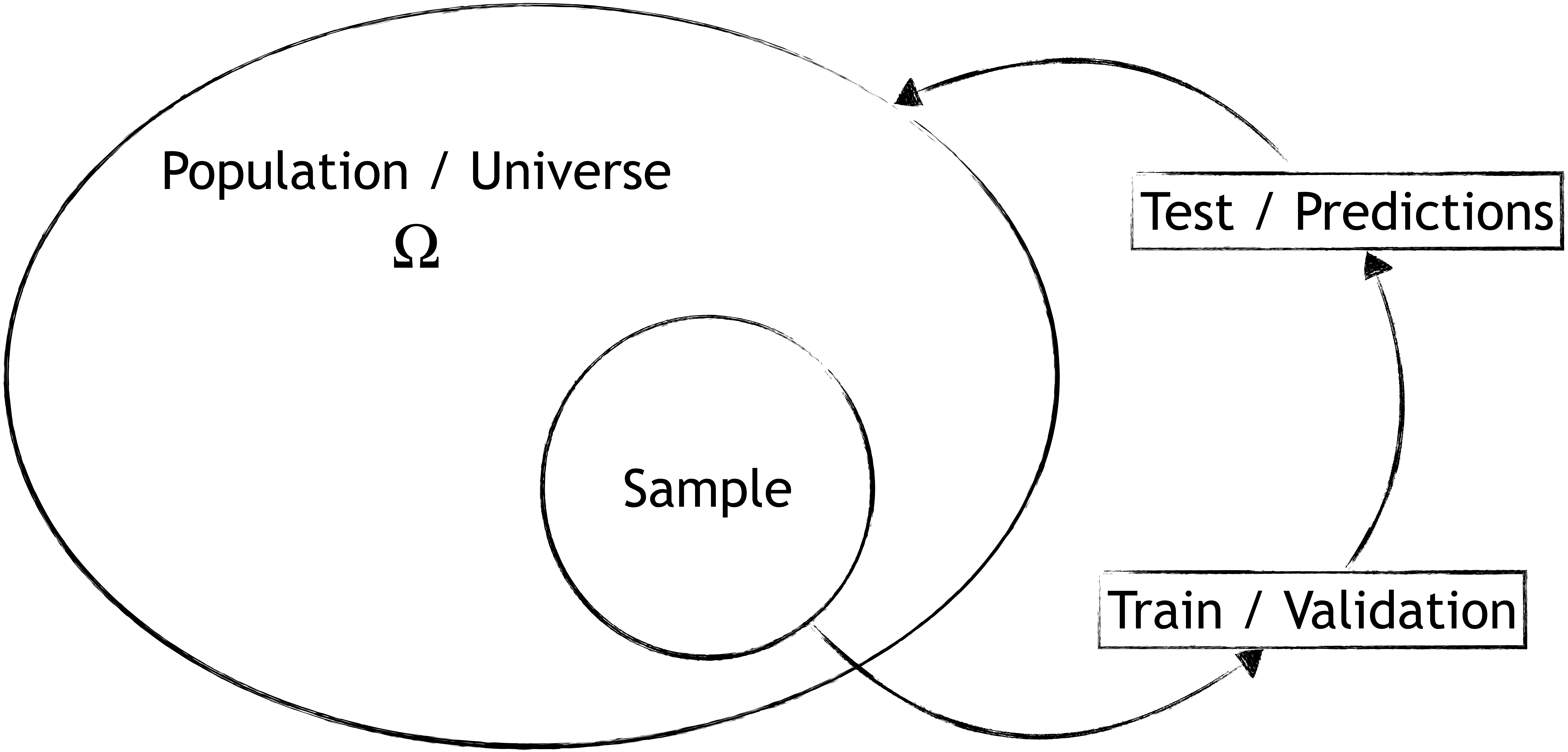
- Tuesday : Introduction to probability theory
  - Fundamental definitions
  - Probability law / distribution (discrete)
  - Discrete vs continuous probability

- Wednesday : Central Limit Theorem, confidence intervals and test hypothesis

- Thursday : Data cleaning methods & pandas manipulations

- Friday : Statistics with scikit-learn

Introduction to probability theory



Some definitions

❖ **Universe :** denoted as Ω, it represents the sample space of all possible elements or outcomes you are studying.

Some definitions

❖ **Universe :** denoted as Ω, it represents the sample space of all possible elements or outcomes you are studying.

❖ **Event :** it represents a subset of outcomes from an experiment in the sample space.

Some definitions

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

- ❖ **<u>Event :</u>** it represents a <u>subset of outcomes</u> from an experiment in the sample space.

- ❖ **<u>Event Space :</u>** denoted as F, it represents a <u>set of events.</u>

Some definitions

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

❖ <u>**Event :**</u> it represents a <u>subset of outcomes</u> from an experiment in the sample space.

❖ <u>**Event Space :**</u> denoted as F, it represents a <u>set of events.</u>

❖ <u>**Probability function :**</u> it assigns for each event from the event space a probability, which is a value in $[0,1]$.

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 :

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

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 :

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

<u>Practice :</u> What would the probability to get a 6 when throwing a 6-figures balanced dice ? And an even number ?

Some practice

<u>Excercice :</u> 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 (2,4,6,8 and 10) ?

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:

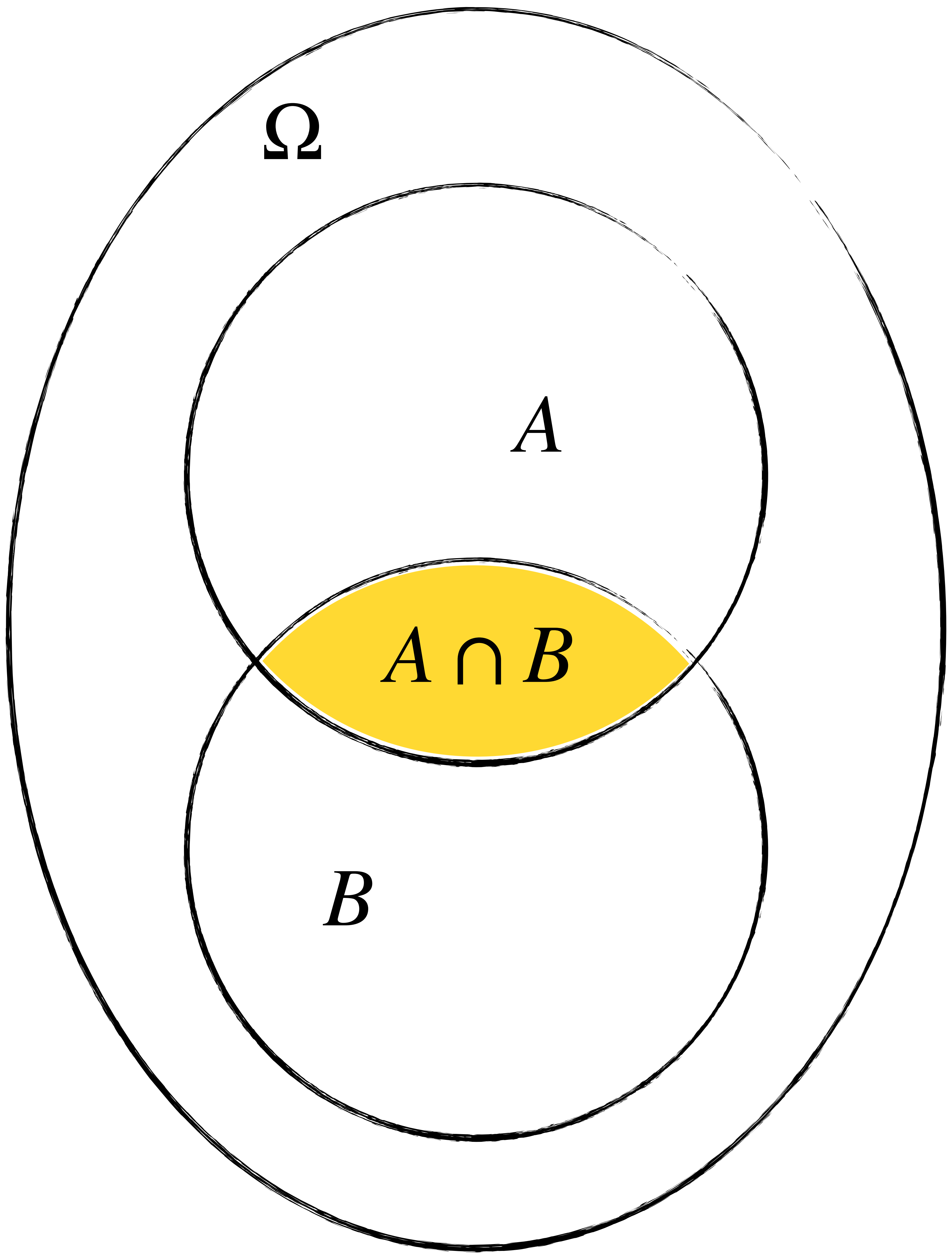
$$\mathbb{P}\left[\text{Pick even card}\right] = \frac{20}{52} \approx 0.38$$

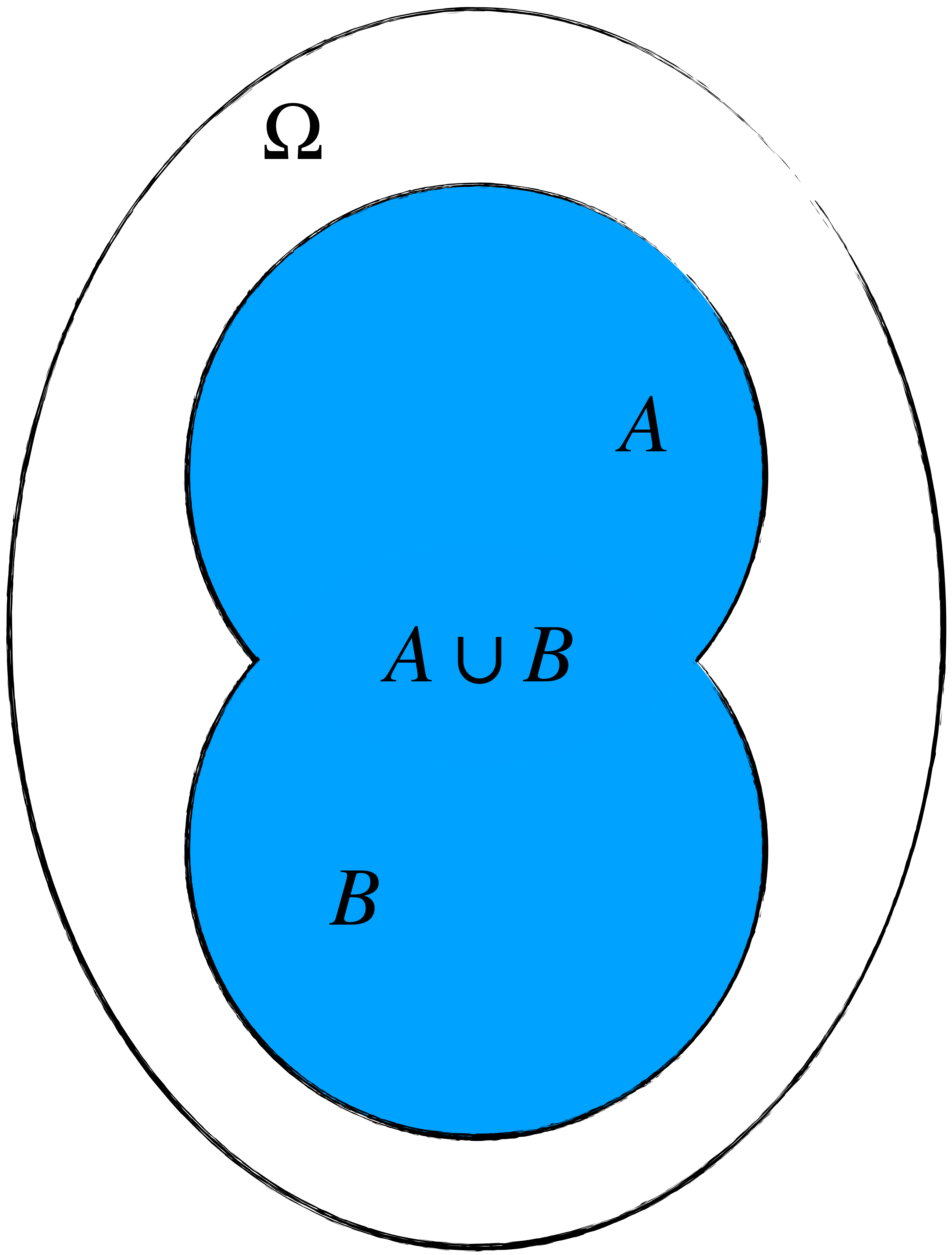
Some properties

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 :

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

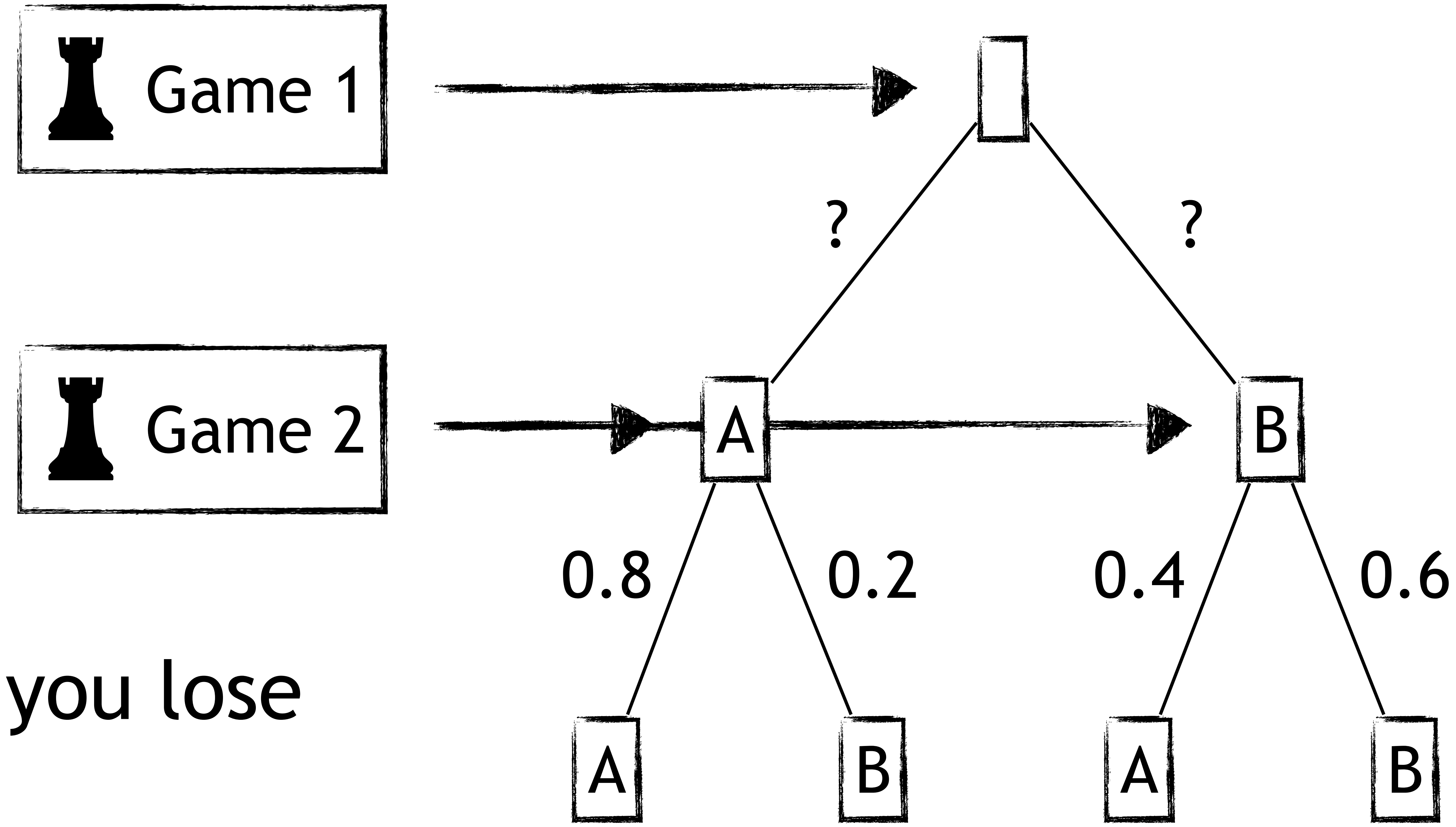
Let's play chess !

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 win ?

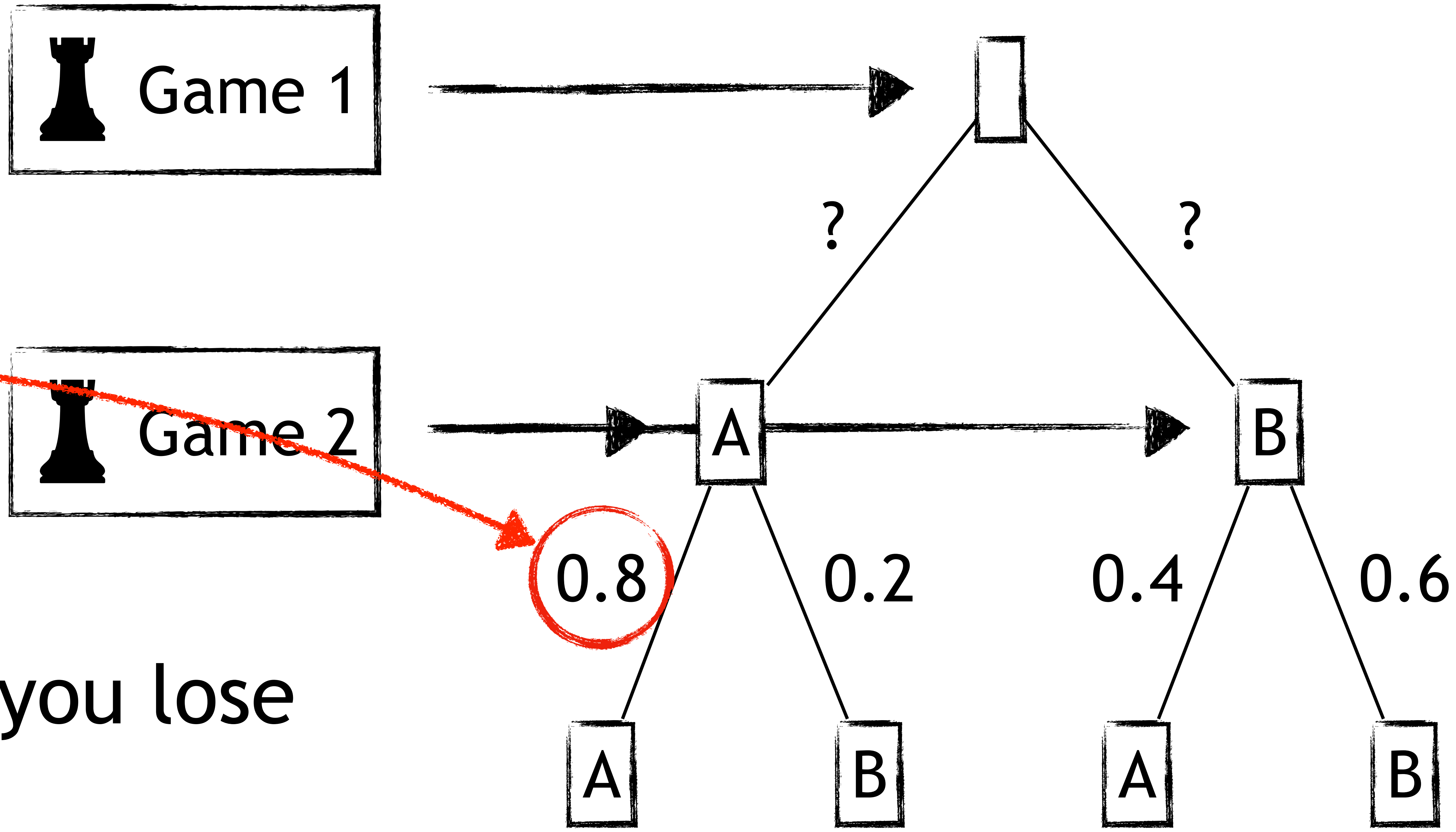
Exo 2 - Your probability to win both games is 0.56, what is the probability you lose the first game ?



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.



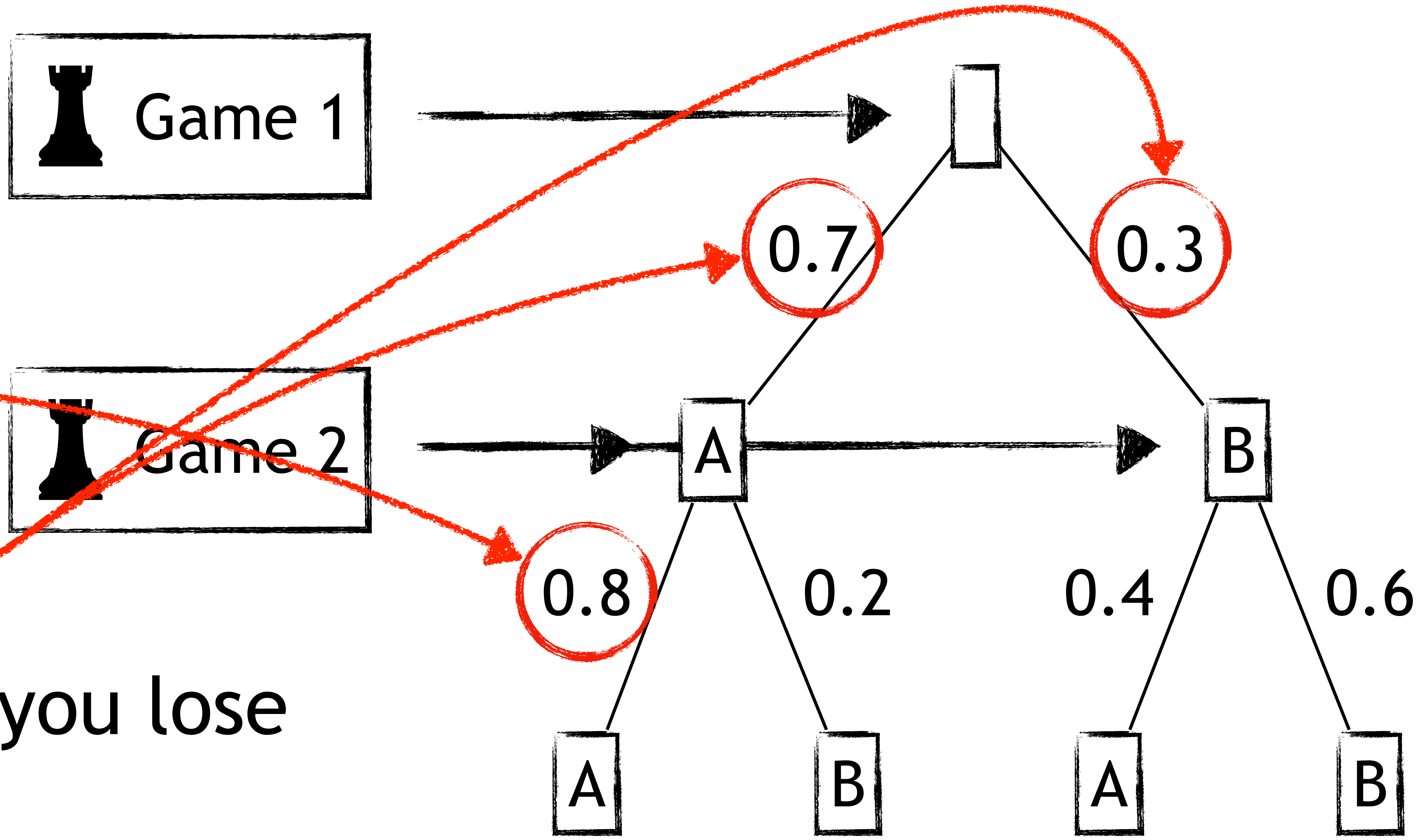
Exo 1 - What is the probability for you to win a game after a win ?

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 :

- <u>Event A :</u> you win the game.
- <u>Event B :</u> I win the game.



<u>Exo 1</u> - What is the probability for you to win a game after a win ?

<u>Exo 2</u> - 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 $q = 1 - p = 0.3$

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

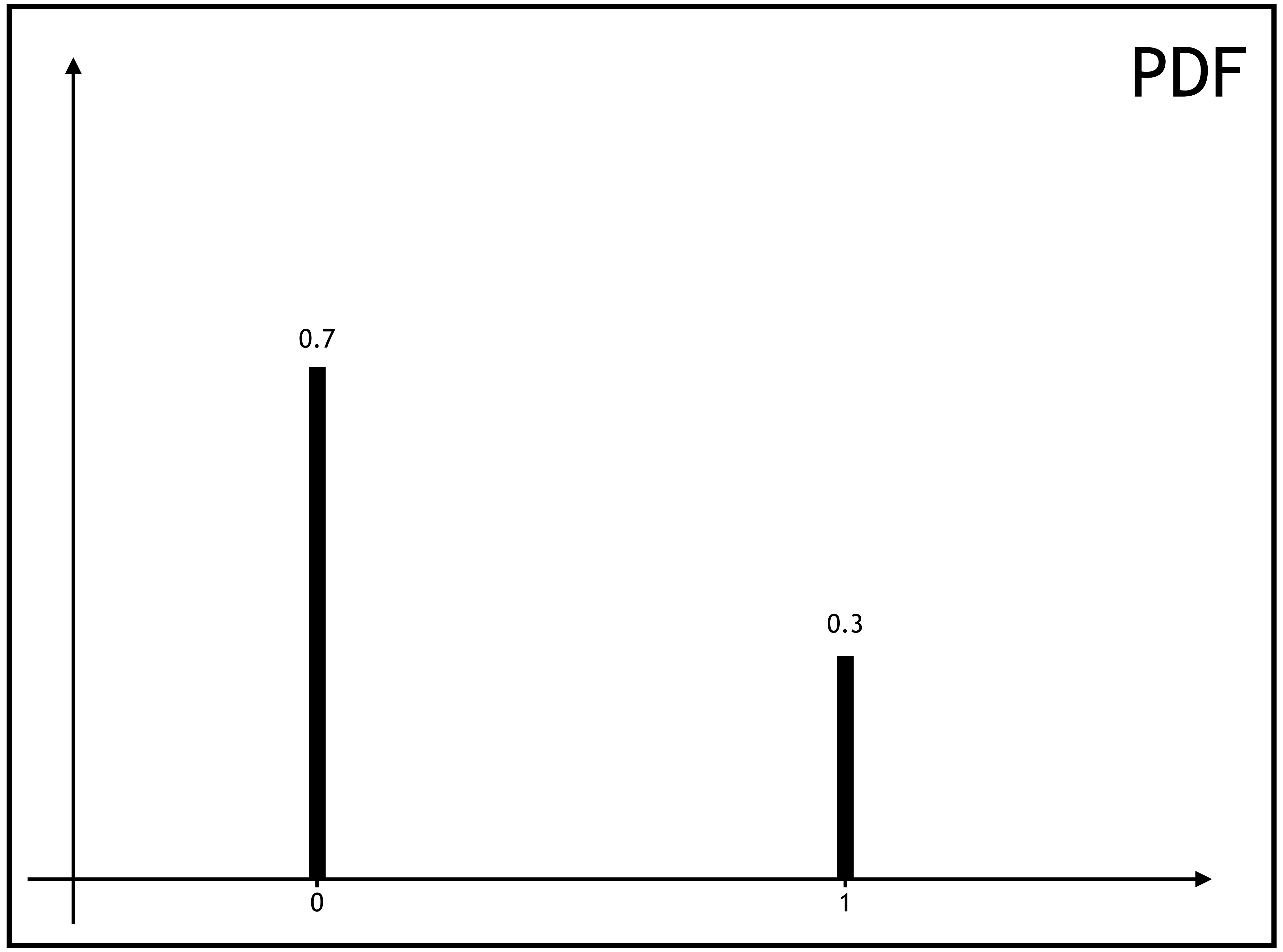
Bernoulli distribution

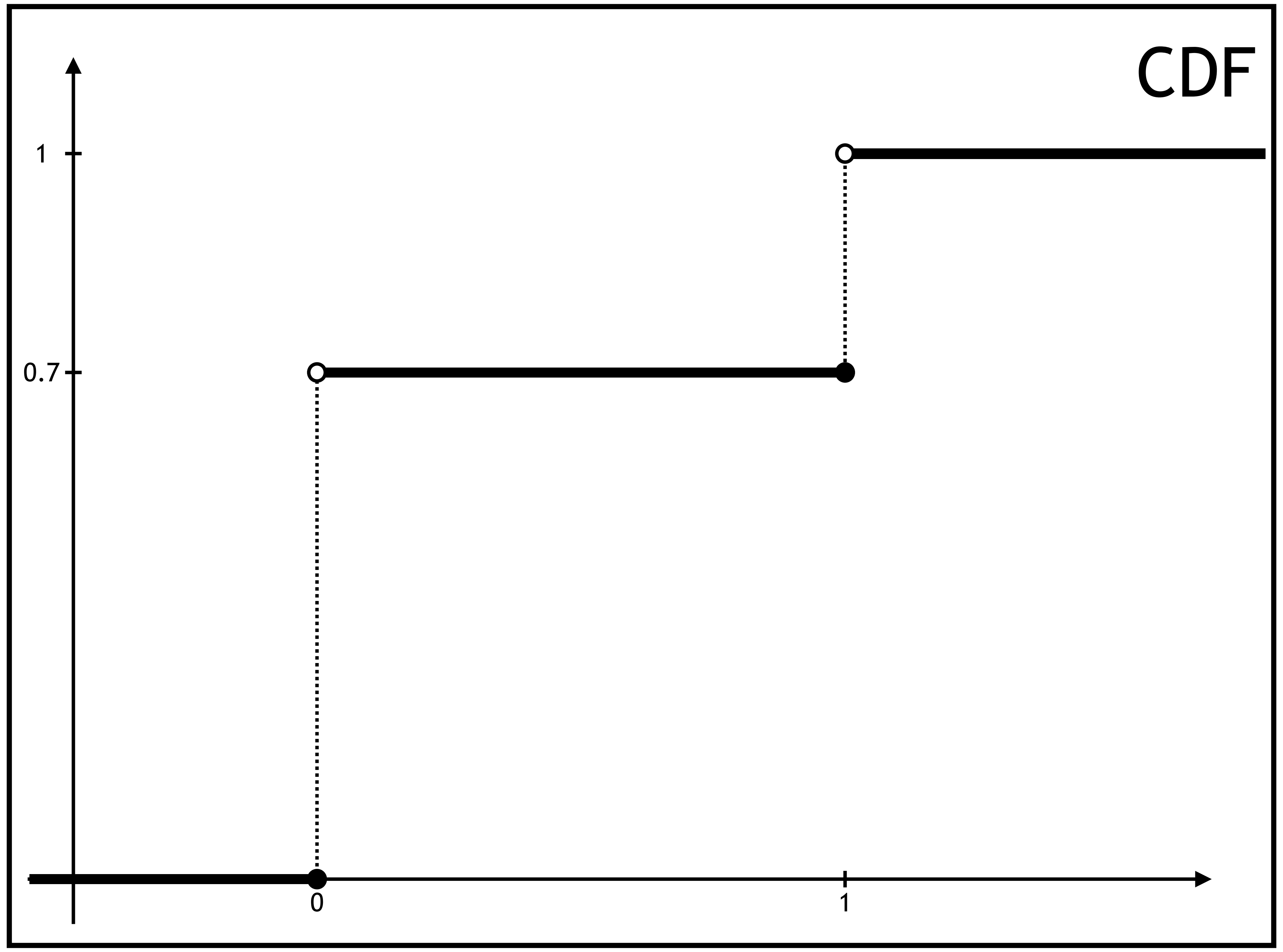
What would be the Cumulative Distribution Function (CDF) for a Bernoulli distribution ? And its Probability Density Function (PDF) ?

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

Bernoulli distribution

What would be the Probability Density Function (PDF) for a Bernoulli distribution ? And its Cumulative Distribution Function (CDF) ?





Binomial distribution

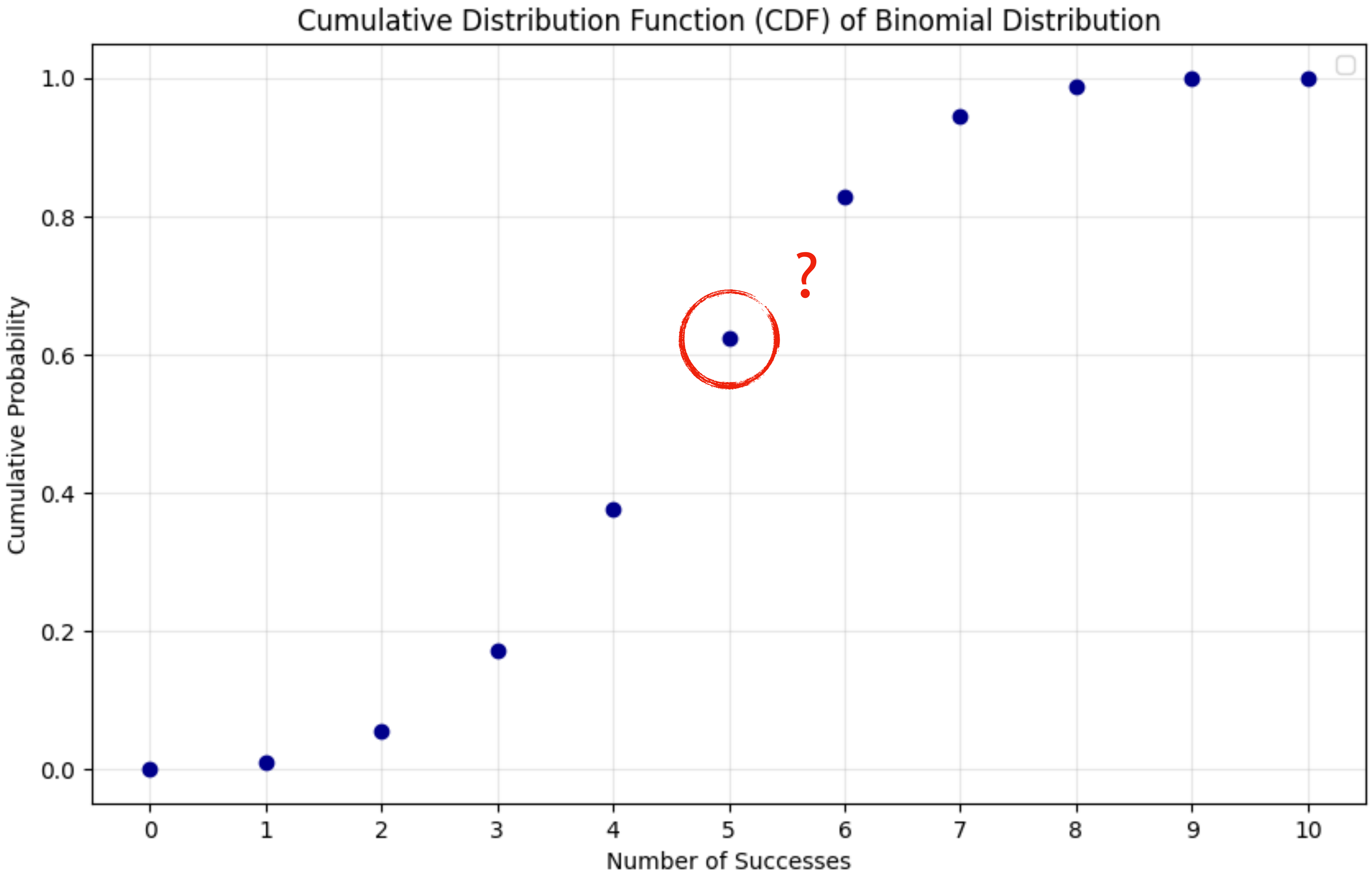
Let's take again the chess example and you play n games against me with the same probability after each game. What's the probability you win k times ?

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 ?

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

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 ?

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

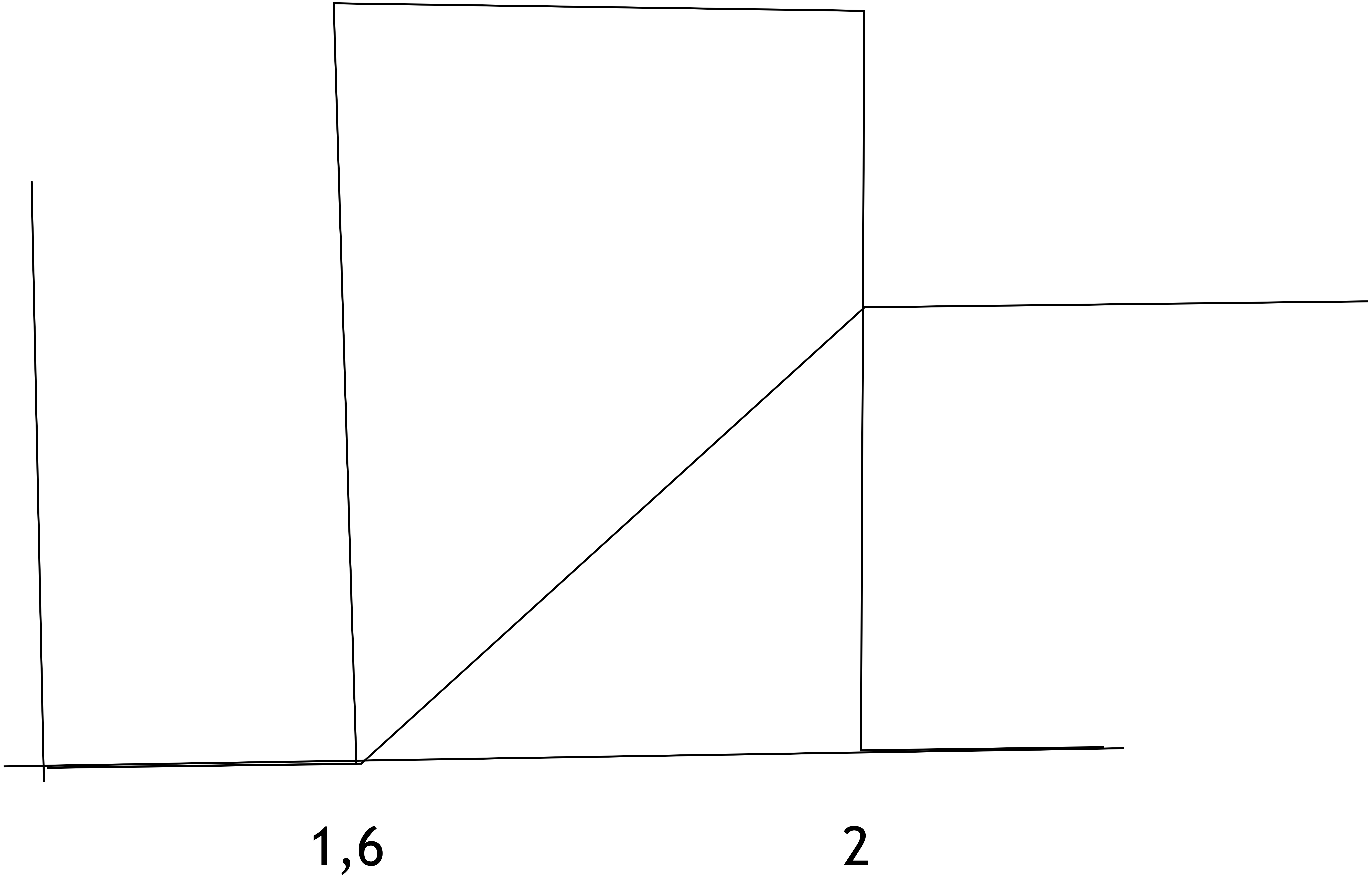
Uniform distribution

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.

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.

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


1,6
2

Normal distribution

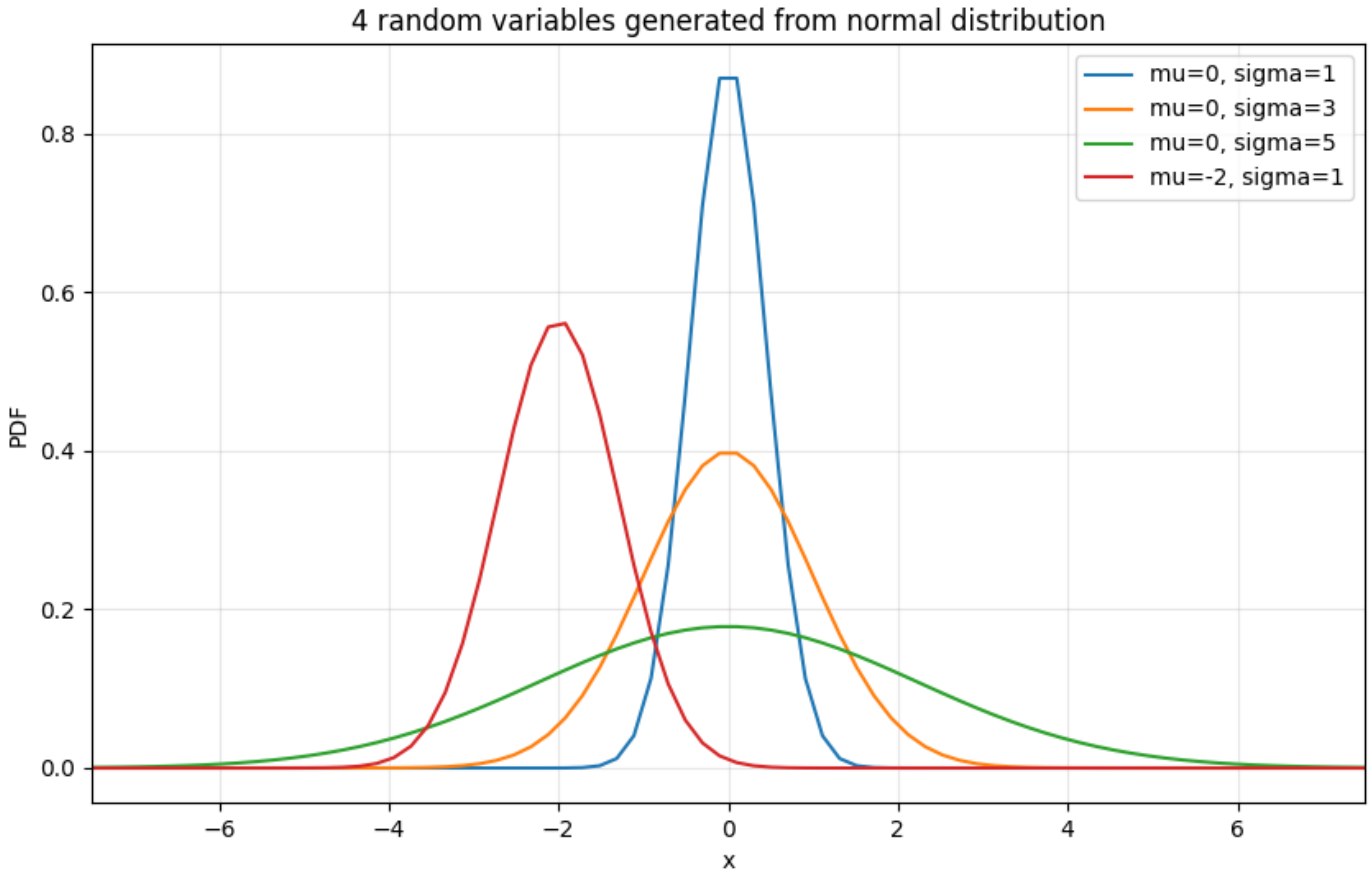
The distance of employees' houses from the company's main office can be represented by a normal distribution.

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, your waiting is an average of 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

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

Exercice for tomorrow

For tomorrow, you have one exercise from the notebook of today's exercises.

The idea for you is to use the Kaggle website to work on a real world dataset (which is already cleaned for you this time).

- Monday : Understand data structures

- Tuesday : Introduction to probability theory

- Wednesday : Central Limit Theorem, confidence intervals and test hypothesis
  - Central Limit Theorem
  - Confidence interval
  - Test hypothesis

- Thursday : Data cleaning methods & pandas manipulations

- Friday : Statistics with scikit-learn

Practice with SciPy

Exercise : To have a better understanding of how a distribution works, let's produce our own "fake data". Using the binomial distribution implementation from SciPy, generate data sample with $N_{\text{size}} = 10$ samples, following a binomial distribution with probability of success of 0.5 and plot the distribution and the cumulative distribution of this dataset.

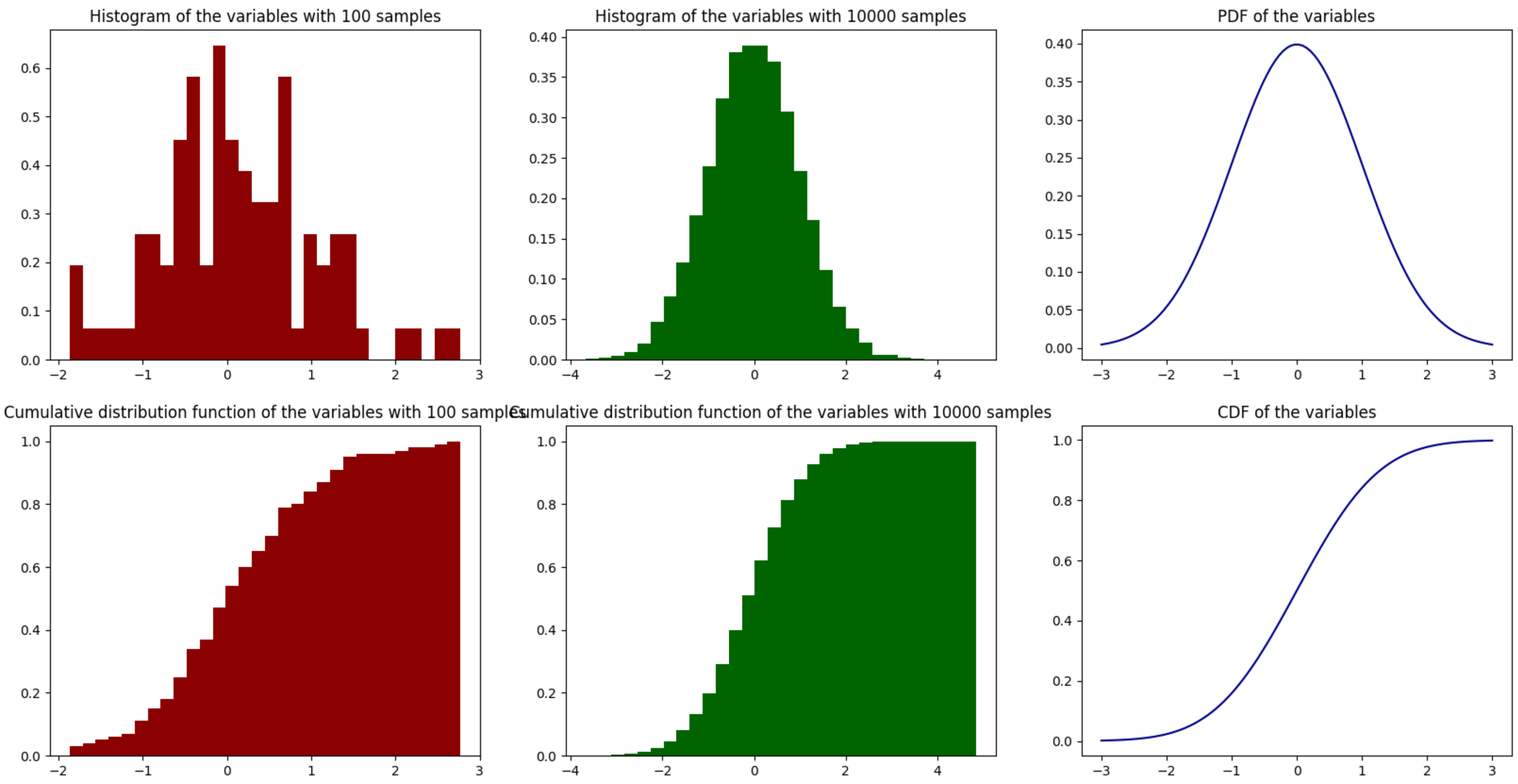
What do you observe ? And with $N_{\text{size}} = 100, 1000$ and 10000 ? What happen if you change the probability ?

Do the same with the normal distribution now and compare the two different distribution.

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: $\bar{X} = \dfrac{1}{5}\sum_{i=1}^{5} X_i$.

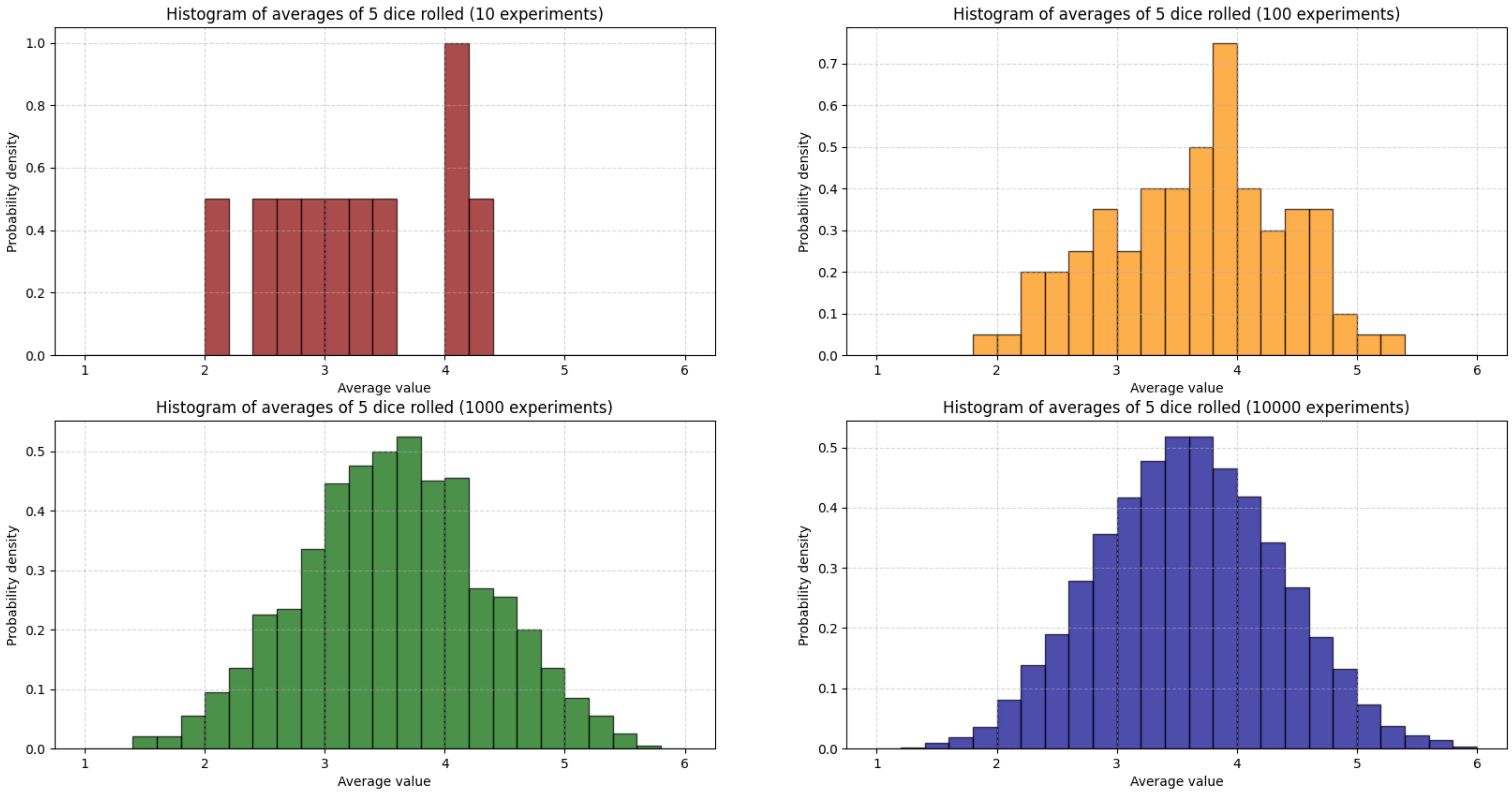
Question 2: Using the library SciPy, build a function that returns the average of the experiment describe above.

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 $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

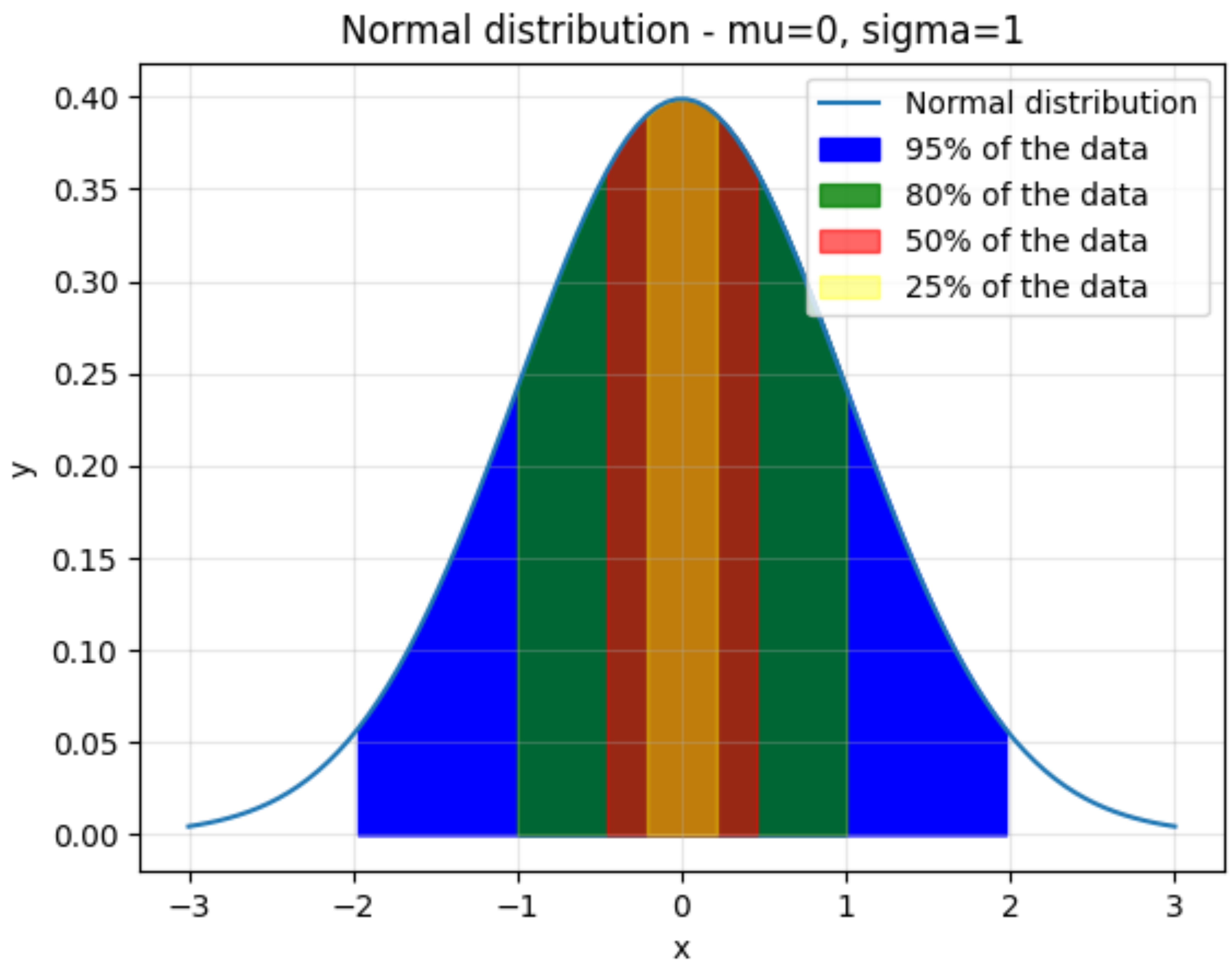
Confidence intervals

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

<u>Definition :</u> 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

<u>Interpretation :</u> « 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

<u>Question :</u> 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` with a standard deviation 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 and n is large enough, which is not always the 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.

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

Test hypothesis

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

$$H_0 : \mu = \mu_0$$

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

Two *sided* - $H_1 : \mu \neq \mu_0$

One 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

Definition : 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.

Exercises

During the course : Only the exercises 1 & 2.

At home : All the remaining exercises.

- Monday : Understand data structures
- Tuesday : Introduction to probability theory
- Wednesday : Central Limit Theorem, confidence intervals and test hypothesis
- Thursday : Data cleaning methods & pandas manipulations
  - Standardisation
  - One-hot encoding & data formatting
  - Remove non-valid values
- 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

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

To delete a row (i.e. a specific data) in pandas : pd.drop(idx)

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.

The one-hot encoding function in pandas is : **pd.get_dummies(df)**

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

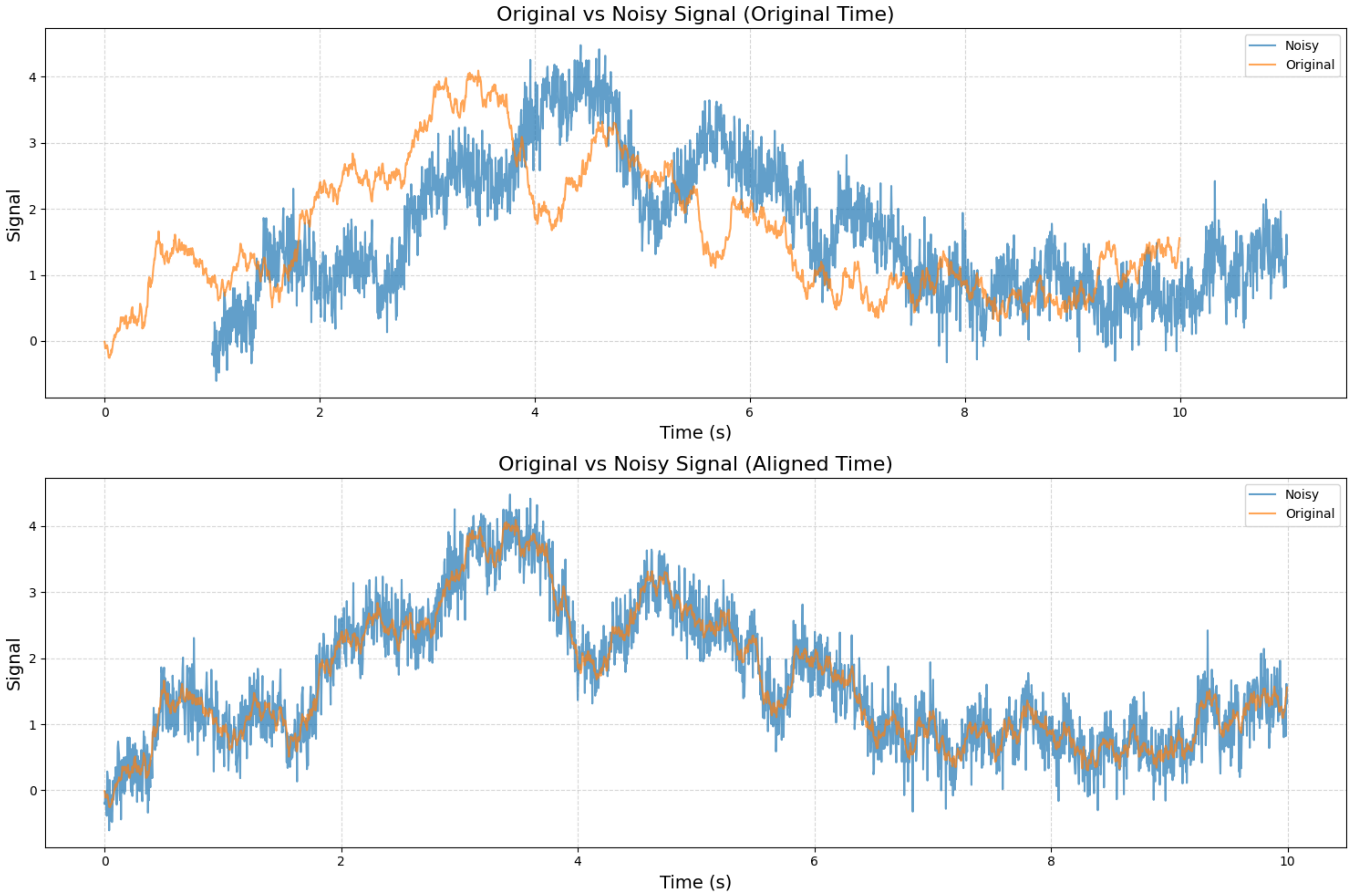
Some normalisation formulas

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

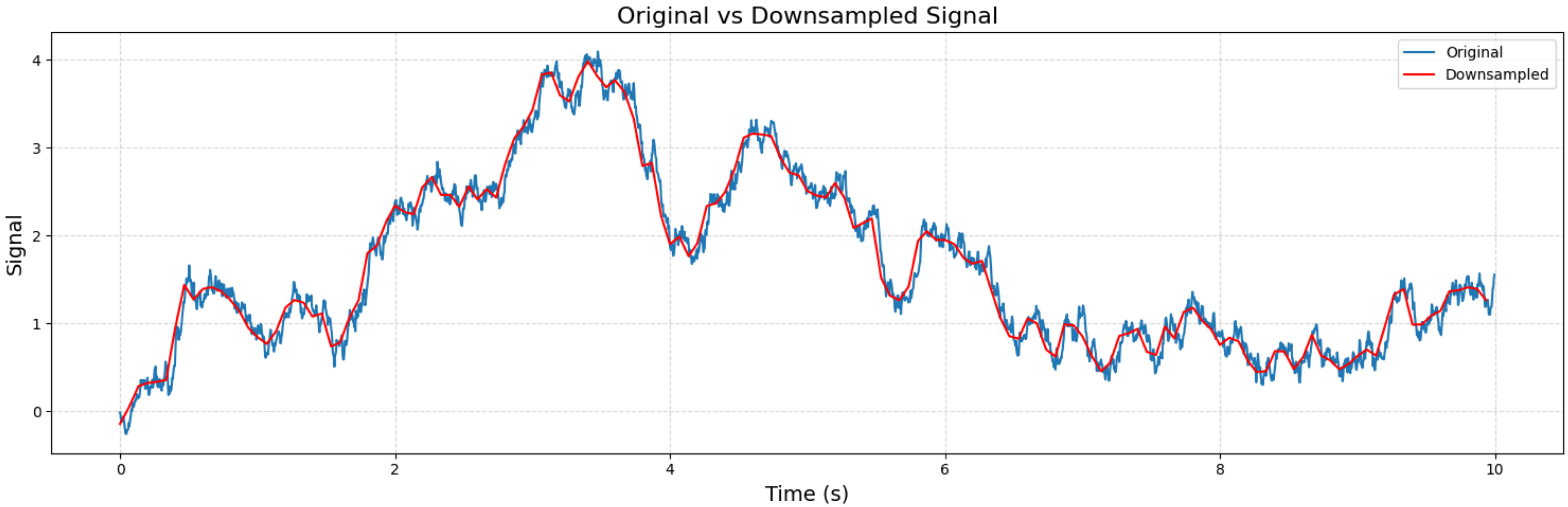
2 - <u>*Standardisation :*</u> $\tilde{x}_i = \dfrac{x_i - \mu}{\sigma}$

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

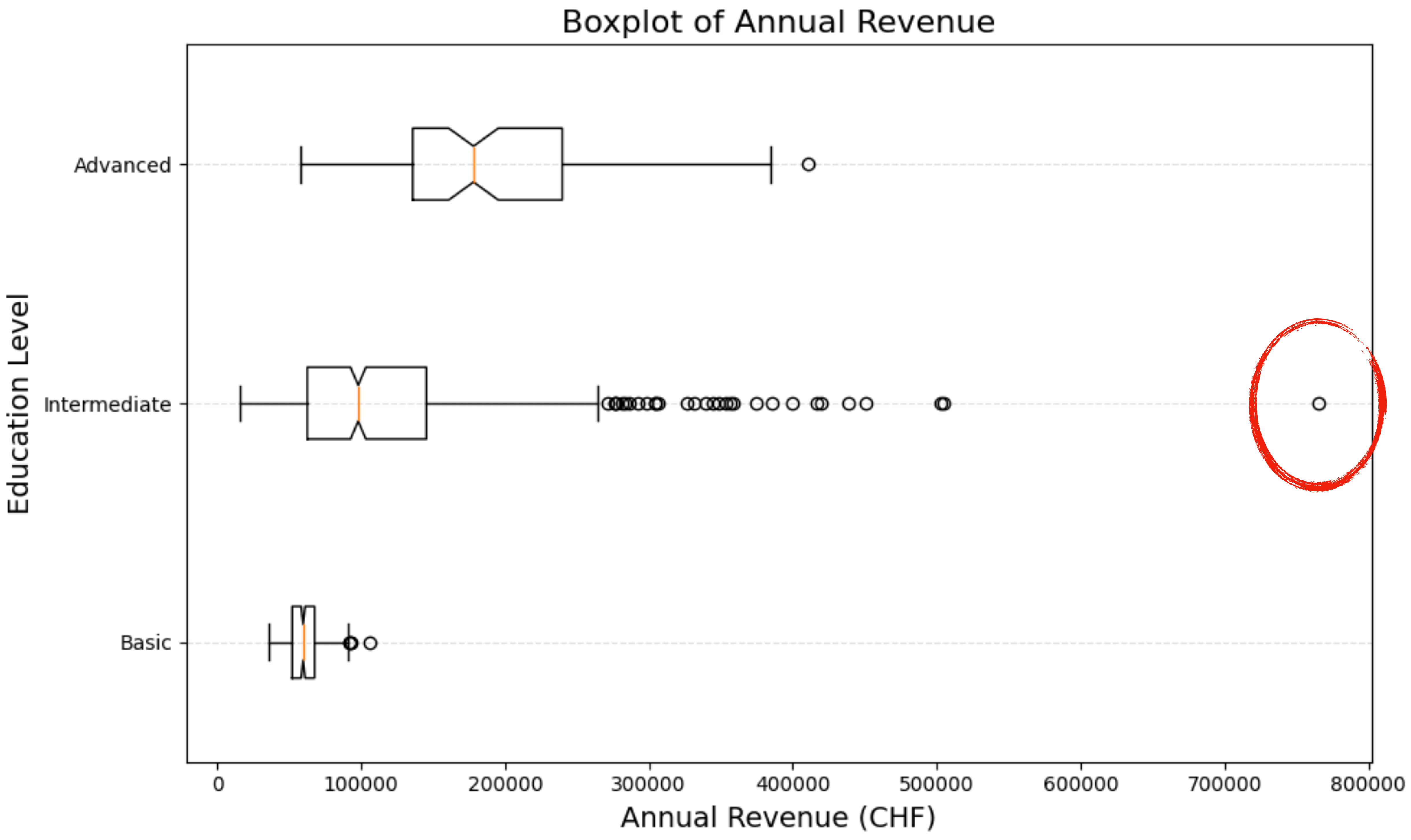
Synchronisation of time series



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

Let’s practice some data cleaning

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:

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

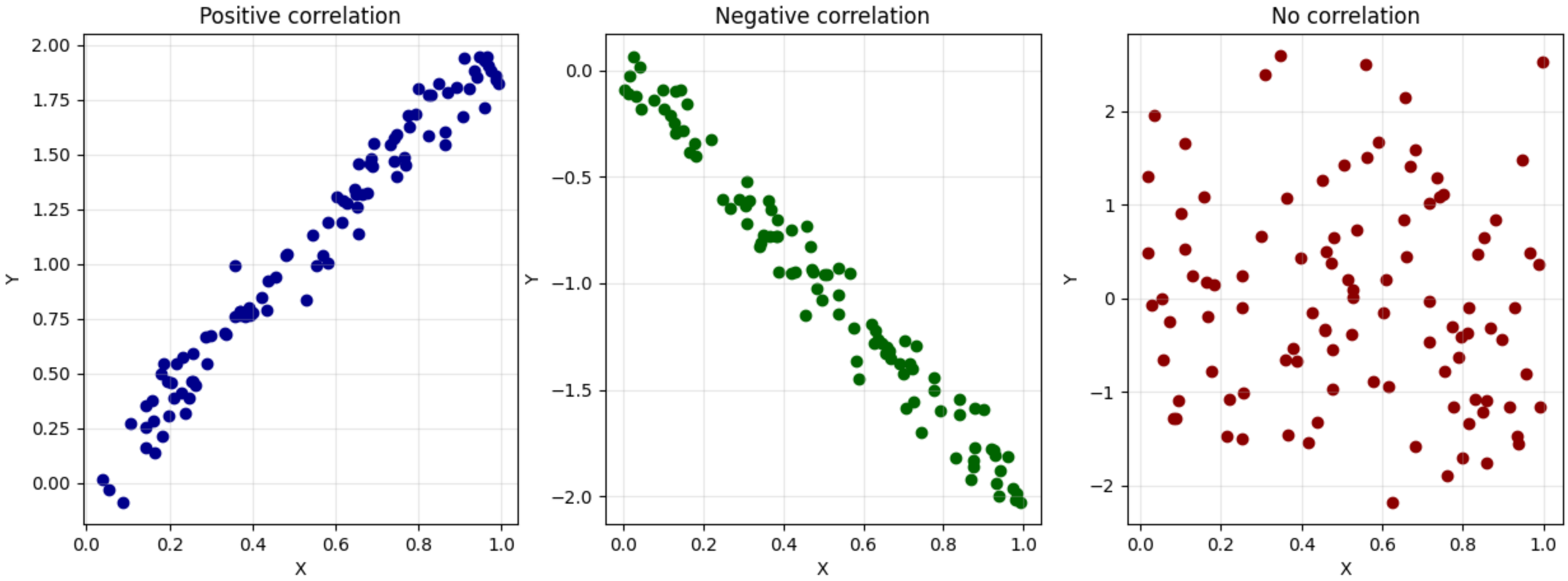
This values is located between -1 and 1.

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 :

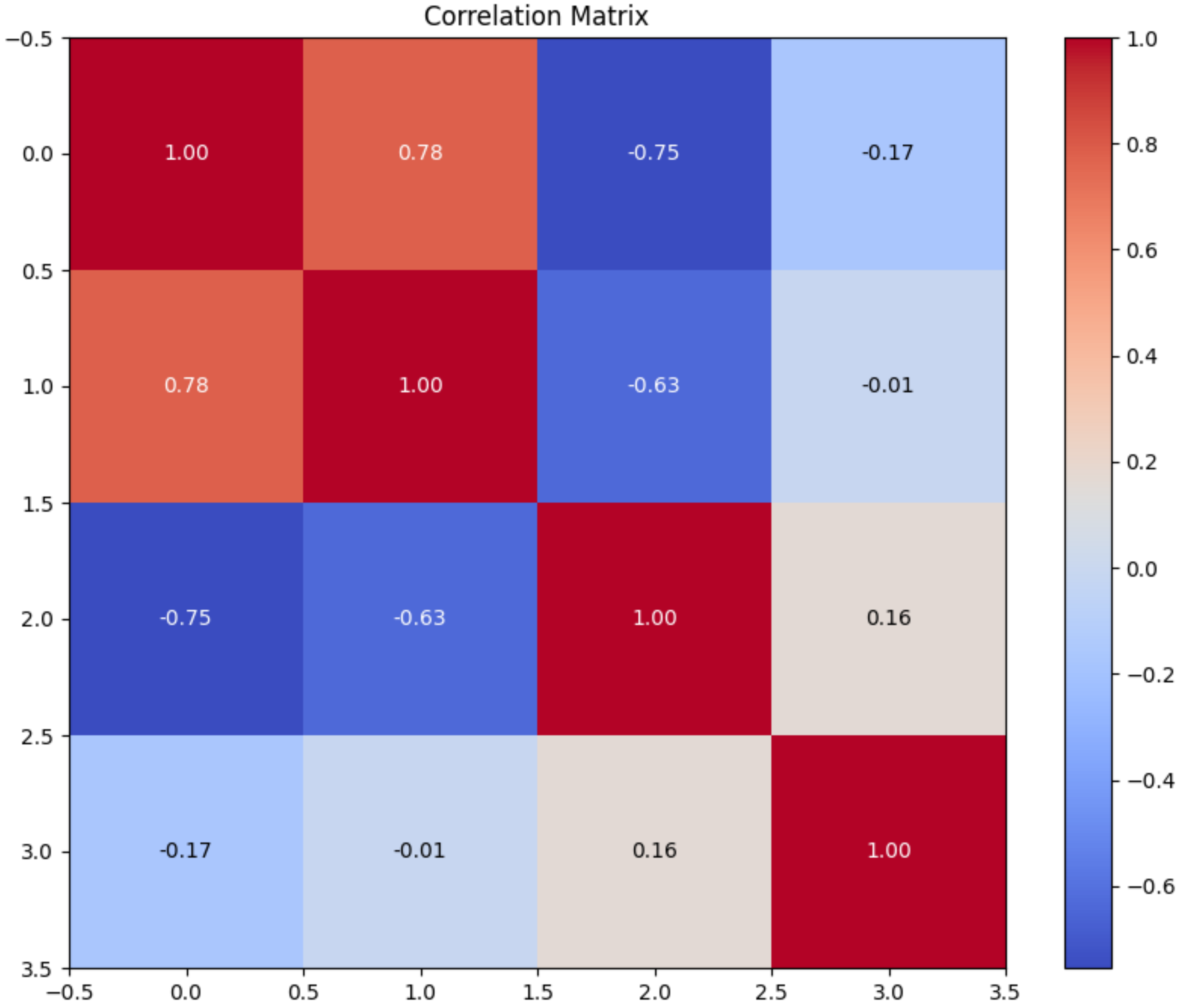
$$\rho_{ij} = \frac{\text{Cov}(X_i, X_j)}{\sigma_{X_i}\sigma_{X_j}},$$

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.

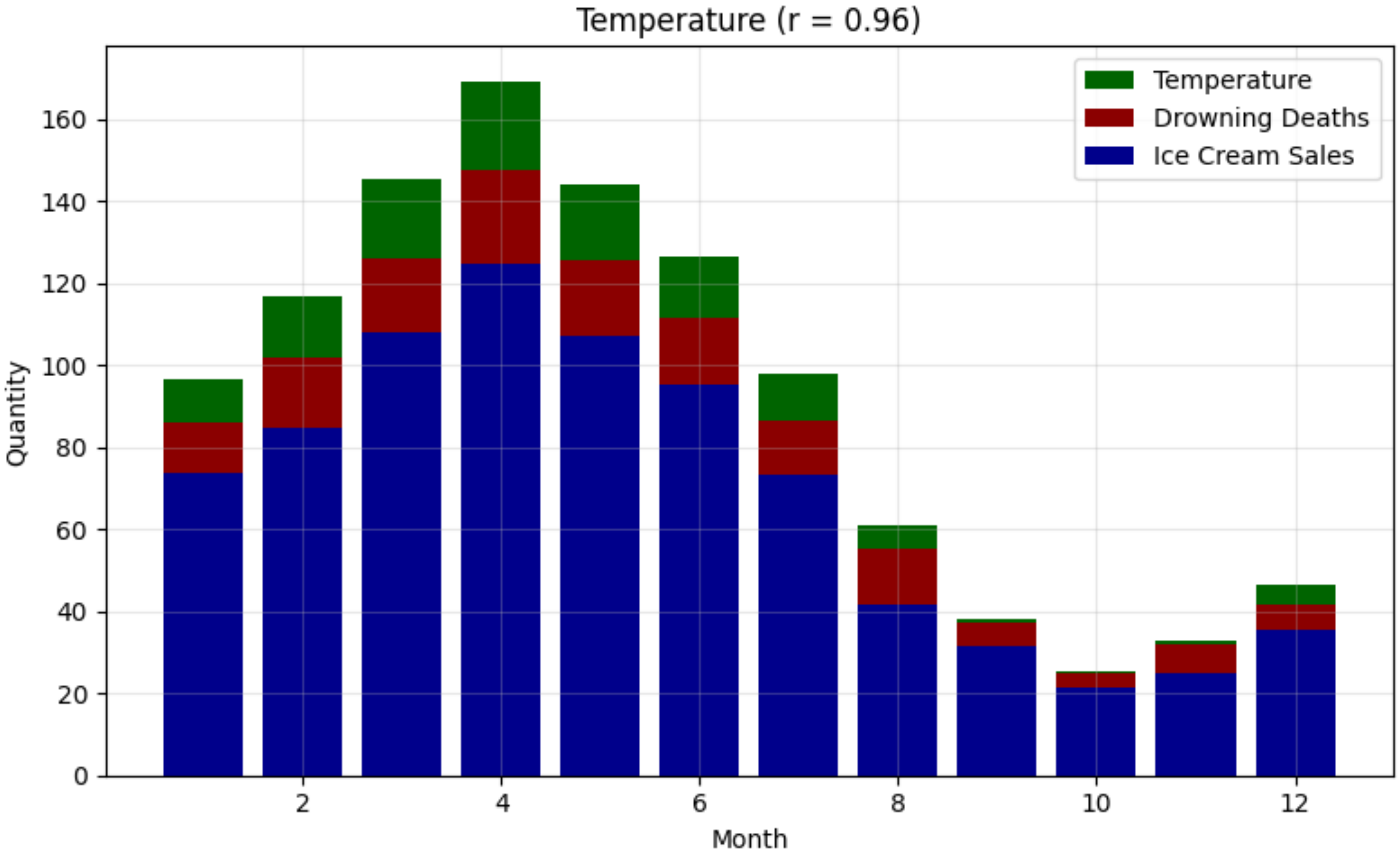
=> Maybe more productive people are just more likely to stay awake with coffee.

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 drowning deaths ?

Question 2: and `ice_cream_sales_drowning_deaths_temperature.csv` ?

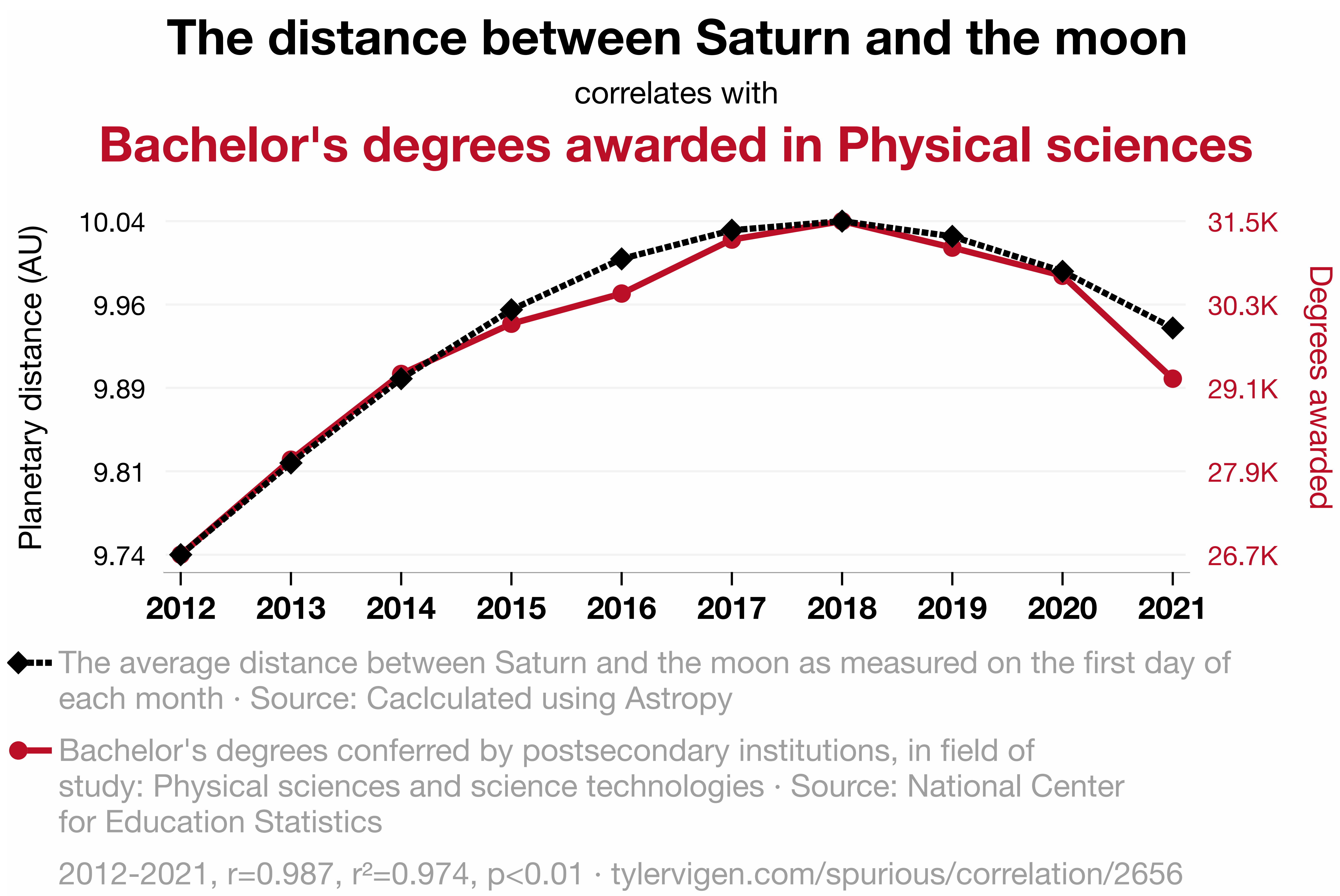
Correlation matrix
- Cofounding variables -



Spurious correlations

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

Spurious correlations



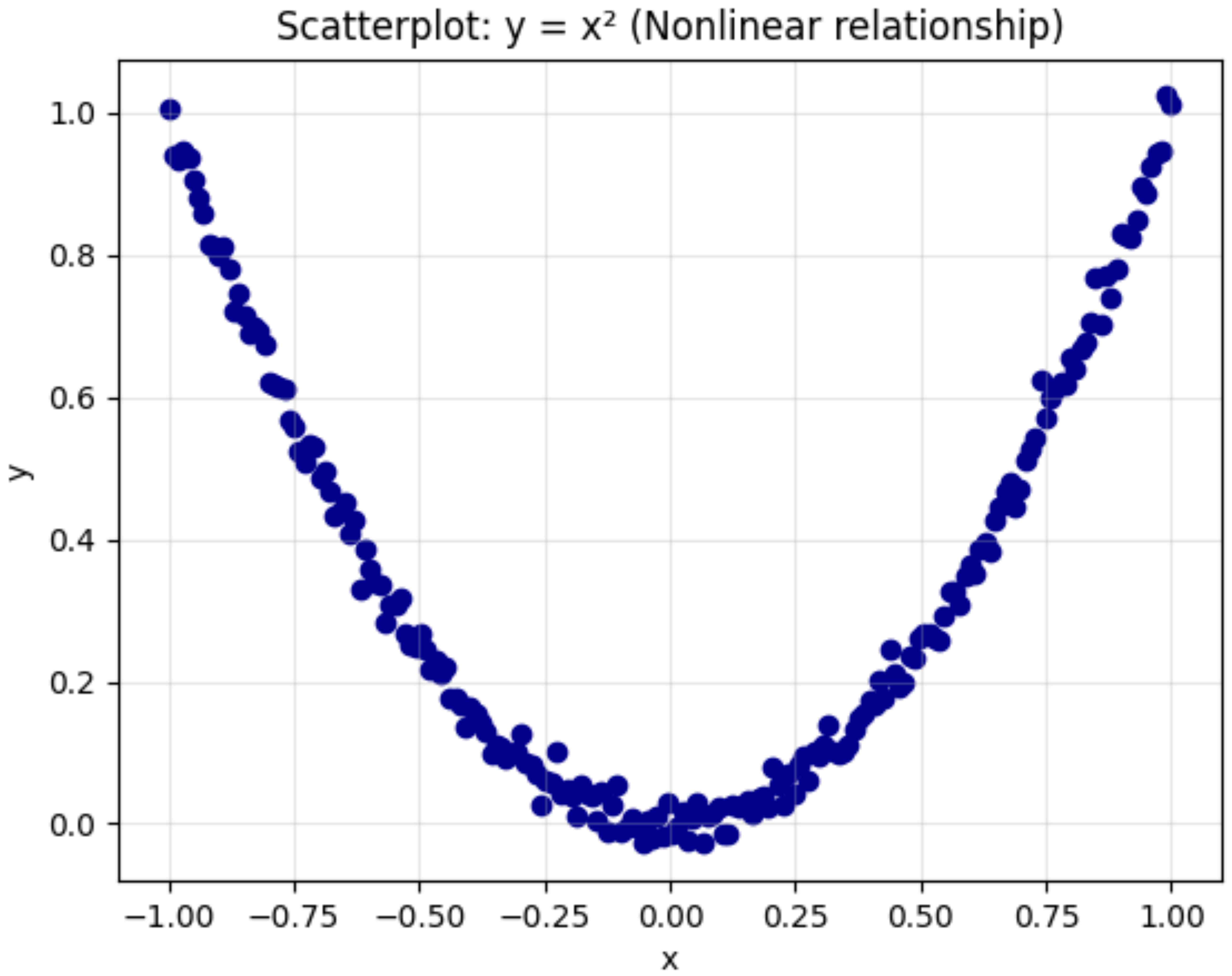
References

Non-linear relationships

Question 1: In the file `non_linear_relationships.csv`, compute the Pearson's correlation coefficient between variables x and y.

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

Non-linear relationships



- Monday : Understand data structures

- Tuesday : Introduction to probability theory

- Wednesday : Central Limit Theorem, confidence intervals and test hypothesis

- Thursday : Data cleaning methods & pandas manipulations

- Friday : Feature selection & model evaluation
  - Feature selection
  - Model evaluation
  - Cross validation

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
- Avoids multicollinearity issue

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 -

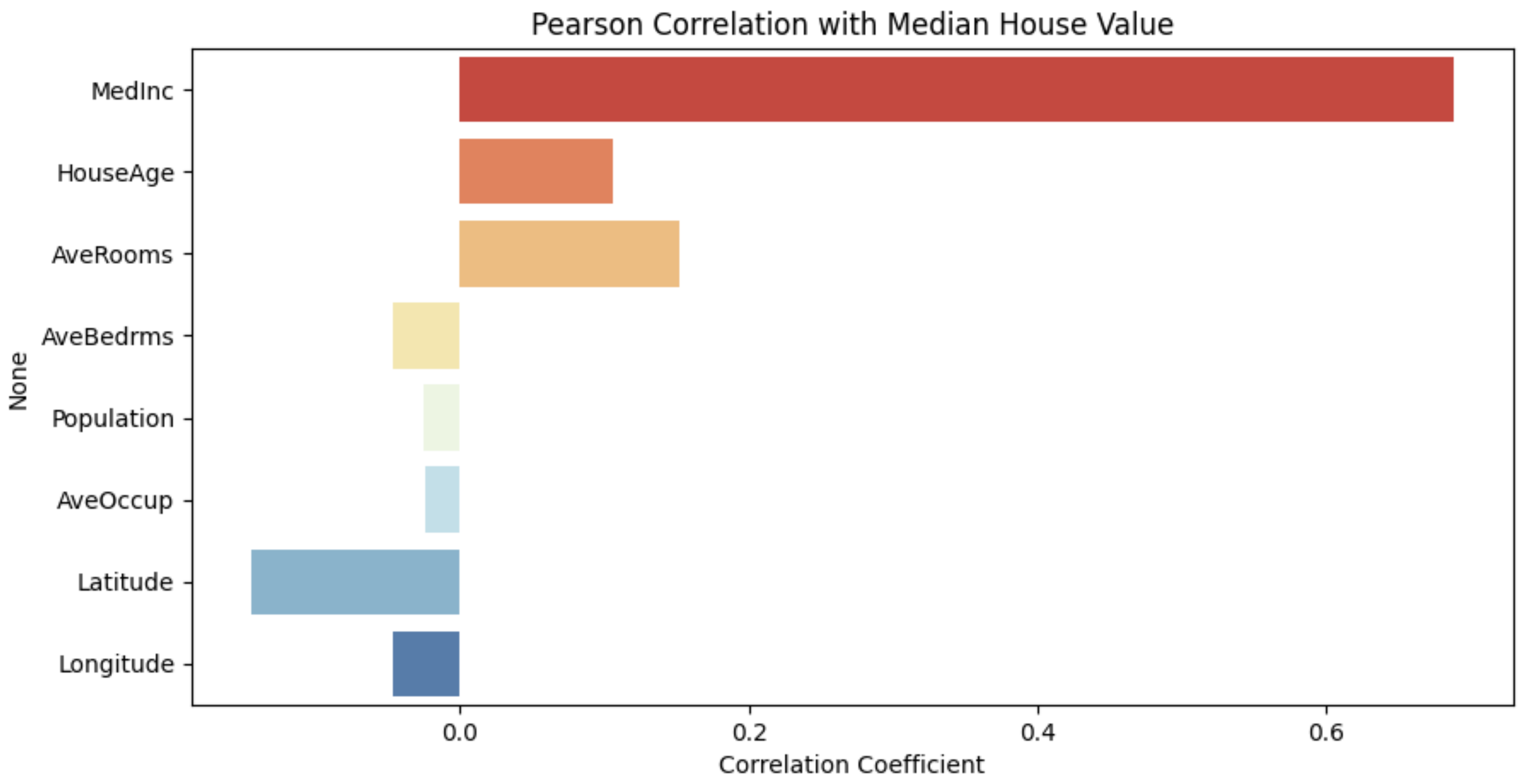
<u>Idea :</u> The idea is to <u>rank all the features</u> using their statistical relationship with the target variable you're trying to predict.

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

Exercice : 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
- Wrapper methods -

Idea : You start with a model, usually easy to understand and train, then 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 -

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 ?

=> Using model performances like accuracy, adjusted R^2 or AIC & BIC !

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 : under-fitting risks !
- ✦ Selection is made only on the train set
- ✦ 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:

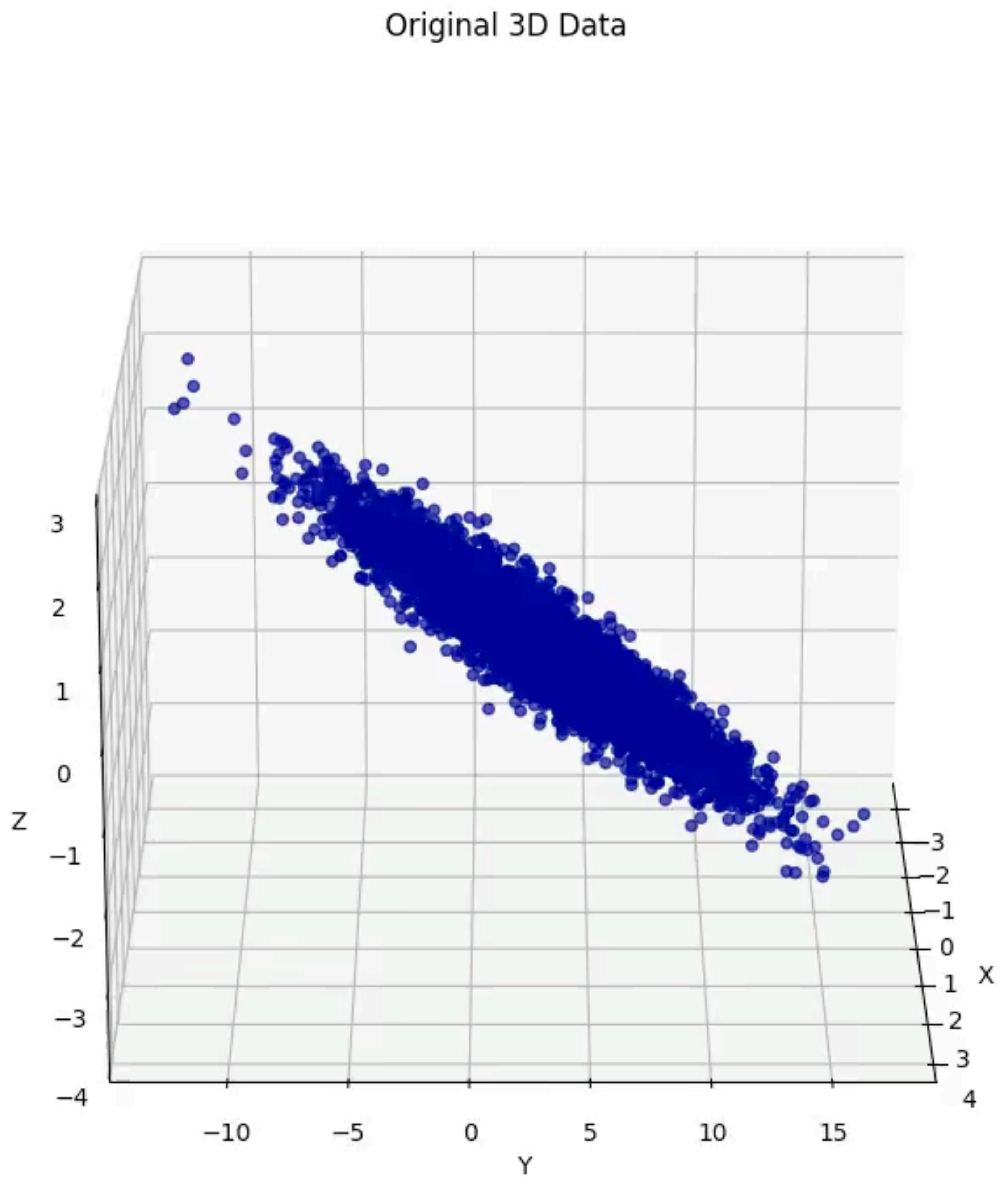
- Filter methods: Pearson’s correlation
- Wrapper methods : Forward
- Wrapper methods : Backward

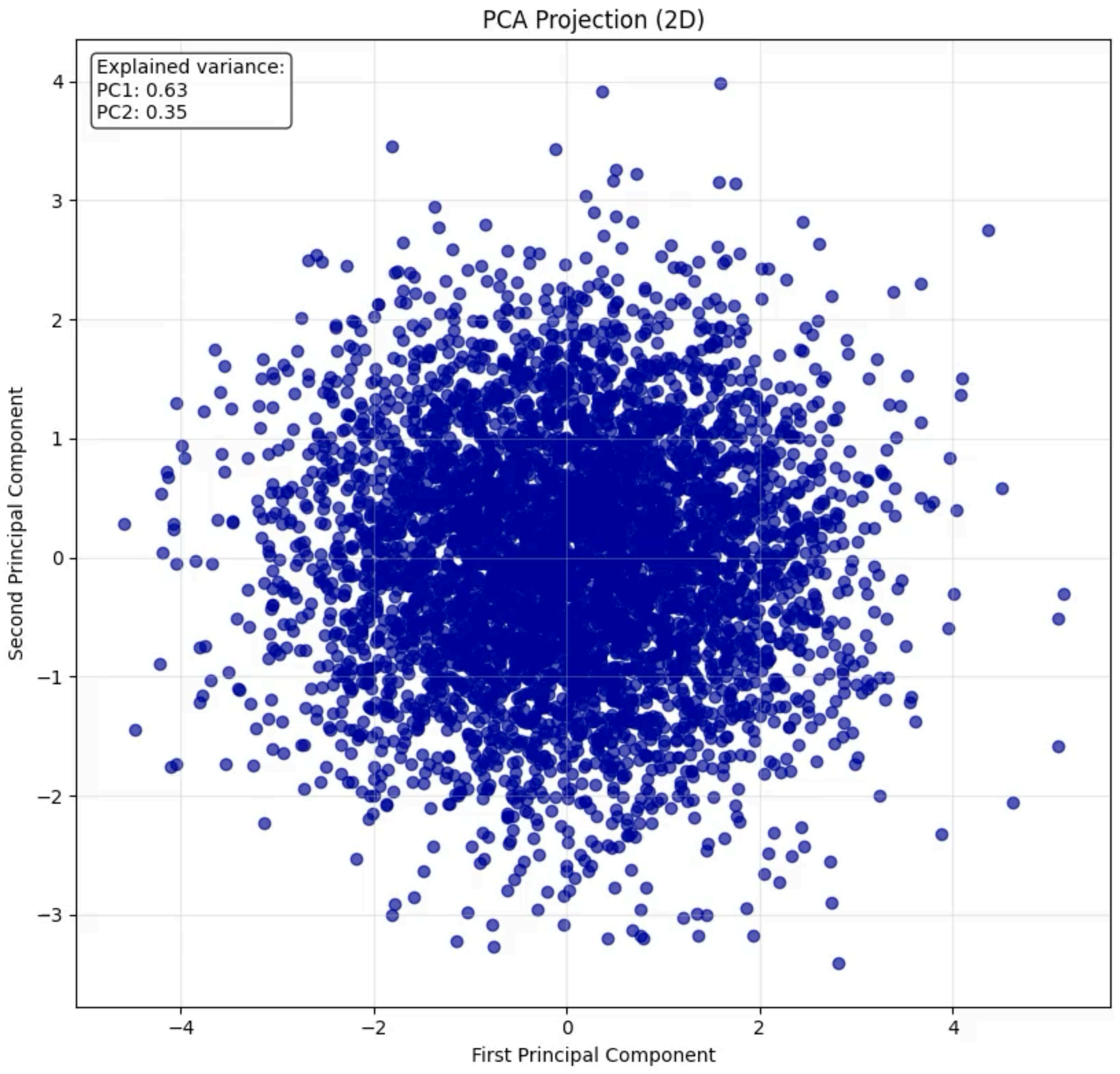
One las feature selection method
- PCA -

<u>Definition :</u> Principal Component Analysis (PCA) is an <u>unsupervised</u> linear transformation technique which <u>projects high-dimensional data into a lower-dimensional space</u>.

- Intuitively it keeps directions (components) with most variance
- Method commonly used in compression, visualisation, denoising and preprocessing data
- It reduces noise and multicollinearity

Feature selection
- PCA -





Feature selection
- PCA -

What is the process behind PCA ?

➡ Center the data by subtracting the mean
➡ Compute the covariance matrix
➡ Compute the eigenvectors and eigenvalues
➡ Project the data onto the subspace spanned by the top-k eigenvectors (i.e., the principal components corresponding to the largest eigenvalues)

Feature selection
- PCA -

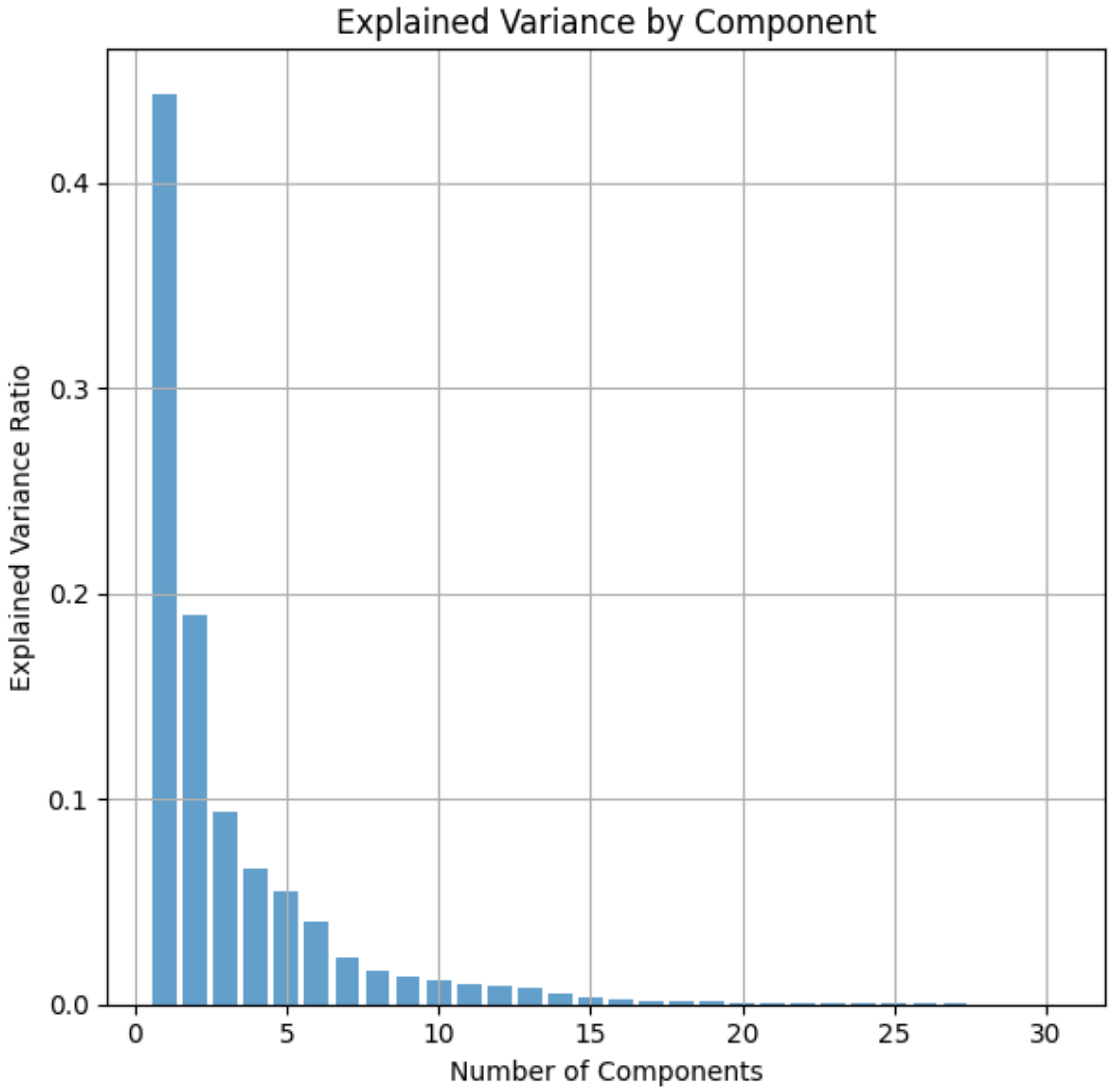
<u>How many compenents should I keep for my model ?</u>

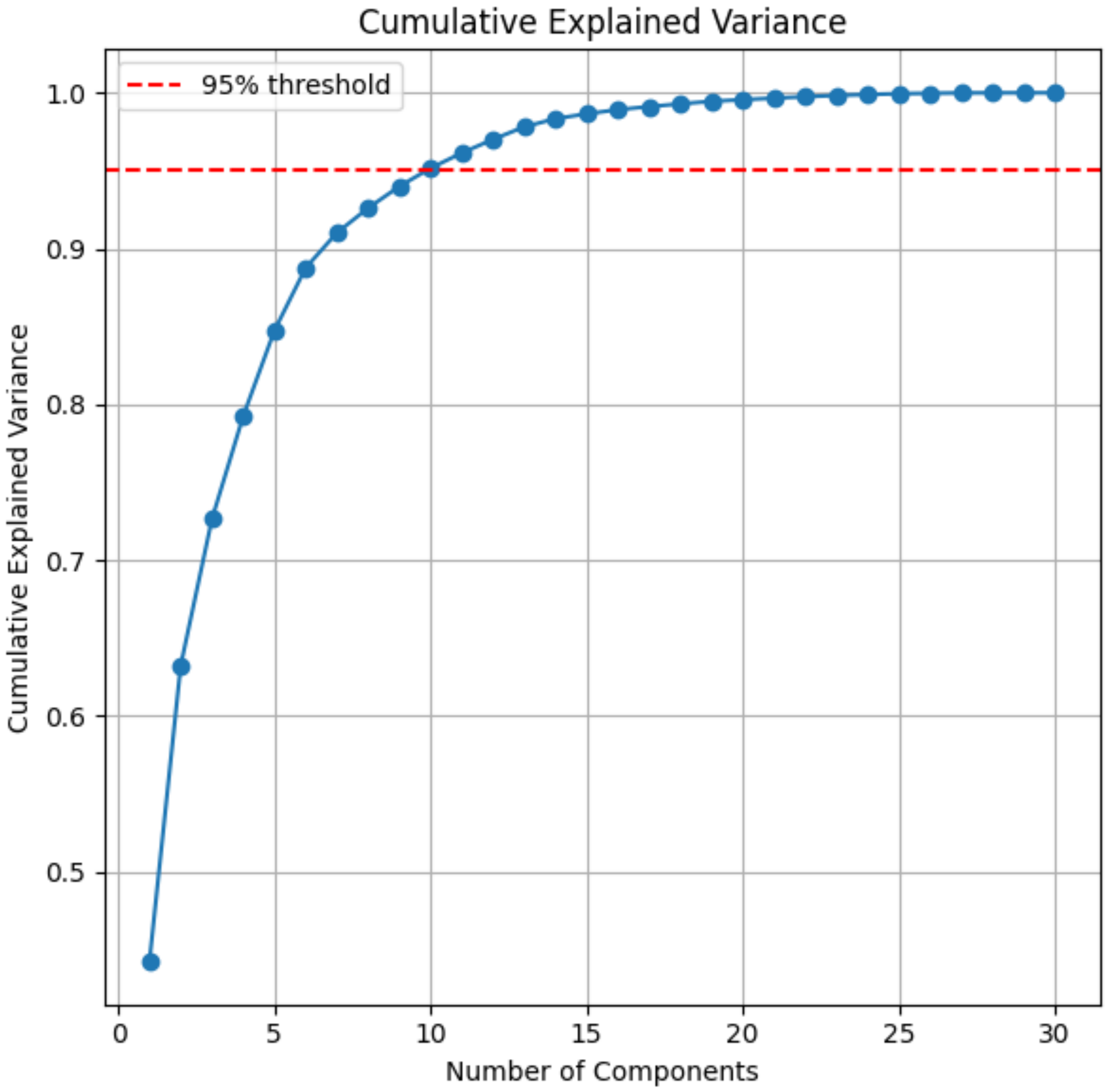
➡ Compute the explained variance ratio (i.e. eigenvalues ratio of each component)

➡ Order the explained variance ratio and select until reaching a satisfying ratio of explained variance ($95\,\%$ or $99\,\%$).

<u>Exercice :</u> What is the number of principal components required to get a $95\,\%$ of cumulative explained variance ration on the `load_breast_cancer` data set ? And for $99\,\%$?

Feature selection
- PCA -





Limitations of PCA method

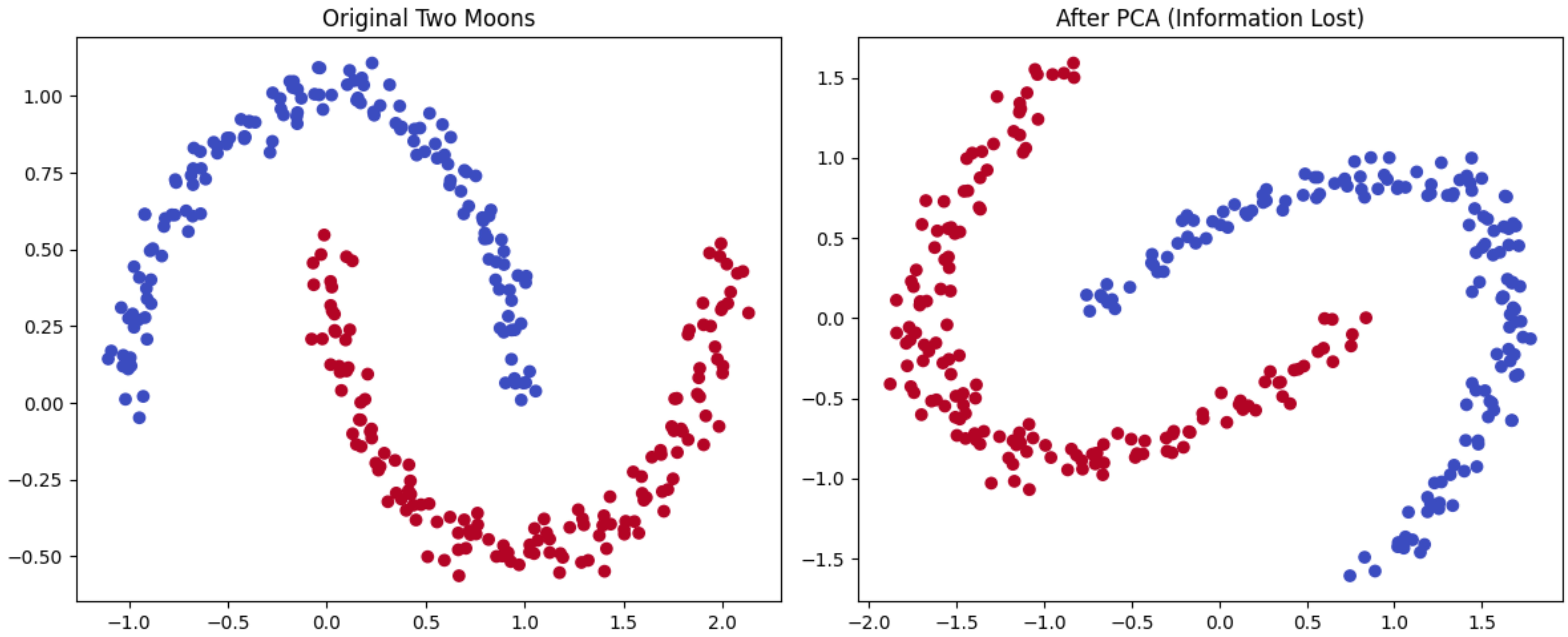
- Assumes linearity and large variance

- Hard to interpret components:

$$PC_i = a_1 x_1 + a_2 x_2 + \ldots + a_n x_n$$

- Not suitable for categorical data => break the assumption of continuous variance

Limitations of PCA method



Cross-validation

Problem : You have a data set with n samples and for the purpose of your study you have to present the results.

But you can not present the results of your model directly from you whole data set ! *Risk :* results biased by your training set.

What is the solution ?

Cross-validation

<u>Solution :</u> You have to divide your data set into <u>training</u> and <u>testing</u> sets.

- What is the ratio to divide your data set ?

- You trained your model on your training set and you go for the final results on the test sets but… your model is crap on testing set ! What do you do ?

Cross-validation

Solution : You have to divide your data set into training and testing sets.

- What is the ratio to divide your data set ? -> 80% - 20%

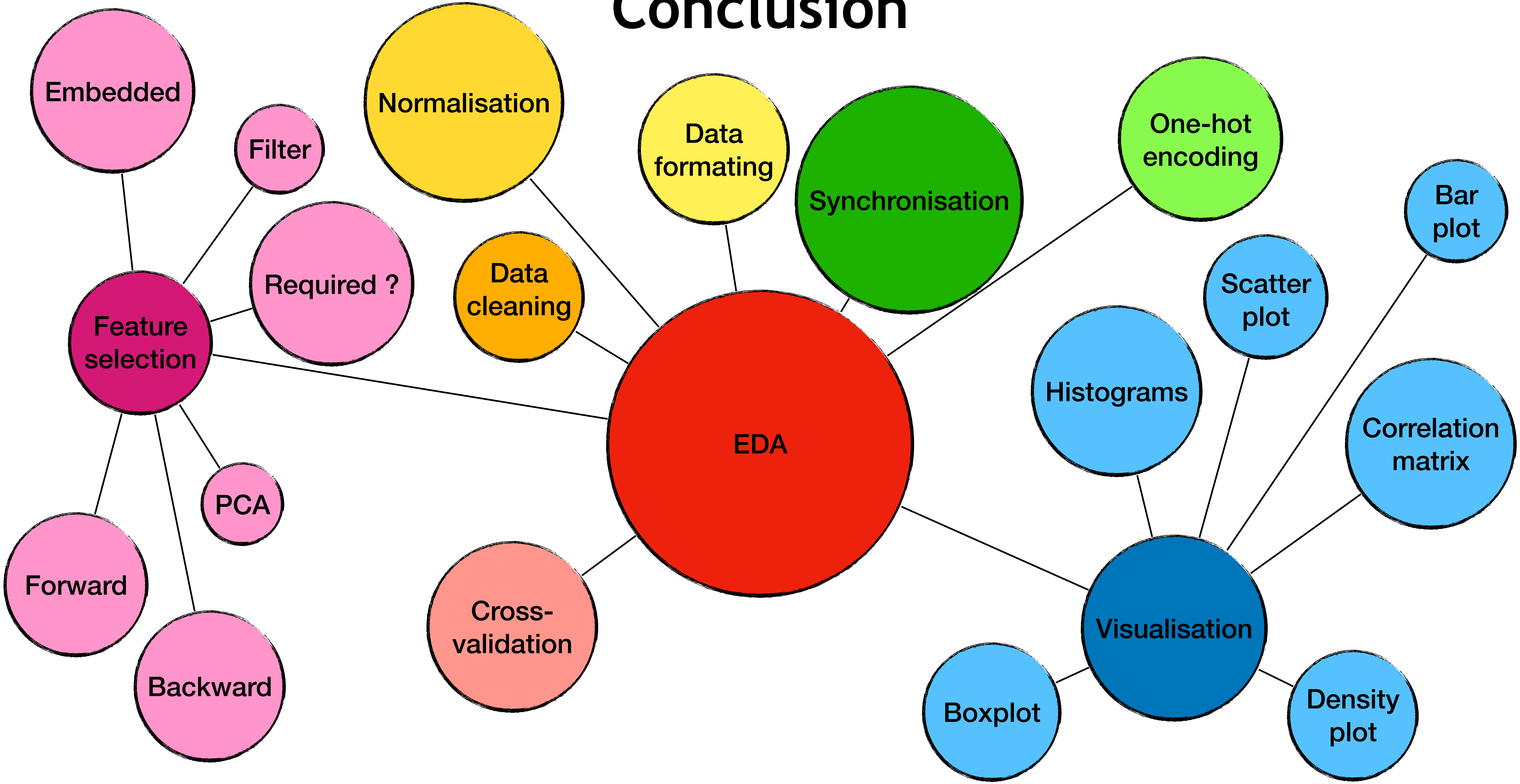
- You trained your model on your training set and you go for the final results on the test sets but… your model is crap on testing set ! What do you do ? -> Add a validation set ! 60% - 20% - 20%

Cross-validation

<u>Question :</u> How do you select your data?

=> We select them randomly, while carefully ensuring that the subclass distribution is preserved or at least taking in consideration.

Conclusion


Embedded
Filter
Normalisation
Data formating
Synchronisation
One-hot encoding
Bar plot
Required ?
Data cleaning
Scatter plot
Feature selection
Histograms
EDA
Correlation matrix
PCA
Forward
Cross-validation
Visualisation
Backward
Boxplot
Density plot