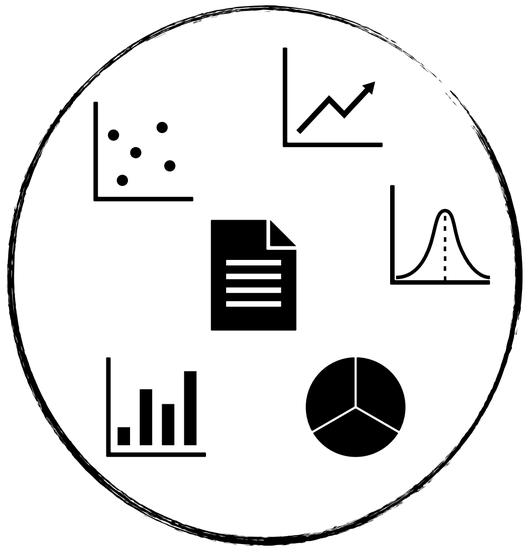
Descriptive statistics

Applied Data Analysis (ADA) - October 2025

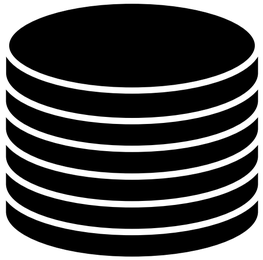
Nomades Advanced Technologies
Gaspard Villa

- 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 : Feature selection and correlation matrix
- Friday : Statistics with scikit-learn

How a project is built ?



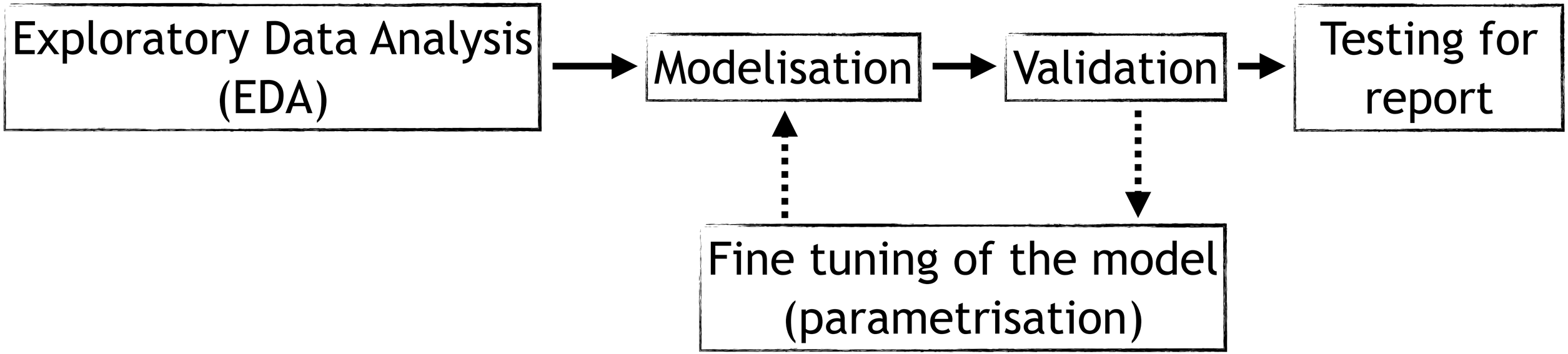
Raw data



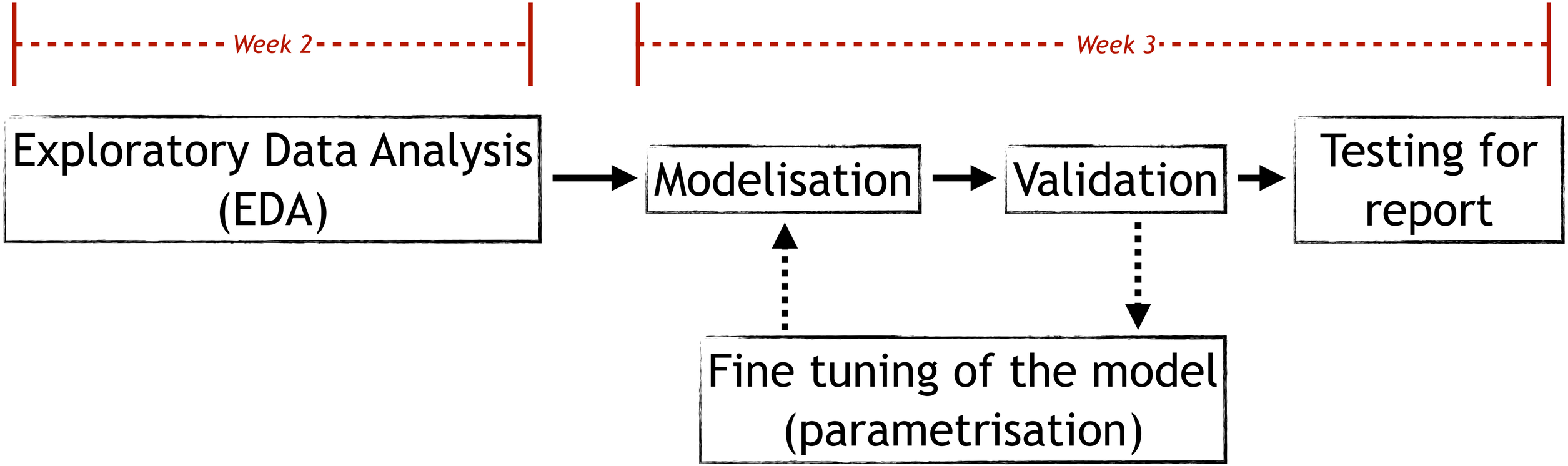
Workable dataset

Modelisation, predictions, hidden informations, etc…

How a project is built ?



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.

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 table.

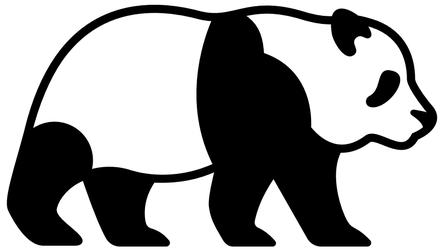
=> Do you know a nice tool/library in Python to manipulate tables ?

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.

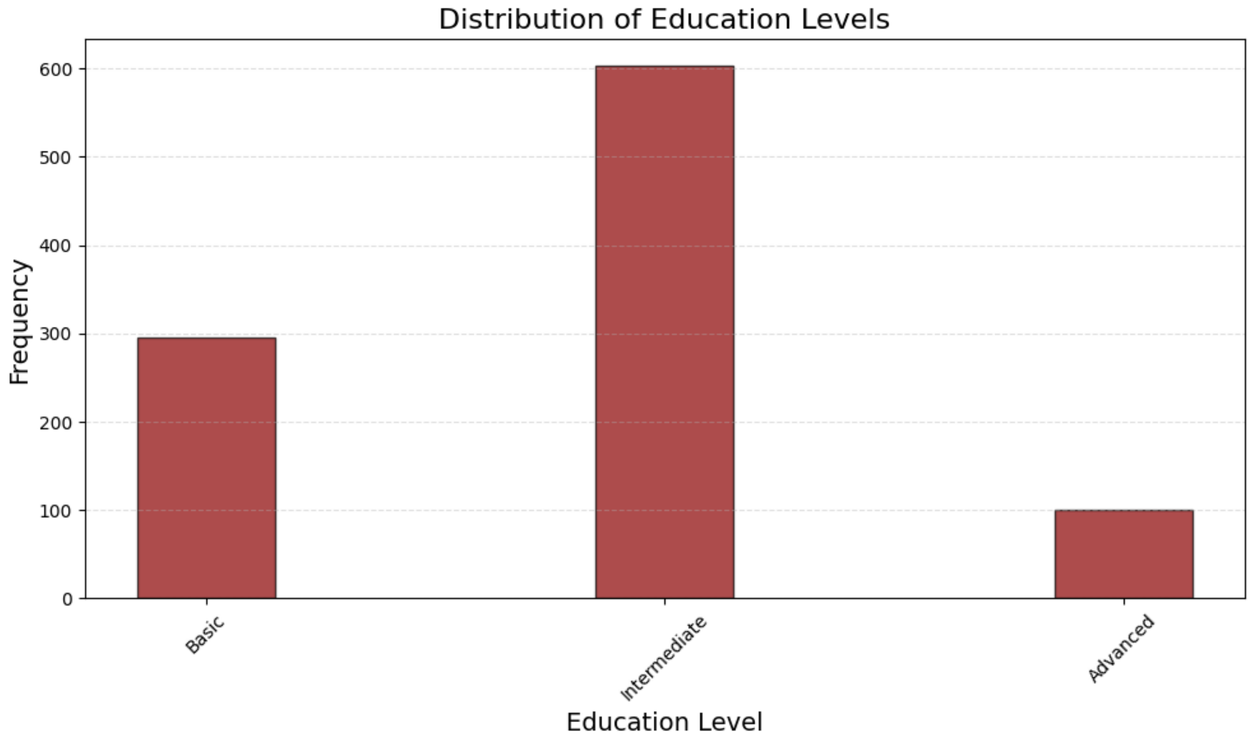
Key for analysis is Visualisation
- See your data -

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

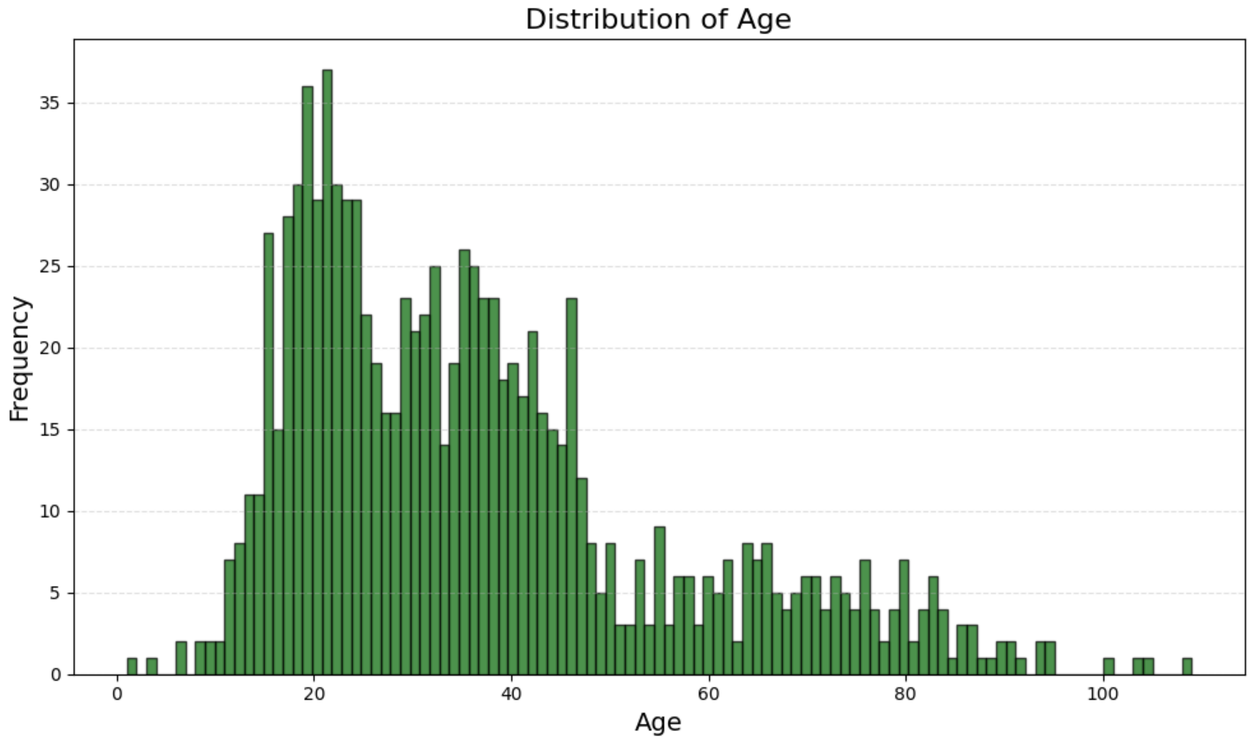


References

Key for analysis is Visualisation
- Univariate analysis -

2 - Univariate analysis for continuous variables

Histograms

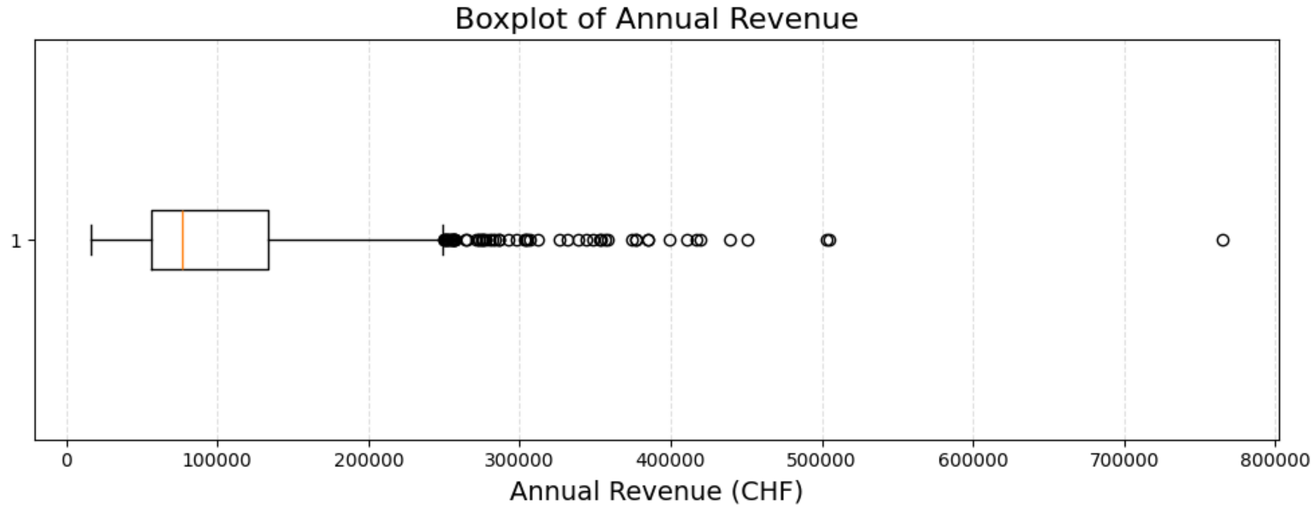


References

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

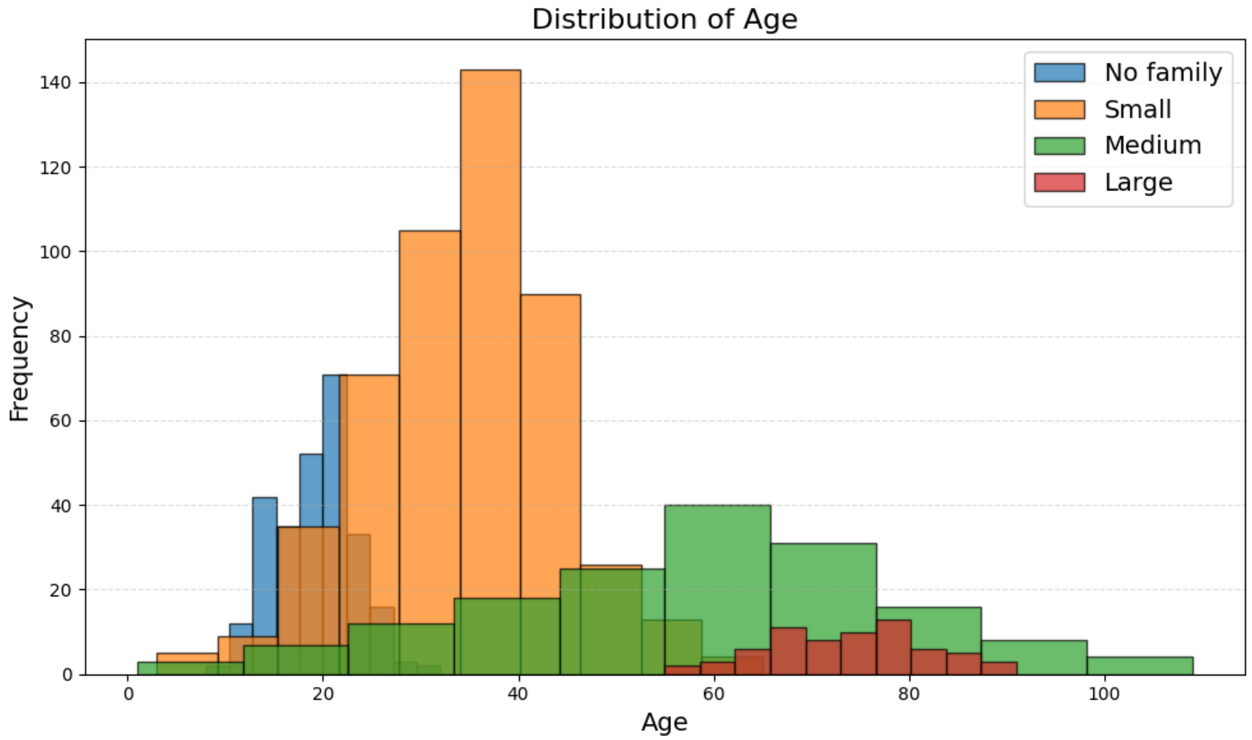
Bivariate scatter plot

References

Key for analysis is Visualisation
- Multivariate analysis -

4 - Multivariate analysis for continuous and/or categorical variables

Multivariate

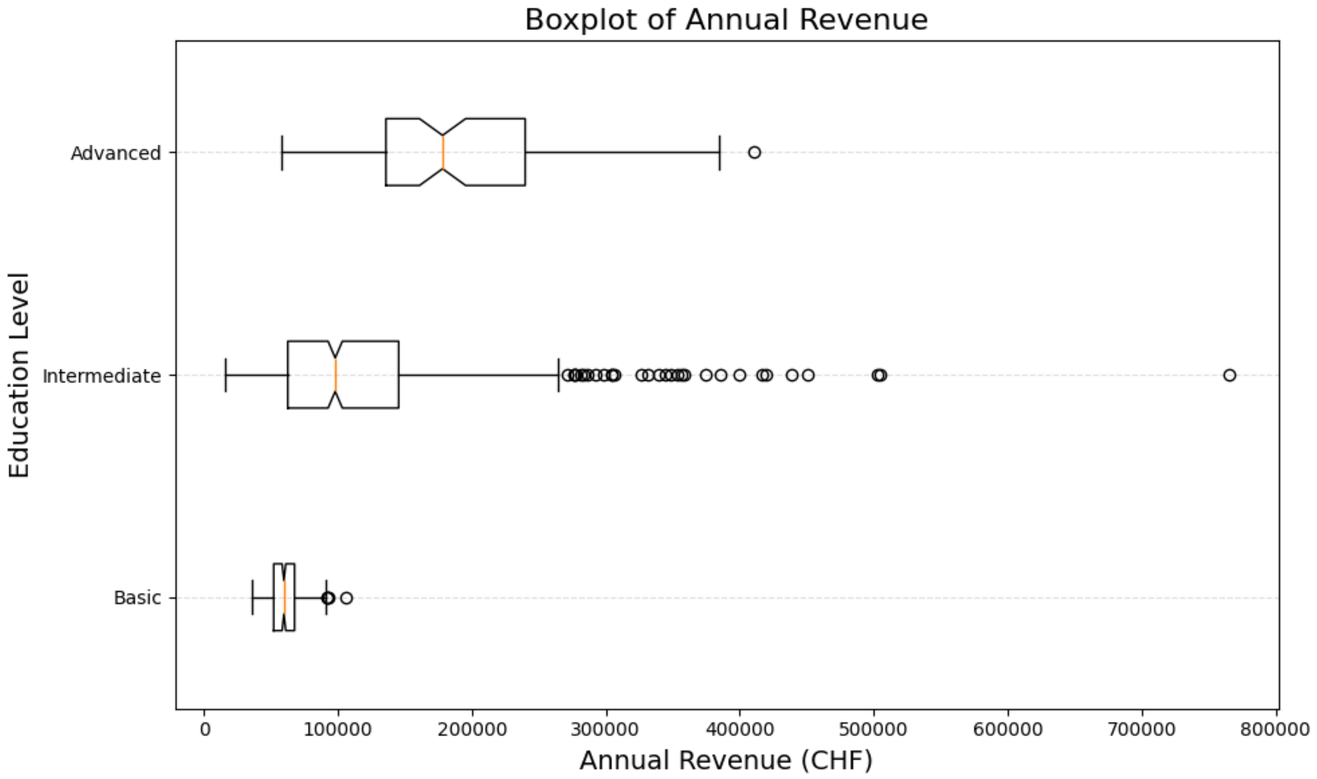


References

Key for analysis is Visualisation
- Multivariate analysis -

5 - Multivariate analysis for continuous and/or categorical variables

Bivariate boxplots

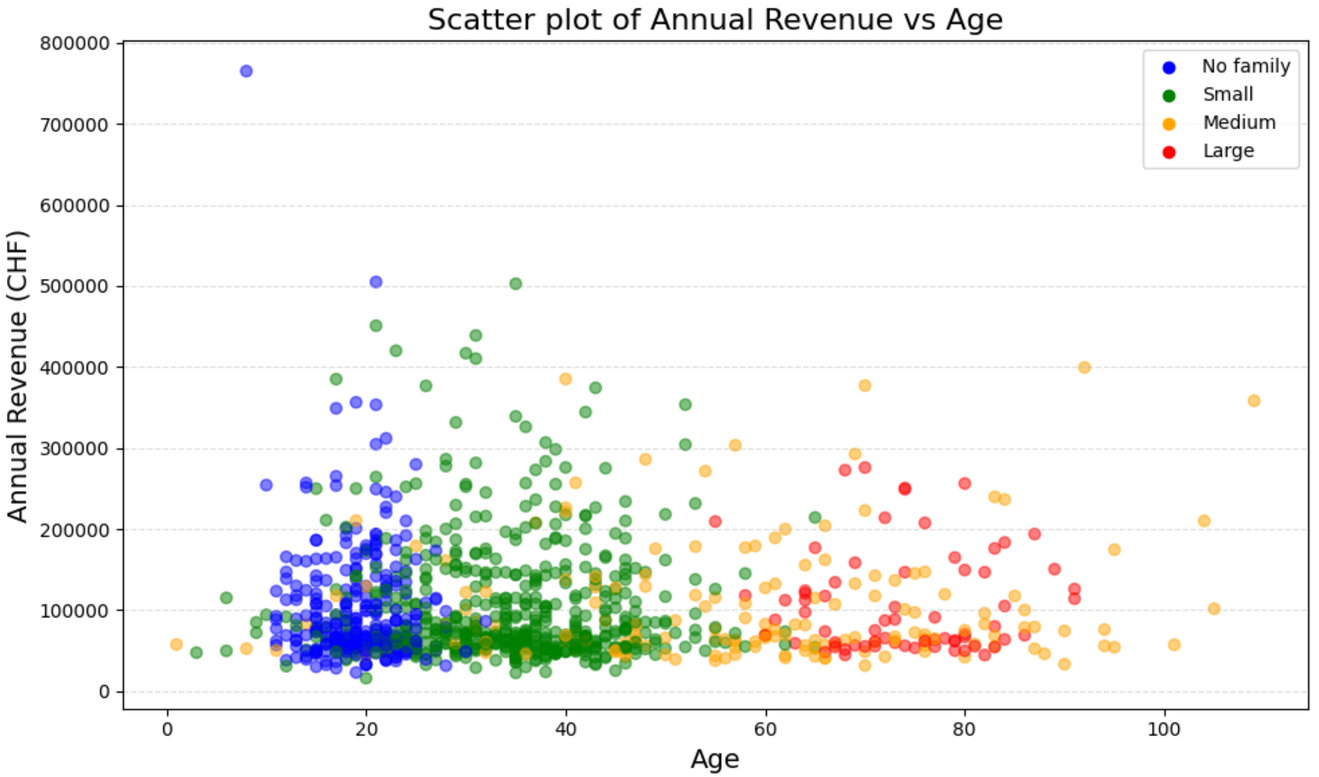


References

Key for analysis is Visualisation
- Multivariate analysis -

6 - Multivariate analysis for continuous and/or categorical variables

Colored scatter plots

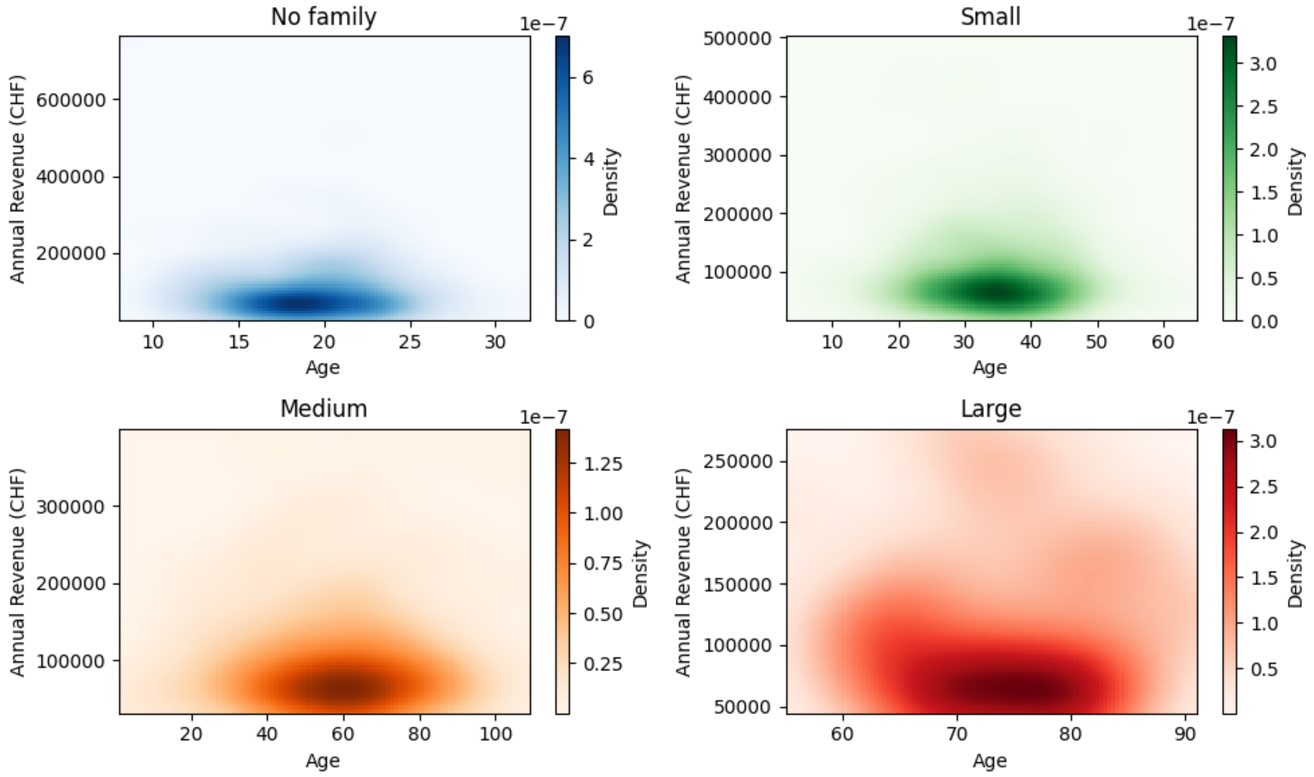


References

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.