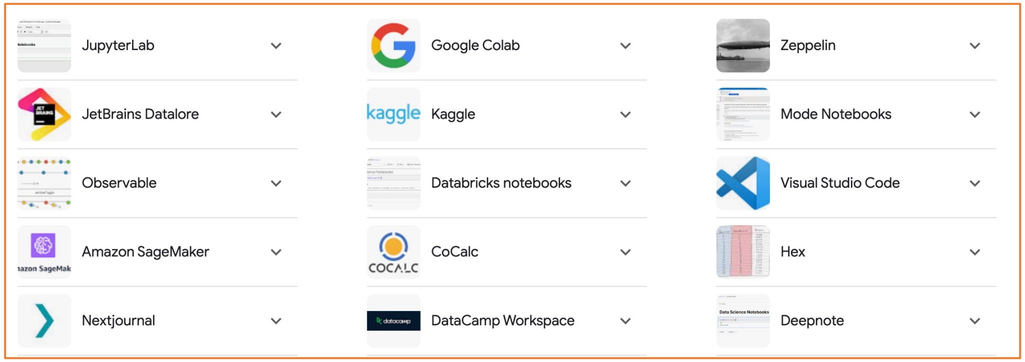
2_2. Hands on

One step back to get inertia

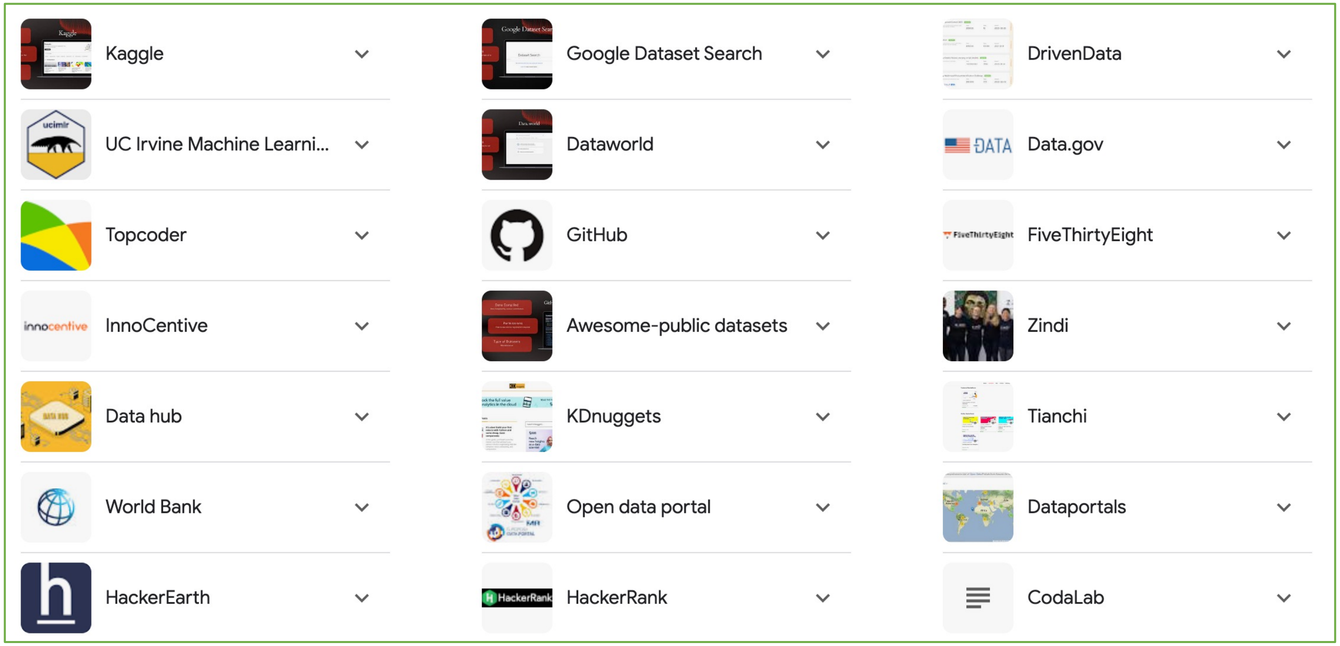
Hardware

Software



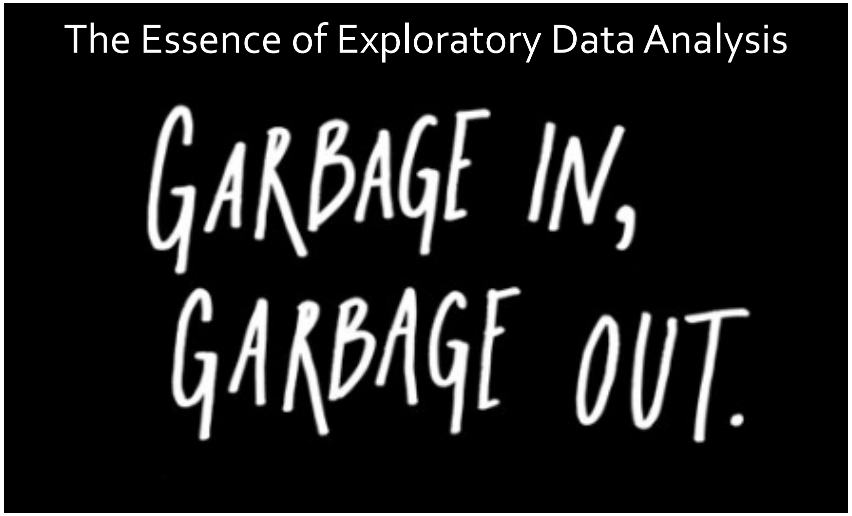
- JupyterLab
- JetBrains Datalore
- Observable
- Amazon SageMaker
- Nextjournal
- Google Colab
- Kaggle
- Databricks notebooks
- CoCalc
- DataCamp Workspace
- Zeppelin
- Mode Notebooks
- Visual Studio Code
- Hex
- Deepnote

Data



- Kaggle
- UC Irvine Machine Learni...
- Topcoder
- InnoCentive
- Data hub
- World Bank
- HackerEarth
- Google Dataset Search
- Dataworld
- GitHub
- Awesome-public datasets
- KDnuggets
- Open data portal
- HackerRank
- DrivenData
- Data.gov
- FiveThirtyEight
- Zindi
- Tianchi
- Dataportals
- CodaLab

Exploratory Data Analysis



The Essence of Exploratory Data Analysis

It is crucial to invest time and effort in **data cleaning, preprocessing, and validation** to minimize the impact of "garbage" data on the overall analysis or decision-making process.

Data Cleaning and Preprocessing Steps

Exploratory Data Analysis

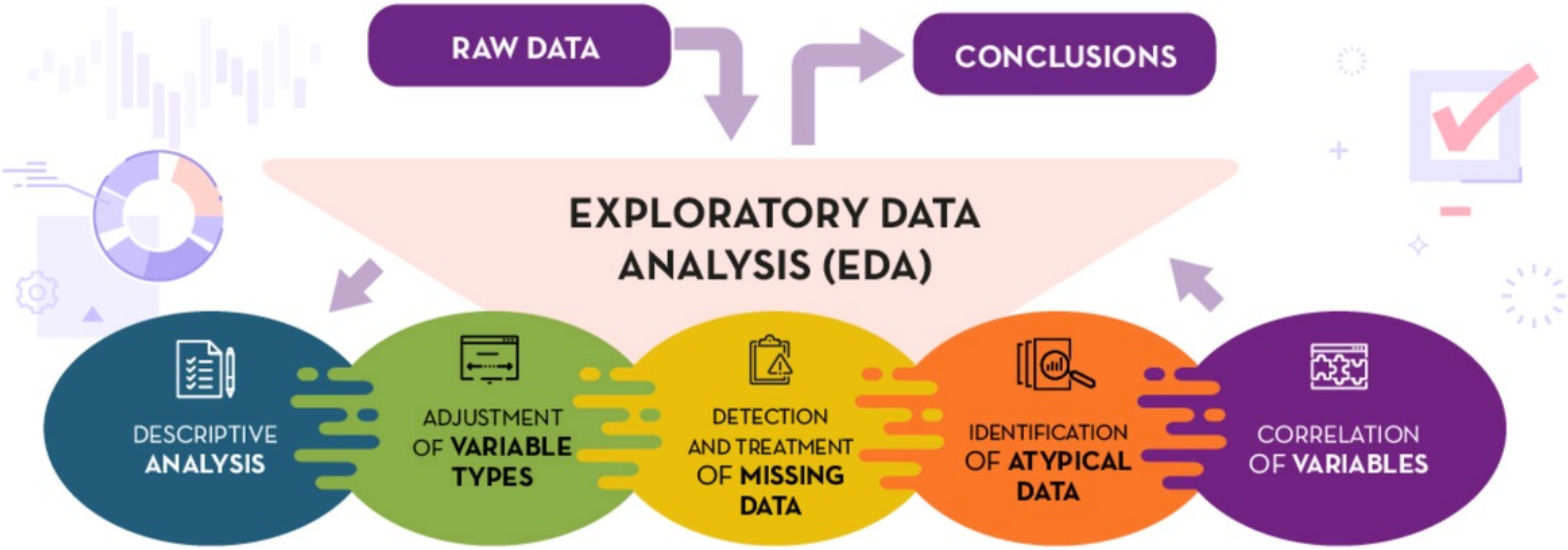
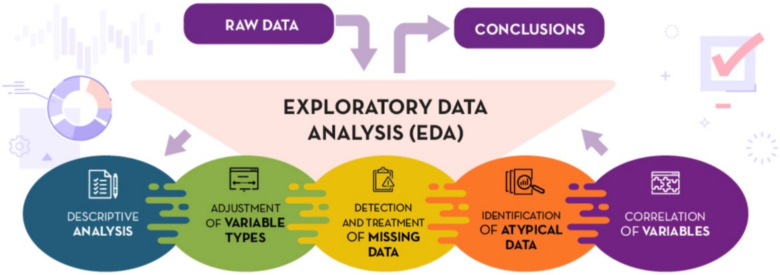

RAW DATA
CONCLUSIONS
EXPLORATORY DATA ANALYSIS (EDA)
DESCRIPTIVE ANALYSIS
ADJUSTMENT OF VARIABLE TYPES
DETECTION AND TREATMENT OF MISSING DATA
IDENTIFICATION OF ATYPICAL DATA
CORRELATION OF VARIABLES

Image Credit: https://datos.gob.es/en/documentacion/practical-introductory-guide-exploratory-data-analysis

Exploratory Data Analysis

0. Methodology


RAW DATA
CONCLUSIONS
EXPLORATORY DATA ANALYSIS (EDA)
DESCRIPTIVE ANALYSIS
ADJUSTMENT OF VARIABLE TYPES
DETECTION AND TREATMENT OF MISSING DATA
IDENTIFICATION OF ATYPICAL DATA
CORRELATION OF VARIABLES

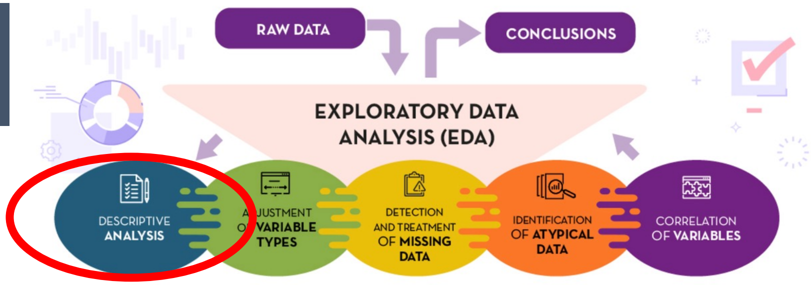
- A practical case
- R / Python / Excel / …
- Not efficient code, but illustrative
- Database:

  url = 'https://github.com/DIAGNijmegen/picai_labels/blob/main/clinical_information/marksheet.csv?raw=true’

- Import libraries and packages

Exploratory Data Analysis

1. Descriptive Analysis

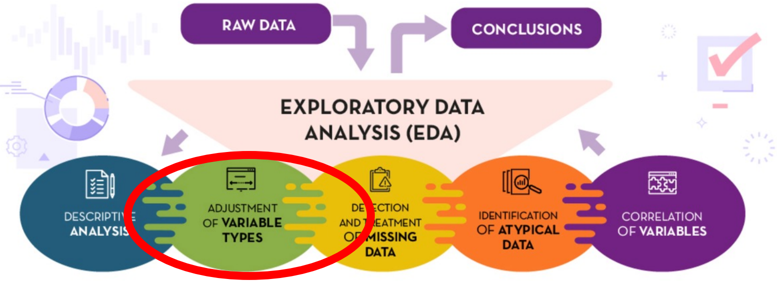

RAW DATA
CONCLUSIONS
EXPLORATORY DATA ANALYSIS (EDA)
DESCRIPTIVE ANALYSIS
ADJUSTMENT OF VARIABLE TYPES
DETECTION AND TREATMENT OF MISSING DATA
IDENTIFICATION OF ATYPICAL DATA
CORRELATION OF VARIABLES

- We will apply **descriptive statistics functions** to explore the structure of the data set and examine the data and variables it presents.

- It will be very useful to use **certain graphic representations** that will help you understand the shape of the data distributions.

Exploratory Data Analysis

2. Variable Types

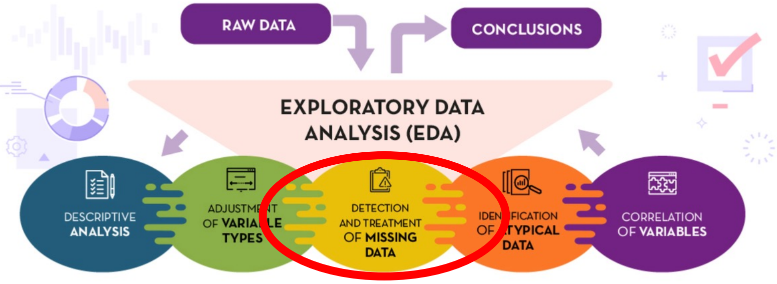

RAW DATA
CONCLUSIONS
EXPLORATORY DATA ANALYSIS (EDA)
DESCRIPTIVE ANALYSIS
ADJUSTMENT OF VARIABLE TYPES
DETECTION AND TREATMENT OF MISSING DATA
IDENTIFICATION OF ATYPICAL DATA
CORRELATION OF VARIABLES

- After loading the data into the work environment, you must verify that each variable has been stored with the **corresponding value type**.

- Usual types of tabular variables:
  - **numeric**: stores numbers that can be float or integers.
  - **character**: holds text strings.
  - **categorical**: contains a limited number of categories.
  - **logical** or boolean: binary variables (TRUE/FALSE or 0/1).
  - **date**: stores specific time intervals.

Exploratory Data Analysis

3. Missing Data

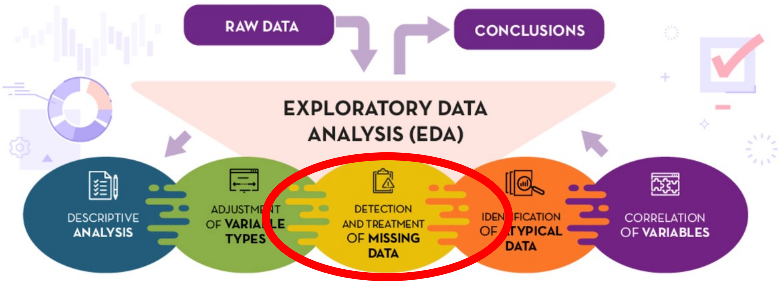

RAW DATA
CONCLUSIONS
EXPLORATORY DATA ANALYSIS (EDA)
DESCRIPTIVE ANALYSIS
ADJUSTMENT OF VARIABLE TYPES
DETECTION AND TREATMENT OF MISSING DATA
IDENTIFICATION OF ATYPICAL DATA
CORRELATION OF VARIABLES

- Dealing with data sets in which there are **missing data** can cause **problems when applying different statistical analysis or generating graphical representations**.

- In order to avoid future problems, it is necessary to learn how to detect and apply some type of **treatment**.

Exploratory Data Analysis

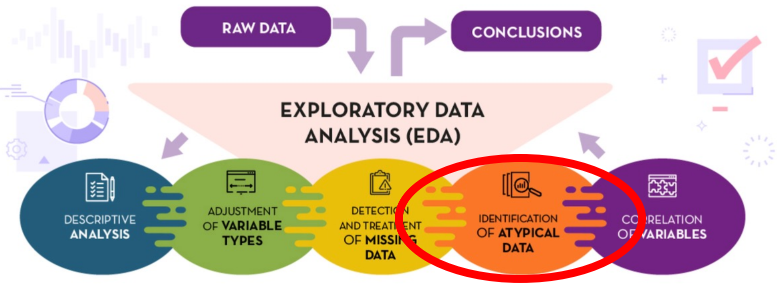
3. Missing Data


RAW DATA
CONCLUSIONS
EXPLORATORY DATA ANALYSIS (EDA)
DESCRIPTIVE ANALYSIS
ADJUSTMENT OF VARIABLE TYPES
DETECTION AND TREATMENT OF MISSING DATA
IDENTIFICATION OF ATYPICAL DATA
CORRELATION OF VARIABLES

- There are several ways to deal with missing values:
  - **Fill in** the values with the mean, median or the most frequent value of the variable.
  - **Fill in** missing values with the value directly before or after them in the row or column.
  - **Replace** all missing data with **0**, if they are numerical values.
  - **Delete rows** that have missing values, as long as the data set is large enough.
  - **Eliminate variables** that present a percentage greater than 50% of missing data.

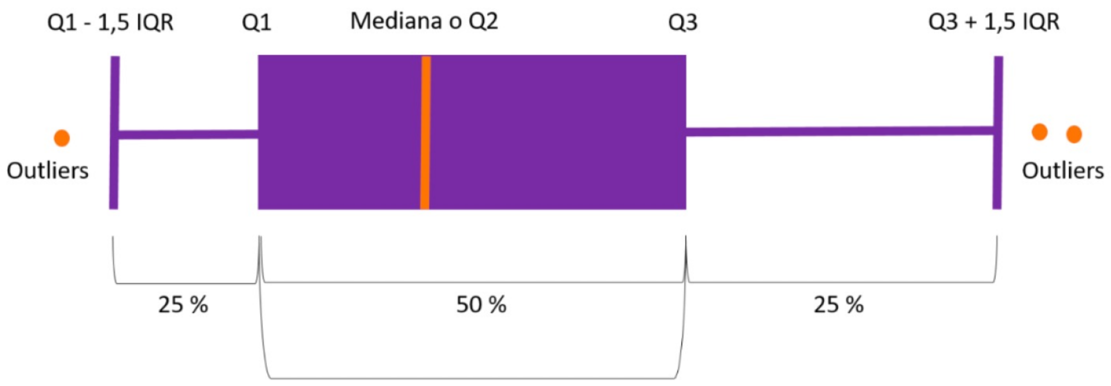
Exploratory Data Analysis

4. Atypical Data. Outliers


RAW DATA
CONCLUSIONS
EXPLORATORY DATA ANALYSIS (EDA)
DESCRIPTIVE ANALYSIS
ADJUSTMENT OF VARIABLE TYPES
DETECTION AND TREATMENT OF MISSING DATA
IDENTIFICATION OF ATYPICAL DATA
CORRELATION OF VARIABLES

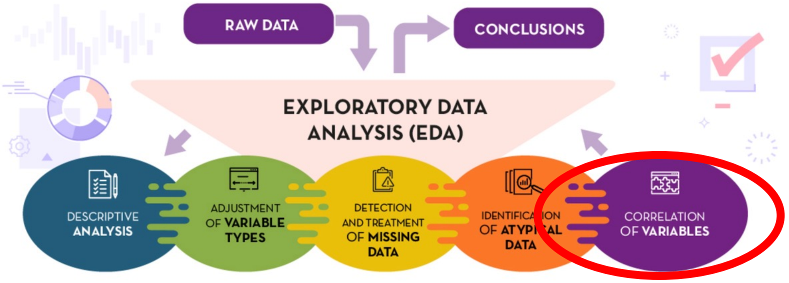
- An **outlier** value is an observation that is significantly different from the rest of the data presented by a variable, of such magnitude that it can be considered an **anomalous value**.

- **IQR-based** or **Percentile** methods.



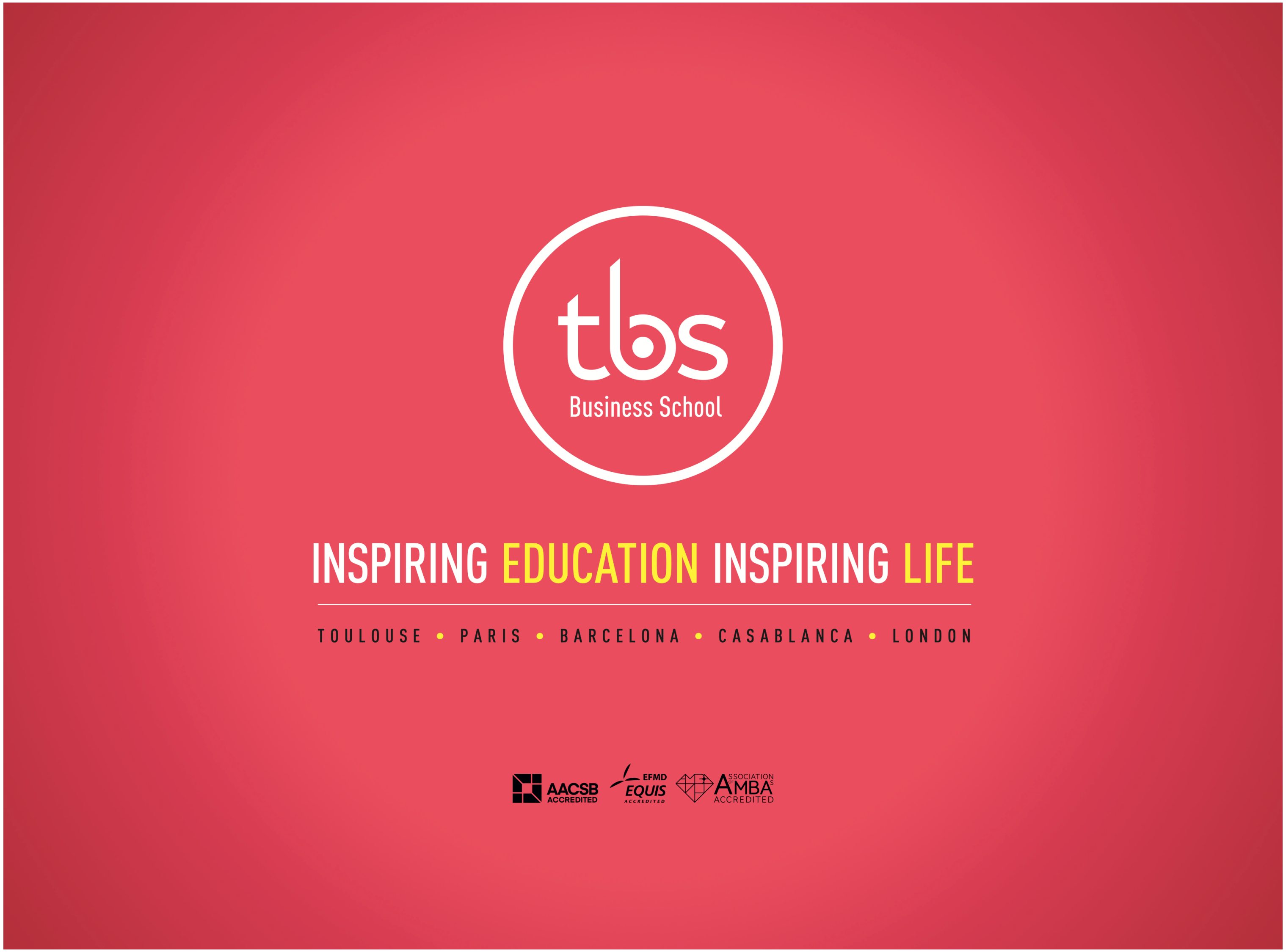
Exploratory Data Analysis

5. Correlation Analysis


RAW DATA
CONCLUSIONS
EXPLORATORY DATA ANALYSIS (EDA)
DESCRIPTIVE ANALYSIS
ADJUSTMENT OF VARIABLE TYPES
DETECTION AND TREATMENT OF MISSING DATA
IDENTIFICATION OF ATYPICAL DATA
CORRELATION OF VARIABLES

- **Correlation** determines the linear relationship between two or more variables, that is, the strength and direction of a possible relationship between variables.

tbs
EDUCATION



tbs
Business School

INSPIRING EDUCATION INSPIRING LIFE

TOULOUSE • PARIS • BARCELONA • CASABLANCA • LONDON

AACSB ACCREDITED
EFMD EQUIS ACCREDITED
ASSOCIATION OF MBAS ACCREDITED