EXPLORATORY DATA ANALYSIS (EDA)

A Practical Introductory Guide to Exploratory Data Analysis

1 WHAT IS?

A set of **statistical techniques** aimed at exploring, describing and summarising the information contained in the data, maximising its understanding.

This allows you to:

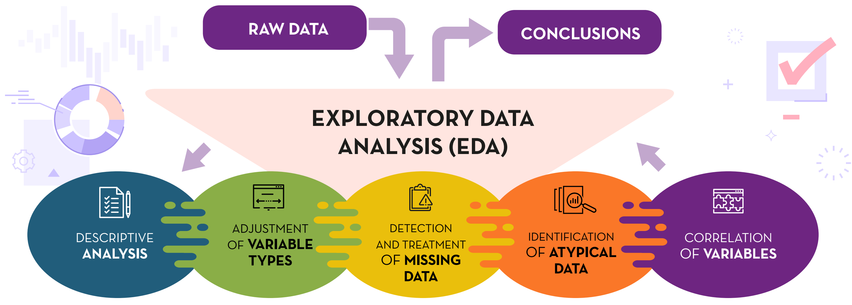
- Perform a descriptive analysis of the data
- Identify possible errors
- Reveal the presence of outliers
- Check the relationship between variables

WHY IS THIS IMPORTANT?

- Statistical **data analysis techniques and machine learning** presuppose the fulfilment of preconditions to ensure **objectivity and interoperability** of data.
- **EDA** is essential to ensure that the results of any statistical analysis are **consistent and truthful**.

2 WHAT ARE THE STEPS TO BE TAKEN?

Source: Based on the book R for Data Science de Wickman y Grolemund (2017)



1 DESCRIPTIVE ANALYSIS

- **What is it?** Synthesis of the information provided by the dataset, extracting its most representative characteristics.
- **Why is it necessary?** To understand the types of data, discover patterns and prepare the data for future analysis.
- **Treatment:** Apply descriptive statistical functions to explore the structure of the dataset, examine the data and the variables it presents.

2 ADJUSTMENT OF VARIABLE TYPES

- **What is it?** Verify that the variables have been stored with the appropriate corresponding value type.
- **Why is it necessary?** Bad coding of variables can negatively influence the grouping of data or the results of the analysis.
- **Treatment:** Apply the appropriate coding for each of the variables.

3 DETECTION AND TREATMENT OF MISSING DATA

- **What is it?** Identify some of the missing data in the variable.
- **Why is it necessary?** Missing data can create problems when applying machine learning techniques, building predictive models, performing statistical analysis or generating graphical representations.
- **Treatment:** There are several ways to treat missing values, such as replacing them with the mean or median, or filling in missing values with the previous or next value in the column.

4 DETECTION AND TREATMENT OF ATYPICAL DATA

- **What is it?** To identify data with values significantly different from those of the variable.
- **Why is it necessary to treat them?** They can modify the results and reduce the power of the statistical analysis or machine learning techniques applied.
- **Treatment:** To reduce their influence in subsequent analyses or, in very extreme cases, to eliminate them from the dataset.

5 CORRELATION OF VARIABLES

- **What is it?** Analysing the relationship between two or more variables.
- **Why is it necessary?** Among other reasons, to discard possible variables that provide redundant information in the dataset, causing noise in the analyses.
- **Treatment:** Calculate the correlation coefficients for the variables to detect coefficients close to 1 or -1.

3 DO YOU WANT TO KNOW MORE?

Discover the "A Practical Introductory Guide to Exploratory Data Analysis"

Learn the above techniques through the development of a **case study**.

Methodology for the practical example:

- **Dataset used**: air quality register in Castilla y León
- **Tools used**: R code in the RStudio programming environment.

red.es iniciativa aporta