2_1. Settings

Our objective:
Using own data, to help companies to impact society and business

The group project is an opportunity for students to demonstrate not just their technical skills but also their ability in presenting results and propose business recommendations.

Data Cleaning and Preprocessing Steps

Track / Module / Course contents
Data Cleaning and Preprocessing Techniques: In-depth exploration of methods to clean and preprocess datasets, addressing data quality issues and preparing data for meaningful analysis.
Statistical Analysis for Business Insights: Application of statistical methods to extract actionable insights from data, with a focus on making informed business decisions.
Data Visualization Tools and Techniques: Introduction to various data visualization tools and techniques, enabling students to create compelling visuals for effective communication of analytical findings.

One step back to get inertia

Hardware

Software

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

Data

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

Data Cleaning and Preprocessing Steps

Exploratory Data Analysis

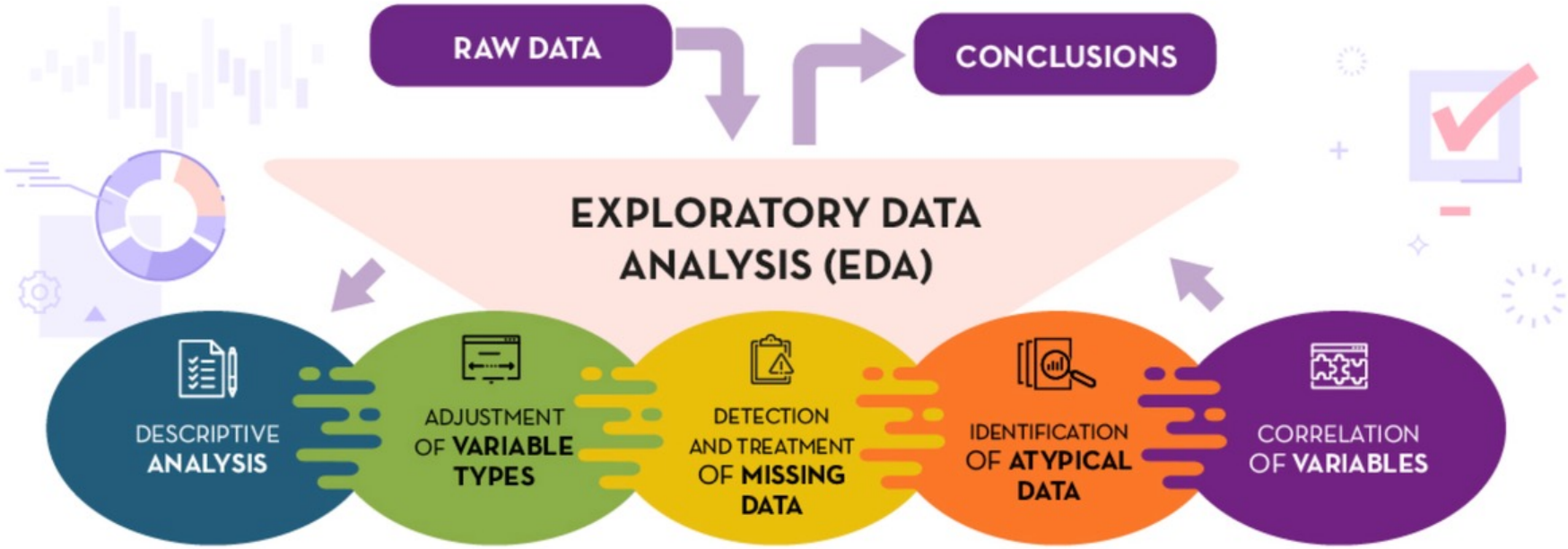


Image Credit: https://datos.gob.es/en/documentacion/practical-introductory-guide-exploratory-data-analysis

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.

Exploratory Data Analysis

1. Visualizing and Summarizing Data

Plots aiding in effectively visualizing and summarizing data:

- **Histograms**: Dividing a numerical variable into bins and displaying its distribution.
- **Box plots**: Offering a visual representation of the distribution, median, quartiles, and outliers.
- **Scatter plots**: Demonstrating the relationship between two numerical variables.
- **Heatmaps**: Visualizing the correlation matrix between multiple variables through color gradients.
- **Summary statistics**: Calculating measures such as mean, median, standard deviation, and quartiles.

Exploratory Data Analysis

2. Uncovering Patterns & Relationships

Visual plots and analysis assisting in unearthing various hidden patterns and relationships within the given data:

- **Correlation analysis**: Quantifying the strength and direction of the linear relationship between variables.
- **Scatter plot matrix**: Displaying multiple scatter plots in a grid to analyze pairwise relationships.
- **Line plots**: Visualizing trends and patterns in time series or sequential data.
- **Bar plots**: Comparing categorical variables and their frequencies or proportions.

Exploratory Data Analysis

3. Detecting Outliers & Anomalies

Addressing outliers ensures that subsequent analyses and models remain unaffected by these exceptional cases

- **Box plots**: Identifying data points lying beyond the whiskers (i.e., beyond 1.5 times the interquartile range).
- **Z-score analysis**: Calculating the number of standard deviations a data point deviates from the mean.
- **Leverage plots**: Assessing influential observations in regression models.

4. Guiding Data Preprocessing

Several measures contribute to guiding effective data preprocessing:

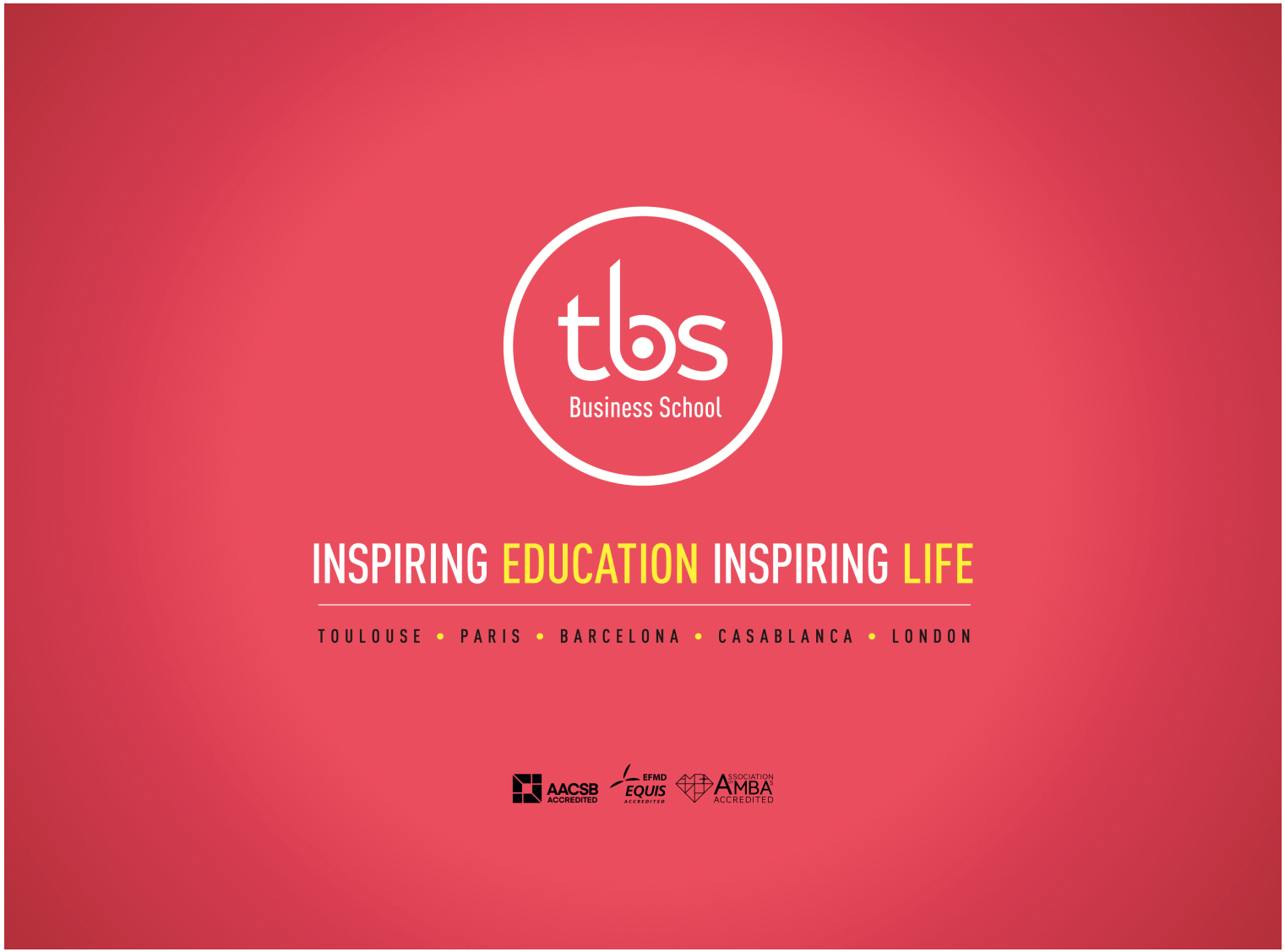
- **Missing data analysis**: Visualizing patterns of missingness using heatmaps or bar plots.
- **Imputation techniques**: Replacing missing values with estimates, such as mean, median, or regression-based imputation.
- **Data validation**: Checking for inconsistent formats, outliers, or unexpected values.

Hands on. In-class Groups!

Exploratory Data Analysis

- Hardware
- Software
- Data

?


tbs
Business School
INSPIRING EDUCATION INSPIRING LIFE
TOULOUSE • PARIS • BARCELONA • CASABLANCA • LONDON
AACSB ACCREDITED
EFMD EQUIS ACCREDITED
ASSOCIATION AMBA ACCREDITED