Preface

WS17. Exploratory Data Analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. EDA is primarily for seeing what the data can tell us beyond the formal modeling or hypothesis testing tasks.

The EDA approach can be broken down into the following steps:

Data Cleaning: This step includes handling missing data, removing outliers, and other data cleansing processes.

Univariate Analysis: Here, each field in the dataset is analyzed independently to better understand its distribution, outliers, and unique values. This could involve statistical plots for measuring central tendency like mean, median, mode, frequency distribution, quartiles, etc.

Bivariate Analysis: This step involves the analysis of two variables to determine the empirical relationship between them. It includes techniques such as scatter plots for continuous variables or crosstabs for categorical data.

Multivariate Analysis: This is an advanced step, involving analysis with more than two variables. It helps to understand the interactions between different fields in the dataset.

Data Visualization: This is the creation of plots such as histograms, box plots, scatter plots, etc., to identify patterns, relationships, or outliers within the dataset. This can be done using visualization tools or libraries.

Insight Generation: After visualizations and some statistical tests, analysts will generate insights that could lead to further questions, hypotheses, and model building.

The EDA process is an important precursor to more complex analyses because it allows for the researcher to confirm or invalidate some initial hypotheses and to formulate a more precise question or hypothesis that can lead to further statistical analysis and testing.

Our focus

- We ignore the Data Cleaning step, although we acknowledge it's practical relevance. We assume that we are working with a clean dataset.

- We emphasize Univariate and Bivariate Analysis of data and the corresponding Data Visualization.
- We cover some basic Multivariate Analysis.
- We emphasize Insight Generation.

We illustrate all of the above using the R programming language.

We further illustrate how to use R programming on a real-world dataset. Our dataset concerns the S&P500 stocks. This will demonstrate a practical aspect of using this book. We have many sample codes regarding this, using real-world data.. We will explore financial metrics such as the Return on Equity , Return on Assets, Return on Invested Capital of S&P500 shares.