Chapter 5: Exploratory Data Analysis

Introduction

The data from experiment/research are generally collected into a spreadsheet or database, most commonly with one row per experimental subject and one column for each subject identifier, outcome variable, and explanatory variable. Each column contains the numeric values for a particular quantitative variable or the levels for a categorical variable. (Some more complicated experiments require a more complex data layout.)

People are not very good at looking at a column of numbers or a whole spreadsheet and then determining important characteristics of the data. They find looking at numbers to be tedious, boring, and/or overwhelming. Exploratory data analysis techniques have been devised as an aid in this situation.

What is EDA?

When it comes to EDA, remember how good you are at this will make huge difference professionally.

Let us take an intuition or

Imagine your wolf pack decides to watch a movie you haven't heard of. There is absolutely no debate about that, it will lead to a state where you find yourself puzzled with lot of questions which needs to be answered in order to make a decision. Being a good chieftain the first question you would ask, what is the cast and crew of the movie? As a regular practice, you would also watch the trailer of the movie on YouTube. Furthermore, you'd find out ratings and reviews the movie has received from the audience.

Whatever investigating measures you would take before finally buying popcorn for your clan in theater, is nothing but what data scientists in their language call 'Exploratory Data Analysis'.

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

It is a good practice to understand the data first and try to gather as many insights from it. EDA is all about making sense of data in hand, before getting them dirty with it.

Why EDA is a significant pick in DS?

The primary objective of EDA assists the look at data before giving rise to any inferences. It enables you to observe noticeable mistakes and reasonable, understand structures within the data, distinguish anomalous events or outliers, and find fascinating associations among the variables.

Data scientists employ exploratory analysis to ensure the outcomes they produce are accurate and acceptable to any desired business findings and objectives. EDA also assists stakeholders by confirming they are inquiring about the moral questions. EDA furthermore helps to answer questions about categorical variables, standard deviations, and confidence intervals. Once EDA is finished, and ideas are brought out, its characteristics employ more sophisticated data analysis or modeling, encompassing machine learning.

Four fundamental kinds of EDA:

Univariate non-graphical:

This is the most straightforward aspect of data analysis. The data is analyzed, consisting of barely one variable. Since it's a sole variable, it does not negotiate with spurs or connections. The univariate analysis's primary objective is to interpret the data and discover structures that occur within it.

Univariate graphical:

Non-graphical techniques do not deliver an entire image of the data. Visual methods are thus employed.

Popular kinds of univariate graphics contain:

- Stem-and-leaf plots, which exhibit all data values and the pattern of the measurement.
- Histograms, a bar plot in which every bar exemplifies the frequency (count) or percentage (count/total count) of trials for a spectrum of values.
- Box plots, which graphically portray the minimum's five-number overview, are the first quartile, median, followed by the third quartile, and the maximum.

Multivariate non-graphical:

Multivariate data rises from additional than one variable. These EDA methods usually exhibit the connection between the two or extra variables of the data through statistics or cross-tabulation.

Multivariate graphical:

Multivariate data employs representations to depict connections between two or extra sets of data. The extensively using graphic is a bar chart or grouped bar plot with every group representing one

level of one of the variables and each bar within an association indicating the degrees of the different variables.

Other popular categories of multivariate graphics contain:

- A Scatter plot is there to conspire data junctures on a vertical and a horizontal axis to indicate how much another influences one variable.
- Multivariate chart, which is a visual manifestation of the connections between response and factors.
- A run chart is a line graph of data conspired over time.
- A bubble chart is a technique in data visualization that exhibits numerous circles (bubbles) in a two-dimensional conspiracy or plot.
- Heat map, which is a visual articulation of data where significances get identified by color.

EDA involves generating summary statistics for numerical data in the dataset and creating various graphical representations to understand the data better. We will now understand EDA with the help of an example dataset.