Caitlin Sizemore

Dr. Sarkar

DS 160-01

02/18/2023

Exploratory Data Analysis

Exploratory Data Analysis, also known as EDA, is the process used to identify patterns and trends in data, spot anomalies, and discover hypotheses. These conclusions are achieved through the use of visual representations and graphs and summary statistics. Because this process is used to discover patterns, it is useful when little information is known about the subject. Exploratory Data Analysis serves as the first few crucial steps to analyzing and understanding data. As with any form of analysis, the first step to Exploratory Data Analysis is to collect the data. This step is largely self-explanatory; it is hard to analyze data that does not exist. Once the data has been collected, it should be processed into spreadsheet software like Excel or csv files. Spreadsheet software allows users to see the layout of columns, rows, and the overall make-up of the data. In addition, the software and subsequent layout of data allow for later visualization techniques including graphs and charts. With the data captured in the spreadsheet software, it is time to clean the data.

Cleaning datasets involves removing information, filling in any missing values, and filtering the remaining data. Data can be very large, including many columns, rows, and information. For analysis purposes, not all information is important. The cleaning process prepares data for analysis, leaving behind the important, correct information. Any unimportant information creates busy, overwhelming data which is not ideal for analysis. Unimportant information is anything that is redundant, unnecessary, or unrelated to the topic. These rows or columns can simply be deleted, removing them from the dataset. While some information may be unimportant, other information is very important but incomplete. To use this data any incomplete information must be filled in. Empty sections within rows and columns will negatively affect analysis. There are a few methods to this process. First, you must consider the purpose of the section of data. Based upon this answer, you can use basic summary statistics like mean, median, mode, etc. to fill in the missing values. Finally, it is crucial to review the data and filter any remaining data. For example, it may be necessary to remove outliers that can affect analysis. Once these steps have been completed, the data has been cleansed and is ready for more in depth, visual analysis.

Exploratory Data Analysis is used to begin to understand data. Therefore, the next steps of the analysis serve the purpose of understanding the most basic information of the data. To begin, it is important to understand the distribution and skew of the dataset. This can be achieved with a histogram graph. Distribution and skew allow the user to understand the layout of the data along with its relationship to the summary statistics, average and median. Another basic visualization important for early analysis is line charts. Line charts show the distribution and trend of the data over time, creating a basic and easy-to-understand layout of information. Through the use of data visualization software like Python’s Matplotlib and Seaborn, these graphs can easily be displayed. Some other summary statistics relevant to the layout of data are the standard deviation, which describes the disbursement of data in relation to the mean, range, which shows the overall layout of the data, and maximum and minimum values, which can help to identify outliers. These summary statistics can display the flow of data from the smallest to largest values.

The next step in Exploratory Data Analysis is correlation. Once the general shape and layout has been determined, correlations must be discovered for further analysis. Correlation describes the relationship between different variables. It is used to discover causation, an important aspect of hypothesis testing. Correlation is best graphed through scatter plots and its line of best fit and heat maps. Similar to the prior step, these can be graphed with Python’s Matplotlib and Seaborn. Correlation can be positive or negative, meaning the variables will increase and decrease together (positive correlation), or opposite of each other (negative correlation). Also, the closer the correlation is to one, the stronger the relationship is. Generally, in statistics, any correlation of 0.7 or greater is considered a strong correlation.

After basic relationships and the overall display of the data has been established, any additional testing can be completed. This includes graphs like bar charts, pie charts, stack plots, box plots, and more. Data visualization is an important tool to understanding data and relationships, making this an incredibly important step for hypothesis testing. As always, these graphs can be created through the use of Matplotlib and Seaborn. Exploratory Data Analysis is used to familiarize yourself with datasets. Its overall goal is to display patterns and trends for further testing. Once the graphs have been completed and you are familiar with the data, Exploratory Data Analysis has been completed and it is time to create hypotheses based upon the observed data and relationships. Without the understanding of information achieved through this analysis, data would be very overwhelming and confusing. Exploratory Data Analysis simplifies data by highlighting important correlation and shapes through the use of data visualization provided by Python and Tableau. This analysis serves as the first process required to truly master a dataset of information.