Assignment No. 1

1.1 Title: Exploratory data analysis (EDA) is a visual approach to understand, analyze dataset to summarize its main characteristics.

1.2 Problem Definition: 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.

1.3 Prerequisite: Install Anaconda Python, Jupyter Notebook, Spyder on Ubuntu 18.04. Add bashrc path.

1.4 Software Requirement: Jupyter Notebook, Spyder on Ubuntu.

1.5 Hardware Requirement: 2GB RAM, 500 GB HDD

1.6 Objectives: Understand the implementation of the EDA model

1.7 Outcomes: After completion of this assignment students can develop and analyze the EDA model and will understand the working.

1.8 Theory Concepts:

- What does EDA mean?

  EDA stands for Exploratory Data Analysis. EDA is a critical first step in analyzing the data from an experiment.
  Here are the main reasons we use EDA:

  • detection of mistakes

  • checking of assumptions

  • preliminary selection of appropriate models

  • determining relationships among the explanatory variables, and

  • assessing the direction and rough size of relationships between explanatory and outcome variables.

- Typical data format

  - The data from an experiment are generally collected into a rectangular array (e.g., 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.)
  - Exploratory data analysis techniques have been devised as an aid in this situation. Most of these techniques work in part by hiding certain aspects of the data while making other aspects clearer.

- Types of EDA

  - Univariate non-graphical
  - Multivariate non-graphical
  - Univariate graphical
  - Multivariate graphical.

1.8.1 Univariate non-graphical EDA

The data that come from making a particular measurement on all of the subjects in a sample represent our observations for a single characteristic such as age, gender, speed at a task, or response to a stimulus. We should think of these measurements as representing a “sample distribution” of the variable, which in turn more or less represents the “population distribution” of the variable. The usual goal of univariate non-graphical EDA is to better appreciate the “sample distribution” and also to make some tentative conclusions about what population distribution(s) is/are compatible with the sample distribution. Outlier detection is also a part of this analysis

1.8.2. Univariate graphical EDA

If we are focusing on data from observation of a single variable on n subjects, i.e., a sample of size n, then in addition to looking at the various sample statistics discussed in the previous section, we also need to look graphically at the distribution of the sample. Non-graphical and graphical methods complement each other. While the non-graphical methods are quantitative and objective, they do not give a full picture of the data; therefore, graphical methods, which are more qualitative and involve a degree of subjective analysis, are also required.

1.8.3. Multivariate non-graphical EDA

Multivariate non-graphical EDA techniques generally show the relationship between two or more variables in the form of either cross-tabulation or statistics.

1.8.4 Multivariate graphical EDA

There are few useful techniques for graphical EDA of two categorical random variables. The only one used commonly is a grouped barplot with each group representing one level of one of the variables and each bar within a group representing the levels of the other variable.

- Degree of Freedom:

Degrees of freedom are numbers that characterize specific distributions in a family of distributions. Often we find that a certain family of distributions is needed in some general situation, and then we need to calculate the degrees of freedom to know which specific distribution within the family is appropriate. The most common situation is when we have a particular statistic and want to know its sampling distribution. If the sampling distribution falls in the "t" family as when performing a t-test, or in the "F" family when performing an ANOVA, or in several other families, we need to find the number of degrees of freedom to figure out which particular member of the family actually represents the desired sampling distribution. One way to think about degrees of freedom for a statistic is that they represent the number of independent pieces of information that go into the calculation of the statistic.

1.8 Conclusion : Thus, after successfully completing this assignment, we were able to understand & implement exploratory data analysis algorithms.