Experiment No.: 5

Title: Write a program to perform Exploratory Data Analysis (EDA) on data set.

Objectives:

1. To learn tools used for EDA.
2. To learn how to apply EDA on the data set.

Theory:

Exploratory Data Analysis:

Exploratory Data Analysis, or in short "EDA", is an approach to analyze data in order to:

- summarize main characteristics of the data

- gain better understanding of the dataset,

- uncover relationships between different variables, and

- extract important variables for the problem we are trying to solve.

The main question we are trying to answer in EDA process is: "What are the characteristics that have the most impact on the target variable?" We can go through a couple of different useful exploratory data analysis techniques in order to answer this question. We can use following to perform EDA:

- Descriptive Statistics
- GroupBy in Python
- ANOVA
- Correlation
- Correlation-Statistics

Descriptive Statistics:

Descriptive Statistics, which describe basic features of a dataset and obtains a short summary about the sample and measures of the data.When you begin to analyze data, it's important to first explore your data before you spendtime building complicated models.One easy way to do so is to calculate somedescriptive statistics for your data.Descriptive statistical analysis helps todescribe basic features of a dataset and obtains a short summary about the sample and measuresof the data. One way in which we can do this is by usingthe describe() function in pandas.

Boxplots are a great way to visualize numericdata, since you can visualize the various distributions of the data.The main features that the boxplot shows are the median of the data, which represents wherethe middle datapoint is. The Upper Quartile shows where the 75th percentileis, the Lower Quartile shows where the 25th percentile is. The data between the Upperand Lower Quartile represents the Interquartile Range.Next, you have the Lower and Upper Extremes. These are calculated as 1.5 times the interquartilerange above the 75th percentile, and as 1.5 times the IQR below the 25th percentile.Finally, boxplots also display outliers as individual dots that occur outside the upperand lower extremes.

Group Byin Python:

Basic of Grouping Data using group by, and how this can help to transform our dataset. If we want to analyze relationship between predictors and target group by can be used.

Analysis of Variance(ANOVA):

ANOVA, the analysis of variance, is a statistical method in which the variation in a set ofobservations is divided into distinct components.Assume that we want to analyze a categorical variable and see the correlation among differentcategories.For example, consider the car dataset, the question we may ask is, how differentcategoriesof the Make feature (as a categorical variable) has impact on the price?To analyze categorical variables, we can use a method suchas the ANOVA method.ANOVA is a statistical test that stands for "Analysis of Variance".ANOVA can be used to find the correlation between different groups of a categoricalvariable.

The ANOVA test returns two values: the F-test score and the p-value.The F-test calculates the ratio of variation between the groups's mean over the variationwithin each of the sample groups.The p-value shows whether the obtained result is statistically significant.Without going too deep into the details, the F-test calculates the ratio of variation betweengroup means over the variation within each of the sample group means.This diagram illustrates a case where the F-test score would be small.The ANOVA test can be performed in Python using the f_oneway method as the built-infunction of the Scipy package.

Correlation:

Correlation is a statistical metric formeasuring to what extent different variables are interdependent.In other words, when we look at two variables over time,if one variablechanges how does this affect change in the other variable?For example, smoking is known to be correlated to lung cancersince you have a higher chance of getting lung cancer if

you smoke.In another example, there is a correlation between umbrella andrain variables where more precipitation means more people use umbrellas.In data science we usually deal more with correlation. We can look at the correlation between two variables.This time we'll visualize these two variables usinga scatter plot and an added linear line called a regression line,which indicates the relationship between the two. We can find positive or negative linear relationshipbetween predictors and target variables.

Correlation-Statistics:

Correlation statistical methods like Pearson Correlation and Correlation Heatmaps can be used identify relationship among the variables. One way to measure the strength of the correlation betweencontinuous numerical variable is by using a method called Pearson correlation.Pearson correlation method will give youtwo values: the correlation coefficient and the P-value.So how do we interpret these values?For the correlation coefficient,a value close to 1 implies a large positive correlation,while a value close to negative 1 implies a large negative correlation,and a value close to zero implies no correlation between the variables.Next, the P-value will tell us howcertain we are about the correlation that we calculated.

Key concepts: Exploratory Data Analysis (EDA)

Algorithm:

- Read data set into Pandas Datafrmes.
- Apply EDA techniques for understanding of data.