Name: Aditya Wavjale
Roll no: 31282

DSBDAL Assignment No. 3

- TITLE: Descriptive Statistics - Measure of central tendancy and variability

- PROBLEM STATEMENT:

Perform the following operations on any open-source dataset (eg:- data.csv)

1. Provide summary statistics (mean, median, minimum, maximum, standard deviation) for a dataset (age, income, etc.) with numeric variables grouped by one of the qualitative (categorical) variable. For example, if your categorical variable is age groups and quantitative variable is income, then provide summary statistics of income grouped by the age groups. Create a list that contains a numeric value for each reponse to the categorical variable.
2. Write a python program to display some basic statistical details like percentile, mean, standard deviation, etc. of the species of 'Iris-setosa', 'Iris-versicolor' and 'Iris-versicolor' of iris.csv dataset.

Provide the codes with outputs and explain everything that you do in this step.

- LEARNING OBJECTIVES:

1. To calculate the statistical summary of the data using OOP concepts.
2. To learn the concepts of percentile, standard deviation, etc.

- S/W AND H/W REQUIREMENTS:

1. H/W - 64-bit Windows OS
2. S/W - Jupyter notebook.

• THEORY :

Statistical data analysis is a procedure of performing various statistical operations. It is a kind of quantitative research, which seeks to quantify the data, and typically applies some of form of statistical analysis.

- Mean :- Mean of a data is the average of the grouped data. It is calculated by dividing the sum of all data values by the total number of values in the data.

- Median :- Median of a sorted data is the middlemost value in the dataset.

- Mode :- Mode of a dataset is the value which occurs for most number of times in the dataset.

- Minimum :- The least value among all values in dataset.
- Maximum :- The highest value among all values in dataset.

- Standard deviation :- It is a statistic that measures the dispersion of a dataset relative to its mean value and is calculated as the square root the variance.
- Percentile :- A percentile is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations fall.
- Quartile :- A quartile devides the data into three points - a lower quatile, median, and upper quartile - to form four groups of the dataset.

i) lower quartile - It is the middle number that falls between the smallest value and median.

ii) middle quartile - It is the median of dataset.

iii) Upper quartile - It is the middle number that falls between the largest value and median.

- Eg:- df = pd.read_csv("nba.csv")
  where, df is the dataframe of the csv file.

Mean :- df['Age'].mean()
Median:- df['Age'].median()
Mode :- df['Age'].mode()
Minimum:- df['Age'].min()
Maximum:- df['Age'].max()
Standard deviation:- df['Age'].std()
Other statistical data:- df['Age'].describe()

- CONCLUSION:

Hence we learned the various statistical terms and calculated them using OOP concepts.