**Exploratory Data Analysis**

Instructions:

Please share your answers filled inline in the word document. Submit Python code and R code files wherever applicable.

Please ensure you update all the details:

**Name: Biswajeet Padhi**

**Batch Id: 280921**

**Topic: Exploratory Data Analysis**

**Problem Statements:**

Q1) Calculate Skewness, Kurtosis using R/Python code & draw inferences on the following data.

**Hint:** [Insights drawn from the data such as data is normally distributed/not, outliers, measures like mean, median, mode, variance, std. deviation]

a. Cars speed and distance

****

b. Top Speed (SP) and Weight (WT)

****

Q2) Draw inferences about the following boxplot & histogram.

**Hint:** [Insights drawn from the plots about the data such as whether data is normally distributed/not, outliers, measures like mean, median, mode, variance, std. deviation]

Ans).

Not normally distributed as observed from histogram

Have outliers as observed from boxplot

Here the data is right skewed as observed from histogram

Here mean>median>mode





Q3) Below are the scores obtained by a student in tests

**34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56**

1. Find mean, median, variance, standard deviation.
2. What can we say about the student marks? [**Hint**: Looking at the various measures calculated above whether the data is normal/skewed or if outliers are present].

Ans) all answers in python code

Q5) What is the nature of skewness when mean, median of data is equal?

Ans) skewness = zero

Normally distributed data

Q6) What is the nature of skewness when mean > median?

Ans) Positively skewed

Q7) What is the nature of skewness when median > mean?

Ans) Negatively skewed

Q8) What does positive kurtosis value indicates for a data?

Ans) Positive kurtosis indicates the data is not normally distributed and has thin peak and thick tails.

Q9) What does negative kurtosis value indicates for a data?

Ans) Negative kurtosis indicates the data is not normally distributed and has wide peak and thin tails.

Q10) Answer the below questions using the below boxplot visualization.



What can we say about the distribution of the data?

Ans) The data is not normally distributed and does not have outliers

What is nature of skewness of the data?

Ans) It is Negatively skewed

What will be the IQR of the data (approximately)?

Ans:)IQR = Q3 – Q1

=18-10

=8

Q11) Comment on the below Boxplot visualizations?



Draw an Inference from the distribution of data for Boxplot 1 with respect Boxplot 2.

**Hint**: [On comparing both the plots, and check if the data is normally distributed/not, outliers present, skewness etc.]

Ans)

Boxplot 1 – normally distributed , no outliers ,skewness is zero

Boxplot 2 – normally distributed , no outliers ,skewness is zero

Q12)



Answer the following three questions based on the boxplot above.

1. What is inter-quartile range of this dataset? [**Hint**: IQR = Q3 – Q1]

In one line, explain what this value implies. (**Hint:** Based on IQR definition)

Ans) IQR= Q3-Q1 = 12-5=7

1. What can we say about the skewness of this dataset?

Ans) The dataset is positively skewed .

1. If it were found that the data point with the value 25 is 2.5, how would the new boxplot be affected?

(**Hint:** On changing the data point from 25 to 2.5 in the data, how is it different from the current one.)

Ans) If the data point is changed from 25 to 2.5 , there will be no outliers in the dataset.

Q13) 

Answer the following three questions based on the histogram above.

1. Where would the mode of this dataset lie? **Hint:** [In terms of values On Y-axis]

Ans) The mode will lie between y interval 5 to 10.

1. Comment on the skewness of the dataset

Ans) It is positively skewed

1. Suppose that the above histogram and the boxplot in question 2 are plotted for the same dataset. Explain how these graphs complement each other in providing information about any dataset. **Hint:** [Visualizing both the plots, draw the insights]

Ans) From the boxplot and histogram plotted , The data set is not normally distributed , it is positively skewed , the dataset have outliers , here mode >median>mean,

**Hints:**

For each assignment, the solution should be submitted in the below format

1. Research and Perform all possible steps for obtaining solution

2.

3. For Statistics calculations, explanation of the solutions should be documented in black and white along with the codes.

Must follow these guidelines:

3.1. Be thorough with the concepts of Probability, Central Limit Theorem and Perform the

calculation stepwise

3.2. For True/False Questions, or short answer type questions explanation is must

3.3. R & Python code for Univariate Analysis (histogram, box plot, bar plots etc.) the data

distribution to be attached

4. All the codes (executable programs) should execute without errors

5. Code modularization should be followed

6. Each line of code should have comments explaining the logic and why you are using that