**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: Gunjan Kumar Gupta**

**Batch ID: DSWDMCSR 300522B**

**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]





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].

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

ANS  **zero skewness**.

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

ANS **Negative Skewnees**

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

ANS **Positive Skewnees**

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

ANS **distribution is peaked and possess thick tails**.

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

ANS Negative excess values of kurtosis (<3) indicate that **a distribution is flat and has thin tails**

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



What can we say about the distribution of the data?

Ans When the median is closer to the top of the box, and if the whisker is shorter on the upper end of the box, then the distribution is negatively skewed (skewed left).

What is nature of skewness of the data?

ANS **Negative Skewnees**

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

ANS The interquartile range (IQR) is the box plot showing the middle 50% of scores and can be calculated by subtracting the lower quartile from the upper quartile (e.g. Q3−Q1). 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 Data is normally distributed, and not outliers , Zero Skewnees

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)

1. What can we say about the skewness of this dataset?
2. 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 1) 12-5 =7

1. +ve Skewnees
2. No Skewed

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]
2. Comment on the skewness of the dataset
3. 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 1) 4-8

1. Right Skewd

30 For Histogram Data is right side stewed, It is not normally distributed

There is outliers in data set.

**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