**Exploratory Data Analysis**

**Problem Statements:**

Q1) Calculate Skewness, Kurtosis using R/Python code & draw inferences on the following data. Refer to the Datasets attachment for data file.

**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]





Answer : In histogram, Data concentration is more in the range 50 to 100 and less in the range 350 to 400. Long tail portion towards the right side, ie, positively skewed.

In boxplot, outliers exist on the upper side of the plot. Median is closer to first quartile. Median is less than mean and positively skewed. Data points are less in between lower limit and first quartile.

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.

Mean = 41, Median=40.5, Variance=25.52941, Standard Deviation=5.052664

1. 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].

Mean is greater than median that means positively skewed.

Skewness = 1.686841 Skewness is greater than 0 that means positively skewed.

Kurtosis = 3.953279 Since the value is greater than 3 the left kurtosis distributions will occur which is very sharp distribution and the values have the tendency to concentrate near or around the mean. There are two outliers

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

Symmetrical distribution

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

Positive skewness and not symmetrical data

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

Negative skewness and not symmetrical data

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

Thin peak and longer tail

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

Wider peak and thinner tail

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



What can we say about the distribution of the data? Not symmetrical

What is nature of skewness of the data? Mode is greater than mean so negatively skewed.

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

No outliers found in both the distribution. Both are symmetrically distributed. Zero skewness observed. Whisker level of 2nd plot is higher than 1st one.

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)

12-5 = 7 ie, Median

1. What can we say about the skewness of this dataset? Positively skewed as median is closer to first quartile.
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.) No outliers.

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] Data distribution is more in between 5 to 7
2. Comment on the skewness of the dataset . Positively skewed .
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] . Histogram deals with frequency of the distribution but boxplot gives insight about q1,q3,median, and outliers.