**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: Vaibhav Chaudhari**

**Batch Id: DSWDEOS 280422**

**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]



The histogram is right skewed showing extreme values are present on the right side. The mode is less than the median, which is less than the mean. Log transformation can be used to convert the data to normal.



The boxplot shows that there are outliers present. Winsorization can be used to treat these outliers.

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].

**Soln:-**

|  |  |
| --- | --- |
| Mean | 41.00 |
| Median | 40.50 |
| Mode | 41.00 |
| StdDev | 5.05 |
| Var | 25.53 |

**The mean = median = mode hence the data is normally distributed with zero skewness. There seems to be some outliers in the dataset as variance is very high.**

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

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

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

Q8) What does positive kurtosis value indicates for a data? **The data is less distributed near the mean and more distributed towards the end (thick tails)**

Q9) What does negative kurtosis value indicates for a data? **The data is more distributed near the mean and less distributed towards the end (thin tails)**

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



What can we say about the distribution of the data? **The data is normally distributed.**

What is nature of skewness of the data? **The data is negatively skewed.**

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

**Soln: The data is normally distributed with zero skewness. There are no outliers in the data.**

Q12)



Answer the following three questions based on the boxplot above.

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

**IQR = 7**

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

**IQR is the measure of how spread-out the values**

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

**Right-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.)

**There will be no outliers in the dataset. The boxplot will now represent normal distribution.**

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]

**6**

1. Comment on the skewness of the dataset

**The data is right 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]

**Both are right-skewed and have outliers. Box plot is used to check median and histogram is used for mode.**

**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