**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: \_\_\_Shubham Kunar Sharma\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Batch Id: \_\_DS WDMCO050121 Blr\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**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

****

**Skewness kurtosis**

**Car\_Speed** -0.1139548 2.422853

**Distance**  0.7824835 3.248019

* **- Here we have data set with two columns (1st speed, 2nd distance). Speed column data are normally distributed. And Distance column’s data are right skewed .We have one outlier in distance column index (49), value 120.**

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

****

**Skewness kurtosis**

**Top\_speed** 1.581454 5.723521

**Weight**  -0.6033099 3.819466

* **Here we have data set with two columns (1st speed, 2nd weight). Speed column data are right skewed. And weight column’s data are in normal distribution.**

**Outliers**

* **Speed**

**high value outlier =** 158.3007 , 164.5985, 169.5985 ,150.5766 , , 151.5985, 167.944

**Weight**

**high value outlier =** 52.99775

**Low value outlier =** 15.84776, 16.35948, 15.75353, 16.19412, 16.04317 , 15.71286, 15.82306, 16.13295, 15.76963

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- According to this boxplot, we can say that Distribution of the data is right skewed with some high value outlier. Mean is much greater median. Same for histogram. Its look like both histogram and boxplot is for same data .

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.

Ans- Mean <-41 , Median <-40.5 , Variance <- 24.11, Standard deviation <- 4.91

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].

Ans – Student marks are normally distributed without any outlier.

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

Ans – If mean ,median of data are equal then it is normal distributed .

No skewness .

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

Ans- Positively/Right skewed

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

Ans- Negatively/Left skewed

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

Ans –**Positive values** of **kurtosis indicate** that a distribution is peaked and possess thick tails.

Data distribution is leptokurtic.

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

And – Data distribution Is platykurtic ().**Negative values** of **kurtosis indicate** that a distribution is flat and has thin tails. Platykurtic distributions have **negative kurtosis values**. A platykurtic distribution is flatter (less peaked) when compared with the normal distribution, with fewer **values** in its shorter (i.e. lighter and thinner) tails.

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



What can we say about the distribution of the data?

Ans- distribution of the data is left skewed, median(approximate 15.5) is greater , than mean .

What is nature of skewness of the data?

Ans - Negatively/Left skewed

What will be the IQR of the data (approximately)?   
Ans – IQR = 8 unit

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.]

* Boxplot 1 is Positively skewed with no outliers.
* Boxplot 2 is normally distributed
* Boxplot 1 is overlap with boxplot 2 .box 1 is completely under box 2 ,then it is likely to be in same group . Median of both box is same. Range of box 1 is smaller than box 2, box 2 imply more variable data than box1.

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 = 7 Unit .

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

Ans- Dataset is 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?

Ans – Then new boxplot does not have any outliers.

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

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 - mode of this dataset lie between 4 to 8 unit .

1. Comment on the skewness of the dataset

Ans – Dataset 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]

Ans - According to both histogram and boxplot, we can say that Both Distribution of the data is right skewed with high value outlier. But we cannot differentiate mode in box plot but we can do that in histogram.

**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