**Exploratory Data Analysis**

**Instructions:**

Please share your answers filled in-line in the word document. Submit code separately wherever applicable.

Please ensure you update all the details:

**Name: \_\_\_\_\_Jaidev Chhabria\_\_\_\_\_\_\_\_ Batch ID:** \_\_\_\_DSWEMON 290521\_\_\_\_\_\_\_

**Topic: Exploratory Data Analysis**

**Grading Guidelines:**

**1. An assignment submission is considered complete only when correct and executable code(s) are submitted along with the documentation explaining the method and results. Failing to submit either of those will be considered an invalid submission and will not be considered for evaluation.**

**2. Assignments submitted after the deadline will affect your grades.**

**Grading:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Ans** | **Date** |  |  | **Ans** | **Date** |
| Correct | On time | A | 100 |  |  |
| 80% & above | On time | B | 85 | Correct | Late |
| 50% & above | On time | C | 75 | 80% & above | Late |
| 50% & below | On time | D | 65 | 50% & above | Late |
|  |  | E | 55 | 50% & below |  |
| Copied/No Submission |  | F | 45 |  |  |

* **Grade A: (>= 90):** When all assignments are submitted on or before the given deadline.
* **Grade B: (>= 80 and < 90):** 
  + When assignments are submitted on time but less than 80% of problems are completed.

(OR)

* + All assignments are submitted after the deadline.
* **Grade C: (>= 70 and < 80):** 
  + When assignments are submitted on time but less than 50% of the problems are completed.

(OR)

* + Less than 80% of problems in the assignments are submitted after the deadline.
* **Grade D: (>= 60 and < 70):**
  + Assignments submitted after the deadline and with 50% or less problems.
* **Grade E: (>= 50 and < 60):** 
  + Less than 30% of problems in the assignments are submitted after the deadline.

(OR)

* + Less than 30% of problems in the assignments are submitted before the deadline.
* **Grade F: (< 50):** No submission (or) malpractice.

**Problem Statements:**

1. Calculate skewness, kurtosis using R/Python & draw inferences on the following data.

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

a. Cars speed and distance

****

**There are 2 outliers in dist column.**

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

****

**SP is normally distributed whereas, WT is positively skewed.**

1. Draw inferences about the given box plot & 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]





Histogram shows data is positively skewed.

Box plot shows there are many outliers in the data. Mean would be greater than Median

1. 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, and standard deviation.

Mean – 41, Median 40.5, variance= 24.1, STD = 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].

Data is positively skewed and there are outliers:

Point 49 and 56

IQR = 4

Q3 = 42

Median = 40.5

Q2 = 38

Upper Fence Limit = 48

Lower Fence Limit = 32

Since 49 and 56 are outside of 1.5 IQR hence they are marginal outliers

1. What is the nature of skewness when the mean and median of the data are equal?

Symmetrical

1. What is the nature of skewness when mean > median?

Positive Skewed

6. What is the nature of skewness when median > mean?

Negative Skewed

1. What does a positive kurtosis value indicate for the data?

Peakness is more

8. What does a negative kurtosis value indicate for the data?

Peakness is little spread

9. Answer the below questions using the below box plot visualization.



1. What can we say about the distribution of the data?

Median value is around 15 or 15.5

1. What is the nature of skewness of the data?

Negative skewed

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

10. Comment on the below boxplot visualization?



Draw an inference from the distribution of data for the first boxplot with respect to the second boxplot.

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

Spread of 2nd chart is bigger than 1st chart

They both are normally distributed and symmetrical

Kurtosis of chart 1 is higher than chart 2

11.



Answer the following three questions based on the boxplot above.

1. What is the interquartile range of this dataset? [**Hint**: IQR = Q3 – Q1]. In one line, explain what this value implies. (**Hint:** Based on IQR definition)

Around 7. It means 50% of the data is within this range

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

Data is positively skewed

1. If it were found that the data point with the value 25 is 2.5, how would the new boxplot look? (**Hint:** On changing the data point from 25 to 2.5 in the data, how is it different from the current one.)

Based on the current values the Lower Fence limit is -5.5 and upper fence limit is 22.5. So if 25 is actually 2.5 then data would not have any outlier.

12.



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]

Between 4 and 8

1. Comment on the skewness of the dataset

Positive 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 the plots shows there an outlier (Assuming we have not corrected 25 to 2.5)

Both the plot shows data is skewed. Box plot has a longer whisker and histogram is showing the density decreases as value on Y axis increases with an outlier

**Note**: For each assignment question, the solution should be submitted in the below format.

**1. Research and perform all possible steps for obtaining a solution.**

**2. For statistical calculations, explanations of the solutions should be documented in black and white along with the codes.**

**3. 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 problems or short answer type questions, explanations are a must.**

**3.3. R & Python code for univariate analysis (histogram, box plot, bar plots etc.) the data distributions need to be attached**

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

**5. Code modularization should be followed.**

**6. The code must have comments explaining what each block of code does.**