**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: \_\_\_\_\_\_\_\_\_MADHUMITA KASINA\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Batch Id: \_\_\_\_\_\_\_\_23092021\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**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

****

**Output:**

|  |  |  |
| --- | --- | --- |
|  | Speed | Distance |
| Skewness | -0.11751 | 0.806895 |
| kurtosis | -0.50899 | 0.405053 |

Chart, histogram

Description automatically generated

**Graph of Histogram – Speed**

Chart, histogram

Description automatically generated

**Graph of Histogram – Distance**

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

****

**Output:**

|  |  |  |
| --- | --- | --- |
|  | SP | WT |
| Skewness | 1.611450 | -0.614753 |
| Kurtosis | 2.977329 | 0.950291 |

**import pandas as pd**

**import matplotlib.pyplot as plt**

#1(a)

# Read data into Python

**data = pd.read\_csv("C:/Users/madhu/OneDrive/Desktop/360digiTMG/Data Science/6 Statistical Data Visualization Plots/Hands-on Material/Statistical Datasets/Q1\_a.csv")**

# Third moment business decision

**data.speed.skew()**

**data.dist.skew()**

# Fourth moment business decision

**data.speed.kurt()**

**data.dist.kurt()**

#Histogram

**plt.hist(data.speed)**

**plt.hist(data.dist)**

# Measures of Central Tendency / First moment business decision

**data.speed.mean()**

**data.speed.median()**

**data.speed.mode()**

# Measures of Dispersion / Second moment business decision

**data.speed.var() # variance**

**data.speed.std() # standard deviation**

**range = max(data.speed) - min(data.speed) # range**

**range**

#1(b)

**import pandas as pd**

**dt = pd.read\_csv("C:/Users/madhu/OneDrive/Desktop/360digiTMG/Data Science/6 Statistical Data Visualization Plots/Hands-on Material/Statistical Datasets/Q2\_b.csv")**

# Third moment business decision

**dt.SP.skew()**

**dt.WT.skew()**

# Fourth moment business decision

**dt.SP.kurt()**

**dt.WT.kurt()**

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: Right side skewed or positively skewed**



**Answer**: **The interface for this box plot is positively skewed.**

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 | 41.0 |
| Median | 40.5 |
| variance | 25.529411764705884 |
| Standard Deviation | 5.05266382858645 |

**import pandas as pd**

**dt1 = pd.read\_excel("C:/Users/madhu/OneDrive/Desktop/360digiTMG/Data Science/6 Statistical Data Visualization Plots/Hands-on Material/Q3.xlsx")**

**dt1.Marks.mean()**

**dt1.Marks.median()**

**dt1.Marks.var()**

**dt1.Marks.std()**

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].

**Answer: - The scores are in uniformly distribution data in Ascending order**

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

**Answer: Normalized Skewness**

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

**Answer: The distribution is positively/ Right skewed**

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

**Answer: Left Skewed**

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

**Answer: - Sharp peak in the plot. less gap between tails to x-axis**

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

**Answer**: Border peak under the curve and more gap between tails and x-axis

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



What can we say about the distribution of the data?

**Answer: Non-Normal Distribution**

What is nature of skewness of the data?

**Answer: Left Skewed**

What will be the IQR of the data (approximately)?   
Answer: **IQR = Q3-Q1 = 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.]

**Answer: Data is Normally Distributed, No Outliers and No Skewness.**

**Boxplot 1 designed with range = 1.5 and Boxplot 2 with range = 3.**

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)

**Answer:**

Q1 = 5

Q3 = 12

Median (Q2) = 7

IQR = Q3 – Q1 = 12 – 5 = 7

**i.e., Second Quartile Range Q2 is the Median Value**

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

**Answer: The 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 be affected?

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

**Answer: In that case there would be no Outliers on the given dataset because of the outlier the data had positive skewness it will reduce and the data will normal distributed**

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]

**Answer: The mode lies on 10, 20 approx. (values of Y)**

1. Comment on the skewness of the dataset

**Answer: Positively/ 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]

**Answer: By checking the Histogram, data is not normally distributed. Outliers are present. Long tail towards right side i.e., Positively Skewed or Right Skewed.**

**By checking the Boxplot, No Symmetry and data has many Outliers.**

**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