**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: S Arunthathi**

**Batch ID:** **DSWDEOD 280622**

**Topic: Exploratory Data Analysis**

**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 as correct submission.**

**2. Ensure that you submit your assignments correctly. Resubmission is not allowed.**

**3. Post the submission you can evaluate your work by referring to keys provided. (will be available only post the submission).**

**Hints: Follow CRISP-ML(Q) methodology steps, where were appropriate.**

1. **Data Understanding: work on each feature of the dataset to create a data dictionary as displayed in the image below:**

Table

Description automatically generated

**Make a table as shown above and provide information about the features such as its data type and its relevance to the model building. And if not relevant, provide reasons and a description of the feature.**

**Problem Statements:**

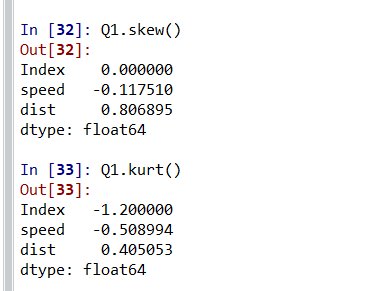
Q1) Calculate Skewness, Kurtosis using R/Python code & draw inferences on the following data. Refer to the Datasets attachment for data file.

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

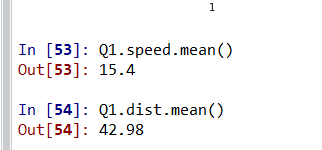
a. Cars speed and distance

****

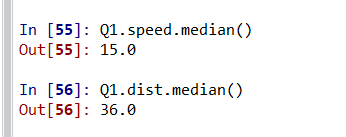
**Answers:**

****

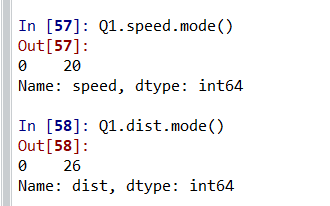
**Mean:**

****

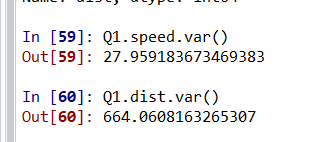
**Median:**

****

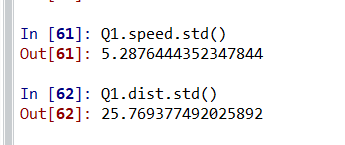
**Mode:**

****

**Variance:**

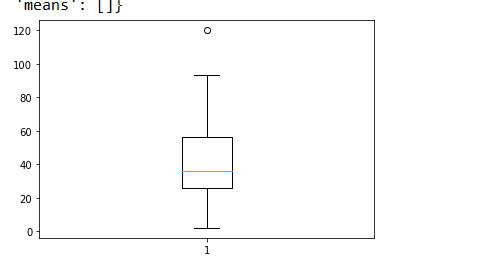
****

**Std. deviation:**

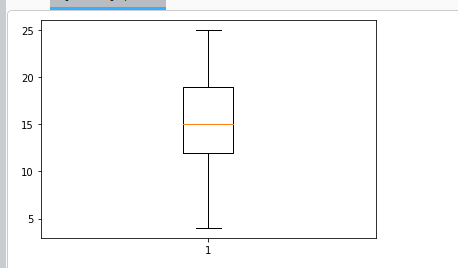
****

**Speed: (Plot)**

**Data is not normally distributed and it has minimum outliers.**

****

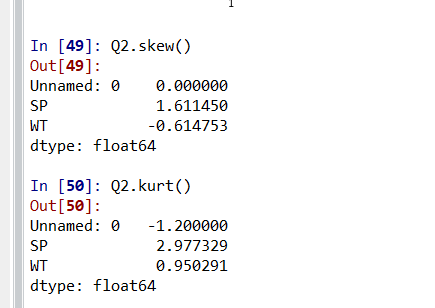
**Distance : (Plot) Data is normally distributed. But it has minimum variation.**

****

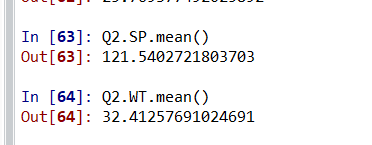
b. Top Speed (SP) and Weight (WT)

****

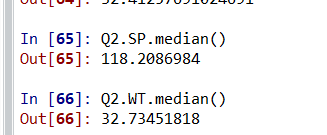
**Answers:**

****

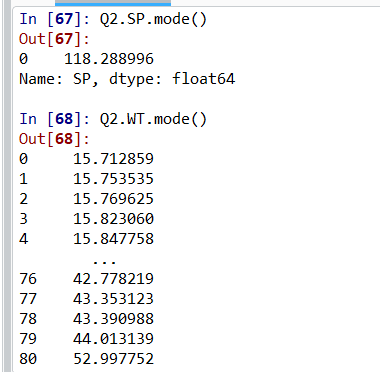
**Mean:**

****

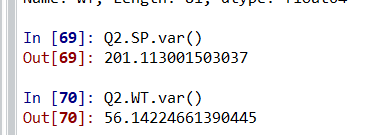
**Median:**

****

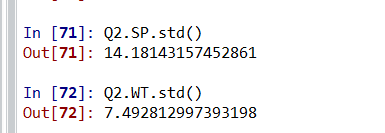
**Mode:**

****

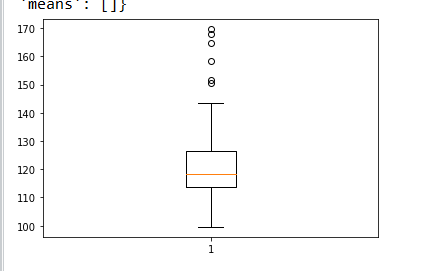
**Variance:**

****

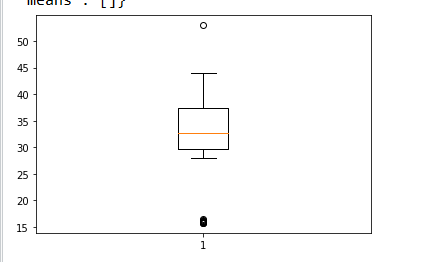
**St. Deviation:**

****

**SP: (Plot) data is not normally distributed. It has lot of outliers. Data is asymmetric it shows positive skewness**

****

**WT: (Plot) data is not normally distributed it has outliers in both positive and negative**

****

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:**

This graph histogram showing that the distribution of the data is asymmetric. This is right skewed and we also say that the this is positively skewed.



**Answer:**

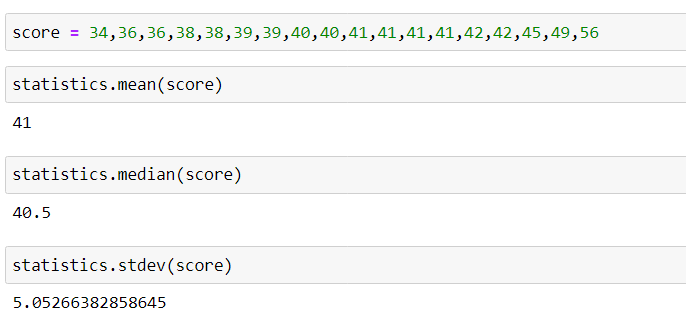
The boxplot showing positively skewness of the distribution.

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].

**Answer:**



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

Answer:- Skewness is zero

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

Answer: positively or right skewed

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

Answer:- Negatively skewed or left skewed

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

Answer:- This shows that the distribution has heavier tails than the normal distribution

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

Answer:- This shows that the distribution has lighter tails than the normal distribution

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



What can we say about the distribution of the data?

=The distribution of data on the high end of the scale

What is nature of skewness of the data?

= The distrubution is skewed left

What will be the IQR of the data (approximately)?   
= The IQR is 18 minus 10

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:- Both the boxplots show some symmetry means that they are normally distributed. There is no outlier, median is around260.

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)

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

**Answer**:

(i) Q1(Lower Quartile) = 5.

Q3(Upper Quartile) = 13

Inter quarter range of the above data set is IQR = 13-5 = 8.

This tells how spread out the middle 50% of values in data set.

(ii) The data set is positively skewed.

(iii) IQR will become 0.8 and the values we get will be in decimals.

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]
2. Comment on the skewness of the dataset
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]

**Answer**:

1. Mode lies between 4 and 8 values on Y axis.
2. The data set is positively skewed.
3. The boxplot and the histogram showing positively skewness of the distribution