**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: \_\_Uppara. Naveen Kumar\_\_\_\_\_\_\_\_\_\_\_ Batch ID:** \_\_\_07092022\_\_\_\_\_\_\_\_

**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

****

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]



Ans: Histogram: - Chick weight data is right skewed or positively skewed--- Yes- More than 50% Chick Weight is between 50 to 150. ---- Yes- Most of the chick weight is between 50 to 100. --- Yes

Box plot: Right skewed (Mean>Median>mode)

More outliers at upperside



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].

Ans: Mean = 41, median = 40.5, mode = 41, Positively Skewed, 49 and 56 are outliers

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

Ans: Symmetrical skewed

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

Ans: Right skewed

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

Ans: Left skewed

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

Ans: Leptokurtic

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

Ans: Platykurtic

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



What can we say about the distribution of the data? Ans: Non-uniformly distributed

What is nature of skewness of the data? Ans: Left skewed

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

Ans:(Equally distributed, No outliers, Symmetrically skewed and Mesokurtic kurtosis)

Q12)



Answer the following three questions based on the boxplot above.

1. What is inter-quartile range of this dataset? [**Hint**: IQR = Q3 – Q1] Ans: 7

In one line, explain what this value implies. (**Hint:** Based on IQR definition) Skewed distributions, to find outliers

Ans: (how spread out the middle 50% of our set of data is)

1. What can we say about the skewness of this dataset? Ans: (Positively Skewed)
2. If it were found that the data point with the value 25 is 2.5, how would the new boxplot be affected? Ans: (the first quartile get affected and thus the interquartile range changes)

(**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: (The mode of this dataset would lie between 4 to 8.

The mode is the data value that occurs the most in a dataset. In the above figure, it is clearly seen that the frequency is maximum somewhere between the values 4 to 8. Hence, the mode will definitely lie between these values.)

1. Comment on the skewness of the dataset

Ans: (The dataset is right skewed.

Skewness refers to the distortion from a normal distribution in a dataset. It can be observed that the dataset is shifted towards the right end with higher peaks than the average. Hence the 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: (Both the graphs are right skewed whose outliers and median are the same.

The boxplot is also right skewed, hence when these are plotted for the same dataset, the graphs will be right skewed. Outliers and median don't change when depiction of a dataset is changed. Since here the same dataset will be represented in boxplot and histogram, they both will be same.)