**2b. Data Visualization**

**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: \_\_Seema Iyengar\_\_\_ Batch ID:** \_\_EDA\_06042023\_10AM\_\_

**Topic: Data Visualization**

**Guidelines:**

**1. An assignment submission is considered complete only when the correct and executable code(s) is 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 a 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 the 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, and Kurtosis using Python code & draw inferences on the following data. Refer to the Datasets attachment for the 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 ->**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name of the Feature** | **Description** | **Type** | **Relevance** | **Explanation** |
| speed | Speed of the Car | Ratio,Quantitative,Structured | Speed is a Relevant Column | Ratio Type because all statistical analysis can be performed Quantitative because it is a numerical data Structured because it can be represented in tabular format |
| dist | Distance travelled by the car | Ratio,Quantitative,Structured | Distance is a Relevant Column | Ratio Type because all statistical analysis can be performed Quantitative because it is a numerical data Structured because it can be represented in tabular format |

**import pandas as pd**

**import numpy as np**

**import matplotlib.pyplot as plt**

**cars = pd.read\_csv(r"C:\Users\seema\OneDrive\Desktop\COURSE\360DigiTMG Course\EDA\_06042023\_10AM\DataSets\cars.csv")**

**cars.info()**

**# Third moment business decision / Skewness**

1. **cars.speed.skew() # skewness is negative(<0)**

**Out[9]: -0.895424912156393**

1. **cars.dist.skew() # skewness is positive(>0)**

**Out[10]: 1.29076266205913**

**# Fourth moment business decision**

1. **cars.speed.kurt() # kurtosis is negative(<3)**

**Out[11]: 0.24956097515622666**

1. **cars.dist.kurt() # kurtosis is negative(<3)**

**Out[12]: 2.464545503089613**

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

****

**Answers ->**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name of the Feature** | **Description** | **Type** | **Relevance** | **Explanation** |
| SP | Speed of the Car | Ratio,Quantitative,Structured | Speed is a Relevant Column | Ratio Type because all statistical analysis can be performed Quantitative because it is a numerical data Structured because it can be represented in tabular format |
| WT | Weight of the Car | Ratio,Quantitative,Structured | Distance is a Relevant Column | Ratio Type because all statistical analysis can be performed Quantitative because it is a numerical data Structured because it can be represented in tabular format |

**cars1 = pd.read\_csv(r"C:\Users\seema\OneDrive\Desktop\COURSE\360DigiTMG Course\EDA\_06042023\_10AM\DataSets\cars1.csv")**

**cars1.info()**

**# Third moment business decision / Skewness**

1. **cars1.SP.skew() # skewness is negative(<0)**

**Out[17]: -0.4267574826349179**

1. **cars1.WT.skew() # skewness is negative(<0)**

**Out[18]: -1.3474621607317778**

**# Fourth moment business decision**

1. **cars1.SP.kurt() # kurtosis is negative(<3)**

**Out[21]: -0.8637279023486579**

1. **cars1.WT.kurt() # kurtosis is negative(<3)**

**Out[22]: 1.1528173104724373**

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]



**Answers ->**

**As per the given above histogram, insights are as follows.**

**The Distribution of ChickWeight is Right Skewed (Positive Skewed) as the tails extend towards right side. The bin range is 100 units.**



**Answers ->**

**Boxplot has outliers as the small circles are present outside the Max (upper whisker) of the boxplot. As the Q2 (50th percentile) or the median lies slightly near to the Q1(25th percentile), the data distribution is left skewed or negative skewed.**

Q3) Below are the scores obtained by a student on tests

**34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56**

1. Find the mean, median, variance, and standard deviation.

**Answers ->**

**a = [34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56]**

**b = pd.Series(a)**

**b.mean()**

**Out[37]: 41.0**

**b.median()**

**Out[52]: 40.5**

**b.var()**

**Out[53]: 25.529411764705884**

**b.std()**

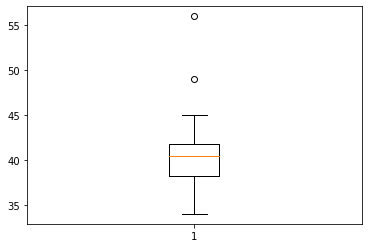
**Out[54]: 5.05266382858645**

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].

**plt.figure()**

**plt.boxplot(b)**

**As per the below boxplot image, the student marks is not normally distributed and is right/positive skewed. Also, the outliers are present as the small circles are present outside the MAX(upper whisker).**



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

**Answers ->**

**In a distribution with zero skew, the mean and median are equal. Zero skew: mean = median.**

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

**Answers ->**

**If the mean is greater than the median, the distribution is positively skewed.**

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

**Answers ->**

**If the median is greater than the mean, the distribution is negatively skewed.**

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

**Answers ->**

**Positive values of kurtosis indicate that distribution is peaked and possesses thick tails.**

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

**Answers ->**

**A distribution with a negative kurtosis value indicates 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?

**Answer-> The data is normally distributed as there are no outliers present in the boxplot.**

What is the nature of the skewness of the data?

**Answer-> The nature of skewness of the data is positive as the median or Q2 is towards the Q3 quartile.**

What will be the IQR of the data (approximately)?   
**Answer-> IQR (InterQuartileRange) = Q3 – Q1**

**Approximately = Q3 = 18, Q1 = 10.**

**Therefore, IQR = 18-10 = 8(answer).**

Q11) Comment on the below Boxplot visualizations.



Draw an Inference from the distribution of data for Boxplot 1 with respect to Boxplot 2.

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

**Answers->**

**Both 1) and 2) boxplots have data normally distributed.**

**Both 1) and 2) boxplots have no outliers.**

**Both 1) and 2) boxplots have skewness at the centre or normally skewed.**

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)

**Answers->**

**IQR = Q3 – Q1 , Q3 = 12, Q1 = 5**

**Therefore, 12 – 5 = 7.**

**IQR = 7. This implies that the Median or Q2(50th percentile) = 7 the IQR.**

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

**Answers ->**

**Skewness is positive as the data after Q2 is distributed towards the MAX (right whisker).**

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.)

**Answers->**

**On changing the data point from 25 to 2.5 in the data, the box plot would not have outliers and 2.5 would be MIN value of boxplot.**

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 the Y-axis]

**Answers ->**

**Mode is the highest peak of the histogram, in this case, 21 is the mode.**

1. Comment on the skewness of the dataset

**Answers ->**

**The skewness of the dataset is positive or right skewed as there are a lot of tails extended towards the right.**

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]

**Answers ->**

**import pandas as pd**

**import numpy as np**

**import matplotlib.pyplot as plt**

**cars = pd.read\_csv(r"C:\Users\seema\OneDrive\Desktop\COURSE\360DigiTMG Course\EDA\_06042023\_10AM\DataSets\cars.csv")**

**cars.info()**

**<class 'pandas.core.frame.DataFrame'>**

**RangeIndex: 27 entries, 0 to 26**

**Data columns (total 2 columns):**

**# Column Non-Null Count Dtype**

**--- ------ -------------- -----**

**0 speed 27 non-null int64**

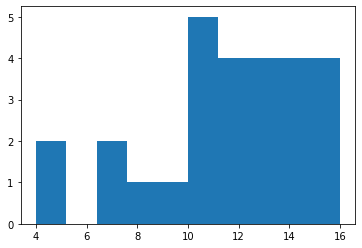
**1 dist 27 non-null int64**

**dtypes: int64(2)**

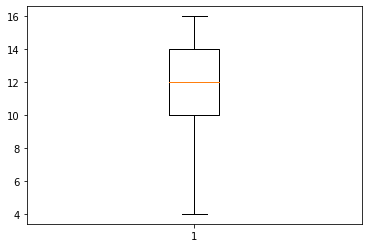
**memory usage: 560.0 bytes**

**plt.figure()**

**plt.hist(cars.speed)**



**plt.boxplot(cars)**



**As per visualization of both the plots, insights are as below.**

**In histogram, the data for speed is left skewed and not normally distributed. Mode is 5. There are outliers present in the speed column of the dataset.**

**In boxplot, the data for speed is normally distributed and skewness is normal. Q2 or the Median is 12 and there are no outliers present in the speed data column of the dataset.**