**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 such as mean, median, mode, variance, std. deviation]

a. Cars Speed and Distance b. Top Speed (SP) and Weight (WT)

**Solutions:-**

**Q1.a.**

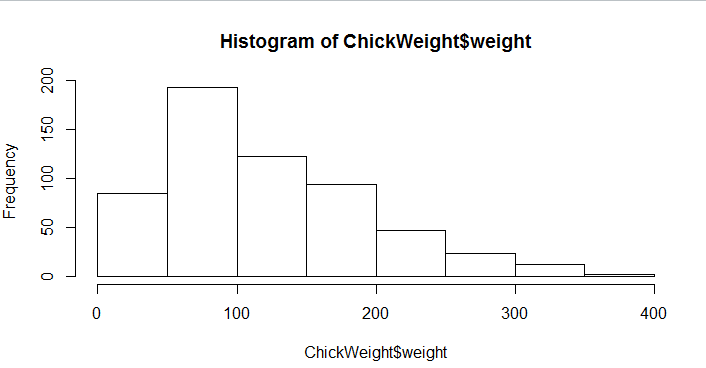
| **Sr. No.** | **Descriptions** | **Speed** | **Distance** |
| --- | --- | --- | --- |
| 1 | Mean | 15.400000 | 42.980000 |
| 2 | Median | 15.0 | 36.0 |
| 3 | Variance | 27.959183673469383 | 664.0608163265307 |
| 4 | Standard Deviation | 5.287644 | 25.769377 |
| 5 | Skewness | -0.117509861446633 | 0.8068949601674215 |
| 6 | Kurtosis | -0.508994420405761 | 0.4050525816795765 |

**Q1.b.**

| **Sr. No.** | **Descriptions** | **Speed** | **Weight** |
| --- | --- | --- | --- |
| 1 | Mean | 121.540272 | 32.412577 |
| 2 | Median | 118.20869840000002 | 32.73451818 |
| 3 | Variance | 201.11300150303705 | 56.14224661390445 |
| 4 | Standard Deviation | 14.181432 | 7.492813 |
| 5 | Skewness | 1.6114501961773555 | -0.6147533255357768 |
| 6 | Kurtosis | 2.9773289437871764 | 0.9502914910300326 |

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]



**Solution:-**

By the **histogram** of ChickWeight$weight the following conclusions (inferences) can be drawn-

1. The tendancy of data concentration is more in the bins or range in between 50-100 as seen.
2. At the same time the least range of weight 350-400 something around 2-10.
3. About skewness, the long tail portions towards the right side. So, it is positively skewed.



By the **boxplot** as shown here in figure, the following inferences can be drawn-

1. The median is closer to first quartile range, and median is less than mean, so it is right skewed.
2. It contains outliers on the upper side of the boxplot.
3. There are a less data sets or points in between the first quartile and lower extreme.

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].

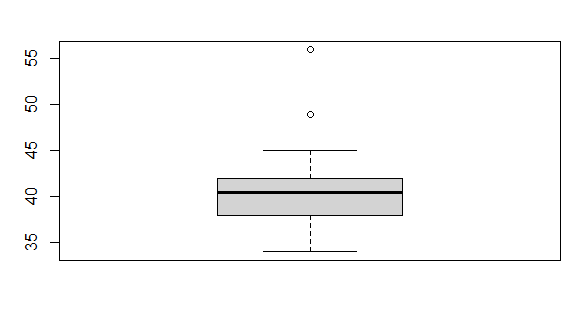
**Solution:-**

1. Mean = 41, Median=40.5, Variance=25.52941, Standard Deviation=5.052664
2. **Skewness=1.542885**, as skewness is greater than zero (0) **right skew distributions** value stands to concentrate on left side, so low values are more frequent than the high values.

As **Mean is greater than Median,** It means **distribution is positively skewed** or positive skewness.

**Kurtosis=5.621631**, as the result value is greater than 3, the **left kurtosis distributions** will occur which is very sharp distribution and the values have the tendency to concentrate near or arround the mean

**Outliers** there are two outliers 49 & 56



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

**Solution:-**

When mean and median of some data is equal, it implies that it is the case of **normal skewness** i.e. **no skew or zero skewness** and it is a perfectly symmetrical distributions

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

**Solution:-**

When mean is greater than median of some data, it implies that it is the case of **positive skewness** and it is not a perfectly symmetrical distributions

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

**Solution:-**

When median is greater than mean of some data, it implies that it is the case of **negative skewness** and it is not a perfectly symmetrical distribution

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

**Solution:-**

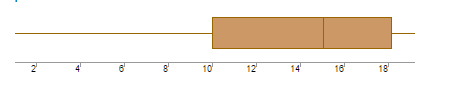
**Positive kurtosis** value indicates or implies **peaked and thick tail.**

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

**Solution:-**

For symmetrical distributions **negative kurtosis** value indicates or implies **wider peak and thinner tails.**

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



What can we say about the distribution of the data?

**Solution:-** As per the boxplot the 50% data is more than 10 and remaining is less. And the data above 15 are approximately 40%.

Negatively Skewed Distributions, as the median is coming closer to the 3rd quartile Q3.

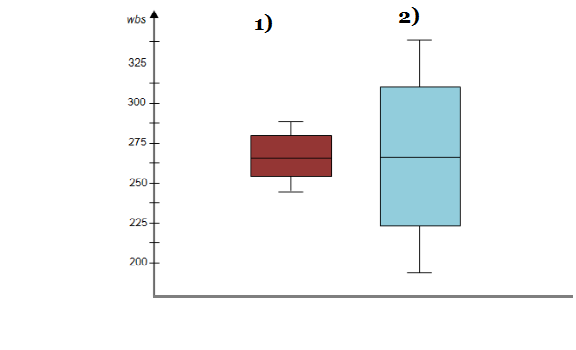
What is nature of skewness of the data?

**Solution:-** Negatively skewed distributions**,** as Median is greater than Mean.

What will be the IQR of the data (approximately)?

**Solution:- 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.]

**Solution:-** On comparing both the box plots have equal properties around the median, so it is the case of symmetric or normal distributions.

The whisker level is high in boxplot 2 as comparing with boxplot 1.

There is no outlier as by observations.

There is zero skewness.

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)

**Solution:-** IQR = Q3 – Q1 = 12-5 = 7

The IQR value i.e. 7 implies the Median

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

**Solution:-** Positively Skewed Distributions, as the median is coming closer to the 1st quartile Q1.

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.)

**Solution:-** Box will help to show, there is no outliers. The Q1 i.e. first quartile will move to left side.

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]

**Solution:-** The mode of the given dataset will lies in between 5 and 7, by observing the histogram. In terms of Y-axis.

1. Comment on the skewness of the dataset

**Solution:-** As per the distributions concerned it is positively skewed i.e.skewness of histogram will move towards right side.

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]

**Solution:-** The histogram will shows the frequency only or histogram deals with the frequency distributions. At the same time boxplot will gives us Lower extreme, First quartile (Q1), Median, Third quartile (Q3), Upper extreme and outliers also.