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: KAVALI RAKESH YADAV

Batch Id: DSWDMCON 180122

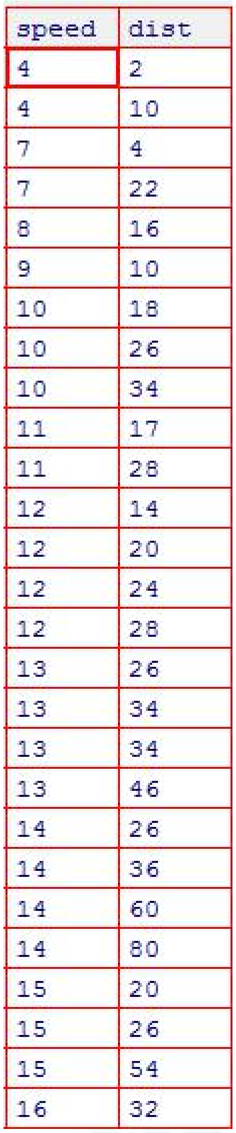
Topic: Exploratory Data Analysis

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]

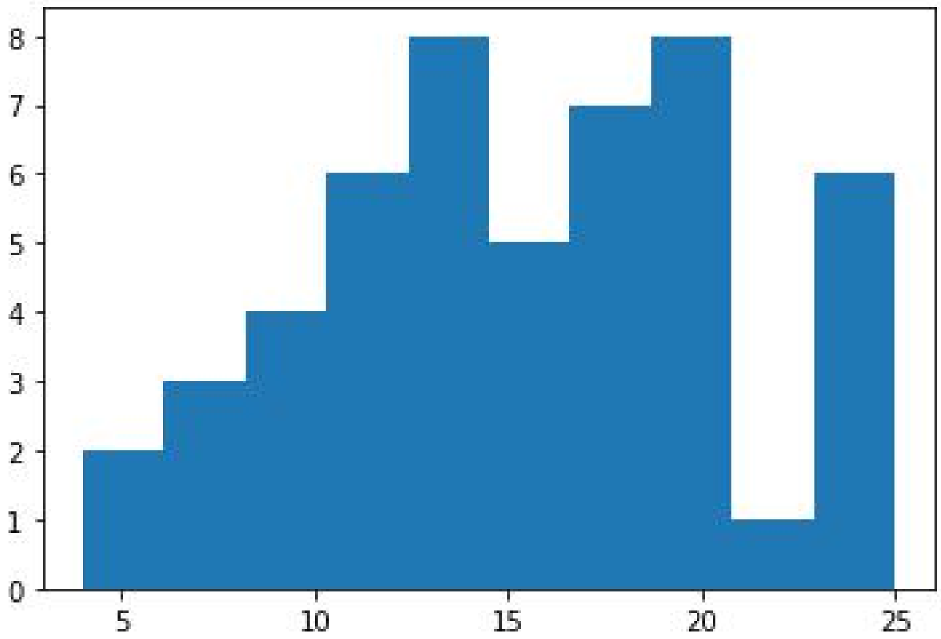
a. Cars speed and distance



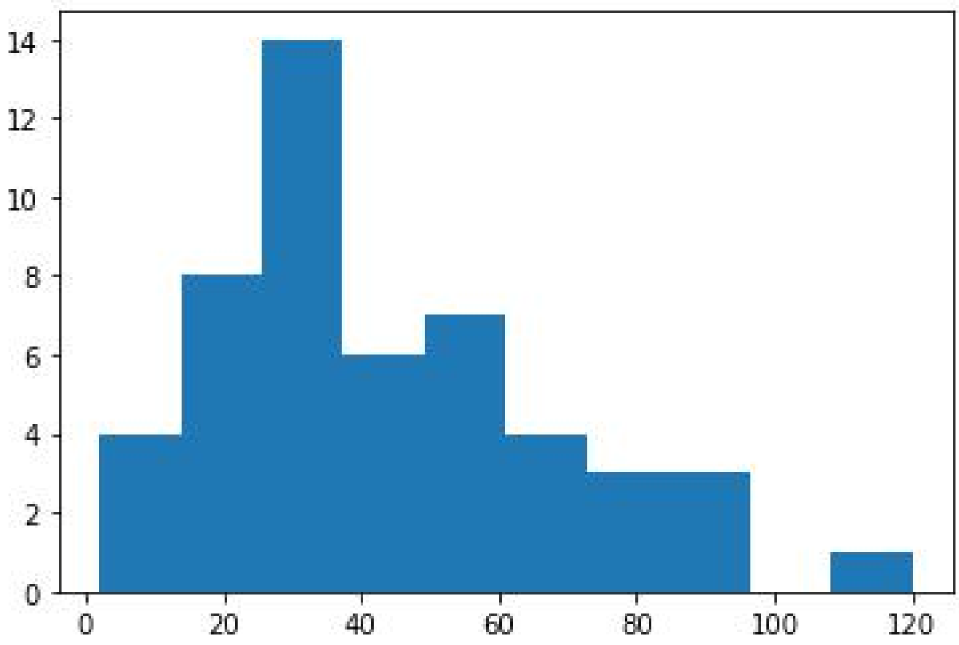
speed	dist
4	2
4	10
7	4
7	22
8	16
9	10
10	18
10	26
10	34
11	17
11	28
12	14
12	20
12	24
12	28
13	26
13	34
13	34
13	46
14	26
14	36
14	60
14	80
15	20
15	26
15	54
16	32

Column	mean	meadian	mode	variance	std.dev	skewness	kurtosis
Speed	15.4	15.1	20	27.9591	5.287	-0.11	-0.5089
Dist	42.98	36	26	664.0608	25.769	0.806	0.405

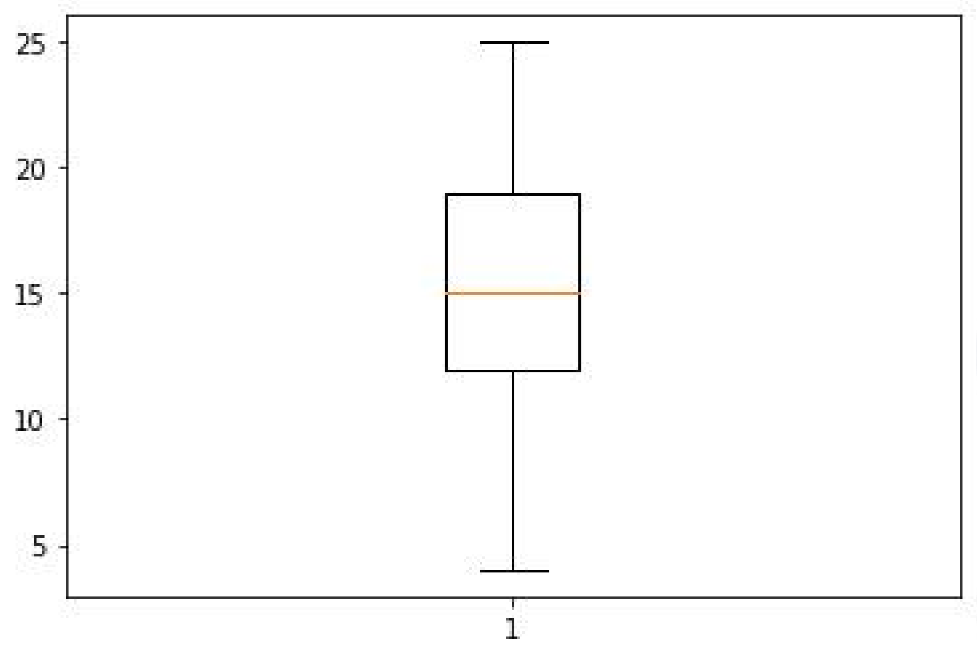
We can infer that speed is negatively skewed and distance is positively skewed.From the boxplot we can say that distance has outliers



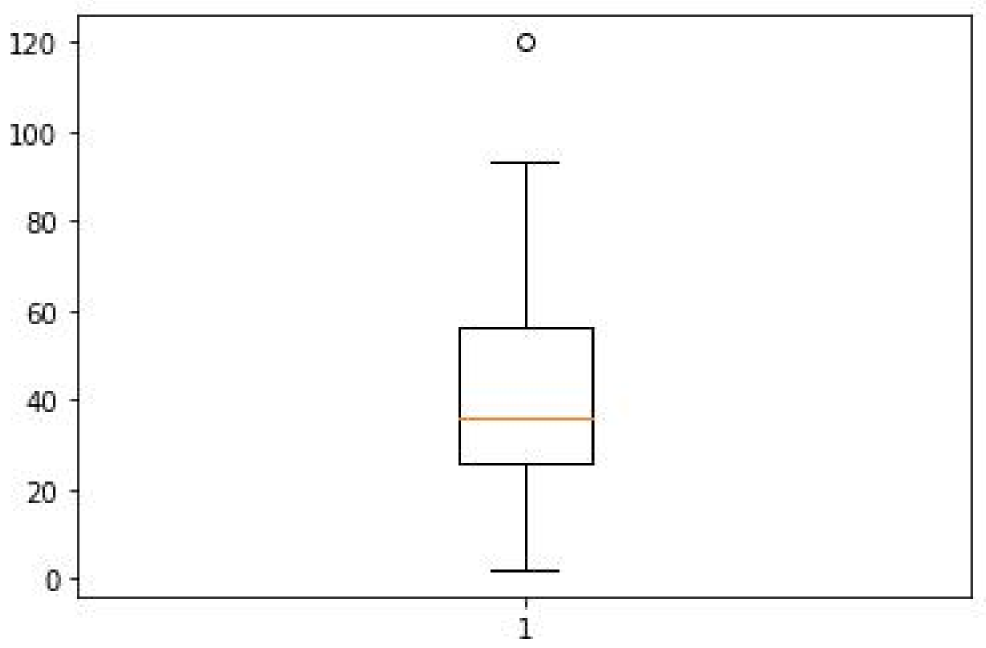
Histogram for speed



Histogram for distance

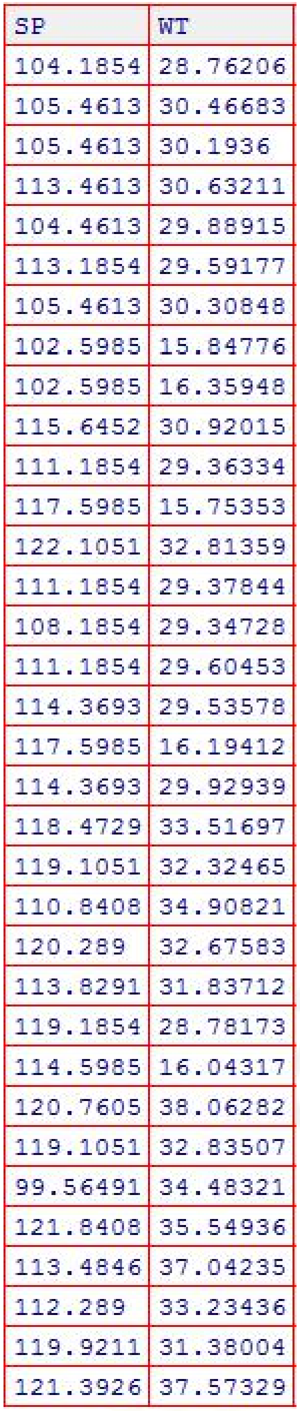


Boxplot for speed



Boxplot for distance

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



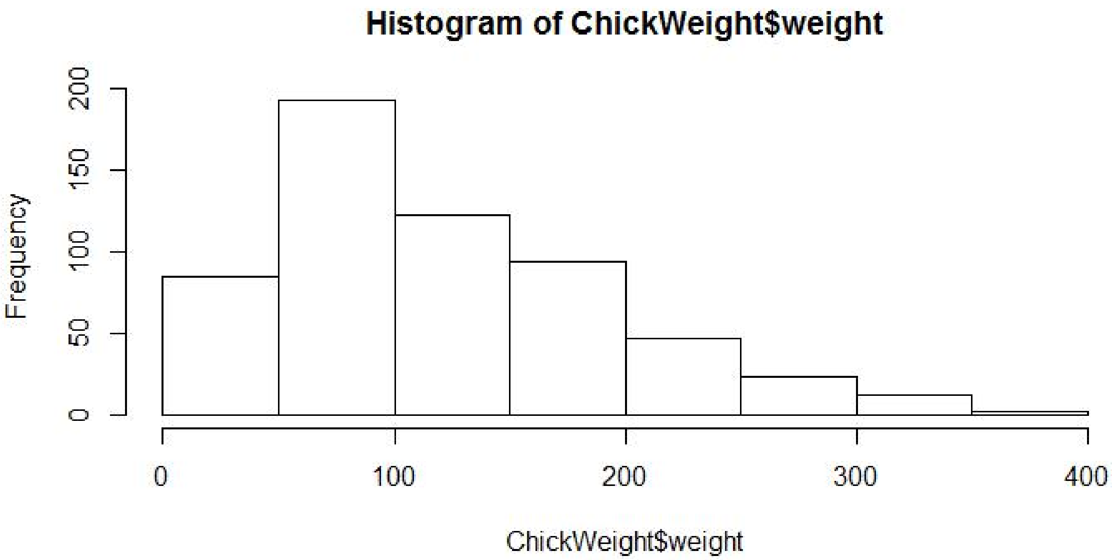
SP	WT
104.1854	28.76206
105.4613	30.46683
105.4613	30.1936
113.4613	30.63211
104.4613	29.88915
113.1854	29.59177
105.4613	30.30848
102.5985	15.84776
102.5985	16.35948
115.6452	30.92015
111.1854	29.36334
117.5985	15.75353
122.1051	32.81359
111.1854	29.37844
108.1854	29.34728
111.1854	29.60453
114.3693	29.53578
117.5985	16.19412
114.3693	29.92939
118.4729	33.51697
119.1051	32.32465
110.8408	34.90821
120.289	32.67583
113.8291	31.83712
119.1854	28.78173
114.5985	16.04317
120.7605	38.06282
119.1051	32.83507
99.56491	34.48321
121.8408	35.54936
113.4846	37.04235
112.289	33.23436
119.9211	31.38004
121.3926	37.57329

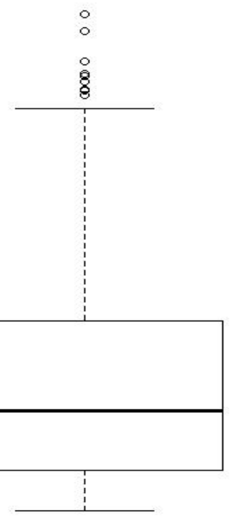
Column	mean	meadian	mode	variance	std.dev	skewness	kurtosis
Speed	121.54	118.208	118.288	201.113	14.181	1.611	2.977
Weight	32.41	32.73	15.712	56.141	7.492	-0.614	0.95

Ans) We can infer from the histogram that speed is right skewed or positively skewed.On the other hand weight is negatively skewed or left skewed.In the python code ,both speed and weight has outliers.

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) From the histogram we can infer that it is positively skewed or right skewed. From the boxplot, we can infer that there are outliers after the outerfence.so we can say it 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.

Ans) Mean=41

Median=40

Mode=41

Variance=27

Standard Deviation=5.7

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) Here mean>median.it is positively skewed. from boxplot we can infer that there are outliers

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

Ans) Normally Distributed

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

Ans) Positive Skew

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

Ans)negative skew

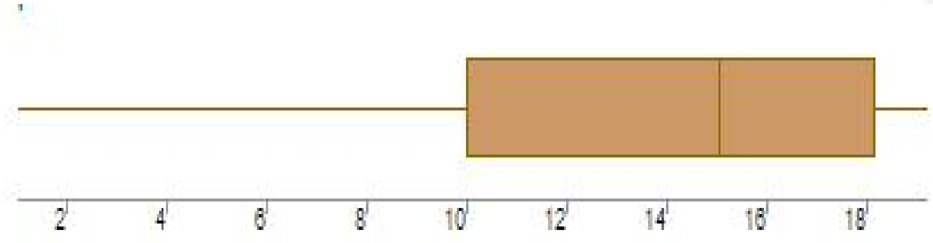
Q8) What does positive kurtosis value indicates for a data?

Ans) Thin peak and thick tail

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

Ans)   Wide peak and thin tails

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



What can we say about the distribution of the data?

Ans) Boxplot shows that it is not normally distributed.Data distributed between 10 and 18.

What is nature of skewness of the data?

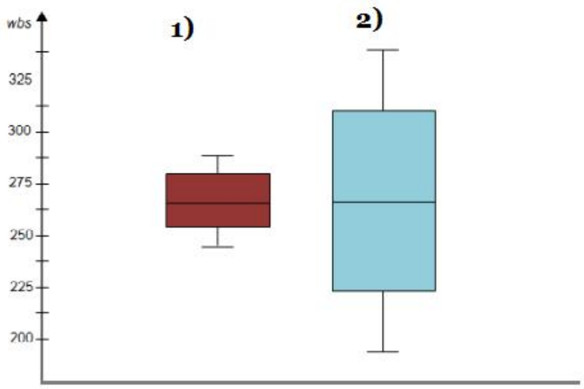
Ans) Positive skew

What will be the IQR of the data (approximately)?
ans) Q1=10,Q3=18

IQR=Q3-Q1

= 8

Q11) Comment on the below Boxplot visualizations?

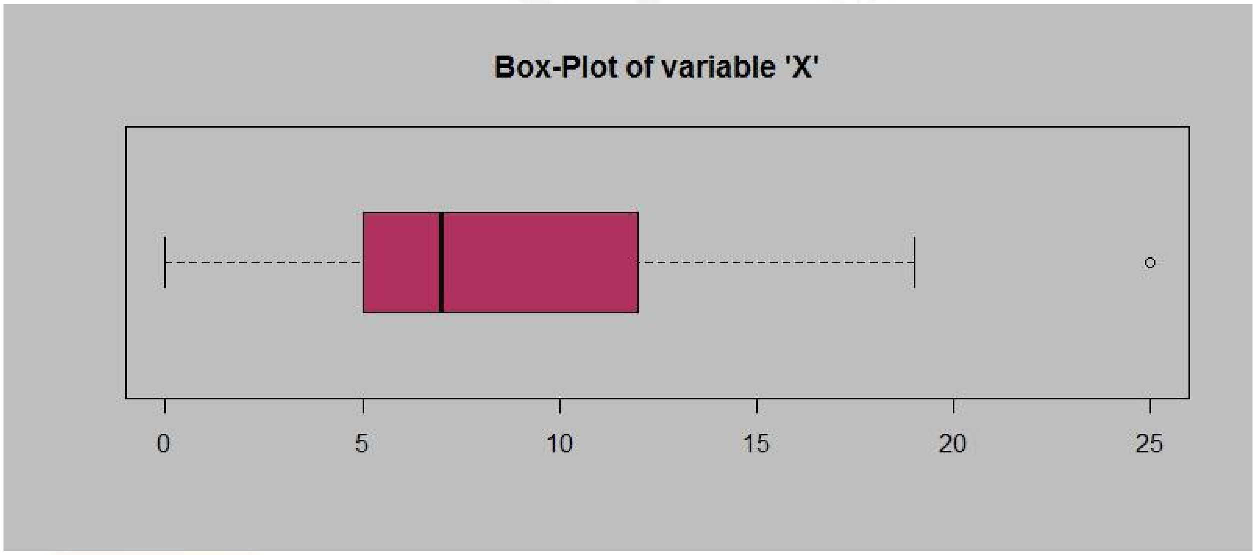


Draw an Inference from the distribution of data for Boxplot 1 with respect Boxplot 2.

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

Ans) From the above boxplots we can infer that both are normally distributed with median 262.5

Q12)



Answer the following three questions based on the boxplot above.

(i) What is inter-quartile range of this dataset? [**Hint**: IQR = Q3 – Q1](**IQR=12-5=7)**

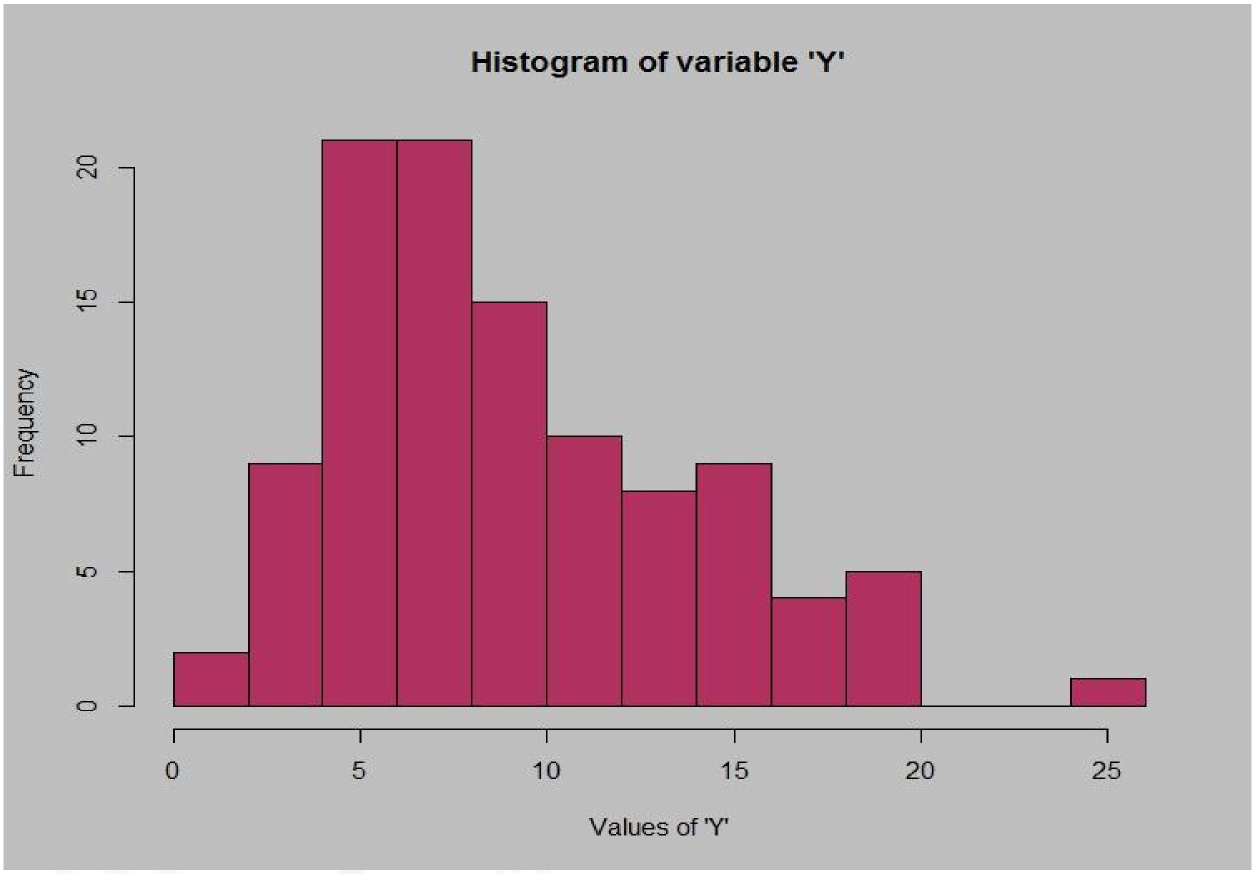
In one line, explain what this value implies. (**Hint:** Based on IQR definition) **(IQR shows the width of the data where date is concentrated)**

(ii) What can we say about the skewness of this dataset?(**positive skew**)

(iii) If it were found that the data point with the value 25 is 2.5, how would the new boxplot be affected? (**there will be no outliers**)

(**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.

(i) Where would the mode of this dataset lie? **Hint:** [In terms of values On Y-axis](**mode is the most repeating value from the histogram it is 20)**

(ii) Comment on the skewness of the dataset(**positive skew as tail is towards right side)**

(iii) 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](**From the histogram there is positive skewness and from the boxplot also we can find the skewness of the data.The presence od outliers can be found using the boxplot**

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