

## Answer 1

I am working in FinTech Sector for loyalty rewards program for the credit and debit cards. Thinking about the statistical question applicable to my job will be:

How many points will be rewarded to the customers during a sale Campaign, swiping/using the card from my company?

To answer this question the details that are required are:

1. First six and last four card number
2. Transaction ID
3. Store ID
4. Transaction Amount
5. Transaction Date
6. Transaction Type
7. Customer ID

These details will help us understand customer spending habits and the type - online or offline – of shopping they are engaging in. This will in turn help prepare for the next sale campaign as:

1. The shopping trends of the customer will help us understand the areas – fashion, electronics, jewellery, etc. – of interest.
2. The Type of shopping will help understand where the customers are spending more money, which in turn will help my company partner with these retailers thereby increasing the usage of the company's card.

Answer 2

Questions are the questions that can be answered by collecting data generated using a variable process generating data over time.

Applying the above concept options: **b, c, e** qualifies as statistical questions.

**Ans: B, C, E**

Answer 3

**Population:** A population consists of all elements whose characteristics are under consideration or study. (Referred Prem S Mann & Class notes)

**Sample:** Sample is a subset of population. A sample used for study represents population as closely as possible. (Referred Prem S Mann & Class notes)

Applying the concept of both the definition, **Option C** is the answer.

**Ans: C**

#### Answer 4

As given in the questionnaire, there are 50 employees and later 4 employees are added, and sample set becomes 54.

Employee	1	2	3	...	25	26	27	28	..	50	51	52	53	54
50	3,20,000	..	...	...	Median = 21,42,000					54,00,000				
54	2,40,000	2,84,000	3,20,000	...	...	...	Median =?			...	...	54,00,000	54,12,000	54,20,000

As 2 salaries less than 3,20,000 is added and other two salaries are added higher than 54,00,000, the middle values in the sample are not affected. So, the previous sample data will be shifted two places after adding 4 new salaries.

In the set of 50 employees the median would be calculated using 25<sup>th</sup> and 26<sup>th</sup> places salaries and in the set of 54 employees the median would be calculated using 27<sup>th</sup> and 28<sup>th</sup> position salaries. As established earlier that previous 50 employee's data position will be shifted two places after new addition, the median would be the same in both the dataset.

**Ans: B (Is exactly equal to Rs. 21,42,000)**

Answer 5

Skewness measures asymmetry in data distribution.

In right skewness, the  $\text{Mode} < \text{Median} < \text{Mean}$ .

Applying the concept, **Option C** is the answer.

**Ans: Option C i.e., Mean is greater than median.**

Answer 6

The data measured are quantitative in nature. Also, the unit of measurement for temperature does not have true zero. So, it is Interval in nature.

**Ans: C Interval**

Answer 7

Interval	Frequency	Cumulative Frequency
501-550	25	25
551-600	40	65
601-650	42	107
651-700	31	138
701-750	26	164
751-800	21	185
Total	185	N = 185

Median Interval is the interval where  $\frac{N}{2} = \frac{185}{2} = 92.5$ .

The next nearest cumulative frequency to 92.5 is 107. So, the Median Interval is 601-650.

**Ans: The median class of above GMAT score id 601-650.**

Answer 8

After updating data table

Interval	Frequency
501-550	25
551-600	53
601-650	42
651-700	31
701-750	48
751-800	21

The modal class is the class with highest frequency i.e., interval 551-600.

**Ans: The modal class of new data is 551-600.**

Answer 9

a. Age of household head.

Type Of Data	Scale Of Measurement
Quantitative	Interval

b. Sex of household head.

Type Of Data	Scale Of Measurement
Qualitative	Nominal

c. Number of people in household.

Type Of Data	Scale Of Measurement
Quantitative	Interval

d. Use of electric heating (yes or no).

Type Of Data	Scale Of Measurement
Qualitative	Nominal

e. Number of large appliances used daily.

Type Of Data	Scale Of Measurement
Quantitative	Ratio

f. Thermostat setting in winter.

Type Of Data	Scale Of Measurement
Qualitative	Ordinal

g. Average number of hours heating is on.

Type Of Data	Scale Of Measurement
Quantitative	Interval

h. Average number of heating days.

Type Of Data	Scale Of Measurement
Quantitative	Interval

i. Household income.

Type Of Data	Scale Of Measurement
Quantitative	Interval

j. Average monthly electric bill.

Type Of Data	Scale Of Measurement
Quantitative	Interval

k. Ranking of this electric company as compared with two previous electricity suppliers.

Type Of Data	Scale Of Measurement
Qualitative	Ordinal

Answer 10

Converting the graphical data into tabular format.

Books	New Child	New Adult	Used Adult	Used child	Total
Classics	300	180	200	180	860
Science Fiction	380	50	90	350	870
Biography	40	380	260	20	700
Textbook	250	210	50	20	530
Total	970	820	600	570	2960

As per the summarized data, the sales of used books of adult classics are greater than that of new books.

**Ans: Option B Adult Classics**

Answer 11

Books	New Child	Used child	Total Child	Used Adult	New Adult	Total Adult	Total
Classics	300	180	480	200	180	380	860
Science Fiction	380	350	730	90	50	140	870
Biography	40	20	60	260	380	640	700
Textbook	250	20	270	50	210	260	530
Total	970	570		600	820		2900

Option A: The Science Fiction in Child category has the largest sale, and the Science Fiction in Adult category has the least sale.

Option B: The Biography Fiction in Adult category has the largest sale, and the Biography Fiction in Child category has the least sale.

**Ans: Option A & B**

Answer 12

Books	New Child	Used child	Used Adult	New Adult	Total
Classics	300	180	200	180	860
Science Fiction	380	350	90	50	870
Biography	40	20	260	380	700
Textbook	250	20	50	210	530
Total	970	570	600	820	2900

Cannot be inferred as the information about the price per book is not provided in the question.

**Ans: Option D**

Answer 13

a.

From the output of the numerical analysis, it can be deduced that Median < Mean. So, to understand the data visually I have opted for Histogram and Boxplot.

For numerical analysis:

I have used the "summary" function in R.

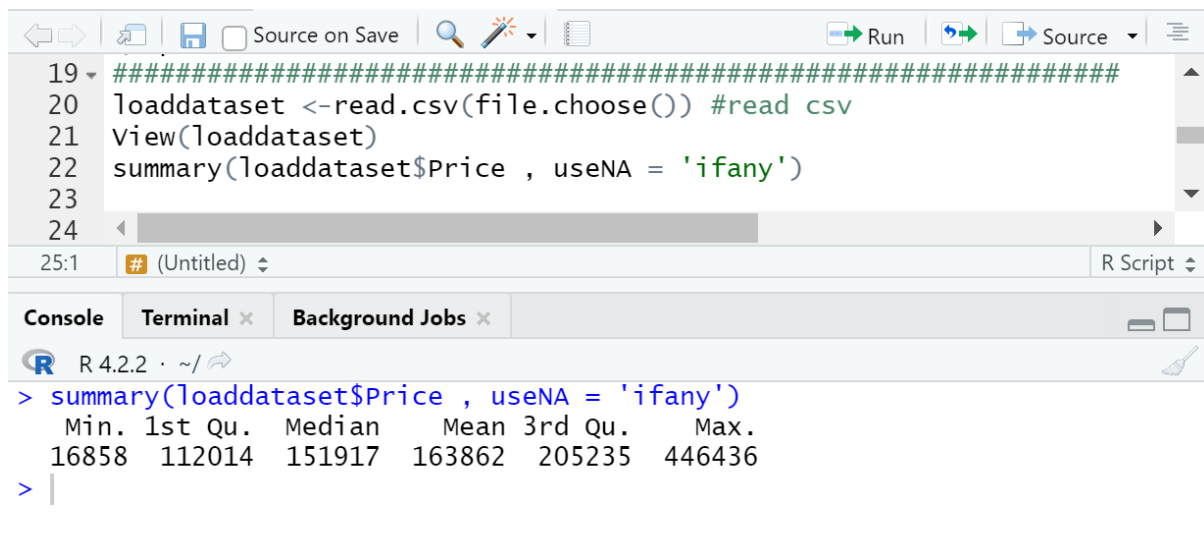
Code:

```
loaddataset <- read.csv(file.choose()) #read csv
```

```
View(loaddataset)
```

```
summary( loaddataset$Price, useNA = 'ifany')
```





```
19 #####
20 loaddataset <- read.csv(file.choose()) #read csv
21 View(loaddataset)
22 summary(loaddataset$Price , useNA = 'ifany')
23
24
```

25:1 |# (Untitled) | R Script

Console | Terminal x | Background Jobs x

```
R 4.2.2 · ~/
> summary(loaddataset$Price , useNA = 'ifany')
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
16858  112014  151917  163862  205235  446436
> |
```

Brief Report:

Min value: 16858

1<sup>st</sup> Quartile:112014

2<sup>nd</sup> Quartile/median:151917

Mean:163862

3<sup>rd</sup> Quartile: 205235

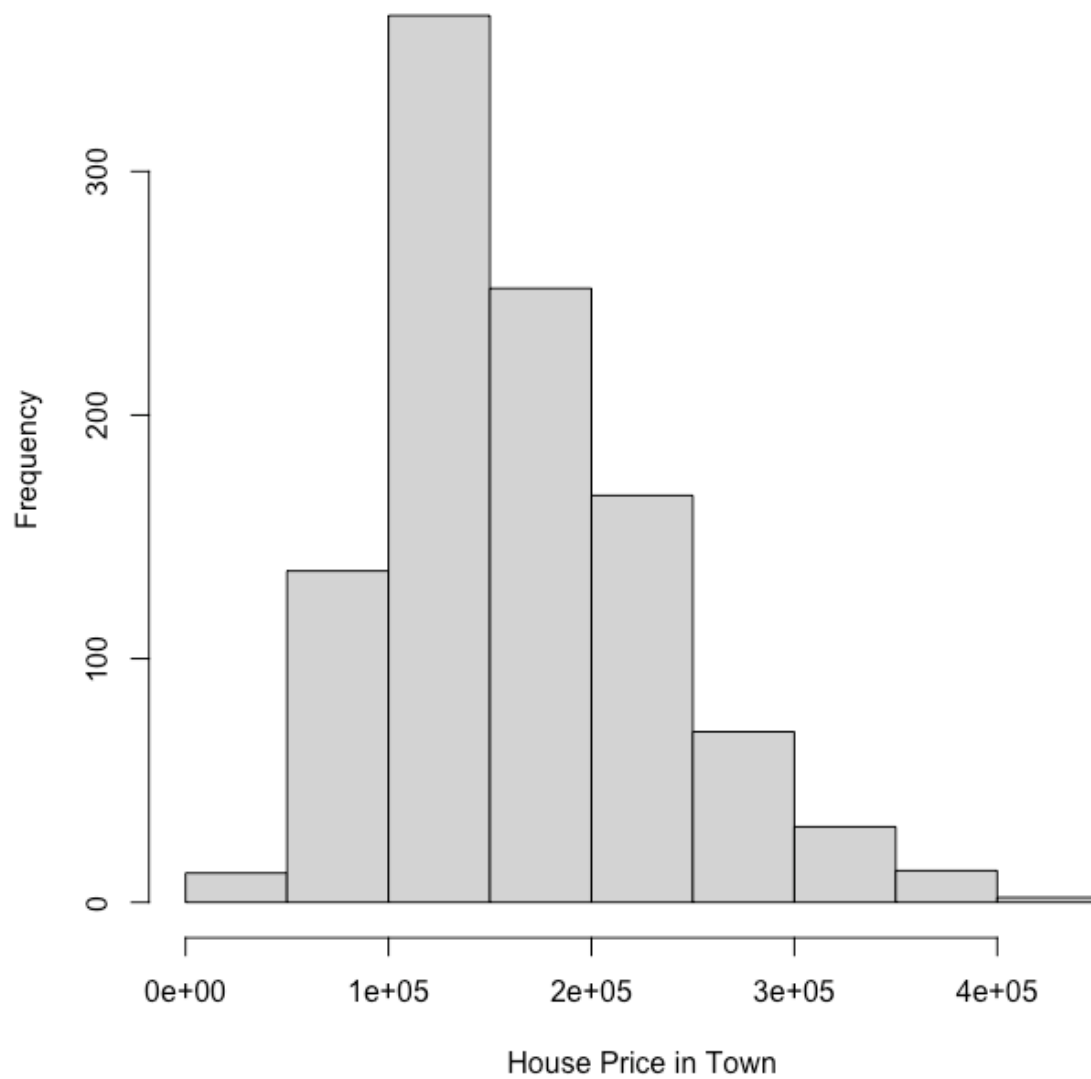
Maximum Value:446436

As Mean > Median, the graphical representation is expected to be Right Skewed with some outliers on the higher side pulling the graph to be skewed. Also we can determine that most of the houses sold are of lesser house value.

Histogram:

hist( loaddataset\$Price, main ='Hist of House Values in the Town', xlab= 'House Price in Town')

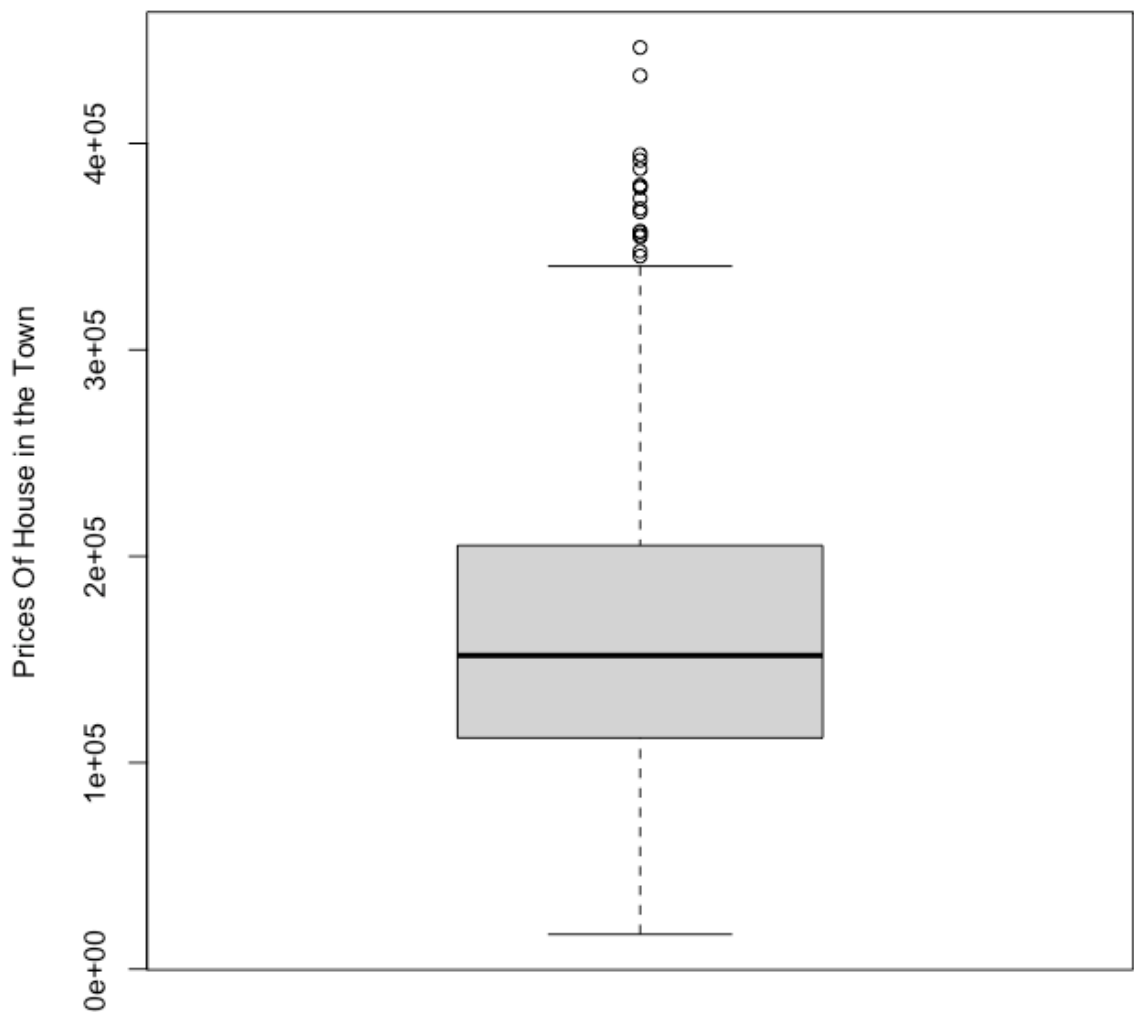
**Hist of House Values in the Town**



Boxplot:

```
boxplot(loaddataset$Price, ylab='Prices Of House in the Town', main ='Box Plot of House Values in the Town')
```

**Box Plot of House Values in the Town**



From Boxplot , we can determine the median house prices.

As we know there are outliers present in the data in the upper end of the boxplot, to find what all values contributing as outlier we can run below code:

```
IQR_Price <- IQR(loaddataset$Price)
Upper_outlier_price <- Q3_price + 1.5 * IQR_Price
Upper_outlier_price
subset(loaddataset$Price ,loaddataset$Price > Upper_outlier_price,)
```

```
31 summary(loaddataset$Price , usena = "Italy")
32 boxplot(loaddataset$Price , ylab='Prices Of House in the Town', main ='Box Plot of House Values
33 hist(loaddataset$Price,main ='Hist of House Values in the Town',xlab='House Price in Town')
34
35 IQR_Price <- IQR(loaddataset$Price)
36 Upper_outlier_price <- Q3_price + 1.5 * IQR_Price
37 subset(loaddataset$Price ,loaddataset$Price > Upper_outlier_price,)
38
39 #####
40
41
35:1 ## (Untitled) R Sci
```

Console Terminal Render Background Jobs

R 4.2.2 · ~/

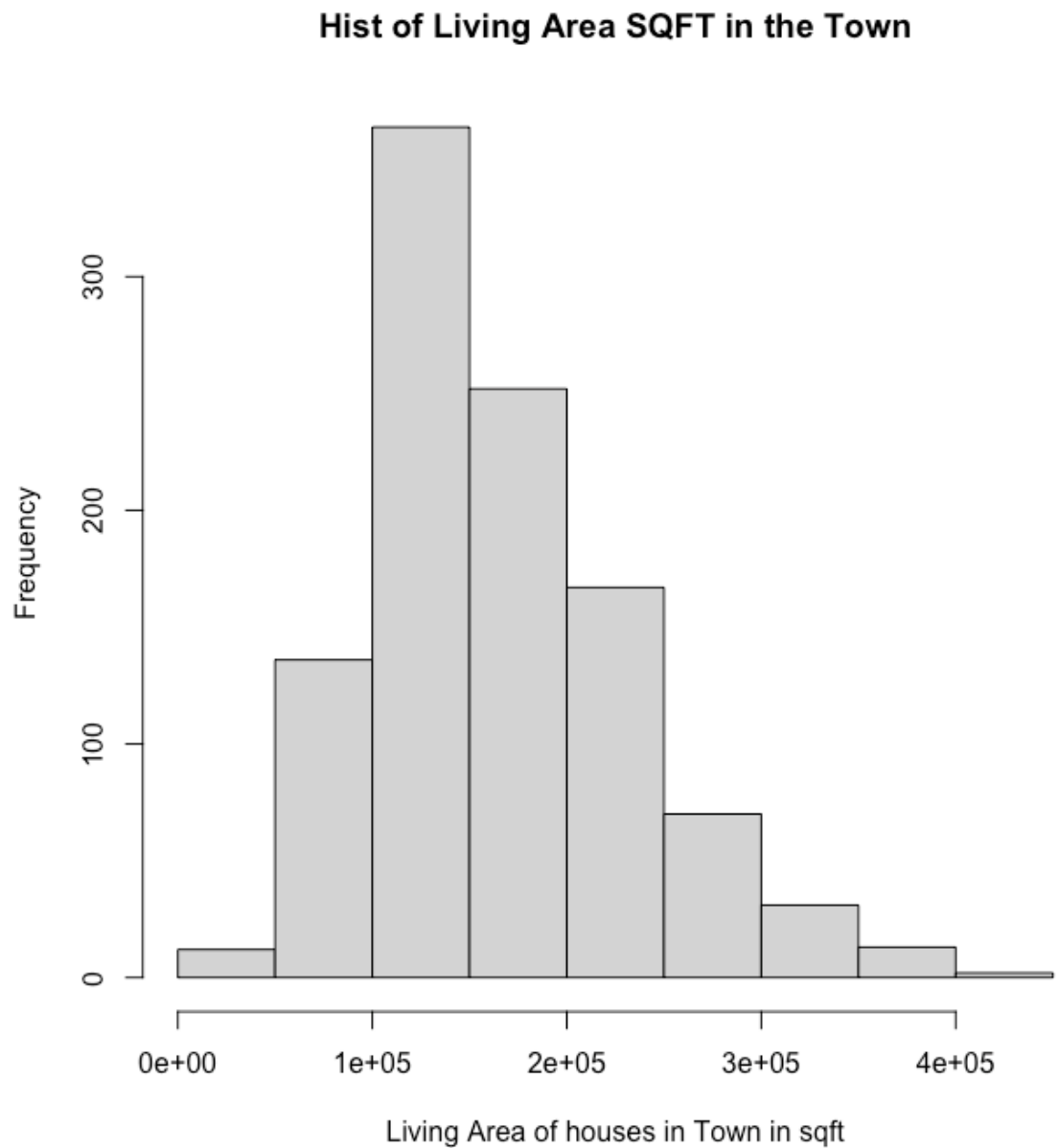
```
> IQR_Price <- IQR(loaddataset$Price)
> Upper_outlier_price <- Q3_price + 1.5 * IQR_Price
> subset(loaddataset$Price ,loaddataset$Price > Upper_outlier_price,)
[1] 387652 368396 379678 391842 355529 446436 378465 357384 373227 394532 345364 366772
[13] 357138 432845 354739 379472 347761
>
```

The above values constitute as outliers.

b.

Code for Histogram for the Living Area variable is:

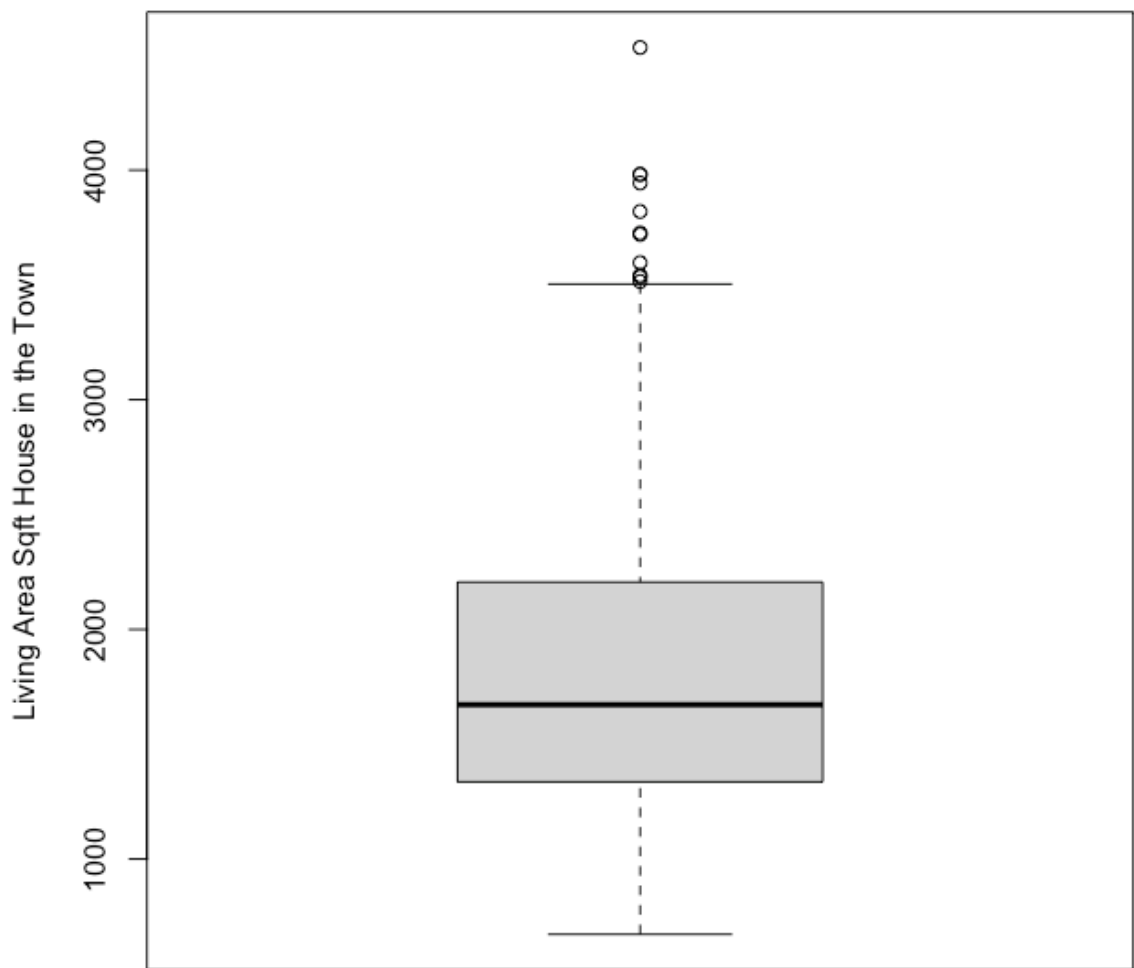
```
hist(loaddataset$Price,main ='Hist of Living Area SQFT in the Town',xlab='Living Area
of houses in Town in sqft')
```



Code for Boxplot for Living Area variable:

```
boxplot(loaddataset$Living.Area, ylab='Living Area Sqft House in the Town', main  
='Box Plot of Living Area of House in the Town in SQFT')
```

**Box Plot of Living Area of House in the Town in SQFT**



From Histogram we can deduce that this variable is Right Skewed and there must be outlier pulling data to be skewed, hence not normally distributed. Histogram also help us determine the frequency of each value/data of the variable, which is difficult to deduce from boxplot. So, from histogram we can analyse that majority sales are for houses that are smaller in sqft.

From Boxplot we can understand the values of outliers using the formula  $Q3 + 1.5 * IQR$ . From Boxplot we can determine that median is around 1650 sq. feet. Also, the interquartile range and the range of this variable will help us understand the distribution of data .