EDA of Amsterdam Housing Prices Dataset-Visualization Semester Project

Students Name: Aniket Santra Roll Number: MDS202106

30/11/2021

- First we import the dataset in R:-

```
#We are importing the dataset and storing it in a dataframe named
"housedataraw"
housedataraw=read.csv("D:/Visualization Project/HousingPrices-Amsterdam-
August-2021.csv",header=TRUE)
dim(housedataraw)

## [1] 924   8

head(housedataraw)

##   Serial.No.                                Address     Zip  Price Area
Room
## 1          1           Blasiusstraat 8 2, Amsterdam 1091 CR 685000   64
3
## 2          2 Kromme Leimuidenstraat 13 H, Amsterdam 1059 EL 475000   60
3
## 3          3              Zaaiersweg 11 A, Amsterdam 1097 SM 850000  109
4
## 4          4           Tenerifestraat 40, Amsterdam 1060 TH 580000  128
6
## 5          5              Winterjanpad 21, Amsterdam 1036 KN 720000  138
5
## 6          6          De Wittenkade 134 I, Amsterdam 1051 AM 450000   53
2
##        Lon      Lat
## 1 4.907736 52.35616
## 2 4.850476 52.34859
## 3 4.944774 52.34378
## 4 4.789928 52.34371
## 5 4.902503 52.41054
## 6 4.875024 52.38223

#We are changing the name of the first column to "Serial No."
colnames(housedataraw)[1]="Serial Number"
```

Now we have to clean our data. If there are some rows having element NA we shall omit those rows:-

```
#We are running na.omit command on the existing dataframe and storing the new
dataset in the dataframe named "housedata"
```

```
housedata=na.omit(housedataraw)
dim(housedata)

## [1] 920   8
```

- In the previous dataframe there were 924 rows now after cleaning in the new dataframe the no of rows is 920. Therefore the raw dataset had 4 rows having element NA which have been omitted. So in the new dataframe we define the first column “Serial No.” from 1 to 920.

```
housedata$`Serial Number`=1:920
head(housedata)

##   Serial Number                                Address     Zip  Price Area
Room
## 1             1           Blasiusstraat 8 2, Amsterdam 1091 CR 685000   64
3
## 2             2 Kromme Leimuidenstraat 13 H, Amsterdam 1059 EL 475000   60
3
## 3             3             Zaaiersweg 11 A, Amsterdam 1097 SM 850000  109
4
## 4             4           Tenerifestraat 40, Amsterdam 1060 TH 580000  128
6
## 5             5             Winterjanpad 21, Amsterdam 1036 KN 720000  138
5
## 6             6         De Wittenkade 134 I, Amsterdam 1051 AM 450000   53
2
##        Lon      Lat
## 1 4.907736 52.35616
## 2 4.850476 52.34859
## 3 4.944774 52.34378
## 4 4.789928 52.34371
## 5 4.902503 52.41054
## 6 4.875024 52.38223
```

Summary and Introduction of the project:-

Amsterdam is the capital and most populous city of Netherlands. It is a urban city known for its artistic heritage, narrow houses with gabled facades, legacies of the city’s 17th-century Golden Age. We have the dataset of Amsterdam Housing Prices. In this dataset we have unique address, Zip code, Area, Price of 920 houses of Amsterdam. We also have no. of rooms, the latitude and longitude of the locations of each houses. In this project it has been attempted to study and analyse the several component of the dataset. We did EDA or exploratory data analysis of this dataset by several statistical graph. We have analysed the each components of the datset and found some statistical characteristics of the components. We have also checked any dependency in between the component by bivariate and multivariate graphical analysis.

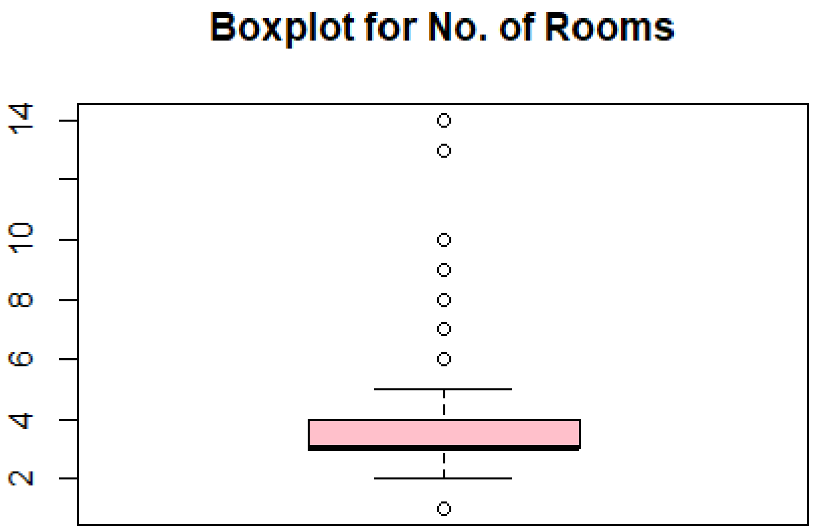
Desciption Of The Variables:-

- **Serial No. -** The variable Serial No. consists of integers starting from 1 which are to mark the all houses serially uniquely. This is a categorical variable of nominal type.
- **Address -** The address variable stores the unique addresses of each of the houses and it is also a categorical variable of nominal type.
- **Zip -** This variable stores the Zip codes of the respective houses' location. The type of the variable Zip is categorical of nominal type.
- **Price -** This is the price variable stores the prices of each of the houses in Euros. The Price variable is a continuous variable.
- **Area –** This variable stores the areas of each of the houses in the unit square metre. The Area variable is a continuous variable.
- **Room -** The room variable is a discrete variable stores the no of rooms in the each respective houses.
- **Lon -** This variable stores the numeric part of longitude of the location of the each houses and it's a continuous variable.
- **Lat -** Like Lon variable it's also a continuous variable stores the numeric part of latitude of the location of each houses.

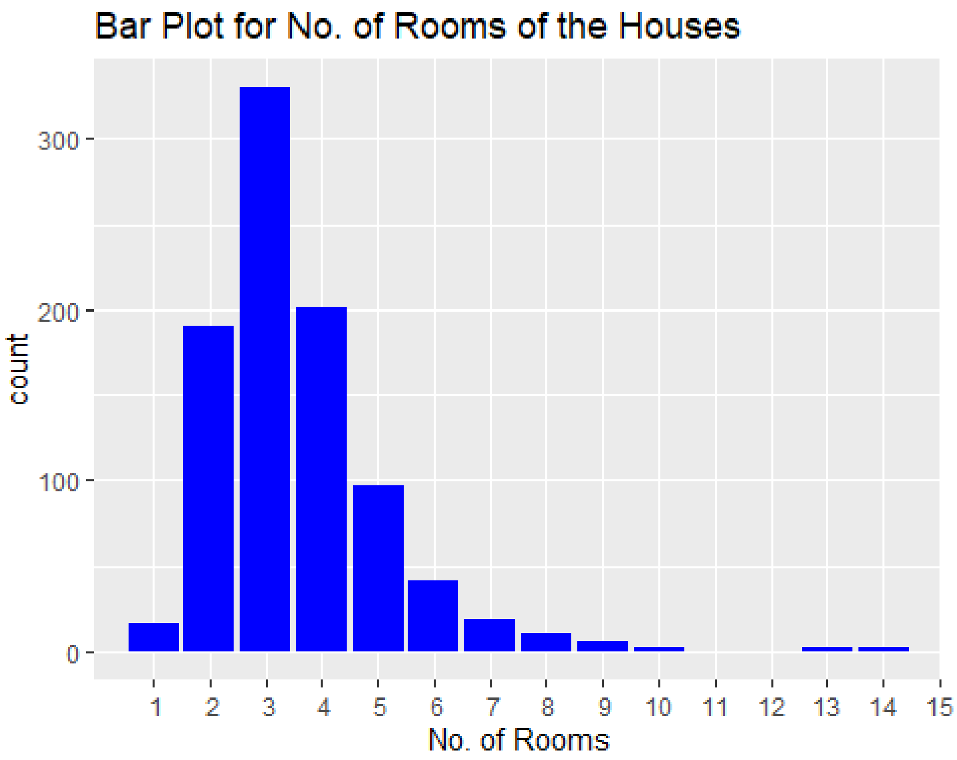
Univariate Graphical Analysis:-

- We consider the discrete and continuous variables for the graphical analysis-
- First we consider the Room variable which is discrete and stores no of rooms in each of the houses:-

```
summary(housedata$Room)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   1.000   3.000   3.000   3.564   4.000  14.000

boxplot(housedata$Room,col="pink",main="Boxplot for No. of Rooms")
library(ggplot2)
```



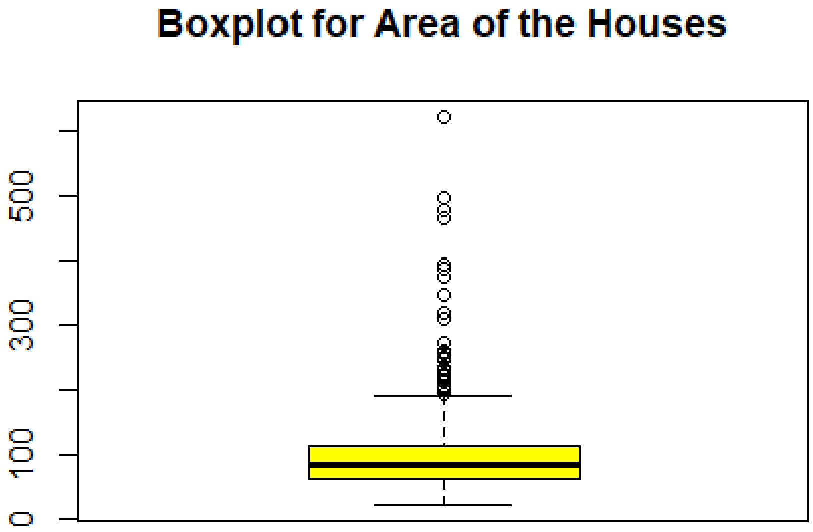
```
ggplot(housedata)+geom_bar(aes(x=Room),
fill="blue")+scale_x_discrete(limits=1:15)+labs(title="Bar Plot for No. of
Rooms of the Houses")+xlab("No. of Rooms")

## Warning: Continuous limits supplied to discrete scale.
## Did you mean `limits = factor(...)` or `scale_*_continuous()`?
```



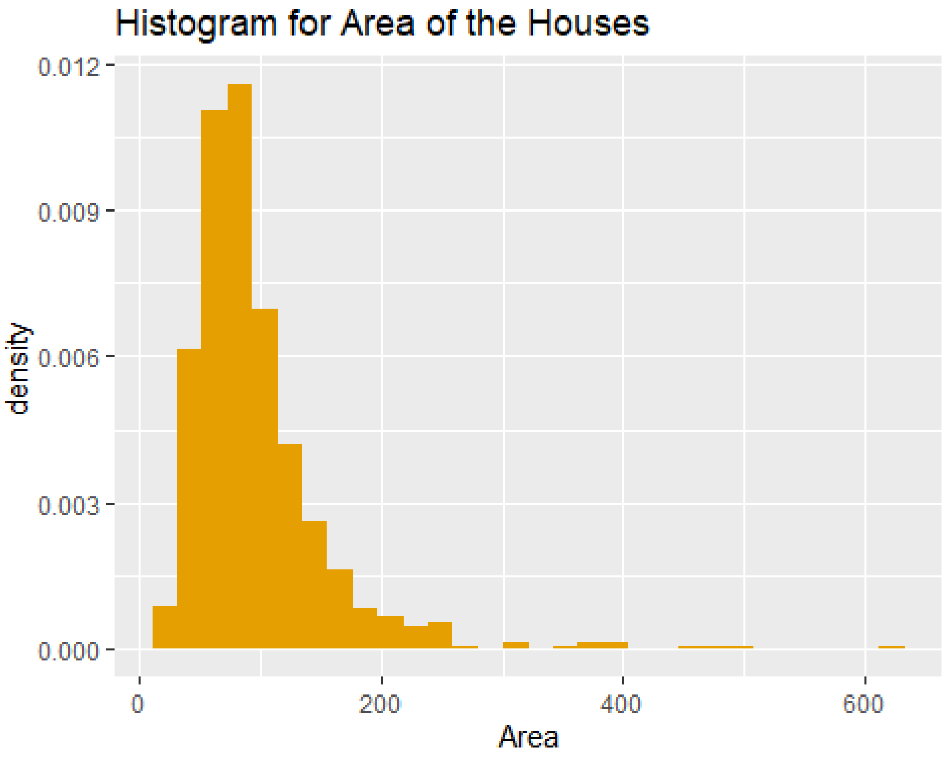
- From the summary we can see the median and mean of the Room is 3, 3.564 respectively and in the bar plot we can see that the house having 3 rooms has the highest frequency. So, we can say that that the average rooms in the houses of Amsterdam is 3. Both the boxplot and histogram show that there are so many outliers example there are two houses having no. of rooms 13, 14. We can ignore these outliers.

- Now consider the Area variable which stores the area of the houses in square meter:-

```
summary(housedata$Area)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   21.00   60.00   83.00   95.61  113.00  623.00

boxplot(housedata$Area, col="yellow", main="Boxplot for Area of the Houses")
```



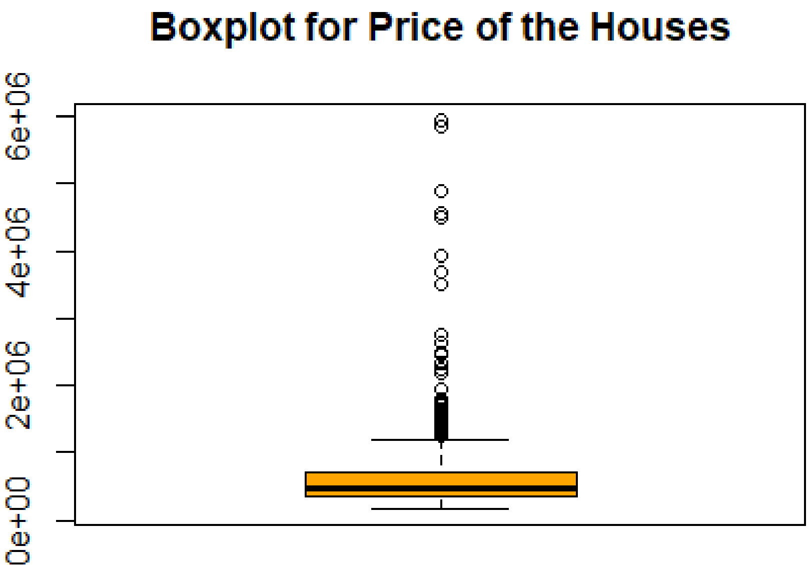
```
library(ggplot2)
ggplot(housedata)+geom_histogram(aes(x=Area, y=..density..),
fill="#E69F00")+labs(title="Histogram for Area of the Houses")+ xlab("Area")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



- From the summary the mean of the Area is 95.61 sq. metres and median is 83 sq. metres. From the histogram we can see that most no of houses having area about 70 to 100 square metres. Also boxplot and histogram show that so many outliers are there in the dataset of area which we can ignore. Hence most of the houses in the Amsterdam have area approximately about 80-95 square metres.The histogram shows that the distribution of the areas of the houses is positively skewed.

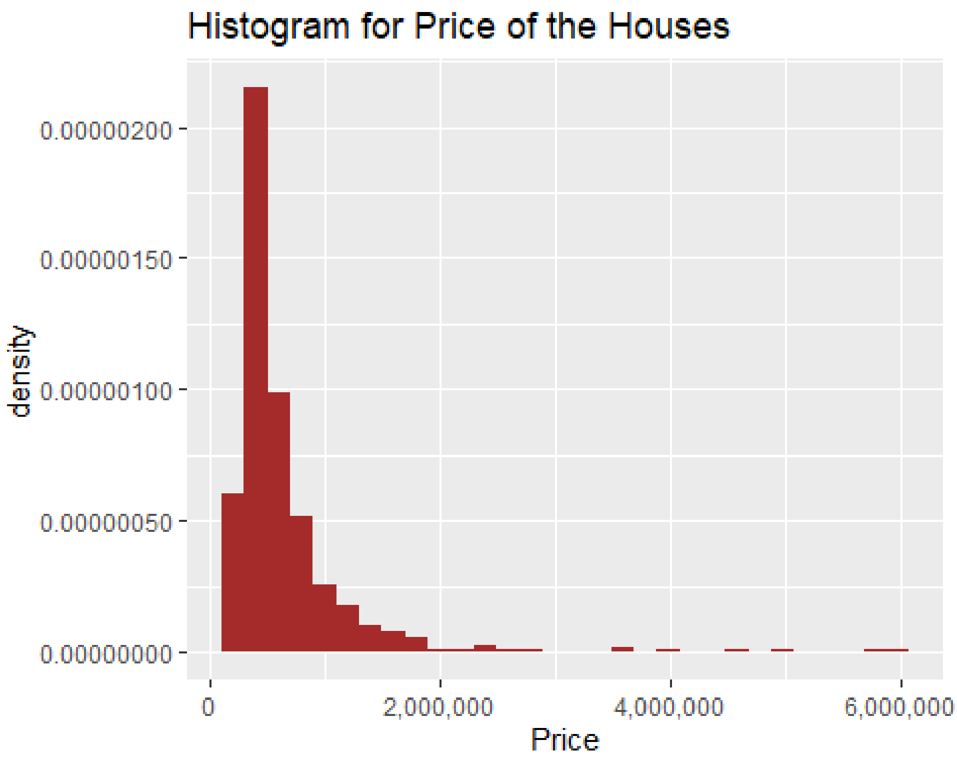
- Consider the Price variable stores the prices of each of the houses in Euros:-

```
summary(housedata$Price)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  175000  350000  467000  622065  700000 5950000

library(scales)
boxplot(housedata$Price, col="orange", main="Boxplot for Price of the
Houses")
```



```
library(ggplot2)
ggplot(housedata)+geom_histogram(aes(x=Price, y=..density..),
fill="#A52A2A")+labs(title="Histogram for Price of the Houses")+
xlab("Price")+scale_x_continuous(labels=comma)+scale_y_continuous(labels=comm
a)
```

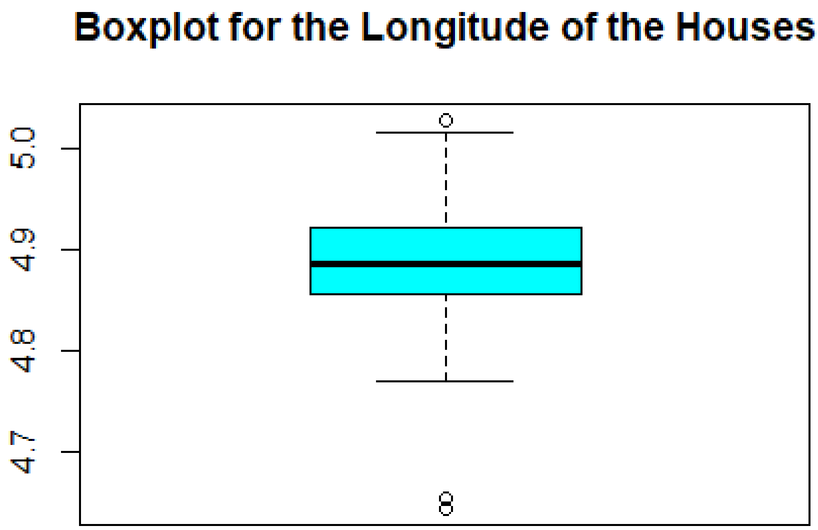
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



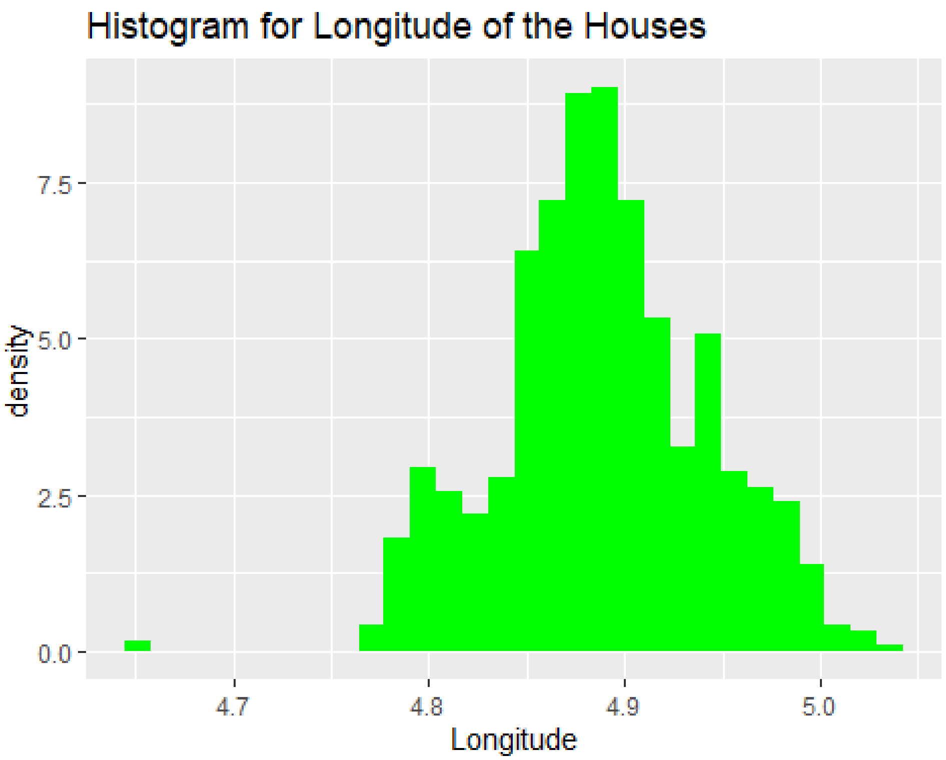
- From the summary mean of the Price is 622065 Euros and the median is 467000 Euros. The histogram shows that the price of most of the house in Amsterdam in between 500000 Euros to 700000 Euros.The histogram shows that the distribution of the prices of the houses is positively skewed.

- Now consider the Lon and Lat variables which stores the longitude and latitude respectively of the locations of each of the houses:-

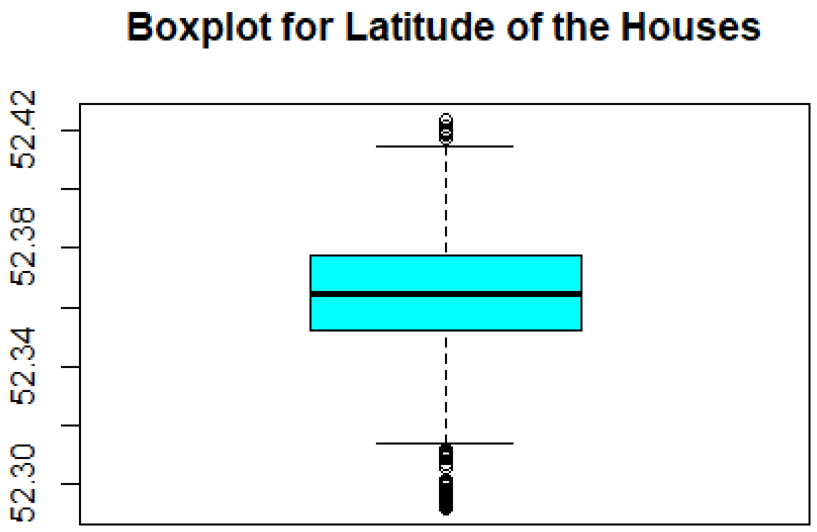
```
summary(housedata$Lon)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.645   4.856   4.887   4.889   4.922   5.029

boxplot(housedata$Lon, col="#00FFFF", main="Boxplot for the Longitude of the
Houses")
```



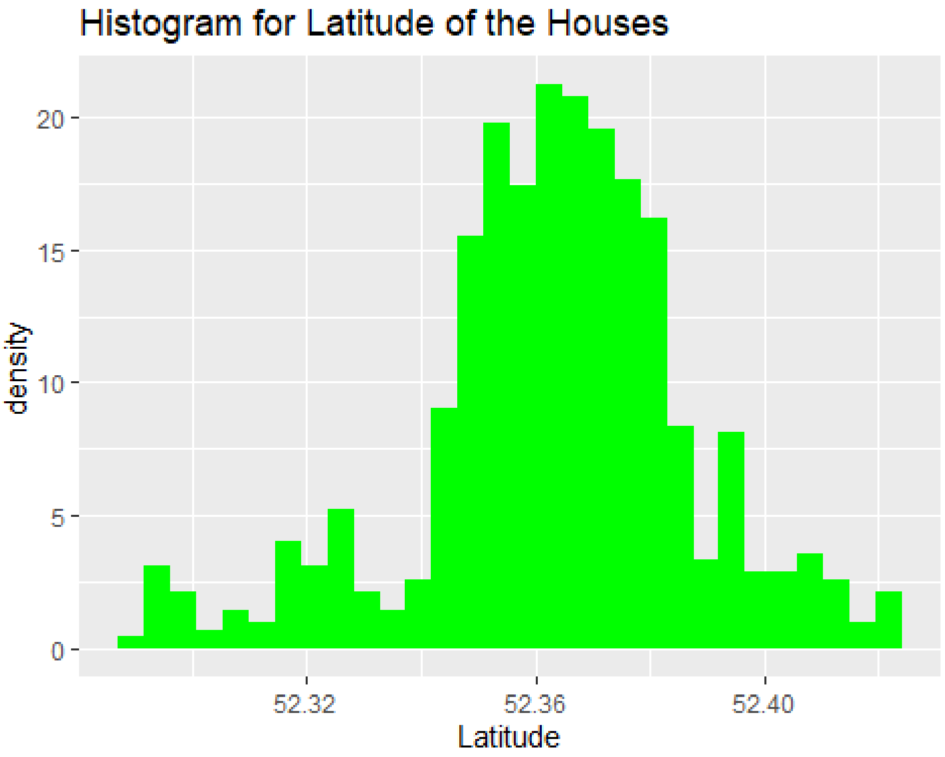
```
library(ggplot2)
ggplot(housedata)+geom_histogram(aes(x=Lon, y=..density..),
fill="green")+labs(title="Histogram for Longitude of the Houses")+
xlab("Longitude")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
#Now consider the Lat variable
summary(housedata$Lat)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   52.29   52.35   52.36   52.36   52.38   52.42

boxplot(housedata$Lat, col="#00FFFF", main="Boxplot for Latitude of the
Houses")
```



```
ggplot(housedata)+geom_histogram(aes(x=Lat, y=..density..),
fill="green")+labs(title="Histogram for Latitude of the Houses")+
xlab("Latitude")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

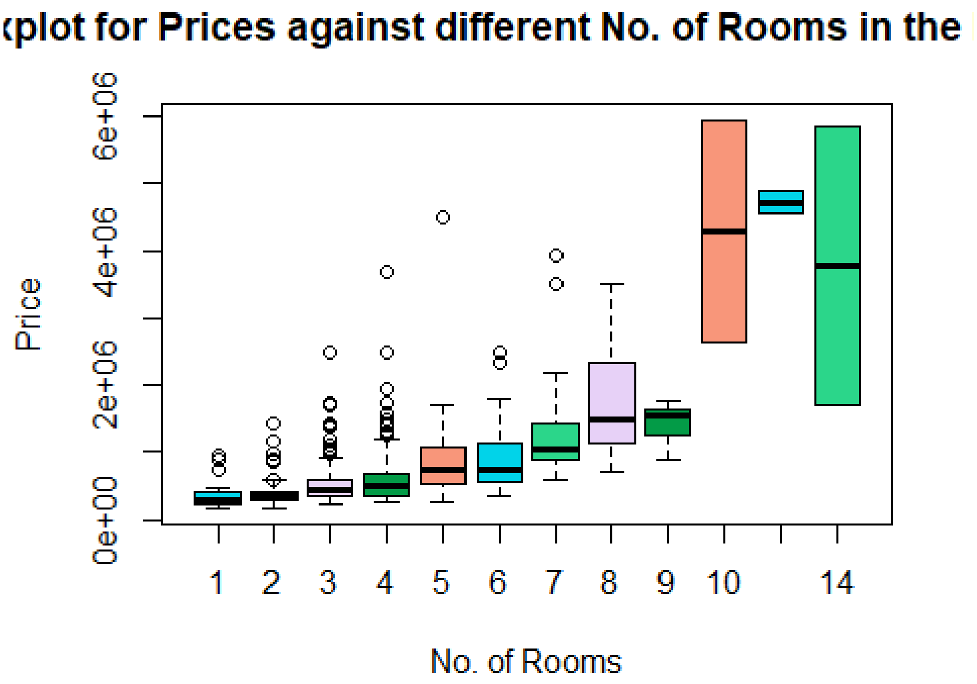


- From google search we know the longitude of Amsterdam 4.9041°E and latitude of Amsterdam 52.3676°N. From summary of Longitude we found mean is 4.889 1st quartile 4.856 and 3rd quartile 4.922 and for Latitude mean is 52.36 1st quartile 52.35 3rd quartile 52.38. From the boxplots we can see there is only one outliers for longitude and some outliers for latitude which we can ignore. From the histogram of longitude we can say the distribution of longitude almost symmetric and from the histogram of latitude the distribution of latitude also almost symmetric ignoring the outliers. So we can say the dataset covers the information of houses from whole the Amsterdam city.

Bivariate and Multivariate Plots

- First we draw side by side boxplot of house prices against different no. of rooms:-

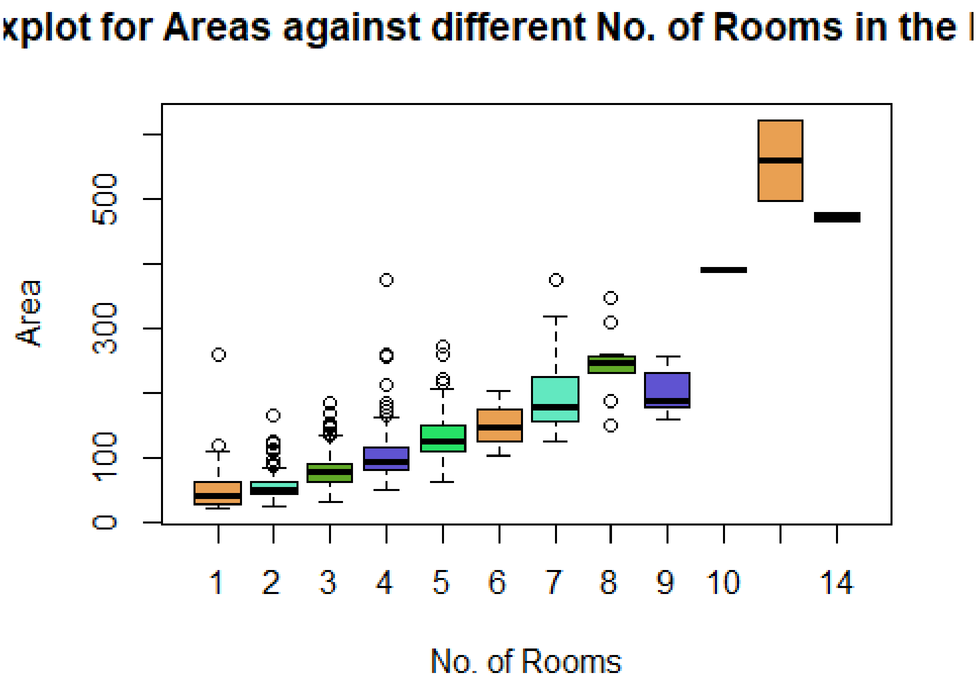
```
boxplot(housedata$Price~housedata$Room,col=rgb(runif(5),runif(5),runif(5)),
main="Boxplot for Prices against different No. of Rooms in the Houses",
xlab="No. of Rooms", ylab="Price")
```



- Previously we saw in bar plot of Room that most of the houses in Amsterdam have no of rooms 2 to 4. So we visualize only the boxplots for those as for other no of rooms we have less data. We can see that for no of rooms 2,3,4 each case there are some outliers, from this we can conclude that there are many luxurious houses for each of the three cases which are expensive.

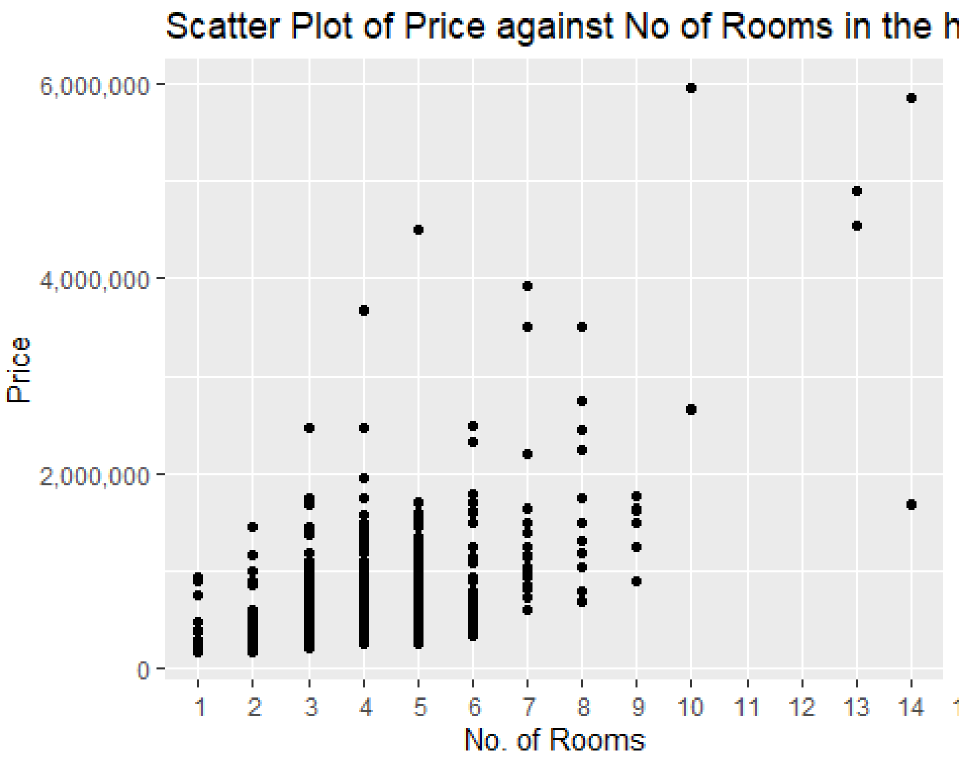
- Now we draw side by side boxplot of house areas against different no of rooms:-

```
boxplot(housedata$Area~housedata$Room,col=rgb(runif(5),runif(5),runif(5)),main="Boxplot for Areas against different No. of Rooms in the houses", xlab="No. of Rooms", ylab="Area")
```

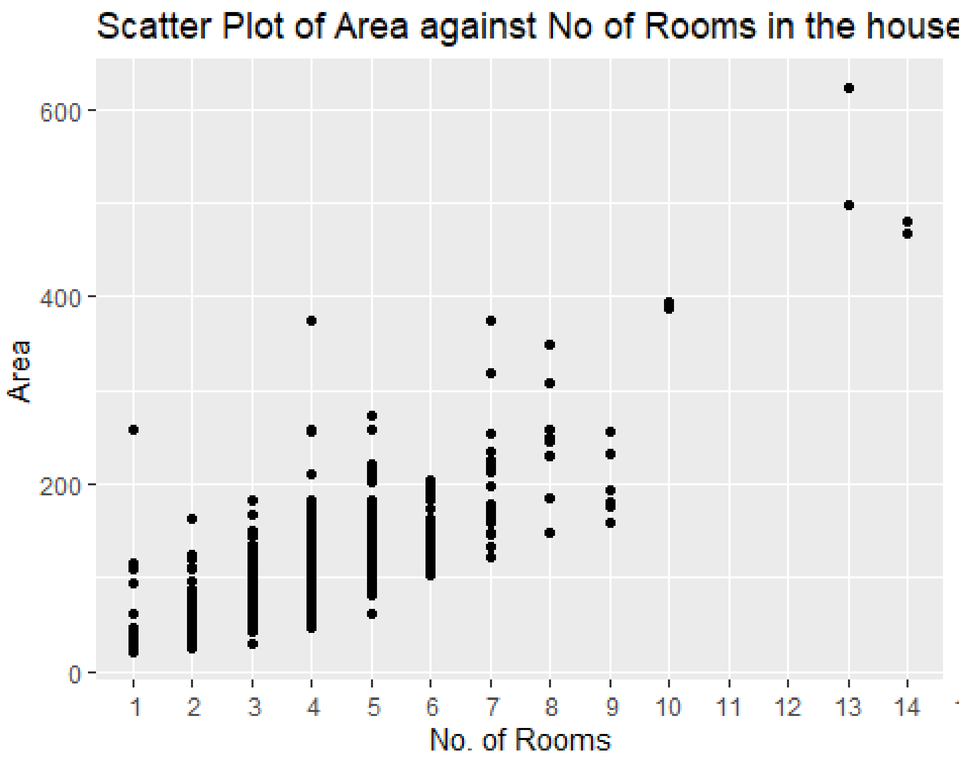


- Previously we saw in bar plot of Room that most of the houses in Amsterdam have no of rooms 2 to 4. So we will visualize only the boxplots for those as for other no of rooms we have less data. We can see that for no of rooms 3,4 the distribution of area is almost symmetric if we ignore the outliers and in each of the cases for no of rooms in the house 2,3,4 there are many outliers which give us the idea that in each case there are many luxurious houses having less no of rooms but large in sizes.
- We draw Scatter Plots of house price and house area against the no. of rooms in the house:-

```
library(ggplot2)
library(scales)
ggplot(housedata)+geom_point(aes(x=Room,y=Price))+scale_x_discrete(limits=1:1
5)+labs(title="Scatter Plot of Price against No of Rooms in the
houses")+xlab("No. of Rooms")+ylab("Price")+scale_y_continuous(labels=comma)

## Warning: Continuous limits supplied to discrete scale.
## Did you mean `limits = factor(...)` or `scale_*_continuous()`?
```



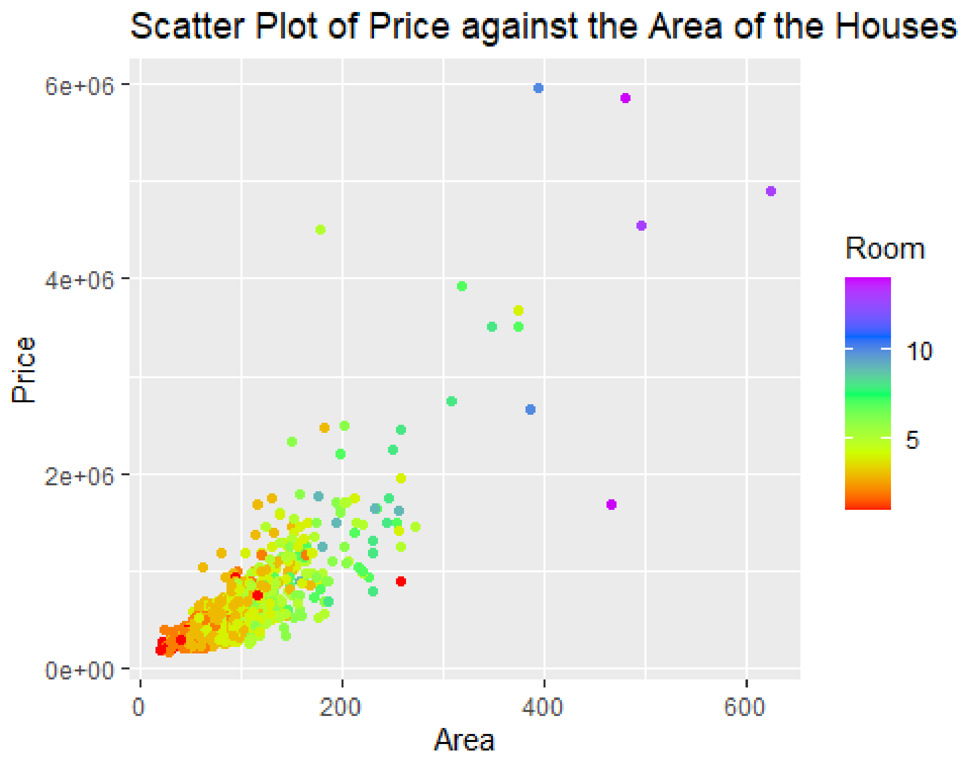
```
ggplot(housedata)+geom_point(aes(x=Room,y=Area))+scale_x_discrete(limits=1:15
)+labs(title="Scatter Plot of Area against No of Rooms in the houses")+
xlab("No. of Rooms")+ylab("Area")

## Warning: Continuous limits supplied to discrete scale.
## Did you mean `limits = factor(...)` or `scale_*_continuous()`?
```



- We have draw the scatter plot of continuous variable Price and Area against the discrete variable Room (no. of rooms). From these scatter plot we can see that in average the price and area of the houses increases as no of rooms increases i.e. there is positive association in both cases but there are some exceptional that no of rooms is less but price and area is high which is because there are some luxurious expensive, big houses in each cases of no of rooms which we can determine as outliers.

- We draw a Scatter Plot of Price against the Area of the houses with distinct no. of rooms:-

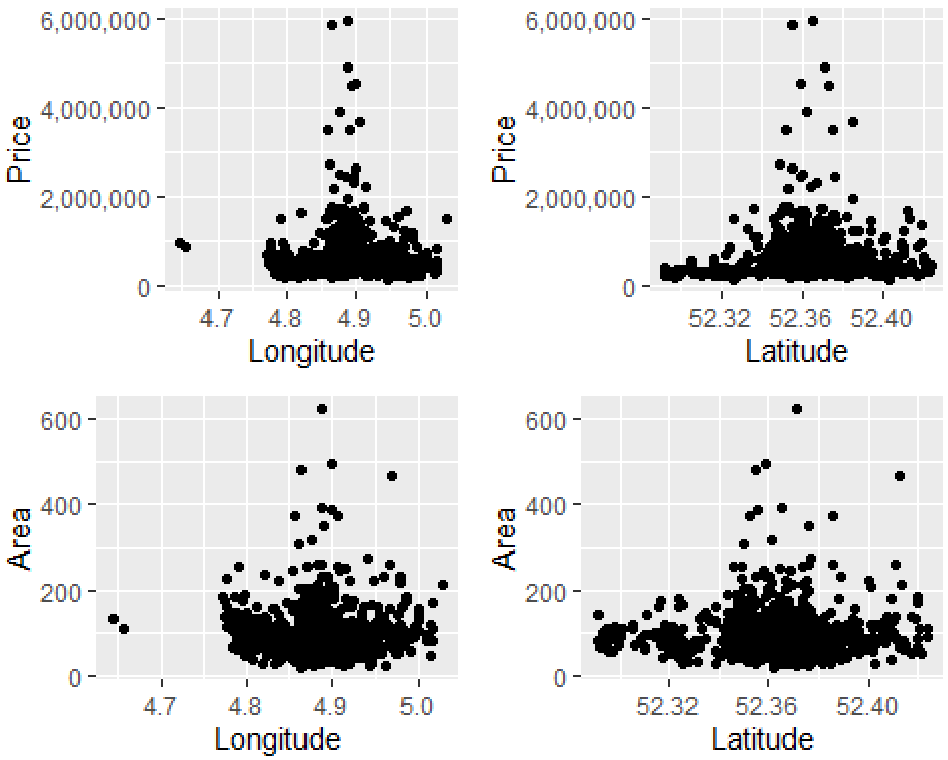
```
ggplot(housedata,)+geom_point(aes(x=Area,y=Price,
col=Room))+scale_color_gradientn(colours=rainbow(5))+labs(title="Scatter Plot
of Price against the Area of the Houses with distinct No. of
Rooms")+xlab("Area")+ylab("Price")
```



- From the scatter plot of Price against area we can see that there is a positive correlation between them i.e. price of house increases as area increases and the points are distinguished by color with the no of rooms the houses have. It helps us to visualize how the no. of rooms are distributed with price and area.

-Now we draw scatter plot of Price and Area against the Longitude and Latitude:-

```
library(ggplot2)
library(gridExtra)
p1=ggplot(housedata)+geom_point(aes(x=Lon,y=Price))+xlab("Longitude")+ylab("Price")+scale_y_continuous(labels=comma)
p2=ggplot(housedata)+geom_point(aes(x=Lat,y=Price))+xlab("Latitude")+ylab("Price")+scale_y_continuous(labels=comma)
p3=ggplot(housedata)+geom_point(aes(x=Lon,y=Area))+xlab("Longitude")+ylab("Area")
p4=ggplot(housedata)+geom_point(aes(x=Lat,y=Area))+xlab("Latitude")+ylab("Area")
grid.arrange(p1,p2,p3,p4)
```



- From the scatter plot of Price and Area against latitude and longitude we can see that almost the houses having areas less than 90 square metres and prices less than 700k euros are equally distributed in the Amsterdam city. Some houses in the middle part of the Amsterdam city are having much price and area which is familiar with a city that every city has some luxurious houses in the central part of the city.

Conclusion:-

- We visualize the whole house price dataset of Amsterdam. We analyse each and every component of the dataset with statistical graphical analysis. From the analysis we found many interesting fact about the dataset- The houses in Amsterdam have on an average 2-4 no. of rooms. There are some luxurious house having rooms greater than 10 which work as outlier in the dataset. We saw in bar plot there are two house having no of rooms 13 and 14 which are pretty much expensive.
- We also found that the average price of house in Amsterdam in between 500k Euros to 700k Euros and in case of area the average area of the houses in Amsterdam is about 80-95 square metres. Both in Price and Area there are some outliers also which influence the statistical measures.
- From the distribution of longitude and latitude visualizing the respective histograms we found the house dataset covers almost the whole Amsterdam city i.e. the house data are collected from the all around the city not a specific area.
- Now come to the bivariate and multivariate analysis part we found some obvious facts that as no of rooms in house increases price and area also increases but there are some exceptional that no of rooms in some houses are less but price and area are high. This is may be for the house position, the rooms of the house is large and

for some another quality characteristics of those house. These houses we can consider as outliers.

- From the scatter plot of price and area against longitude and latitude we found that the houses having price less than 700k euros and area less than 90 square metres are equally distributed in the city. There is not the case that a specific area of the has houses of high price and area and specific area of the city has houses of low price and area.The houses having excessively large no. of rooms, large area and of higher price are situated in the central part of the Amsterdam city.

These are some findings we got from the dataset.

At the end we can conclude that if we ignore and omitted the outliers in the dataset which influence several statistical measures, the dataset would be ideal to predict the price of the house of Amsterdam depends on no of rooms, area and location and also could do some statistical hypothesis test.