Assignement 1

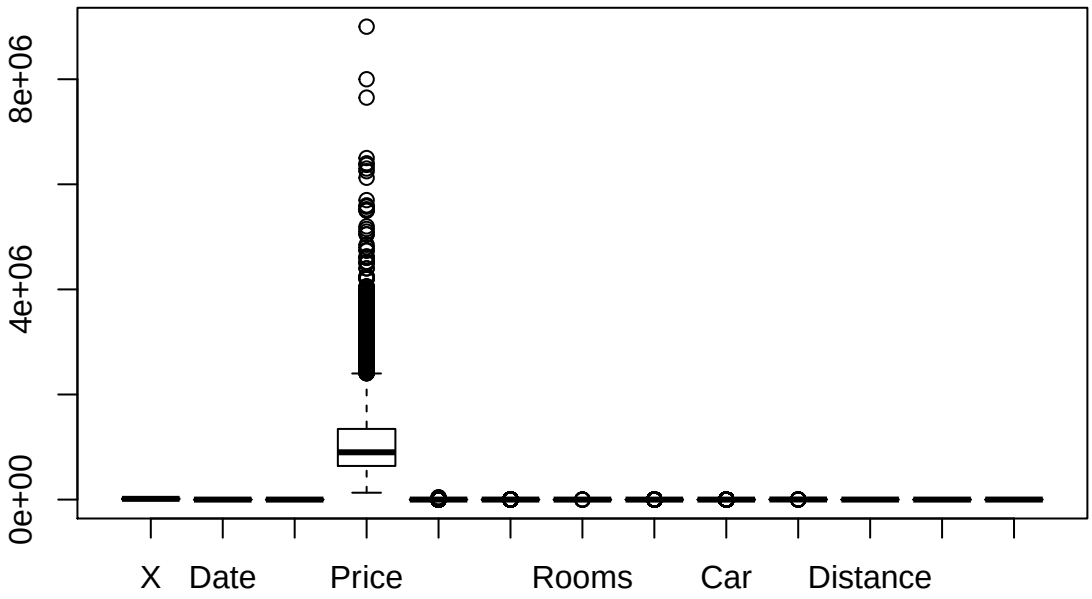
Louve Le Bronec

R Markdown

```
library(tidyr)

# read the csv file
housing.dataset <- read.csv("melbourne_data.csv")

# See the structure of the data
str(housing.dataset)
```

```
## 'data.frame':    34857 obs. of  13 variables:
##  $ X            : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Date         : Factor w/ 78 levels "1/07/2017","10/02/2018",..: 59 61 64 64 65 65 66 70 70 70 ...
##  $ Type         : Factor w/ 3 levels "h","t","u": 1 1 1 3 1 1 1 1 1 1 ...
##  $ Price        : int  NA 1480000 1035000 NA 1465000 850000 1600000 NA NA NA ...
##  $ Landsize     : int  126 202 156 0 134 94 120 400 201 202 ...
##  $ BuildingArea : num  NA NA 79 NA 150 NA 142 220 NA NA ...
##  $ Rooms        : int  2 2 2 3 3 3 4 4 2 2 ...
##  $ Bathroom     : int  1 1 1 2 2 2 1 2 1 2 ...
##  $ Car          : int  1 1 0 1 0 1 2 2 2 1 ...
##  $ YearBuilt    : int  NA NA 1900 NA 1900 NA 2014 2006 1900 1900 ...
##  $ Distance     : Factor w/ 216 levels "#N/A","0","0.7",..: 82 82 82 82 82 82 82 82 82 82 ...
##  $ Regionname   : Factor w/ 9 levels "#N/A","Eastern Metropolitan",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ Propertycount: Factor w/ 343 levels "#N/A","1008",..: 191 191 191 191 191 191 191 191 191 191 ...
```

```
# 3. Task

# Question 1

# remove the na values
housing.dataset <- housing.dataset %>% drop_na()

# remove the outliers and incorrect values

#Display boxplot to see the outliers
boxplot(housing.dataset)
```

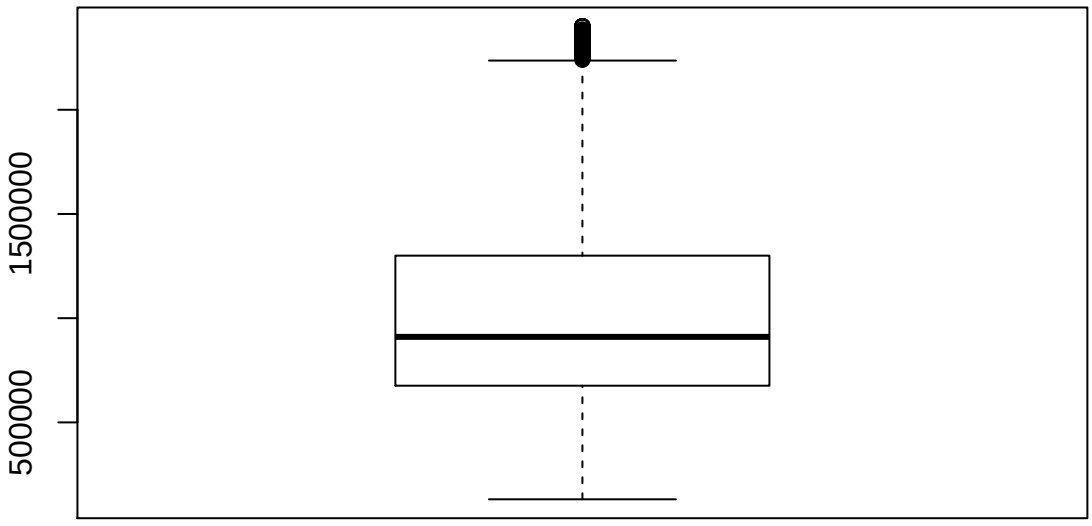



```
#Remove the outliers in function of the quantiles

#Price
outliers <- boxplot(housing.dataset$Price, plot=FALSE)$out
x<-housing.dataset
housing.dataset<- x[-which(x$Price %in% outliers),]

#Landsize
outliers <- boxplot(housing.dataset$Landsize, plot=FALSE)$out
x<-housing.dataset
housing.dataset<- x[-which(x$Landsize %in% outliers),]

#BuildingArea
outliers <- boxplot(housing.dataset$BuildingArea, plot=FALSE)$out
x<-housing.dataset
housing.dataset<- x[-which(x$BuildingArea %in% outliers),]

#Rooms
outliers <- boxplot(housing.dataset$Rooms, plot=FALSE)$out
x<-housing.dataset
housing.dataset<- x[-which(x$Rooms %in% outliers),]

#Bathrooms
outliers <- boxplot(housing.dataset$Bathroom, plot=FALSE)$out
x<-housing.dataset
housing.dataset<- x[-which(x$Bathroom %in% outliers),]

#Car
outliers <- boxplot(housing.dataset$Car, plot=FALSE)$out
x<-housing.dataset
housing.dataset<- x[-which(x$Car %in% outliers),]

#Removing incorrect values
housing.dataset <- housing.dataset[!(housing.dataset$Landsize == 0 | housing.dataset$BuildingArea == 0
```

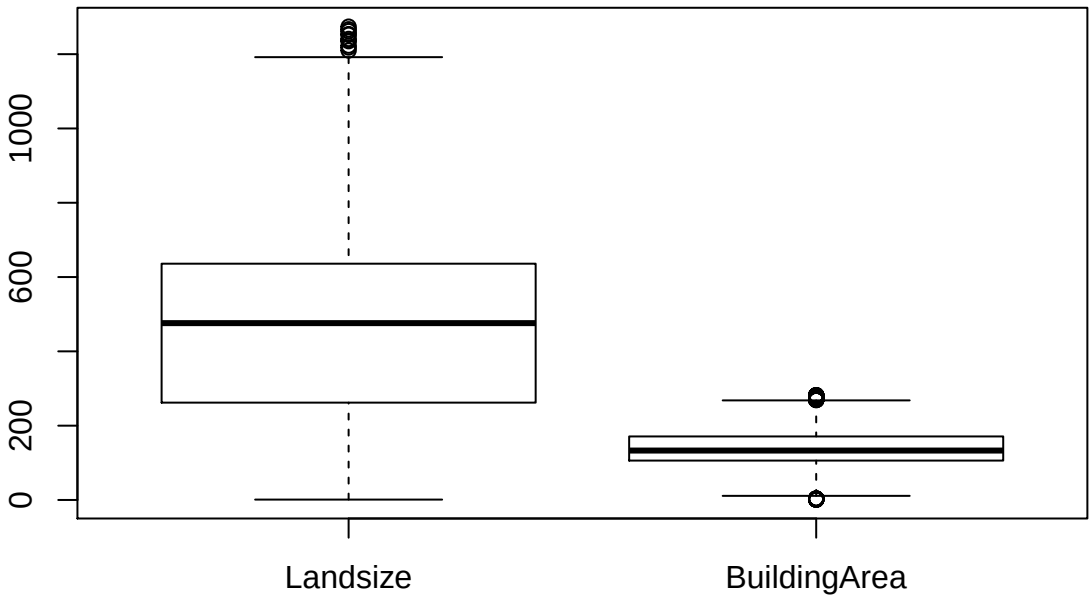
```r
#Summary and boxplot to check if it worked
# We can see that the outliers and incorrect values are not here anymore
summary(housing.dataset)
```

```
##        X                Date      Type         Price             Landsize
##  Min.   :    3   24/02/2018: 179   h:5350   Min.   : 131000   Min.   :   1.0
##  1st Qu.: 7202   17/03/2018: 171   t: 612   1st Qu.: 676000   1st Qu.: 262.0
##  Median :14838   27/05/2017: 170   u: 560   Median : 910250   Median : 476.0
##  Mean   :15726   3/03/2018 : 166            Mean   :1024266   Mean   : 465.6
##  3rd Qu.:23672   28/10/2017: 155            3rd Qu.:1300000   3rd Qu.: 636.0
##  Max.   :34857   3/06/2017 : 151            Max.   :2400000   Max.   :1274.0
##                  (Other)   :5530
##   BuildingArea       Rooms          Bathroom          Car          YearBuilt
##  Min.   :  1.0   Min.   :1.000   Min.   :1.000   Min.   :0.000   Min.   :1850
##  1st Qu.:106.0   1st Qu.:3.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1940
##  Median :133.0   Median :3.000   Median :2.000   Median :2.000   Median :1970
##  Mean   :140.9   Mean   :3.124   Mean   :1.587   Mean   :1.564   Mean   :1964
##  3rd Qu.:171.0   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:1996
##  Max.   :282.0   Max.   :6.000   Max.   :3.000   Max.   :3.000   Max.   :2018
##
##     Distance                          Regionname   Propertycount
##  11.2   : 286   Northern Metropolitan     :2016   21650  : 163
##  14.7   : 141   Western Metropolitan      :1666   8870   : 154
##  7.8    : 139   Southern Metropolitan     :1612   10969  : 121
##  13.9   : 136   Eastern Metropolitan      : 797   11918  : 116
##  5.2    : 136   South-Eastern Metropolitan: 309   11204  : 110
##  9.2    : 125   Northern Victoria         :  49   14577  : 106
##  (Other):5559   (Other)                   :  73   (Other):5752
```

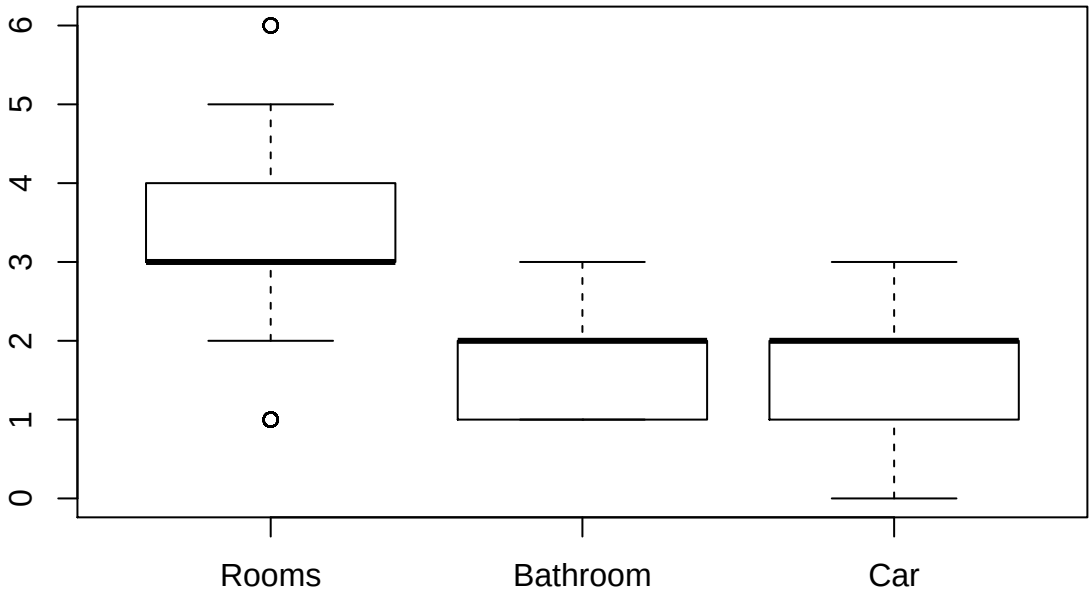
```r
boxplot(housing.dataset$Price,  main="Boxplot for the price")
```



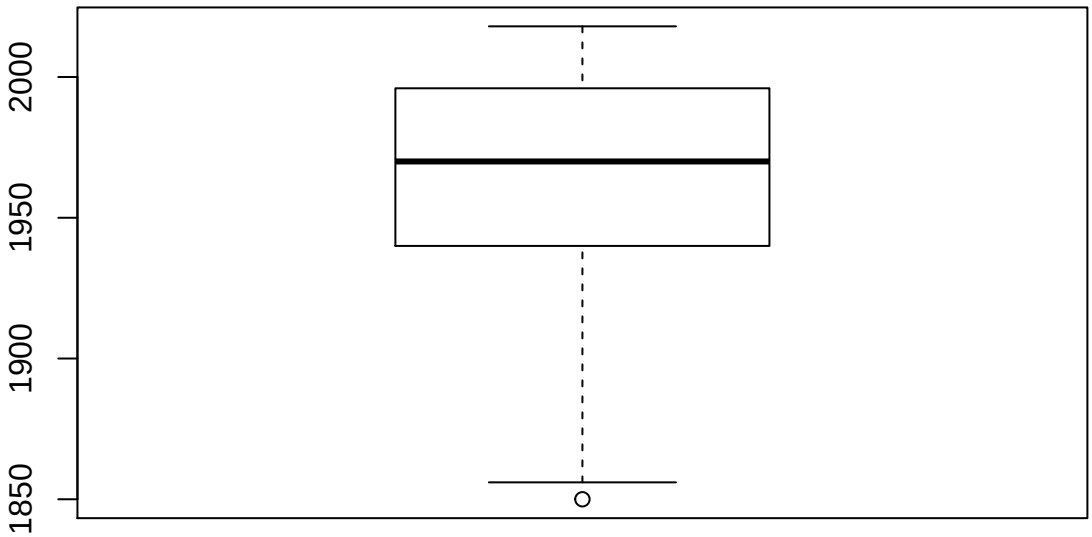

```r
boxplot(housing.dataset[5:6])
```




```
boxplot(housing.dataset[7:9])
```



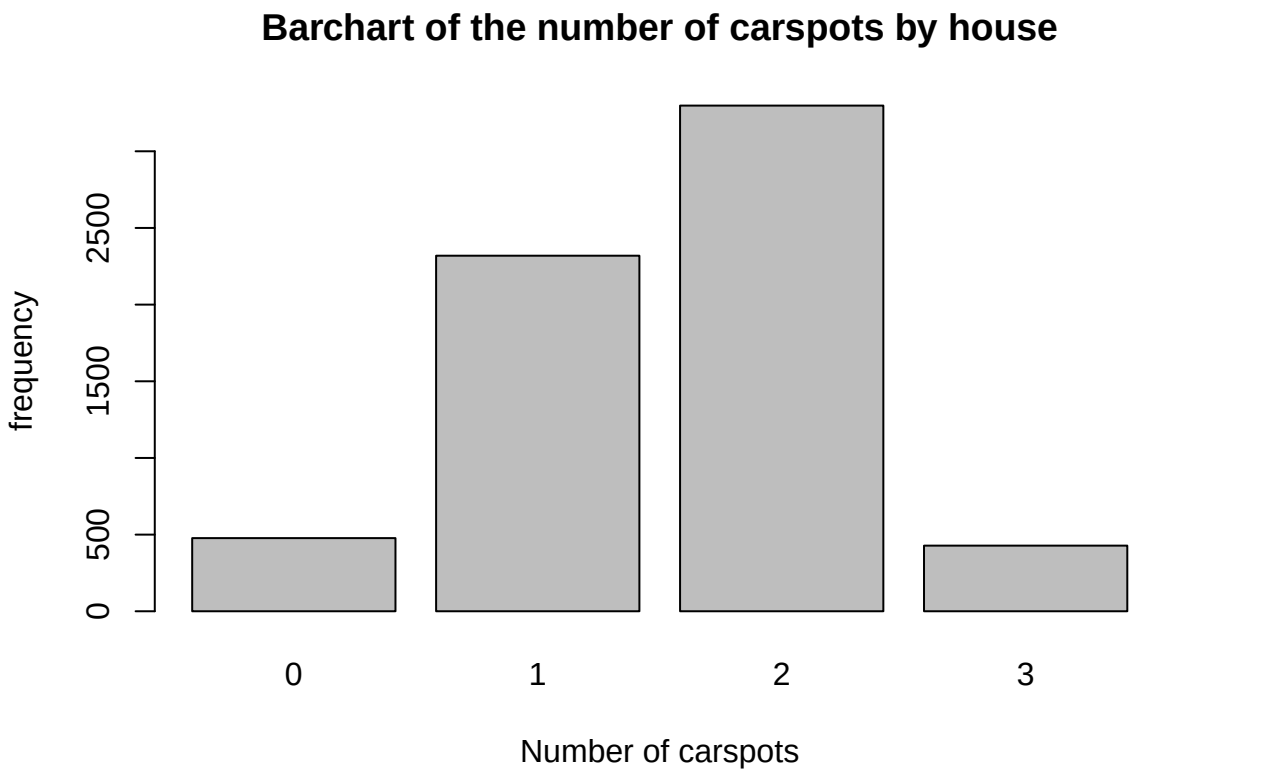

```
boxplot(housing.dataset$YearBuilt, main="Boxplot for the year the house was built")
```



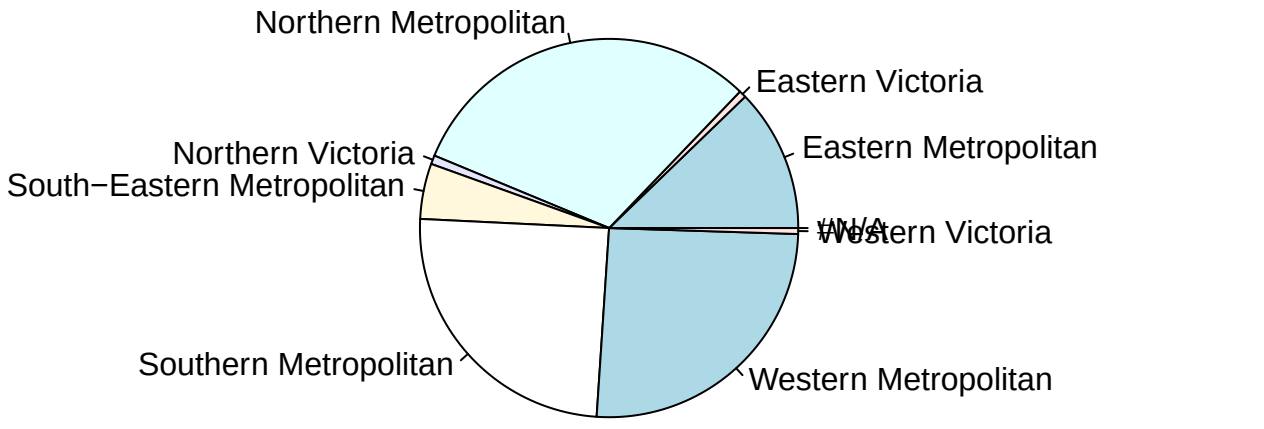

```
# Question 2

#summarise the datas
summary(housing.dataset)
```

```
##        X                  Date      Type         Price             Landsize
##  Min.   :    3   24/02/2018: 179   h:5350   Min.   : 131000   Min.   :   1.0
##  1st Qu.: 7202   17/03/2018: 171   t: 612   1st Qu.: 676000   1st Qu.: 262.0
##  Median :14838   27/05/2017: 170   u: 560   Median : 910250   Median : 476.0
##  Mean   :15726   3/03/2018 : 166            Mean   :1024266   Mean   : 465.6
##  3rd Qu.:23672   28/10/2017: 155            3rd Qu.:1300000   3rd Qu.: 636.0
##  Max.   :34857   3/06/2017 : 151            Max.   :2400000   Max.   :1274.0
##                  (Other)   :5530
##   BuildingArea       Rooms          Bathroom          Car          YearBuilt
##  Min.   :  1.0   Min.   :1.000   Min.   :1.000   Min.   :0.000   Min.   :1850
##  1st Qu.:106.0   1st Qu.:3.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1940
##  Median :133.0   Median :3.000   Median :2.000   Median :2.000   Median :1970
##  Mean   :140.9   Mean   :3.124   Mean   :1.587   Mean   :1.564   Mean   :1964
##  3rd Qu.:171.0   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:1996
##  Max.   :282.0   Max.   :6.000   Max.   :3.000   Max.   :3.000   Max.   :2018
##
##     Distance                          Regionname   Propertycount
##  11.2   : 286   Northern Metropolitan     :2016   21650  : 163
##  14.7   : 141   Western Metropolitan      :1666   8870   : 154
##  7.8    : 139   Southern Metropolitan     :1612   10969  : 121
##  13.9   : 136   Eastern Metropolitan      : 797   11918  : 116
##  5.2    : 136   South-Eastern Metropolitan: 309   11204  : 110
##  9.2    : 125   Northern Victoria         :  49   14577  : 106
##  (Other):5559   (Other)                   :  73   (Other):5752
```

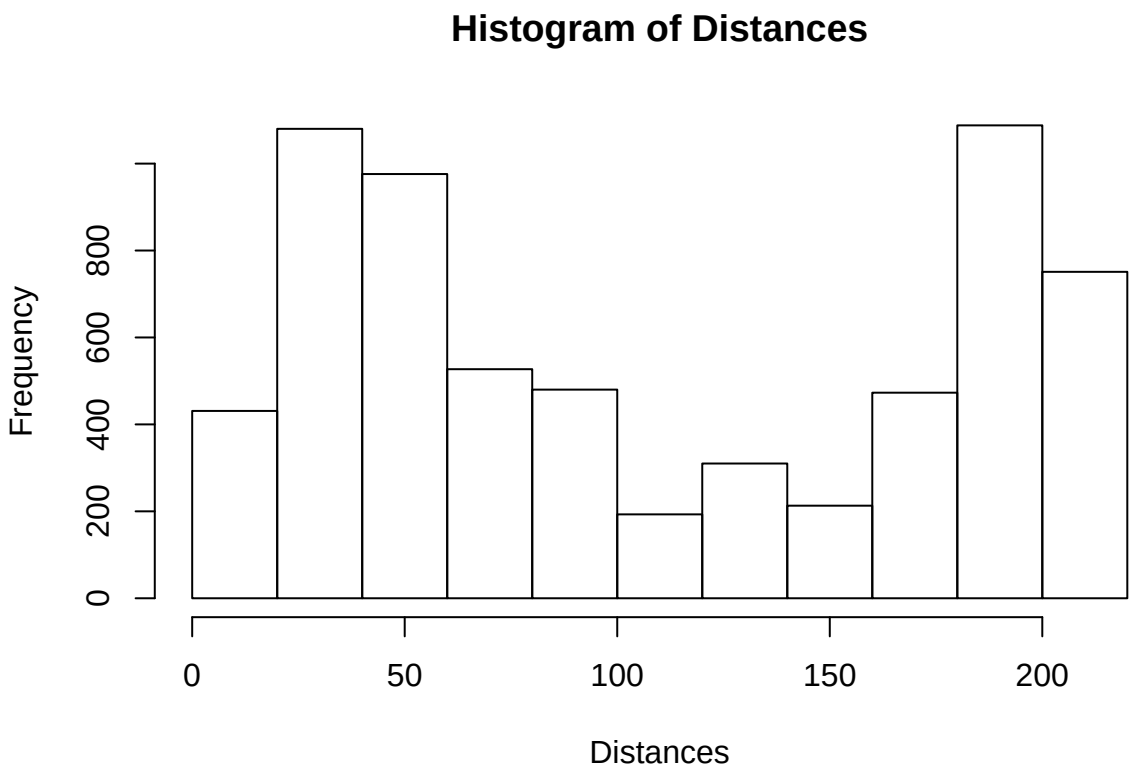
```
#With the summary we can see the minimum, the maximum, the mean and the median of each variables. For e

#Barplot
count <- table(housing.dataset$Car)
barplot(count, main = "Barchart of the number of carspots by house", xlab="Number of carspots", ylab="f
```



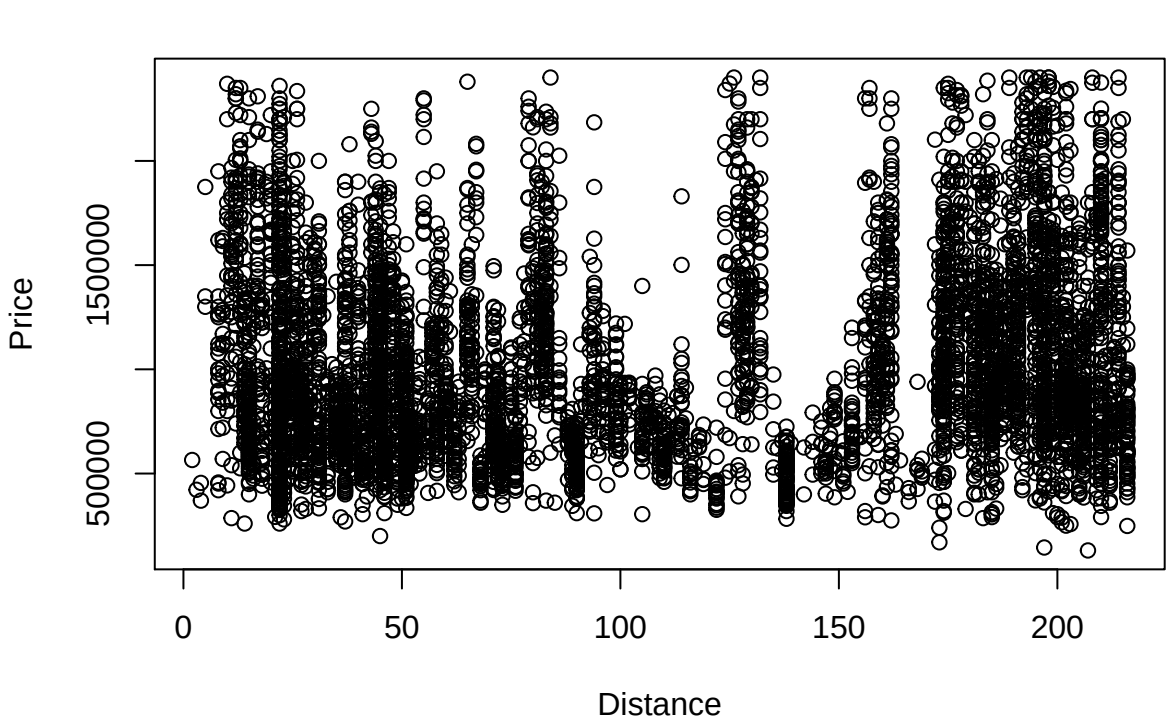

```
#Piechart
count <- table(housing.dataset$Regionname)
pie(count, main = "Pie chart of the regions")
```




```
#Histogram
distance <- as.numeric(housing.dataset$Distance)
hist(distance, xlab = "Distances", main = "Histogram of Distances")
```




```
#Scatter plot
plot(distance, housing.dataset$Price, main="Price in function of distance from CBD", xlab="Distance", yl
```




```
# Question 3.a

library(ggplot2)
```

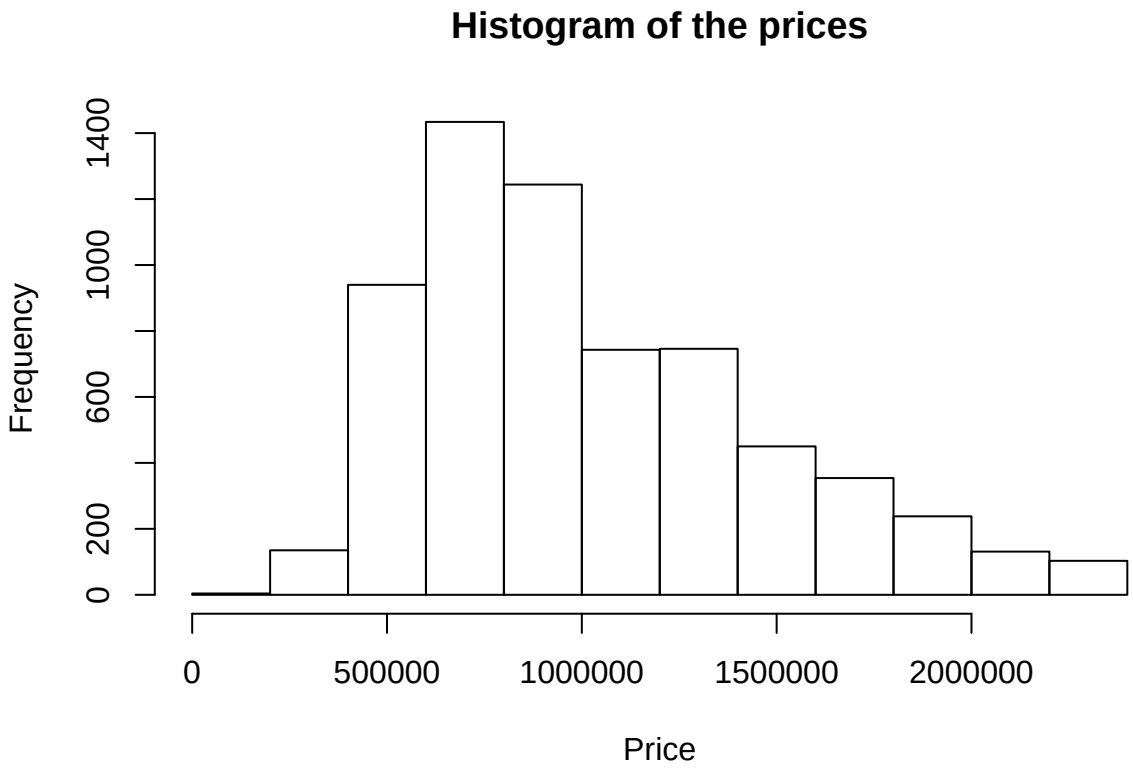
```r
#summary of the price to have more infos
summary(housing.dataset$Price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  131000  676000  910250 1024266 1300000 2400000
```

```r
#calcul of the variance
var(housing.dataset$Price)
```

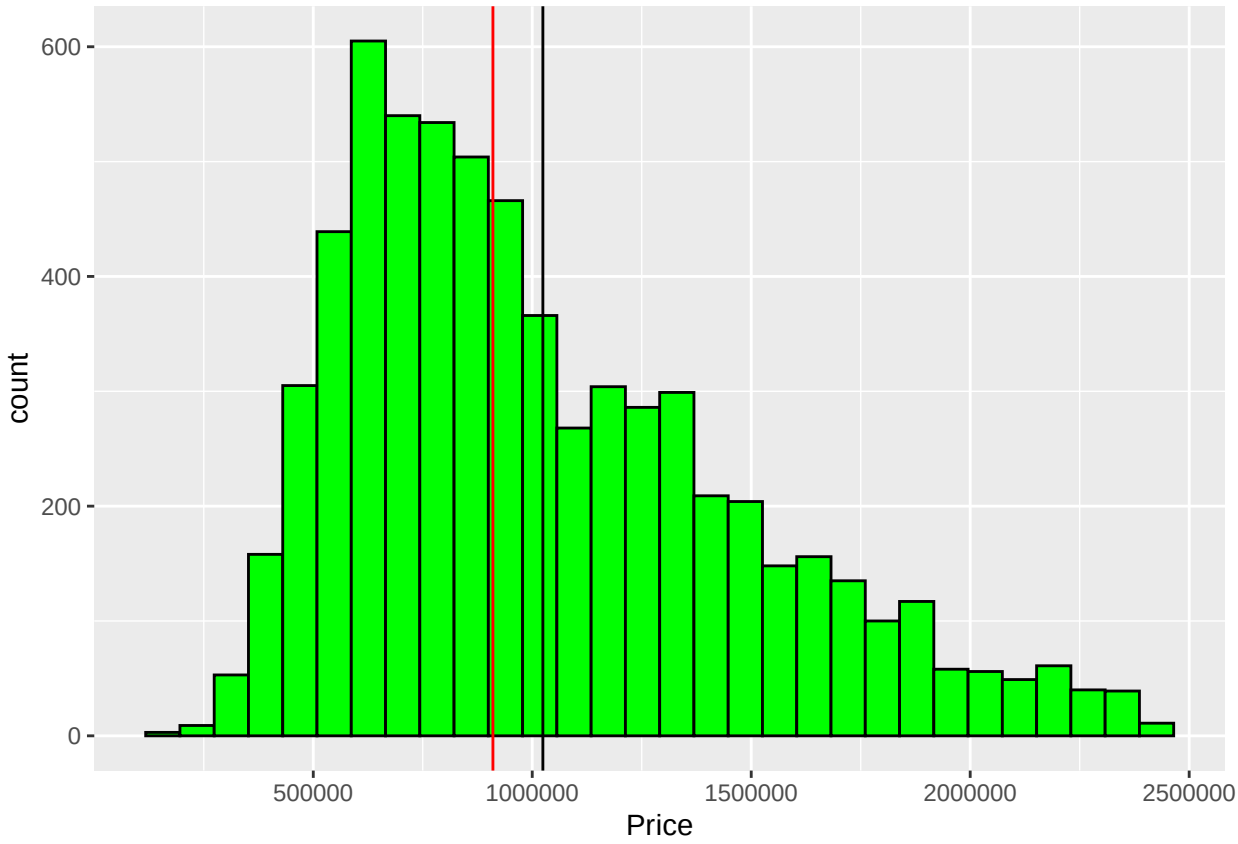
```
## [1] 203427309529
```

```r
#Histogram methode 1
hist(housing.dataset$Price, xlab = "Price", main = "Histogram of the prices")
```




```r
#Methode 2
ggplot(housing.dataset, aes(Price)) +
    geom_histogram(color = "black", fill = "green") +
    geom_vline(xintercept = mean(housing.dataset$Price)) +
    geom_vline(xintercept = median(housing.dataset$Price), color ="red")
```

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




```
#We can see with this graph and the summary that the most of the houses cost around 910250$ but the ave
```

```
# Question 3

#Different separations for the prices (in function of quantiles and median)
# low: < 695000$
#medium: 695000$ to 955000$
#medium high: 955000$ to 1401000$
#high : 1401000$ to 9000000$
summary(housing.dataset$Price)
```

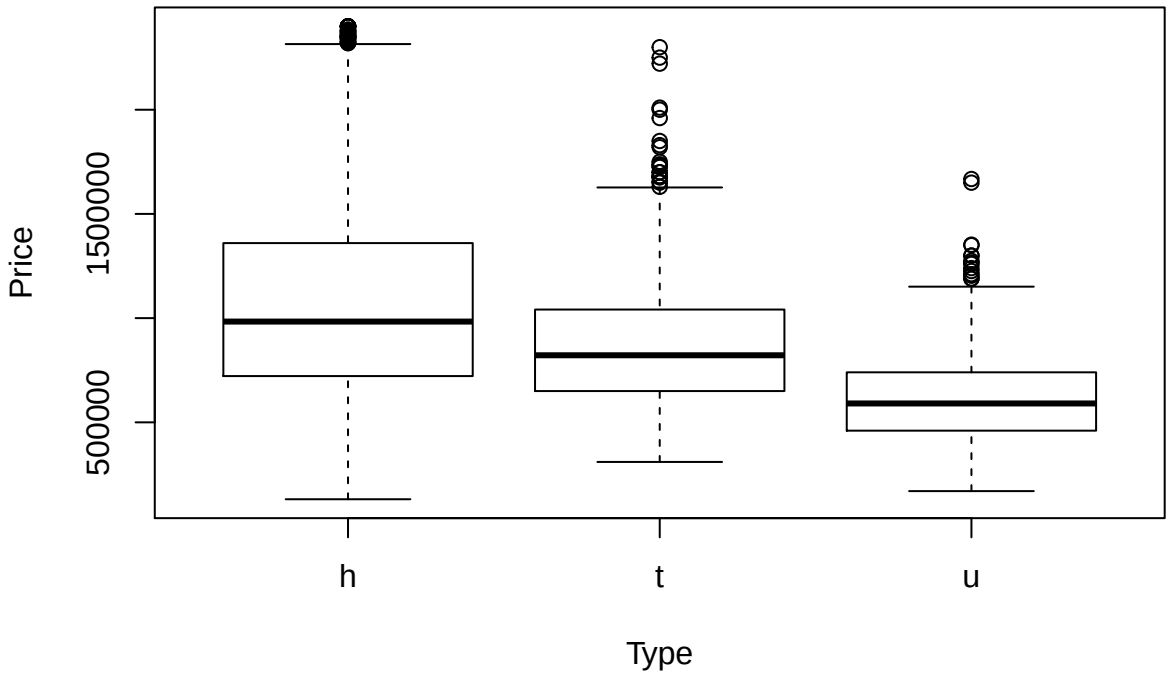
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  131000  676000  910250 1024266 1300000 2400000
```

```
#Cuting the datas
housing.dataset$PriceCategory <- cut(housing.dataset$Price, breaks = quantile(housing.dataset$Price), la

#summary for the different categories
summary(housing.dataset$PriceCategory)
```

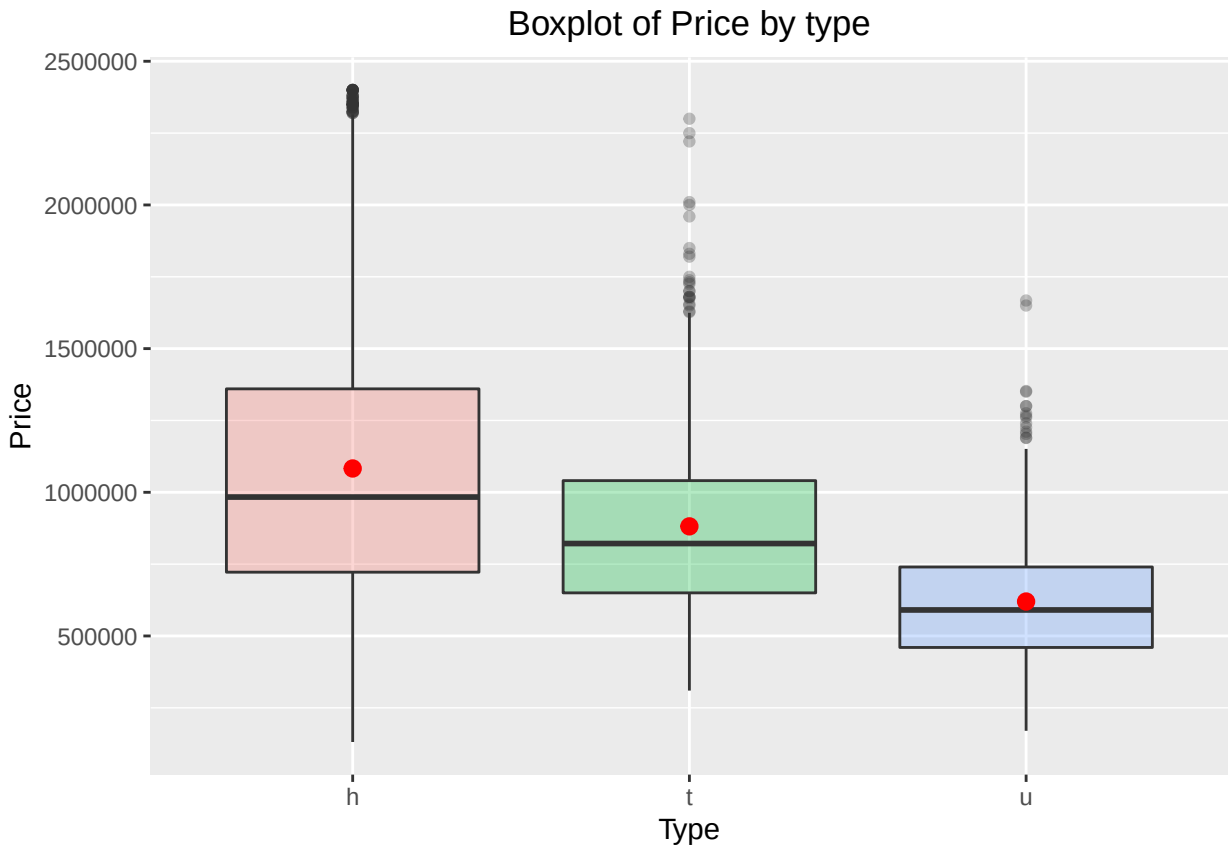
```
##         low  medium low medium high        high        NA's
##        1633        1627        1656        1605           1
```

```
#1633 houses have a price less than 131000$; 1627 houses are between 131000$ and  676000$; 1656 houses
```

```
# Question 3.c

# Prices by type of houses
boxplot(housing.dataset$Price~housing.dataset$Type, xlab="Type", ylab="Price", main="Price by type of h
```




```
#Method 2
ggplot(housing.dataset, aes(x=Type, y=Price, fill=Type)) +
  geom_boxplot(alpha=0.3) +
  stat_summary(fun=mean, geom="point", shape=20, size=4, color="red", fill="red")+
  theme(legend.position="none")+
  ggtitle("Boxplot of Price by type")+
  theme(plot.title = element_text(hjust = 0.5))
```




```
# We can see that houses type h are the more expensive and the houses type u are the cheaper.
```

```
# Question 3.d
# display the different variables
head(housing.dataset)
```

```
##    X      Date Type   Price Landsize BuildingArea Rooms Bathroom Car YearBuilt
## 1  3 4/02/2016    h 1035000      156           79     2        1   0      1900
## 2  5 4/03/2017    h 1465000      134          150     3        2   0      1900
## 3  7 4/06/2016    h 1600000      120          142     4        1   2      2014
## 4 12 7/05/2016    h 1876000      245          210     3        2   0      1910
## 5 15 8/10/2016    h 1636000      256          107     2        1   2      1890
## 6 19 8/10/2016    h 1097000      220           75     2        1   2      1900
##   Distance            Regionname Propertycount PriceCategory
## 1      2.5 Northern Metropolitan          4019   medium high
## 2      2.5 Northern Metropolitan          4019          high
## 3      2.5 Northern Metropolitan          4019          high
## 4      2.5 Northern Metropolitan          4019          high
## 5      2.5 Northern Metropolitan          4019          high
## 6      2.5 Northern Metropolitan          4019   medium high
```

```
#calcul and dosplay of the correlation values
round(cor(housing.dataset[,c(1,4:10)]),2)
```

```
##                  X Price Landsize BuildingArea Rooms Bathroom  Car YearBuilt
## X             1.00 -0.08     0.19         0.13  0.20     0.12 0.16      0.16
## Price        -0.08  1.00     0.09         0.43  0.32     0.29 0.08     -0.44
## Landsize      0.19  0.09     1.00         0.29  0.36     0.11 0.32      0.04
```

```
## BuildingArea  0.13  0.43     0.29          1.00  0.68     0.60 0.35      0.15
## Rooms         0.20  0.32     0.36          0.68  1.00     0.54 0.35      0.08
## Bathroom      0.12  0.29     0.11          0.60  0.54     1.00 0.29      0.27
## Car           0.16  0.08     0.32          0.35  0.35     0.29 1.00      0.26
## YearBuilt     0.16 -0.44     0.04          0.15  0.08     0.27 0.26      1.00
```

```
#We can see that the price is most correlated with the buildingArea, the rooms and bathrooms because th
```

```
# Question 4

#frequencies
#display number of houses of each type
table(unlist(housing.dataset$Type))
```

```
##
##    h    t    u
## 5350  612  560
```

```
#calcul the total number of houses
sum <- 5350+612+560
#calcul the frequencies
housesH <- 5350/sum * 100
housesH
```

```
## [1] 82.03005
```

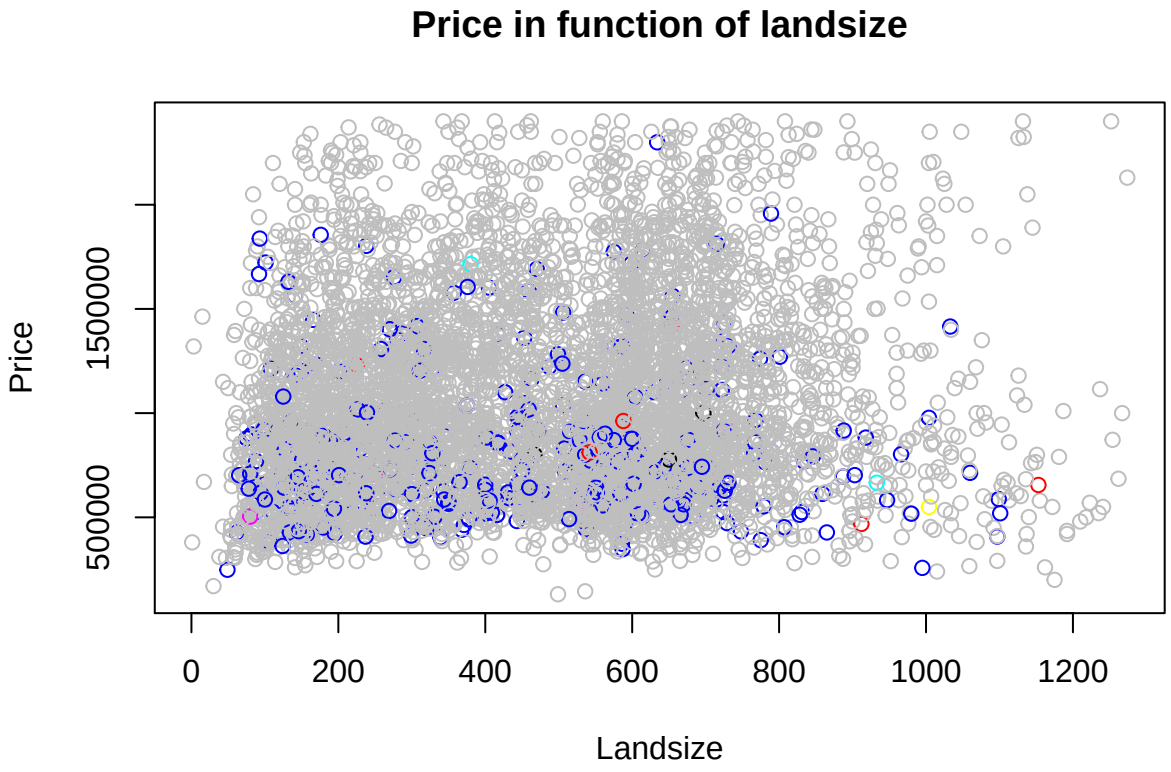
```
housesT <- 612/sum * 100
housesT
```

```
## [1] 9.383625
```

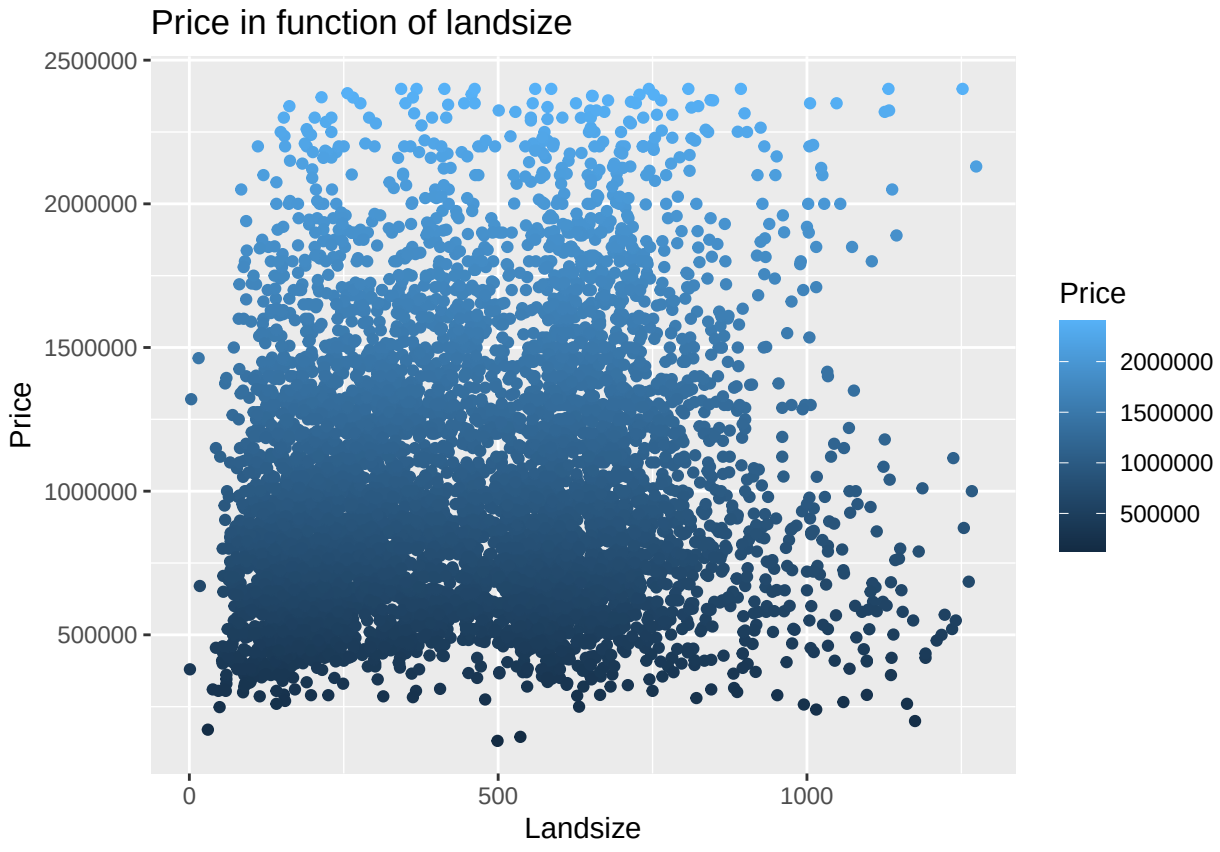
```
housesU <- 560/sum * 100
housesU
```

```
## [1] 8.586323
```

```
#82.03% of houses are type h, 9.38% of the houses are type t and 8.59% are type u.

# Scatterplot for the price and the landsize
#Scatter plot 1
plot(housing.dataset$Landsize, housing.dataset$Price, main="Price in function of landsize", xlab="Landsi
```

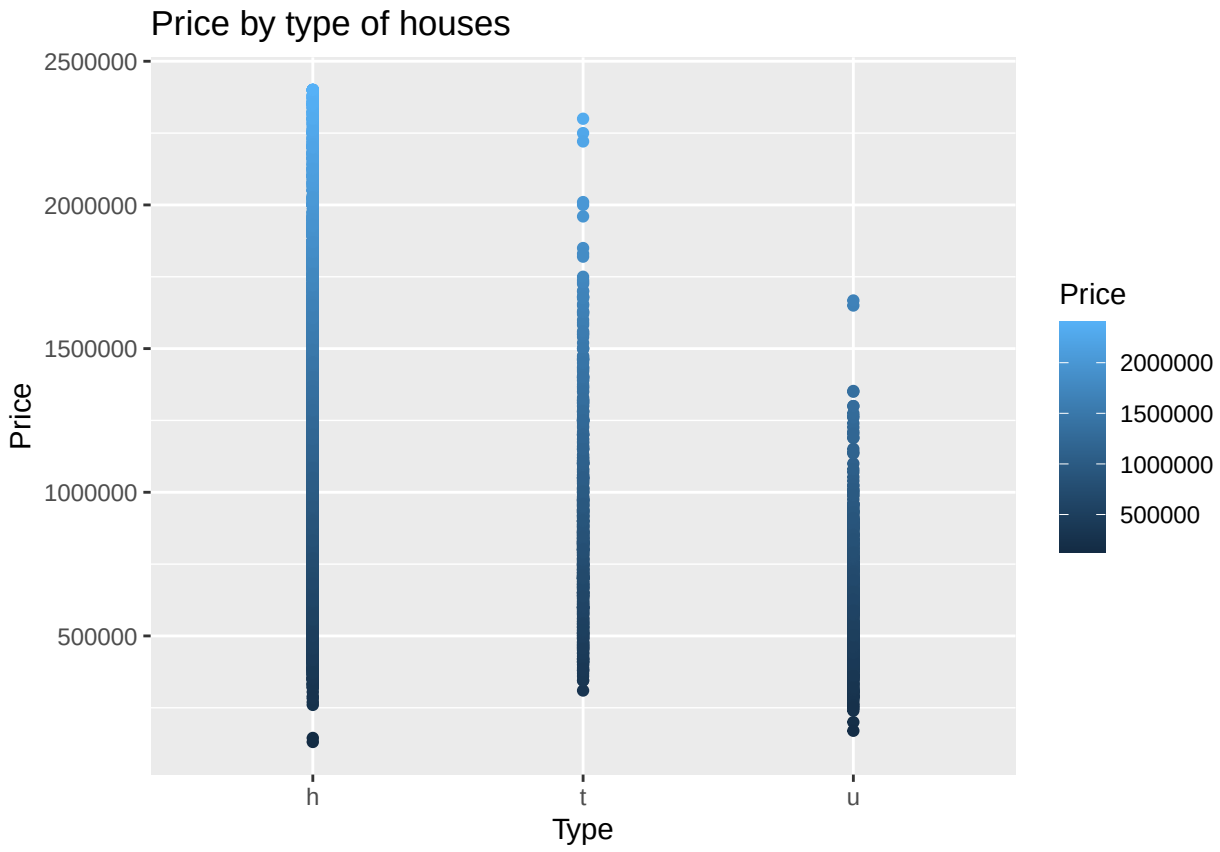



```
#Methode 2
ggplot(housing.dataset, aes(x=Landsize, y=Price, color=Price)) + geom_point() +
  ggtitle("Price in function of landsize")
```



```r
# Scatterplot for the price and the type of house
#Scatter plot 2
plot(housing.dataset$Type, housing.dataset$Price, main="Price in function of Type of houses", xlab="Dis
```


```r
#Methode 2
ggplot(housing.dataset, aes(x=Type, y=Price, color=Price)) + geom_point() +
  ggtitle("Price by type of houses")
```


Price by type of houses
2500000
2000000
1500000
1000000
500000
Price
h
t
u
Type
Price
2000000
1500000
1000000
500000