EDA

2023-10-22

housing =read.csv('data.csv')

head(housing)

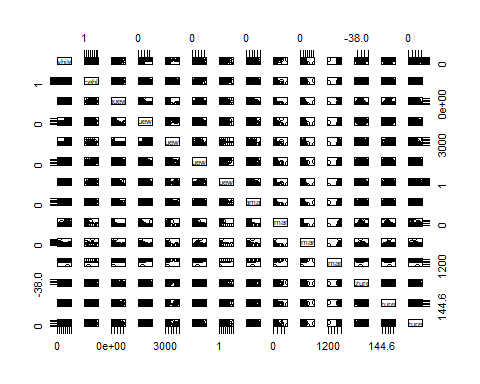
## X Suburb Address Rooms Type Price Method SellerG Date  
## 1 1 Abbotsford 25 Bloomburg St 2 h 1035000 S Biggin 4/2/2016  
## 2 2 Abbotsford 5 Charles St 3 h 1465000 SP Biggin 4/3/2017  
## 3 3 Abbotsford 55a Park St 4 h 1600000 VB Nelson 4/6/2016  
## 4 4 Abbotsford 124 Yarra St 3 h 1876000 S Nelson 7/5/2016  
## 5 5 Abbotsford 98 Charles St 2 h 1636000 S Nelson 8/10/2016  
## 6 6 Abbotsford 10 Valiant St 2 h 1097000 S Biggin 8/10/2016  
## Distance Postcode Bedroom2 Bathroom Car Landsize BuildingArea YearBuilt  
## 1 2.5 3067 2 1 0 156 79 1900  
## 2 2.5 3067 3 2 0 134 150 1900  
## 3 2.5 3067 3 1 2 120 142 2014  
## 4 2.5 3067 4 2 0 245 210 1910  
## 5 2.5 3067 2 1 2 256 107 1890  
## 6 2.5 3067 3 1 2 220 75 1900  
## CouncilArea Lattitude Longtitude Regionname Propertycount  
## 1 Yarra -37.8079 144.9934 Northern Metropolitan 4019  
## 2 Yarra -37.8093 144.9944 Northern Metropolitan 4019  
## 3 Yarra -37.8072 144.9941 Northern Metropolitan 4019  
## 4 Yarra -37.8024 144.9993 Northern Metropolitan 4019  
## 5 Yarra -37.8060 144.9954 Northern Metropolitan 4019  
## 6 Yarra -37.8010 144.9989 Northern Metropolitan 4019

our depend variable is sale price called as price in dataset

str(housing)

## 'data.frame': 6830 obs. of 22 variables:  
## $ X : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Suburb : chr "Abbotsford" "Abbotsford" "Abbotsford" "Abbotsford" ...  
## $ Address : chr "25 Bloomburg St" "5 Charles St" "55a Park St" "124 Yarra St" ...  
## $ Rooms : int 2 3 4 3 2 2 3 2 2 3 ...  
## $ Type : chr "h" "h" "h" "h" ...  
## $ Price : num 1035000 1465000 1600000 1876000 1636000 ...  
## $ Method : chr "S" "SP" "VB" "S" ...  
## $ SellerG : chr "Biggin" "Biggin" "Nelson" "Nelson" ...  
## $ Date : chr "4/2/2016" "4/3/2017" "4/6/2016" "7/5/2016" ...  
## $ Distance : num 2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5 ...  
## $ Postcode : int 3067 3067 3067 3067 3067 3067 3067 3067 3067 3067 ...  
## $ Bedroom2 : int 2 3 3 4 2 3 3 2 2 3 ...  
## $ Bathroom : int 1 2 1 2 1 1 2 2 1 2 ...  
## $ Car : int 0 0 2 0 2 2 2 1 2 1 ...  
## $ Landsize : int 156 134 120 245 256 220 214 0 238 113 ...  
## $ BuildingArea : num 79 150 142 210 107 75 190 94 97 110 ...  
## $ YearBuilt : int 1900 1900 2014 1910 1890 1900 2005 2009 1890 1880 ...  
## $ CouncilArea : chr "Yarra" "Yarra" "Yarra" "Yarra" ...  
## $ Lattitude : num -37.8 -37.8 -37.8 -37.8 -37.8 ...  
## $ Longtitude : num 145 145 145 145 145 ...  
## $ Regionname : chr "Northern Metropolitan" "Northern Metropolitan" "Northern Metropolitan" "Northern Metropolitan" ...  
## $ Propertycount: int 4019 4019 4019 4019 4019 4019 4019 4019 4019 4019 ...

# Assuming 'housing' is a data frame in R  
# Select only the numerical columns  
numerical\_data <- housing[sapply(housing, is.numeric)]  
  
# Create a scatterplot matrix  
pairs(numerical\_data,colors(distinct = TRUE))



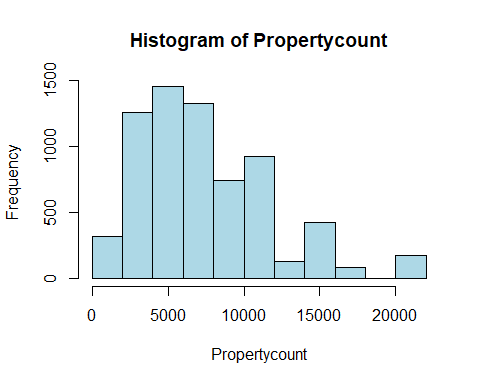
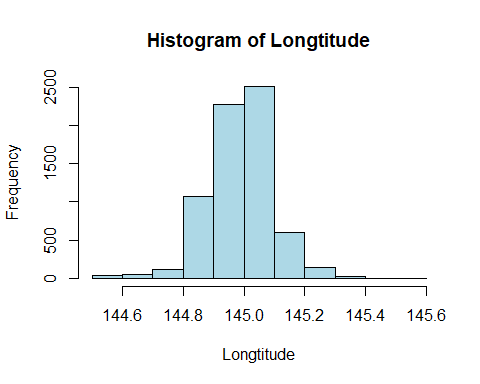
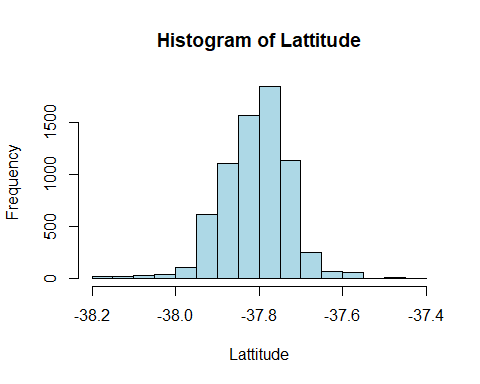
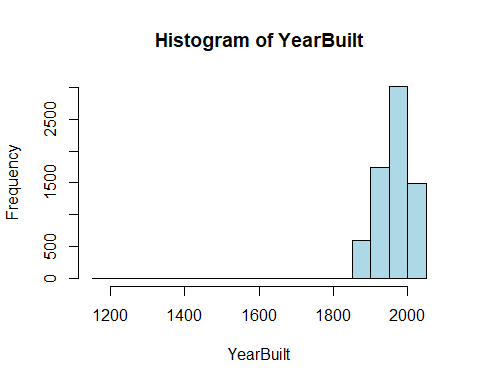
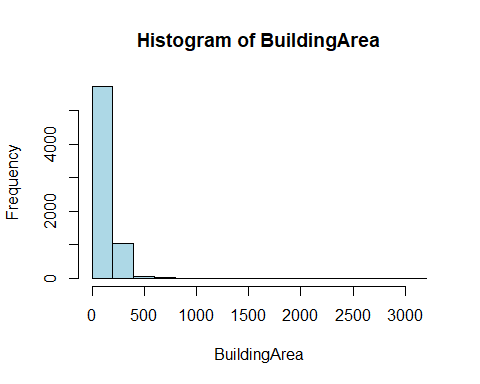
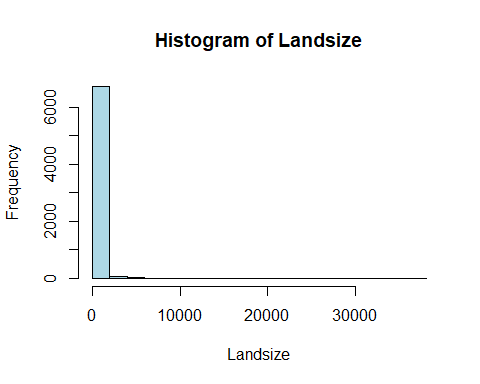
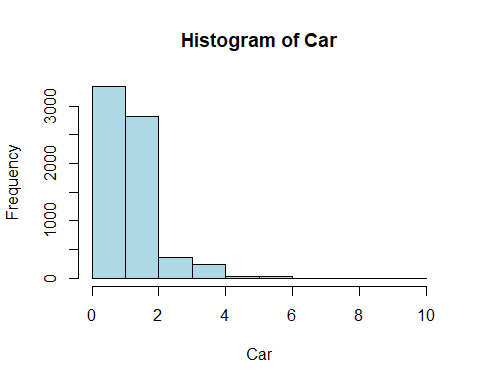
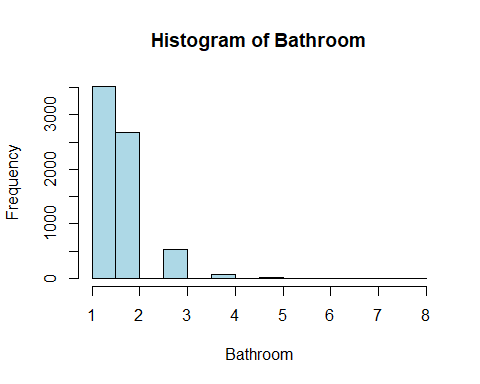
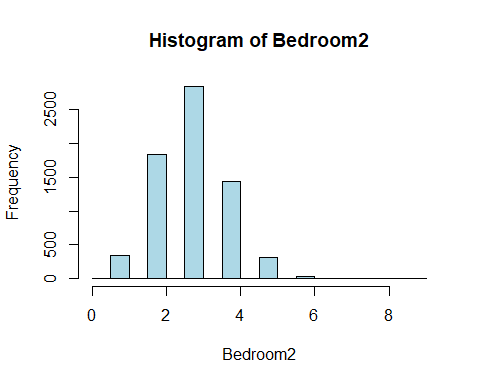
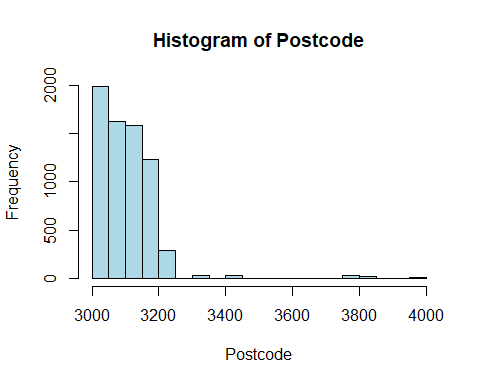
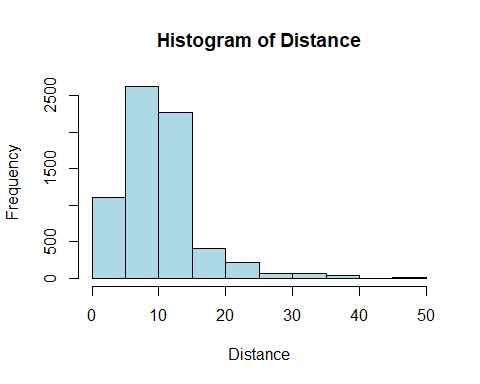
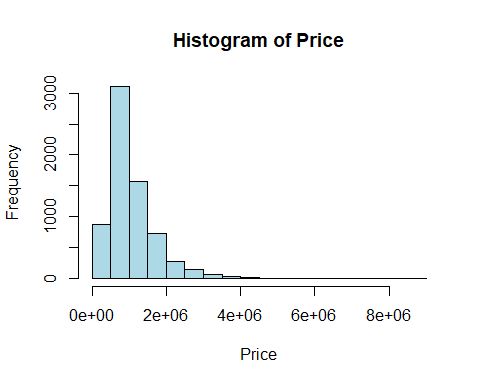
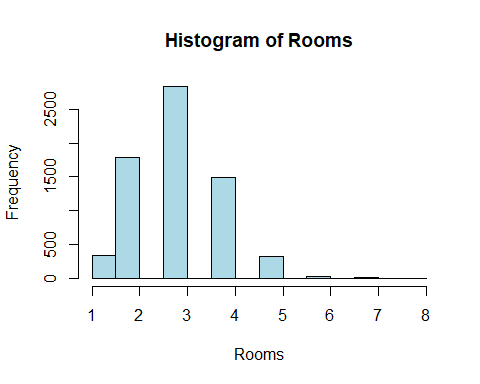
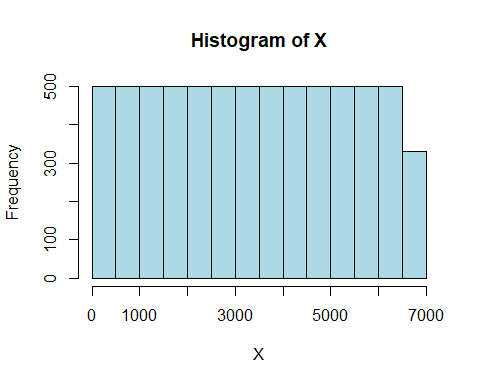
summary(housing)

## X Suburb Address Rooms   
## Min. : 1 Length:6830 Length:6830 Min. :1.000   
## 1st Qu.:1708 Class :character Class :character 1st Qu.:2.000   
## Median :3416 Mode :character Mode :character Median :3.000   
## Mean :3416 Mean :2.978   
## 3rd Qu.:5123 3rd Qu.:4.000   
## Max. :6830 Max. :8.000   
## Type Price Method SellerG   
## Length:6830 Min. : 131000 Length:6830 Length:6830   
## Class :character 1st Qu.: 630000 Class :character Class :character   
## Mode :character Median : 890000 Mode :character Mode :character   
## Mean :1077604   
## 3rd Qu.:1334000   
## Max. :9000000   
## Date Distance Postcode Bedroom2   
## Length:6830 Min. : 0.00 Min. :3000 Min. :0.000   
## Class :character 1st Qu.: 6.10 1st Qu.:3044 1st Qu.:2.000   
## Mode :character Median : 9.20 Median :3083 Median :3.000   
## Mean :10.15 Mean :3104 Mean :2.951   
## 3rd Qu.:13.00 3rd Qu.:3147 3rd Qu.:4.000   
## Max. :47.40 Max. :3977 Max. :9.000   
## Bathroom Car Landsize BuildingArea   
## Min. :1.000 Min. : 0.000 Min. : 0.0 Min. : 0.0   
## 1st Qu.:1.000 1st Qu.: 1.000 1st Qu.: 167.0 1st Qu.: 93.0   
## Median :1.000 Median : 2.000 Median : 404.0 Median : 126.0   
## Mean :1.594 Mean : 1.607 Mean : 487.5 Mean : 143.4   
## 3rd Qu.:2.000 3rd Qu.: 2.000 3rd Qu.: 641.0 3rd Qu.: 173.0   
## Max. :8.000 Max. :10.000 Max. :37000.0 Max. :3112.0   
## YearBuilt CouncilArea Lattitude Longtitude   
## Min. :1196 Length:6830 Min. :-38.16 Min. :144.5   
## 1st Qu.:1940 Class :character 1st Qu.:-37.86 1st Qu.:144.9   
## Median :1970 Mode :character Median :-37.80 Median :145.0   
## Mean :1964 Mean :-37.81 Mean :145.0   
## 3rd Qu.:2000 3rd Qu.:-37.76 3rd Qu.:145.1   
## Max. :2018 Max. :-37.41 Max. :145.5   
## Regionname Propertycount   
## Length:6830 Min. : 389   
## Class :character 1st Qu.: 4381   
## Mode :character Median : 6567   
## Mean : 7434   
## 3rd Qu.:10175   
## Max. :21650

str(housing)

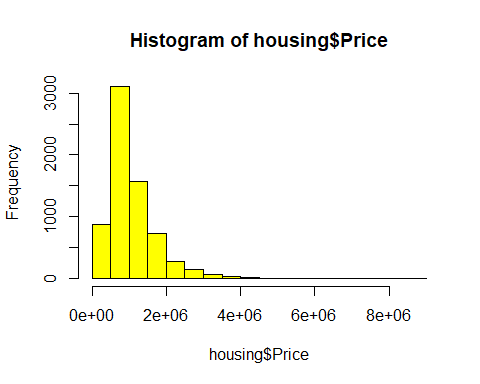
## 'data.frame': 6830 obs. of 22 variables:  
## $ X : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Suburb : chr "Abbotsford" "Abbotsford" "Abbotsford" "Abbotsford" ...  
## $ Address : chr "25 Bloomburg St" "5 Charles St" "55a Park St" "124 Yarra St" ...  
## $ Rooms : int 2 3 4 3 2 2 3 2 2 3 ...  
## $ Type : chr "h" "h" "h" "h" ...  
## $ Price : num 1035000 1465000 1600000 1876000 1636000 ...  
## $ Method : chr "S" "SP" "VB" "S" ...  
## $ SellerG : chr "Biggin" "Biggin" "Nelson" "Nelson" ...  
## $ Date : chr "4/2/2016" "4/3/2017" "4/6/2016" "7/5/2016" ...  
## $ Distance : num 2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5 ...  
## $ Postcode : int 3067 3067 3067 3067 3067 3067 3067 3067 3067 3067 ...  
## $ Bedroom2 : int 2 3 3 4 2 3 3 2 2 3 ...  
## $ Bathroom : int 1 2 1 2 1 1 2 2 1 2 ...  
## $ Car : int 0 0 2 0 2 2 2 1 2 1 ...  
## $ Landsize : int 156 134 120 245 256 220 214 0 238 113 ...  
## $ BuildingArea : num 79 150 142 210 107 75 190 94 97 110 ...  
## $ YearBuilt : int 1900 1900 2014 1910 1890 1900 2005 2009 1890 1880 ...  
## $ CouncilArea : chr "Yarra" "Yarra" "Yarra" "Yarra" ...  
## $ Lattitude : num -37.8 -37.8 -37.8 -37.8 -37.8 ...  
## $ Longtitude : num 145 145 145 145 145 ...  
## $ Regionname : chr "Northern Metropolitan" "Northern Metropolitan" "Northern Metropolitan" "Northern Metropolitan" ...  
## $ Propertycount: int 4019 4019 4019 4019 4019 4019 4019 4019 4019 4019 ...

# Assuming 'housing' is a data frame in R  
# Get a list of all column names  
column\_names <- names(housing)  
  
# Create separate histograms for each numerical variable  
for (col in column\_names) {  
 if (is.numeric(housing[[col]])) {  
 hist(housing[[col]],  
 main = paste("Histogram of", col),  
 xlab = col,  
 ylab = "Frequency",  
 col = "lightblue")  
 }  
}

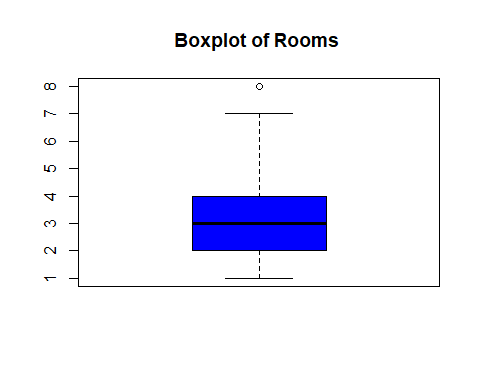


histogram of our predict variable

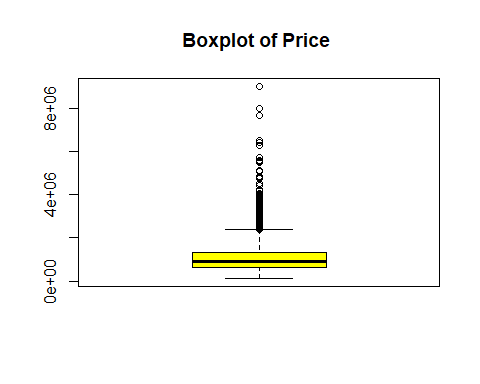
hist(housing$Price,col = 'yellow')



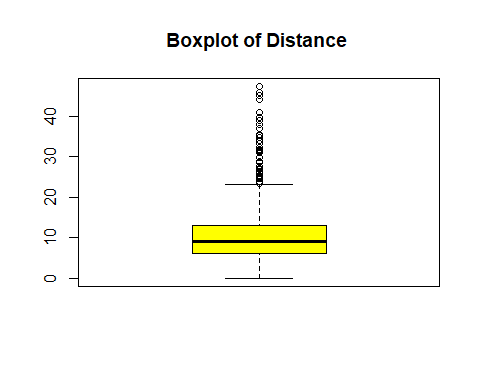
boxplot(housing$Rooms,col = 'blue',main='Boxplot of Rooms ')



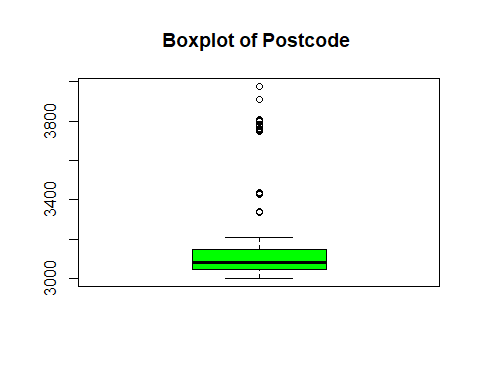
boxplot(housing$Price,col = 'yellow',main='Boxplot of Price ')



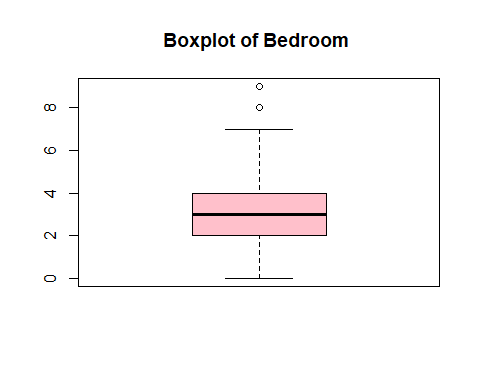
boxplot(housing$Distance,col='yellow',main='Boxplot of Distance ')



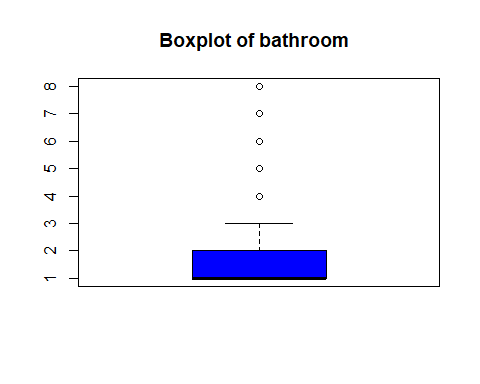
boxplot(housing$Postcode,main='Boxplot of Postcode ',col = 'green')



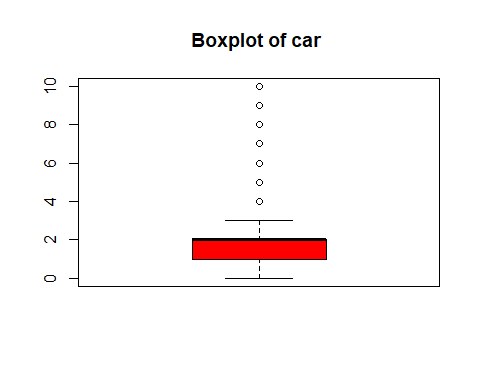
boxplot(housing$Bedroom2,main='Boxplot of Bedroom ',col = 'pink')



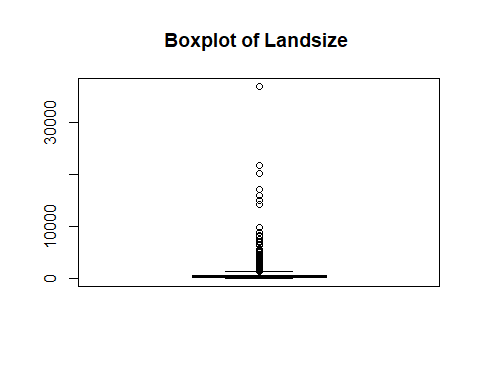
boxplot(housing$Bathroom,main='Boxplot of bathroom ',col = 'blue')



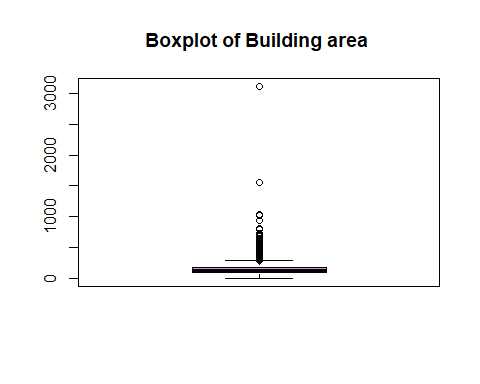
boxplot(housing$Car,main='Boxplot of car ',col = 'red')



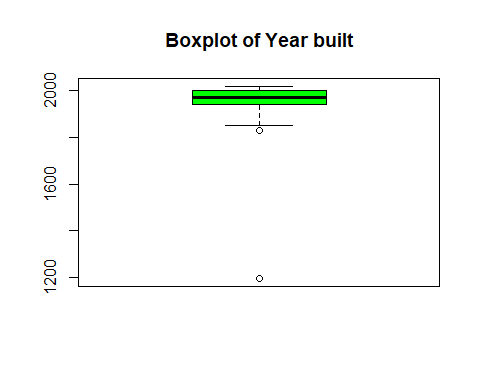
boxplot(housing$Landsize,main='Boxplot of Landsize ',col = 'violet')



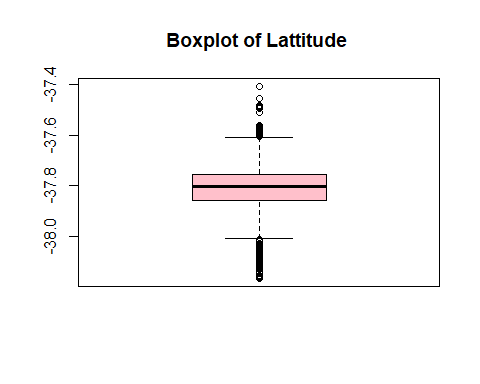
boxplot(housing$BuildingArea,main='Boxplot of Building area ',col='violet')



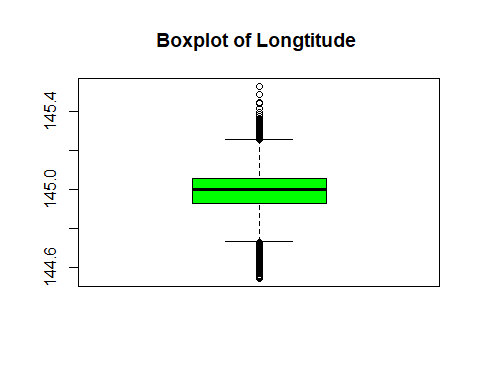
boxplot(housing$YearBuilt,col='green',main='Boxplot of Year built')



boxplot(housing$Lattitude,main='Boxplot of Lattitude ',col = 'pink')



boxplot(housing$Longtitude,main='Boxplot of Longtitude ',col = 'green')



boxplot(housing$Propertycount,main='Boxplot of Property count ',col = 'orange')

