Quantitative Data Analysis of House Dataset

26 February, 2023

1. Organise and clean the data

1.1 Subset the data into the specific dataset allocated

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.2
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.2.2
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## v purrr   0.3.5
```

```
## Warning: package 'tidyr' was built under R version 4.2.2
```

```
## Warning: package 'readr' was built under R version 4.2.2
```

```
## Warning: package 'purrr' was built under R version 4.2.2
```

```
## Warning: package 'dplyr' was built under R version 4.2.2
```

```
## Warning: package 'forcats' was built under R version 4.2.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
mydata<-read.csv("Mydata.csv")
```

Removing 'id' variable as it is not needed for further analysis.

```
mydata<-subset(mydata,select=-c(id))
str(mydata)
```

```
## 'data.frame':    903 obs. of  11 variables:
##  $ price               : int  109000 84000 149000 43000 52000 35000 59000 24000 115000 300000 ...
##  $ mq                  : int  190 150 60 73 52 50 60 30 120 750 ...
##  $ floor               : int  1 1 2 1 1 1 1 2 1 3 ...
##  $ n_rooms             : int  5 5 2 4 4 3 5 2 5 3 ...
##  $ n_bathrooms         : int  1 2 1 1 1 1 1 1 2 1 ...
##  $ has_terrace         : int  0 1 1 1 0 0 0 0 1 0 ...
##  $ has_alarm           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ heating             : chr  "autonomous" "autonomous" "autonomous" "autonomous" ...
##  $ has_air_conditioning: int  0 0 1 0 0 0 0 0 0 1 ...
##  $ has_parking         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ is_furnished        : int  0 0 0 0 0 0 0 0 0 0 ...
```

1.2 Data quality analysis

To check the quality of the dataset:

1. The summary() function automatically calculates summary statistics for the vector. such as, the minimum value,The value of the 1st quartile, median value, The value of the 2nd quartile and the maximum value.

```
summary(mydata)
```

```
##      price              mq             floor          n_rooms     
##  Min.   :  6500   Min.   :  0.0   Min.   :1.000   Min.   :-1.000  
##  1st Qu.: 75000   1st Qu.: 75.0   1st Qu.:1.000   1st Qu.: 3.000  
##  Median :120000   Median :100.0   Median :2.000   Median : 3.000  
##  Mean   :145262   Mean   :117.3   Mean   :1.837   Mean   : 3.467  
##  3rd Qu.:190000   3rd Qu.:136.5   3rd Qu.:2.000   3rd Qu.: 4.000  
##  Max.   :500000   Max.   :821.0   Max.   :9.000   Max.   : 5.000  
##   n_bathrooms     has_terrace      has_alarm          heating         
##  Min.   :1.000   Min.   :0.000   Min.   :0.00000   Length:903        
##  1st Qu.:1.000   1st Qu.:0.000   1st Qu.:0.00000   Class :character  
##  Median :1.000   Median :0.000   Median :0.00000   Mode  :character  
##  Mean   :1.444   Mean   :0.124   Mean   :0.01107                     
##  3rd Qu.:2.000   3rd Qu.:0.000   3rd Qu.:0.00000                     
##  Max.   :3.000   Max.   :1.000   Max.   :1.00000                     
##  has_air_conditioning  has_parking       is_furnished    
##  Min.   :0.0000       Min.   :0.00000   Min.   :0.00000  
##  1st Qu.:0.0000       1st Qu.:0.00000   1st Qu.:0.00000  
##  Median :0.0000       Median :0.00000   Median :0.00000  
##  Mean   :0.3079       Mean   :0.01218   Mean   :0.08306  
##  3rd Qu.:1.0000       3rd Qu.:0.00000   3rd Qu.:0.00000  
##  Max.   :1.0000       Max.   :1.00000   Max.   :1.00000  
```

2. Checking if there is any ‘NA’ in the dataset using is.na function.

```
sum(is.na(mydata))
```

```
## [1] 0
```

3. str() gives us the structure of the object and information about the class, length and content of each column.

```
str(mydata)
```

```
## 'data.frame':    903 obs. of  11 variables:
##  $ price               : int  109000 84000 149000 43000 52000 35000 59000 24000 115000 300000 ...
##  $ mq                  : int  190 150 60 73 52 50 60 30 120 750 ...
##  $ floor               : int  1 1 2 1 1 1 1 2 1 3 ...
##  $ n_rooms             : int  5 5 2 4 4 3 5 2 5 3 ...
##  $ n_bathrooms         : int  1 2 1 1 1 1 1 1 2 1 ...
##  $ has_terrace         : int  0 1 1 1 0 0 0 0 1 0 ...
##  $ has_alarm           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ heating             : chr  "autonomous" "autonomous" "autonomous" "autonomous" ...
##  $ has_air_conditioning: int  0 0 1 0 0 0 0 0 0 1 ...
##  $ has_parking         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ is_furnished        : int  0 0 0 0 0 0 0 0 0 0 ...
```

4. With table function we can get frequency of the variables.

```
table(mydata$floor)
```

```
## 
##   1   2   3   4   5   6   7   8   9 
## 431 294 116  39  12   4   5   1   1
```

```
table(mydata$n_rooms)
```

```
## 
##  -1   2   3   4   5 
##   1 159 322 257 164
```

```
table(mydata$n_bathrooms)
```

```
## 
##   1   2   3 
## 538 329  36
```

```
table(mydata$has_terrace)
```

```
## 
##   0   1 
## 791 112
```

```
table(mydata$has_alarm)
```

```
##
##   0   1
## 893  10
```

```
table(mydata$heating)
```

```
##
## autonamous autonomous      other
##          1        804         98
```

```
table(mydata$has_air_conditioning)
```

```
##
##   0   1
## 625 278
```

```
table(mydata$has_parking)
```

```
##
##   0   1
## 892  11
```

```
table(mydata$is_furnished)
```

```
##
##   0   1
## 828  75
```

Summary of the findings variable wise:

1. n_room: It has '-1' value. The number of rooms cannot be negative value.
2. mq: It has '0' value. Total square meters of the property cannot be zero value.
3. Heating: There is a typographical error in it.
4. Price: No issues found in this variable.
5. Floor: Fewer levels are there in it. For better analysis typecast is required.
6. n_bathroom: Fewer levels are there in it. For better analysis typecast is required.
7. has_terrace: Fewer levels are there in it. For better analysis typecast is required.
8. has_alarm: Fewer levels are there in it. For better analysis typecast is required.
9. has_air_conditioning: Fewer levels are there in it. For better analysis typecast is required.
10. has_parking: Fewer levels are there in it. For better analysis typecast is required.
11. is_furnished: Fewer levels are there in it. For better analysis typecast is required.

1.3 Data cleaning

There are number of issues found in the data set:

1. There is a typographical error in the 'heating' variable. The data 'autonamous' is replaced with the correct data 'autonomous'

```r
#checking the unique values
unique(mydata$heating) #autonomous is spelled wrong
```

```
## [1] "autonomous" "other"      "autonamous"
```

```r
mydata$heating <-ifelse(mydata$heating=='autonamous','autonomous',mydata$heating)
```

2. n_room variable has '-1' value. The number of rooms cannot be negative value. So, we are replacing it with the mode of n_room.

```r
# Replacing the n_room ='-1' with the mode of n_room
mydata$n_rooms[mydata$n_rooms==-1]<-as.integer(names(sort(-table(factor(mydata$n_rooms))))[1])
```

3. mq variable has '0' value. Total square meters of the property cannot be zero value. So, we are replacing it with the mean of mq.

```r
# Replacing the mq ='0' with the mean of mq
mydata$mq[mydata$mq==0]<-round(mean(mydata$mq))
```

4. The 8 variables n_bathrooms, n_rooms,floor, has_terrace,has_alarm, has_air_conditioning, has_parking, and is_furnished has very few levels and can be typecast into factors for better analysis, so that R treats them as a grouping variable.

```r
#changing numerical to categorical
mydata$n_bathrooms<-as.factor(mydata$n_bathrooms)
mydata$has_terrace<-as.factor(mydata$has_terrace)
mydata$has_alarm<-as.factor(mydata$has_alarm)
mydata$has_air_conditioning<-as.factor(mydata$has_air_conditioning)
mydata$has_parking<-as.factor(mydata$has_parking)
mydata$is_furnished<-as.factor(mydata$is_furnished)
mydata$n_rooms<-as.factor(mydata$n_rooms)
mydata$floor<-as.factor(mydata$floor)
mydata$heating<-as.factor(mydata$heating)
```

---

2. Exploratory Data Analysis (EDA)

2.1 EDA plan

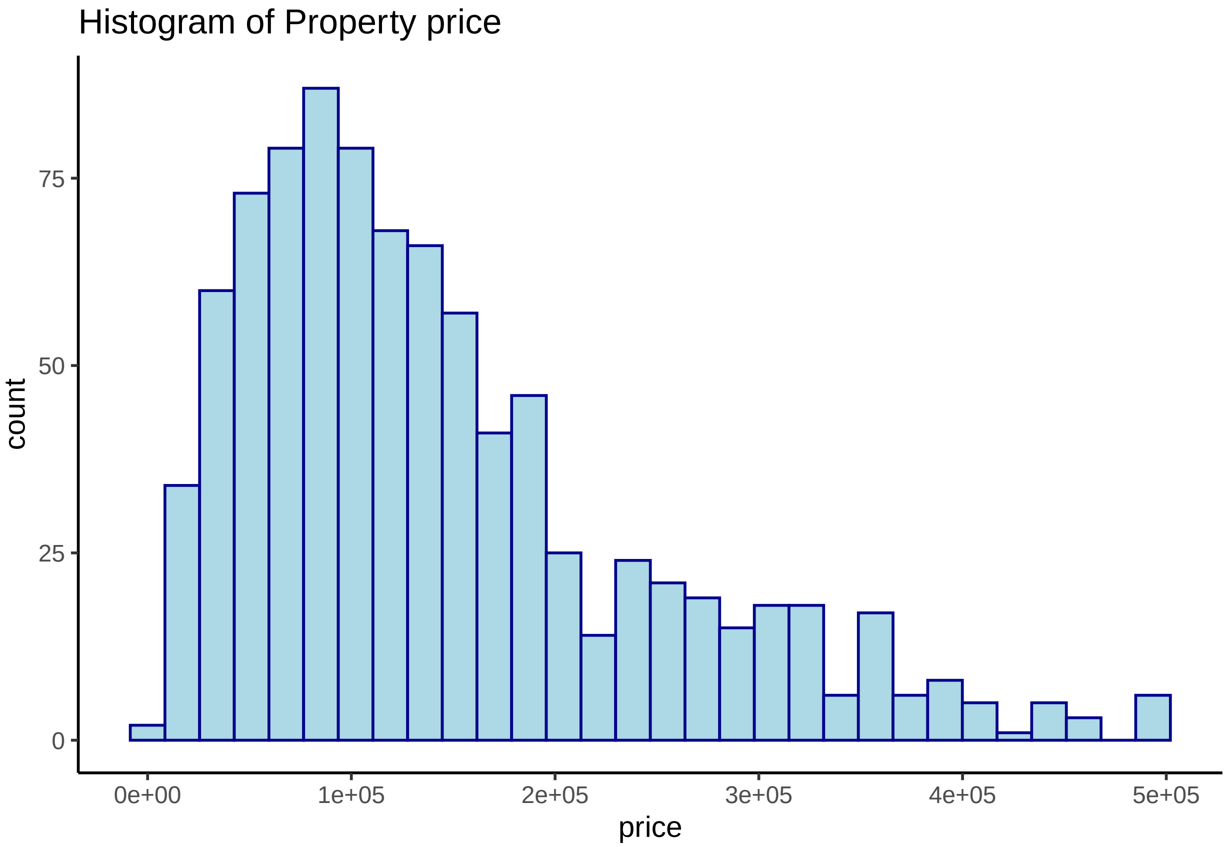
1. Analyzing the continuous variable using histogram.
2. Analyzing the categorical variable using bar plot.
3. Explore the property price and mq with scattered plot.
4. Explore the relationship between the property price and other categorical variables with box plot

2.2 EDA and summary of results

```
ggplot(mydata, aes(x=price)) + geom_histogram(color="darkblue", fill="lightblue") + ggtitle("Histogram
```

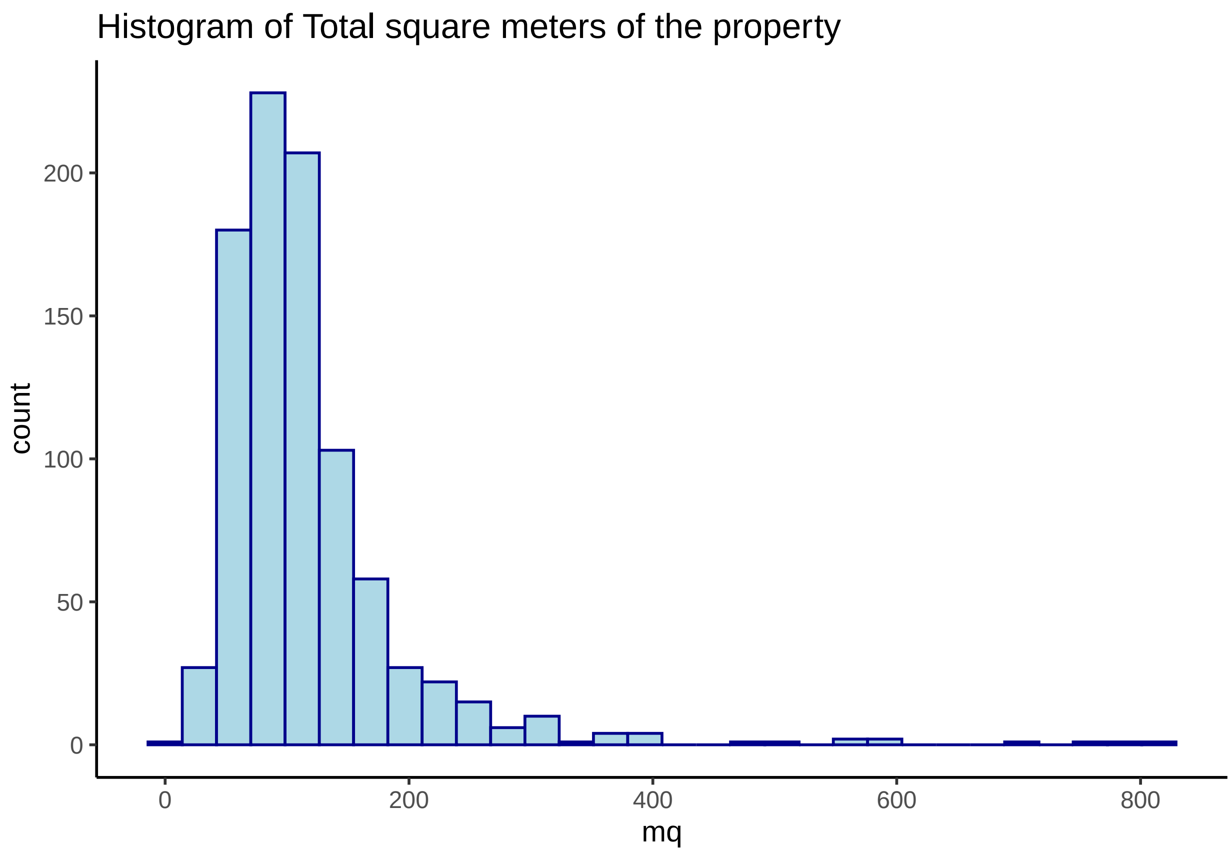
Analyzing the continues variable:

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



```
ggplot(mydata, aes(x=mq)) + geom_histogram(color="darkblue", fill="lightblue") + ggtitle("Histogram of
```

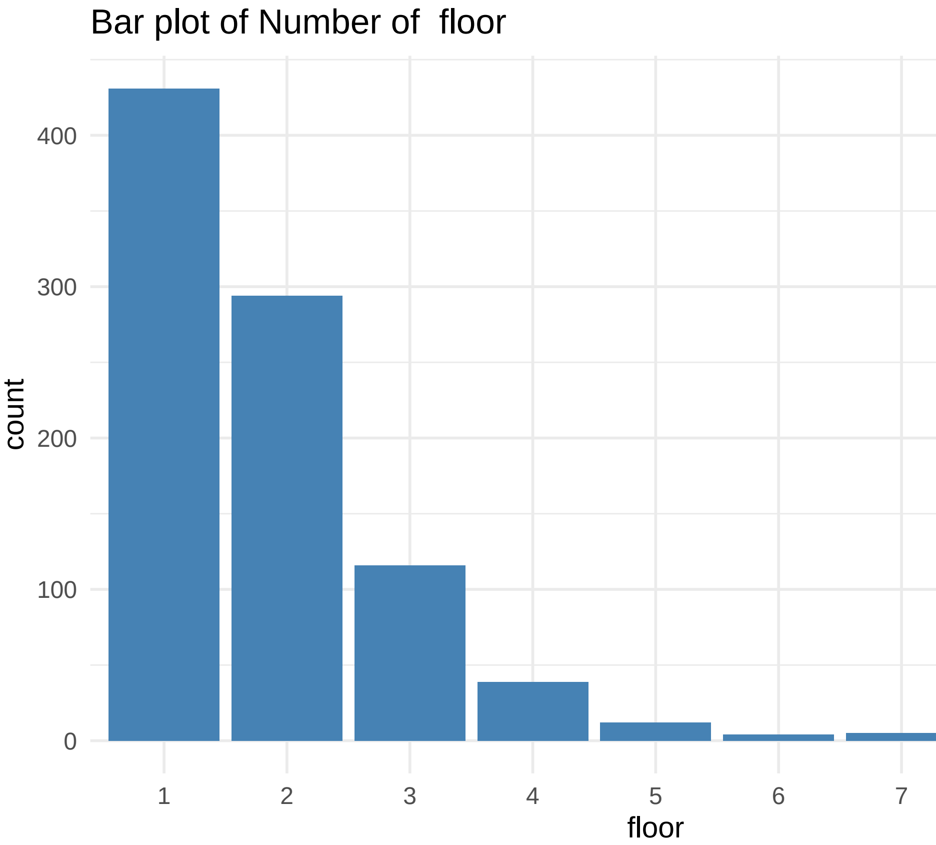
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Summary of the findings:

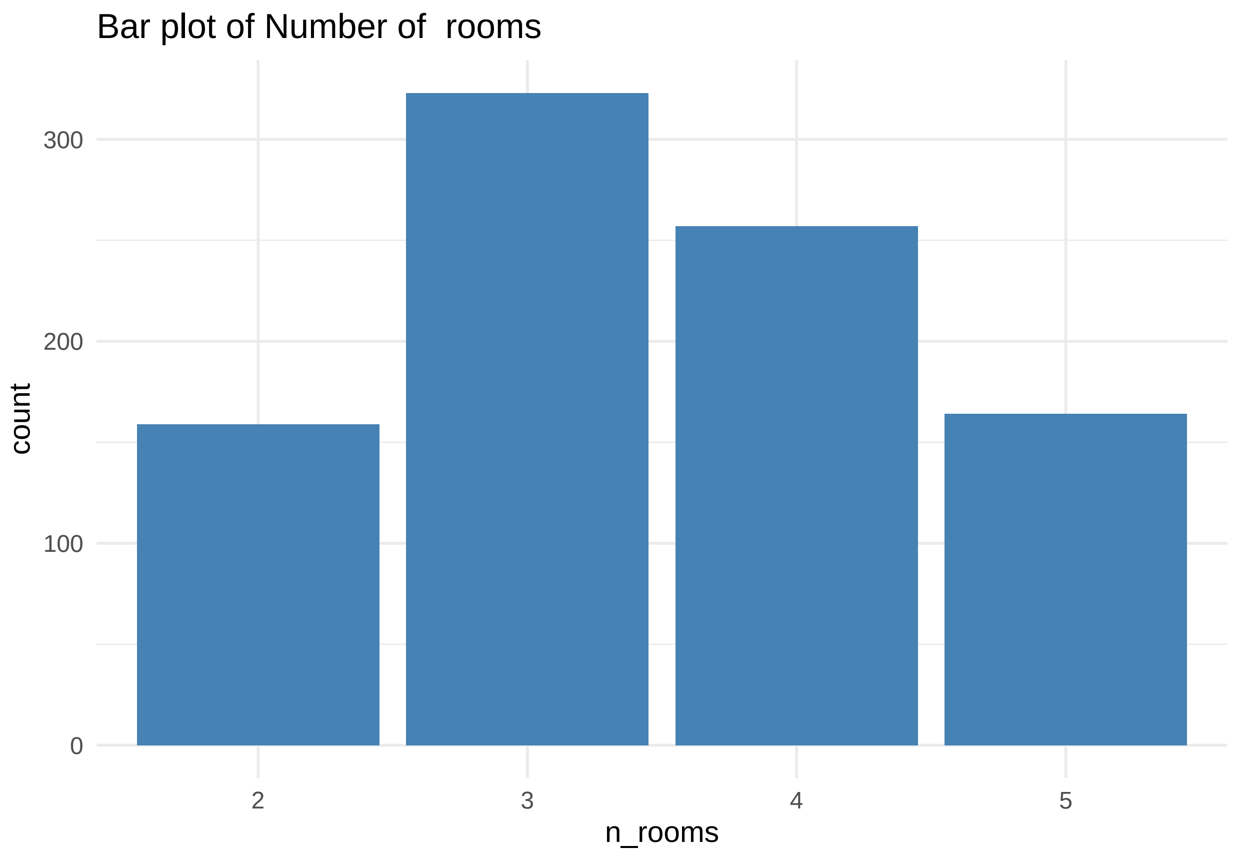
1. Both the price and total square meters of the property histogram is skewed right.
2. In price histogram, there are few outliers like the one in 500000, which may affect the results

```
ggplot(mydata, aes(x=floor))+ geom_bar( fill="steelblue")+ theme_minimal()+ggtitle("Bar plot of Number
```

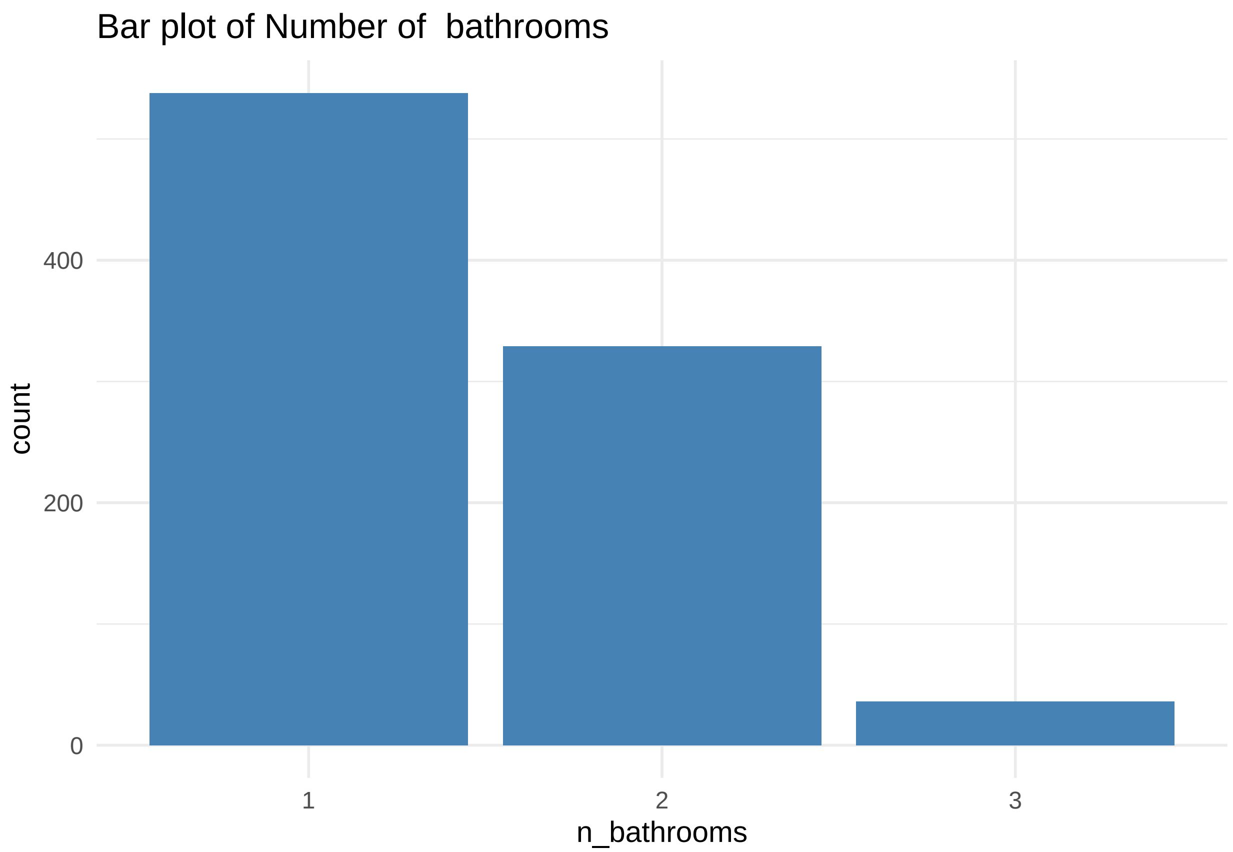


Analyzing the categorical variables:

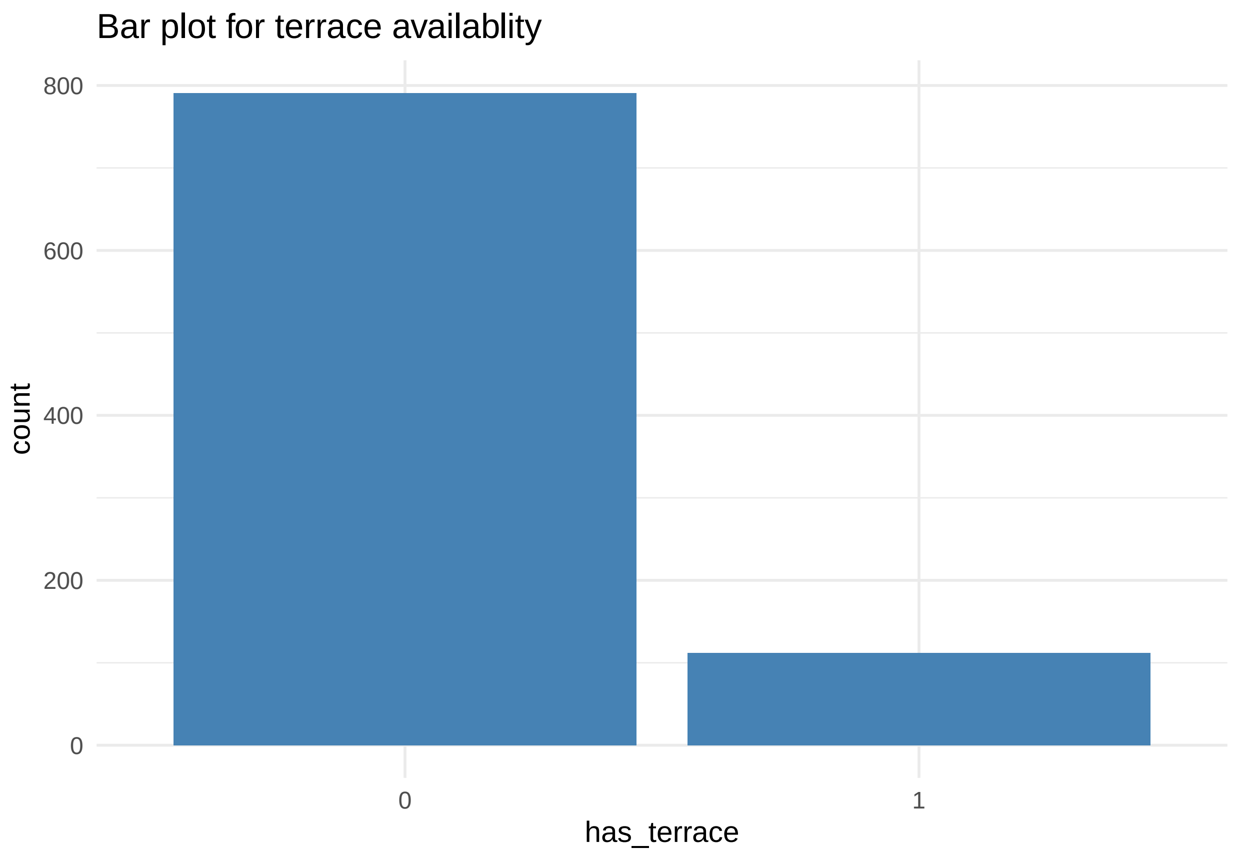
```
ggplot(mydata, aes(x=n_rooms))+ geom_bar( fill="steelblue")+ theme_minimal()+ggtitle("Bar plot of Numbe
```



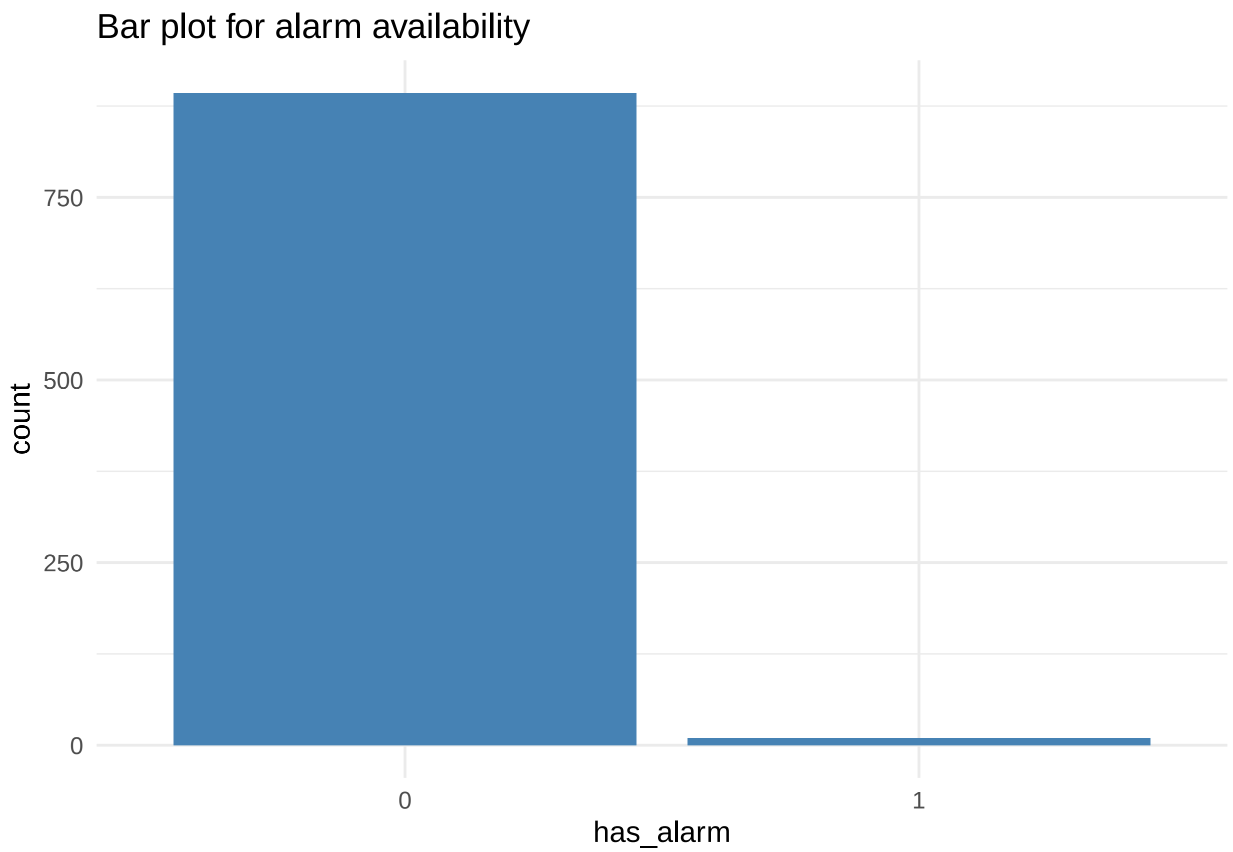
```
ggplot(mydata, aes(x=n_bathrooms))+ geom_bar( fill="steelblue")+ theme_minimal()+ggtitle("Bar plot of Nu
```



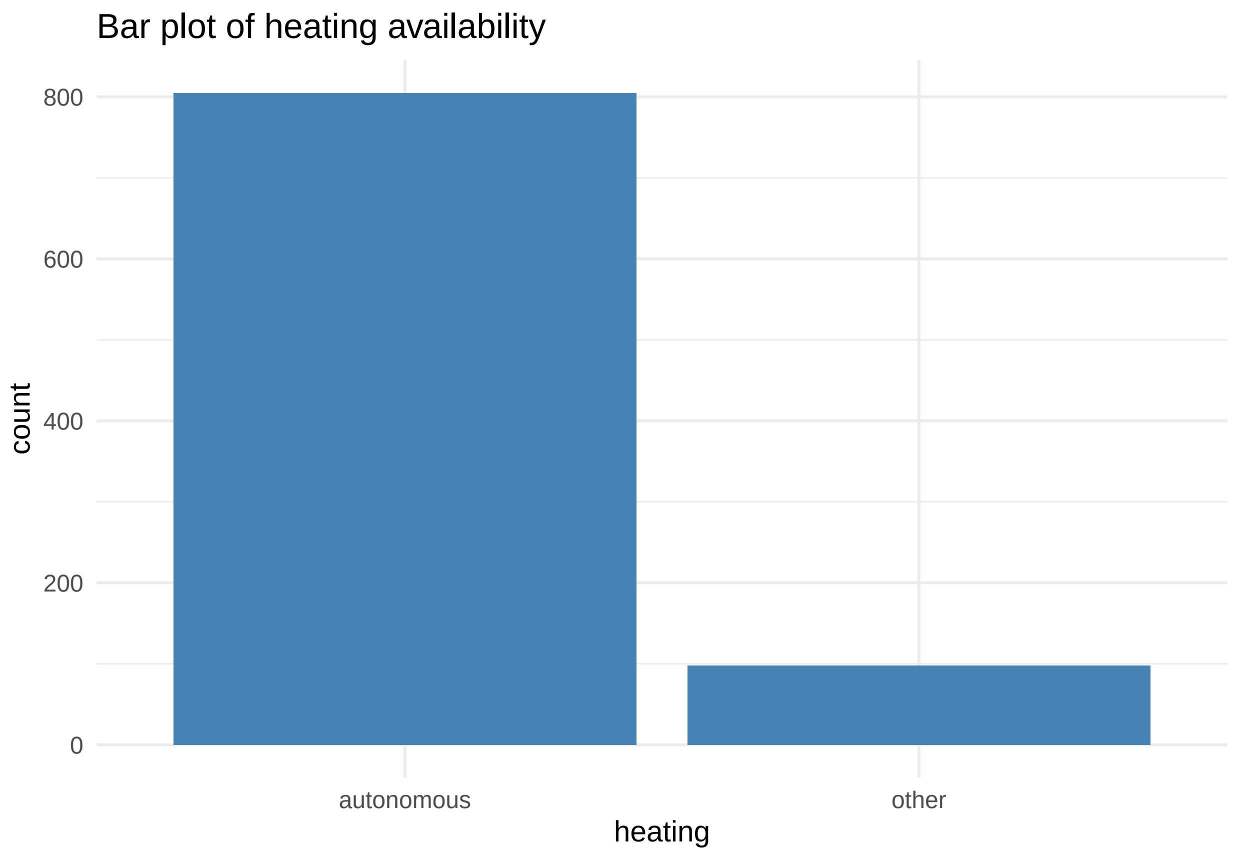
```
ggplot(mydata, aes(x=has_terrace))+ geom_bar( fill="steelblue")+ theme_minimal()+ggtitle("Bar plot for t
```



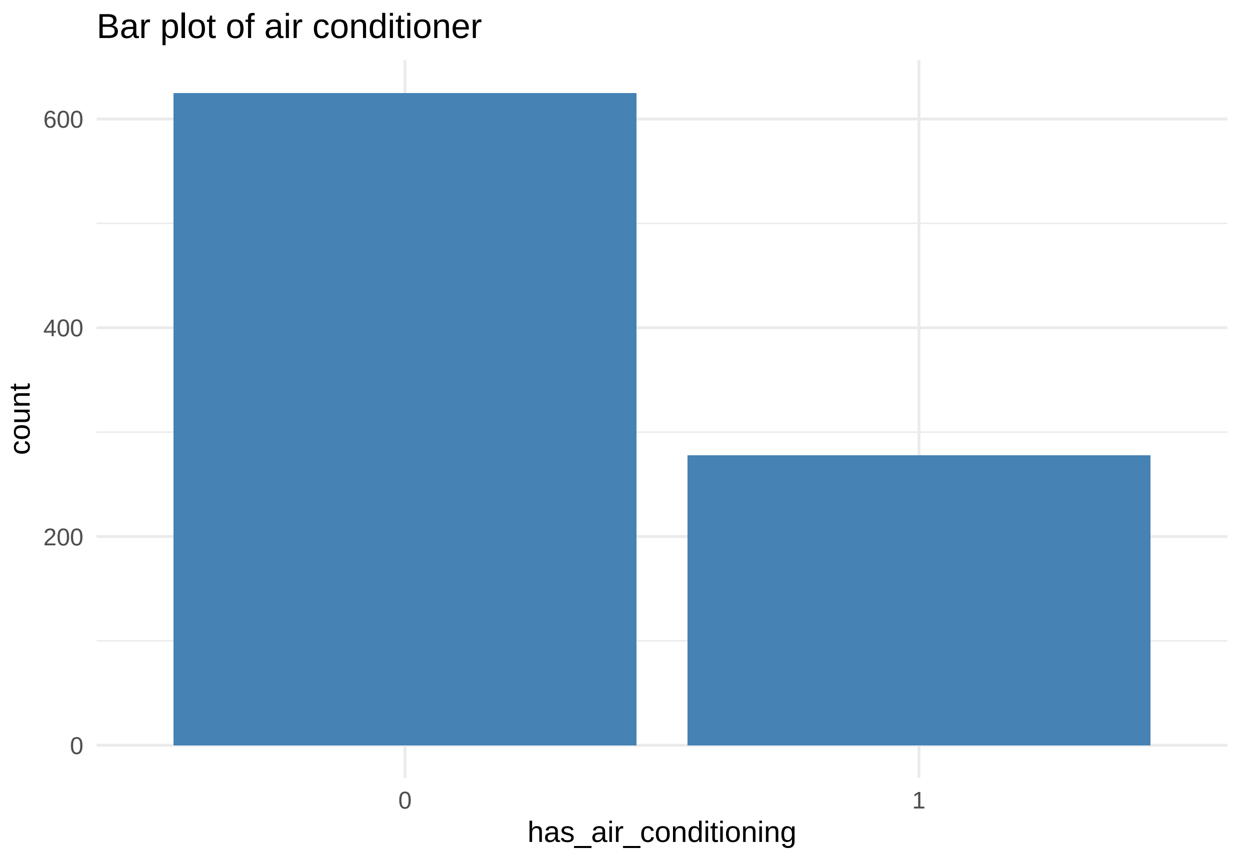
```
ggplot(mydata, aes(x=has_alarm))+ geom_bar( fill="steelblue")+ theme_minimal()+ggtitle("Bar plot for ala
```



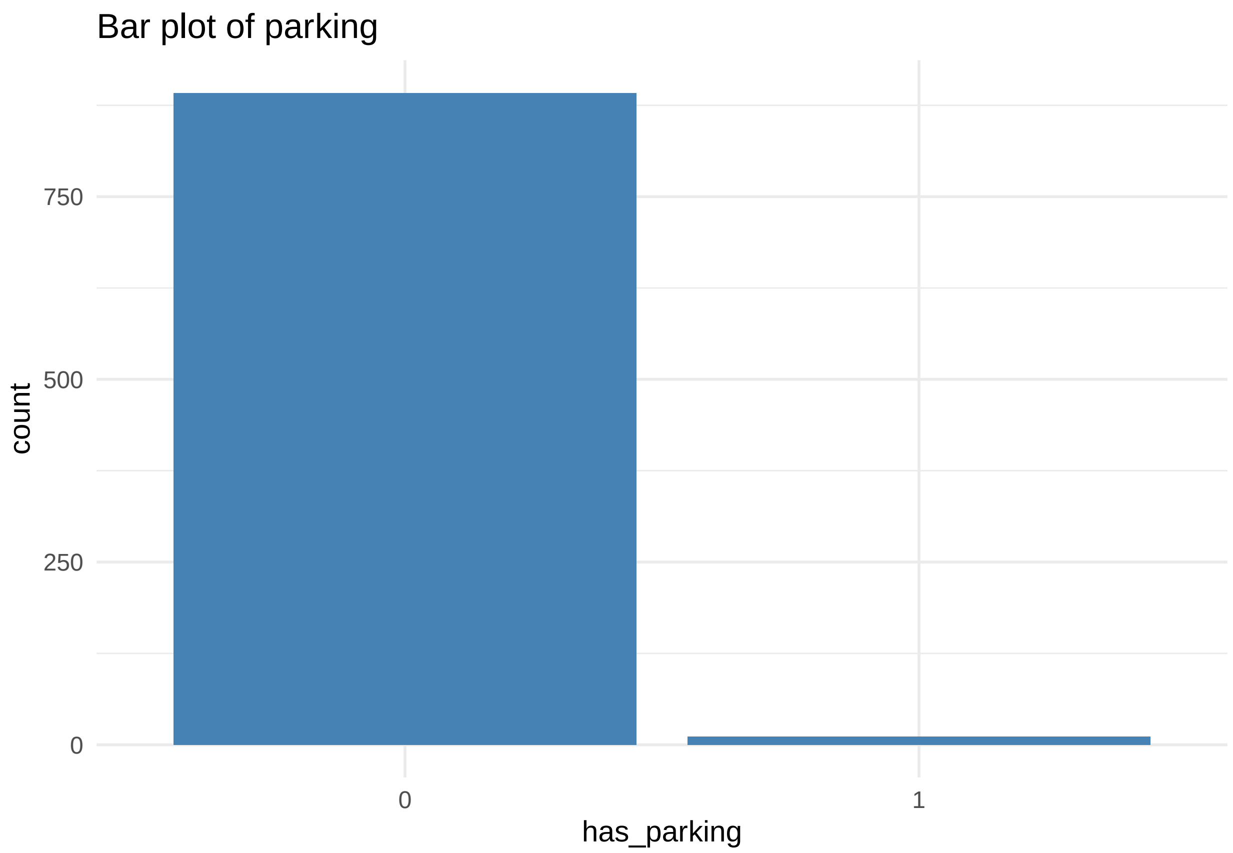
```
ggplot(mydata, aes(x=heating))+ geom_bar( fill="steelblue")+ theme_minimal()+ggtitle("Bar plot of heati
```



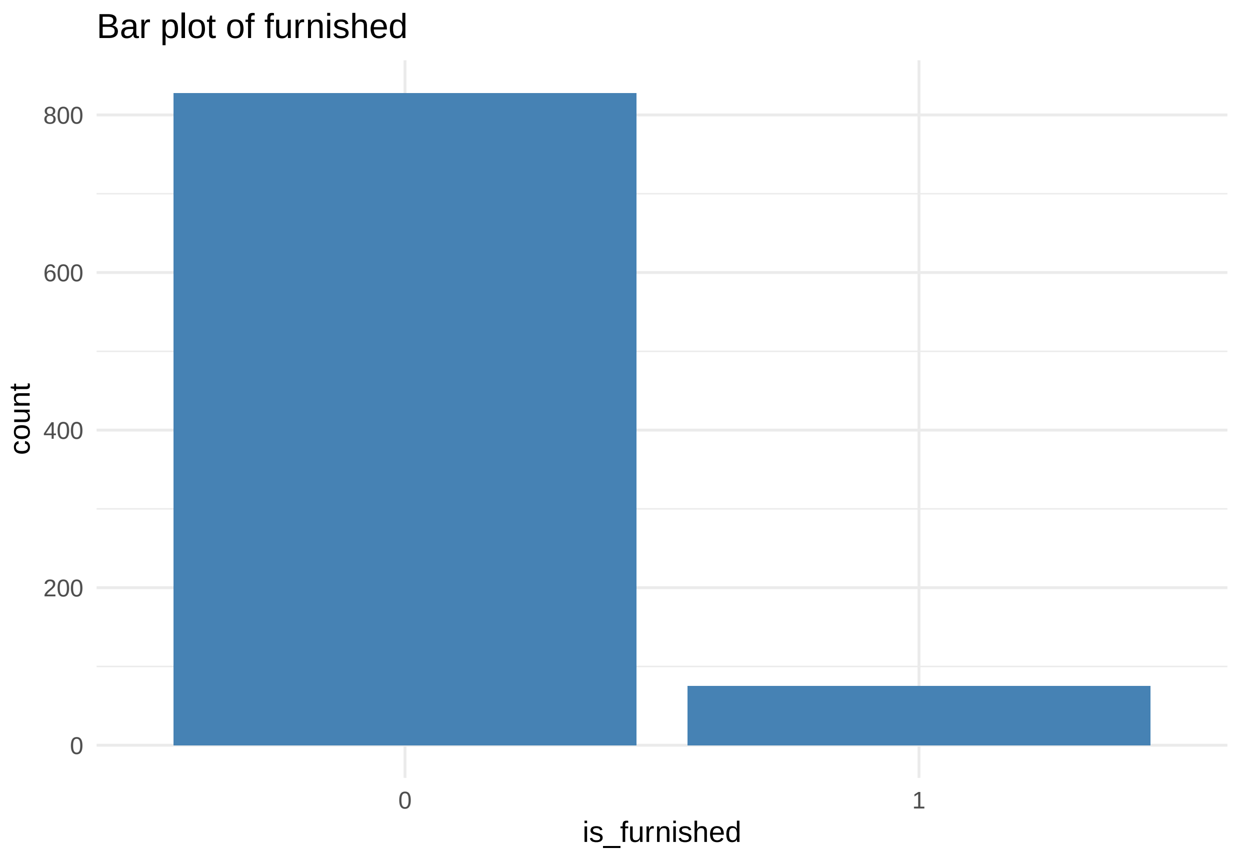
```
ggplot(mydata, aes(x=has_air_conditioning))+ geom_bar( fill="steelblue")+ theme_minimal()+ggtitle("Bar
```



```
ggplot(mydata, aes(x=has_parking))+ geom_bar( fill="steelblue")+ theme_minimal()+ggtitle("Bar plot of pa
```



```
ggplot(mydata, aes(x=is_furnished))+ geom_bar( fill="steelblue")+ theme_minimal()+ggtitle("Bar plot of
```

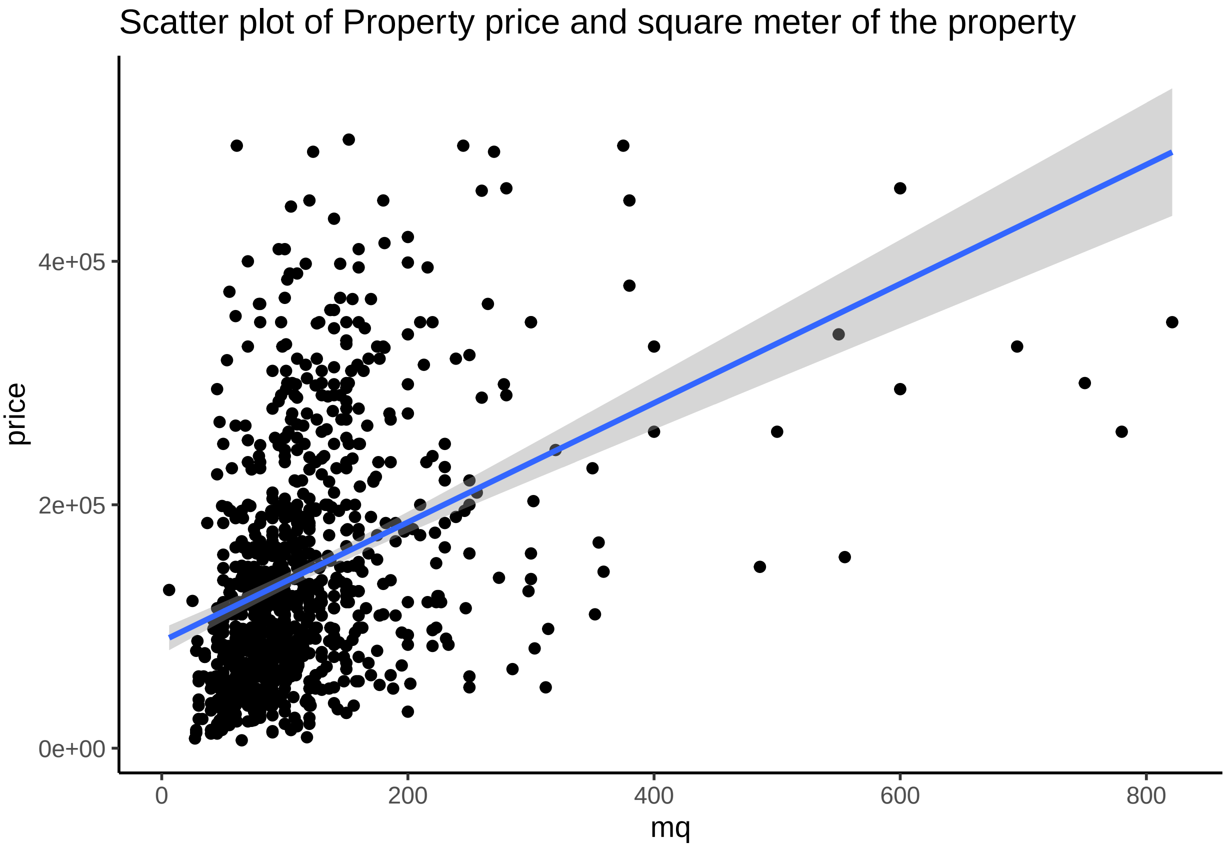


The count is high when there is no alarm,terrace, heating,air conditioner, parking, and furniture.

```
ggplot(mydata, aes(x=mq, y=price)) + geom_point() +  geom_smooth(method=lm)+ ggtitle("Scatter plot of P
```

Exploring the relationship between property price and mq with scattered plot:

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```

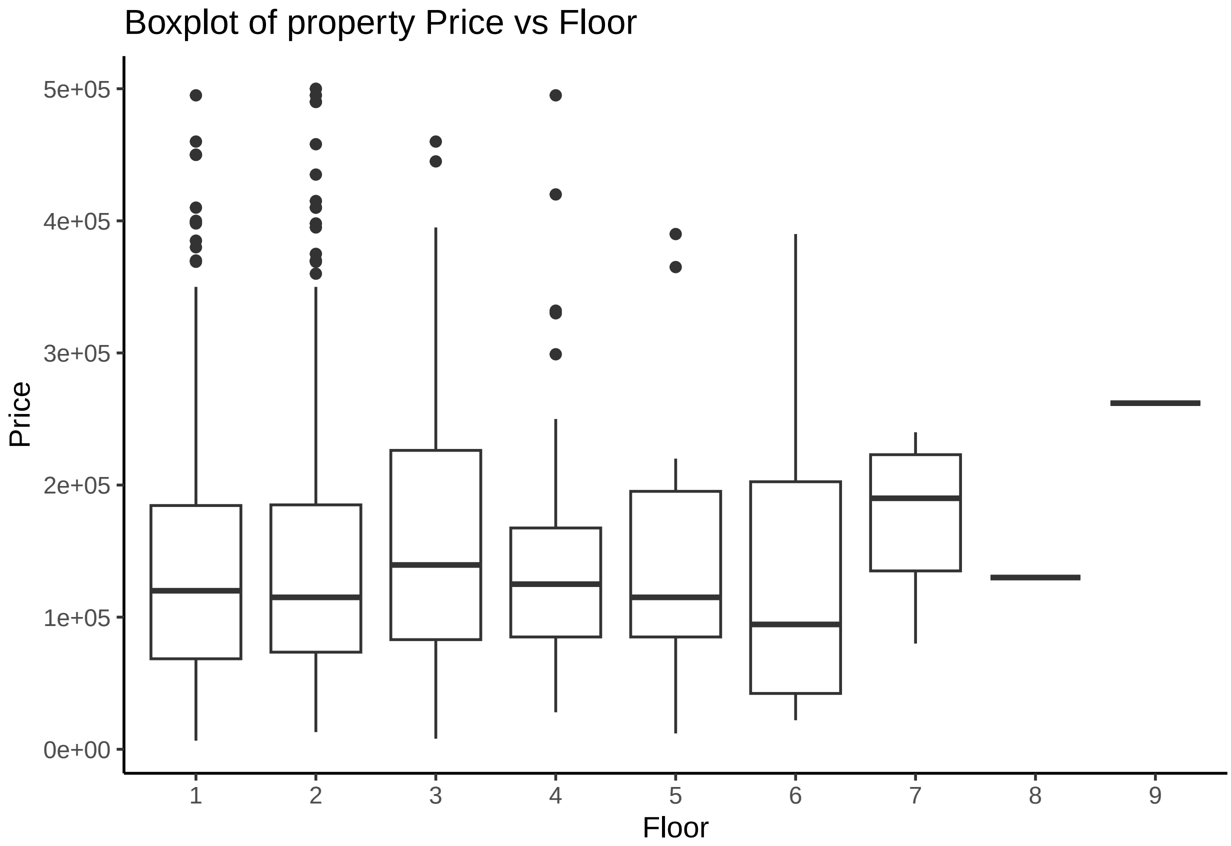


Summary of the findings:

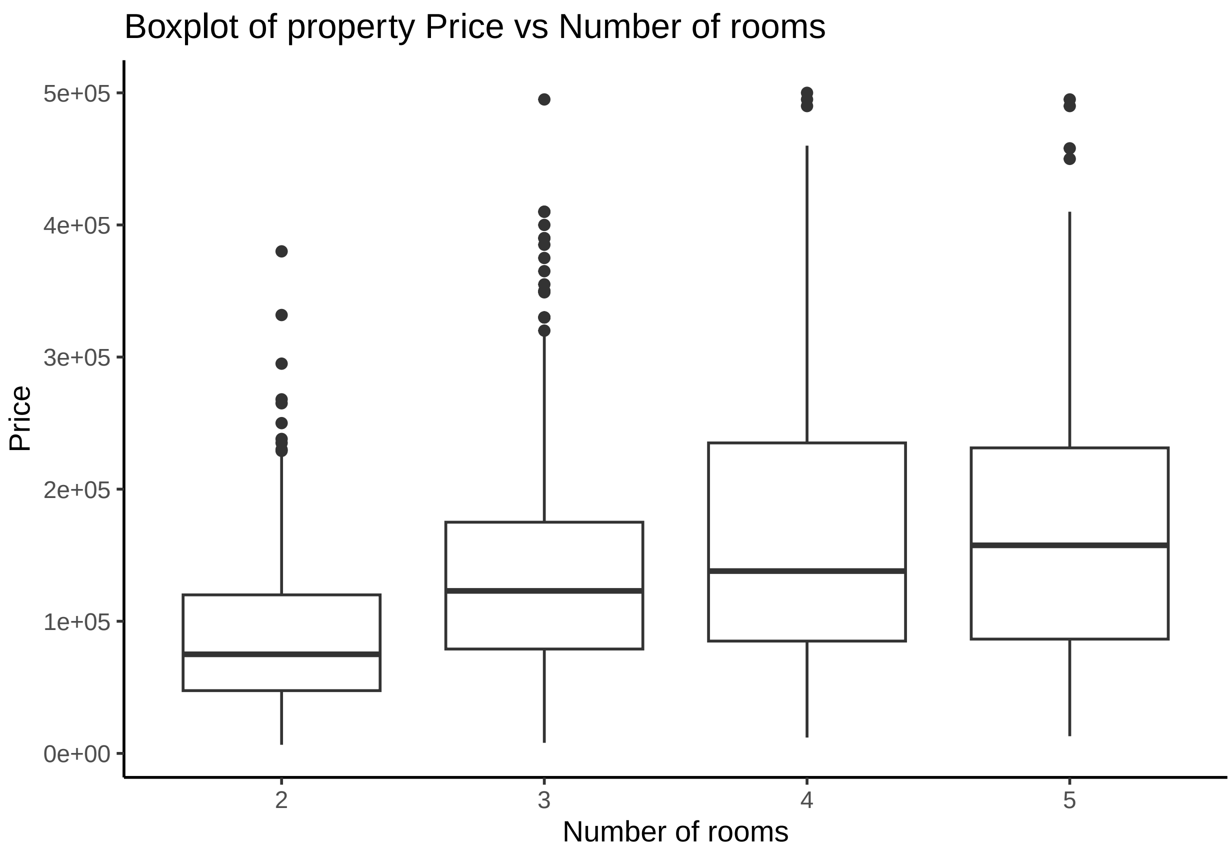
1. The data have positive linear relationship, as the price of the property increases the square meter increases.
2. There are many outliers in the graph.
3. It likely have positive co relation.

```
ggplot(mydata, aes(x=floor, y=price))+geom_boxplot()+theme_classic() + ggtitle("Boxplot of property Pri
```

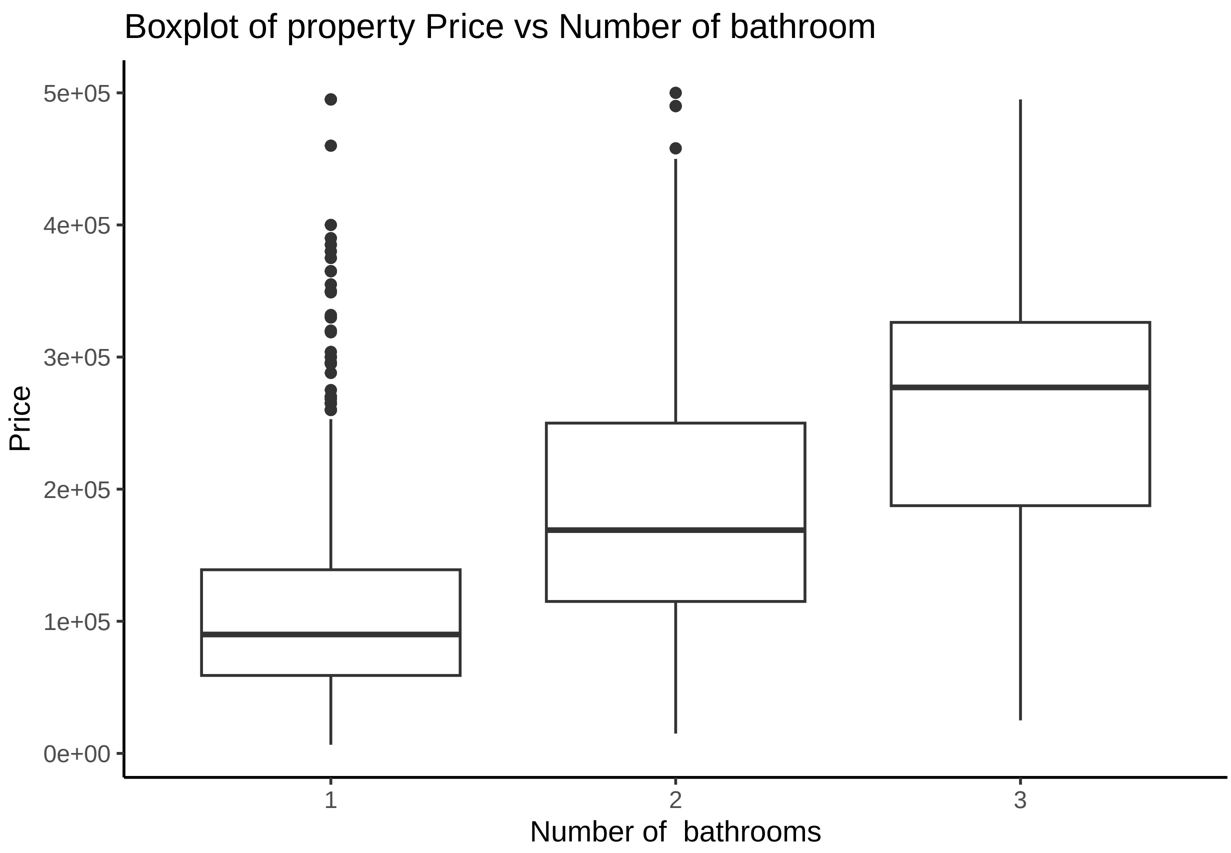
Exploring the relationship between the property price and other categorical variables:



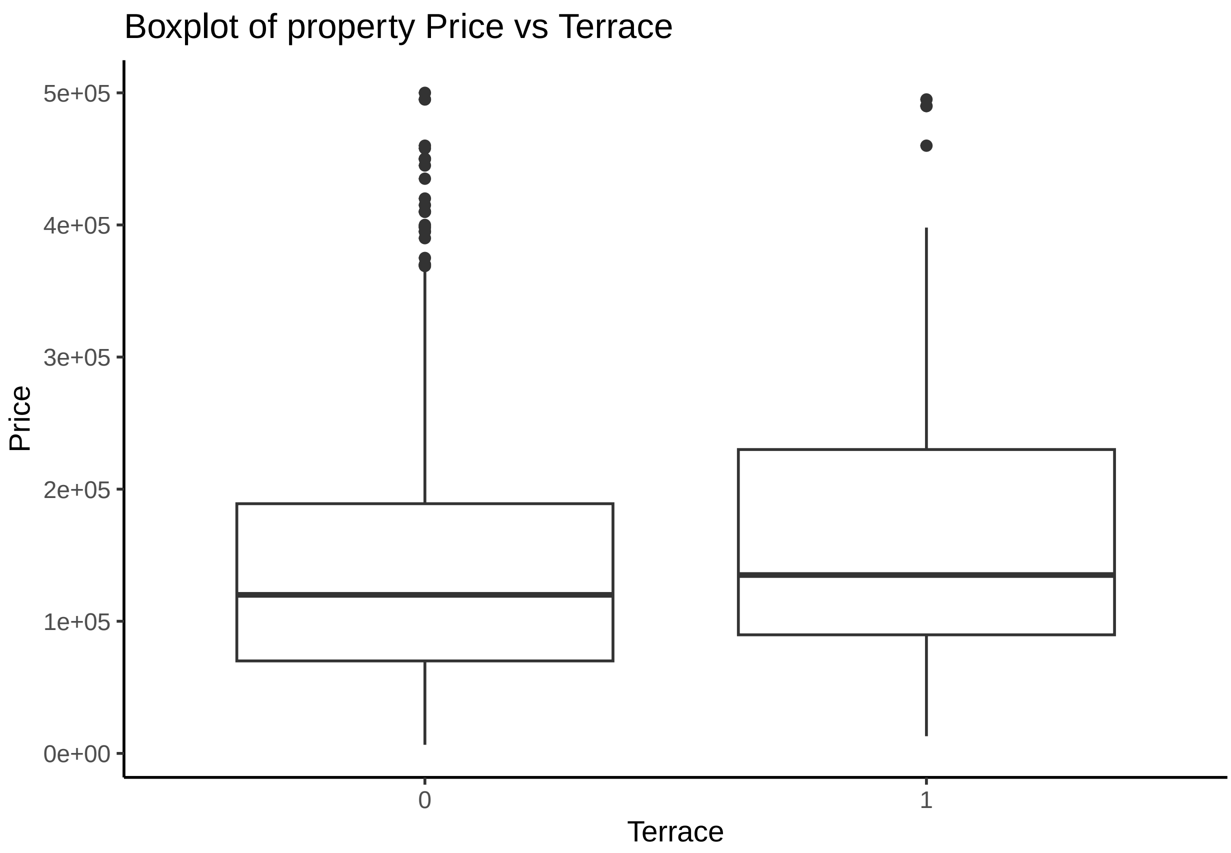
```
ggplot(mydata, aes(x=n_rooms, y=price))+geom_boxplot()+theme_classic() + ggtitle("Boxplot of property P
```



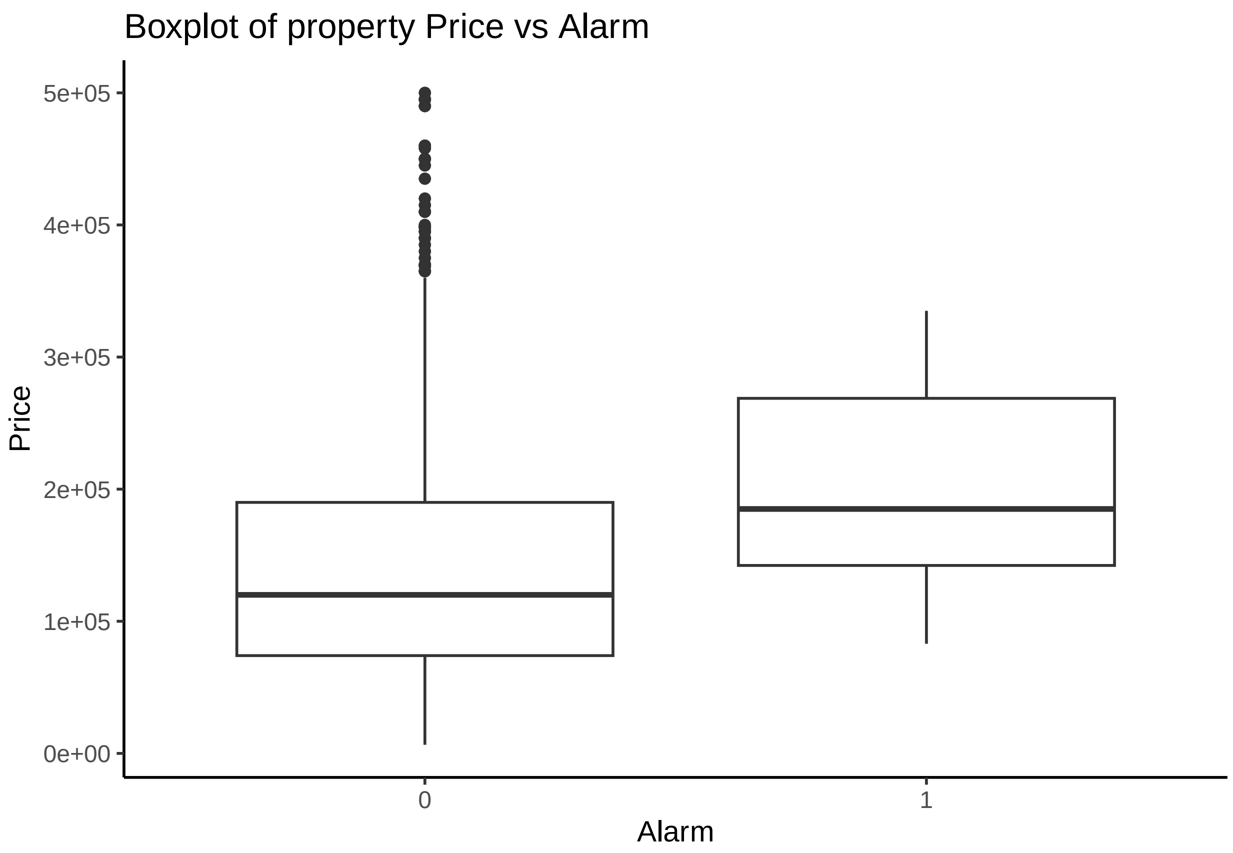
```
ggplot(mydata, aes(x=n_bathrooms, y=price))+geom_boxplot()+theme_classic() + ggtitle("Boxplot of proper
```



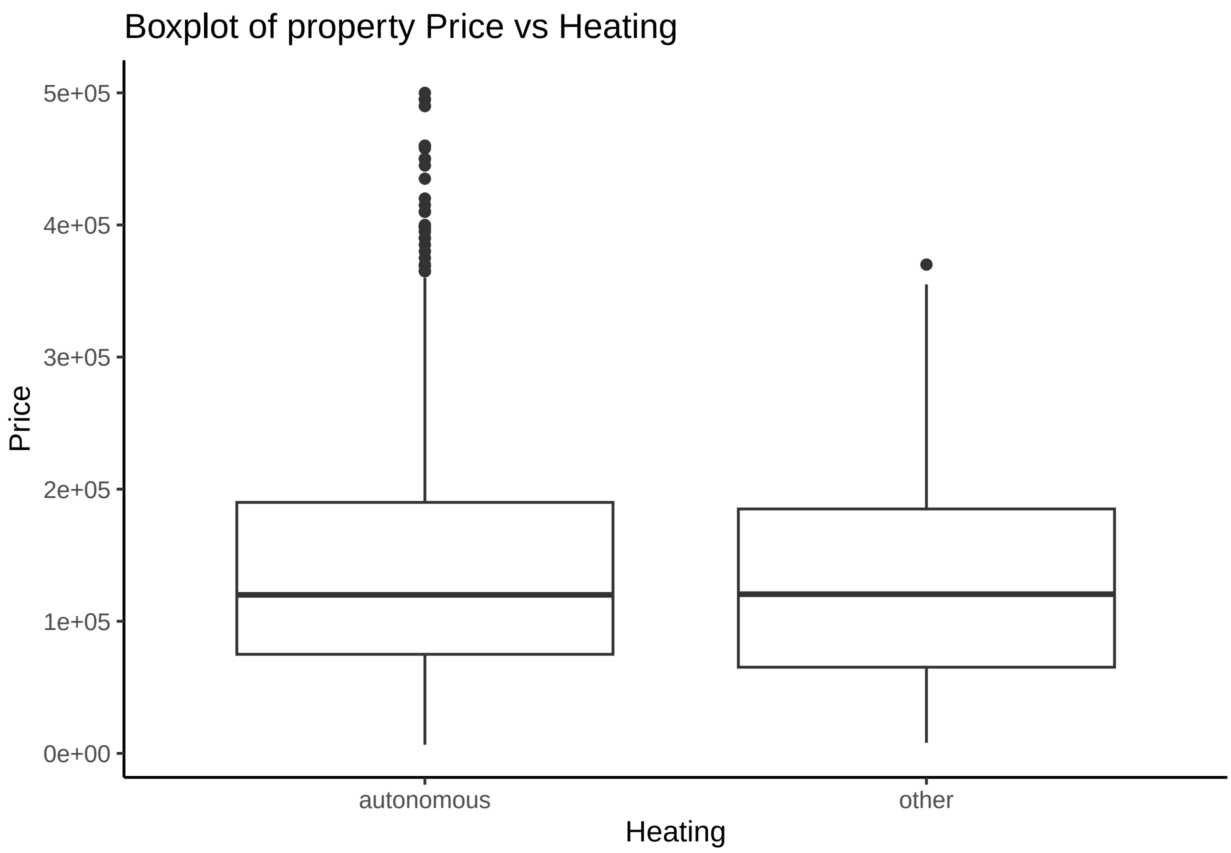
```
ggplot(mydata, aes(x=has_terrace, y=price))+geom_boxplot()+theme_classic() + ggtitle("Boxplot of proper
```



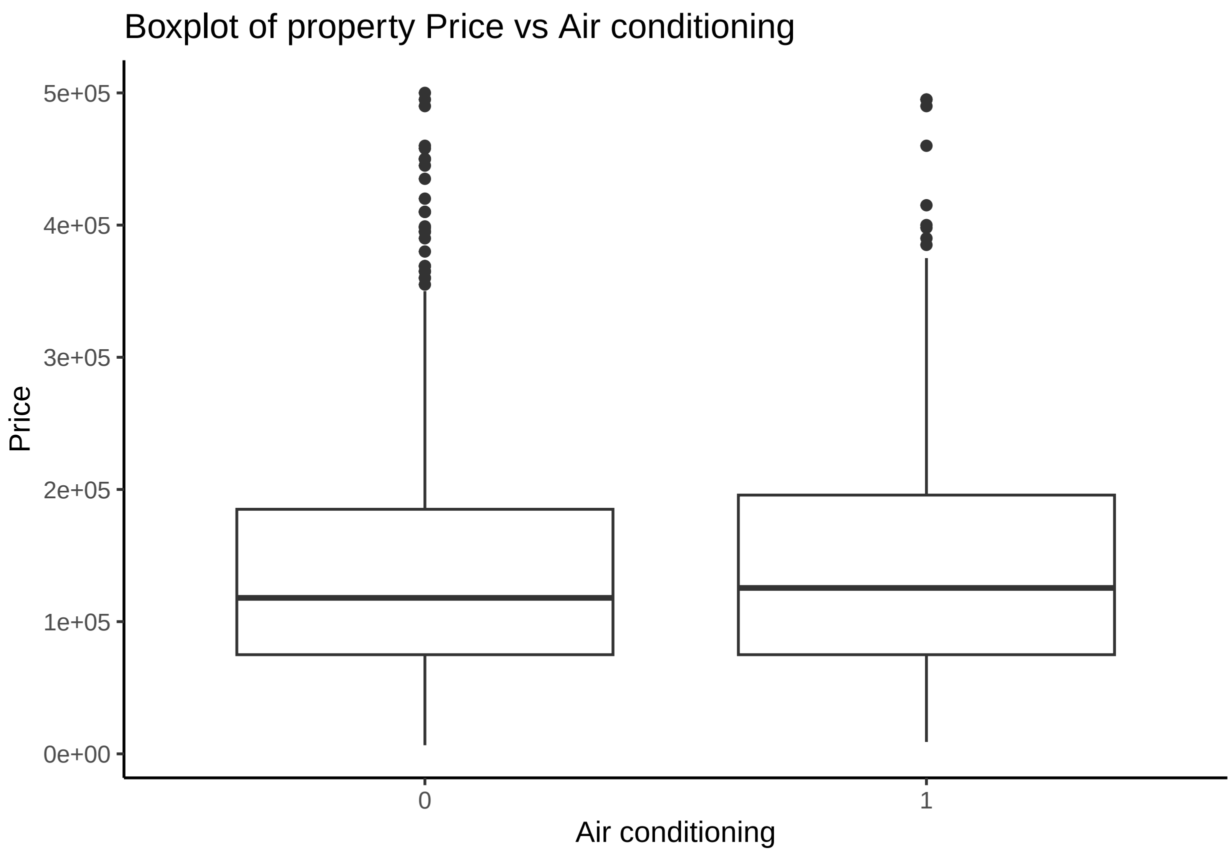
```
ggplot(mydata, aes(x=has_alarm, y=price))+geom_boxplot()+theme_classic() + ggtitle("Boxplot of property
```



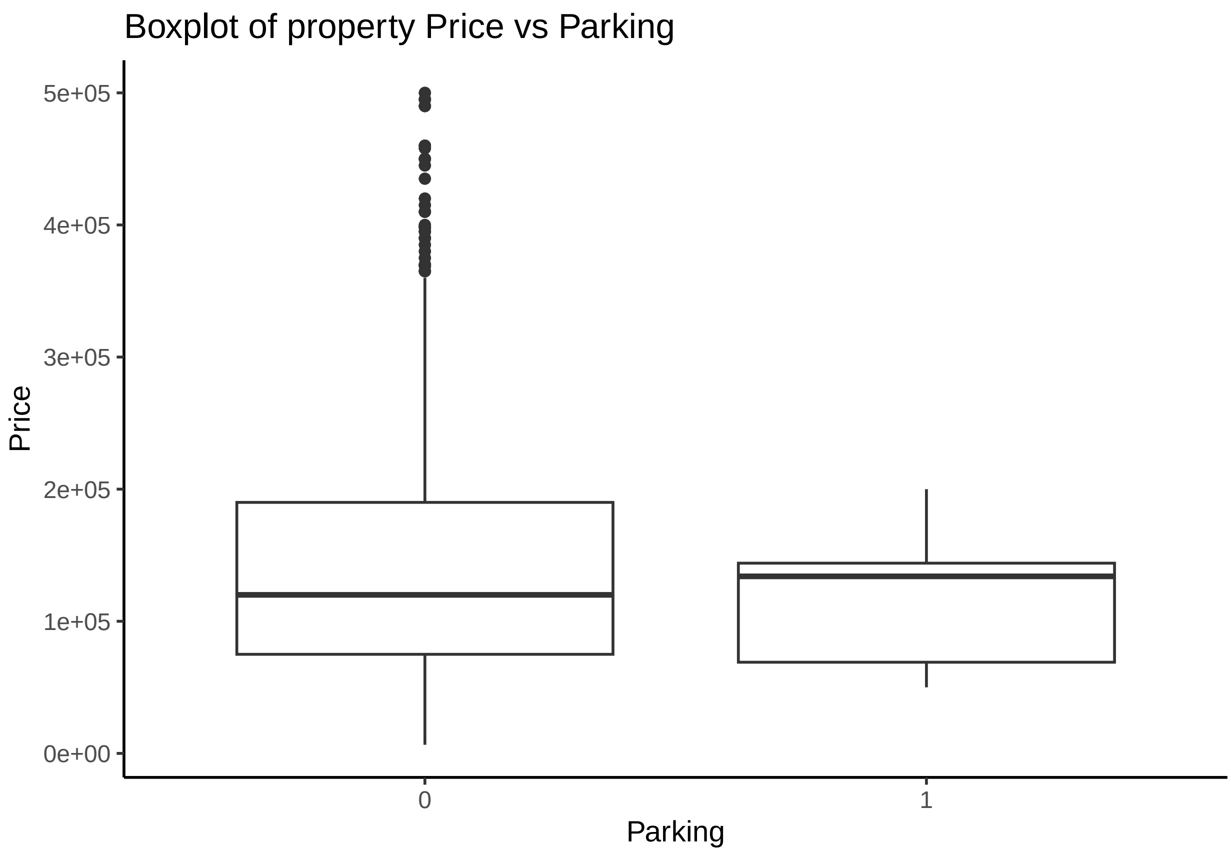
```
ggplot(mydata, aes(x=heating, y=price))+geom_boxplot()+theme_classic() + ggtitle("Boxplot of property P
```



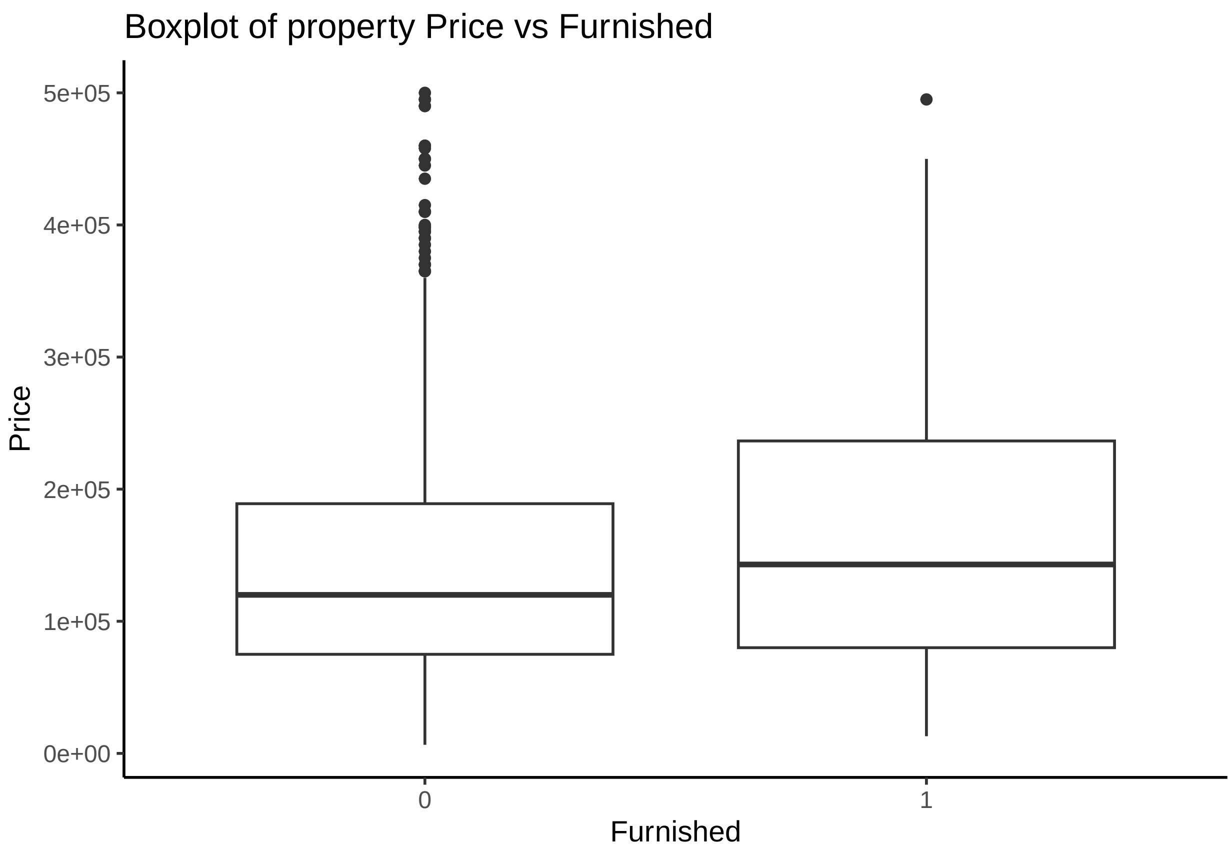
```
ggplot(mydata, aes(x=has_air_conditioning, y=price))+geom_boxplot()+theme_classic() + ggtitle("Boxplot
```



```
ggplot(mydata, aes(x=has_parking, y=price))+geom_boxplot()+theme_classic() + ggtitle("Boxplot of proper
```



```
ggplot(mydata, aes(x=is_furnished, y=price))+geom_boxplot()+theme_classic() + ggtitle("Boxplot of prope
```



Summary of the findings:

1. Box plot of Property Price vs Terrace:

- Median of property price with terrace is slightly higher than the median of property price without terrace, so the price of property is related to terrace presence.
- Both the data have suspicious out liners, which may require a closer look.
- both batches of data appear to be right-skewed.
- The interquartile range is greater for property with terrace.

2. Box plot of Property Price vs Alarm:

- Median of property price with alarm is much greater than the median of property price without alarm, so the price of property is related to alarm presence.
- Property with the alarm has many out-liner, which may require a closer look.
- Both batches of data appear to be right-skewed.
- The interquartile range is slightly greater for property with alarm, though the overall range for the data set is higher for property without alarm.

3. Box plot of property Price vs Heating:

- Median of property price with autonomous heating is same as the median of property price with other heating, so the price of property is related to both.

- Property with the autonomous heating has many out-liner, which may require a closer look.
- Autonomous heating property appear to be right-skewed.
- The interquartile range and overall range for the data set is same for both the type of property.

4. Box plot of property Price vs Air conditioning:

- Median of property price with air conditioning is similar as the median of property price without air conditioning, so the price of property is related to both.
- Both type of properties has many out-liner, which may require a closer look.
- Both type of properties appears to be right-skewed.
- The interquartile range and overall range for the data set is higher for property with air conditioning.

5. Box plot of property Price vs Parking:

- Median of property price with Parking is higher than the median of property price without Parking, so the price of property is related to Parking presence.
- Property without Parking has many out-liner, which may require a closer look.
- Property without Parking appear to be right-skewed and property with Parking is left skewed.
- The interquartile range and overall range are higher for property without Parking.

6. Box plot of property Price vs Furnished:

- Median of property price which furnished is higher than the median of property price which is not furnished, so the price of property is related to with furnished
- Property which is furnished has many out-liner, which may require a closer look.
- Both the property appears to be right-skewed.
- The interquartile range and overall range are higher for property which is furnished.

7. Box plot of Property Price vs number of bathroom:

- Median of property price with three bathroom is much greater than the others, so the price of property is related to property with three bathrooms.
- Property with the one and two bathroom has many out-liner, which may require a closer look.
- Property with one and two bathrooms appear to be right-skewed and property with three bathroom is left skewed.
- The interquartile range of three and two bathroom is same and greater than the range of one bathroom.

8. Box plot of property price vs number of rooms:

- Median of property price with five rooms is much greater than the others, so the price of property is related to property with five rooms.
- property with two and four room appear to be right-skewed.
- The interquartile range of property with four room greater than others.

2.3 Additional insights and issues

Issues found

1. There are many outliers found in the scatter plot between property price and square meter of the property. Outliers in a scatterplot can be a problem because they can distort the overall pattern of the data, making it difficult to accurately interpret the relationship between the variables being plotted.

2. The histogram of property price and mq is right skewed. This can be a problem because it can make it difficult to accurately interpret the distribution of the data.

---

3. Modelling

3.1 Explain your analysis plan

1. The Property price is the dependent variable which is a numerical value and the other independent variables are mix of categorical and numerical.We are implementing multilevel regression model to get the significant values.
2. Used a model selection approach to achieve a minimal adequate model to identify the best model with significant covariance.
3. The ‘mq’ and ‘price’ variable are linearly correlated.

3.2 Build a model for property price

ANCOVA MODEL: We are implementing multilevel regression model to get the significant values.

```
#0.28
ancova<-lm(mydata$price~mydata$mq+mydata$floor+mydata$n_rooms+mydata$n_bathrooms+mydata$has_terrace+myda
summary(ancova)
```

```
##
## Call:
## lm(formula = mydata$price ~ mydata$mq + mydata$floor + mydata$n_rooms +
##     mydata$n_bathrooms + mydata$has_terrace + mydata$has_alarm +
##     mydata$heating + mydata$has_air_conditioning + mydata$has_parking +
##     mydata$is_furnished, data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -201808  -55391  -14948   38405  393877
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   49467.07    8237.41   6.005 2.79e-09 ***
## mydata$mq                       369.35      38.91   9.493  < 2e-16 ***
## mydata$floor2                  8405.32    6455.42   1.302  0.19324
## mydata$floor3                 23368.82    8969.37   2.605  0.00933 **
## mydata$floor4                 19091.99   14259.31   1.339  0.18094
## mydata$floor5                 28715.70   24912.96   1.153  0.24937
## mydata$floor6                 32432.43   42751.40   0.759  0.44828
## mydata$floor7                 19922.91   38492.06   0.518  0.60488
## mydata$floor8                 20668.69   84953.05   0.243  0.80783
## mydata$floor9                 84310.28   84951.33   0.992  0.32125
## mydata$n_rooms3               18005.13    8487.57   2.121  0.03417 *
## mydata$n_rooms4               15542.04    9416.94   1.650  0.09921 .
## mydata$n_rooms5               -2349.58   10783.68  -0.218  0.82757
## mydata$n_bathrooms2           63187.49    6543.76   9.656  < 2e-16 ***
## mydata$n_bathrooms3          116280.53   15446.78   7.528 1.27e-13 ***
## mydata$has_terrace1            8995.08    8747.23   1.028  0.30407
## mydata$has_alarm1             31022.53   27437.26   1.131  0.25850
## mydata$heatingother           11470.04    9444.80   1.214  0.22491
## mydata$has_air_conditioning1  11119.97    6287.87   1.768  0.07733 .
## mydata$has_parking1          -25103.87   26092.23  -0.962  0.33625
## mydata$is_furnished1          14288.29   10367.75   1.378  0.16851
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 84630 on 882 degrees of freedom
## Multiple R-squared:  0.296,  Adjusted R-squared:   0.28
## F-statistic: 18.54 on 20 and 882 DF,  p-value: < 2.2e-16
```

Minimal adequate model: Used a model selection approach to achieve a minimal adequate model.

```
#0.2782
m<-step(ancova)
```

```
## Start:  AIC=20511.7
## mydata$price ~ mydata$mq + mydata$floor + mydata$n_rooms + mydata$n_bathrooms +
##     mydata$has_terrace + mydata$has_alarm + mydata$heating +
##     mydata$has_air_conditioning + mydata$has_parking + mydata$is_furnished
##
##                                Df  Sum of Sq        RSS   AIC
## - mydata$floor                  8 7.1537e+10 6.3885e+12 20506
## - mydata$has_parking            1 6.6298e+09 6.3236e+12 20511
## - mydata$has_terrace            1 7.5738e+09 6.3246e+12 20511
## - mydata$has_alarm              1 9.1562e+09 6.3262e+12 20511
## - mydata$heating                1 1.0563e+10 6.3276e+12 20511
## - mydata$is_furnished           1 1.3603e+10 6.3306e+12 20512
## <none>                                      6.3170e+12 20512
## - mydata$has_air_conditioning  1 2.2400e+10 6.3394e+12 20513
## - mydata$n_rooms                3 6.7719e+10 6.3847e+12 20515
## - mydata$mq                     1 6.4540e+11 6.9624e+12 20598
## - mydata$n_bathrooms            2 8.6998e+11 7.1870e+12 20624
##
## Step:  AIC=20505.86
## mydata$price ~ mydata$mq + mydata$n_rooms + mydata$n_bathrooms +
##     mydata$has_terrace + mydata$has_alarm + mydata$heating +
##     mydata$has_air_conditioning + mydata$has_parking + mydata$is_furnished
##
##                                Df  Sum of Sq        RSS   AIC
## - mydata$has_parking            1 6.7080e+09 6.3953e+12 20505
## - mydata$has_terrace            1 7.8772e+09 6.3964e+12 20505
## - mydata$has_alarm              1 9.4407e+09 6.3980e+12 20505
## - mydata$is_furnished           1 1.2783e+10 6.4013e+12 20506
## <none>                                      6.3885e+12 20506
## - mydata$heating                1 2.1176e+10 6.4097e+12 20507
## - mydata$has_air_conditioning  1 2.4291e+10 6.4128e+12 20507
## - mydata$n_rooms               3 7.3981e+10 6.4625e+12 20510
## - mydata$mq                    1 6.5041e+11 7.0390e+12 20591
## - mydata$n_bathrooms           2 8.6572e+11 7.2543e+12 20617
##
## Step:  AIC=20504.81
## mydata$price ~ mydata$mq + mydata$n_rooms + mydata$n_bathrooms +
##     mydata$has_terrace + mydata$has_alarm + mydata$heating +
##     mydata$has_air_conditioning + mydata$is_furnished
##
##                               Df  Sum of Sq        RSS   AIC
## - mydata$has_terrace           1 7.2837e+09 6.4025e+12 20504
## - mydata$has_alarm             1 8.3672e+09 6.4036e+12 20504
## - mydata$is_furnished          1 1.2785e+10 6.4080e+12 20505
## <none>                                      6.3953e+12 20505
## - mydata$heating               1 2.1432e+10 6.4167e+12 20506
## - mydata$has_air_conditioning  1 2.3349e+10 6.4186e+12 20506
## - mydata$n_rooms               3 7.4968e+10 6.4702e+12 20509
## - mydata$mq                    1 6.5272e+11 7.0480e+12 20591
## - mydata$n_bathrooms           2 8.6049e+11 7.2557e+12 20615
##
## Step:  AIC=20503.84
## mydata$price ~ mydata$mq + mydata$n_rooms + mydata$n_bathrooms +
##     mydata$has_alarm + mydata$heating + mydata$has_air_conditioning +
##     mydata$is_furnished
##
##                               Df  Sum of Sq        RSS   AIC
## - mydata$has_alarm             1 9.4146e+09 6.4120e+12 20503
## - mydata$is_furnished          1 1.3722e+10 6.4163e+12 20504
## <none>                                      6.4025e+12 20504
## - mydata$heating               1 1.9879e+10 6.4224e+12 20505
## - mydata$has_air_conditioning  1 2.7394e+10 6.4299e+12 20506
## - mydata$n_rooms               3 7.7509e+10 6.4800e+12 20509
## - mydata$mq                    1 6.5587e+11 7.0584e+12 20590
## - mydata$n_bathrooms           2 8.6682e+11 7.2694e+12 20615
##
## Step:  AIC=20503.17
## mydata$price ~ mydata$mq + mydata$n_rooms + mydata$n_bathrooms +
##     mydata$heating + mydata$has_air_conditioning + mydata$is_furnished
##
##                               Df  Sum of Sq        RSS   AIC
## <none>                                      6.4120e+12 20503
## - mydata$is_furnished          1 1.4290e+10 6.4262e+12 20503
## - mydata$heating               1 1.9061e+10 6.4310e+12 20504
## - mydata$has_air_conditioning  1 3.2519e+10 6.4445e+12 20506
## - mydata$n_rooms               3 7.6683e+10 6.4886e+12 20508
## - mydata$mq                    1 6.5762e+11 7.0696e+12 20589
## - mydata$n_bathrooms           2 8.7790e+11 7.2899e+12 20615
```

```
summary(m)
```

```
##
## Call:
## lm(formula = mydata$price ~ mydata$mq + mydata$n_rooms + mydata$n_bathrooms +
##     mydata$heating + mydata$has_air_conditioning + mydata$is_furnished,
##     data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -208723  -56285  -15368   37846  384596
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  54886.62    7716.95   7.112 2.34e-12 ***
## mydata$mq                      372.40      38.91   9.570  < 2e-16 ***
## mydata$n_rooms3              19689.53    8444.78   2.332   0.0199 *
## mydata$n_rooms4              18010.20    9338.07   1.929   0.0541 .
## mydata$n_rooms5               -670.52   10717.54  -0.063   0.9501
## mydata$n_bathrooms2          63615.68    6490.63   9.801  < 2e-16 ***
## mydata$n_bathrooms3         113795.47   15368.64   7.404 3.05e-13 ***
## mydata$heatingother          15018.08    9217.55   1.629   0.1036
## mydata$has_air_conditioning1  13111.39    6160.97   2.128   0.0336 *
## mydata$is_furnished1          14597.97   10347.83   1.411   0.1587
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 84740 on 893 degrees of freedom
## Multiple R-squared:  0.2854, Adjusted R-squared:  0.2782
## F-statistic: 39.62 on 9 and 893 DF,  p-value: < 2.2e-16
```

The step function has ended with this minimal adequate model:

> mydata$price\ mydata$mq + mydata$n_rooms + mydata$n_bathrooms + mydata$heating + mydata$has_air_conditioning + mydata$is_furnished

Summary of the findings:

1. the F statistic is significant but the r^2 is very low.
2. The covariance is_furnished1,heatingother and n_rooms5 are not significant.
3. From the summary function we can see that there is a weak negative relationship between price and n_rooms5. This is reflected in the value of the estimate for the effect of n_rooms5 which is -670.52.

Exploring the interaction between the variables: The Total square meters of the property 'mq' is related to the number of room 'n_room' and number of bath rooms 'n_bathrooms'. We are proceeding with the interaction method to check the best fit for the model.

```
# 0.2947
an<-lm(mydata$price~mydata$floor+mydata$mq*mydata$n_rooms*mydata$n_bathrooms+mydata$has_terrace+mydata$h
summary(an)
```

```
##
## Call:
## lm(formula = mydata$price ~ mydata$floor + mydata$mq * mydata$n_rooms *
##     mydata$n_bathrooms + mydata$has_terrace + mydata$has_alarm +
##     mydata$heating + mydata$has_air_conditioning + mydata$has_parking +
##     mydata$is_furnished, data = mydata)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -199015  -54972  -12783   37954  386737
##
## Coefficients: (3 not defined because of singularities)
##                                                 Estimate Std. Error t value
## (Intercept)                                     54528.58   11359.63   4.800
## mydata$floor2                                    8819.62    6408.86   1.376
## mydata$floor3                                   24026.45    8961.86   2.681
## mydata$floor4                                   18327.00   14141.29   1.296
## mydata$floor5                                   39026.45   24829.00   1.572
## mydata$floor6                                   35107.69   42353.40   0.829
## mydata$floor7                                   12412.22   38374.40   0.323
## mydata$floor8                                   34086.02   84199.21   0.405
## mydata$floor9                                   71490.01   84237.06   0.849
## mydata$mq                                         353.16     123.44   2.861
## mydata$n_rooms3                                 27356.13   14304.80   1.912
## mydata$n_rooms4                                -37661.85   19882.07  -1.894
## mydata$n_rooms5                                  5402.28   20071.23   0.269
## mydata$n_bathrooms2                             21201.01   50301.23   0.421
## mydata$n_bathrooms3                             68198.06   81958.22   0.832
## mydata$has_terrace1                             10236.41    8707.85   1.176
## mydata$has_alarm1                               24146.17   27223.16   0.887
## mydata$heatingother                             12218.40    9414.83   1.298
## mydata$has_air_conditioning1                    11285.53    6285.01   1.796
## mydata$has_parking1                            -26541.32   25890.94  -1.025
## mydata$is_furnished1                            14814.88   10317.24   1.436
## mydata$mq:mydata$n_rooms3                        -105.75     145.82  -0.725
## mydata$mq:mydata$n_rooms4                         305.56     178.71   1.710
## mydata$mq:mydata$n_rooms5                         -30.94     146.25  -0.212
## mydata$mq:mydata$n_bathrooms2                    -233.08     306.52  -0.760
## mydata$mq:mydata$n_bathrooms3                     241.40     383.82   0.629
## mydata$n_rooms3:mydata$n_bathrooms2            -14415.32   56403.01  -0.256
## mydata$n_rooms4:mydata$n_bathrooms2            105662.22   57285.32   1.844
## mydata$n_rooms5:mydata$n_bathrooms2             15586.56   56962.81   0.274
## mydata$n_rooms3:mydata$n_bathrooms3           -180250.51   96385.66  -1.870
## mydata$n_rooms4:mydata$n_bathrooms3             95145.93   91468.07   1.040
## mydata$n_rooms5:mydata$n_bathrooms3                   NA         NA      NA
## mydata$mq:mydata$n_rooms3:mydata$n_bathrooms2     707.17     382.27   1.850
## mydata$mq:mydata$n_rooms4:mydata$n_bathrooms2     -76.54     359.86  -0.213
## mydata$mq:mydata$n_rooms5:mydata$n_bathrooms2     334.55     335.19   0.998
## mydata$mq:mydata$n_rooms3:mydata$n_bathrooms3         NA         NA      NA
## mydata$mq:mydata$n_rooms4:mydata$n_bathrooms3    -492.51     438.46  -1.123
## mydata$mq:mydata$n_rooms5:mydata$n_bathrooms3         NA         NA      NA
##                                               Pr(>|t|)
## (Intercept)                                   1.87e-06 ***
## mydata$floor2                                  0.16913
## mydata$floor3                                  0.00748 **
## mydata$floor4                                  0.19532
## mydata$floor5                                  0.11636
## mydata$floor6                                  0.40738
## mydata$floor7                                  0.74643
## mydata$floor8                                  0.68571
## mydata$floor9                                  0.39630
## mydata$mq                                      0.00433 **
## mydata$n_rooms3                                0.05616 .
## mydata$n_rooms4                                0.05852 .
## mydata$n_rooms5                                0.78787
## mydata$n_bathrooms2                            0.67351
## mydata$n_bathrooms3                            0.40558
## mydata$has_terrace1                            0.24010
## mydata$has_alarm1                              0.37534
## mydata$heatingother                            0.19471
## mydata$has_air_conditioning1                   0.07290 .
## mydata$has_parking1                            0.30559
## mydata$is_furnished1                           0.15138
## mydata$mq:mydata$n_rooms3                      0.46852
## mydata$mq:mydata$n_rooms4                      0.08766 .
## mydata$mq:mydata$n_rooms5                      0.83251
## mydata$mq:mydata$n_bathrooms2                  0.44721
## mydata$mq:mydata$n_bathrooms3                  0.52955
## mydata$n_rooms3:mydata$n_bathrooms2            0.79834
## mydata$n_rooms4:mydata$n_bathrooms2            0.06545 .
## mydata$n_rooms5:mydata$n_bathrooms2            0.78444
## mydata$n_rooms3:mydata$n_bathrooms3            0.06181 .
## mydata$n_rooms4:mydata$n_bathrooms3            0.29853
## mydata$n_rooms5:mydata$n_bathrooms3                 NA
## mydata$mq:mydata$n_rooms3:mydata$n_bathrooms2  0.06466 .
## mydata$mq:mydata$n_rooms4:mydata$n_bathrooms2  0.83162
## mydata$mq:mydata$n_rooms5:mydata$n_bathrooms2  0.31851
## mydata$mq:mydata$n_rooms3:mydata$n_bathrooms3       NA
## mydata$mq:mydata$n_rooms4:mydata$n_bathrooms3  0.26163
## mydata$mq:mydata$n_rooms5:mydata$n_bathrooms3       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 83760 on 868 degrees of freedom
## Multiple R-squared:  0.3213, Adjusted R-squared:  0.2947
## F-statistic: 12.09 on 34 and 868 DF,  p-value: < 2.2e-16
```

Minimal adequate model for interactions:

```
#0.2905
r<-step(an)
```

```
## Start:  AIC=20506.61
## mydata$price ~ mydata$floor + mydata$mq * mydata$n_rooms * mydata$n_bathrooms +
##     mydata$has_terrace + mydata$has_alarm + mydata$heating +
##     mydata$has_air_conditioning + mydata$has_parking + mydata$is_furnished
##
##                                               Df  Sum of Sq        RSS   AIC
## - mydata$floor                                 8 7.7591e+10 6.1673e+12 20502
## - mydata$has_alarm                             1 5.5195e+09 6.0952e+12 20505
## - mydata$has_parking                           1 7.3727e+09 6.0971e+12 20506
## - mydata$has_terrace                           1 9.6951e+09 6.0994e+12 20506
## - mydata$heating                               1 1.1816e+10 6.1015e+12 20506
## <none>                                                      6.0897e+12 20507
## - mydata$is_furnished                          1 1.4466e+10 6.1042e+12 20507
## - mydata$mq:mydata$n_rooms:mydata$n_bathrooms  4 5.9757e+10 6.1495e+12 20507
## - mydata$has_air_conditioning                  1 2.2621e+10 6.1123e+12 20508
##
## Step:  AIC=20502.04
## mydata$price ~ mydata$mq + mydata$n_rooms + mydata$n_bathrooms +
##     mydata$has_terrace + mydata$has_alarm + mydata$heating +
##     mydata$has_air_conditioning + mydata$has_parking + mydata$is_furnished +
##     mydata$mq:mydata$n_rooms + mydata$mq:mydata$n_bathrooms +
##     mydata$n_rooms:mydata$n_bathrooms + mydata$mq:mydata$n_rooms:mydata$n_bathrooms
##
##                                                Df  Sum of Sq        RSS   AIC
## - mydata$has_alarm                             1 5.5973e+09 6.1729e+12 20501
## - mydata$has_parking                           1 7.4519e+09 6.1748e+12 20501
## - mydata$has_terrace                           1 1.0002e+10 6.1773e+12 20502
## - mydata$is_furnished                          1 1.3259e+10 6.1806e+12 20502
## <none>                                                      6.1673e+12 20502
## - mydata$mq:mydata$n_rooms:mydata$n_bathrooms  4 5.5312e+10 6.2226e+12 20502
## - mydata$heating                               1 2.2069e+10 6.1894e+12 20503
## - mydata$has_air_conditioning                  1 2.4687e+10 6.1920e+12 20504
##
## Step:  AIC=20500.86
## mydata$price ~ mydata$mq + mydata$n_rooms + mydata$n_bathrooms +
##     mydata$has_terrace + mydata$heating + mydata$has_air_conditioning +
##     mydata$has_parking + mydata$is_furnished + mydata$mq:mydata$n_rooms +
##     mydata$mq:mydata$n_bathrooms + mydata$n_rooms:mydata$n_bathrooms +
##     mydata$mq:mydata$n_rooms:mydata$n_bathrooms
##
##                                                Df  Sum of Sq        RSS   AIC
## - mydata$has_parking                           1 6.5603e+09 6.1795e+12 20500
## - mydata$has_terrace                           1 1.0991e+10 6.1839e+12 20501
## - mydata$is_furnished                          1 1.3646e+10 6.1866e+12 20501
## <none>                                                      6.1729e+12 20501
## - mydata$mq:mydata$n_rooms:mydata$n_bathrooms  4 5.6509e+10 6.2294e+12 20501
## - mydata$heating                               1 2.1449e+10 6.1944e+12 20502
## - mydata$has_air_conditioning                  1 2.7975e+10 6.2009e+12 20503
##
## Step:  AIC=20499.82
## mydata$price ~ mydata$mq + mydata$n_rooms + mydata$n_bathrooms +
##     mydata$has_terrace + mydata$heating + mydata$has_air_conditioning +
##     mydata$is_furnished + mydata$mq:mydata$n_rooms + mydata$mq:mydata$n_bathrooms +
##     mydata$n_rooms:mydata$n_bathrooms + mydata$mq:mydata$n_rooms:mydata$n_bathrooms
##
##                                                Df  Sum of Sq        RSS   AIC
## - mydata$has_terrace                           1 1.0283e+10 6.1898e+12 20499
## - mydata$is_furnished                          1 1.3659e+10 6.1931e+12 20500
## <none>                                                      6.1795e+12 20500
## - mydata$mq:mydata$n_rooms:mydata$n_bathrooms  4 5.6281e+10 6.2358e+12 20500
## - mydata$heating                               1 2.1680e+10 6.2012e+12 20501
## - mydata$has_air_conditioning                  1 2.6670e+10 6.2061e+12 20502
##
## Step:  AIC=20499.32
## mydata$price ~ mydata$mq + mydata$n_rooms + mydata$n_bathrooms +
##     mydata$heating + mydata$has_air_conditioning + mydata$is_furnished +
##     mydata$mq:mydata$n_rooms + mydata$mq:mydata$n_bathrooms +
##     mydata$n_rooms:mydata$n_bathrooms + mydata$mq:mydata$n_rooms:mydata$n_bathrooms
##
##                                                Df  Sum of Sq        RSS   AIC
## - mydata$mq:mydata$n_rooms:mydata$n_bathrooms  4 5.4635e+10 6.2444e+12 20499
## <none>                                                      6.1898e+12 20499
## - mydata$is_furnished                          1 1.4586e+10 6.2043e+12 20499
## - mydata$heating                               1 1.9801e+10 6.2096e+12 20500
## - mydata$has_air_conditioning                  1 3.2156e+10 6.2219e+12 20502
##
## Step:  AIC=20499.25
## mydata$price ~ mydata$mq + mydata$n_rooms + mydata$n_bathrooms +
##     mydata$heating + mydata$has_air_conditioning + mydata$is_furnished +
##     mydata$mq:mydata$n_rooms + mydata$mq:mydata$n_bathrooms +
##     mydata$n_rooms:mydata$n_bathrooms
##
##                                      Df  Sum of Sq        RSS   AIC
## - mydata$mq:mydata$n_rooms           3 1.3357e+10 6.2577e+12 20495
## - mydata$mq:mydata$n_bathrooms       2 5.4415e+08 6.2449e+12 20495
## - mydata$is_furnished                1 1.3840e+10 6.2582e+12 20499
## <none>                                            6.2444e+12 20499
## - mydata$heating                     1 1.8212e+10 6.2626e+12 20500
## - mydata$has_air_conditioning        1 3.0123e+10 6.2745e+12 20502
## - mydata$n_rooms:mydata$n_bathrooms  5 1.2635e+11 6.3707e+12 20507
##
## Step:  AIC=20495.18
## mydata$price ~ mydata$mq + mydata$n_rooms + mydata$n_bathrooms +
##     mydata$heating + mydata$has_air_conditioning + mydata$is_furnished +
##     mydata$mq:mydata$n_bathrooms + mydata$n_rooms:mydata$n_bathrooms
##
##                                      Df  Sum of Sq        RSS   AIC
## - mydata$mq:mydata$n_bathrooms       2 3.5531e+09 6.2613e+12 20492
## - mydata$is_furnished                1 1.3281e+10 6.2710e+12 20495
## <none>                                            6.2577e+12 20495
## - mydata$heating                     1 1.7276e+10 6.2750e+12 20496
## - mydata$has_air_conditioning        1 2.9112e+10 6.2869e+12 20497
## - mydata$n_rooms:mydata$n_bathrooms  5 1.4371e+11 6.4015e+12 20506
##
## Step:  AIC=20491.7
## mydata$price ~ mydata$mq + mydata$n_rooms + mydata$n_bathrooms +
##     mydata$heating + mydata$has_air_conditioning + mydata$is_furnished +
##     mydata$n_rooms:mydata$n_bathrooms
##
##                                      Df  Sum of Sq        RSS   AIC
## - mydata$is_furnished                1 1.2628e+10 6.2739e+12 20492
## <none>                                            6.2613e+12 20492
## - mydata$heating                     1 1.7713e+10 6.2790e+12 20492
## - mydata$has_air_conditioning        1 2.8284e+10 6.2896e+12 20494
## - mydata$n_rooms:mydata$n_bathrooms  5 1.5065e+11 6.4120e+12 20503
## - mydata$mq                          1 6.4073e+11 6.9020e+12 20578
##
## Step:  AIC=20491.52
## mydata$price ~ mydata$mq + mydata$n_rooms + mydata$n_bathrooms +
##     mydata$heating + mydata$has_air_conditioning + mydata$n_rooms:mydata$n_bathrooms
##
##                                    Df  Sum of Sq        RSS   AIC
## <none>                                           6.2739e+12 20492
## - mydata$heating                    1 1.7733e+10 6.2917e+12 20492
## - mydata$has_air_conditioning       1 3.3266e+10 6.3072e+12 20494
## - mydata$n_rooms:mydata$n_bathrooms  5 1.5232e+11 6.4262e+12 20503
## - mydata$mq                         1 6.4408e+11 6.9180e+12 20578
```

```
summary(r)
```

```
##
## Call:
## lm(formula = mydata$price ~ mydata$mq + mydata$n_rooms + mydata$n_bathrooms +
##     mydata$heating + mydata$has_air_conditioning + mydata$n_rooms:mydata$n_bathrooms,
##     data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -196103  -53558  -14368   35677  379847
##
## Coefficients: (1 not defined because of singularities)
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        60366.13    7788.95   7.750 2.50e-14 ***
## mydata$mq                            371.00      38.83   9.553  < 2e-16 ***
## mydata$n_rooms3                    18921.38    9014.71   2.099 0.036103 *
## mydata$n_rooms4                     -432.25   10529.53  -0.041 0.967264
## mydata$n_rooms5                      237.94   13764.03   0.017 0.986211
## mydata$n_bathrooms2               -14651.58   29014.77  -0.505 0.613705
## mydata$n_bathrooms3               113879.83   24558.57   4.637 4.06e-06 ***
## mydata$heatingother                14520.36    9160.09   1.585 0.113282
## mydata$has_air_conditioning1       13234.87    6095.92   2.171 0.030187 *
## mydata$n_rooms3:mydata$n_bathrooms2   68934.22   30646.16   2.249 0.024734 *
## mydata$n_rooms4:mydata$n_bathrooms2  106448.34   30919.73   3.443 0.000603 ***
## mydata$n_rooms5:mydata$n_bathrooms2   70441.61   32291.65   2.181 0.029414 *
## mydata$n_rooms3:mydata$n_bathrooms3 -213502.04   87867.70  -2.430 0.015304 *
## mydata$n_rooms4:mydata$n_bathrooms3   20349.06   31889.64   0.638 0.523567
## mydata$n_rooms5:mydata$n_bathrooms3         NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 84010 on 889 degrees of freedom
## Multiple R-squared:  0.3008, Adjusted R-squared:  0.2905
## F-statistic: 29.41 on 13 and 889 DF,  p-value: < 2.2e-16
```

The step function has ended with this minimal adequate model in interactions:

> mydata*price mydata*mq + mydatan$_r$*ooms* + *mydata*n_bathrooms + mydata*heating* + *mydata*has_air_conditioning + mydatan$_r$*ooms* : *mydata*n_bathrooms

Summary of the findings:

1. The r^2 is low = 0.2905, but significant enough to prove this model is good fit.

2. F is significant and p value = 2.2e-16.
3. Most of the variables are significant.

3.3 Critique model using relevant diagnostics

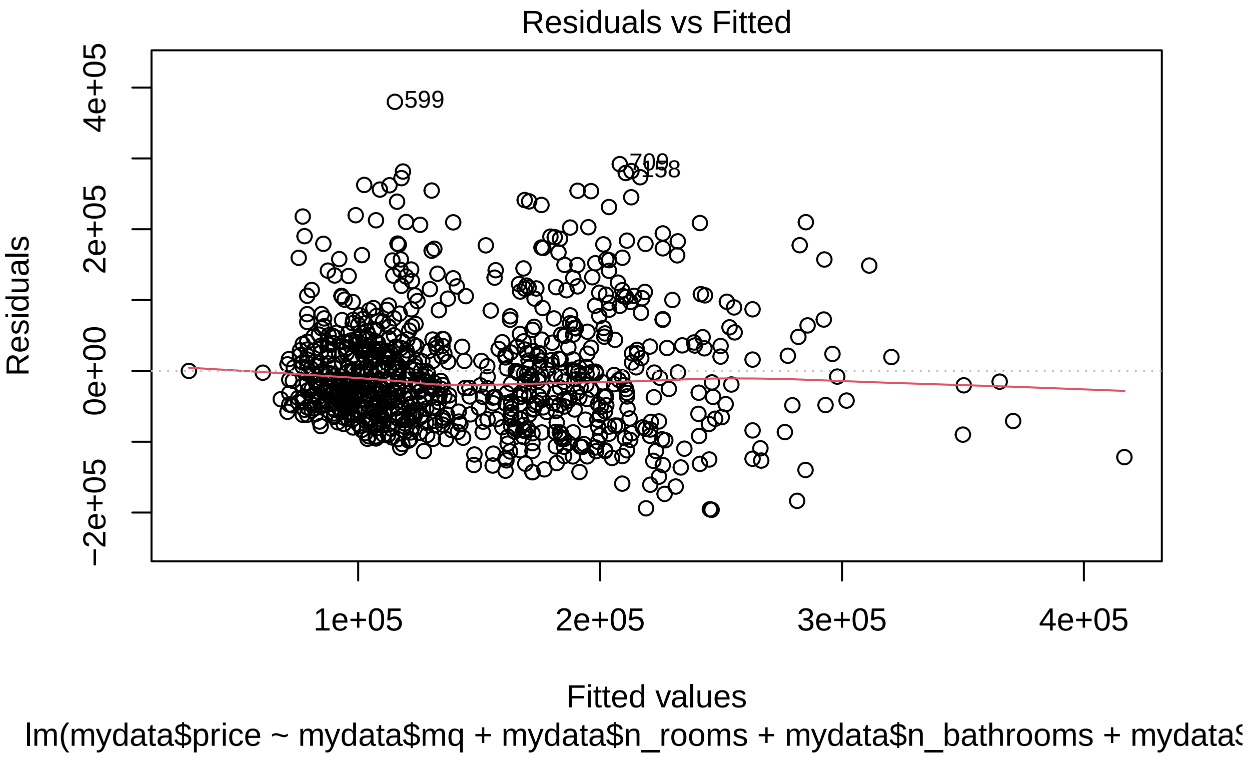
Summary of the findings in minimal adequate model in interactions:

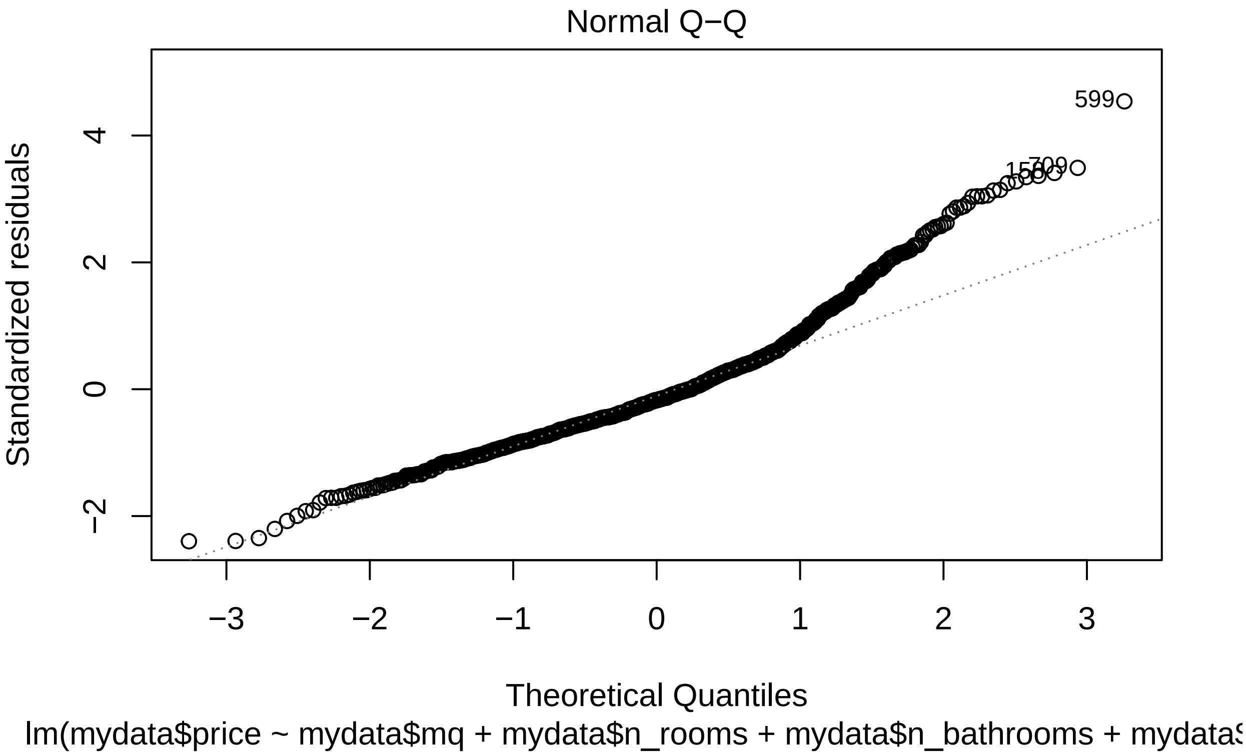
1. The F statistics is good, p value is significant and r^2 is high compared to other models, which indicates the goodness of regression model.
2. From the summary function we can see that there is a negative relationship between price and n_rooms4. This is reflected in the value of the estimate for the effect of n_rooms4 which is -432.25.
3. There is a strong negative relationship between price and n_bathrooms2 = -14651.58.

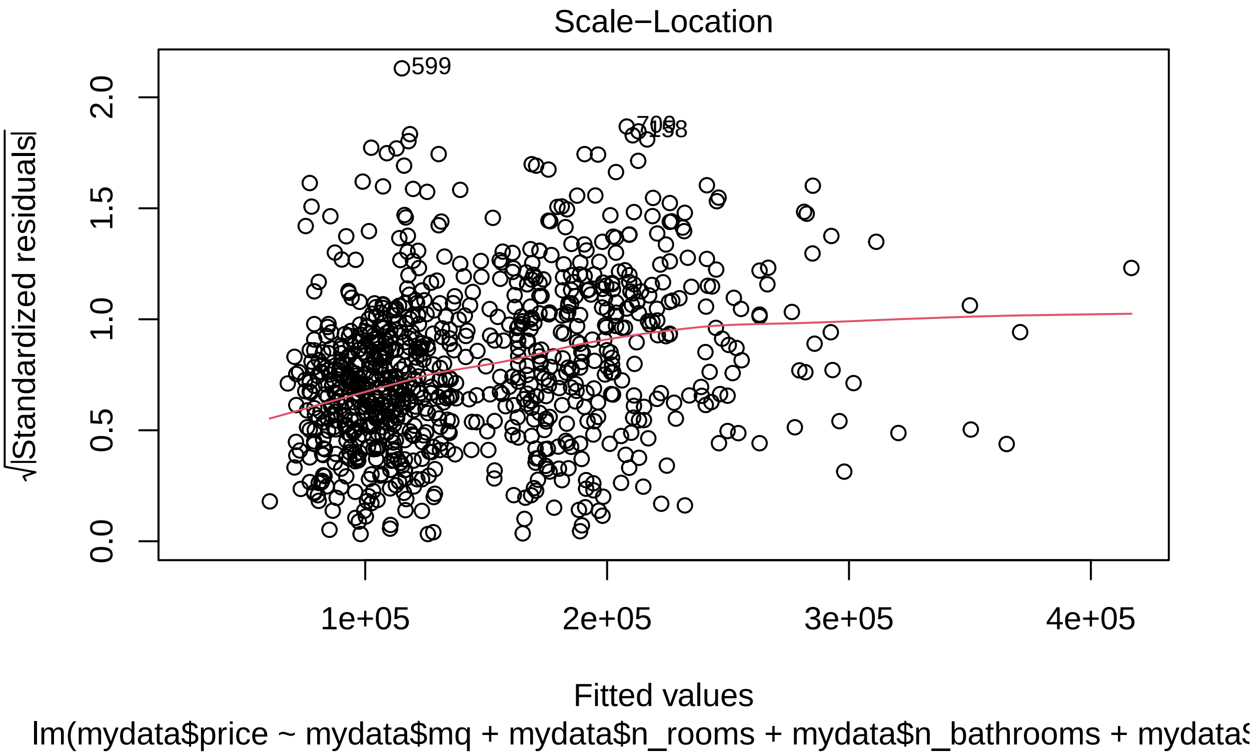
and check its residuals are obtained using:

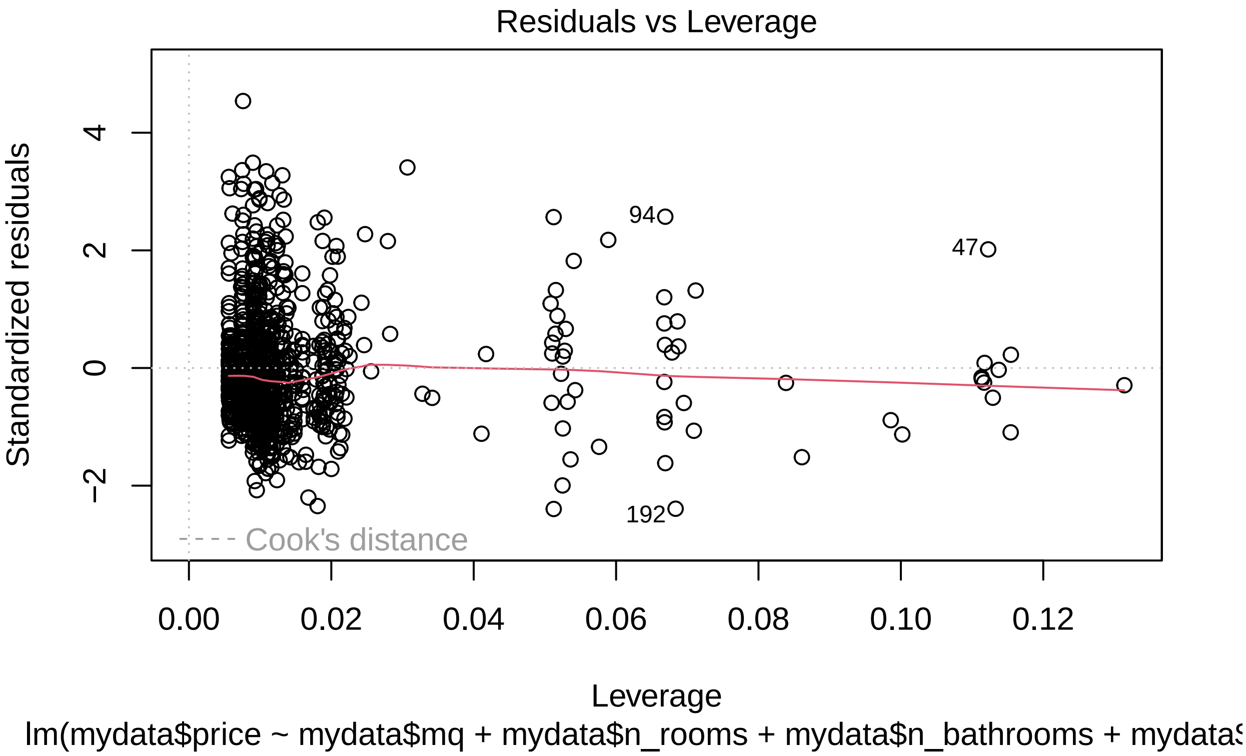
```
plot(r)
```

```
## Warning: not plotting observations with leverage one:
##   288
```



Normal Q−Q
599
1709
15
4
2
0
−2
Standardized residuals
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(mydata$price ~ mydata$mq + mydata$n_rooms + mydata$n_bathrooms + mydata$


Scale−Location
599
709
158
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
1e+05
2e+05
3e+05
4e+05
Fitted values
lm(mydata$price ~ mydata$mq + mydata$n_rooms + mydata$n_bathrooms + mydata$



1. All of the four residual diagnostic plots are looking better.
2. The diagnostics for this model do not point to major issues, but there are some outliers in QQ plot (158, 709,599) that can be considered for further investigation.
3. No heterosedacity present in this model

3.4 Suggest improvements to your model

From the above plots of the data there is reason to assume that some polynomial relation is possible.

```
pol<-lm(formula= mydata$price~poly(mydata$mq,2)+mydata$floor+mydata$n_rooms+mydata$n_bathrooms+mydata$ha
summary(pol)
```

```
##
## Call:
## lm(formula = mydata$price ~ poly(mydata$mq, 2) + mydata$floor +
##     mydata$n_rooms + mydata$n_bathrooms + mydata$has_terrace +
##     mydata$has_alarm + mydata$heating + mydata$has_air_conditioning +
##     mydata$has_parking + mydata$is_furnished, data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -203346  -54796  -13511   38111  396735
##
## Coefficients:
```

```
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                      98398       8354  11.779  < 2e-16 ***
## poly(mydata$mq, 2)1             945584      97329   9.715  < 2e-16 ***
## poly(mydata$mq, 2)2            -200002      98870  -2.023  0.04339 *  
## mydata$floor2                     8100       6446   1.257  0.20922    
## mydata$floor3                    23377       8954   2.611  0.00918 ** 
## mydata$floor4                    16988      14272   1.190  0.23426    
## mydata$floor5                    29357      24871   1.180  0.23818    
## mydata$floor6                    33496      42680   0.785  0.43277    
## mydata$floor7                    17394      38445   0.452  0.65107    
## mydata$floor8                    18838      84809   0.222  0.82428    
## mydata$floor9                    84554      84803   0.997  0.31901    
## mydata$n_rooms3                  15093       8594   1.756  0.07940 .  
## mydata$n_rooms4                   9406       9878   0.952  0.34121    
## mydata$n_rooms5                  -9184      11283  -0.814  0.41585    
## mydata$n_bathrooms2              59514       6780   8.778  < 2e-16 ***
## mydata$n_bathrooms3             110405      15691   7.036 3.97e-12 ***
## mydata$has_terrace1               9094       8732   1.041  0.29795    
## mydata$has_alarm1                29106      27406   1.062  0.28850    
## mydata$heatingother              12439       9440   1.318  0.18797    
## mydata$has_air_conditioning1     11333       6278   1.805  0.07137 .  
## mydata$has_parking1             -24261      26050  -0.931  0.35193    
## mydata$is_furnished1             15096      10357   1.457  0.14534    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 84480 on 881 degrees of freedom
## Multiple R-squared:  0.2992, Adjusted R-squared:  0.2825 
## F-statistic: 17.91 on 21 and 881 DF,  p-value: < 2.2e-16
```

Minimal adequate model:

Used a model selection approach to achieve a minimal adequate model.

```
pol1<-step(pol)
```

```
## Start:  AIC=20509.51
## mydata$price ~ poly(mydata$mq, 2) + mydata$floor + mydata$n_rooms + 
##     mydata$n_bathrooms + mydata$has_terrace + mydata$has_alarm + 
##     mydata$heating + mydata$has_air_conditioning + mydata$has_parking + 
##     mydata$is_furnished
## 
##                                Df  Sum of Sq        RSS   AIC
## - mydata$floor                  8 6.9913e+10 6.3577e+12 20504
## - mydata$has_parking            1 6.1907e+09 6.2940e+12 20508
## - mydata$has_terrace            1 7.7410e+09 6.2955e+12 20509
## - mydata$has_alarm              1 8.0504e+09 6.2959e+12 20509
## - mydata$heating                1 1.2391e+10 6.3002e+12 20509
## <none>                                       6.2878e+12 20510
## - mydata$is_furnished           1 1.5161e+10 6.3030e+12 20510
## - mydata$has_air_conditioning   1 2.3260e+10 6.3111e+12 20511
## - mydata$n_rooms                3 6.8859e+10 6.3567e+12 20513
## - poly(mydata$mq, 2)            2 6.7461e+11 6.9624e+12 20598
## - mydata$n_bathrooms            2 7.1080e+11 6.9986e+12 20602
```

```
##
## Step:  AIC=20503.5
## mydata$price ~ poly(mydata$mq, 2) + mydata$n_rooms + mydata$n_bathrooms +
##     mydata$has_terrace + mydata$has_alarm + mydata$heating +
##     mydata$has_air_conditioning + mydata$has_parking + mydata$is_furnished
##
##                               Df  Sum of Sq        RSS   AIC
## - mydata$has_parking           1 6.3942e+09 6.3641e+12 20502
## - mydata$has_terrace           1 8.1173e+09 6.3658e+12 20503
## - mydata$has_alarm             1 8.2069e+09 6.3659e+12 20503
## <none>                                      6.3577e+12 20504
## - mydata$is_furnished          1 1.4224e+10 6.3719e+12 20504
## - mydata$heating               1 2.3588e+10 6.3813e+12 20505
## - mydata$has_air_conditioning  1 2.5312e+10 6.3830e+12 20505
## - mydata$n_rooms               3 7.4461e+10 6.4322e+12 20508
## - poly(mydata$mq, 2)           2 6.8123e+11 7.0390e+12 20591
## - mydata$n_bathrooms           2 7.0691e+11 7.0646e+12 20595
##
## Step:  AIC=20502.4
## mydata$price ~ poly(mydata$mq, 2) + mydata$n_rooms + mydata$n_bathrooms +
##     mydata$has_terrace + mydata$has_alarm + mydata$heating +
##     mydata$has_air_conditioning + mydata$is_furnished
##
##                               Df  Sum of Sq        RSS   AIC
## - mydata$has_alarm             1 7.2247e+09 6.3713e+12 20501
## - mydata$has_terrace           1 7.5304e+09 6.3716e+12 20502
## <none>                                      6.3641e+12 20502
## - mydata$is_furnished          1 1.4233e+10 6.3783e+12 20502
## - mydata$heating               1 2.3864e+10 6.3880e+12 20504
## - mydata$has_air_conditioning  1 2.4379e+10 6.3885e+12 20504
## - mydata$n_rooms               3 7.4415e+10 6.4385e+12 20507
## - poly(mydata$mq, 2)           2 6.8386e+11 7.0480e+12 20591
## - mydata$n_bathrooms           2 7.0211e+11 7.0662e+12 20593
##
## Step:  AIC=20501.43
## mydata$price ~ poly(mydata$mq, 2) + mydata$n_rooms + mydata$n_bathrooms +
##     mydata$has_terrace + mydata$heating + mydata$has_air_conditioning +
##     mydata$is_furnished
##
##                               Df  Sum of Sq        RSS   AIC
## - mydata$has_terrace           1 8.5242e+09 6.3799e+12 20501
## <none>                                      6.3713e+12 20501
## - mydata$is_furnished          1 1.4704e+10 6.3860e+12 20502
## - mydata$heating               1 2.3231e+10 6.3946e+12 20503
## - mydata$has_air_conditioning  1 2.8301e+10 6.3996e+12 20503
## - mydata$n_rooms               3 7.3781e+10 6.4451e+12 20506
## - poly(mydata$mq, 2)           2 6.8644e+11 7.0578e+12 20590
## - mydata$n_bathrooms           2 7.0773e+11 7.0791e+12 20593
##
## Step:  AIC=20500.64
## mydata$price ~ poly(mydata$mq, 2) + mydata$n_rooms + mydata$n_bathrooms +
##     mydata$heating + mydata$has_air_conditioning + mydata$is_furnished
##
##                               Df  Sum of Sq        RSS   AIC
## <none>                                      6.3799e+12 20501
## - mydata$is_furnished          1 1.5821e+10 6.3957e+12 20501
## - mydata$heating               1 2.1431e+10 6.4013e+12 20502
## - mydata$has_air_conditioning  1 3.3525e+10 6.4134e+12 20503
## - mydata$n_rooms               3 7.5933e+10 6.4558e+12 20505
## - poly(mydata$mq, 2)           2 6.8971e+11 7.0696e+12 20589
## - mydata$n_bathrooms           2 7.1443e+11 7.0943e+12 20593
```

```
summary(pol1)
```

```
##
## Call:
## lm(formula = mydata$price ~ poly(mydata$mq, 2) + mydata$n_rooms +
##     mydata$n_bathrooms + mydata$heating + mydata$has_air_conditioning +
##     mydata$is_furnished, data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -205823  -56289  -15539   37551  387736
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    104223       7809  13.347  < 2e-16 ***
## poly(mydata$mq, 2)1            954532      97237   9.817  < 2e-16 ***
## poly(mydata$mq, 2)2           -208752      98550  -2.118   0.0344 *
## mydata$n_rooms3                 16695       8546   1.953   0.0511 .
## mydata$n_rooms4                 11594       9800   1.183   0.2371
## mydata$n_rooms5                 -7822      11217  -0.697   0.4858
## mydata$n_bathrooms2             59793       6725   8.891  < 2e-16 ***
## mydata$n_bathrooms3            107797      15598   6.911 9.16e-12 ***
## mydata$heatingother             15942       9210   1.731   0.0838 .
## mydata$has_air_conditioning1    13314       6150   2.165   0.0307 *
## mydata$is_furnished1            15370      10334   1.487   0.1373
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 84570 on 892 degrees of freedom
## Multiple R-squared:  0.289,  Adjusted R-squared:  0.281
## F-statistic: 36.25 on 10 and 892 DF,  p-value: < 2.2e-16
```

The step function has ended with this minimal adequate model in interactions:

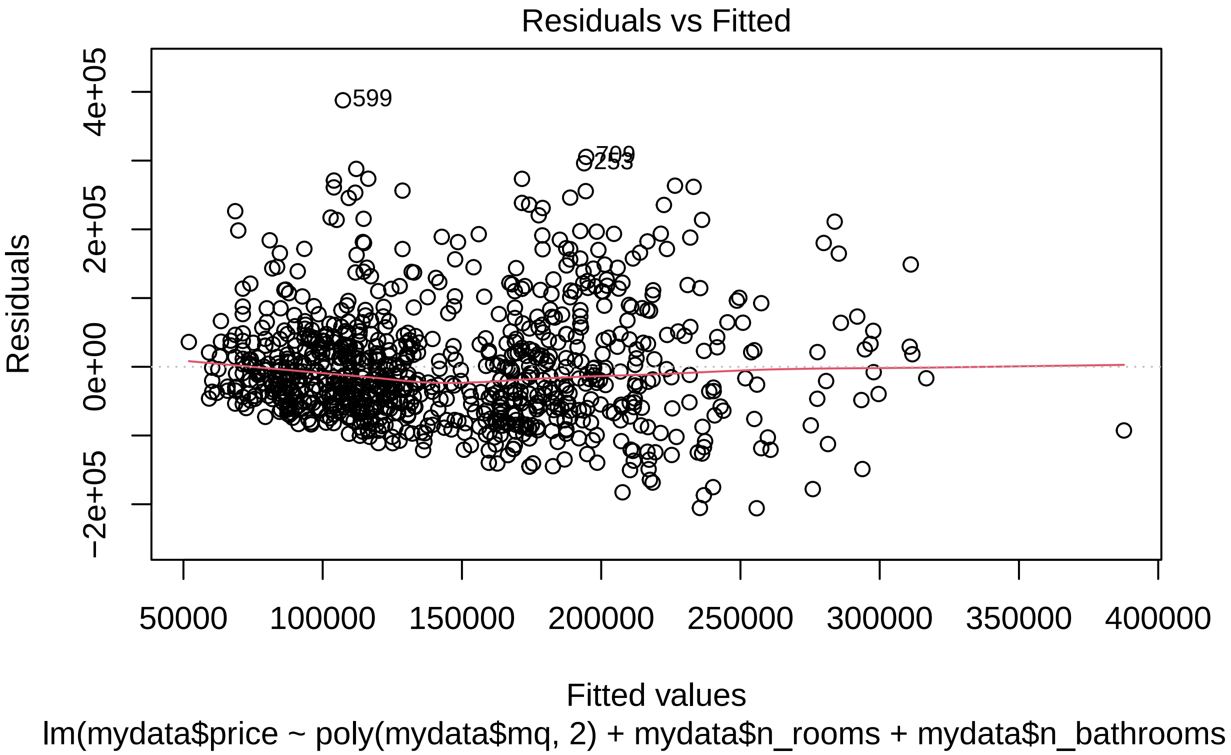
> lm(formula = mydata$price\ poly(mydata$mq, 2) + mydata$n_rooms + mydata$n_bathrooms + mydata$heating + mydata$has_air_conditioning + mydata$is_furnished)

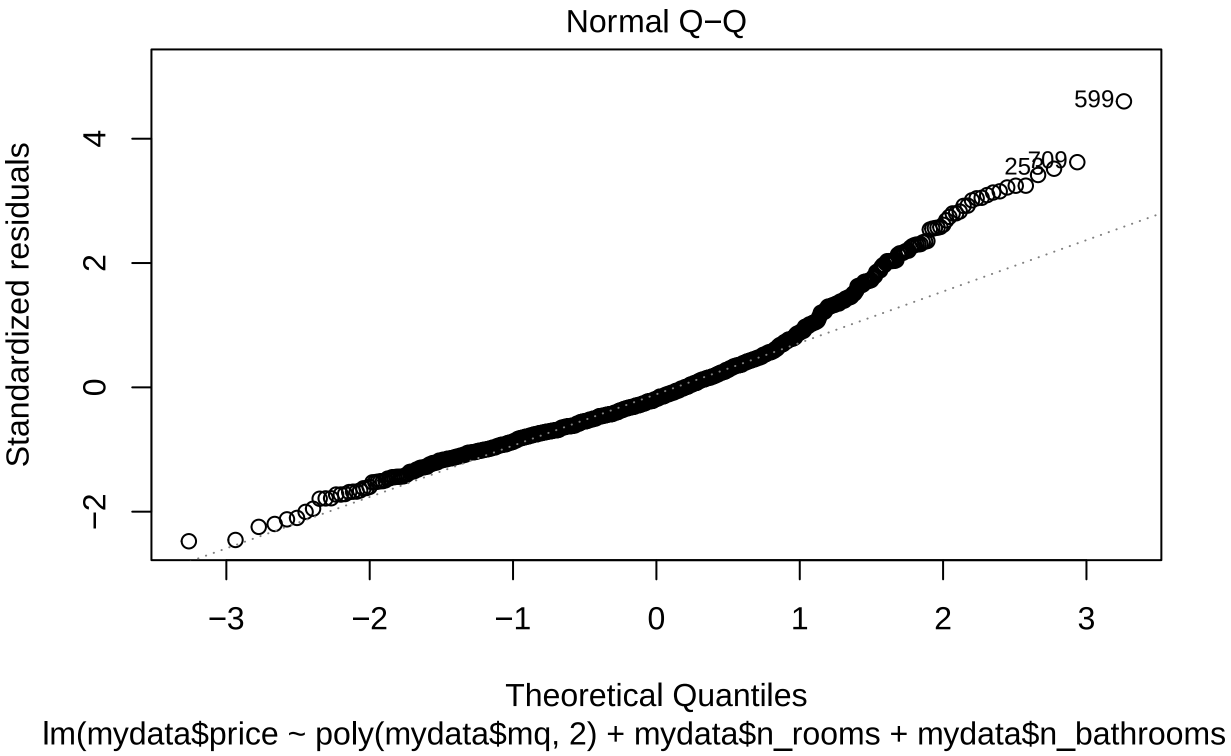
Summary of the findings:

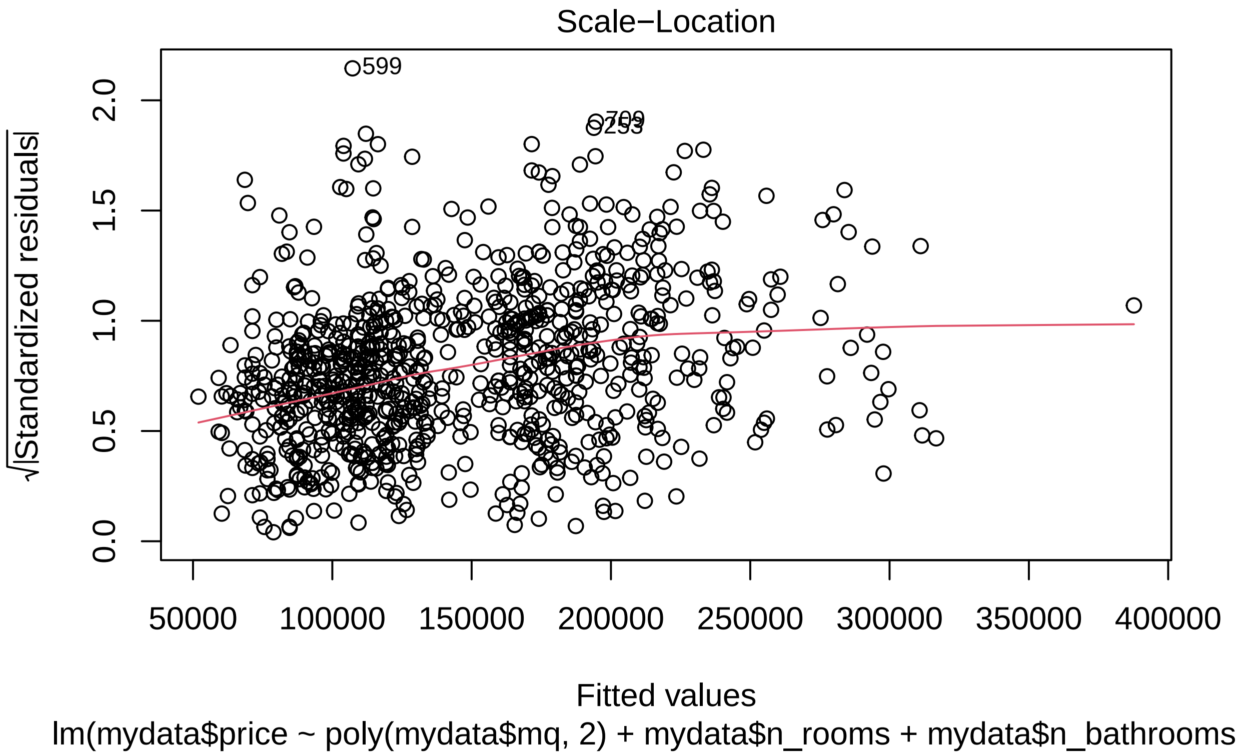
1. Other then n_room and is_furnished, all the other variables are significant.
2. R- squared value is 0.281, which is low, but significant to provide good fit for the model.
3. F values is significant which p value = 2.2e-16.

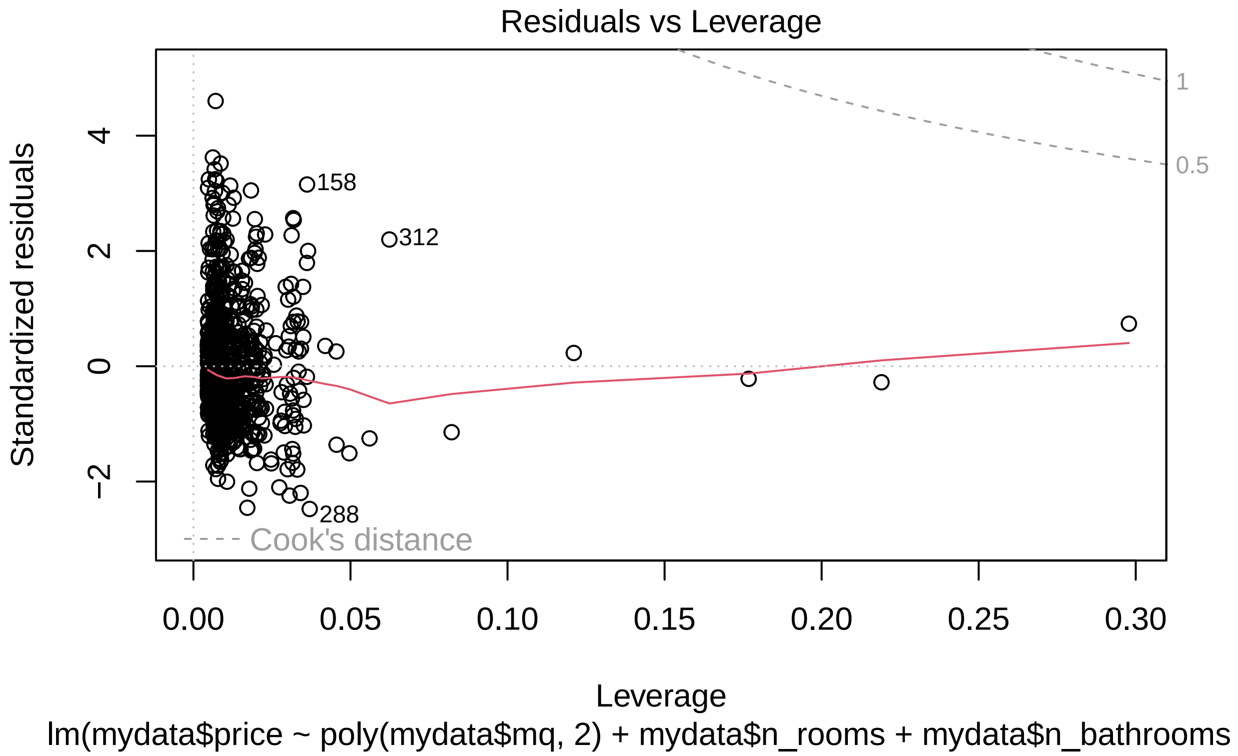
Graphical representation:

```
plot(pol1)
```


Residuals vs Fitted
Residuals
4e+05
2e+05
0e+00
−2e+05
599
709
253
50000
100000
150000
200000
250000
300000
350000
400000
Fitted values
lm(mydata$price ~ poly(mydata$mq, 2) + mydata$n_rooms + mydata$n_bathrooms


Normal Q−Q
Standardized residuals
−2
0
2
4
599
709
253
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(mydata$price ~ poly(mydata$mq, 2) + mydata$n_rooms + mydata$n_bathrooms


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
599
709
253
50000
100000
150000
200000
250000
300000
350000
400000
Fitted values
lm(mydata$price ~ poly(mydata$mq, 2) + mydata$n_rooms + mydata$n_bathrooms



1. The plots of residual vs fitted and QQ plot does not raise any concerns, although the QQ plot have some outliers. 2. No heteroscedasticity present in this model

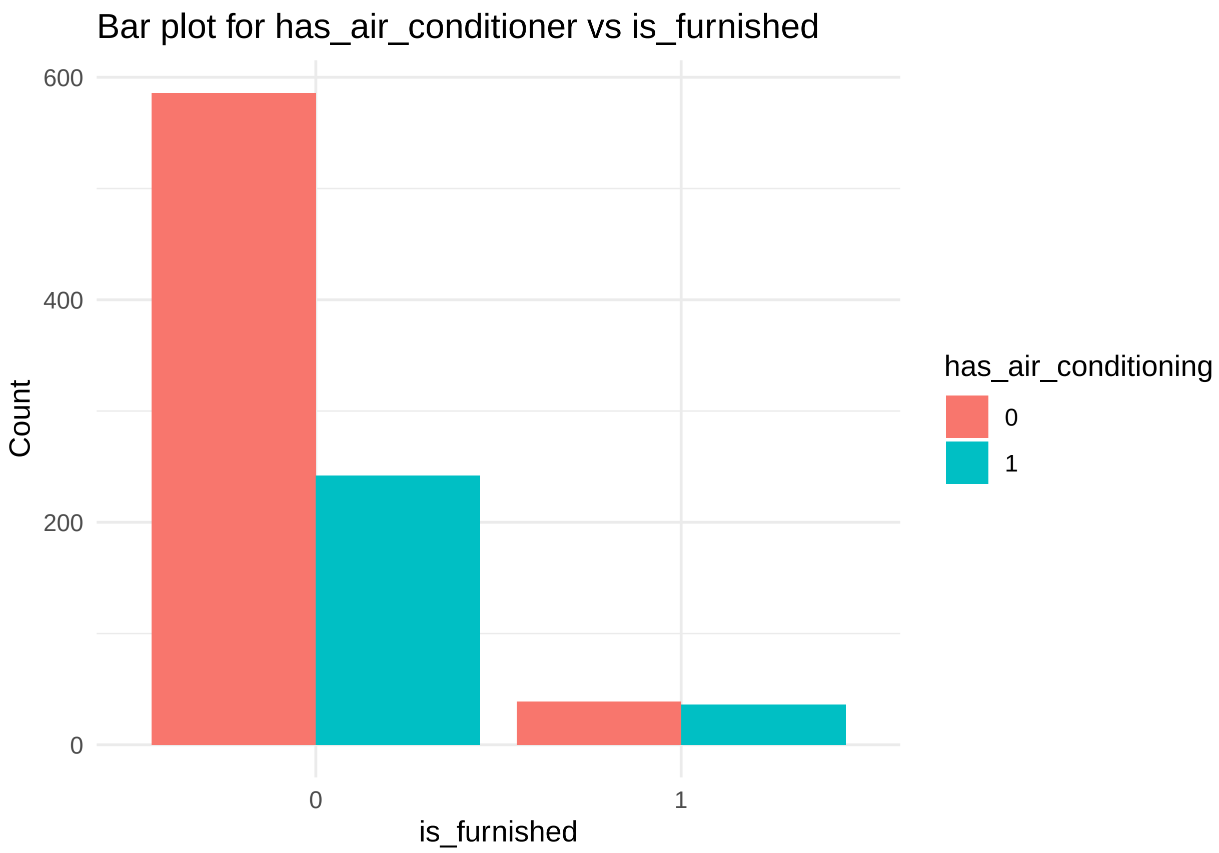
4. Extension work

4.1 Model the likelihood of a property being furnished (using the is_furnished variable provided).

EDA: The variables has_air_condiotioing, has_alarm and n_room have logical relationship with the is_furnished. They are more likely will have a co relationship with the dependent variable. so lets examine them with the bar chart.

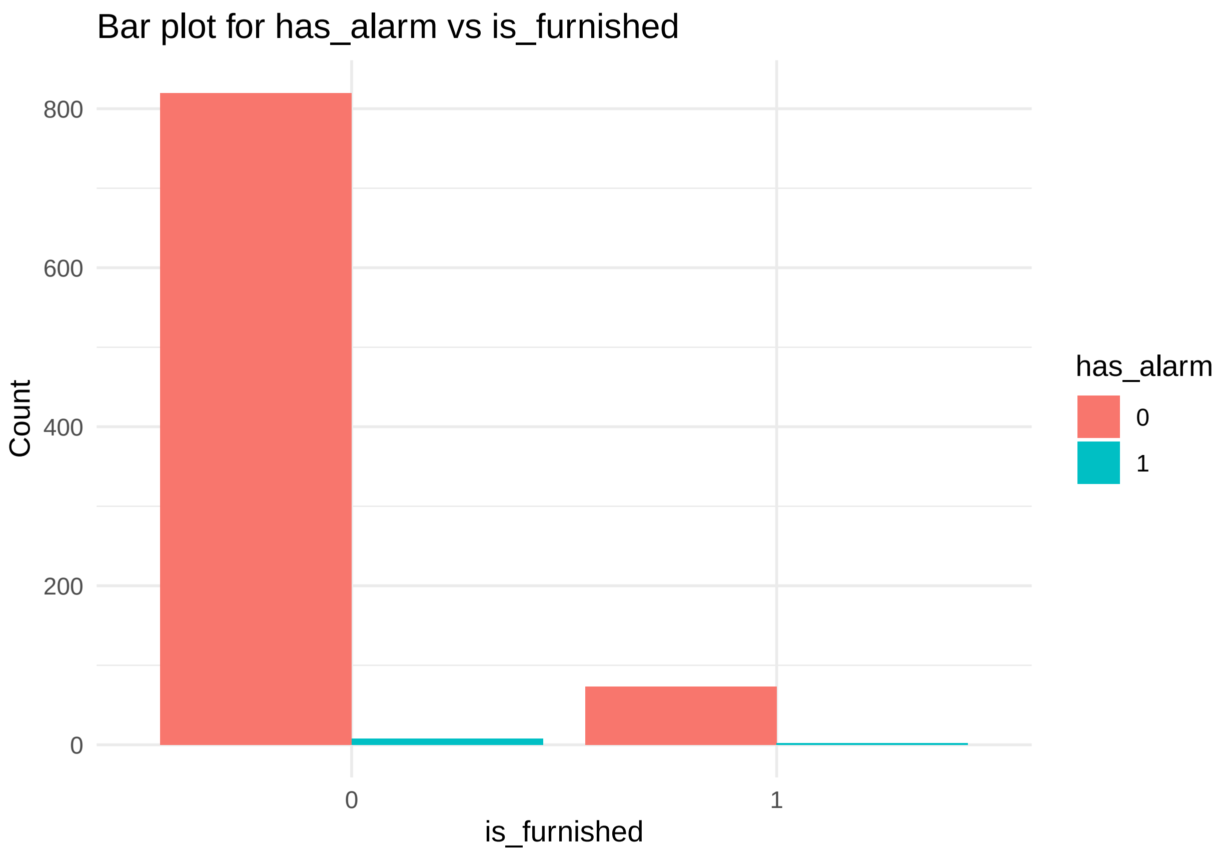
```
ggplot(mydata,
       aes(x = (is_furnished),
           fill = (has_air_conditioning)))+
  geom_bar(position = "dodge") +

  labs(y = "Count",
       fill = "has_air_conditioning",
       x = "is_furnished",
       title = "Bar plot for has_air_conditioner vs is_furnished") +
  theme_minimal()
```



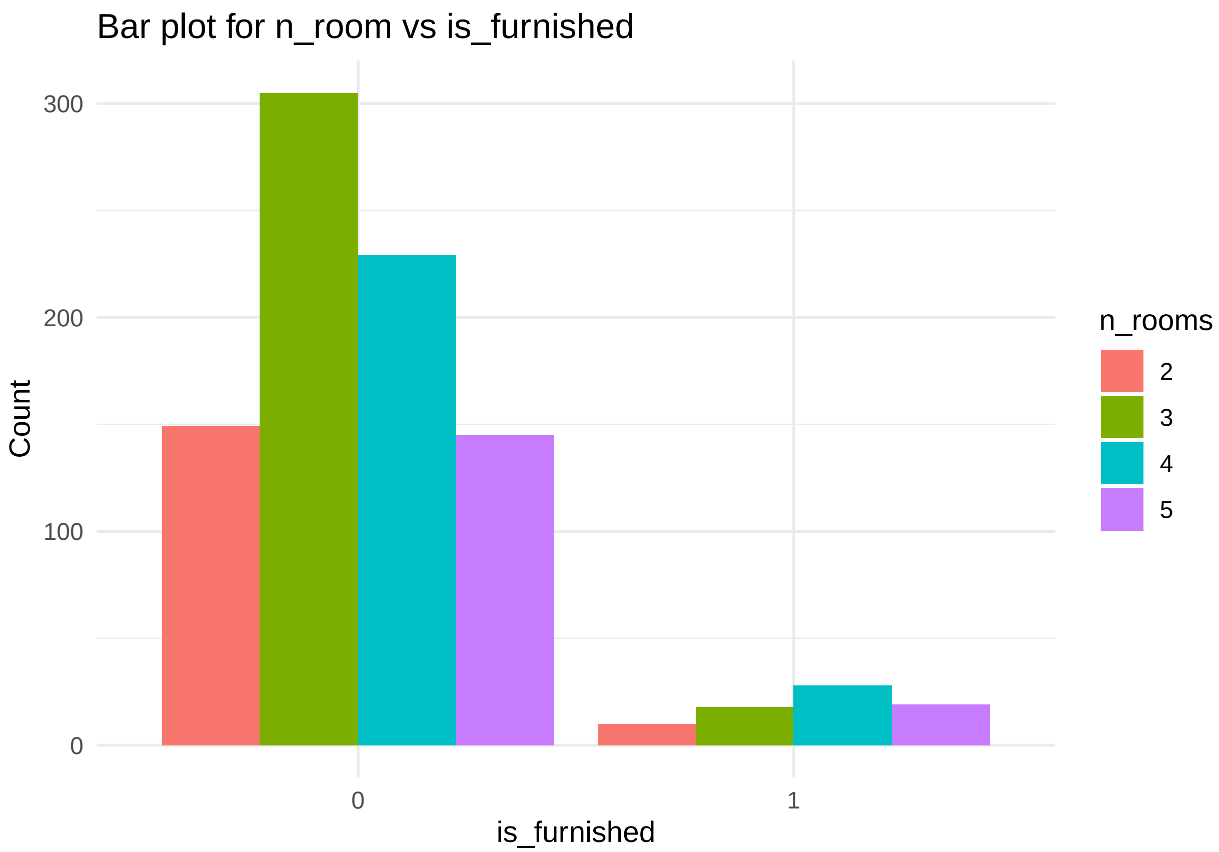
```
ggplot(mydata,
       aes(x = (is_furnished),
           fill = (has_alarm)))+
  geom_bar(position = "dodge") +

  labs(y = "Count",
       fill = "has_alarm",
       x = "is_furnished",
       title = "Bar plot for has_alarm vs is_furnished") +
  theme_minimal()
```



```
ggplot(mydata,
       aes(x = (is_furnished),
           fill = (n_rooms)))+
  geom_bar(position = "dodge") +

  labs(y = "Count",
       fill = "n_rooms",
       x = "is_furnished",
       title = "Bar plot for n_room vs is_furnished") +
  theme_minimal()
```



From the above graphs we can able to find the count is more when the value is 0 for is_furnished and n_rooms.

we can find the dependency between these categorical values with count data and chi square test:

The null hypothesis that we are testing is: H_0: The variables are independent.
The alternative hypothesis is: H_1: There is a relationship between the variables.

Since the count data of is_furnished and has_alarm contains value less the 5, we can use fisher with the same hypothesis to determine the dependency.

```
fisher.test(table(mydata$is_furnished,mydata$has_alarm))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(mydata$is_furnished, mydata$has_alarm)
## p-value = 0.1987
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##   0.2849639 14.4055216
## sample estimates:
## odds ratio
##   2.803305
```

```
chisq.test(table(mydata$is_furnished,mydata$n_rooms))
```

```
##
```

```
##  Pearson's Chi-squared test
##
## data:  table(mydata$is_furnished, mydata$n_rooms)
## X-squared = 8.5952, df = 3, p-value = 0.03519
```

```
chisq.test(table(mydata$is_furnished,mydata$has_air_conditioning))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(mydata$is_furnished, mydata$has_air_conditioning)
## X-squared = 10.51, df = 1, p-value = 0.001187
```

Summary of the finding:

1. We can see that the chi square is significant, so the is_furnished is dependent on has_air_conditioning and n_rooms
2. The fishes test indicates that there is no evident to conclude any relationshion between the variables.

The dependent variable is 'is_furnished', which is a binary attribute, and the independent variables are mix of numerical and categorical. So we are proceeding with the logical regression model.

```
fur.glm<- glm(mydata$is_furnished~mydata$price*mydata$mq+mydata$floor+mydata$n_rooms+mydata$n_bathrooms+
summary.lm(fur.glm)
```

```
##
## Call:
## glm(formula = mydata$is_furnished ~ mydata$price * mydata$mq +
##     mydata$floor + mydata$n_rooms + mydata$n_bathrooms + mydata$has_terrace +
##     mydata$has_alarm + mydata$heating + mydata$has_air_conditioning +
##     mydata$has_parking, family = "binomial", data = mydata)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7126 -0.3192 -0.2740 -0.2024  5.5493
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -3.167e+00  5.082e-01  -6.232 7.12e-10 ***
## mydata$price                1.460e-06  2.250e-06   0.649  0.51649
## mydata$mq                  -7.172e-04  3.801e-03  -0.189  0.85041
## mydata$floor2               1.496e-02  2.754e-01   0.054  0.95671
## mydata$floor3              -8.449e-02  3.979e-01  -0.212  0.83190
## mydata$floor4               4.216e-01  5.257e-01   0.802  0.42282
## mydata$floor5              -1.533e+01  1.092e+03  -0.014  0.98881
## mydata$floor6              -1.508e+01  1.886e+03  -0.008  0.99362
## mydata$floor7              -1.495e+01  1.728e+03  -0.009  0.99310
## mydata$floor8              -1.512e+01  3.908e+03  -0.004  0.99691
## mydata$floor9              -1.532e+01  3.908e+03  -0.004  0.99687
## mydata$n_rooms3            -1.823e-01  4.262e-01  -0.428  0.66894
## mydata$n_rooms4             5.730e-01  4.338e-01   1.321  0.18696
## mydata$n_rooms5             6.438e-01  4.741e-01   1.358  0.17479
## mydata$n_bathrooms2          -3.153e-02  2.983e-01  -0.106  0.91583
## mydata$n_bathrooms3          -9.369e-01  7.870e-01  -1.190  0.23419
## mydata$has_terrace1           3.727e-01  3.226e-01   1.156  0.24816
## mydata$has_alarm1             1.896e-01  8.343e-01   0.227  0.82024
## mydata$heatingother          -1.057e-02  4.286e-01  -0.025  0.98033
## mydata$has_air_conditioning1  7.747e-01  2.526e-01   3.067  0.00223 **
## mydata$has_parking1           6.775e-02  1.075e+00   0.063  0.94976
## mydata$price:mydata$mq        2.625e-09  1.267e-08   0.207  0.83586
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9877 on 881 degrees of freedom
## Multiple R-squared:  0.0003231,  Adjusted R-squared:  -0.02351
## F-statistic: 0.01356 on 21 and 881 DF,  p-value: 1
```

Minimal adequate model: Used a model selection approach to achieve a minimal adequate model.

```
a<-step(fur.glm)
```

```
## Start:  AIC=530.67
## mydata$is_furnished ~ mydata$price * mydata$mq + mydata$floor +
##     mydata$n_rooms + mydata$n_bathrooms + mydata$has_terrace +
##     mydata$has_alarm + mydata$heating + mydata$has_air_conditioning +
##     mydata$has_parking
##
##                               Df Deviance    AIC
## - mydata$floor                 8   491.60 519.60
## - mydata$n_bathrooms           2   488.44 528.44
## - mydata$heating               1   486.67 528.67
## - mydata$has_parking           1   486.67 528.67
## - mydata$price:mydata$mq       1   486.71 528.71
## - mydata$has_alarm             1   486.72 528.72
## - mydata$has_terrace           1   487.91 529.91
## <none>                             486.67 530.67
## - mydata$n_rooms               3   493.64 531.64
## - mydata$has_air_conditioning  1   495.61 537.61
##
## Step:  AIC=519.6
## mydata$is_furnished ~ mydata$price + mydata$mq + mydata$n_rooms +
##     mydata$n_bathrooms + mydata$has_terrace + mydata$has_alarm +
##     mydata$heating + mydata$has_air_conditioning + mydata$has_parking +
##     mydata$price:mydata$mq
##
##                               Df Deviance    AIC
## - mydata$n_bathrooms           2   493.61 517.61
## - mydata$heating               1   491.61 517.61
## - mydata$has_parking           1   491.62 517.62
## - mydata$has_alarm             1   491.68 517.68
## - mydata$price:mydata$mq       1   491.70 517.70
## - mydata$has_terrace           1   492.85 518.85
## <none>                             491.60 519.60
## - mydata$n_rooms               3   498.91 520.91
## - mydata$has_air_conditioning  1   500.38 526.38
```

```
##
## Step:  AIC=517.61
## mydata$is_furnished ~ mydata$price + mydata$mq + mydata$n_rooms +
##     mydata$has_terrace + mydata$has_alarm + mydata$heating +
##     mydata$has_air_conditioning + mydata$has_parking + mydata$price:mydata$mq
##
##                                Df Deviance    AIC
## - mydata$heating                1   493.61 515.61
## - mydata$has_parking            1   493.63 515.63
## - mydata$price:mydata$mq        1   493.69 515.69
## - mydata$has_alarm              1   493.73 515.73
## - mydata$has_terrace            1   494.88 516.88
## <none>                              493.61 517.61
## - mydata$n_rooms                3   500.39 518.39
## - mydata$has_air_conditioning  1   502.45 524.45
##
## Step:  AIC=515.61
## mydata$is_furnished ~ mydata$price + mydata$mq + mydata$n_rooms +
##     mydata$has_terrace + mydata$has_alarm + mydata$has_air_conditioning +
##     mydata$has_parking + mydata$price:mydata$mq
##
##                                Df Deviance    AIC
## - mydata$has_parking            1   493.63 513.63
## - mydata$price:mydata$mq        1   493.69 513.69
## - mydata$has_alarm              1   493.74 513.74
## - mydata$has_terrace            1   494.90 514.90
## <none>                              493.61 515.61
## - mydata$n_rooms                3   500.43 516.43
## - mydata$has_air_conditioning  1   502.46 522.46
##
## Step:  AIC=513.63
## mydata$is_furnished ~ mydata$price + mydata$mq + mydata$n_rooms +
##     mydata$has_terrace + mydata$has_alarm + mydata$has_air_conditioning +
##     mydata$price:mydata$mq
##
##                                Df Deviance    AIC
## - mydata$price:mydata$mq        1   493.71 511.71
## - mydata$has_alarm              1   493.76 511.76
## - mydata$has_terrace            1   494.93 512.93
## <none>                              493.63 513.63
## - mydata$n_rooms                3   500.44 514.44
## - mydata$has_air_conditioning  1   502.50 520.50
##
## Step:  AIC=511.71
## mydata$is_furnished ~ mydata$price + mydata$mq + mydata$n_rooms +
##     mydata$has_terrace + mydata$has_alarm + mydata$has_air_conditioning
##
##                                Df Deviance    AIC
## - mydata$mq                     1   493.71 509.71
## - mydata$has_alarm              1   493.84 509.84
## - mydata$price                  1   494.92 510.92
## - mydata$has_terrace            1   495.04 511.04
## <none>                              493.71 511.71
## - mydata$n_rooms                3   500.46 512.46
## - mydata$has_air_conditioning  1   502.54 518.54
##
## Step:  AIC=509.71
## mydata$is_furnished ~ mydata$price + mydata$n_rooms + mydata$has_terrace +
##     mydata$has_alarm + mydata$has_air_conditioning
##
##                                Df Deviance    AIC
## - mydata$has_alarm              1   493.84 507.84
## - mydata$has_terrace            1   495.04 509.04
## - mydata$price                  1   495.07 509.07
## <none>                              493.71 509.71
## - mydata$n_rooms                3   500.92 510.92
## - mydata$has_air_conditioning  1   502.62 516.62
##
## Step:  AIC=507.84
## mydata$is_furnished ~ mydata$price + mydata$n_rooms + mydata$has_terrace +
##     mydata$has_air_conditioning
##
##                                Df Deviance    AIC
## - mydata$has_terrace            1   495.26 507.26
## - mydata$price                  1   495.27 507.27
## <none>                              493.84 507.84
## - mydata$n_rooms                3   501.14 509.14
## - mydata$has_air_conditioning  1   503.24 515.24
##
## Step:  AIC=507.26
## mydata$is_furnished ~ mydata$price + mydata$n_rooms + mydata$has_air_conditioning
##
##                                Df Deviance    AIC
## - mydata$price                  1   496.86 506.86
## <none>                              495.26 507.26
## - mydata$n_rooms                3   502.84 508.84
## - mydata$has_air_conditioning  1   506.29 516.29
##
## Step:  AIC=506.86
## mydata$is_furnished ~ mydata$n_rooms + mydata$has_air_conditioning
##
##                                Df Deviance    AIC
## <none>                              496.86 506.86
## - mydata$n_rooms                3   506.20 510.20
## - mydata$has_air_conditioning  1   508.20 516.20
```

```
summary.lm(a)
```

```
##
## Call:
## glm(formula = mydata$is_furnished ~ mydata$n_rooms + mydata$has_air_conditioning,
##     family = "binomial", data = mydata)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4768 -0.3137 -0.2982 -0.2054  4.8694
##
## Coefficients:
```

```
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -3.0463     0.3481  -8.750  < 2e-16 ***
## mydata$n_rooms3                 -0.1195     0.4061  -0.294 0.768609
## mydata$n_rooms4                  0.6264     0.3830   1.636 0.102269
## mydata$n_rooms5                  0.7279     0.4081   1.784 0.074832 .
## mydata$has_air_conditioning1     0.8372     0.2437   3.436 0.000618 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.993 on 898 degrees of freedom
## Multiple R-squared:  0.0001753,  Adjusted R-squared:  -0.004278
## F-statistic: 0.03935 on 4 and 898 DF,  p-value: 0.9971
```

The step function has ended with this minimal adequate model in interactions:

> glm(formula = mydata$is_{f}urnished\ mydata$n_rooms + mydata$has_air_conditioning, family = “binomial”, data = myd)

Summary of the findings:

1. we can say that only has_air_conditioning variable is significant.
2. R-squared value is very small.
3. F value is not significant with p value =1.

As has_air_conditioning is the only significant variable in it, We can create a binomial model with only that variable.

```
d<-glm(formula = (mydata$is_furnished)~mydata$has_air_conditioning,family = "binomial",data=mydata)
summary.lm (d)
```

```
##
## Call:
## glm(formula = (mydata$is_furnished) ~ mydata$has_air_conditioning,
##     family = "binomial", data = mydata)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3857 -0.3857 -0.2580 -0.2580  3.8763
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -2.7098     0.1656 -16.368   <2e-16 ***
## mydata$has_air_conditioning1     0.8043     0.2437   3.301    0.001 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.001 on 901 degrees of freedom
## Multiple R-squared:  8.413e-05,  Adjusted R-squared:  -0.001026
## F-statistic: 0.0758 on 1 and 901 DF,  p-value: 0.7831
```

Summary of the findings:

1. The F values is not significant with p value =0.7831.
2. The R-squared values is very low, shows it is not fit.

Since the model is not significant, analyzing the model plot will not be helpful.