Final Analysis

1 Preliminary analysis

1.1 Loading the Data and the first look of the dataset

```
dtrain <- read.csv("training.csv")
```

Loading the data for model training for specific dictionary. the variable "dtrain" has the training dataset. In specific, column "prices" in the dataset will be the response variable of linear model and other columns will be predictors of the linear model.

```
colnames(dtrain)
```

```
##  [1] "X"          "bathrm"     "hf_bathrm"  "heat"       "ac"
##  [6] "rooms"      "bedrm"      "ayb"        "yr_rmdl"    "eyb"
## [11] "stories"    "saledate"   "price"      "gba"        "style"
## [16] "grade"      "extwall"    "kitchens"   "fireplaces" "landarea"
```

The above listed names of each columns in the training dataset. There are 19 variables in total and (total - response) 18 predictors for the linear model.

1.2 Filling the missing datas and some general transformations

Before examining more about the dataset, there are some variable need preprocessing. Data clearing is an important process. There are two steps in this procedure: filling the missing data and transforming some categorical data to numerical.

1.2.1 Missing Data

First finding the data which has NA value in it.

```
colnames(dtrain)[colSums(is.na(dtrain)) > 0]
```

```
## [1] "yr_rmdl" "stories"
```

From the above result, both "yr_rmdl" and "stories" fields have missing data. Since the values of missing data are unknown, using the mean of other values ("yr_rmdl" = 2006, "stories" = 2) as its value to minimize the independent effect of NA values.

```
dtrain$yr_rmdl[is.na(dtrain$yr_rmdl)] <- 2006
dtrain$stories[is.na(dtrain$stories)] <- 2
```

1.2.2 Categorical Data

From the dataset, "saledate" field has an irregular format which can convert to numerical data

```
dtrain$saledate <- as.integer(substr(as.Date(dtrain$saledate),1,4))
```

Since the main difference between the values in the "saledate" field is years, ignoring month and times in the value will decrease the distractions in the linear model. Therefore, the "saledate" in char format is converted into its year in int format.
Besides, "ac" field has only two categories "Yes" and "No". Therefore, it can be transformed into "1" and "0".

```
for (row in 1:length(dtrain$ac)) {
   if (dtrain$ac[row] == "Y") {
    dtrain$ac[row] = as.integer(1)
   }
   if (dtrain$ac[row] == "N") {
    dtrain$ac[row] = as.integer(0)
  }
}
dtrain$ac = as.integer(dtrain$ac)
```

1.3 Correlation coefficient between response and predictors

In order to learn more details about the given dataset, correlation coefficient will be an essential element to be analysis. It can provide an evaluation on the relationship between different predictors and the response variables.
First of all,filtering out the numerical predictors for the dataset

```
nums <- unlist(lapply(dtrain, is.numeric))
colnames(dtrain[ , nums])
```

```
##  [1] "X"          "bathrm"     "hf_bathrm"  "ac"         "rooms"
##  [6] "bedrm"      "ayb"        "yr_rmdl"    "eyb"        "stories"
## [11] "saledate"   "price"      "gba"        "kitchens"   "fireplaces"
## [16] "landarea"
```

The above listed all the numerical predictors in the given dataset. Next, finding the correlation Coefficients between response and these predictors

```
cor(dtrain[ , nums])[,"price"]
```

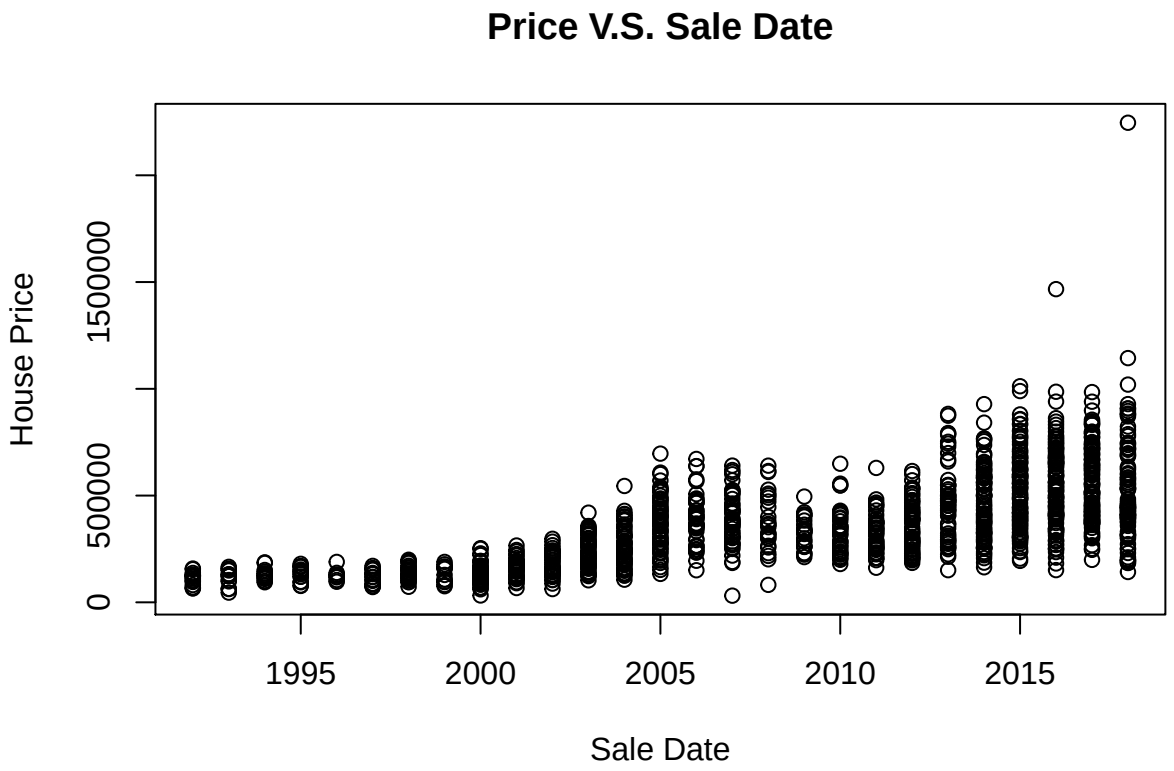
```
##           X      bathrm   hf_bathrm          ac       rooms       bedrm
## -0.39949090  0.53201366  0.18517412  0.25323748  0.36836444  0.46695990
##         ayb     yr_rmdl         eyb     stories    saledate       price
## -0.06812407  0.21799058  0.23270733  0.24321614  0.67242944  1.00000000
##         gba    kitchens  fireplaces    landarea
##  0.49482898  0.14529565  0.20414081  0.14668340
```

Using function cor(), it shows the correlation coefficient between responses and all numerical predictors. From the result, there are few fields that have a higher correlation than others. They are “bathrm”, “bedrm”, “saledate”, and “gba”. Therefore, these fields might have a stronger relationship with the response variable. This result is very important to the linear model. In specific, “saledate” has the highest correlation coefficient among all predictors.

1.4 Plots between response and some usefull predictors

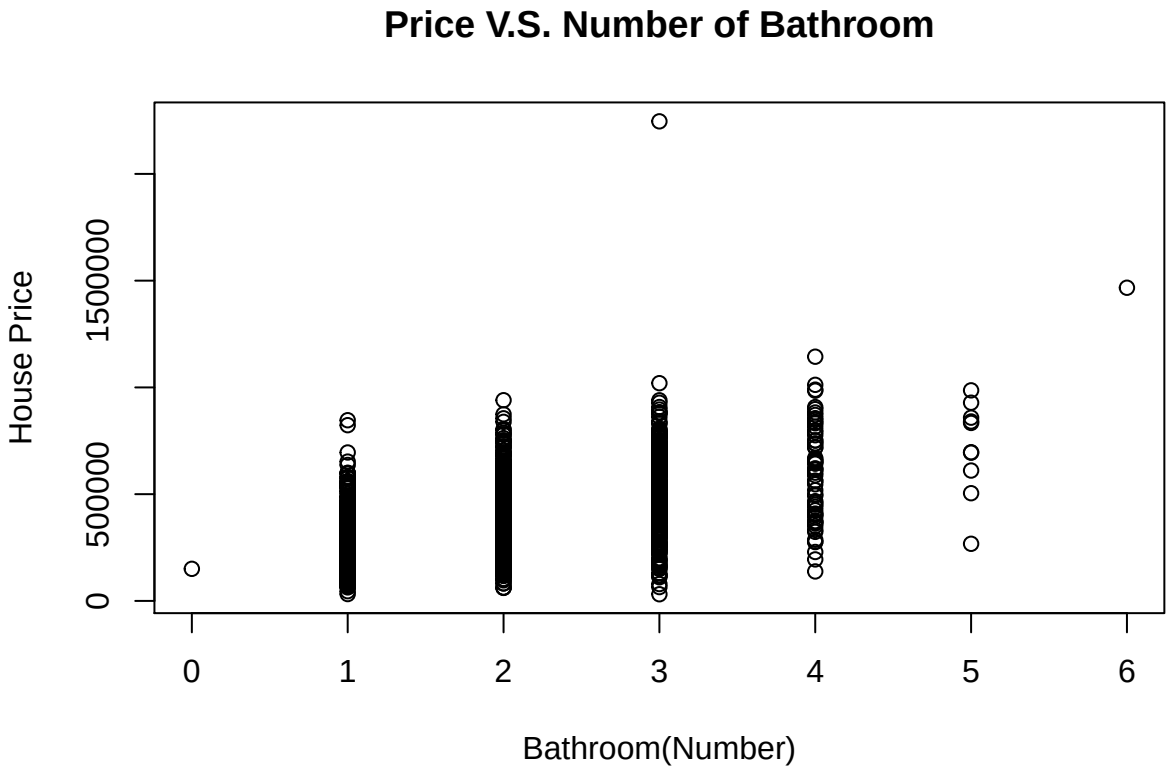
To further justify the relationship between the response variable and the higher correlation predictors, this section will show their plots with the response (price) to examine them in visualization.

```
plot(dtrain$saledate, dtrain$price, main = "Price V.S. Sale Date",
     xlab = "Sale Date", ylab = "House Price")
```



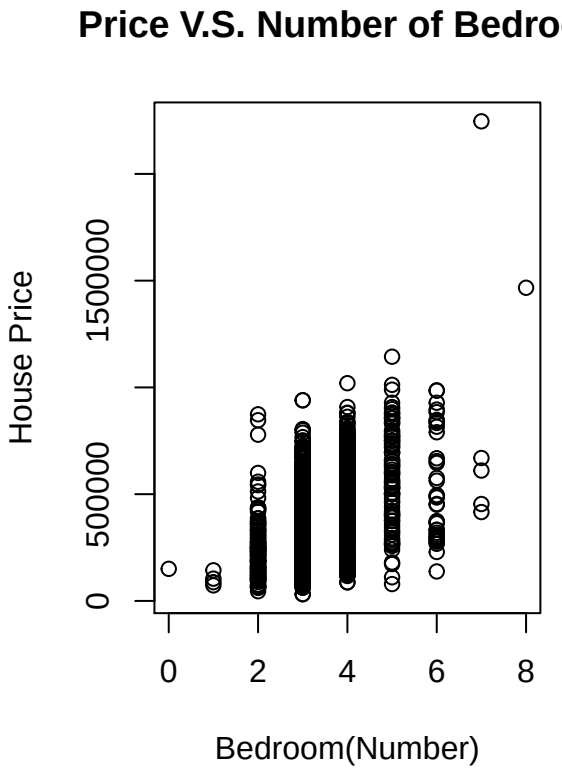
This is the plot of "saledate" v.s. "price". As the saledate increase, the price increase too. Therefore, this plot proves that there exists a linear (or higher-order) relationship between these two variables.

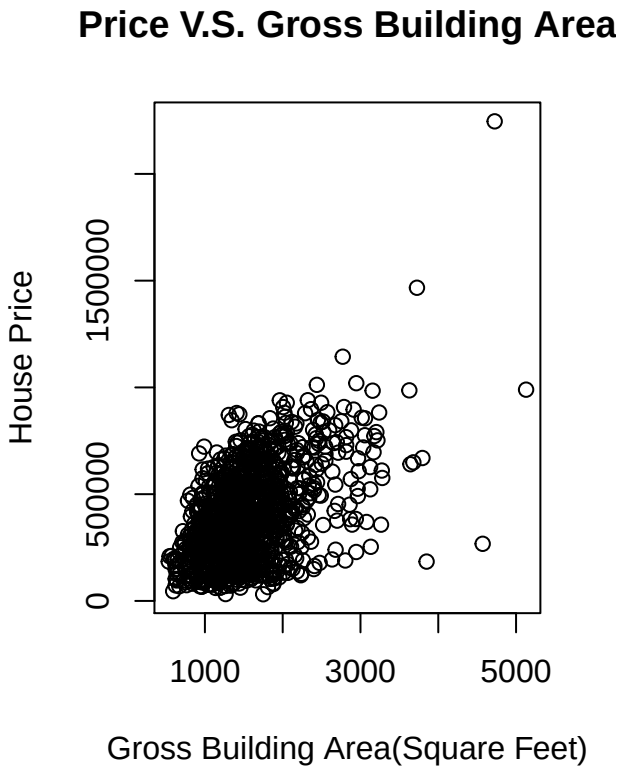
```
plot(dtrain$bathrm, dtrain$price, main = "Price V.S. Number of Bathroom",
     xlab = "Bathroom(Number)", ylab = "House Price")
```



Similar to the above, when the "bathrm" increase, the price increase too. Therefore, this plot proves that there exists a linear (or higher-order) relationship between these two variables.

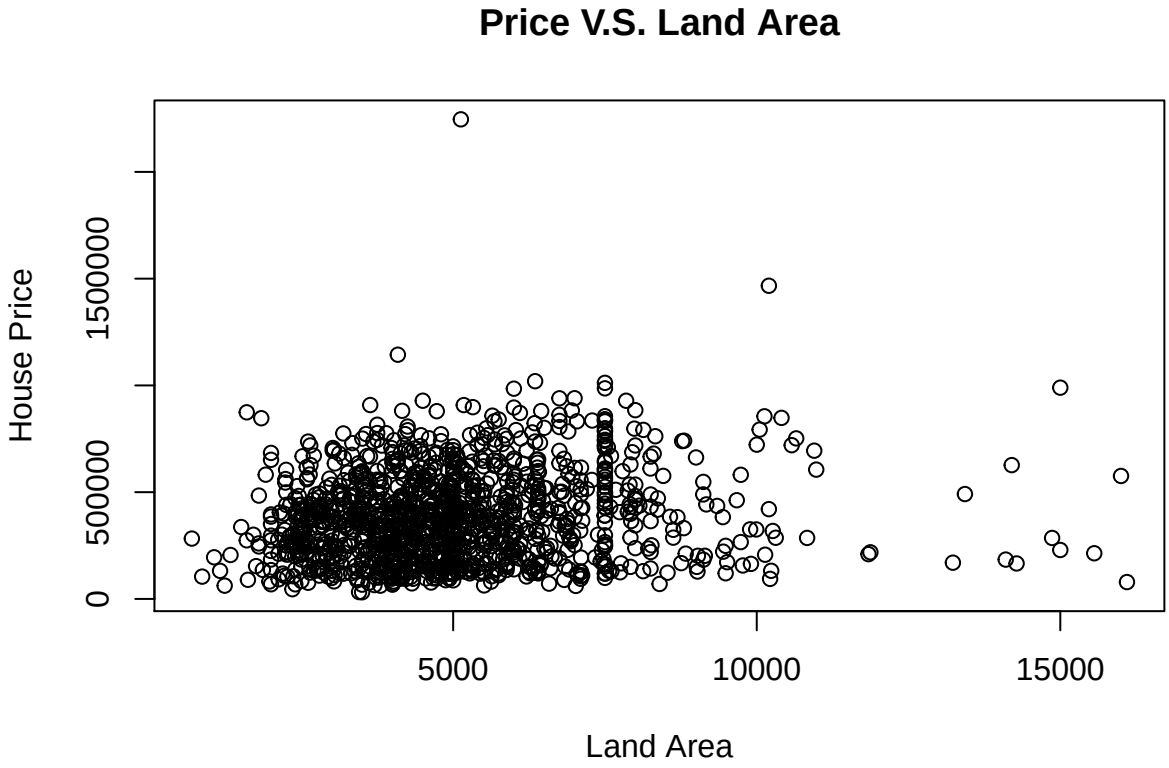
```
par(mfrow = c(1:2))
plot(dtrain$bedrm, dtrain$price, main = "Price V.S. Number of Bedroom",
     xlab = "Bedroom(Number)", ylab = "House Price")
plot(dtrain$gba, dtrain$price, main = "Price V.S. Gross Building Area",
     xlab = "Gross Building Area(Square Feet)", ylab = "House Price")
```





Compare to the previous two plots, the plots for both bedroom v.s. price and gross building area v.s. price cannot significantly show strong relationships. However, there still an increasing pattern in the distribution of dots of the plots. Also, in the "gba" vs. "price", there are some outliers in the gba fields that effect the visulization of the pattern.
At last, the plot of lower correlation predictors versus prices might further support this result

```
plot(dtrain$landarea, dtrain$price, main = "Price V.S. Land Area",
     xlab = "Land Area", ylab = "House Price")
```



The correlation coefficient between land area and price is 0.1452, which is much smaller than previous predictors. As shown in the plots, the dots are in a random distribution which is extremely hard to tell whether there exists a relationship or not. This result proves that higher correlation coefficients are related to a strong relationship.

1.4 Correlation coefficient within predictors

Next, analyzing the correlation coefficient within the predictors is also important to a better linear model. If there exists a strong relationship between two predictors, avoiding both of them into the linear model.

```
cor(dtrain[ , nums])[,"bedrm"]
```

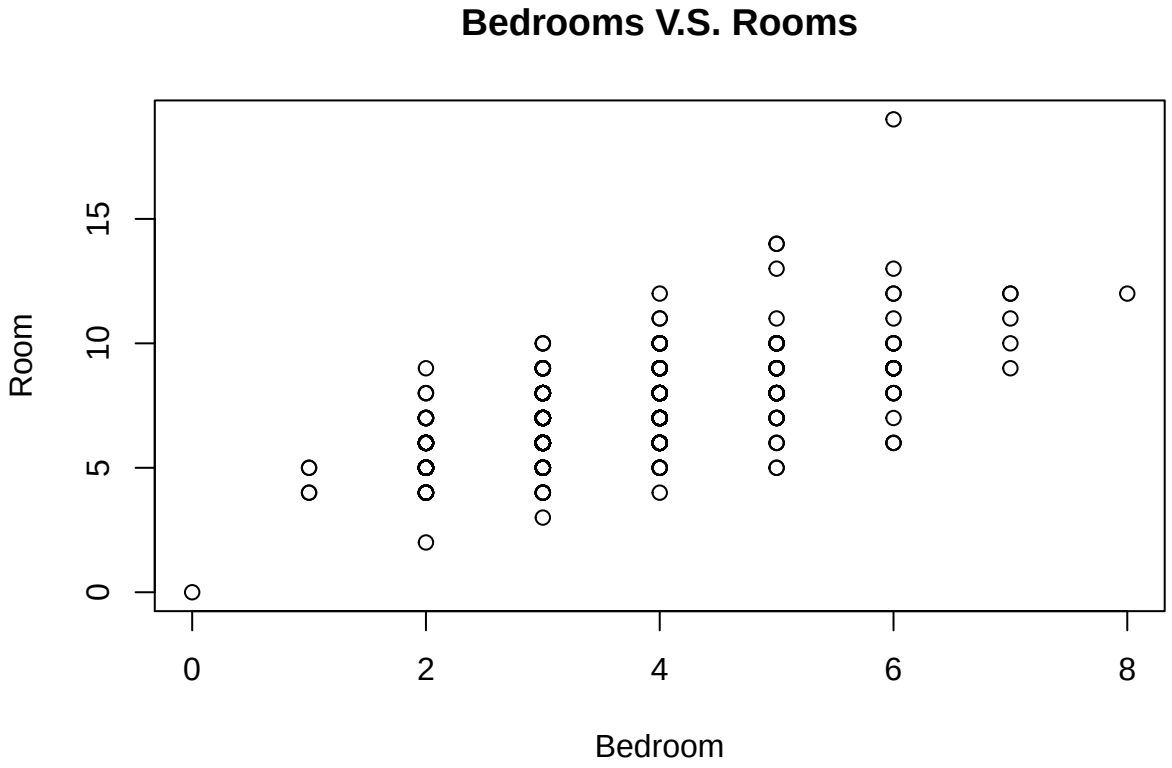
```
##          X     bathrm  hf_bathrm         ac      rooms      bedrm        ayb
## -0.0661952  0.6177090  0.1511272  0.2333872  0.6446478  1.0000000  0.1024059
##    yr_rmdl        eyb    stories   saledate      price        gba   kitchens
##  0.1654267  0.3209646  0.2357411  0.2353410  0.4669599  0.5824250  0.1628645
## fireplaces   landarea
##  0.1026266  0.1924927
```

the result above is the correlation of predictor “bedrm” with also predictors. This list of values includes the two highest correlation coefficients among all values in the dataset. Avoiding using “bathrm”, “rooms”, and “bedroom” together in the linear model will decrease the distraction.

1.5 Plots between predictors

Same as the response variable, the plot between predictors will also support the result that founded in the correlation coefficient.

```
plot(dtrain$bedrm, dtrain$rooms, main = "Bedrooms V.S. Rooms",
     xlab = "Bedroom", ylab = "Room")
```



This is the plot of "bedroom" v.s. "rooms". As the bedroom increase, the total number of room increase too. This result make sense in real life. Therefore, this plot proves that there exists a linear relationship between these two variables.

2 Transformation

In the Preliminary analysis, the "saledate" and "ac" field already transformed into numerical variables. There still some of the categorical variables which can be transformed into different factor level to perform a better prediction in the result.

```
levels(factor(dtrain$grade))
```

```
## [1] "Above Average" "Average"       "Fair Quality"  "Good Quality"
## [5] "Low Quality"   "Superior"      "Very Good"
```

This are the seven values in the categorical variables "grade". The meaning behinds these values represent a ranking that cannot interpret by lm function directly. Therefore, resorting to the order of the factor level will definitely increase the accuracy of the linear model.

```
dtrain$grade <- factor(dtrain$grade,
levels = c("Low Quality","Fair Quality",
           "Average","Above Average","Good Quality","Very Good","Superior"))
```

Therefore, all variables in the dataset that need the transformation have finished.
The summary of transformation:
Categorical Field: "saledate(YEAR-MO-DA xx:xx:xx)" -> Numerical Field :"saledate(YEAR)"
Categorical Field: "ac(Y/N)" -> Numerical Field :"ac(1/0)"
Categorical Field: "grades" -> Categorical Field :"grades(reranking)"

3 Model Checking

The final model from prediction and model building process is

$$price \sim saledate^4 + gba + grade + bathrm^2 + fireplaces + hf\_bathrm^2$$

In this section, the model checking process will apply to this final model to check whether there is any issue encounter.

3.1 Fitted the final model into the linear model function lm

```
fit <- lm(price ~ poly(saledate,4) + gba + grade + poly(bathrm,2) + fireplaces +
            poly(hf_bathrm, 2), data=dtrain)
```

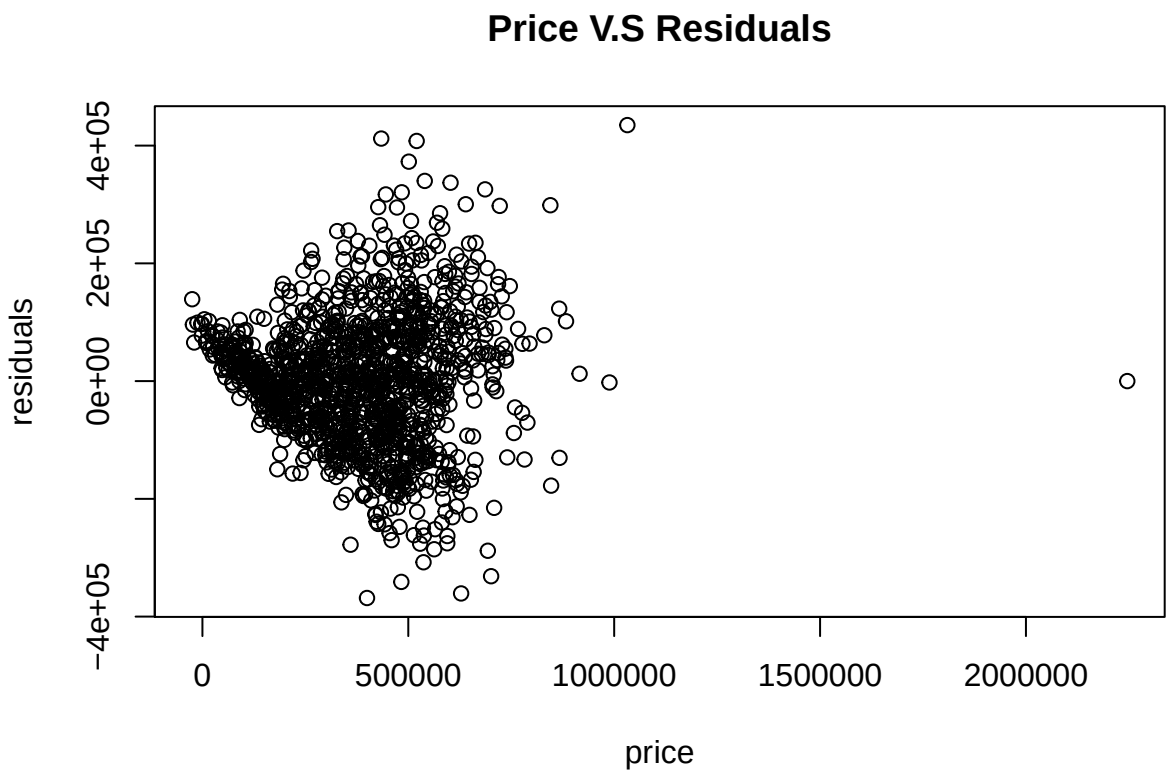
3.2 Checking mean of residual E(r) = 0

The most important assumption for linear regression models is $\mathrm{E}(\epsilon_i) = 0$. This assumption might lead to non-linear relationships between predictors and response. Since the existing of higher-order terms (residual versus predictors) already checked in the model building process, this section will mainly focus on the residual plot with the fitted variable.

3.2.1 Simple residual plot with fitted variable

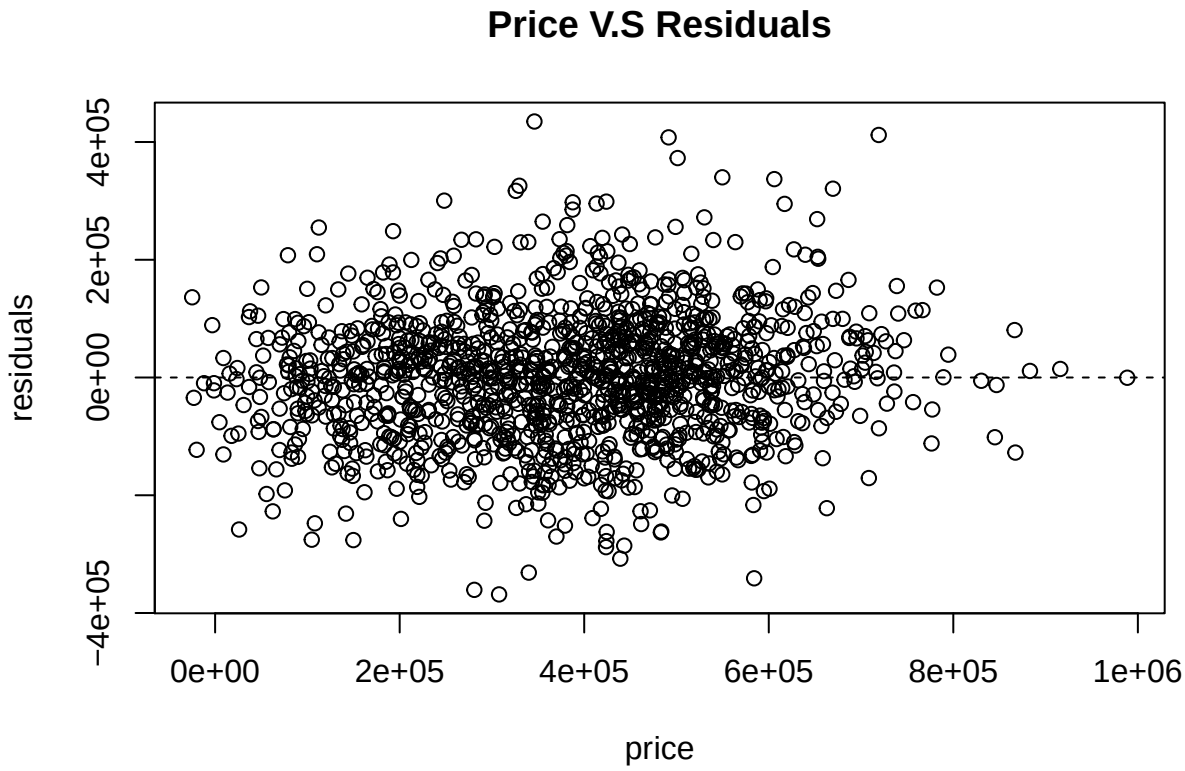
Since the assumption $\mathrm{E}(\epsilon_i) = 0$ also means $cov(\bar{r} + \hat{Y}) = 0$, the plot of residual versus response should be random distributed.

```
plot(fitted(fit), residuals(fit), xlab = "price",
     ylab ="residuals", main = "Price V.S Residuals" )
```



The above graph is the general residual plot with the fitted variable "price" without any transformation. There exist some outliers in the graph which hard to tell the distributions of the residual.

```r
price_filt = fitted(fit)[fitted(fit) < 1000000]
plot(price_filt, residuals(fit)[1:length(price_filt)], xlab = "price",
     ylab ="residuals", main = "Price V.S Residuals" )
abline(h=c(0), lty=2)
```

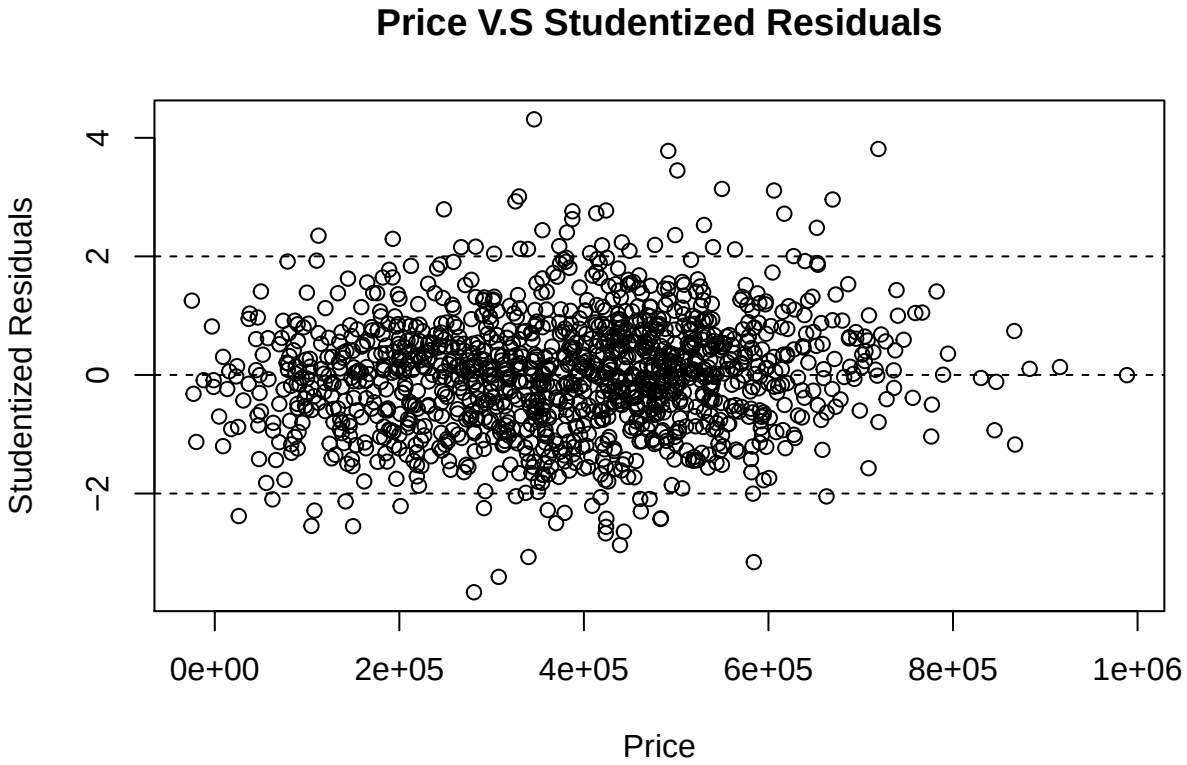


After removing the outliers, the distributions of the graph shows more clearly than before. As shown in the plot, there is no significant linear relationship between residual and fitted value.

3.2.2 Studentized residual plot with fitted variable

The studentized residual plot can also examine this assumption.

```r
plot(price_filt, rstudent(fit)[1:length(price_filt)], xlab = "Price",
     ylab ="Studentized Residuals", main = "Price V.S Studentized Residuals" )
abline(h=c(-2,0,2), lty=2)
```



The graph above is the studentized residual plot with fitted value without outliers. As shown in the plot, a large percent of data is in the range of (-2,2) with no significant distribution. However, there still some of the data outside this range. This implies response transformation might be needed for this model.

3.3 Checking the importance of model predictors

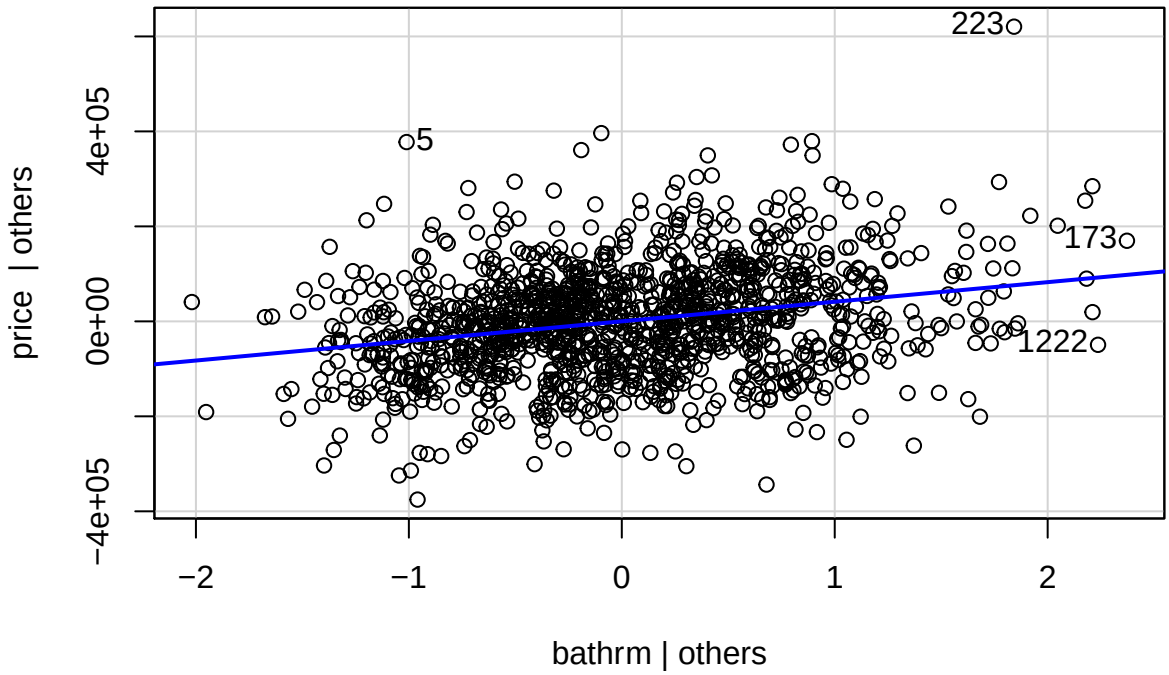
In this section, the model checking will focus on how effective the predictors is to the whole model. This process is important because the predictors which have a minor effect might increase the complexity of the linear model without improvement in the accuracy. This section will mainly examine the "saledate" and "bathrm" predictor which further justify the result from correlation coefficient.

The function that used to examine this question is the avPlots() function is the "car" library.

```
library(car)
```

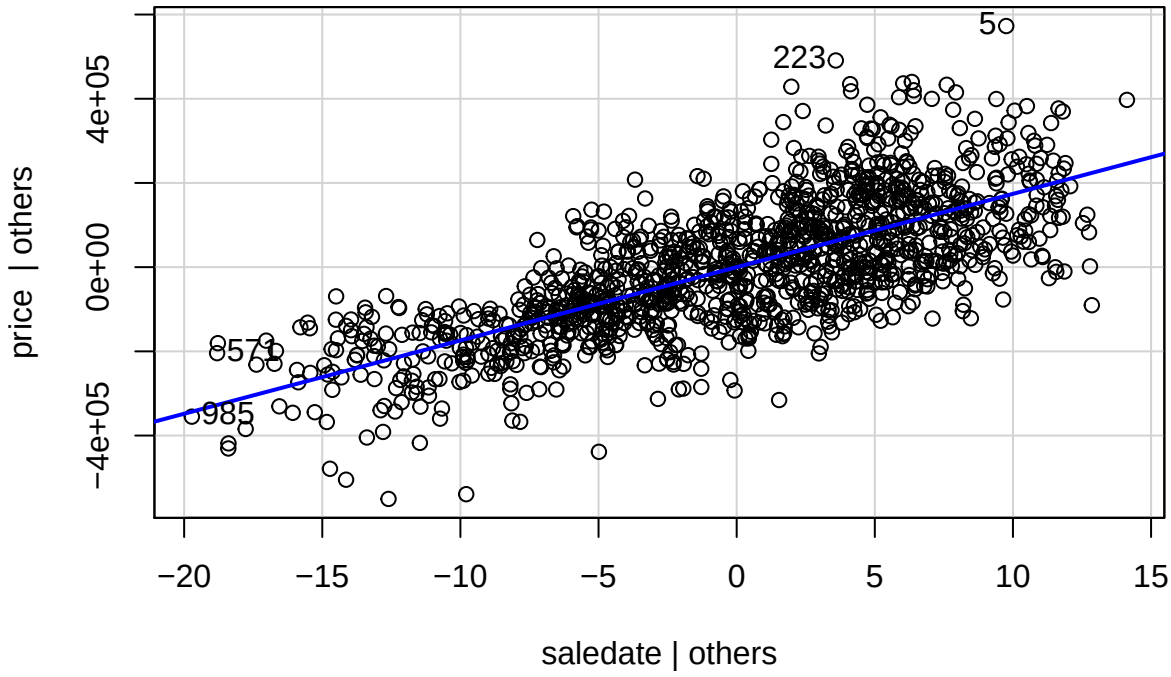
```
## Loading required package: carData
```

```
fitt <- lm(price ~ poly(saledate,4) + gba + grade + bathrm + fireplaces +
            poly(hf_bathrm, 2), data=dtrain)
avPlots(fitt,~bathrm)
```



This plot is the add-variable plot of "bathrm" predictor in the final linear model. As shown in the plot, the plot shows a linear pattern which indicates there exist a relationship between response variable price and predictor "bathrm"

```
fitt <- lm(price ~ saledate + gba + grade + poly(bathrm,2) + fireplaces +
             poly(hf_bathrm, 2), data=dtrain)
avPlots(fitt,~saledate)
```



This plot is the add-variable plot of "saledate" predictor in the final linear model. As shown in the plot, the plot shows a linear pattern and the pattern is more significant than the previous plot. Therefore, it indicates that predictor "bathrm" is important to the linear model.

```r
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: price
##                       Df     Sum Sq    Mean Sq  F value    Pr(>F)
## poly(saledate, 4)      4 2.6274e+13 6.5686e+12 552.8459 < 2.2e-16 ***
## gba                    1 1.0557e+13 1.0557e+13 888.5300 < 2.2e-16 ***
## grade                  6 3.4131e+12 5.6884e+11  47.8767 < 2.2e-16 ***
## poly(bathrm, 2)        2 1.1510e+12 5.7548e+11  48.4349 < 2.2e-16 ***
## fireplaces             1 5.5520e+11 5.5520e+11  46.7286 1.256e-11 ***
## poly(hf_bathrm, 2)     2 1.1277e+11 5.6385e+10   4.7456  0.008843 **
## Residuals           1286 1.5280e+13 1.1881e+10
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Applying F-test to the model that selected from above, all the p-value is much lower than 0.05. Therefore, there is a strong evidence against the null hypothesis (There is no relationship between predictor and response variable). All of these evidences approve the final linear model will be a good model for this dataset.

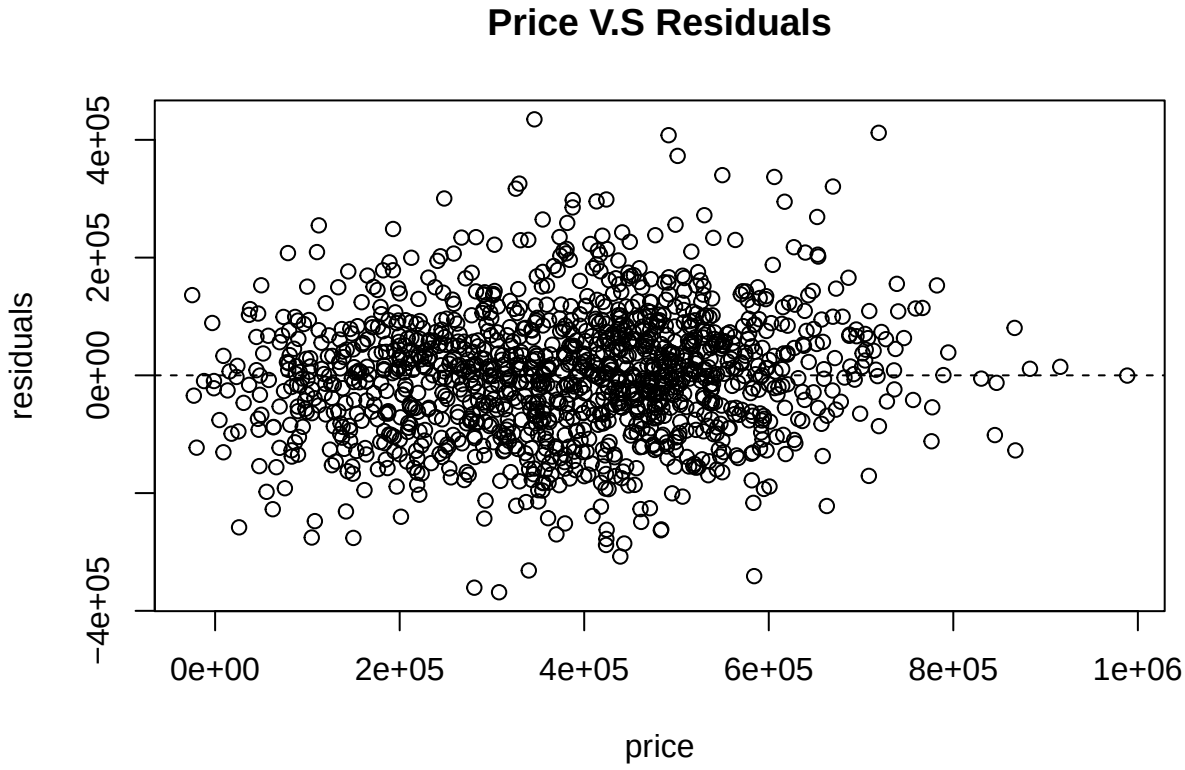
3.4 Checking Variance of the residuals

In this section, the model checking process will check the assumption of variance in the final linear model $V(\epsilon_i) = \sigma^2$. This assumption will affect the distribution of residual, which further affects the accuracy of the final linear model.

3.4.1 Constant Variance

In order to check whether the variances of residuals are constant or not, the residuals versus fitted value plot will be used.

```r
plot(price_filt, residuals(fit)[1:length(price_filt)], xlab = "price",
     ylab ="residuals", main = "Price V.S Residuals" )
abline(h=c(0), lty=2)
```

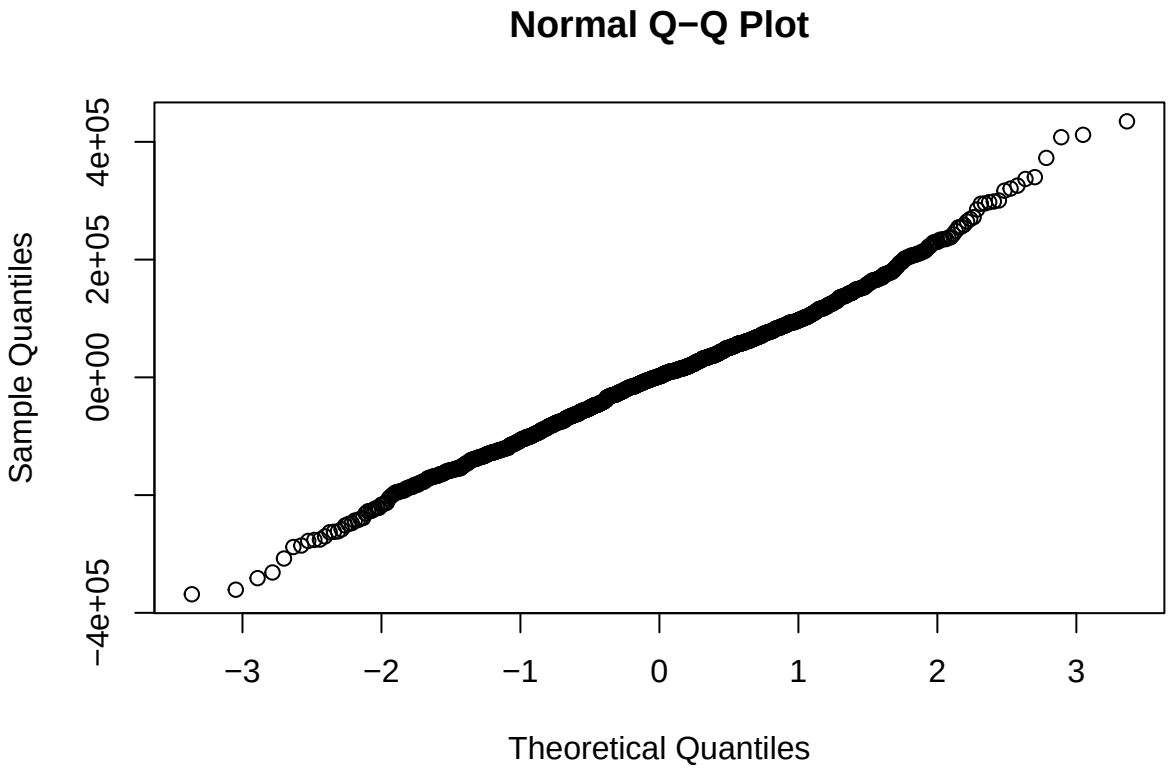


To examine the distance between residuals and zero. the plot without outliers will be used. From the plot, the distance between residuals dots to the zero lines are all approximately equal. There is no significant increase in distance as the price change. Therefore, this linear model has constant residual variance.

3.5 Normality of Residual

One of the assumptions for the residual of the linear model is that the residual follows the normal distribution. This section is going to check whether this assumption holds for the final linear model.

```
qqnorm(residuals(fit))
```



From the QQ-Plot of residual plot (Theoretical Quantiles V.S. Sample Quantiles), the pattern of the plot follows the forms for normal distribution. Therefore, this assumption holds.

3.6 Outliers

In the previous section, some of the plots are distracting by the outliers of the dataset. Since examing the pattern of the plot do not need detailed information about outliers, these outliers have been roughly filtered out using the rank function. In this section, model checking will focus on the outliers in the final data model. Using leverage, Cook's distances, and other techniques to exam the effect of outliers in the final model.

3.6.1 Leverage

Since categorical is hard to have any outliers, this section will mainly focus on the numerical variable in this predictors. Therefore, we first filters out the numerical predictors.

```
nums <- unlist(lapply(dtrain, is.numeric))
dim(dtrain[ , nums])
```

```
## [1] 1303   16
```

From the result, there are 14 (15 - price) numerical predictor and n = 1303. Next, fiting the final model will only numerical predictors. Since "grade" is a categorical predictor, it will not included in this model.

```
fit_num <- lm(price ~ poly(saledate,4) + gba + poly(bathrm,2) + fireplaces +
            poly(hf_bathrm, 2), data=dtrain[ , nums])
```

Finding the leverages of final model using the hatvalues() function.

```
head(hatvalues(fit_num))
```

```
##           1           2           3           4           5           6
## 0.008307986 0.072702120 0.008034303 0.008231444 0.005549019 0.007447531
```

To find the data that has irregular leverages, the mean of leverages is an important checker. All the irregular leverages should have a value higher than $2 \times \frac{p+1}{n}$.

```
hat_mean = (14 + 1) / 1303
leverage = data.frame(seq.int(1303) ,hatvalues(fit_num))[
hatvalues(fit_num) > 2 * hat_mean,1 ]
leverage
```

```
##  [1]    2   54   78  119  125  143  173  186  223  283  284  313  343  352  401
## [16]  471  509  527  535  556  557  558  571  603  666  671  716  723  793  858
## [31]  869  899  985 1065 1155 1169 1246
```

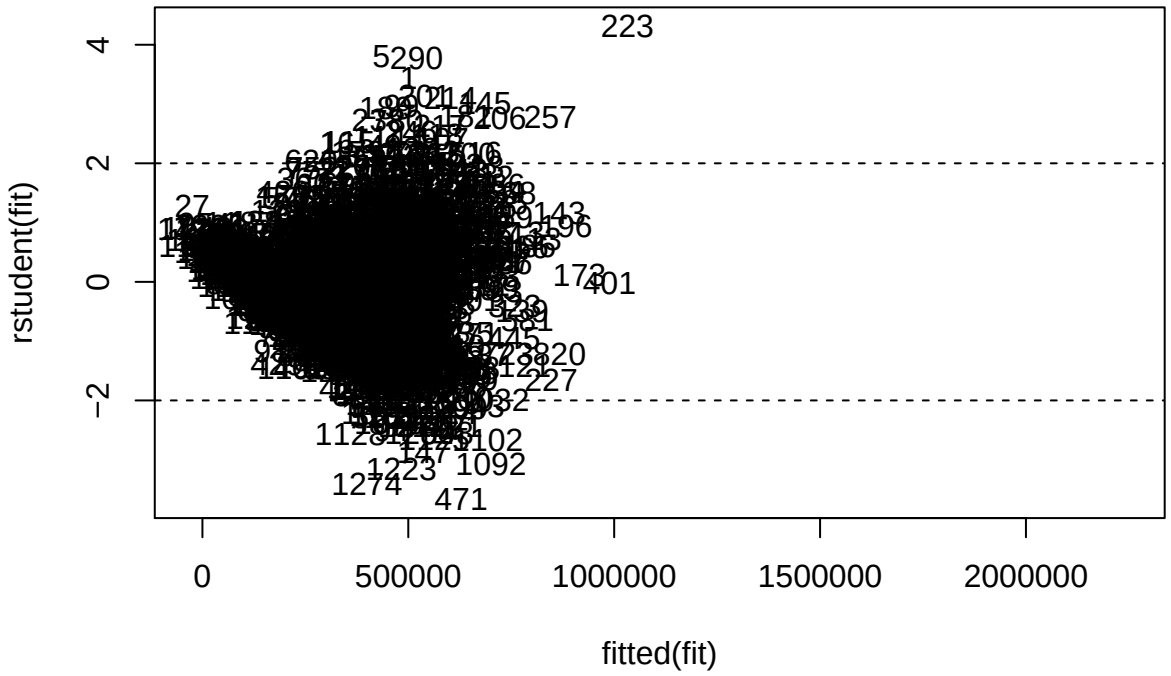
These are the indexes that has irregular leverages. However, data will be an influential case only if both leverage and residual are irregular. Therefore, student residual that larger than 2 or smaller than -2 should be listed out.

```
residual = data.frame(seq.int(1303) ,rstudent(fit_num))[
  rstudent(fit_num) > 2 | rstudent(fit_num) < -2, 1]
intersect(leverage,residual)
```

```
## [1]   2 223 471 557
```

These are the indexes that have both leverages larger than twice of mean and residual outside the range (-2,2). These are the outliers that have higher chances to effect the final prediction of the model.

```
plot(fitted(fit), rstudent(fit), type="n")
text(fitted(fit), rstudent(fit))
abline(h=c(-2, 2), lty=2)
```

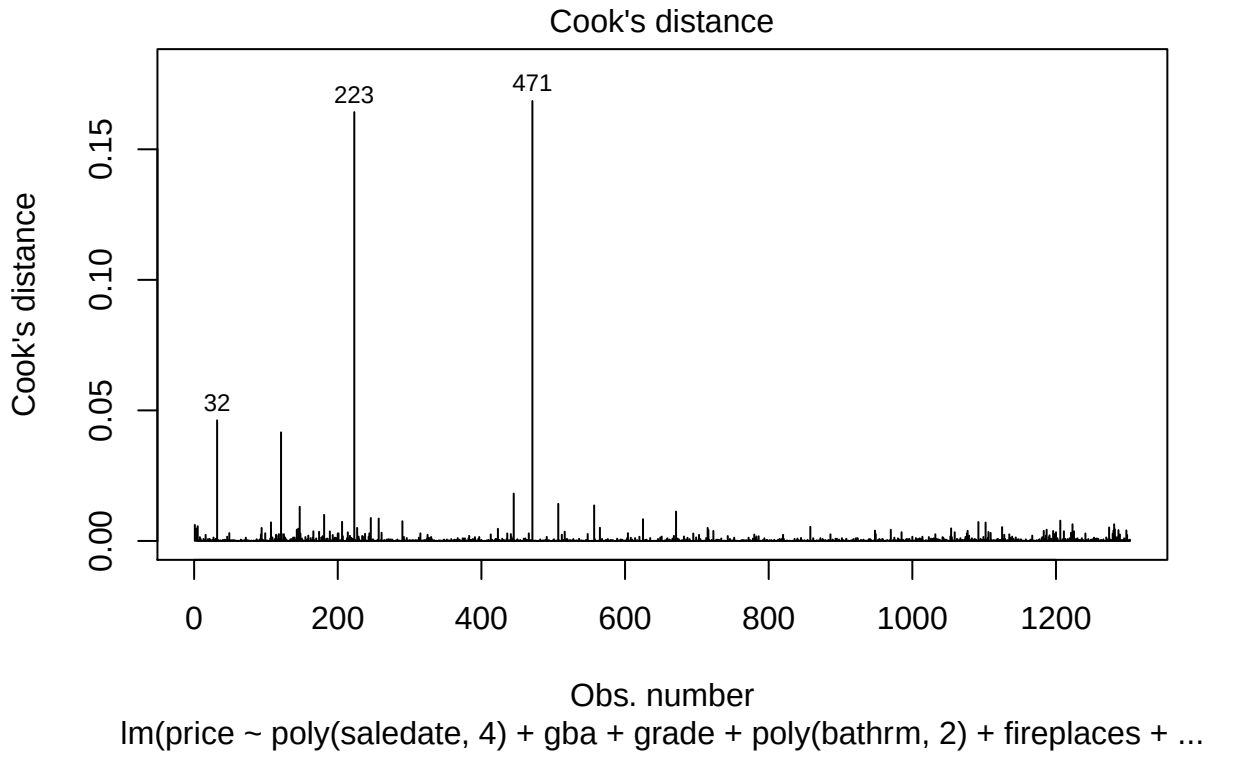


The studentized residual plot also shows the outliers in the previous calculation. The index 471 (bottom) and 223 (top right) are the significant outliers of this dataset in this plot.

3.6.2 Cook's Distance

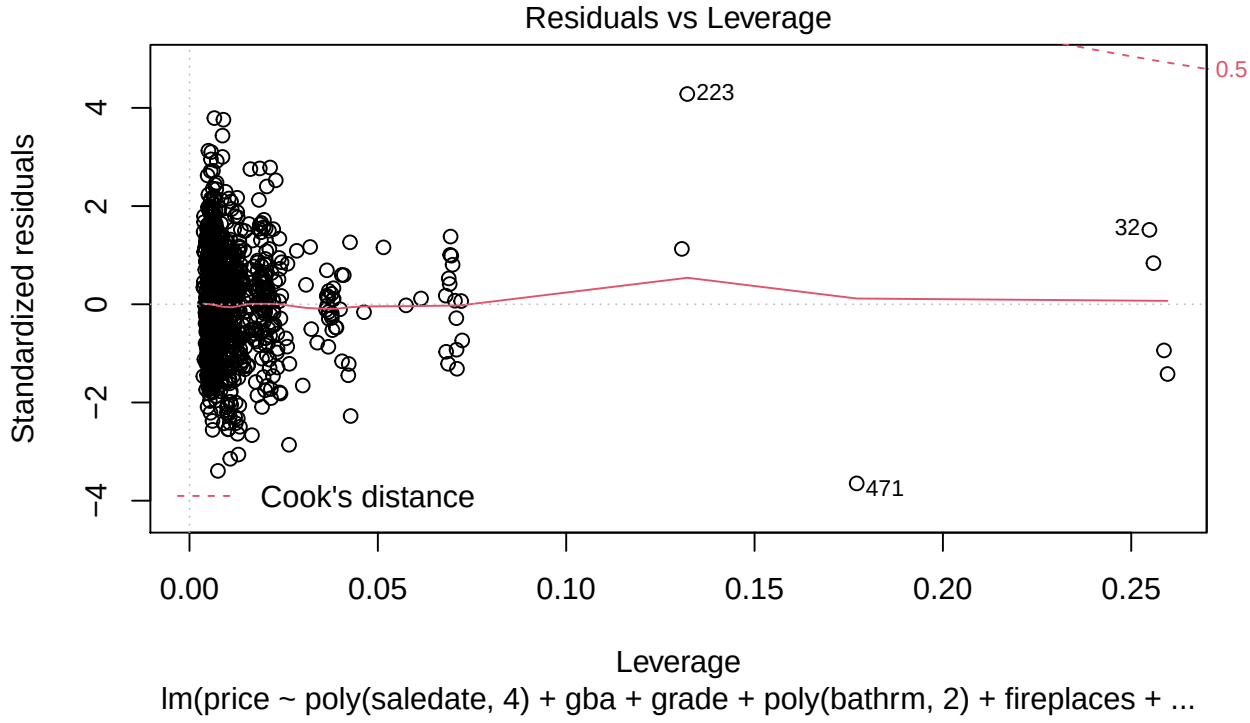
To further prove the accurancy of outliers in pervious section, this section will used cook's distance plot to find the influencial case of the dataset.

```
plot(fit, which = 4)
```



This Cook's distance plot also shows that 223 and 471 will be the outliers in the dataset. At the end, using another Cook's distance plot to check whether these outliers have significant effect to the accuracy of linear model.

```
plot(fit, which = 5)
```

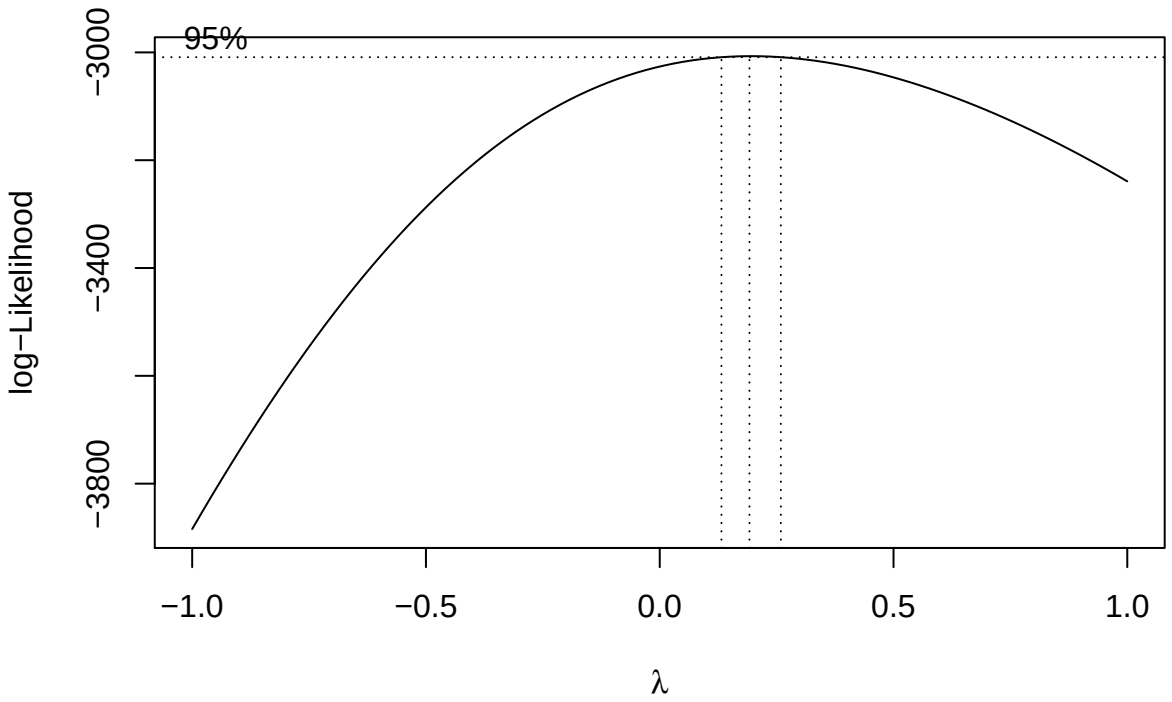


In this plot, {32,223,471}are listed out to be the outliers of the dataset. This result matches with previous calculation. However, all outliers are lying inside the range of 0.5. Therefore, all outliers will not make a significant effect on the accuracy of the model. Removing the outliers from the dataset will be optional.

4 Response transformation

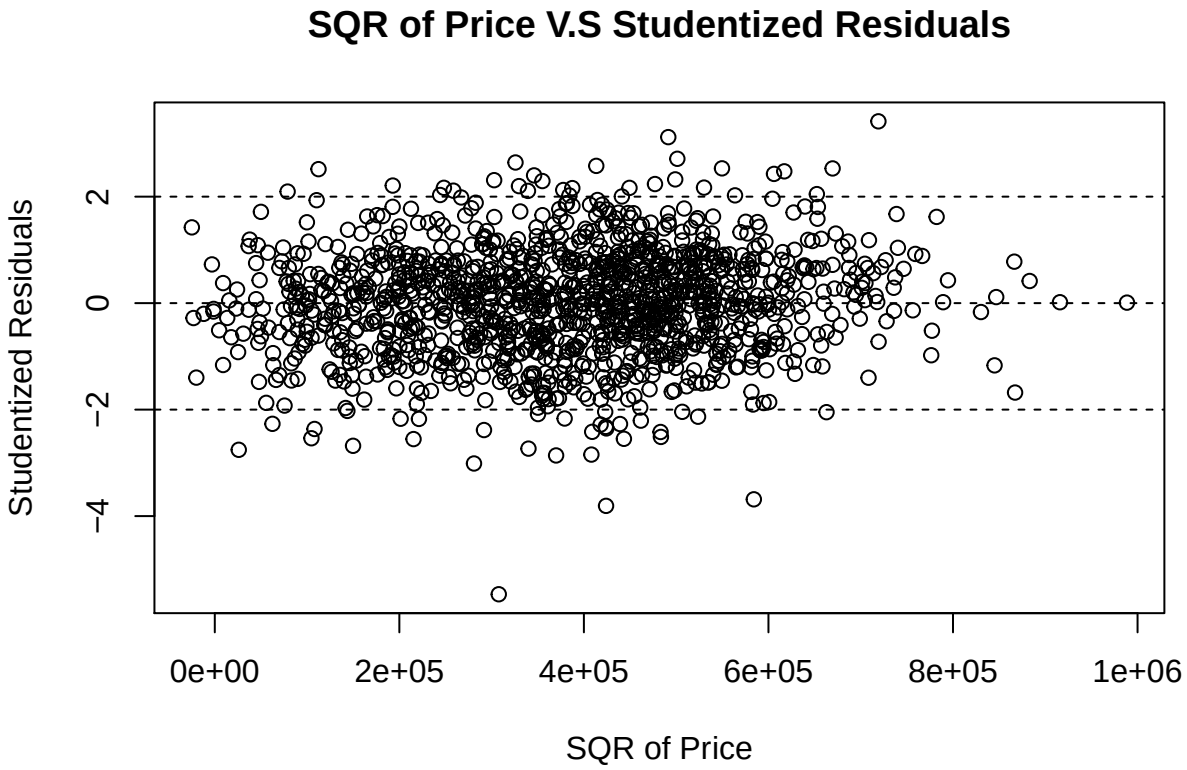
In the previous section, there are some data outside the range (-2,2) in the Studentized Residuals plot. One of the method to decrease these number is using response transformation.

```
library(MASS)
boxcox(fit, lambda = seq(-1,1,1/20))
```



The Boxcox plot shows the 95% confidence interval of lambda. In this case, lambda is close to 0.5. Therefore, a square-root transformation to the response transformation will be the fittest solution.

```
fit <- lm(sqrt(price) ~ poly(saledate,4) + gba + grade + poly(bathrm,2) + fireplaces +
            poly(hf_bathrm, 2), data=dtrain)
plot(price_filt, rstudent(fit)[1:length(price_filt)], xlab = "SQR of Price",
     ylab ="Studentized Residuals", main = "SQR of Price V.S Studentized Residuals" )
abline(h=c(-2,0,2), lty=2)
```



From the new studentized Residuals plot, the number of data inside the (-2,2) range has significantly increase. Therefore, square-root of response variable is a good transformation.

5 Model interpretation

In this report, we checked the distribution of the data to know the relationship between predictors and response. We also did the model checking and response transformation, the final linear model agrees on the assumption on mean, variance, and normality. Also, the final model's studentized residual holds in the range from -2 to 2. All of the results agree that the final linear model will be a good fit for this dataset. Therefore, the technique part of the model analysis finished. In this section, we will mainly focus on what does this linear model means in real life and how can this model support the improvements in the related field.

5.1 Interpretation on the predictors

$$\sqrt{Price} = saledate^4 + bathrm^2 + +hf\_bathrm^2 + gba + grade + fireplaces$$

```
summary(fit)$coefficients
```

```
##                          Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)           442.0443957 8.349529e+01  5.2942437  1.404108e-07
## poly(saledate, 4)1   3953.2801378 8.872265e+01 44.5577329 5.301355e-263
## poly(saledate, 4)2   -152.6970771 8.287412e+01 -1.8425182  6.562945e-02
## poly(saledate, 4)3   -272.9976507 8.383890e+01 -3.2562169  1.158507e-03
## poly(saledate, 4)4    706.0006471 8.309039e+01  8.4967783  5.319803e-17
## gba                     0.0598883 5.906597e-03 10.1392222  2.705651e-23
## gradeFair Quality     -45.3287624 8.569875e+01 -0.5289314  5.969443e-01
## gradeAverage           19.6676866 8.334884e+01  0.2359683  8.134948e-01
## gradeAbove Average     61.5031254 8.357346e+01  0.7359170  4.619154e-01
## gradeGood Quality     117.4833307 8.437861e+01  1.3923355  1.640615e-01
```

```
## gradeVery Good       129.6447719 9.344330e+01  1.3874164  1.655552e-01
## gradeSuperior        427.4161217 1.209577e+02  3.5335999  4.244706e-04
## poly(bathrm, 2)1    1025.7649160 1.049257e+02  9.7761105  8.026587e-22
## poly(bathrm, 2)2      44.1309739 8.538695e+01  0.5168351  6.053601e-01
## fireplaces            33.0932496 4.103584e+00  8.0644744  1.670043e-15
## poly(hf_bathrm, 2)1  257.0180248 8.874603e+01  2.8961073  3.842137e-03
## poly(hf_bathrm, 2)2 -128.1671233 8.316155e+01 -1.5411824  1.235184e-01
```

In this final linear model, 6 predictors affect the change of prices.

5.1.1 Sale Date

"sale date" represents the selling data of the house. From the previous section, there exists a strong relationship (correlation and add-variable plot) between the selling time of the house and the price of the house. Therefore, when predicting the price of the house in real life, we should not ignore the current situation in the economy. For example, inflation in the market, the global price of the house, and other factors might affect the prices of houses. From the coefficients of selling date predictors, a one year increase in date will result in 3953 dollars in averages price of houses. Besides, the result may also be affected by some big events that happened that year. For example, the subprime crisis in 2008.

5.1.2 Bathroom

"bathrm" represents the number of bathrooms in the house. The relationship between the number of bathrooms with house prices is a surprise result found in this linear model. In the preliminary analysis, we found that there exists a relationship between the number of bathrooms and the number of bedrooms. This result related to the pattern of house building. It implies as number bedrooms increase, the population in the house will increase too. Therefore, the number of bathrooms should also increase. However, in the model building process, we found that the correlation of bathroom with prices is higher than the bedroom and price 's correlation. This result implies that buyers care more about the environment of daily caring rather than the sleeping environment. As the number of bathroom increase by 1, the price of the house increases by 1376502.1662 dollars. The designer of houses should notice this result in future planning.

5.1.3 Half-Bathroom & Fireplace

Similar to "bathrm", "hf_bathrm" represents the number of half bathrooms and "fireplaces" represents the number of places for fire in the houses. The general definition of half bathroom is the bathroom that cannot be used for showering. The effect of half bathrooms is weaker than the full bathroom which means buyers think showering is an essential feature of a bathroom.

5.1.4 Gross Building Area

"gba" represents gross building area in square feet of the house. This is an important factor because the area of the houses will significantly affect the prices of the houses in real life. Since higher area means higher building cost and land used, the effect of gross building area on the prices is affecting by the construction company of the house. Also, this result represents buyers are willing to pay more money to buy a house that has a larger area. In real life, the price of the house often labels in the price per square feet. This result indicates that the final linear model agrees with this fact in the markets.

5.1.5 House quality

"grade" provides a ranking on how good the quality of the house is. This categorical data shows the rate of house quality affects the price of the house. From the coefficient, a Low of Fair quality will decrease 45 (in the rate) of house price, and a house with superior quality will increase the price by 427. The grade predictor is the only predictor in the model that might decrease the house price. This result implies that a house with a lower quality might have a strong negative effect on the final price of houses. The buyer of the house should also think about whether the given prices match with its house quality or not.

5.2 Summary

In summary, when the buyer decides to buy houses or agents are going to sell a house, six things might affect the price of the houses. First, when a buyer compares with the price in the past, they should think about current prices in the market and inflation. All of this might affect the final price of the house since the sale date is very important. The seller should also think about whether its a fair price in today's market. Second, when buyers visit the houses, they should focus on the number of bathrooms, half bathrooms, and fireplaces. All of these facilities will affect the price of the house. When agents are selling the house, they should explain to the buyers about these facilities. Third, the gross building area is a big factor that might affect house prices. Buyers should calculate the average price per square feet when they decided on the final price and agents should explain more about how big the house is in order to raise the final price. At last, both buyers and agents should not ignore the effect of house quality to the final prices. Since bad house quality might affect the living environment of the houses, the buyer should be ranking on how good the quality is in their minds to match with the final price. Therefore, all of these factors might affect the final price of the house, buyers and agents should learn more information and decides the price of the house base on these situations. Besides, some special conditions might apply to individual houses and buyers, it might also affect the final price.