HOUSING PRICE PREDICTION USING R

Course: ALY 6020

Name: Yuanying Li

Table of content

1.INTRODUCTION

Housing prices are an important reflection of the economy, and housing price ranges are of great interest for both buyers and sellers. In this project. housing price will be predicted given explanatory variables that cover many aspects of residential houses. As housing price, they will be predicted with various regression techniques including basic Linear, Lasso, Ridge, SVM regression, and Random Forest regression; We will also perform PCA to improve the prediction accuracy.

The goal of this project is to create a regression model and a classification model that are able to accurately estimate the price of the house given the features. In this project, we use the dataset named "housing.csv" and choose Linear Regression Model and Random Forest to predict the price.

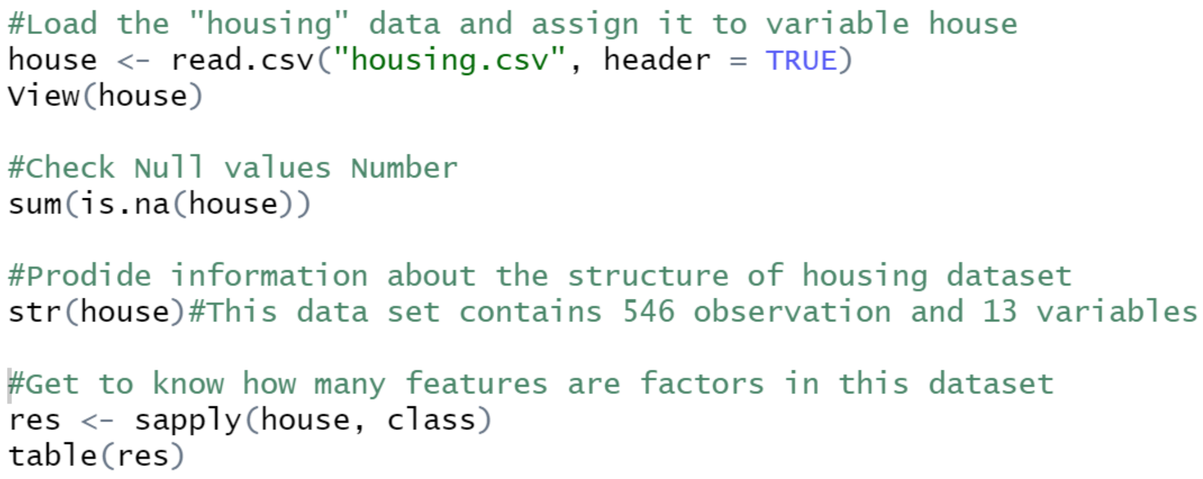
2.DATA AND PREPROCESSING

2.1 Define the decision variables

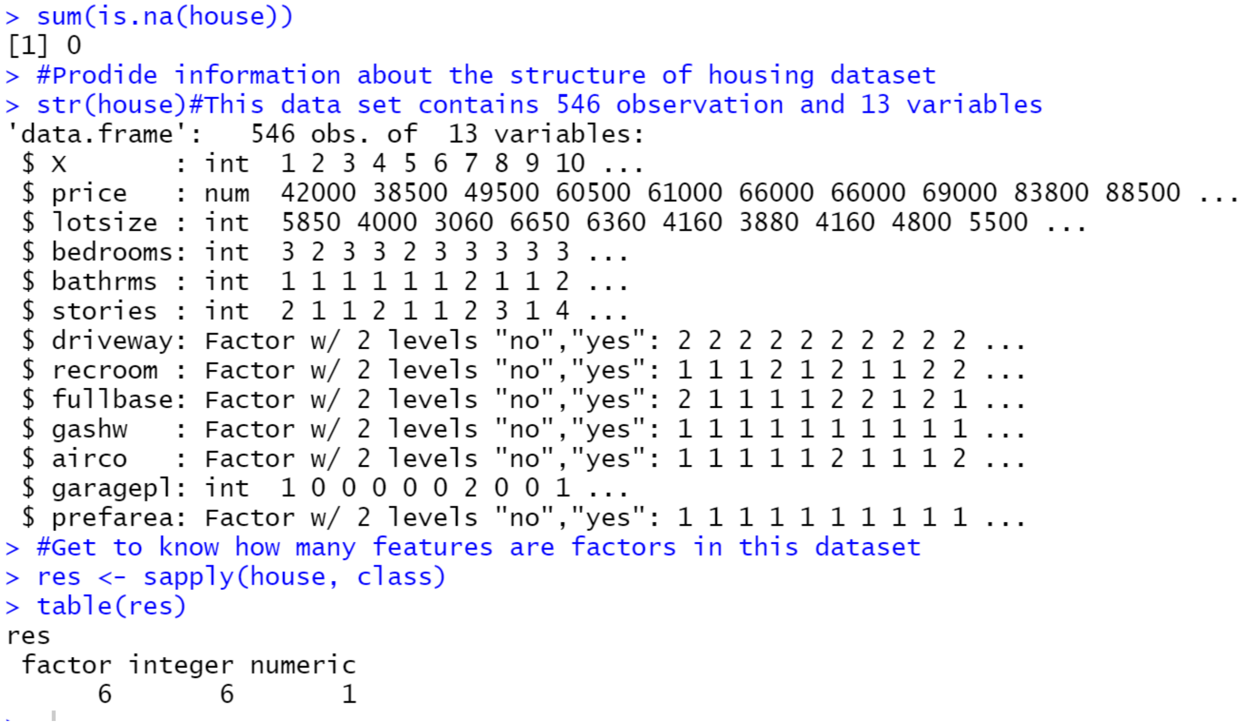
Here is the explanation of features' name in dataset.

- Price: the price of house
- Lotsize: the size of housr
- Bedrooms: the number of bedrooms
- Bathrms: the number of bathrooms
- Stories: the number of storages
- Driveway: if the house has the driveway
- Recroom: if the house has the recreational room
- Fullbase: : if the house has the full finished basement
- Gashow: if the house has the gas for heating
- Airco: if the house has the gas for heating
- Prefarea: if the house located in preferred area

2.2 Exploratory Data Analysis



```
#Load the "housing" data and assign it to variable house
house <- read.csv("housing.csv", header = TRUE)
View(house)

#Check Null values Number
sum(is.na(house))

#Prodide information about the structure of housing dataset
str(house)#This data set contains 546 observation and 13 variables

#Get to know how many features are factors in this dataset
res <- sapply(house, class)
table(res)
```



```
> sum(is.na(house))
[1] 0
> #Prodide information about the structure of housing dataset
> str(house)#This data set contains 546 observation and 13 variables
'data.frame':	546 obs. of  13 variables:
 $ X       : int  1 2 3 4 5 6 7 8 9 10 ...
 $ price   : num  42000 38500 49500 60500 61000 66000 66000 69000 83800 88500 ...
 $ lotsize : int  5850 4000 3060 6650 6360 4160 3880 4160 4800 5500 ...
 $ bedrooms: int  3 2 3 3 2 3 3 3 3 3 ...
 $ bathrms : int  1 1 1 1 1 1 2 1 1 2 ...
 $ stories : int  2 1 1 2 1 1 2 3 1 4 ...
 $ driveway: Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
 $ recroom : Factor w/ 2 levels "no","yes": 1 1 1 2 1 2 1 1 2 2 ...
 $ fullbase: Factor w/ 2 levels "no","yes": 2 1 1 1 1 2 2 1 2 1 ...
 $ gashw   : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ airco   : Factor w/ 2 levels "no","yes": 1 1 1 1 1 2 1 1 1 2 ...
 $ garagepl: int  1 0 0 0 0 0 2 0 0 1 ...
 $ prefarea: Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
> #Get to know how many features are factors in this dataset
> res <- sapply(house, class)
> table(res)
res
 factor integer numeric 
      6       6       1 
```

Observations:

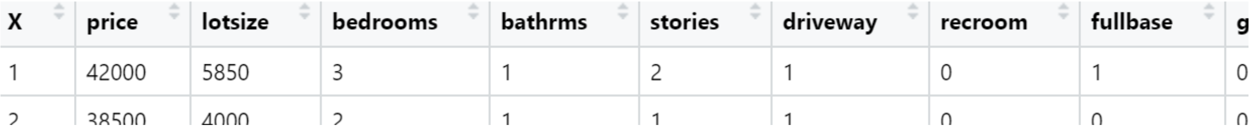
1. Sample size is not that large (within 546), with 12 variables, the dependent variable is called price. There is not missing value in this dataset.

2. The dataset contain some features which are factor data type(6 factor, 6 integer, 1 numeric), we should covert those columns to numeric and fit the model.

2.3 Data preprocessing

In this study we firstly focus on six factors that possibly influence housing price, including driveway, recreational room, full finished basement, gas for heating, central air conditioning and whether located in preffered neighbourhood of the city. Each of them has 2 levels, namely 'yes' or 'no'. We need to covert features driveway, recroom, fullbase, gashw, airco into numeric value to fit the model, in data set '1' represents 'yes', no represents 'no'. Here is the new dataset.

```
#Convert features which are belongs to factor into a numberic value column.
a <- sub("no","0", house$driveway)
b <- sub("yes","1",a)
house$driveway <- b
house$driveway <- as.numeric(house$driveway)
```



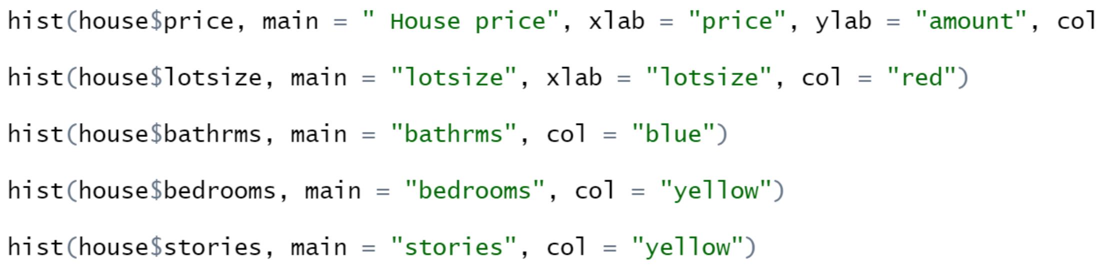
X	price	lotsize	bedrooms	bathrms	stories	driveway	recroom	fullbase	g
1	42000	5850	3	1	2	1	0	1	0
2	38500	4000	2	1	1	1	0	0	0

Since the requirement of analysis has not mentioned the preferred area, so we drop the column prefarea.

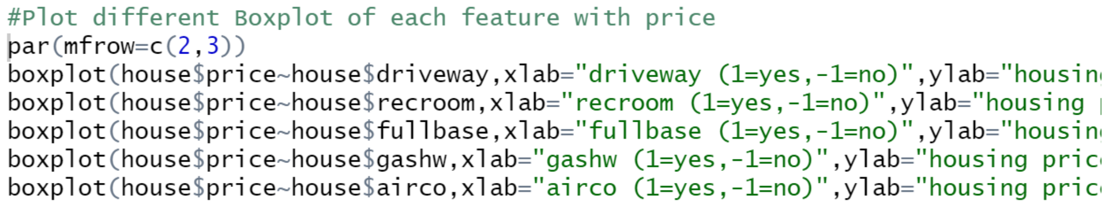
2.4 Date visualization

2.4.1 Histogram and Boxplot

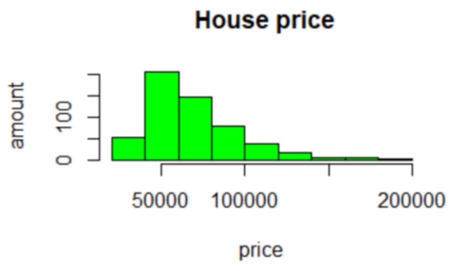
we would like to know something about price range and how much is the average price for a house, so we plot histogram to see the distribution of each of each feature and plot different Boxplot of each feature with price.

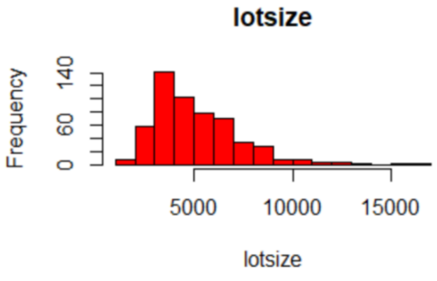


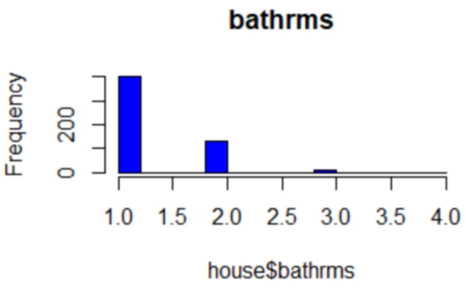
```
hist(house$price, main = " House price", xlab = "price", ylab = "amount", col

hist(house$lotsize, main = "lotsize", xlab = "lotsize", col = "red")

hist(house$bathrms, main = "bathrms", col = "blue")

hist(house$bedrooms, main = "bedrooms", col = "yellow")

hist(house$stories, main = "stories", col = "yellow")
```

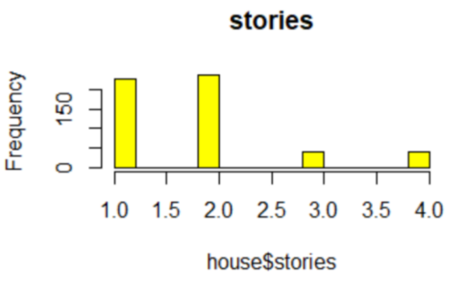


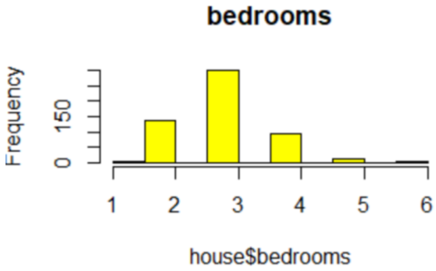
```
#Plot different Boxplot of each feature with price
par(mfrow=c(2,3))
boxplot(house$price~house$driveway,xlab="driveway (1=yes,-1=no)",ylab="housin
boxplot(house$price~house$recroom,xlab="recroom (1=yes,-1=no)",ylab="housing
boxplot(house$price~house$fullbase,xlab="fullbase (1=yes,-1=no)",ylab="housin
boxplot(house$price~house$gashw,xlab="gashw (1=yes,-1=no)",ylab="housing pric
boxplot(house$price~house$airco,xlab="airco (1=yes,-1=no)",ylab="housing pric
```

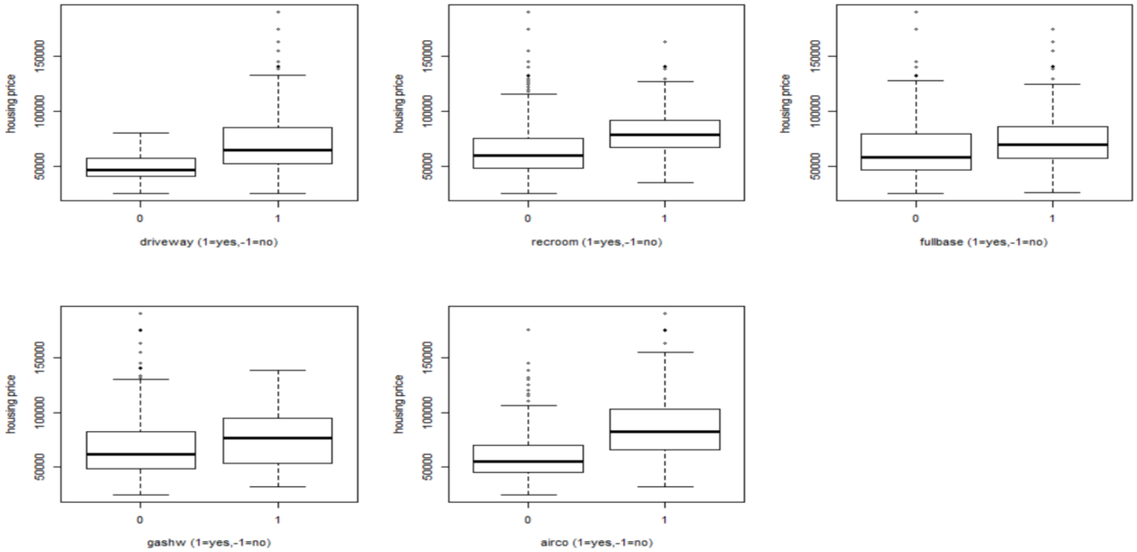












Observation:

1. According to the distribution of house price, we could see the majority of the house price is $50,000 - $80,000, so if we would to buy a normal house, we can buy the price range between $50,000 - $80,000, otherwise, it would be too high or low.

2. When Check the lotsize, the range of house size would be 2,000 - 6,000, the rest of size is outlier.

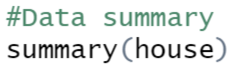
3. Combining the histogram of bathroom, storage, bedroom, it is easily to see the mode of those 3 features are 1, 2, 3. However, we assume that we would buy a house that has 4 bedrooms, 2 bathrooms, 2 stories so maybe this type of house is not usual to see.

4. From boxplots above, it seems all five factors could influence the price of a house. Especially driveway, recreational room and gas supply, means are different whether these are provided or not. However for other 2 features means are only slightly different.

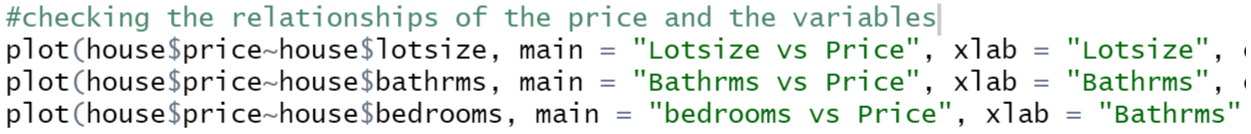
2.4.2 Summary Data

Use the summary command to calculate each feature's mathematics detail.

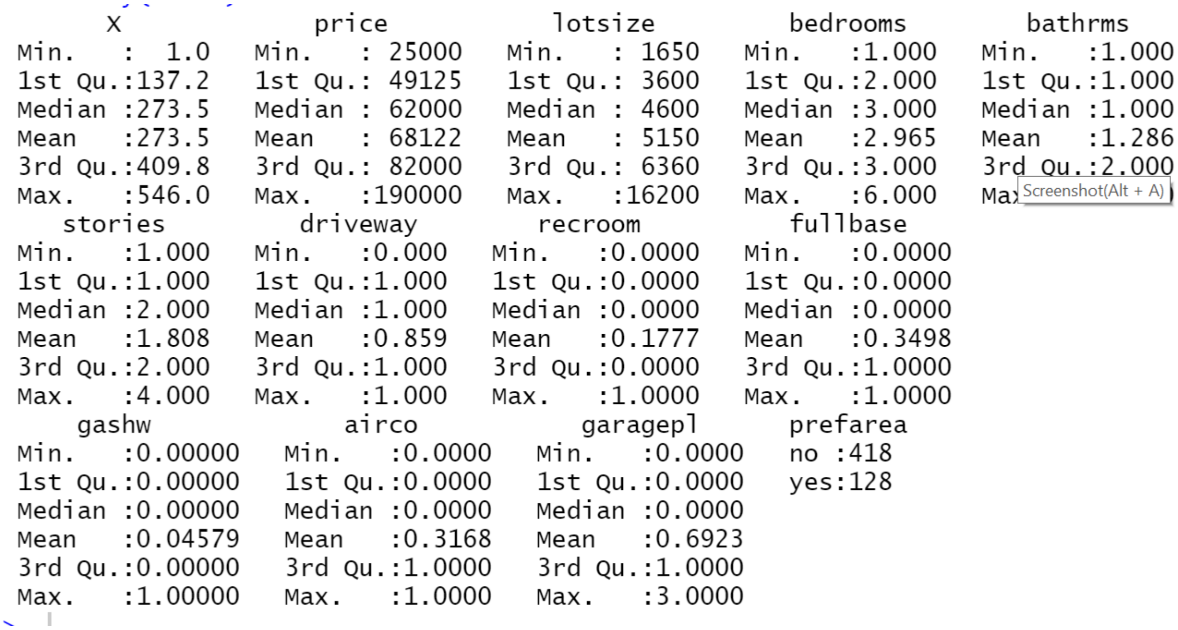
As we could see, the mean of the price is $62,000, we would use this number to compare with the housing price after prediction.



```
#Data summary
summary(house)
```



```
#checking the relationships of the price and the variables
plot(house$price~house$lotsize, main = "Lotsize vs Price", xlab = "Lotsize",
plot(house$price~house$bathrms, main = "Bathrms vs Price", xlab = "Bathrms",
plot(house$price~house$bedrooms, main = "bedrooms vs Price", xlab = "Bathrms"
```

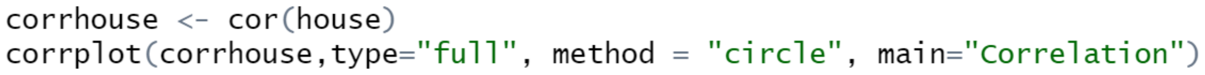


```
       X              price           lotsize         bedrooms         bathrms     
 Min.   :  1.0   Min.   : 25000   Min.   : 1650   Min.   :1.000   Min.   :1.000  
 1st Qu.:137.2   1st Qu.: 49125   1st Qu.: 3600   1st Qu.:2.000   1st Qu.:1.000  
 Median :273.5   Median : 62000   Median : 4600   Median :3.000   Median :1.000  
 Mean   :273.5   Mean   : 68122   Mean   : 5150   Mean   :2.965   Mean   :1.286  
 3rd Qu.:409.8   3rd Qu.: 82000   3rd Qu.: 6360   3rd Qu.:3.000   3rd Qu.:2.000  
 Max.   :546.0   Max.   :190000   Max.   :16200   Max.   :6.000   Ma
    stories         driveway         recroom          fullbase     
 Min.   :1.000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:0.0000  
 Median :2.000   Median :1.000   Median :0.0000   Median :0.0000  
 Mean   :1.808   Mean   :0.859   Mean   :0.1777   Mean   :0.3498  
 3rd Qu.:2.000   3rd Qu.:1.000   3rd Qu.:0.0000   3rd Qu.:1.0000  
 Max.   :4.000   Max.   :1.000   Max.   :1.0000   Max.   :1.0000  
     gashw             airco           garagepl      prefarea 
 Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   no :418  
 1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   yes:128  
 Median :0.00000   Median :0.0000   Median :0.0000            
 Mean   :0.04579   Mean   :0.3168   Mean   :0.6923            
 3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:1.0000            
 Max.   :1.00000   Max.   :1.0000   Max.   :3.0000            
```

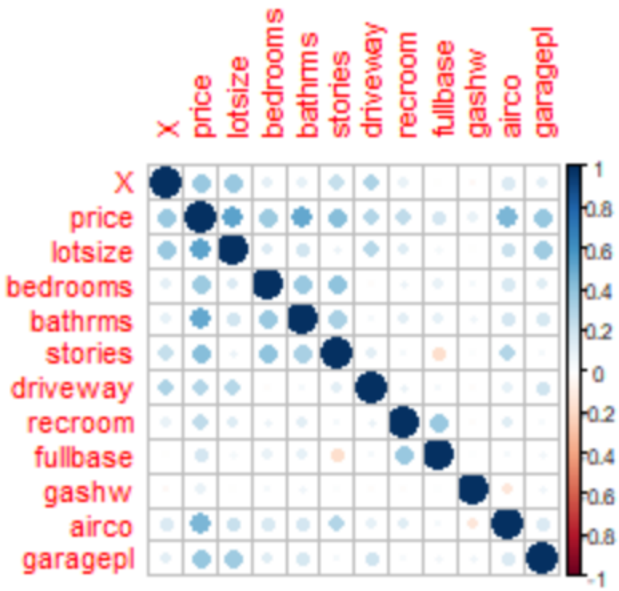
3.BUILD THE MODEL

3.1 Select variables that may have great impact on house price

Firstly, find correlation amongst all the features and price and plot.

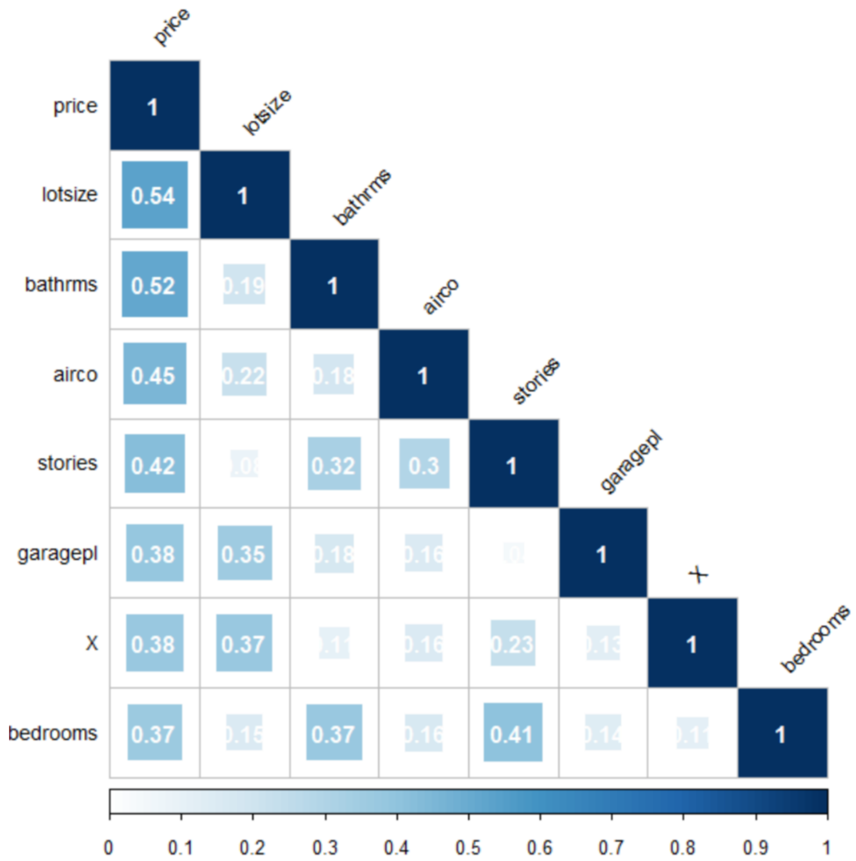


```
corrhouse <- cor(house)
corrplot(corrhouse,type="full", method = "circle", main="Correlation")
```



However, it is difficult to visualize the correlation of each column, so we sort the order with descend order, and keep the correlation coefficient which are over 0.3 then drop the rest of variables. The following is the code and result.

```
corr<-cor(house)
name <- names(which(sort(abs(corr[, "price"]), decreasing = T) > 0.3))
corrplot(cor(house[,name]),title = "Correlation Plot",method="square",typ
```



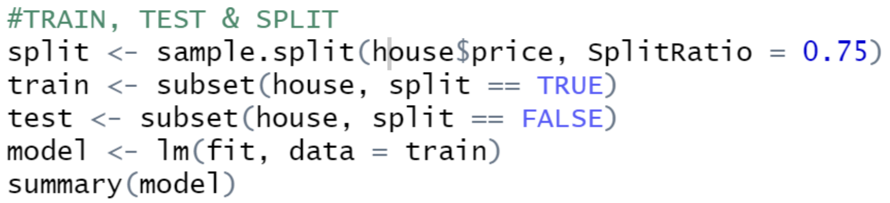
Observation:

1. the greatest impact is lotsize, the following is bathrms, airco, stories, garagepl, bedrooms, so there are 6 variable we would use to build the model.

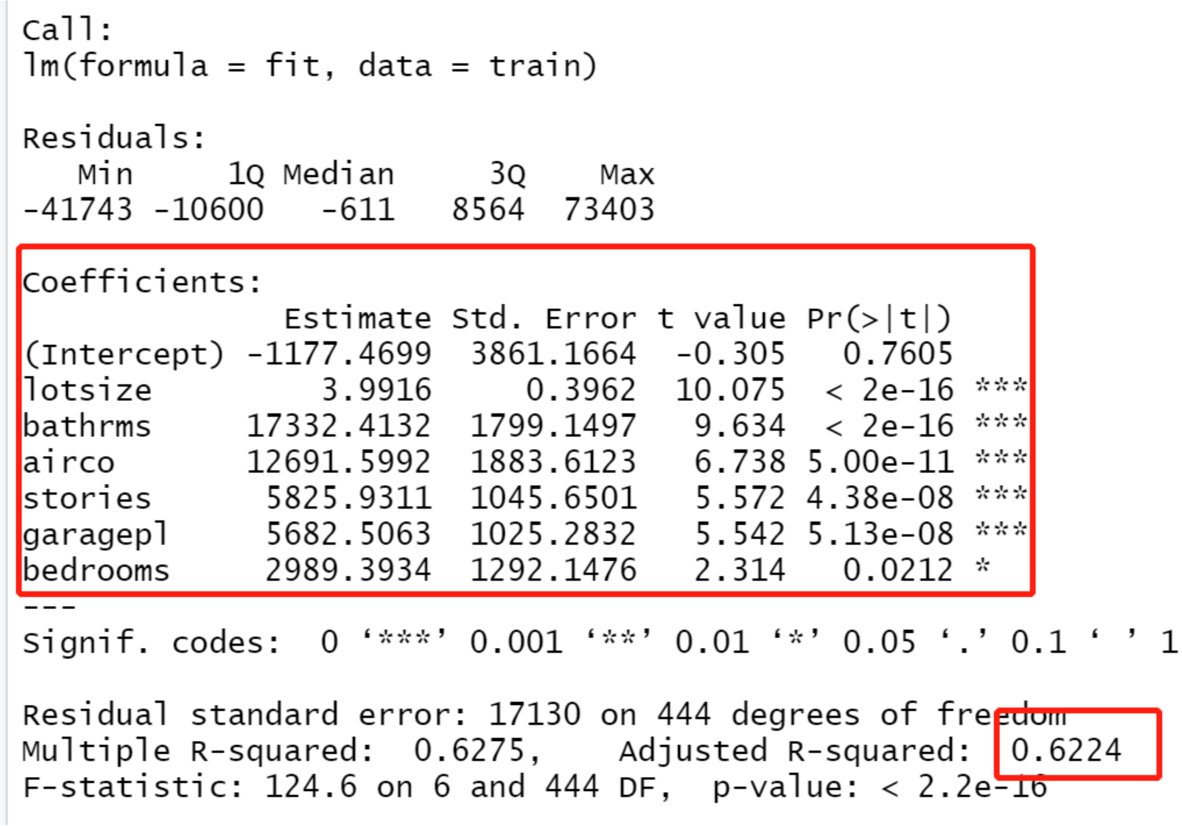
3.2 Use Linear Regression

Create Multiple Linear Regression Model, Choose the features of correlation coefficient what are over 0.3 to be the independent variable of the model. Also, split the dataset into train dataset and test dataset. The code and result would be showed as below.

```
fit <- price ~ lotsize + bathrms + airco + stories + garagepl + bedrooms
```



```
#TRAIN, TEST & SPLIT
split <- sample.split(house$price, SplitRatio = 0.75)
train <- subset(house, split == TRUE)
test <- subset(house, split == FALSE)
model <- lm(fit, data = train)
summary(model)
```



```
Call:
lm(formula = fit, data = train)

Residuals:
   Min     1Q Median     3Q    Max 
-41743 -10600   -611   8564  73403 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1177.4699  3861.1664  -0.305   0.7605    
lotsize         3.9916     0.3962  10.075  < 2e-16 ***
bathrms     17332.4132  1799.1497   9.634  < 2e-16 ***
airco       12691.5992  1883.6123   6.738 5.00e-11 ***
stories      5825.9311  1045.6501   5.572 4.38e-08 ***
garagepl     5682.5063  1025.2832   5.542 5.13e-08 ***
bedrooms     2989.3934  1292.1476   2.314   0.0212 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17130 on 444 degrees of freedom
Multiple R-squared:  0.6275,	Adjusted R-squared:  0.6224 
F-statistic: 124.6 on 6 and 444 DF,  p-value: < 2.2e-16
```

Observation:

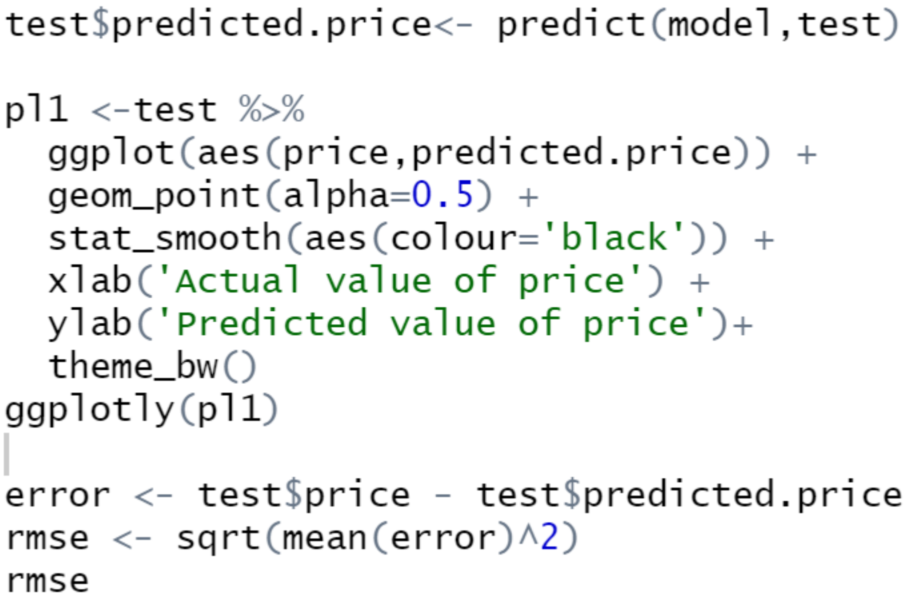
1. Coefficients: The model would be price = 3.9916lotsize + 17332.4132bathrms + 12691.5992airco + 5825.9311stories + 5682.5063garagepl + 2989.3934bedrooms

2. Residuals: residuals represent the difference between real value and prediction, the max residual is 73403, it means the max residual would be 73403, and we would plot the residual plot to check the difference.

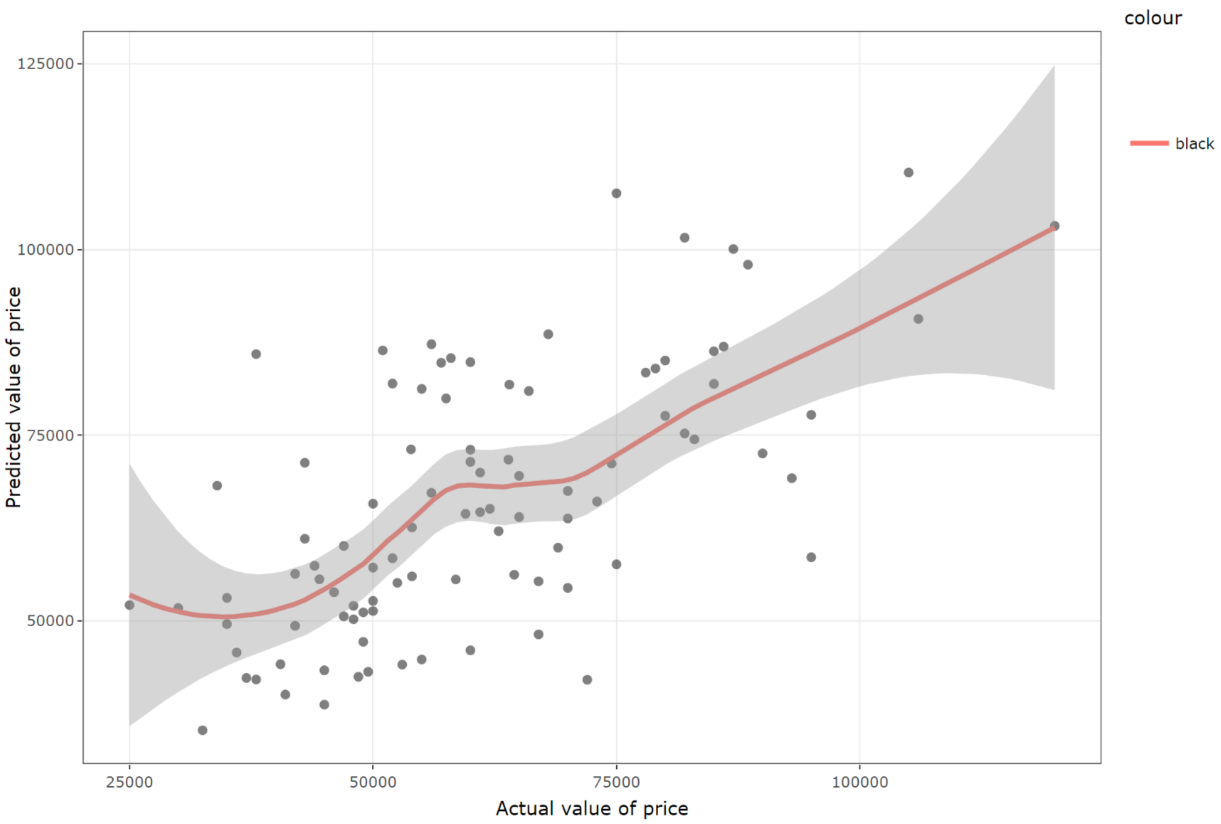
3. R-squared (Coefficient of determination):
Also known as the determination coefficient of model fitting, the value between 0 and 1 is closer to 1, indicating that the model's dependent variable has a stronger explanatory ability to the response variable Y.
Adjusted R-squared
When the number of independent variables increases, R square will also increase even if the linear relationship between some independent variables and Y is not significant.Since Adjusted R Square increases the penalty for increasing variables, we use it as the basic criterion for determining whether the model is good or bad. So the R-square of this model is 0.6224.

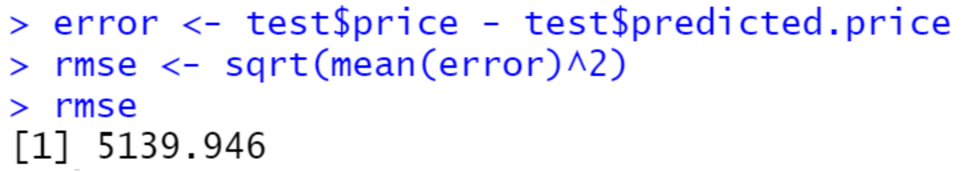
3.3 Evaluation Model - Linear Regression

After building the model, evaluate the model is one of the important parts to do, so we predict the price from test model and check the accuracy comparing with the real price and visualize the residuals through plot the difference



```
test$predicted.price<- predict(model,test)

pl1 <-test %>%
  ggplot(aes(price,predicted.price)) +
  geom_point(alpha=0.5) +
  stat_smooth(aes(colour='black')) +
  xlab('Actual value of price') +
  ylab('Predicted value of price')+
  theme_bw()
ggplotly(pl1)

error <- test$price - test$predicted.price
rmse <- sqrt(mean(error)^2)
rmse
```

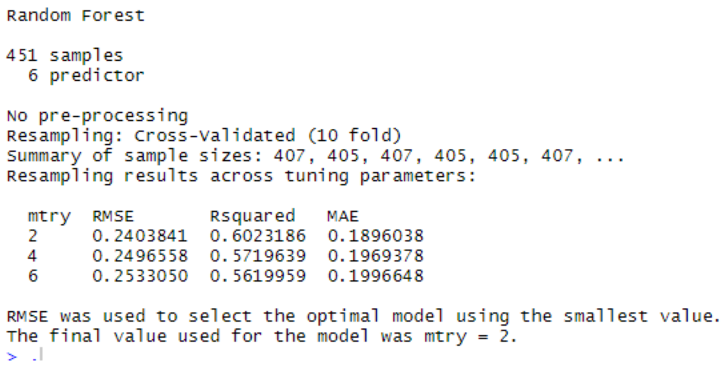




```
> error <- test$price - test$predicted.price
> rmse <- sqrt(mean(error)^2)
> rmse
[1] 5139.946
```

According to the plot, we could see the residual is a little large, and RMSE is 5139, it means the model is not so good. So we change the another model to predict.

3.4 Random Forest Model



```
Random Forest

451 samples
  6 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 407, 405, 407, 405, 405, 407, ...
Resampling results across tuning parameters:

  mtry  RMSE       Rsquared   MAE
  2     0.2403841  0.6023186  0.1896038
  4     0.2496558  0.5719639  0.1969378
  6     0.2533050  0.5619959  0.1996648

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was mtry = 2.
>
```

We use Random Forest to predict the price, and the result shows Adjusted R-square is 0.6023.

4.CONCLUSION

We used two different models above to predict the price, and the final result of our submitted answers are as follows.

Model	Adjusted R-squared
Linear Regression	0.6224
Random Forest	0.6023

According to the Adjusted R- squared value, we could choose Linear Regression as our model. If we would like to buy 4 bedrooms, 2 bathrooms, 2 storied house with approx. lot size of 5500 SFT using this following formula, we could get the house price is $80,228.05, so if house price is about $80,228.05. it is a reasonable price, if it is not, we should think the price is too high or too low.

price = 3.9916lotsize + 17332.4132bathrms + 12691.5992airco + 5825.9311stories + 5682.5063garagepl + 2989.3934bedrooms

However, if we just use mean to predict the price, the prediction price is 68122, which is much lower than the price using model. It may make a mistake prediction when buy a house in real life. What's more, when Linear Regression was implemented, we could also use PCA to cross-validated its number of dimension given the tuned regularization coefficient. To obtain an lower test error, we might need to cross-validate both the number of dimension after PCA as well as regularization coefficient together, though this would be more computationally intensive.
For Random Forest Classification/Regression, besides the depth, we might need to examine further variations to optimize this algorithm, such as considering the splits of nodes, the requirements of leaf nodes, etc.

5.REFERENCE

[1] De Cook, Dean. "Ames, Iowa: Alternative to theBoston Housing Data as an End of Semester Regression Project." Journal of Statistics
Education, vol. 19, no. 3, 2011.