CS 520 Mathematical Modeling Project

A Statistician's Priority List for Boosting Home Value

To: Prof. Allali, Mohamed

By Li, Shuo (Olin)

Introduction

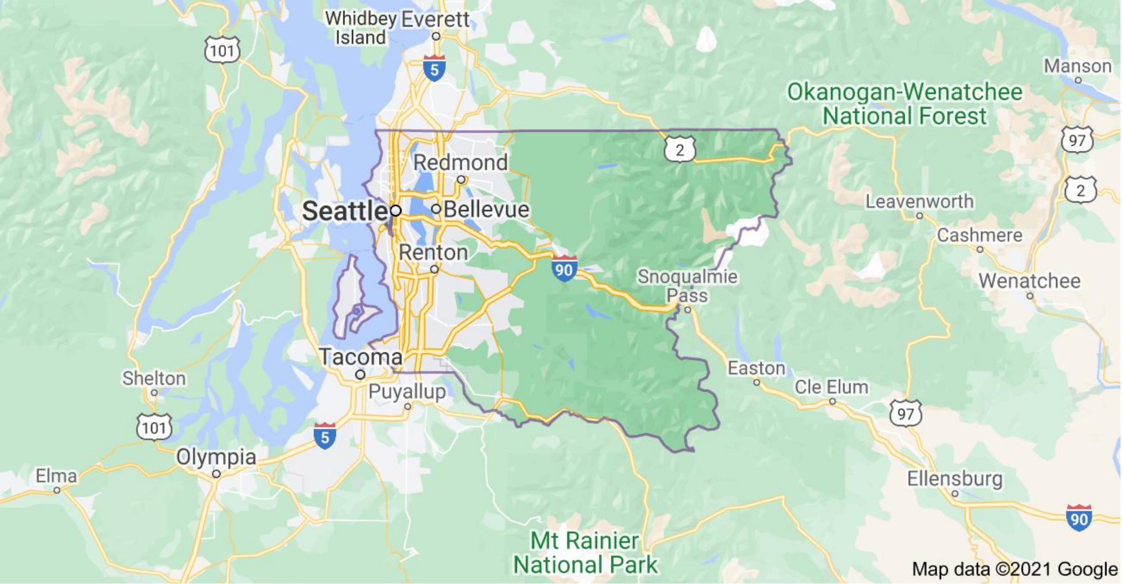
- Investment. Hence, how to maintain and boost home value during their stay has been a question for many house owners.
- Factors that would help increase home values with a limited budget.
- The home owners can do the things that really matter and bring up the sale prices of the house by prioritizing the controllable things.
- To obtain a priority list for house improvement and home value bringing up, statistical methods like linear regression and random forest would be utilized in this project to analyze a Kaggle dataset containing house sale prices of King County, Washington from May 2014 to May 2015.
- The primary objective of this project is to find out a priority list for home improvement that can be helpful for boosting home values from a statistician′ s view.

Be rich!

Background

- As the most populous county in Washington with the largest city of the state, Seattle, sitting in the west, King County embraces quite a few large corporations (e.g., Boeing, Microsoft, Amazon). The map below highlights the exact location of King County.

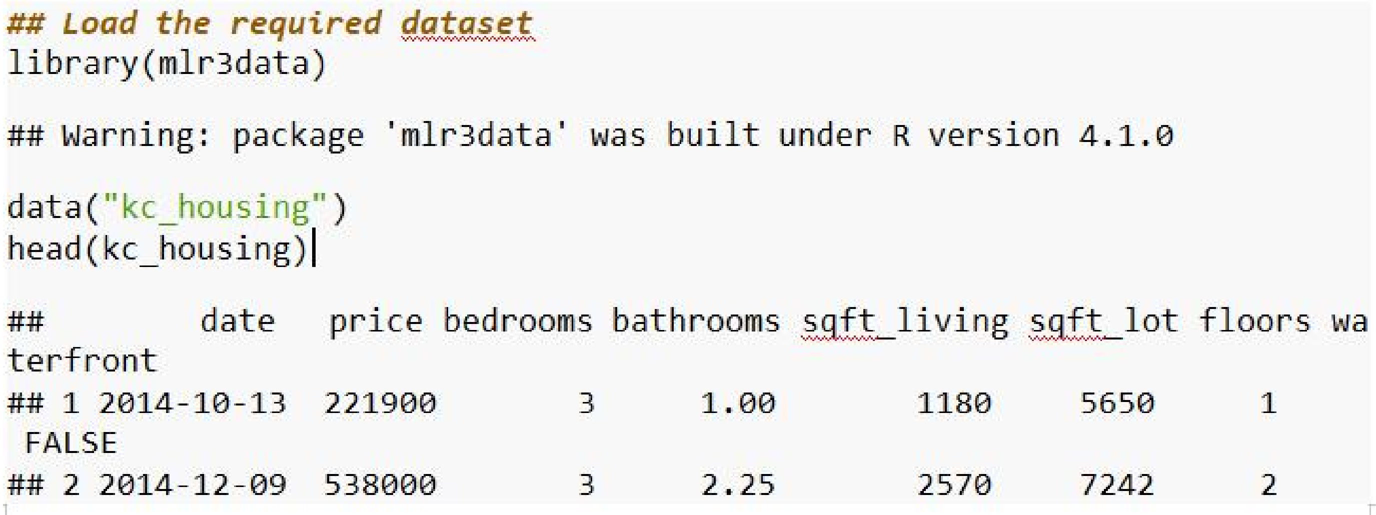
- The strong local economy has been driving the increase in the number of households in this area, which has been creating the demand from the real estate market. The graphic below displays the right-skewed distribution of house sale prices in the time window we analyzed. The house price in King County ranges from $75,000 to $7,700,000, and the average price is slightly above $540,000.



Map data ©2021 Google

Data

- A real-world dataset that contains house sale price information and the corresponding house features of King County, Washington from 2014 to 2015 will be used. It is originated from Kaggle, and can be imported to R from mlr3data package.
- Basically, there are 21,613 observations along with 19 house features such as the number of bathrooms, bedrooms, floors, and square footage of the house in the original data. The code that help us load the data and print the first few lines of the original data is shown as below



```
## Load the required dataset
library(mlr3data)

## Warning: package 'mlr3data' was built under R version 4.1.0

data("kc_housing")
head(kc_housing)

##          date   price bedrooms bathrooms sqft_living sqft_lot floors wa
terfront
## 1 2014-10-13  221900        3      1.00        1180     5650      1
 FALSE
## 2 2014-12-09  538000        3      2.25        2570     7242      2
```

Variables

- id: unique ID of the house
- date: the sale date of the house
- price: the final sale price of the house
- bedrooms: count of bedrooms in the house
- bathrooms: count of bathrooms in the house
- sqft_living: square footage of the living area in the house
- sqft_lot: square footage of the lot for the house
- floors: total levels in the house
- waterfront: whether the house has a waterfront view. If yes, the value is 1. Otherwise, the value is 0.
- view: how many times the house has been viewed
- condition: the overall condition of the house
- grade: the overall grade given to the housing unit by King County grading system. According to King County Assessor′ s webpage, this represents the construction quality of improvements. Grades run from grade 1 to 13.
- sqft_basement: square footage of the basement
- sqft_above: square footage of the house apart from the basement
- yr_built: which year the house was built
- yr_renovated: which year the house was renovated. If no renovation has been done, the value is 0
- zipcode: the zip code for the house address
- lat: latitude coordinate of the house location
- long: longitude coordinate of the house location
- sqft_living15: square footage of the living area in the house measured in 2015
- sqft_lot15: square footage of the lot for the house measured in 2015
- renovated: whether the house has been renovated. If yes, the value is 1. Otherwise, the value is 0.
- basemt: whether the house has basement. If yes, the value is 1. Otherwise, the value is 0.

Packages

ggplot2 and **lattice**: Data visualization and randomForest for random forest models that can be helpful to analyze the effects of the house factors on the house price;

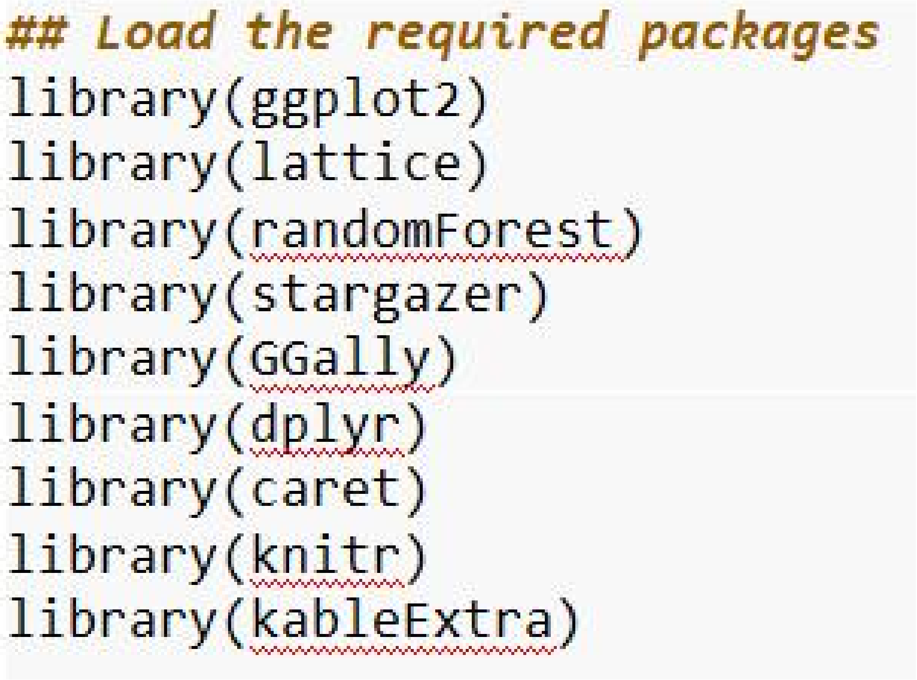
stargazer: Offer neat and more readable model results of linear regressions;

GGally: Obtain the correlation matrix;

dplyr: Manipulate and modify data frames;

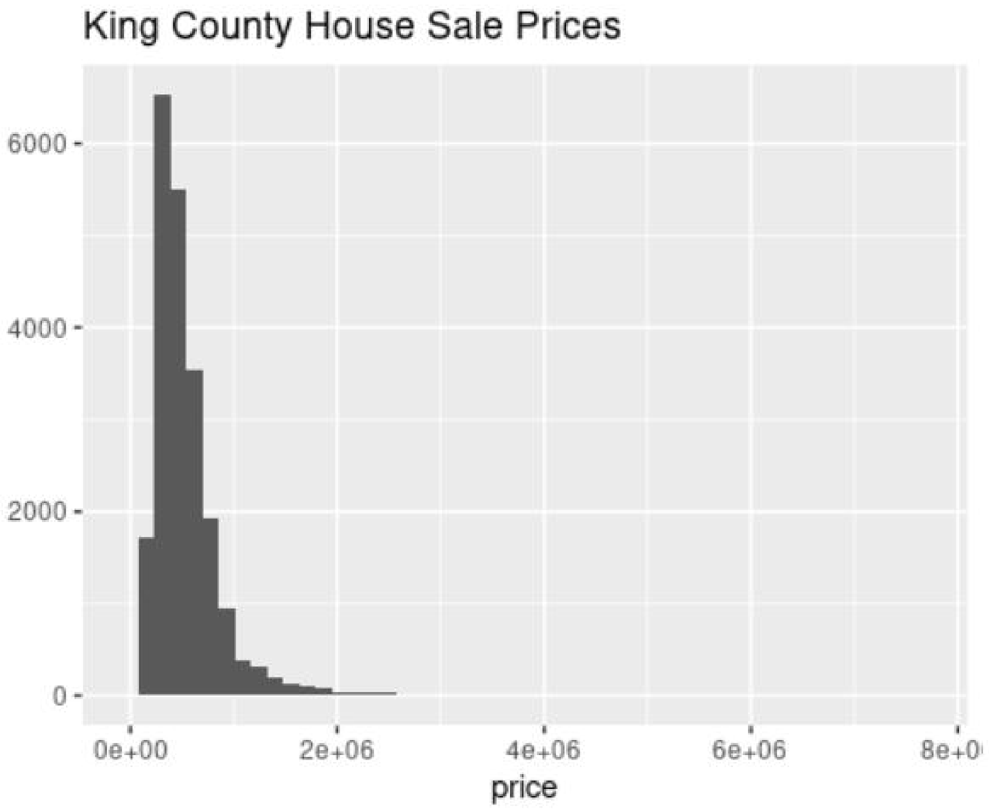
caret: Tune random forest models so that the model performs best regarding cross-validation root mean square errors (CV RMSEs) can be found;

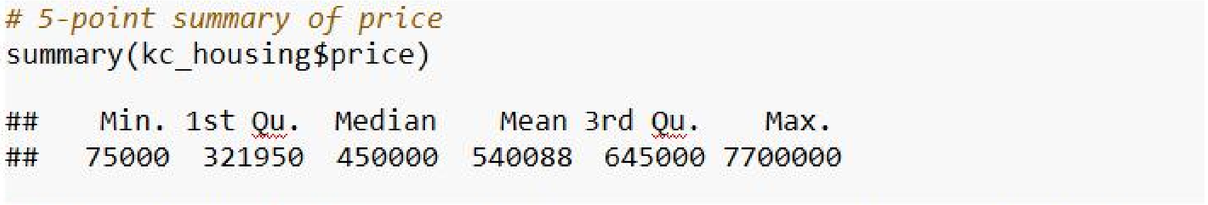
knitr and **kableExtra**: Generate a neat and well-formatted final document via R markdown.



```
## Load the required packages
library(ggplot2)
library(lattice)
library(randomForest)
library(stargazer)
library(GGally)
library(dplyr)
library(caret)
library(knitr)
library(kableExtra)
```

Exploratory Data Analysis (EDA)

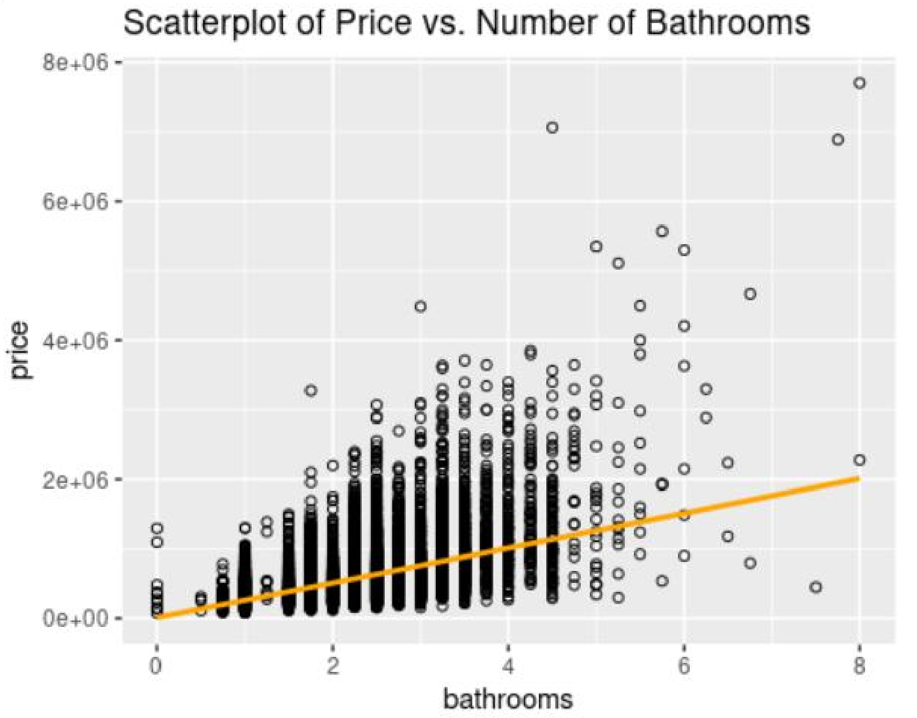


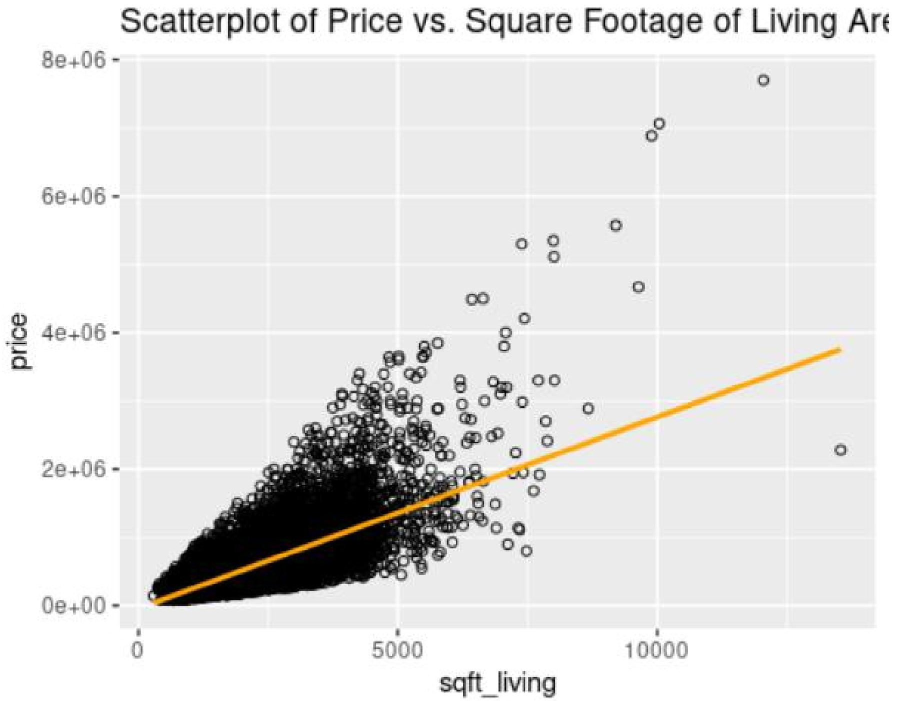


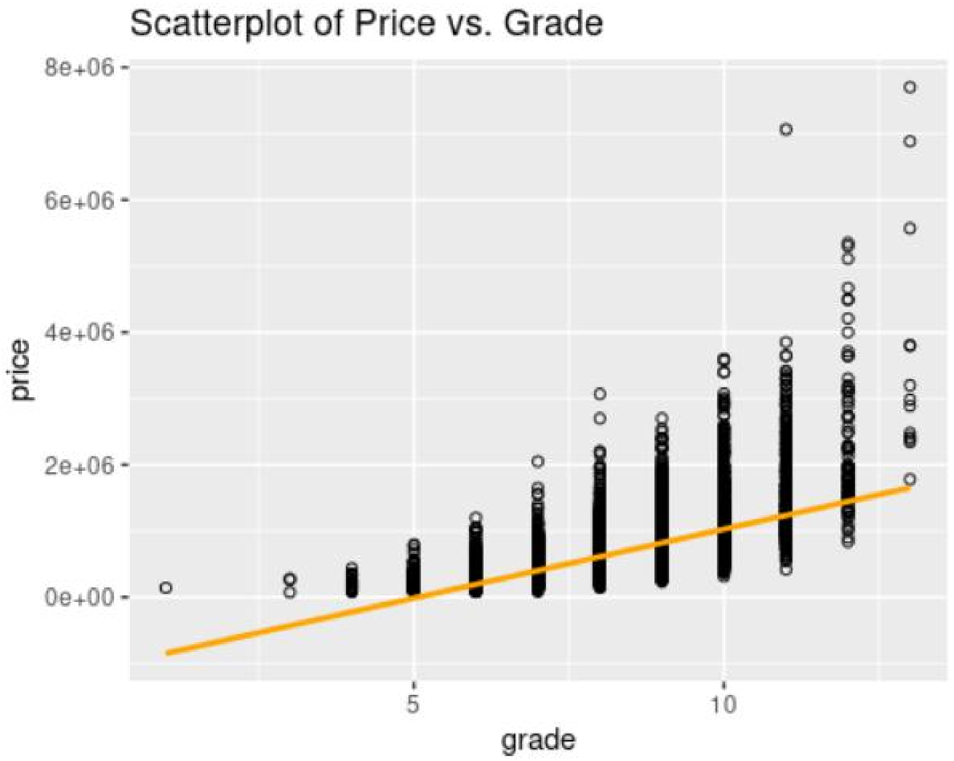
```
# 5-point summary of price
summary(kc_housing$price)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   75000  321950  450000  540088  645000 7700000
```

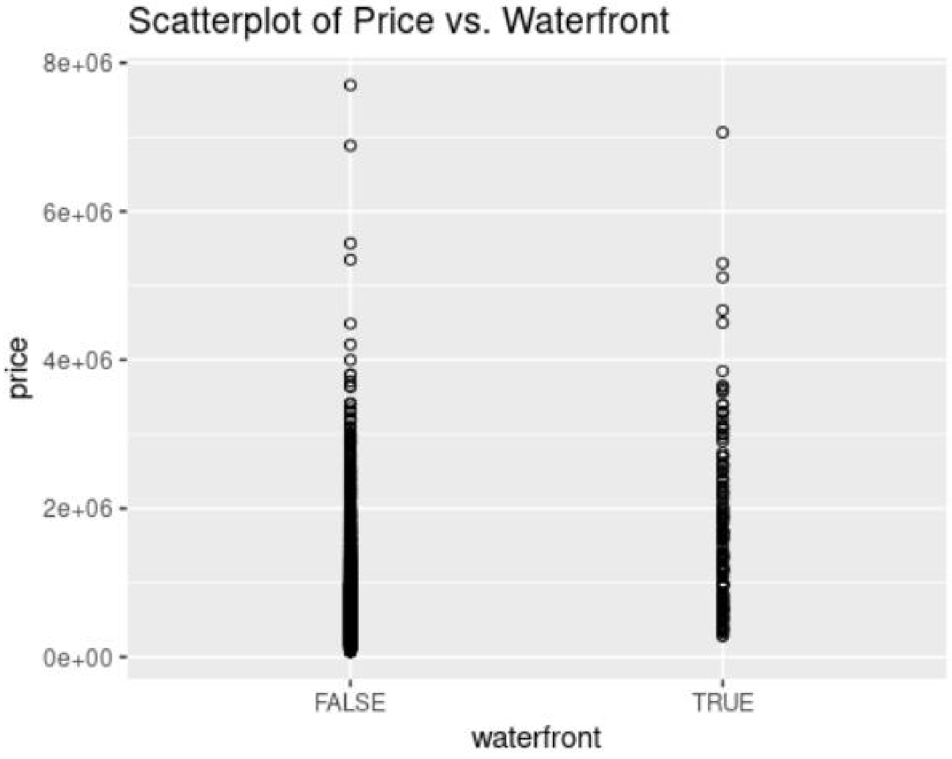
It is clear that the distribution of the home price is positively skewed with a long right tail, which implies that some houses are expected to have higher values than others.





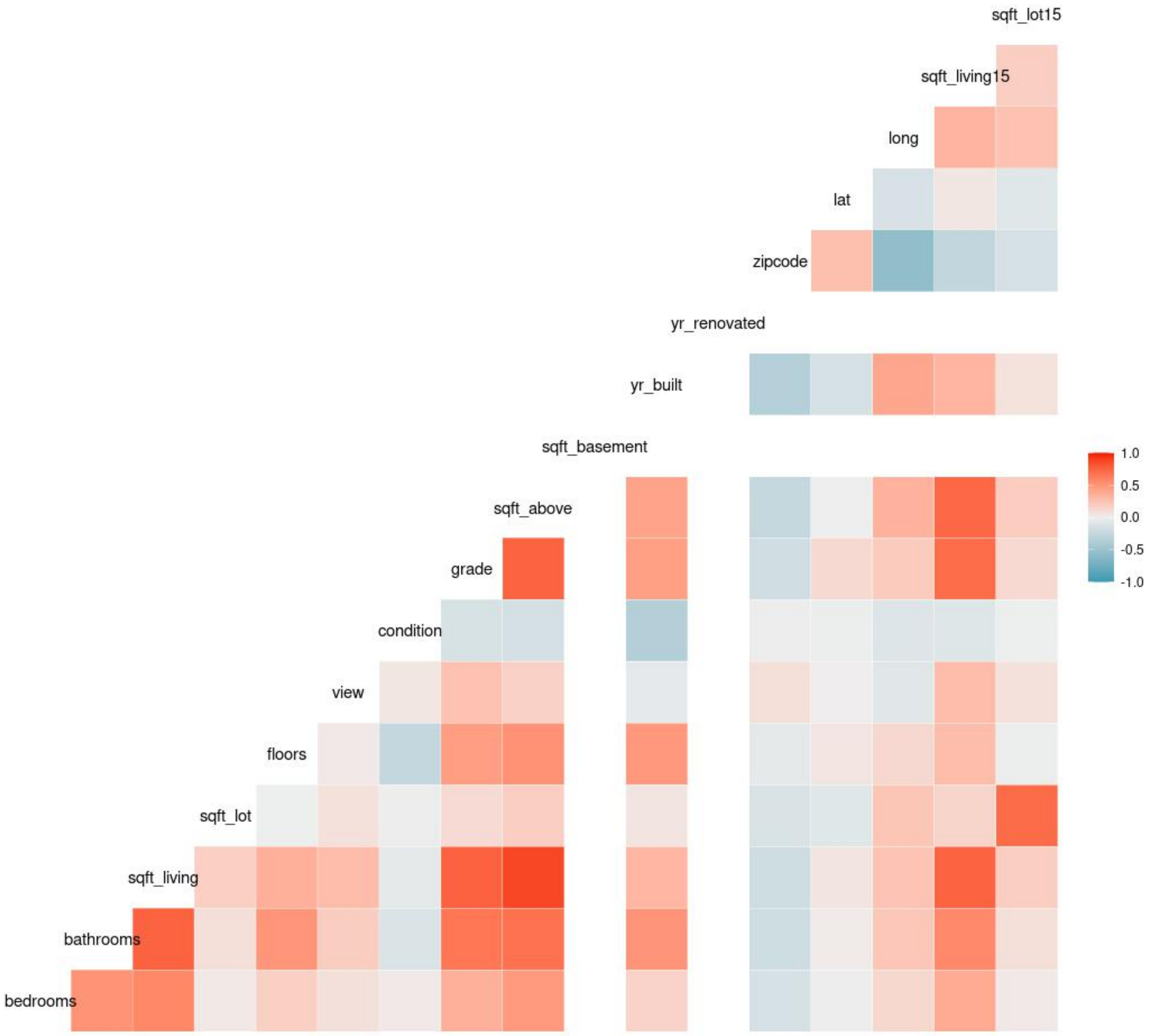


Scatterplot of Price vs. Grade



Scatterplot of Price vs. Waterfront

We also include the correlation matrix that reflects how variables are correlated with each other.



We noticed that sqft_living and sqft_above are highly correlated with a correlation of 0.8765966. This makes a lot of sense because most of living area is usually above the basement. The univariate correlation between sqft_living and price (0.7020351) is higher than that between sqft_above and price (0.6055673). Similarly, sqft_living and sqft_living15 are highly correlated with a correlation of 0.7564203. The univariate correlation between sqft_living and price (0.7020351) is higher than that between sqft_living15 and price (0.5853789).

We consider to remove the variables that are not likely to affect the house price later.

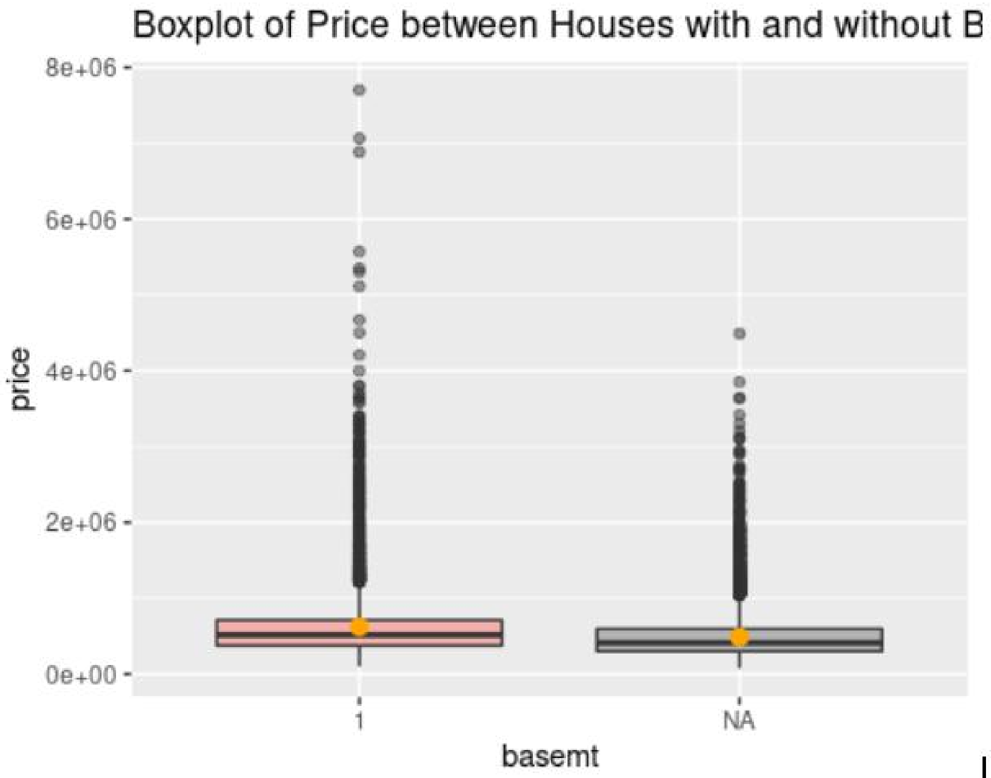
Data Modification

- Based on above findings, we modify our dataset by introducing two new binary features:

  1). renovated: Equal to 1 if the house have been renovated and 0 otherwise

  2). basemt: Equal to 1 if a house has basement and 0 otherwise.

- We also exclude the useless information from the original set and remain the variables that are most relevant to the house price. Moreover, we would drop the variables of sqft_above and sqft_living15 in the further analysis to avoid the issue of multicollinearity. The new modified version of the dataset is named as house, and its summary statistics are printed.



However, as we noticed in the boxplot of basemt, this new variable does not likely to have significant influence on the home prices. So this is also a needless variable to make predictive model later.

Combined with the variable modification investigation with the previous EDA part, we then create a new dataset that contains the necessary predictors only, and summarize the final dataset as below.

```
# Create a new dataset for further analysis
house = subset( kc_housing, select = -c(  date,   sqft_basement,
sqft_living15,   sqft_above,   sqft_lot,   sqft_lot15,   yr_built,
yr_renovated,   zipcode  ))
house$waterfront = as.factor(house$waterfront)
house = na.omit(house)
summary(house)
```

```
##      price            bedrooms        bathrooms       sqft_living
##  Min.   : 186000   Min.   : 1.00   Min.   :0.750   Min.   :  980
##  1st Qu.: 526975   1st Qu.: 3.00   1st Qu.:2.000   1st Qu.: 2065
##  Median : 780000   Median : 4.00   Median :2.500   Median : 2640
##  Mean   : 941443   Mean   : 3.76   Mean   :2.626   Mean   : 2764
##  3rd Qu.:1114000   3rd Qu.: 4.00   3rd Qu.:3.000   3rd Qu.: 3190
##  Max.   :7700000   Max.   :11.00   Max.   :8.000   Max.   :12050
##      floors       waterfront       view           condition          grade
##  Min.   :1.000   FALSE:437   Min.   :0.0000   Min.   :2.000   Min.   : 5.000
##  1st Qu.:1.000   TRUE : 26   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.: 7.000
##  Median :1.500               Median :0.0000   Median :3.000   Median : 8.000
##  Mean   :1.506               Mean   :0.8834   Mean   :3.218   Mean   : 8.058
##  3rd Qu.:2.000               3rd Qu.:2.0000   3rd Qu.:3.000   3rd Qu.: 9.000
##  Max.   :3.000               Max.   :4.0000   Max.   :5.000   Max.   :13.000
##       lat             long         renovated basemt
##  Min.   :47.21   Min.   :-122.5   1:463     1:463
##  1st Qu.:47.55   1st Qu.:-122.4
##  Median :47.62   Median :-122.3
##  Mean   :47.60   Mean   :-122.3
##  3rd Qu.:47.67   3rd Qu.:-122.2
##  Max.   :47.77   Max.   :-121.8
```

Modeling

- Before fitting the models, we partitioned the dataset in which 70% are used as train sets and 30% are used as test sets. And in order to make our results consistent every time we run the data, I set a seed of 913 to remove the sampling randomness.

```
# Test-Train split
set.seed(913)
house_idx = createDataPartition(house$price, p = 0.7, list = FALSE)
house_trn = house[house_idx, ]
house_tst = house[-house_idx, ]
```

- Then we use the training set house_trn to tuning the models and reporting the cross-validated errors measured as RMSEs. Here, we will use a 5-fold cross validation

```
# Create a utility function for calculating RMSE later
rmse = function(actual, predicted) {  sqrt(mean((actual - predicted) ^ 2))}
```

Random Forest Model

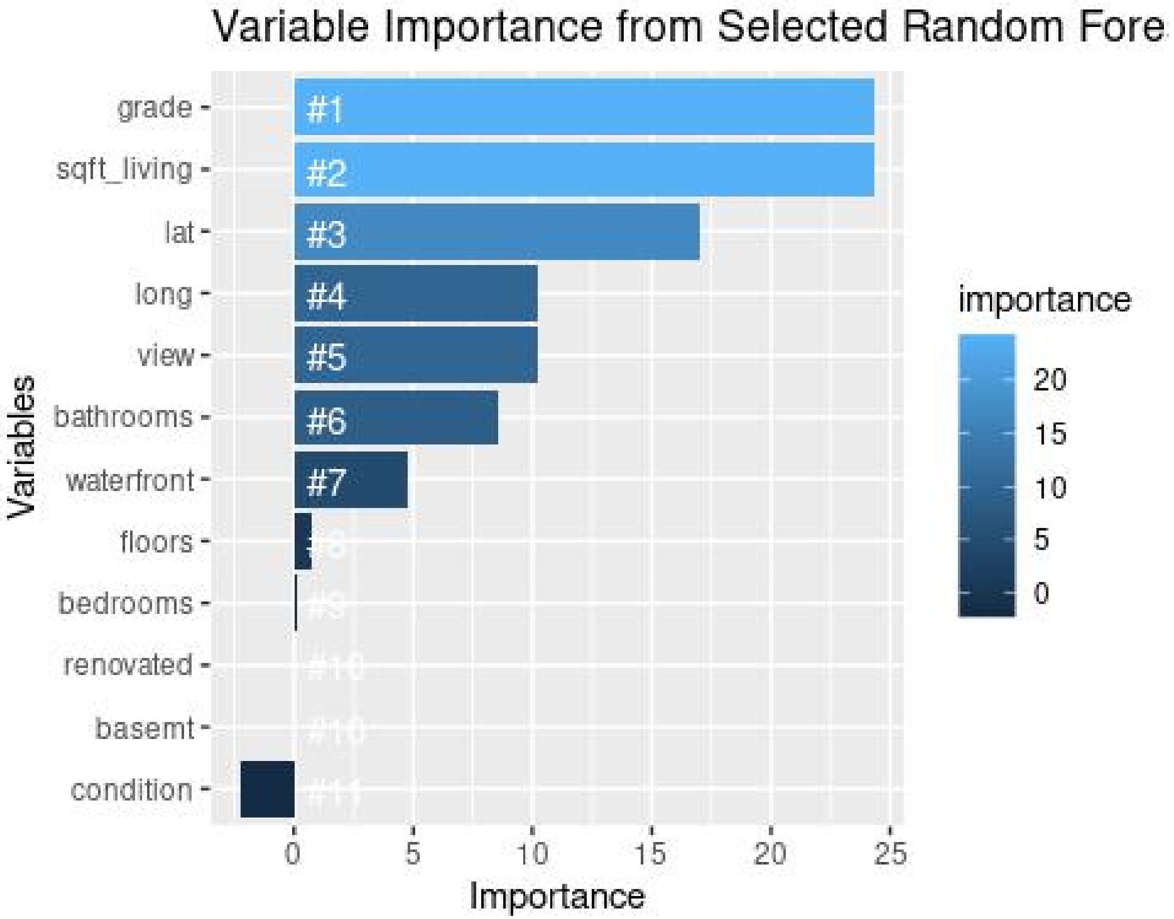
The best random forest model with scaling uses 7 for mtry, meaning that 7 features will be randomly chosen every time a tree is grown. The number of trees is 500 by default. We find that about 88% of the variation in prices can be explained by this model.

It turns out that the grade (ranging from 1 to 13), which represents the construction quality, is the most important variable regarding house sale price.

As for the houses graded over 12, they are thought to have excellent designs and use the best materials while construction. As a result, their prices are expected to be highest too.

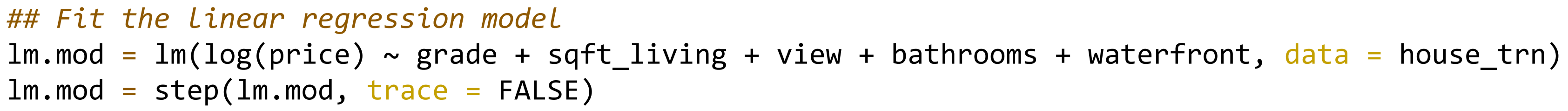
The spacing of the living rooms, turns out to be the next most important feature that is related to home values.

The further next significant factor that is highly correlated with the home prices is the location, consistent with the latitude and longitude ranked as third and fourth important variables.

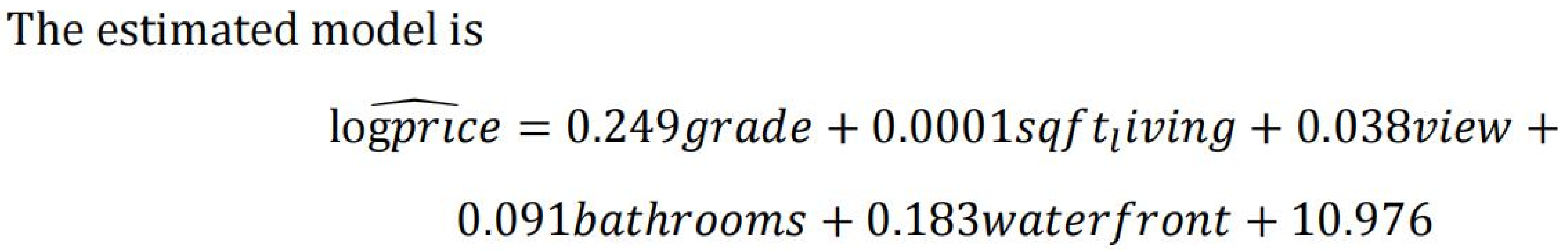


Additive Linear Regression Model

- Although, we learned the effects of location on house price are likely to be significant, we are not going to interpret them in the multiple linear regression model because location is not a factor that could be changed for home owners after the house being purchased. Hence, in this linear model, only the variables of grade, sqft_living, view, waterfront and bathrooms will be included.



```
## Fit the linear regression model
lm.mod = lm(log(price) ~ grade + sqft_living + view + bathrooms + waterfront, data = house_trn)
lm.mod = step(lm.mod, trace = FALSE)
```



The estimated model is

$$\log\widehat{price} = 0.249grade + 0.0001sqf\,t_l iving + 0.038view +$$

$$0.091bathrooms + 0.183waterf\,ront + 10.976$$

Model comparisons

- The reported test rmses for random forest model is 322215.6 and for the linear regression model is 351018.4. That means the best random forest model we found by caret performs better in predicting than the linear model. Therefore, it is a better model that can help us predict the house values.

```
rf.prediction = predict(rf_mod, house_tst)
rf.rmse = rmse(rf.prediction, house_tst$price)
rf.rmse
## [1] 319255.8

lm.prediction = exp(predict(lm.mod, house_tst))
lm.rmse = rmse(lm.prediction, house_tst$price)
lm.rmse
## [1] 351018.4
```

Limitations

- 1. This project comes from the fact that some unchangable external factors such as the location of the house and the real estate market such as unexpected economic crisis could impact the house price unwantedly, even if the priority list is completed by the owner;
- 2. It is quite time consuming to tuning the parameters in the random forest model, which limited my attempts of a wider range of possible values in the parameters should be tuned. That means, the random forest model may still be improved if we can take more time and tuning more models;
- 3. Only two methods are considered in predicting the house price so there may be better models using different methods;
- 4. Ones should notice that the omitted variable biases may still exist, and there are other potential factors that are not included in the data may also affect the home values significantly.

Conclusion

- From the view of statisticians, a priority list for boosting home value should contain the following two things:
- 1).Try to maintain the house in a good condition and add more custom design so that it can be graded higher;

  2).Try to expand the space of living room in the house.
- If we want to use the easily collected house features to predict the house prices, a random forest model with 7 features being selected per tree and considers 500 trees at one time is recommended. Although it may not be precise, it can offer a general idea of how valuable the homes would be for the house owners.

Thank you!