A Statistician's Priority List for Boosting Home Value

Introduction

For many people, a house is not only a residence but also a place where they have been investing throughout their stay. Hence, how to maintain and boost home value during their stay has been a question for many house owners. Generally speaking, there are many commonly known factors that would help increase home values, however, for most people with a limited budget, it is hard to take everything into consideration when they want to boost their home values. Therefore, it is of great significance to learn what should be prioritized during the home improvement with a purpose of value boost. Although home onwers are unable to obtain everything they want with a tight budget, they 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. Statistical models would be constructed to find out the most significant house attributes that are related to home prices.

Purpose

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. Commonly used house attributes would be analyzed and the project would be planned to figure out the most relevant features of a house regarding sale prices. Hopefully, this project could offer some suggestions on house improvement and home value boosting for home investors during their stay.

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.

Map of King County, Washington

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*.

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 housein the original data. The code that help us load the data and print the fisrt few lines of the original data is shown as below

```
## Load the required dataset
library(mlr3data)
data("kc_housing")
head(kc_housing)
```

```
##         date   price bedrooms bathrooms sqft_living sqft_lot floors waterf
ront
## 1 2014-10-13  221900        3      1.00        1180     5650      1       F
ALSE
## 2 2014-12-09  538000        3      2.25        2570     7242      2       F
ALSE
## 3 2015-02-25  180000        2      1.00         770    10000      1       F
ALSE
## 4 2014-12-09  604000        4      3.00        1960     5000      1       F
```

```
ALSE
## 5 2015-02-18  510000        3      2.00        1680     8080      1       F
ALSE
## 6 2014-05-12 1225000        4      4.50        5420   101930      1       F
ALSE
##   view condition grade sqft_above sqft_basement yr_built yr_renovated zipc
ode
## 1    0         3     7       1180            NA     1955           NA   98
178
## 2    0         3     7       2170           400     1951         1991   98
125
## 3    0         3     6        770            NA     1933           NA   98
028
## 4    0         5     7       1050           910     1965           NA   98
136
## 5    0         3     8       1680            NA     1987           NA   98
074
## 6    0         3    11       3890          1530     2001           NA   98
053
##       lat     long sqft_living15 sqft_lot15
## 1 47.5112 -122.257          1340       5650
## 2 47.7210 -122.319          1690       7639
## 3 47.7379 -122.233          2720       8062
## 4 47.5208 -122.393          1360       5000
## 5 47.6168 -122.045          1800       7503
## 6 47.6561 -122.005          4760     101930
```

```r
## Print out the summary statistics
summary(kc_housing)
```

```
##       date                          price            bedrooms
##  Min.   :2014-05-02 00:00:00   Min.   :  75000   Min.   : 0.000
##  1st Qu.:2014-07-22 00:00:00   1st Qu.: 321950   1st Qu.: 3.000
##  Median :2014-10-16 00:00:00   Median : 450000   Median : 3.000
##  Mean   :2014-10-29 03:58:09   Mean   : 540088   Mean   : 3.371
##  3rd Qu.:2015-02-17 00:00:00   3rd Qu.: 645000   3rd Qu.: 4.000
##  Max.   :2015-05-27 00:00:00   Max.   :7700000   Max.   :33.000
##
##    bathrooms      sqft_living       sqft_lot           floors
##  Min.   :0.000   Min.   :  290   Min.   :    520   Min.   :1.000
##  1st Qu.:1.750   1st Qu.: 1427   1st Qu.:   5040   1st Qu.:1.000
##  Median :2.250   Median : 1910   Median :   7618   Median :1.500
##  Mean   :2.115   Mean   : 2080   Mean   :  15107   Mean   :1.494
##  3rd Qu.:2.500   3rd Qu.: 2550   3rd Qu.:  10688   3rd Qu.:2.000
##  Max.   :8.000   Max.   :13540   Max.   :1651359   Max.   :3.500
##
##  waterfront           view          condition         grade
##  Mode :logical   Min.   :0.0000   Min.   :1.000   Min.   : 1.000
##  FALSE:21450     1st Qu.:0.0000   1st Qu.:3.000   1st Qu.: 7.000
##  TRUE :163       Median :0.0000   Median :3.000   Median : 7.000
```

```
##                   Mean   :0.2343   Mean   :3.409   Mean   : 7.657
##                   3rd Qu.:0.0000   3rd Qu.:4.000   3rd Qu.: 8.000
##                   Max.   :4.0000   Max.   :5.000   Max.   :13.000
##
##    sqft_above    sqft_basement      yr_built     yr_renovated      zipcode
##  Min.   : 290   Min.   :  10.0   Min.   :1900   Min.   :1934    Min.   :98
001
##  1st Qu.:1190   1st Qu.: 450.0   1st Qu.:1951   1st Qu.:1987    1st Qu.:98
033
##  Median :1560   Median : 700.0   Median :1975   Median :2000    Median :98
065
##  Mean   :1788   Mean   : 742.4   Mean   :1971   Mean   :1996    Mean   :98
078
##  3rd Qu.:2210   3rd Qu.: 980.0   3rd Qu.:1997   3rd Qu.:2007    3rd Qu.:98
118
##  Max.   :9410   Max.   :4820.0   Max.   :2015   Max.   :2015    Max.   :98
199
##                 NA's   :13126                   NA's   :20699
##       lat             long         sqft_living15    sqft_lot15
##  Min.   :47.16   Min.   :-122.5   Min.   : 399   Min.   :   651
##  1st Qu.:47.47   1st Qu.:-122.3   1st Qu.:1490   1st Qu.:  5100
##  Median :47.57   Median :-122.2   Median :1840   Median :  7620
##  Mean   :47.56   Mean   :-122.2   Mean   :1987   Mean   : 12768
##  3rd Qu.:47.68   3rd Qu.:-122.1   3rd Qu.:2360   3rd Qu.: 10083
##  Max.   :47.78   Max.   :-121.3   Max.   :6210   Max.   :871200
##
```

```
## Print out the data dimensions
dim(kc_housing)
```

```
## [1] 21613    20
```

Variables

The full variable dictionary is summarized as below:

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

The first package that will be used in this project is `mlr3data`, which offers the dataset that we are going to analyze. Besides, we will use the `ggplot2` and `lattice` packages for the purpose of data visualization and `randomForest` for random forest models that can be helpful to analyze the effects of the house factors on the house price. Also, we will utilize the `stargazer` package to offer neat and more readable model results of linear regressions. Anotehr important package that can be useful in this project is `GGally`, in which the `ggcorr` function can help us obtain the correlation matrix. We also use the `dplyr` package to manipulate and modify data frames.

And after receiving the feedback on the first draft of the project, I decide to add new packages of `caret`, which is used to tuning random forest models so that the model performs best regarding cross-validation root mean square errors (CV RMSEs) can be found.

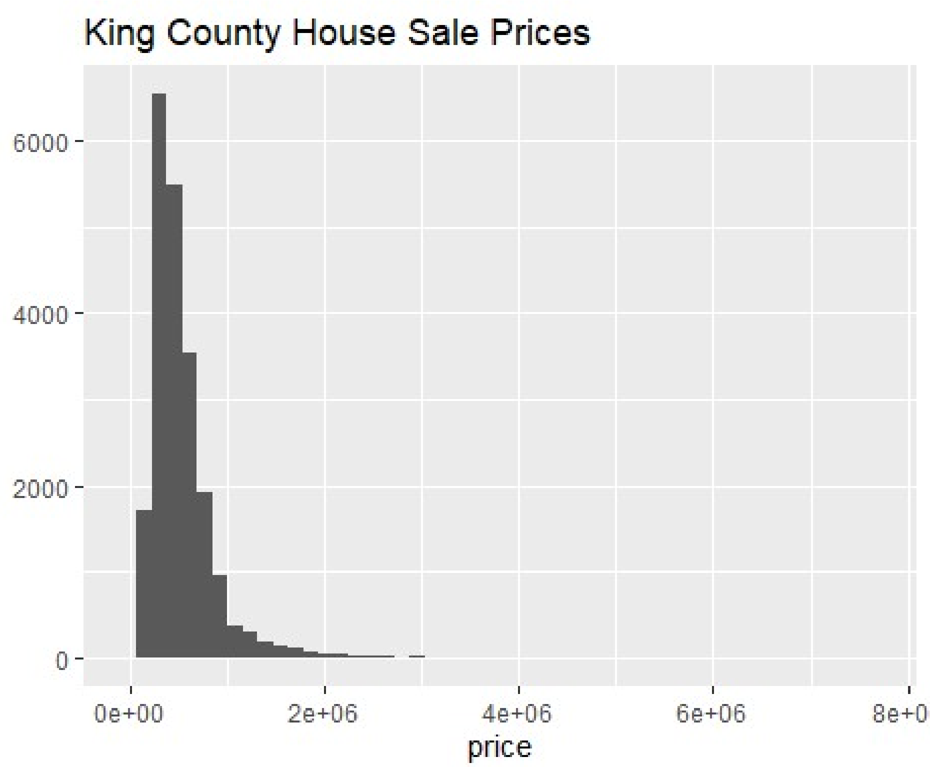
Also, the packages of `knitr` and `kableExtra` are also considered to help us generate a neat and well-formatted final document via R markdown.

```r
## 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

We start our data exploration with the variable of interest `price`.

```r
# Check the distribution of house sale price
qplot(x = price, data = kc_housing, bins = 50,
      main = "King County House Sale Prices")
```



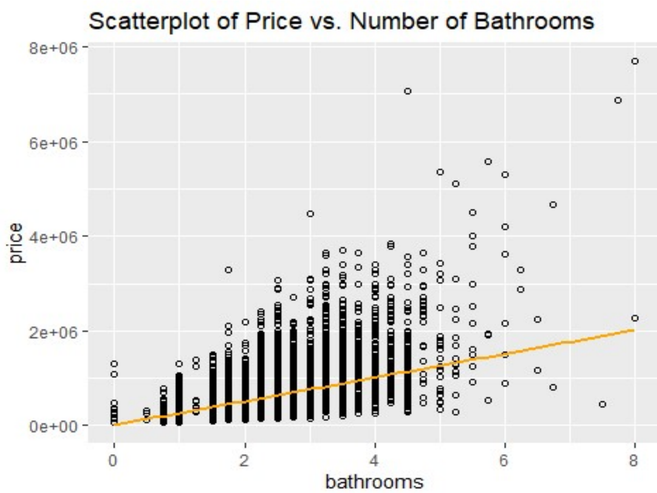

```r
# 5-point summary of price
summary(kc_housing$price)

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

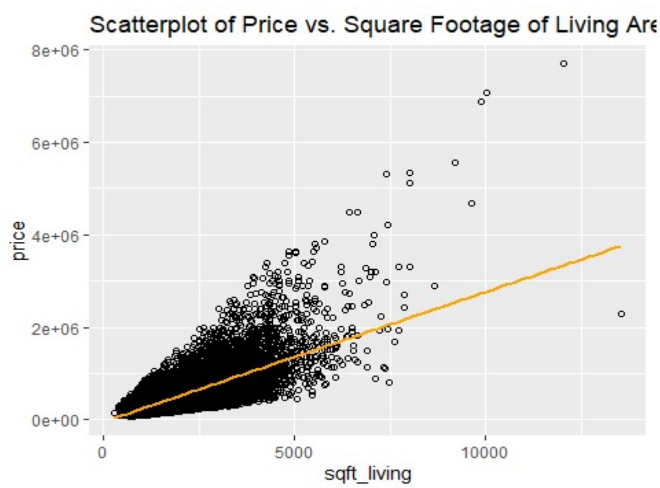
It is clear that the distribution of the home price is positively skewed with a long right tail, which implies that some houses are expected to have higehr values than others.

Next, we explore the relationship between the features of the house and the home prices by plotting `price` with each feature. We find that some features like `bathrooms`, `sqft_living`, `grade` and `waterfront` have relatively stronger relationships with `price` than others.

```
# Create scatterplot for price and bathrooms
ggplot(kc_housing, aes(x = bathrooms, y = price)) +
  geom_point(shape = 1) +
  geom_smooth(method = lm, color = "orange", se = FALSE) +
  ggtitle("Scatterplot of Price vs. Number of Bathrooms")
```

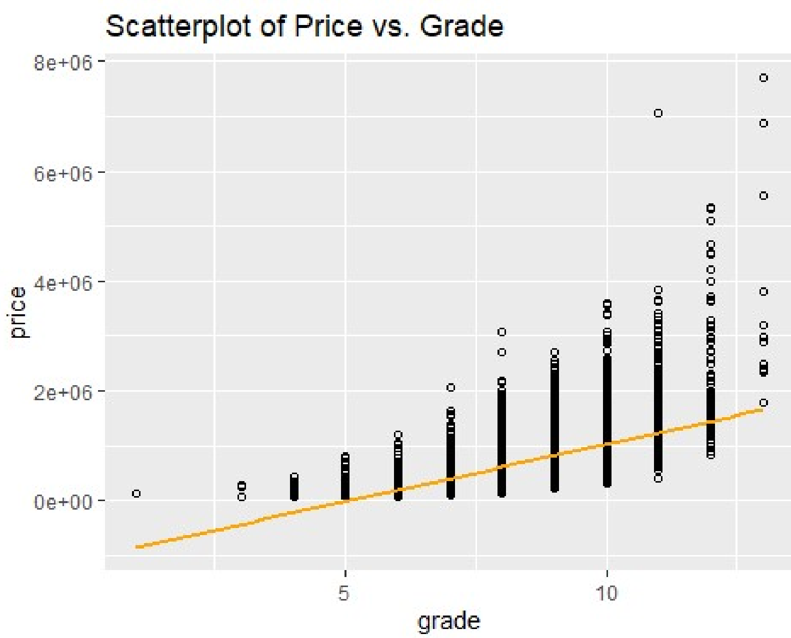



```
# Create scatterplot for price and sqft_living
ggplot(kc_housing, aes(x = sqft_living, y = price)) +
  geom_point(shape = 1) +
  geom_smooth(method = lm, color = "orange", se = FALSE) +
  ggtitle("Scatterplot of Price vs. Square Footage of Living Area")
```

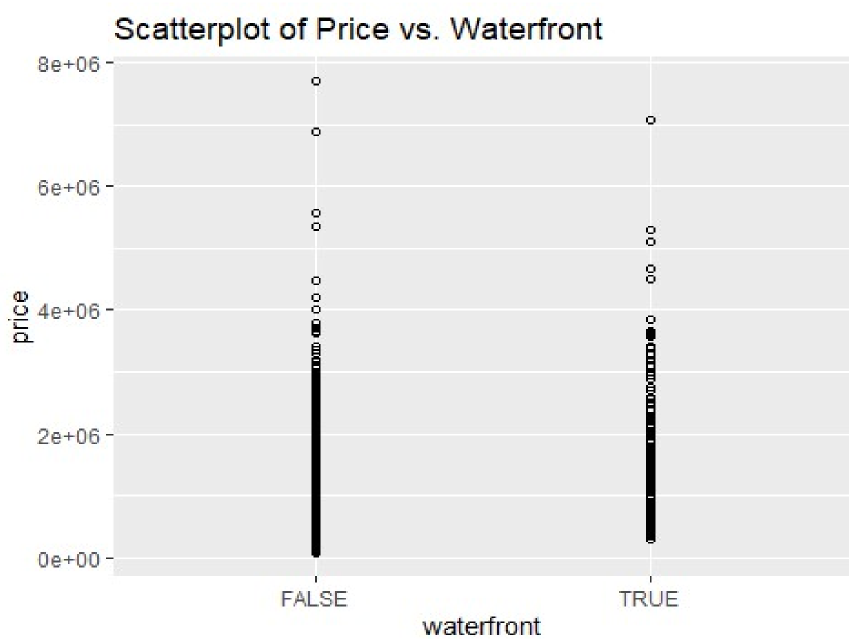



```
# Create scatterplot for price and grade
ggplot(kc_housing, aes(x = grade, y = price)) +
  geom_point(shape = 1) +
```

```
  geom_smooth(method = lm, color = "orange", se = FALSE) +
  ggtitle("Scatterplot of Price vs. Grade")
```

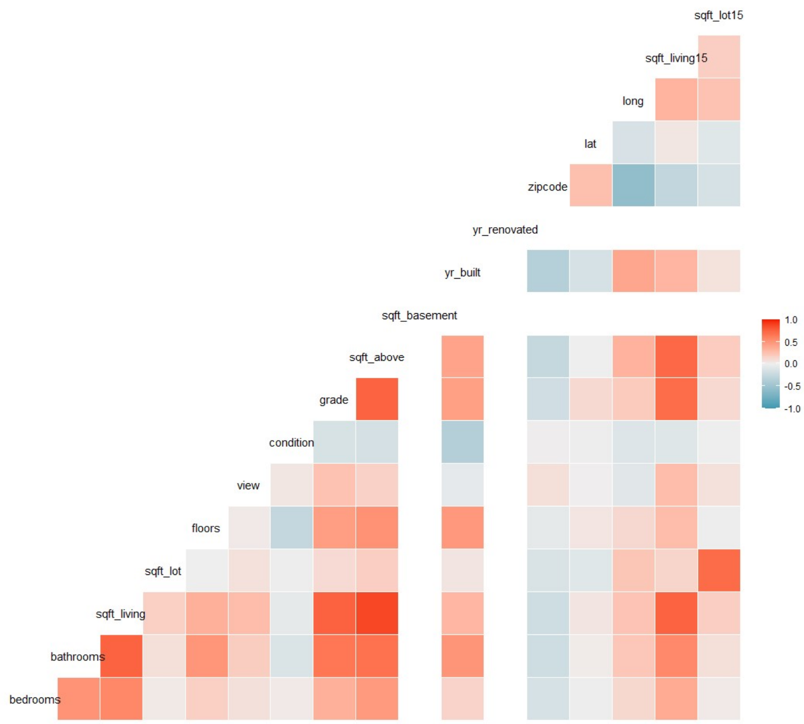



```
# Create scatterplot for price and waterfront
ggplot(kc_housing, aes(x = waterfront, y = price)) +
  geom_point(shape = 1) +
  geom_smooth(method = lm, color = "orange", se = FALSE) +
  ggtitle("Scatterplot of Price vs. Waterfront")
```



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

```r
# Correlation matrix for numeric variables
ggcorr(kc_housing[, -c(1:2)], method = c("everything","pearson"))
```



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

Data Modification

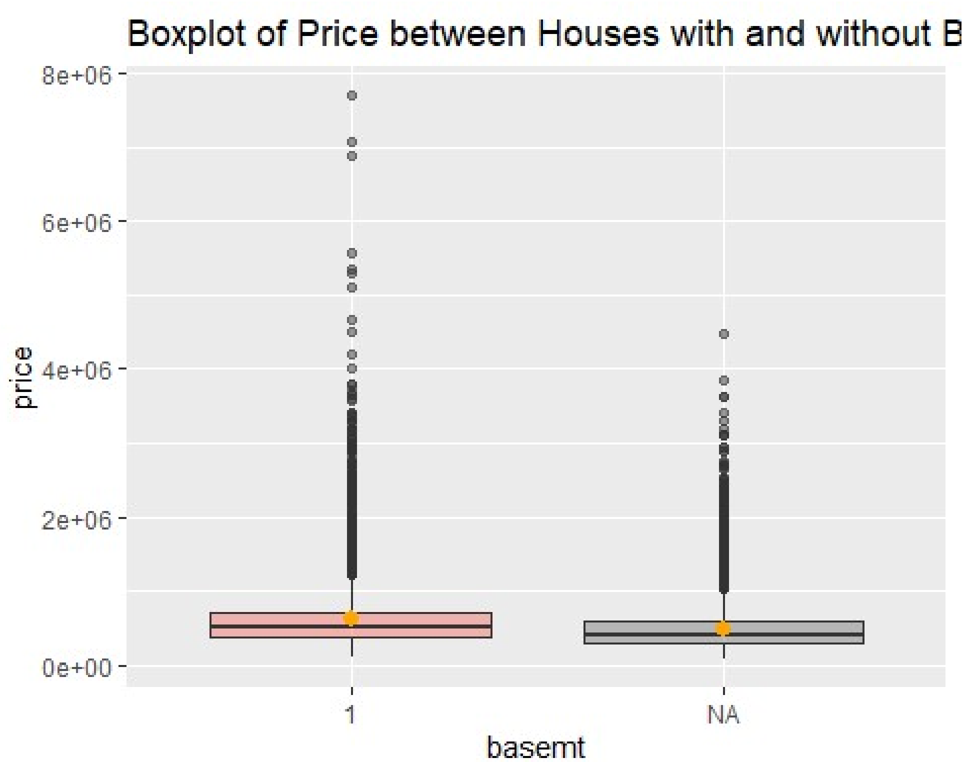
Based on above findings, we will modify our dataset by introducing two new binary features:

- renovated: Equal to 1 if the house have been renovated and 0 otherwise
- 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 it's summary statistics are printed.

```r
# Create new variable "renovated" based on the existing variable "yr_renovate
d"
```

```
kc_housing$renovated = as.factor(ifelse(kc_housing$yr_renovated > 0, "1", "0"
))
# Create new variable "basemt" based on the existing variable "sqft_basement"
kc_housing$basemt = as.factor(ifelse(kc_housing$sqft_basement > 0, "1", "0"))
# Create boxplot for price and basemt
ggplot(kc_housing,
       aes(x = basemt,y = price,fill = basemt)) +
geom_boxplot(alpha = 0.5) +
stat_summary(fun.y = mean,
             geom = "point",
             shape = 20,
             size = 4,
             color= "orange",
             fill= "orange") +
theme(legend.position = "none") +
ggtitle("Boxplot of Price between Houses with and without Basements")
```



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.   : 5.0
00
##  1st Qu.:1.000   TRUE : 26   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.: 7.0
00
##  Median :1.500               Median :0.0000   Median :3.000   Median : 8.0
00
##  Mean   :1.506               Mean   :0.8834   Mean   :3.218   Mean   : 8.0
58
##  3rd Qu.:2.000               3rd Qu.:2.0000   3rd Qu.:3.000   3rd Qu.: 9.0
00
##  Max.   :3.000               Max.   :4.0000   Max.   :5.000   Max.   :13.0
00
##       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.

```r
# 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.

Next we construct the functions that we may need during the modeling procedure.

- Create a utility function for calculating RMSEs

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

Random Forest Model

First, we trained random forest models using all the predictors in the `house` dataset with `price` as response variable. The default tuning parameters chosen by the `caret` package would be used.

```r
# Tune a random forest model without normalizing predictors
set.seed(913)
rf_unscale_mod = train(
  price ~ ., data = house_trn,
  trControl = trainControl(method = "cv", number = 5),
  method = "rf"
)
```

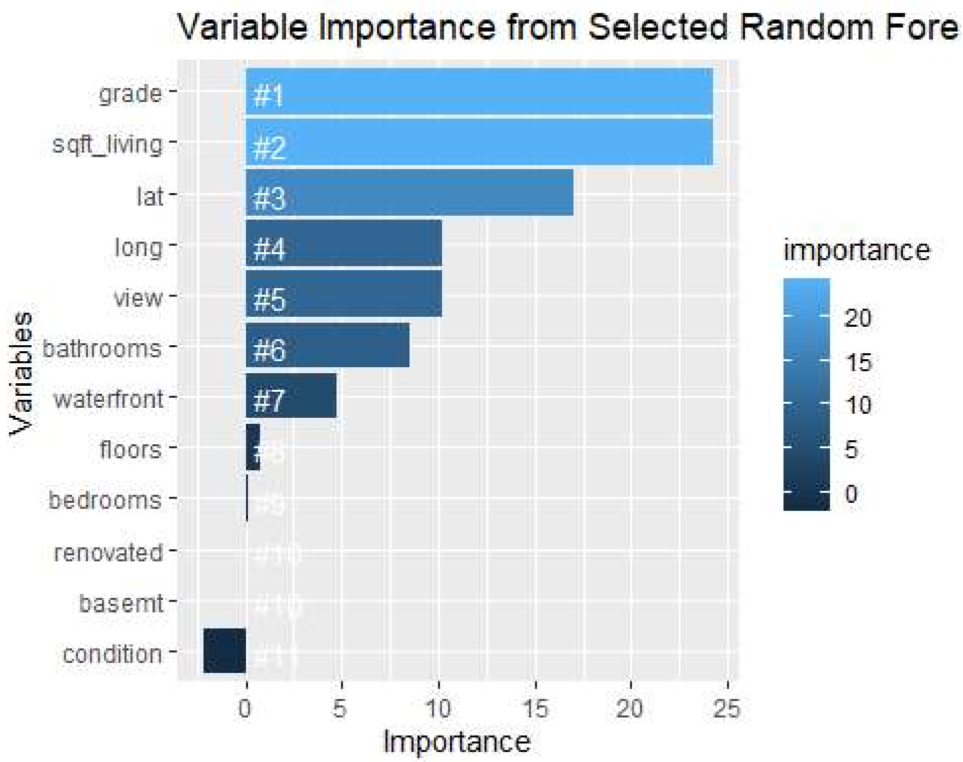
The best random forest model with scaling uses 7 for `mtry`, meaning that 7 features will be randomly chosen every time a tree is grown. The number of trees is 500 by default. We find that about 88% of the variation in prices can be explained by this model.

```r
## Fit the best random forest model
rf_mod = randomForest(price ~ .,
                      data = house_trn,
                      mtry = 7,
                      importance = T)
importance = importance(rf_mod)
VarImportance = data.frame(variables = row.names(importance),
                           importance = round(importance[, '%IncMSE'], 2))

## Rank variables by importance
rank = VarImportance %>% mutate(rank = paste0('#', dense_rank(desc(importance
))))

ggplot(rank, aes(
  x = reorder(variables, importance),
```

```
  y = importance,
  fill = importance
)) +
  geom_bar(stat = 'identity') +
  geom_text(
    aes(x = variables, y = 0.5, label = rank),
    hjust = 0,
    vjust = 0.6,
    size = 4,
    color = 'white'
  ) +
  labs(x = 'Variables', y = 'Importance') +
  ggtitle("Variable Importance from Selected Random Forest Model") +
  coord_flip()
```



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. For houses only meet the minimum building standards, their grades are low and ranged from 1 to 3, and their prices are expected to be the lowest on average. For houses that have achieved average performance in terms of construction and design, they are graded as 7. And their averaged prices are expected to be moderate among all houses. 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.

`sqft_living`, the spacing of the living rooms, turns out to be the next most important feature that is related to home values. And 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)
## Print the model results
stargazer(lm.mod, type = "text")
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                             log(price)
## -----------------------------------------------
## grade                        0.249***
##                               (0.022)
##
## sqft_living                  0.0001***
##                              (0.00002)
##
## view                          0.038**
##                               (0.016)
##
## bathrooms                    0.091***
##                               (0.030)
##
## waterfront                    0.183**
##                               (0.093)
##
## Constant                     10.976***
##                               (0.131)
##
## -----------------------------------------------
## Observations                    326
## R2                             0.728
## Adjusted R2                    0.724
## Residual Std. Error      0.314 (df = 320)
## F Statistic          171.686*** (df = 5; 320)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

The estimated model is

$$\log\widehat{price} = 0.249 grade + 0.0001 sqft_l iving + 0.038 view +$$

$$0.091 bathrooms + 0.183 waterfront + 10.976$$

The R-squared of this model is 0.728, which means this linear model containing the most relevant predictors could explain about 72.8% of the total variation in the house prices. And F statistic of this model is 0.314 (p = 0.000), which means the overall model is statistically significant. Also, at 5% level of significance, we notice all predictors are statistically significant individually.

As we notice, here we used the log-transformation on the dependent variable because the variable is right skewed based on previous analysis. Hence, the model results imply that one additional point in the house grade is expected to increase the house price by 24.9% on average when other factors are assumed to be the same. And the house with a waterfront is expected to be 18.3% higher in price than the house without a waterfront when other conditions are the same. And an additional bathroom in a house is expected to bring up the home value by 9.1%, and with a one-sqft increase in the living room is expected to enhanced the house value by 0.01% on average conditional to all other factors respectively. As for the time of the house being viewed, it is shown that each time the house is viewed, the house value is expected to be boost by 3.8% on average.

Model comparisons

Finally we compare the predictive performance of the two models.

```
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
```

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. So we decide this as a better model that can help us predict the house values.

Conclusion

In conclusion, from a view of statistician, 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

And 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.

Limitations

The first limitation of 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.

Besides, 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.

Thirdly, we only considered two methods in predicting the house price so there may be better models using different methods.

Lastly, we 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.