House Price Prediction

Harry Shao

11/22/2020

Abstract

We use conditions from training dataset to investigate the house price from test dataset. I create some new variables which I think may help us to predict the price and then create 7 different models to predict price and log(price). By using RMPSE, I find that the multiple regression model which contains sqft_living, bedrooms, floors, sqft_basement, has_view(variable I create), and waterfront perform best for both price and log(price). Therefore, I use this model for the new dataset to predict the house price.

Introduction

In this project, I am going to create some new variables from the training dataset. I will use original variables and the variables I create to compare some different models. I use a correlation plot to choose my explanatory variables. The explanatory variables should be correlated with price but not highly correlated with other explanatory variables. I am then using RMPSE to get the best model to predict price and the best model to predict log(price), choosing the better model between the best model for price and the best model for log(price) as our predicting model. Finally, using the predicting model to predict the house price for the test dataset.

Exploratory Analysis

Create graphs and tables and add discussion paragraphs.

#The summary table doesn't show any NA's, so there is no missing values in our train dataset.

```
##        id                date               price            bedromms
##  Min.   :1.120e+07   Length:1000        Min.   : 100000   Min.   :0.000
##  1st Qu.:2.122e+09   Class :character   1st Qu.: 320000   1st Qu.:3.000
##  Median :3.838e+09   Mode  :character   Median : 452000   Median :3.000
##  Mean   :4.531e+09                      Mean   : 553334   Mean   :3.377
##  3rd Qu.:7.209e+09                      3rd Qu.: 650862   3rd Qu.:4.000
##  Max.   :9.839e+09                      Max.   :4489000   Max.   :8.000
##    bathrooms      sqft_living      sqft_lot          floors
##  Min.   :0.500   Min.   : 570   Min.   :   572   Min.   :1.000
##  1st Qu.:1.750   1st Qu.:1430   1st Qu.:  5000   1st Qu.:1.000
##  Median :2.250   Median :1930   Median :  7492   Median :1.500
##  Mean   :2.138   Mean   :2099   Mean   : 15356   Mean   :1.501
##  3rd Qu.:2.500   3rd Qu.:2532   3rd Qu.: 10804   3rd Qu.:2.000
##  Max.   :4.750   Max.   :6510   Max.   :920423   Max.   :3.000
##  waterfront           view         condition         grade         sqft_above
##  Mode :logical   Min.   :0.000   Min.   :1.000   Min.   : 4.00   Min.   : 570
##  FALSE:993       1st Qu.:0.000   1st Qu.:3.000   1st Qu.: 7.00   1st Qu.:1180
##  TRUE :7         Median :0.000   Median :3.000   Median : 7.00   Median :1560
##                  Mean   :0.237   Mean   :3.379   Mean   : 7.65   Mean   :1794
##                  3rd Qu.:0.000   3rd Qu.:4.000   3rd Qu.: 8.00   3rd Qu.:2230
##                  Max.   :4.000   Max.   :5.000   Max.   :12.00   Max.   :6430
##  sqft_basement       yr_built     yr_renovated        zipcode
##  Min.   :   0.0   Min.   :1900   Min.   :   0.00   Min.   :98001
##  1st Qu.:   0.0   1st Qu.:1953   1st Qu.:   0.00   1st Qu.:98033
##  Median :   0.0   Median :1976   Median :   0.00   Median :98059
##  Mean   : 304.9   Mean   :1972   Mean   :  79.78   Mean   :98076
##  3rd Qu.: 600.0   3rd Qu.:1999   3rd Qu.:   0.00   3rd Qu.:98116
##  Max.   :3260.0   Max.   :2015   Max.   :2014.00   Max.   :98199
##       lat             long        sqft_living15    sqft_lot15
##  Min.   :47.18   Min.   :-122.5   Min.   : 840   Min.   :   817
##  1st Qu.:47.47   1st Qu.:-122.3   1st Qu.:1520   1st Qu.:  5000
##  Median :47.57   Median :-122.2   Median :1830   Median :  7422
##  Mean   :47.56   Mean   :-122.2   Mean   :2004   Mean   : 13452
##  3rd Qu.:47.68   3rd Qu.:-122.1   3rd Qu.:2380   3rd Qu.:  9942
##  Max.   :47.78   Max.   :-121.7   Max.   :5080   Max.   :411962
```

```
##       id                date              price           bedroom
##  Min.   :7.600e+06   Length:1000        Mode:logical   Min.   :0.000
##  1st Qu.:2.062e+09   Class :character   NA's:1000      1st Qu.:3.000
##  Median :4.026e+09   Mode  :character                  Median :3.000
##  Mean   :4.659e+09                                     Mean   :3.374
##  3rd Qu.:7.300e+09                                     3rd Qu.:4.000
##  Max.   :9.834e+09                                     Max.   :8.000
##    bathrooms      sqft_living      sqft_lot          floors
##  Min.   :0.000   Min.   : 560   Min.   :   520   Min.   :1.000
##  1st Qu.:1.500   1st Qu.:1398   1st Qu.:  5098   1st Qu.:1.000
##  Median :2.250   Median :1933   Median :  7687   Median :1.500
##  Mean   :2.135   Mean   :2079   Mean   : 13956   Mean   :1.511
##  3rd Qu.:2.500   3rd Qu.:2562   3rd Qu.: 10361   3rd Qu.:2.000
##  Max.   :6.250   Max.   :8020   Max.   :426450   Max.   :3.000
##  waterfront           view          condition         grade
##  Mode :logical   Min.   :0.000   Min.   :1.000   Min.   : 4.000
##  FALSE:990       1st Qu.:0.000   1st Qu.:3.000   1st Qu.: 7.000
##  TRUE :10        Median :0.000   Median :3.000   Median : 7.000
##                  Mean   :0.261   Mean   :3.378   Mean   : 7.685
##                  3rd Qu.:0.000   3rd Qu.:4.000   3rd Qu.: 8.000
##                  Max.   :4.000   Max.   :5.000   Max.   :12.000
##    sqft_above   sqft_basement       yr_built     yr_renovated
##  Min.   : 560   Min.   :   0.0   Min.   :1900   Min.   :   0.00
##  1st Qu.:1210   1st Qu.:   0.0   1st Qu.:1953   1st Qu.:   0.00
##  Median :1567   Median :   0.0   Median :1975   Median :   0.00
##  Mean   :1801   Mean   : 278.4   Mean   :1972   Mean   :  75.92
##  3rd Qu.:2250   3rd Qu.: 520.0   3rd Qu.:1997   3rd Qu.:   0.00
##  Max.   :8020   Max.   :2250.0   Max.   :2015   Max.   :2015.00
##     zipcode           lat              long         sqft_living15
##  Min.   :98001   Min.   :47.18   Min.   :-122.5   Min.   : 700
##  1st Qu.:98033   1st Qu.:47.46   1st Qu.:-122.3   1st Qu.:1470
##  Median :98074   Median :47.58   Median :-122.3   Median :1820
##  Mean   :98080   Mean   :47.56   Mean   :-122.2   Mean   :1985
##  3rd Qu.:98118   3rd Qu.:47.68   3rd Qu.:-122.1   3rd Qu.:2370
##  Max.   :98199   Max.   :47.78   Max.   :-121.4   Max.   :5790
##    sqft_lot15
##  Min.   :   794
##  1st Qu.:  5175
##  Median :  7700
##  Mean   : 11841
##  3rd Qu.: 10042
##  Max.   :253519
```

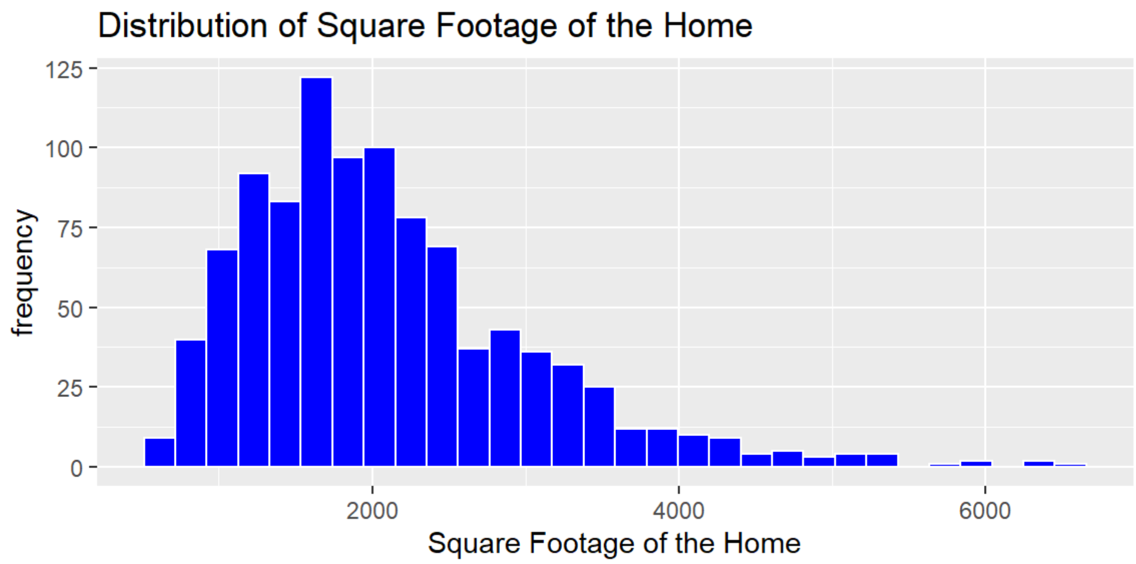
#The only categorical variable is 'waterfront', and the rest variables are quantitive variables. There are no missing values in both test dataset and train dataset.

```
##                  id price bedrooms bathrooms sqft_living sqft_lot floors  view
## id             1.00  0.01     0.03      0.01        0.00    -0.12   0.00  0.01
## price          0.01  1.00     0.32      0.52        0.72     0.14   0.27  0.34
## bedrooms       0.03  0.32     1.00      0.50        0.59    -0.05   0.13  0.00
## bathrooms      0.01  0.52     0.50      1.00        0.73     0.04   0.49  0.15
## sqft_living    0.00  0.72     0.59      0.73        1.00     0.16   0.33  0.26
## sqft_lot      -0.12  0.14    -0.05      0.04        0.16     1.00   0.02  0.06
## floors         0.00  0.27     0.13      0.49        0.33     0.02   1.00  0.07
## view           0.01  0.34     0.00      0.15        0.26     0.06   0.07  1.00
## condition     -0.06 -0.01     0.08     -0.09       -0.05    -0.03  -0.27  0.00
## grade         -0.01  0.70     0.35      0.67        0.76     0.14   0.47  0.27
## sqft_above    -0.02  0.64     0.46      0.65        0.87     0.19   0.50  0.19
## sqft_basement  0.04  0.29     0.36      0.31        0.44    -0.01  -0.26  0.19
## yr_built       0.01  0.07     0.13      0.51        0.30     0.08   0.48 -0.05
## yr_renovated   0.05  0.11     0.02      0.05        0.06    -0.02   0.02  0.10
## zipcode       -0.01 -0.11    -0.13     -0.21       -0.22    -0.13  -0.02  0.01
## lat            0.02  0.29     0.01      0.03        0.05    -0.08   0.07  0.01
## long           0.01  0.05     0.12      0.22        0.27     0.25   0.11 -0.05
## sqft_living15 -0.03  0.60     0.39      0.55        0.76     0.13   0.26  0.31
## sqft_lot15    -0.13  0.10    -0.05      0.01        0.15     0.82  -0.01  0.06
##               condition grade sqft_above sqft_basement yr_built yr_renovated
## id                -0.06 -0.01      -0.02          0.04     0.01         0.05
## price             -0.01  0.70       0.64          0.29     0.07         0.11
## bedrooms           0.08  0.35       0.46          0.36     0.13         0.02
## bathrooms         -0.09  0.67       0.65          0.31     0.51         0.05
## sqft_living       -0.05  0.76       0.87          0.44     0.30         0.06
## sqft_lot          -0.03  0.14       0.19         -0.01     0.08        -0.02
## floors            -0.27  0.47       0.50         -0.26     0.48         0.02
## view               0.00  0.27       0.19          0.19    -0.05         0.10
## condition          1.00 -0.14      -0.15          0.18    -0.30        -0.03
## grade             -0.14  1.00       0.77          0.14     0.46         0.00
## sqft_above        -0.15  0.77       1.00         -0.05     0.41         0.02
## sqft_basement      0.18  0.14      -0.05          1.00    -0.14         0.09
## yr_built          -0.30  0.46       0.41         -0.14     1.00        -0.24
## yr_renovated      -0.03  0.00       0.02          0.09    -0.24         1.00
## zipcode           -0.04 -0.21      -0.28          0.05    -0.35         0.08
## lat               -0.05  0.10       0.00          0.11    -0.15         0.05
## long              -0.10  0.22       0.37         -0.13     0.43        -0.10
## sqft_living15     -0.09  0.71       0.74          0.19     0.30         0.01
## sqft_lot15         0.00  0.11       0.17          0.00     0.06        -0.02
##               zipcode   lat  long sqft_living15 sqft_lot15
## id              -0.01  0.02  0.01         -0.03      -0.13
## price           -0.11  0.29  0.05          0.60       0.10
## bedrooms        -0.13  0.01  0.12          0.39      -0.05
## bathrooms       -0.21  0.03  0.22          0.55       0.01
## sqft_living     -0.22  0.05  0.27          0.76       0.15
## sqft_lot        -0.13 -0.08  0.25          0.13       0.82
## floors          -0.02  0.07  0.11          0.26      -0.01
## view             0.01  0.01 -0.05          0.31       0.06
## condition       -0.04 -0.05 -0.10         -0.09       0.00
## grade           -0.21  0.10  0.22          0.71       0.11
## sqft_above      -0.28  0.00  0.37          0.74       0.17
## sqft_basement    0.05  0.11 -0.13          0.19       0.00
## yr_built        -0.35 -0.15  0.43          0.30       0.06
## yr_renovated     0.08  0.05 -0.10          0.01      -0.02
## zipcode          1.00  0.23 -0.58         -0.33      -0.14
## lat              0.23  1.00 -0.15          0.04      -0.10
```

```
## long           -0.58 -0.15  1.00           0.35        0.18
## sqft_living15  -0.33  0.04  0.35           1.00        0.14
## sqft_lot15     -0.14 -0.10  0.18           0.14        1.00
```

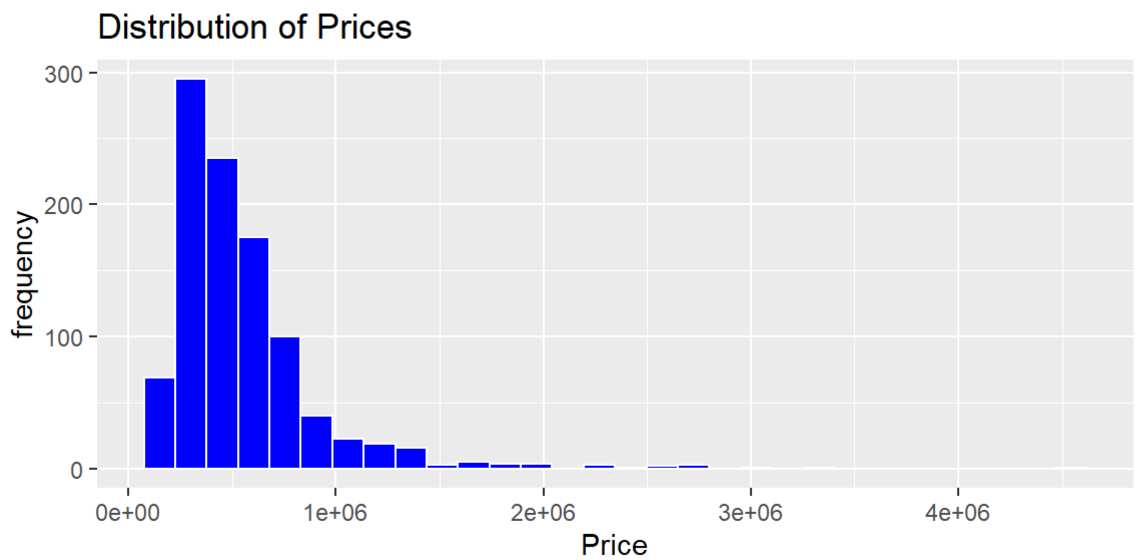
#By observing the correaltion plot, I find square footage of the home and grade are highly correlated with our response varibale price.Number of Bathrooms is highly correlated with square footage of the home; Square footage of the home is highly correlated with grade, square footage of house apart from basement and living room area in 2015; square footage of the lot is highly correlated with lotSize area in 2015; grade is highly correlated with square footage of house apart from basement and living room area in 2015; square footage of house apart from basement is highly correlated with living room area in 2015.

#Variable Choosing: First I ignore variables that has small correlation with our response variable(Correlation below 0.2). Hence I remove id(a notation for a house), sqft_lot(square footage of the lot),condition(How good the condition is), yr_built(Built Year), yr_renovated(Year when house was renovated),zipcode, long(Longitude coordinate), and sqft_lot15(lotSize area in 2015). Then choosing explanatory variables from the rest, the explanatory variable can't have high correlation with the other explanatory.



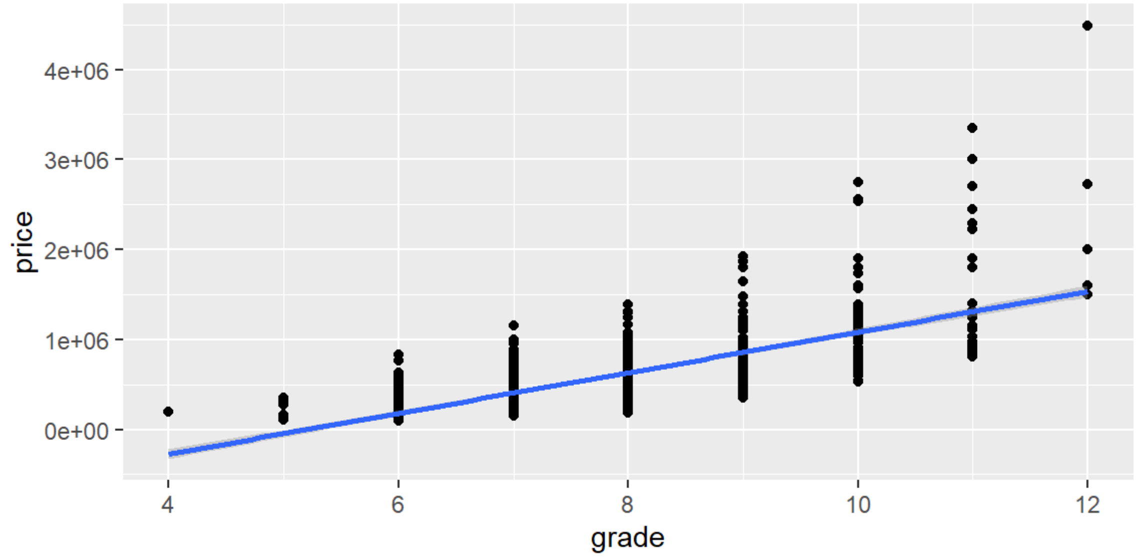
Distribution of Square Footage of the Home

#The distribution of square footage of the home skews to the right.

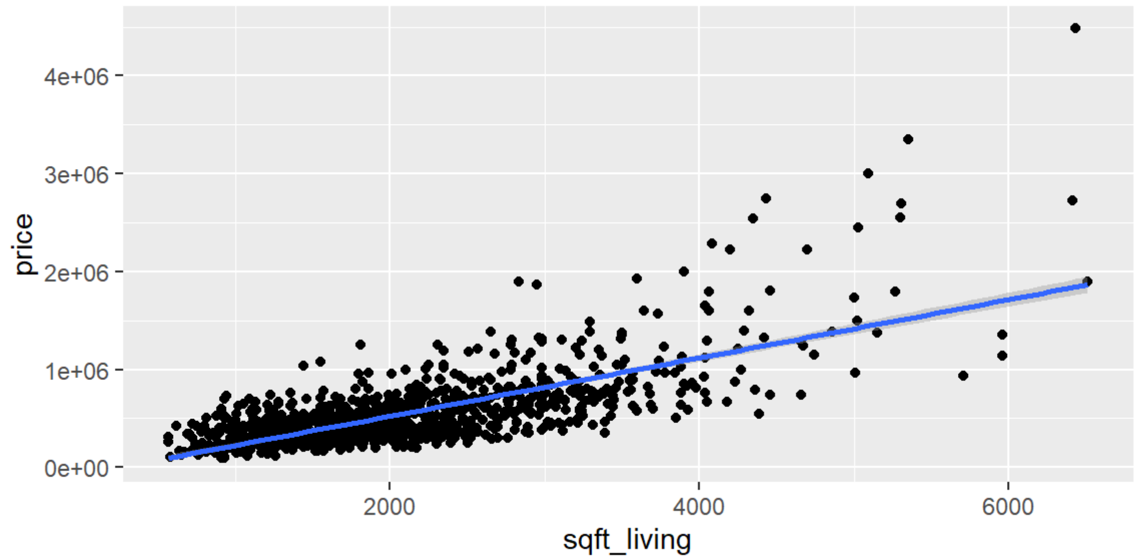


AirBnB Prices

#The distribution of price skew to the right.



Price by Grade



Price by Square Footage of the Home

Average Price by Waterfront

waterfront	Mean_Price	SD_Price	N
FALSE	548096.8	381146.7	993
TRUE	1296285.7	715572.9	7

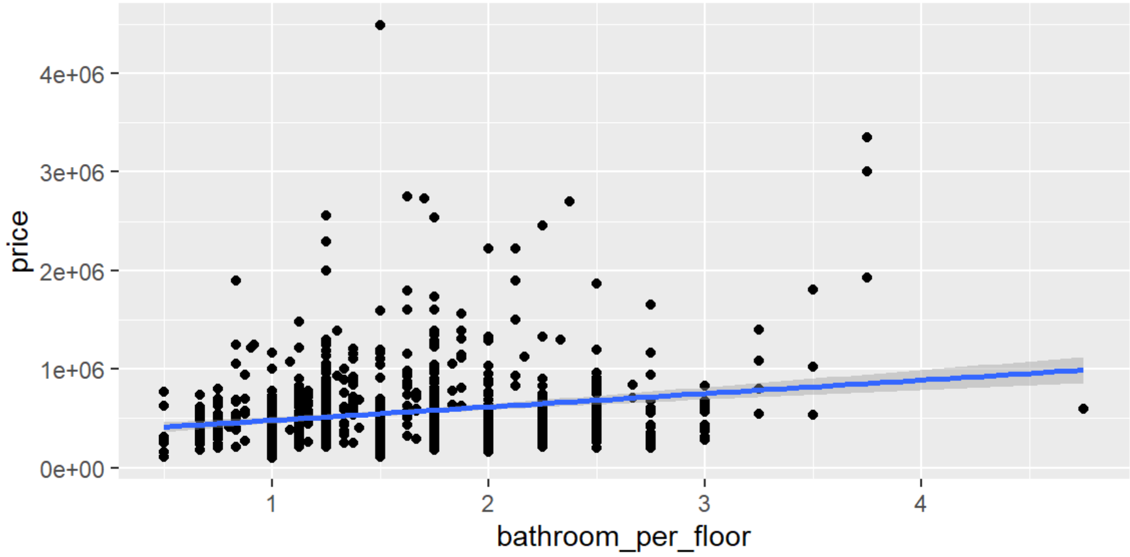
Feature Engineering

Create new variables, or modify existing variables. Include description of each variable you change and create, and relevant table or graph.

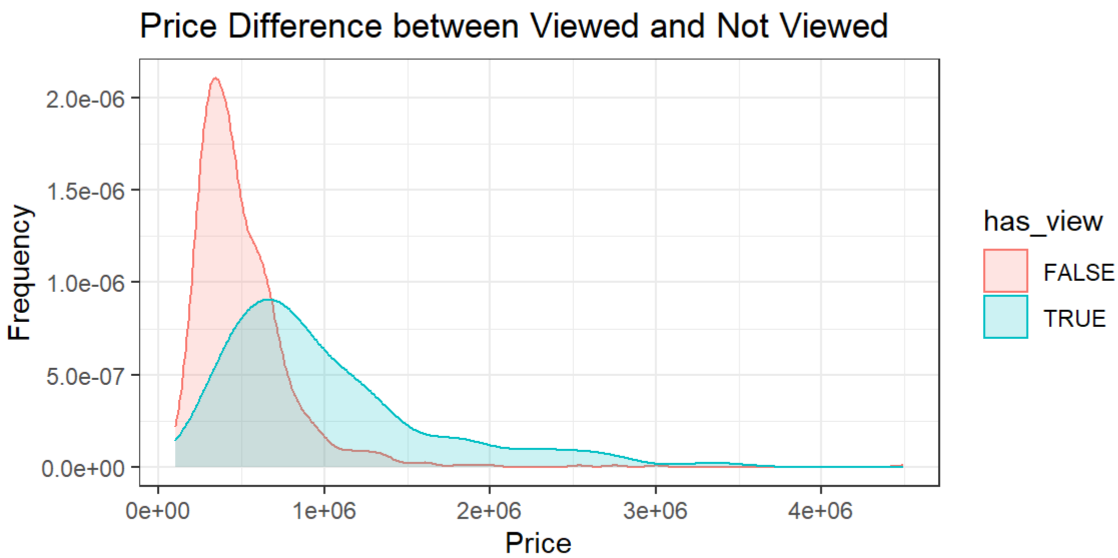
More variable engineering:

1. group together infrequent property types into a category called “other”. We keep the four most frequent categories: apartment, house, townhouse, and condominium.
2. create a yes/no variable for whether or not there was an online review.
3. modify the `host_has_profile_pic` and `host_identity_verified` variables to group missing values and false’s together.
4. create a variable to tell whether the last review was made on a weekend.

#The reason I create bathroom per floor is it is more convinient if there are more bathrooms per floor. Hence, the price will increase if we make the house more convinient. Since both 'view' and 'sqrt_basement' have high correlation with price, I want to see whether the house has been viewed or whether there is a basement can also affect the price.

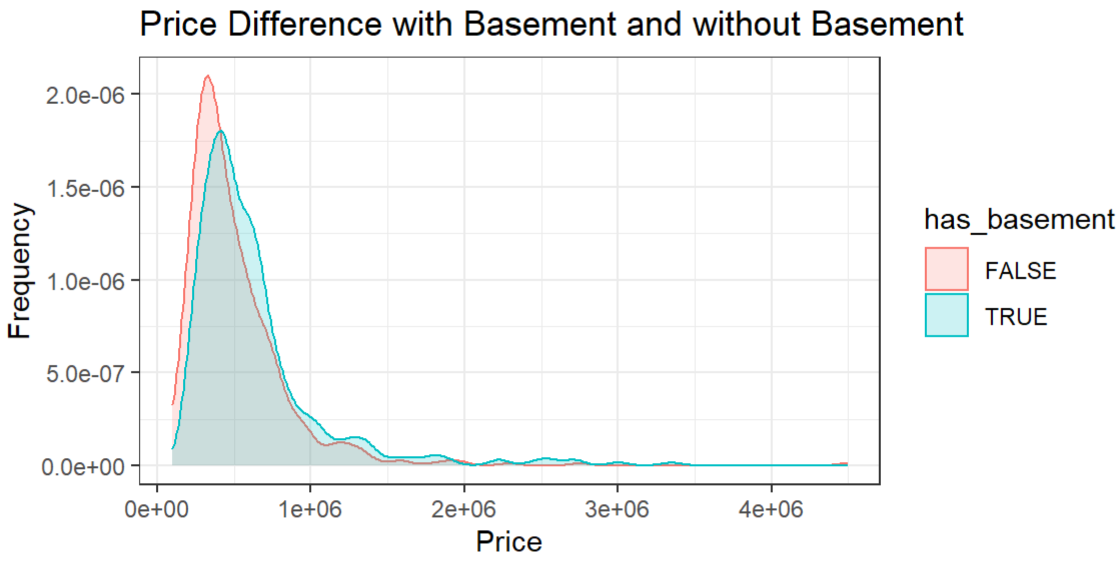


Price Distribution by bathroom per floor



Average Price by View

has_view	Mean_Price	SD_Price	N
FALSE	504756.6	319751.2	901
TRUE	995438.4	619256.4	99



Average Price by Basement

has_basement	Mean_Price	SD_Price	N
FALSE	505919.1	347458.7	598
TRUE	623866.9	434357.6	402

Model Evaluation

```
##                       id price bedrooms bathrooms sqft_living sqft_lot floors
## id                  1.00  0.01     0.03      0.01        0.00    -0.12   0.00
## price               0.01  1.00     0.32      0.52        0.72     0.14   0.27
## bedrooms            0.03  0.32     1.00      0.50        0.59    -0.05   0.13
## bathrooms           0.01  0.52     0.50      1.00        0.73     0.04   0.49
## sqft_living         0.00  0.72     0.59      0.73        1.00     0.16   0.33
## sqft_lot           -0.12  0.14    -0.05      0.04        0.16     1.00   0.02
## floors              0.00  0.27     0.13      0.49        0.33     0.02   1.00
## view                0.01  0.34     0.00      0.15        0.26     0.06   0.07
## condition          -0.06 -0.01     0.08     -0.09       -0.05    -0.03  -0.27
## grade              -0.01  0.70     0.35      0.67        0.76     0.14   0.47
## sqft_above         -0.02  0.64     0.46      0.65        0.87     0.19   0.50
## sqft_basement       0.04  0.29     0.36      0.31        0.44    -0.01  -0.26
## yr_built            0.01  0.07     0.13      0.51        0.30     0.08   0.48
## yr_renovated        0.05  0.11     0.02      0.05        0.06    -0.02   0.02
## zipcode            -0.01 -0.11    -0.13     -0.21       -0.22    -0.13  -0.02
## lat                 0.02  0.29     0.01      0.03        0.05    -0.08   0.07
## long                0.01  0.05     0.12      0.22        0.27     0.25   0.11
## sqft_living15      -0.03  0.60     0.39      0.55        0.76     0.13   0.26
## sqft_lot15         -0.13  0.10    -0.05      0.01        0.15     0.82  -0.01
## bathroom_per_floor  0.01  0.21     0.35      0.51        0.35     0.01  -0.45
##                     view condition grade sqft_above sqft_basement yr_built
## id                  0.01     -0.06 -0.01      -0.02          0.04     0.01
## price               0.34     -0.01  0.70       0.64          0.29     0.07
## bedrooms            0.00      0.08  0.35       0.46          0.36     0.13
## bathrooms           0.15     -0.09  0.67       0.65          0.31     0.51
## sqft_living         0.26     -0.05  0.76       0.87          0.44     0.30
## sqft_lot            0.06     -0.03  0.14       0.19         -0.01     0.08
## floors              0.07     -0.27  0.47       0.50         -0.26     0.48
## view                1.00      0.00  0.27       0.19          0.19    -0.05
## condition           0.00      1.00 -0.14      -0.15          0.18    -0.30
## grade               0.27     -0.14  1.00       0.77          0.14     0.46
## sqft_above          0.19     -0.15  0.77       1.00         -0.05     0.41
## sqft_basement       0.19      0.18  0.14      -0.05          1.00    -0.14
## yr_built           -0.05     -0.30  0.46       0.41         -0.14     1.00
## yr_renovated        0.10     -0.03  0.00       0.02          0.09    -0.24
## zipcode             0.01     -0.04 -0.21      -0.28          0.05    -0.35
## lat                 0.01     -0.05  0.10       0.00          0.11    -0.15
## long               -0.05     -0.10  0.22       0.37         -0.13     0.43
## sqft_living15       0.31     -0.09  0.71       0.74          0.19     0.30
## sqft_lot15          0.06      0.00  0.11       0.17          0.00     0.06
## bathroom_per_floor  0.05      0.18  0.18       0.07          0.58     0.07
##                    yr_renovated zipcode   lat  long sqft_living15 sqft_lot15
## id                         0.05   -0.01  0.02  0.01         -0.03      -0.13
## price                      0.11   -0.11  0.29  0.05          0.60       0.10
## bedrooms                   0.02   -0.13  0.01  0.12          0.39      -0.05
## bathrooms                  0.05   -0.21  0.03  0.22          0.55       0.01
## sqft_living                0.06   -0.22  0.05  0.27          0.76       0.15
## sqft_lot                  -0.02   -0.13 -0.08  0.25          0.13       0.82
## floors                     0.02   -0.02  0.07  0.11          0.26      -0.01
## view                       0.10    0.01  0.01 -0.05          0.31       0.06
## condition                 -0.03   -0.04 -0.05 -0.10         -0.09       0.00
## grade                      0.00   -0.21  0.10  0.22          0.71       0.11
## sqft_above                 0.02   -0.28  0.00  0.37          0.74       0.17
## sqft_basement              0.09    0.05  0.11 -0.13          0.19       0.00
## yr_built                  -0.24   -0.35 -0.15  0.43          0.30       0.06
## yr_renovated               1.00    0.08  0.05 -0.10          0.01      -0.02
```

```
## zipcode                0.08    1.00  0.23 -0.58          -0.33      -0.14
## lat                    0.05    0.23  1.00 -0.15           0.04      -0.10
## long                  -0.10   -0.58 -0.15  1.00           0.35       0.18
## sqft_living15          0.01   -0.33  0.04  0.35           1.00       0.14
## sqft_lot15            -0.02   -0.14 -0.10  0.18           0.14       1.00
## bathroom_per_floor     0.02   -0.18 -0.02  0.08           0.24       0.02
##                   bathroom_per_floor
## id                              0.01
## price                           0.21
## bedrooms                        0.35
## bathrooms                       0.51
## sqft_living                     0.35
## sqft_lot                        0.01
## floors                         -0.45
## view                            0.05
## condition                       0.18
## grade                           0.18
## sqft_above                      0.07
## sqft_basement                   0.58
## yr_built                        0.07
## yr_renovated                    0.02
## zipcode                        -0.18
## lat                            -0.02
## long                            0.08
## sqft_living15                   0.24
## sqft_lot15                      0.02
## bathroom_per_floor              1.00
```

We consider 6 models:

1. simple linear regression model using only sqft_living as explanatory variable. I create this model because 'sqft_living' is one of the highly correlated variable with price.
2. multiple regression model with the five quantitative explanatory variables most highly correlated with price but not correlated with other varibles. I create this model because these five quantitative variables highly correlated with the price. If I choose explanatory variables which highly correlated with the other, then as one explanatory variable changes, other variables will also change, so our model may not be accurate.
3. same variables as in (2), with interactions included. Since in model 2, my explanatory variables do not highly correlated with each other, so interactions term may create better model.
4. multiple regression model with sqrt_living, and two categorical variables: waterfront, and has_view. Since previous models doesn't include any categorical variables, so I want to create a model that include all important categorical variables.
5. same model as in (4), with interactions included. Since all categorical variables are not correlated with each other, we can create a model which include interactions.
6. multiple regression model with combination of categorical and quantitative variables mentioned so far(exclude view). Since all variables I mention so far are important for prediction, I create a model include all variables. However, I exclude view because I include has_view, and these two variables are correlated.

7.same model as in (6), with interactions included. Since explanatory variables in model 6 are not highly correlated with each other, we can create the model which include interactions.

Cross Validation Results

Model	RMSPE
1	269675.5
2	261472.6
3	277455.5
4	260001.8
5	260362.3
6	259453.8
7	2910675.4

We also consider predicting log(price), using the same 7 models.

Cross Validation Results for Log Model

Model	RMSPE
1	0.3838850
2	0.3713272
3	0.3766700
4	0.3735125
5	0.3728334
6	0.3702307
7	2.8422749

Model 6 performs better on both price and log(price) but we can't compare these directly using the R output from `caret` because RMSPE is computed on different scales.

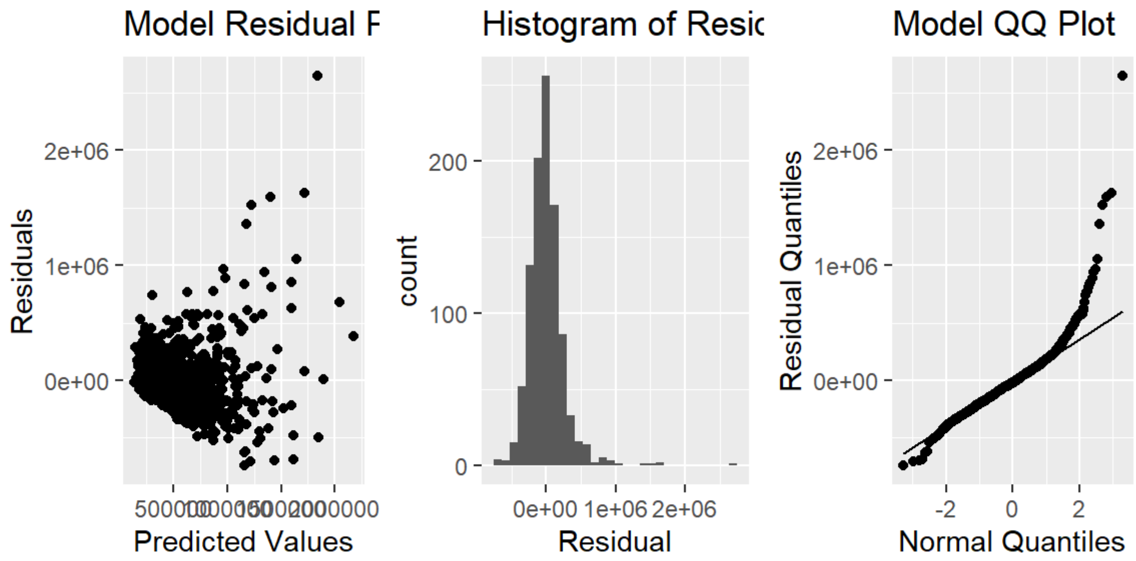
Instead, we'll convert the predictions for log(price) back to price, and calculate RMSPE ourselves for these two models. We partition the data into a training set, containing 80% of the data, and a test set, containing the remaining 20%, and repeat this procedure 10 times. This is not true cross-validation, since we aren't dividing into distinct folds, and withholding each fold once, but it has the same effect of evaluating the model on data not used to train it.

Comparsion of MSPE for Log and Original Response Scales

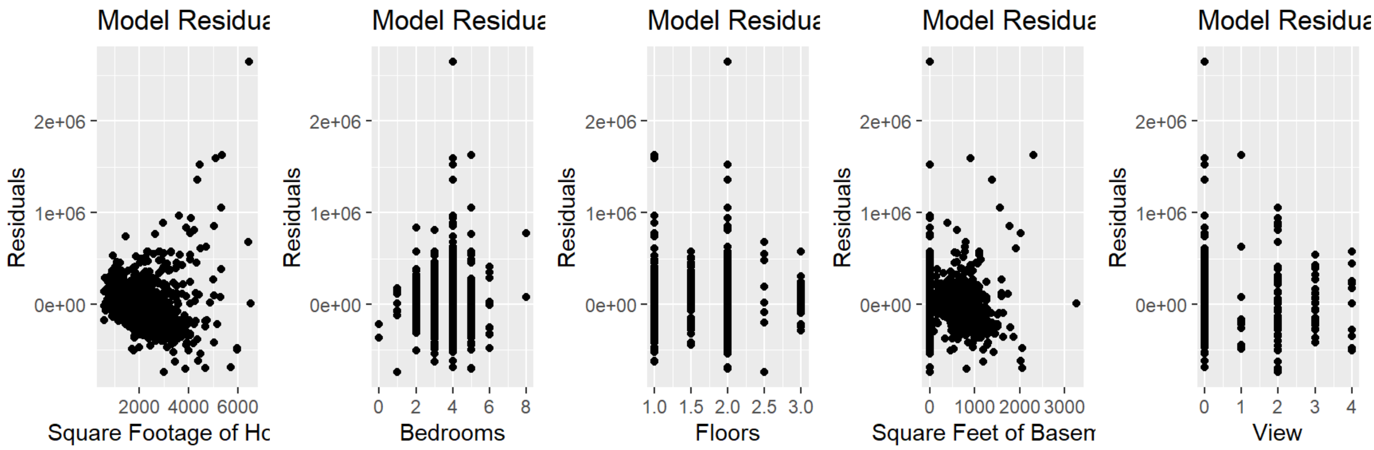
Model	RMSPE
Original	257503.4
Log of Exp. Vars	271066.7

We see that the model that did not use the log transfrom performed better.

We'll create residual plots for this model.



Plots for Model Check



Residual by Explanatory Variable Plot

The residual by explanatory variable plots, show that some of variables have some large outliers. This is not a big problem, model assumption violations are sometimes ok when we’re only interested in prediction, but it’s possible that correcting these will improve predictions.

Now, we make the predictions on the new data.

Conclusions

In this project, I create seven different models, the most useful model is model 6 which includes sqft_living, bedrooms, floors, sqft_basement, has_view, and waterfront. The least useful model is model 7 which also includes sqft_living, bedrooms, floors, sqft_basement, has_view, and waterfront, but includes interaction. Since the best model and the worst model contain the same explanatory variables, we can’t determine which variables are useful and which are not useful.

I create three variables but I think only one of them is useful, ‘has_view’. However, since ‘view’ and ‘has_view’ correlated in some way, I choose ‘has_view’ and ignore ‘view’. If it is possible, I would like to know which cities houses are located in, because it is obviously that if fixed all other conditions, the houses’ price in Washington is higher than houses’ price in Appleton.