Predicitng a House’s Price

Toan Pham

11/22/2020

Abstract

In this project, we will predict the price of 1000 houses in King County, Washington in 2014 and 2015, using a training data set of 1000 other houses sold in King County during the same time period. We attempted to build a model that can give the best prediction of house prices. Using cross-validation, we choose one amongst 7 different models using different variables. The final model have 16 variables: number of bedrooms, number of bathrooms, living area, lot area(2015 and original), number of floor, waterfront, view, condition, grade, area above, basement area, year built, year renovated, day passed.

Introduction

We developed model to predict the prices of 1000 houses sold in King County, Washington in 2014 and 2015 , using another training data set of 1000 houses sold in King County during the same time period. Using different graphs and tables, we determine the strong variables that will be put into the model. There are a total of 7 different models using different variables in this report. Using cross-validation and testing RMSPE, we choose the model that return the smallest RMSPE. Finally, we use the model to predict the price for the testing data. and engineered some of those variables to give them better ability to predict house prices. We then came up with 6 models using different explanatory variables, and compared their test error (RMSPE) using cross-validation method to get the model with the smallest RMSPE value. We would ultimately use that model to predict the prices of 1000 houses in Test data. In this project, we strive to make the best prediction model for the two data set, without making an overly complicated model. This information will be very useful for people interested in buying or selling a house near the King County area.

Exploratory Analysis

```
##        id                    date         price            bedrooms
##  Min.   :  11200290   2015-04-14: 13   Min.   : 100000   Min.   :0.000
##  1st Qu.:2121634330   2015-03-11: 11   1st Qu.: 320000   1st Qu.:3.000
##  Median :3838100398   2014-06-18:  9   Median : 452000   Median :3.000
##  Mean   :4531346313   2014-12-03:  9   Mean   : 553334   Mean   :3.377
##  3rd Qu.:7209087798   2015-04-01:  9   3rd Qu.: 650862   3rd Qu.:4.000
##  Max.   :9839300775   2014-06-20:  8   Max.   :4489000   Max.   :8.000
##                       (Other)   :941
##    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
##
```

```
## Observations: 1,000
## Variables: 21
## $ id            <dbl> 6403510090, 7879600070, 1025069192, 2487200680, 76044...
## $ date          <fct> 2014-11-11, 2014-10-24, 2014-11-05, 2015-02-24, 2014-...
## $ price         <dbl> 437500, 269950, 929000, 447000, 450000, 550000, 54300...
## $ bedrooms      <int> 4, 4, 4, 2, 4, 3, 3, 4, 4, 4, 4, 3, 2, 4, 5, 3, 2, 3,...
## $ bathrooms     <dbl> 2.50, 2.50, 3.25, 1.00, 2.50, 1.00, 2.25, 3.00, 2.50,...
## $ sqft_living   <int> 2680, 1960, 4030, 720, 2290, 1010, 1240, 2880, 3570, ...
## $ sqft_lot      <int> 7513, 7230, 57499, 7500, 5515, 6120, 949, 5500, 17411...
## $ floors        <dbl> 2.0, 2.0, 2.0, 1.0, 2.0, 1.0, 3.0, 2.0, 2.0, 2.0, 1.0...
## $ waterfront    <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALS...
## $ view          <int> 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ condition     <int> 3, 3, 3, 3, 3, 3, 3, 4, 3, 3, 4, 4, 3, 3, 4, 3, 4, 3,...
## $ grade         <int> 8, 8, 9, 6, 8, 6, 8, 9, 10, 7, 8, 7, 7, 8, 6, 7, 7, 7...
## $ sqft_above    <int> 2680, 1960, 4030, 720, 2290, 1010, 1240, 1920, 3570, ...
## $ sqft_basement <int> 0, 0, 0, 0, 0, 0, 0, 960, 0, 250, 0, 710, 0, 1220, 0,...
## $ yr_built      <int> 1998, 2002, 2002, 1925, 2006, 1942, 2008, 1926, 1990,...
## $ yr_renovated  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ zipcode       <int> 98059, 98023, 98053, 98136, 98106, 98033, 98103, 9810...
## $ lat           <dbl> 47.4956, 47.2855, 47.6617, 47.5185, 47.5518, 47.6648,...
## $ long          <dbl> -122.161, -122.360, -122.026, -122.392, -122.357, -12...
## $ sqft_living15 <int> 2640, 1850, 3470, 1390, 1380, 1260, 1310, 2110, 3510,...
## $ sqft_lot15    <int> 7243, 7208, 57499, 5000, 5515, 5977, 1140, 5500, 1615...
```

In the original Train data set, we have 21 variables and a total of 1000 observations. There are only 1 categorical variable “waterfront”, and the rest are quantitative. Also, there are no missing values in our dataset, so we are going to use all 1000 observations.

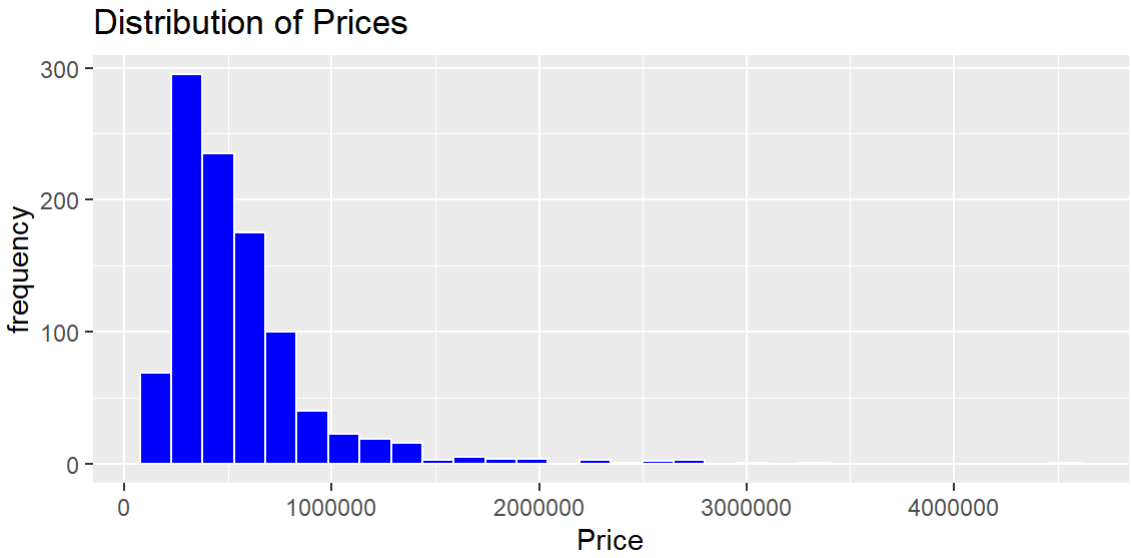


Figure 1 :Price Distribution

Figure 1 show the distribution of prices to be right-skewed. The majority of the houses are under 1 millions USD in price, with most being around 400-500 thousands USD.However, as the histogram is right-skewed, there are a few outliers that are way higher than 1 millions, with the highest being around 4.5 millions.

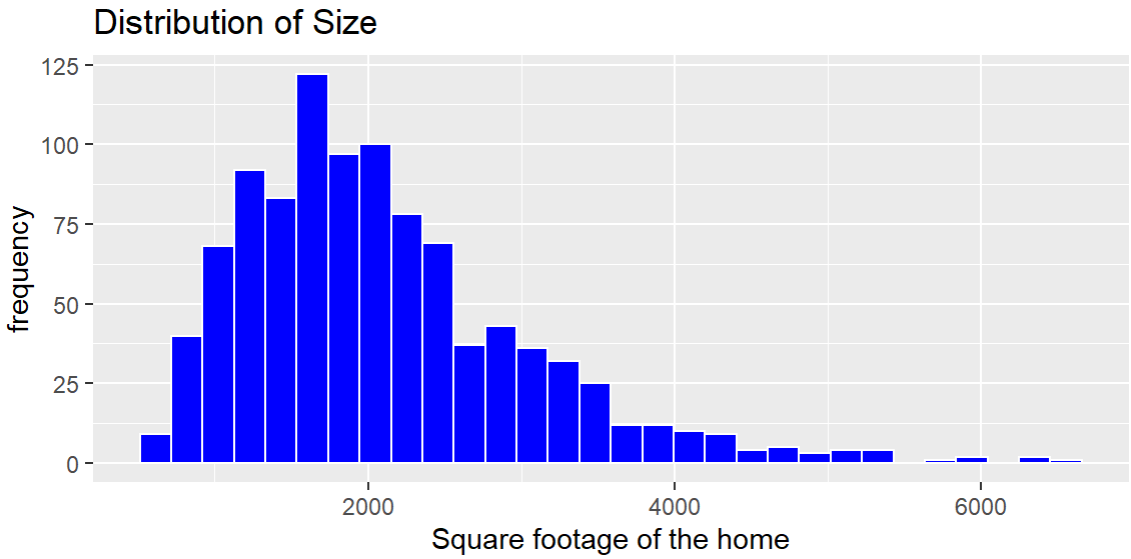


Figure 2: Size of House Distribution

Figure 2 show the distribution of size to also be right-skewed, although the range are not as extreme as price. The majoirty of the houses have size between 1000 and 3500 square feet. Houses that are bigger than 3500 square feet are less common the bigger they get, with houses that exceed 6000 square feet being near non-existant.

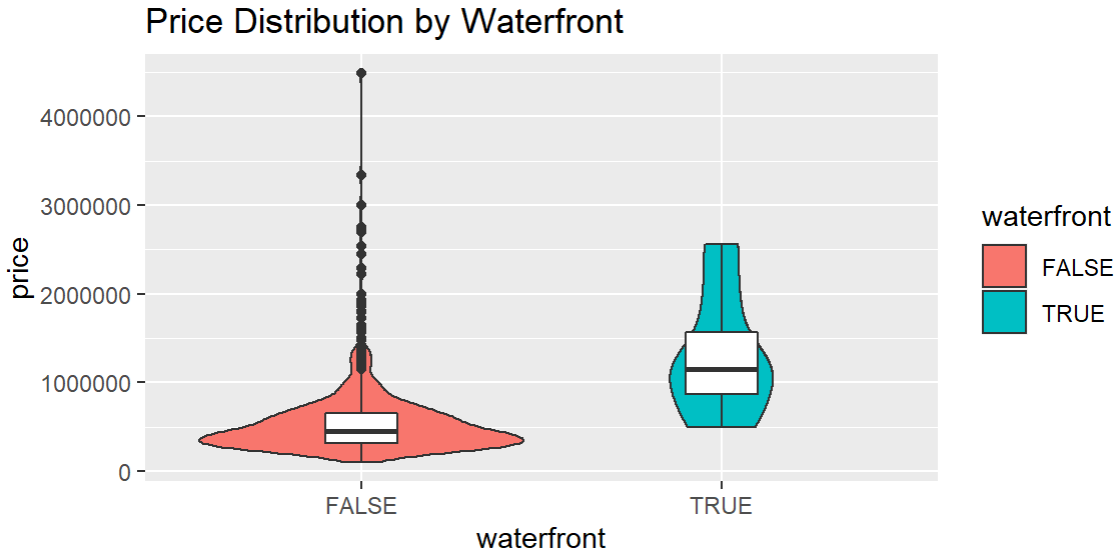


Figure 3: Price Distribution by Waterfront

Figure 3 gives us some interesting results about the waterfront and prices. In general, houses with waterfront are expected to be more expensive than houses without. However, the distribution of price between these two types are quite varied. In non-waterfront houses, the price are extremely consistent. In other words, most houses without waterfront are very similar in price: a bit less than $500000. Houses without waterfront rarely exceed the 1.5 millions USD mark. In constrast, the price distribution for house with waterfront are quite even across all price range, from 500000 USD to 3500000USD. There are no price range that are way more common than others, although the higher end houses are somewhat rarer than cheaper ones. However, one thing that should be noted is that there are only 7 houses without waterfront in our data set, so this observation may not be accurate to real life.

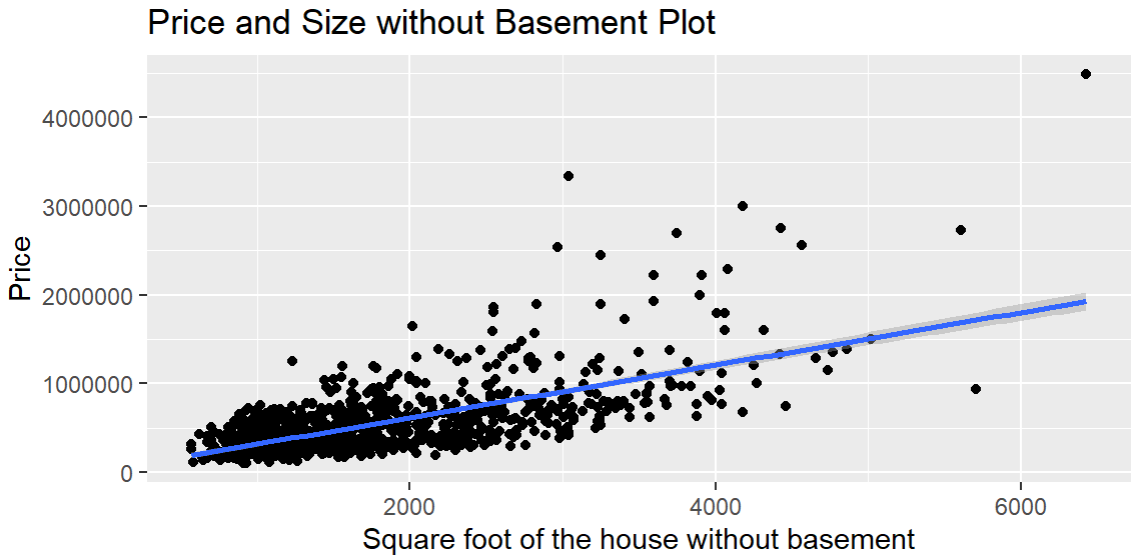


Figure 4: Price and Square Feet without Basement Plot

Figure 4 shows that there is an obvious direct correlation between price and size of house not counting basement. In other words, generally, the bigger the house, the more expensive it gets. One thing to note is that one house is responsible for the strongest outlier for both price and size, costing more than 4.5 millions USD and bigger than 6000 square feet.

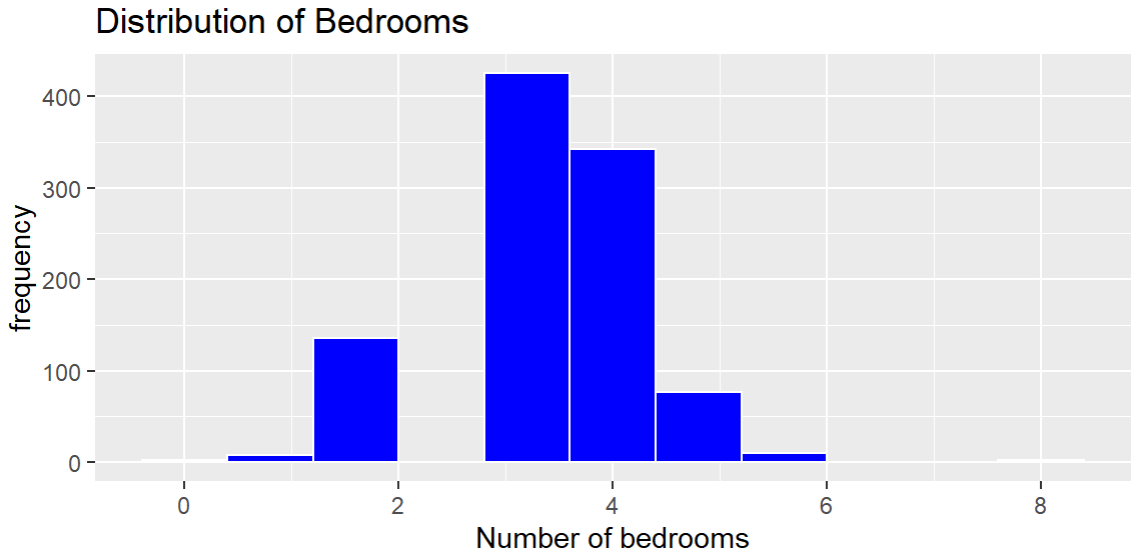


Figure 5: Bedrooms Distribution

Figure 5 show the majority of the houses have 3 to 5 bedrooms. 2 bedrooms houses are also quite common, although not as abundant as the previous range. There are almost no houses that exceed 6 bedrooms, with only one single house with 8 bedrooms.

Average, Standard Deviation, Median Price by Condition

condition	Mean_Price	SD_Price	Median_Price	N
1	105500.0	NA	105500	1
2	410229.2	681650.8	217875	12
3	568446.2	401355.5	458000	655

condition	Mean_Price	SD_Price	Median_Price	N
4	517116.3	340711.5	430000	271
5	587460.2	370747.5	495000	61

Table 1 shows that “condition” is not as straightforward as initially expected. While condition 5 has the highest mean and median price, which implies a direct correlation between condition and price, condition 4 and condition 5 suggest otherwise. More specifically, condition 3 actually have both higher mean price and median price compared to condition 4. Although the difference is not extreme as condition 2 vs condition 3, this may be worth taking under consideration when building our model. There is only one house with condition 1 so we do not consider it.

Average, Standard Deviation, Median Price by View

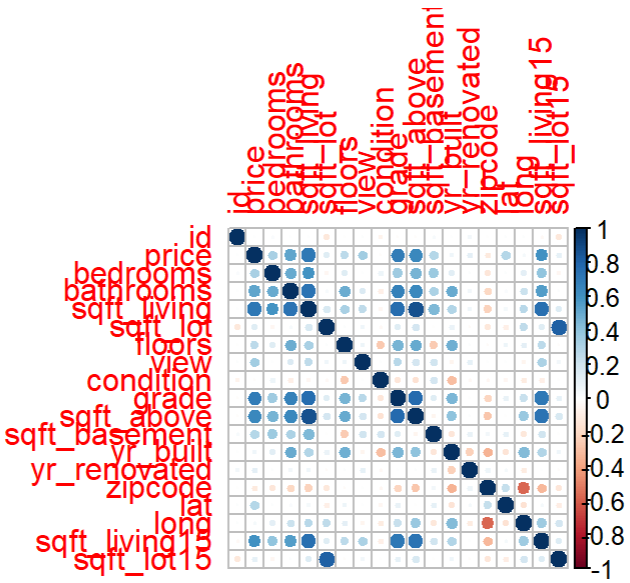
view	Mean_Price	SD_Price	Median_Price	N
0	504756.6	319751.2	429900	901
1	1046227.3	994506.2	735000	11
2	1039065.6	636953.9	795900	48
3	900691.7	476157.3	775000	30
4	1014400.0	428743.7	1067500	10

Table 2 shows that houses that have never been viewed are cheaper than houses that have been viewed. Also, the most majority of the houses also have never been viewed. On the other hand, the number of time viewed does not seem to have much effect on the price. The mean price across 1-4 time viewed does not have a lot of difference.

Average, Standard Deviation, Median Price by Floors

floors	Mean_Price	Mean_sqftliving	Mean_grade	N
1.0	440031.5	1745.458	7.112450	498
1.5	581581.0	1952.321	7.107143	84
2.0	684117.4	2616.606	8.415550	373
2.5	1116475.0	3497.500	9.250000	8
3.0	574000.0	1660.568	8.054054	37

Table 3 shows us the relationship between floors and mean price, size, and grade of the house. There are only 8 houses that have 2.5 floors so we do not consider it. In general, 2 floors seems to be the best deal when it comes to choosing houses. Houses with 2 floors boast highest mean price and size, as well as highest grade. Houses with 3 floors actually perform worse than 2 floors according to these metrics.



Correlation Plot for Quantitative Variables

The correlation plot show us which variables to include/avoid in our model.“sqft living”, “grade”, “sqft_above” and “sqft_living15” are highly correlated with the response variable “price”. “sqft_living”, “sqft_living15”, “grade”, “sqft_above” and “bathrooms” are highly correlated with each other. “price”. Furthermore, “sqft_living15”, “grade”, “sqft_above” and “bathrooms”. “sqft_lot” and “sqft_lot15” are also highly correlated with each other. In general, we should avoid using the variables that are highly correlated, as well as avoid choosing variables that does not seem to belong in our model, such as “id”, “lat”, and “long”. The rest of the variables can be considered to be put into the model. ## Feature Engineering

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

As seen by table 1, while having no view drastically lower the price, the number of times views does not have a huge impact on the price of the houses. Thus, I decide to turn the quantitative variable “view” to a logical categorical variable “viewBoolean”, which indicates whether people have viewed the house or not.

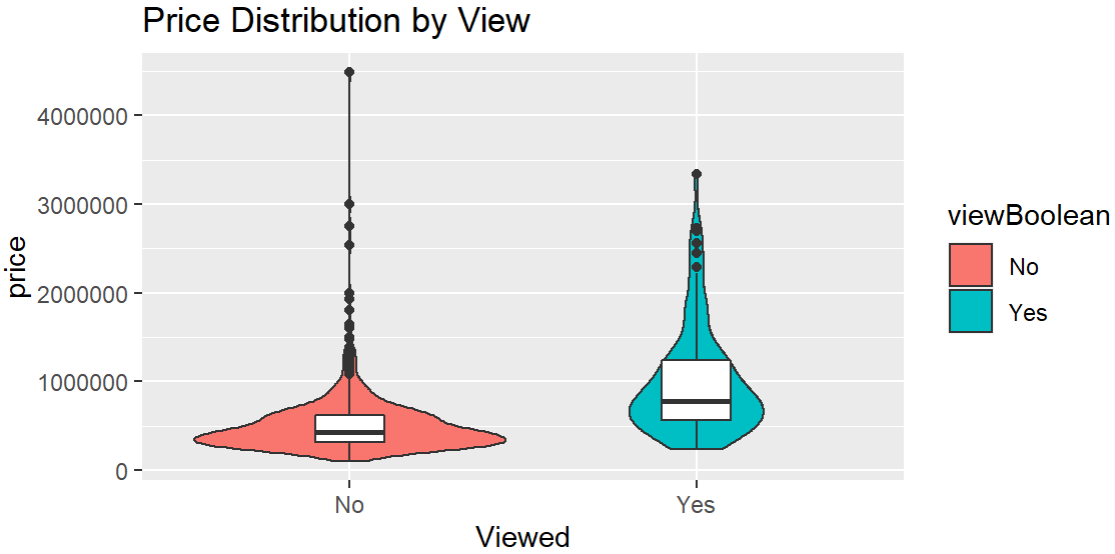


Figure 6: Price by View

Consistent with the hypothesis, figure 6 shows houses that haven’t been viewed have lower price than its counterpart. It also shows how houses with no view have more consistent price, with most houses fall right

under the 500000 USD price mark. In contrast, houses with views are more evenly distributed across price ranges, with a little bit above 500000 USD seems to be the most popular range.

As the number of bathrooms and bedrooms both have direct correlation with price, I combine them into one variable “necessity” to simplify the model and avoid using two similar variables. “Necessity” should also have direct correlation with price.

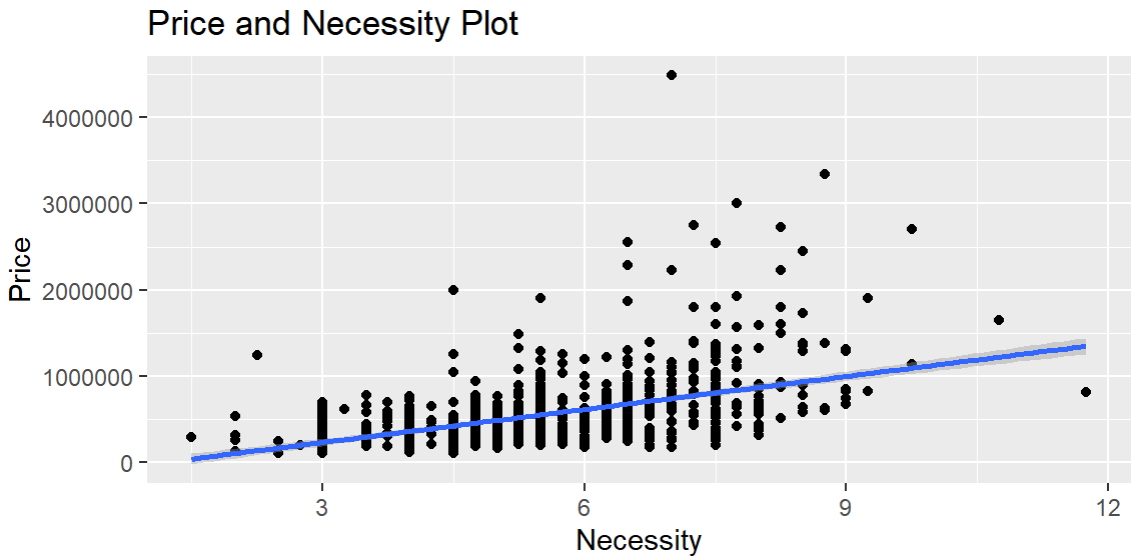


Figure 7: Price and Necessity Plot

True to our hypothesis, “necessity” have a direct correlation with price. With this, we can include “necessity” in our model instead of “bathrooms” and “bedrooms”.

Similar to “view”, a new logical variable “renovated” is necessary to determine whether the house has been renovated or not. I believe a renovated house should cost more than a normal house (as well as easier to use than knowing the year it has been renovated), that’s why I decide to make this variable.

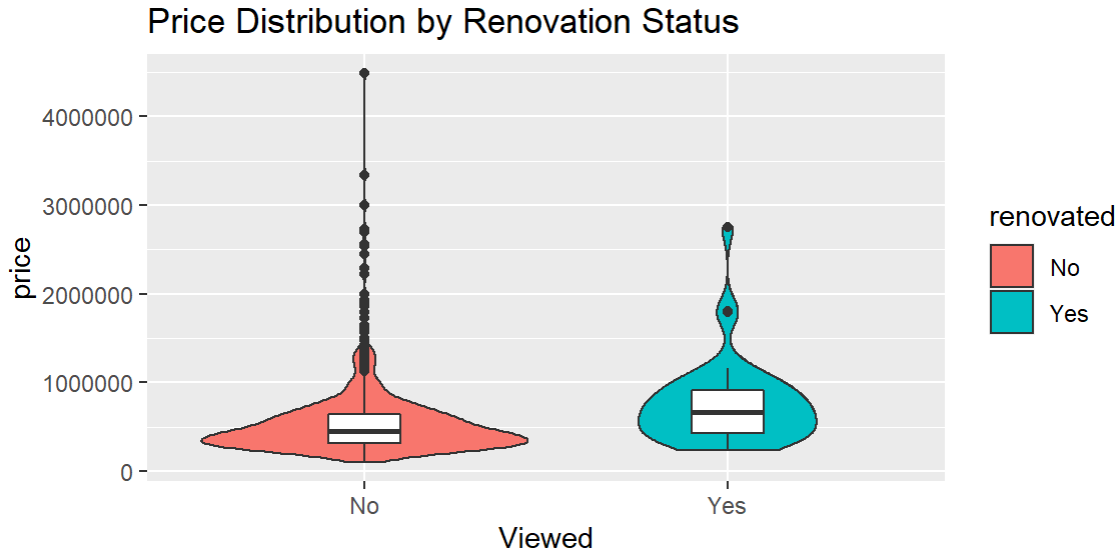


Figure 8: Price by Renovated

Renovation Table

renovated	Mean_Price	N

renovated	Mean_Price	N
No	544621.6	960
Yes	762433.8	40

Again, true to our hypothesis, renovated houses have higher mean price. Although there are only 40 renovated houses in our data set, it should be sufficent to include “renovated” in our model.

Model Evaluation

We’ll perform 5 repeats of 5-fold cross validation.

Finding optimal lambda for Ridge Regression:

```
## [1] 1000
```

Finding optimal cp for decision tree:

	cp <dbl>
3	0.001321941

1 row

We consider 7 models:

1. simple linear regression model using only grade as explanatory variable. “grade” is overall grade given to the housing unit, based on King County grading system, so it should play an important role in determining the price of the house.
2. Similar to model 1, but with polynomial regression. I want to fit a model with “grade” but with higher powers, with this model being 3.
3. Multiple regression model with “waterfront”,“necessity”,“renovated”,“viewBoolean”, and “sqft_living15”. A model with two strong quantitative variables that are not highly correlated as well as some categorical variables. I believe this is enough to determine the price of a house, without being worried about overfitting.
4. Similar to model 3, but include interaction. As I did not check for interaction before, I would like to see if including interaction will help predict better or not
5. model including almost all variables, and leaving out only those that we wouldn’t expect to have much relationship with price ((latitude, longitude, id, zipcode). By including all the variables that have effect on the response variable, I can get a high RMSPE.
6. A more experimental model using ridge regression, with the same variables as model 5, and optimal lambda = 1000.
7. A more experimental model using decision tree, using optimal cp = 0.001 and model 5’s variables.

Cross Validation Results

Model	RMSPE
1	277042.2

Model	RMSPE
2	260817.0
3	288660.6
4	875225.4
5	231598.5
6	231062.7
7	256876.4

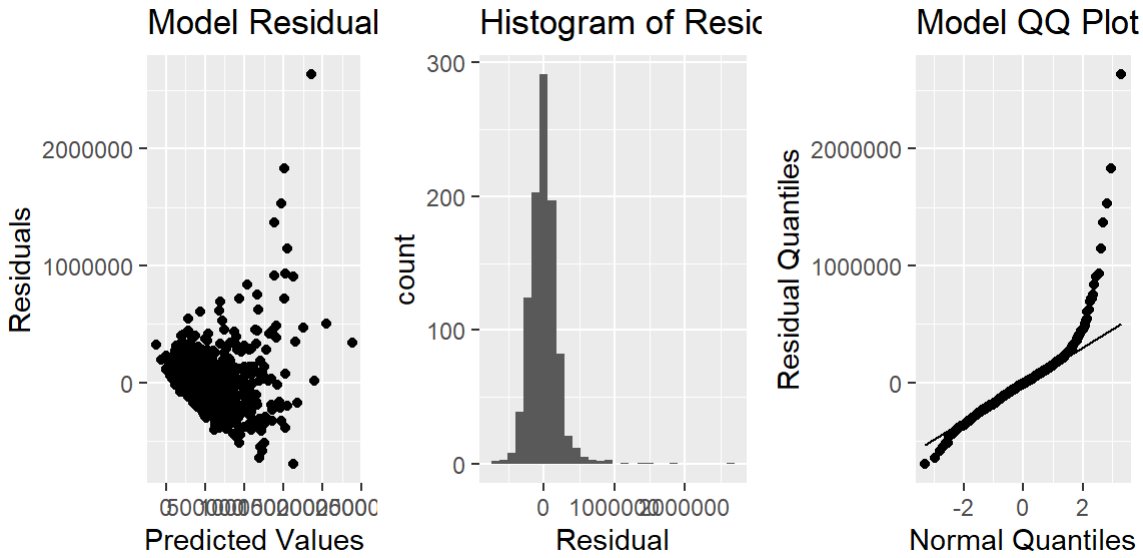
We also consider predicting log(price), using the same 8 models.

Cross Validation Results for Log Model

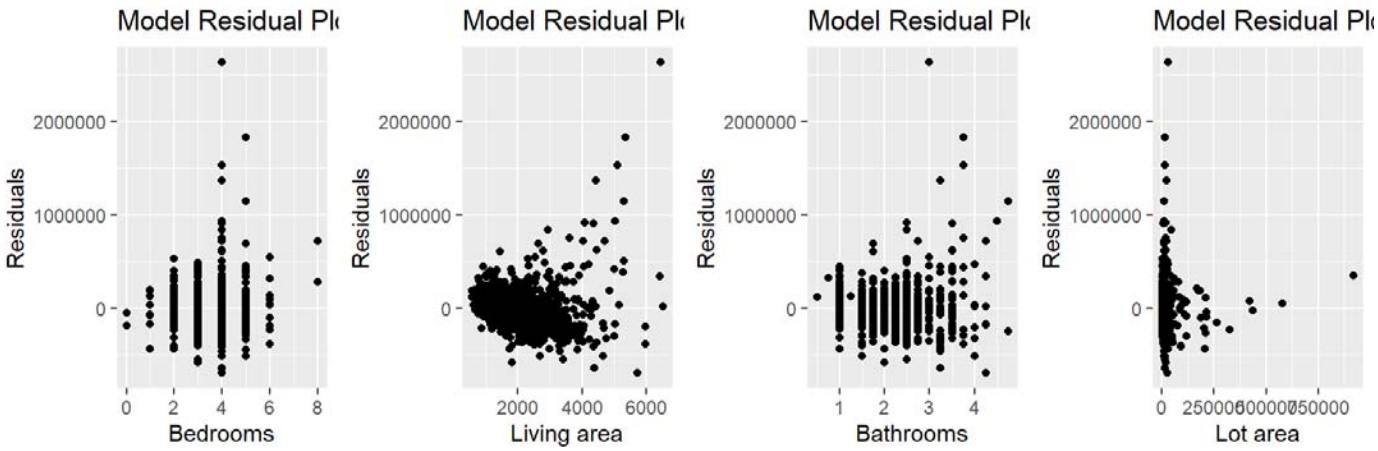
Model	RMSPE
1	0.3786307
2	0.3793743
3	0.3927296
4	0.8559155
5	0.3108689
6	0.5463043
7	0.3699935

We see that model 6 and model 5 is quite similar in performance in predicting direct price, while model 5 was best at predicting log(price). For the sake of simplicity, I dedide to use model 5 to predict model.

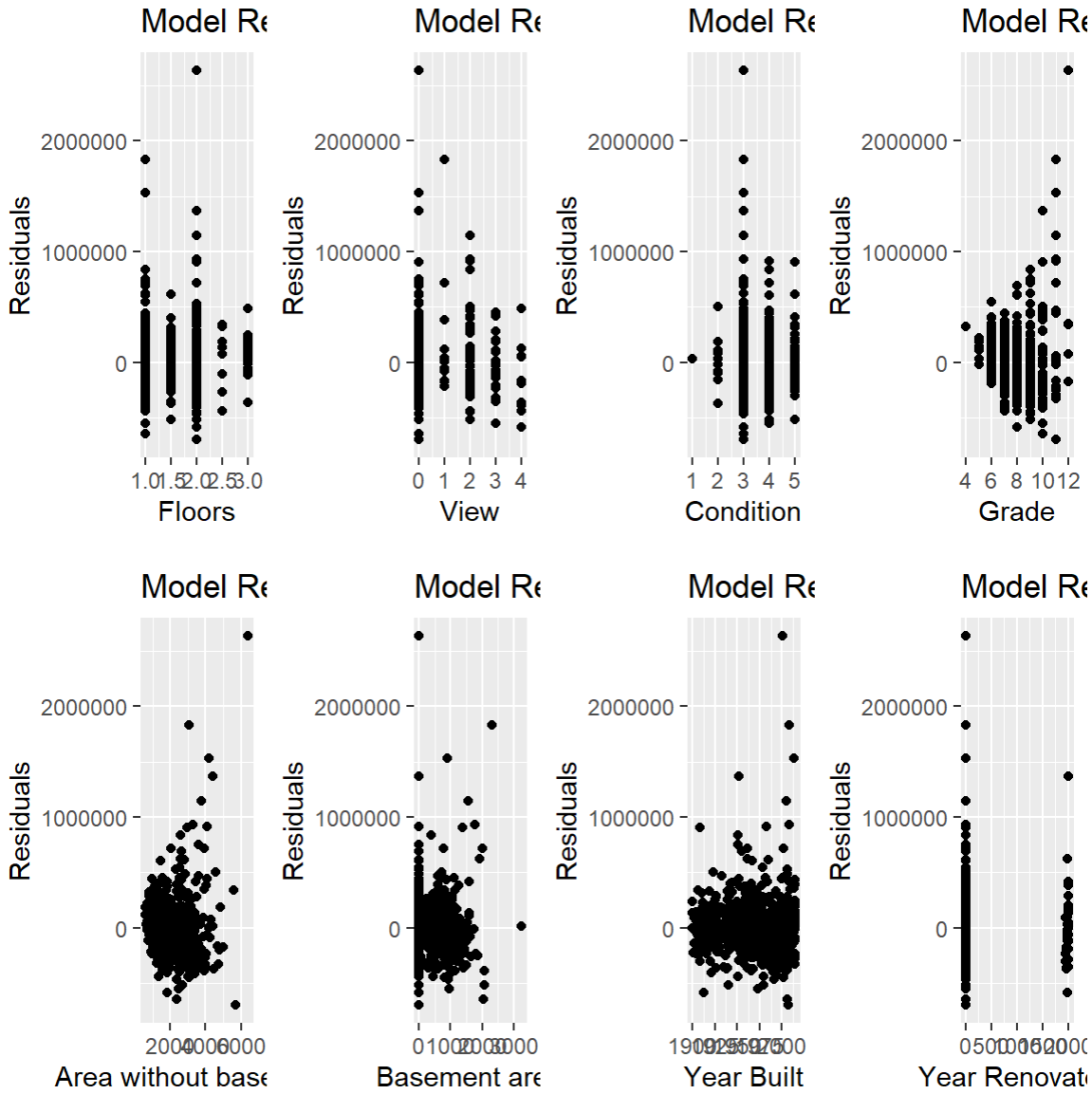
We'll create residual plots for this model.

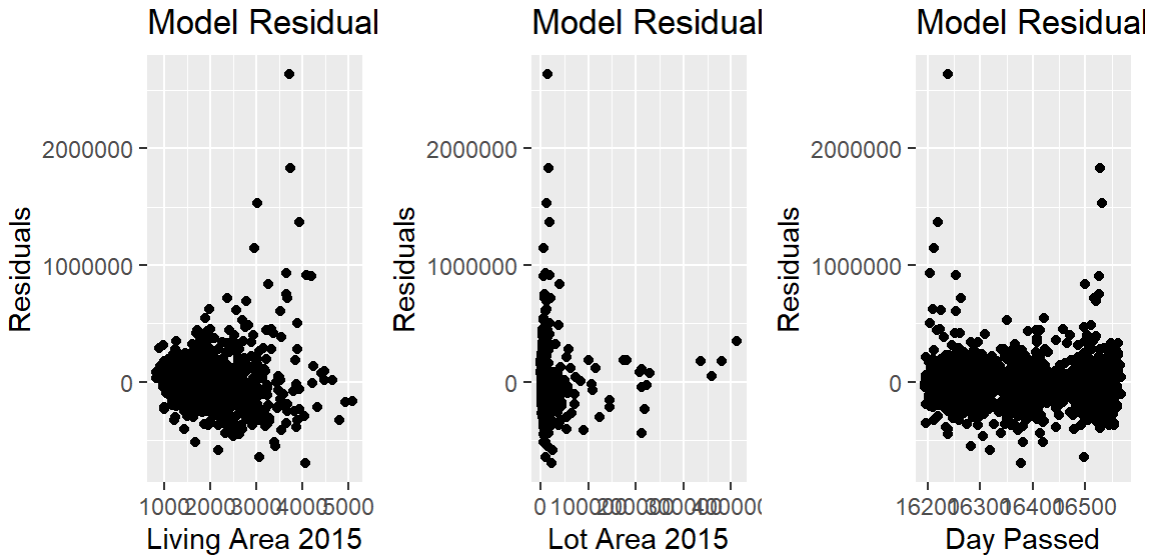


Plots for Model Check



Residual by Explanatory Variable Plot





The residual by explanatory variable plots, show that there are some outliers in our data.This is in line with our graphs ealier and should not be a problem in our prediction. On that same note, there are also some constance variance and normality assumption violation in our models. Again, because we are building this model for predicting purposes, this should not have a big impact on our prediction. However, it's possible that correcting these will improve predictions.

Now, we make the predictions on the new data.

Conclusions

Among the 7 models we made, model 5 of multiple regression using 17 explanatory variables returns the best result. This is rather surprising, as model 5 includes some highly correlated variable like sqft_living, sqft_lot and sqft_living15, sqft_lot15, Log transformation is not necessary in our report, as the RMSPE does not differ much. Engineereed variables turn out to not help much, as model 4 which includes all of them does not return as good of a result. Although, this may be caused by the low number of variables in model 4. Model 2 also performs surprisingly good with only one explanatory variables. This shows that "grade" is a great indication of price in our data set, and should always be considered when making another model.

The variables that were not very helpful are zipcode, lat, long, and id. This is to be expected, as these are values given to a house and not very related to price. I did not include them in my models, although even if including them returns a lower RMSPE, the model would hardly be reliable.

Ridge Regression and Decision Tree did not help much with our model. Although Ridge Regression did return a lower RMSPE, the difference is negligible and I do not use it for the sake of simplicty of the model. Decision tree is a lot more interesting, as using decision tree with all variables and cp = 0.001 did return a low RMSPE value of ~213000 (better than all 7 included model), I find it too unreliable due to including ID, lat, long as variables in the tree. Due to my lack of expertise with decision tree, I decided to not include it in our 7 models, although it may be worth taking under consideration.

This report provides a good model for predicting price in King County, Washington. Although the location is rather limited, we can easily expand this to use on different areas as well, not just King County.