Predicting Housing Prices in King County

Alisa Sima, Andy Yu, August Chen, Harrison Lian

King County is the most populous county in Washington with over 2 million residents and the 12th most populous county in the United States. It is also home to Microsoft and Amazon, two of the largest tech companies. Given that Seattle is a hub of innovation and attractive place to live, the population has increased over time, thus increasing demand for housing. Realtors interested in selling their houses may want to pay close attention to what factors influence the price of a house. Thus, we are interested in analyzing what variables affect housing prices in King County and ultimately try to predict the price.

**The Data**

We obtained our dataset from Kaggle. Our data consists of 21,613 observations and 21 variables, where each observation represents a home sold between May 2014 and May 2015.

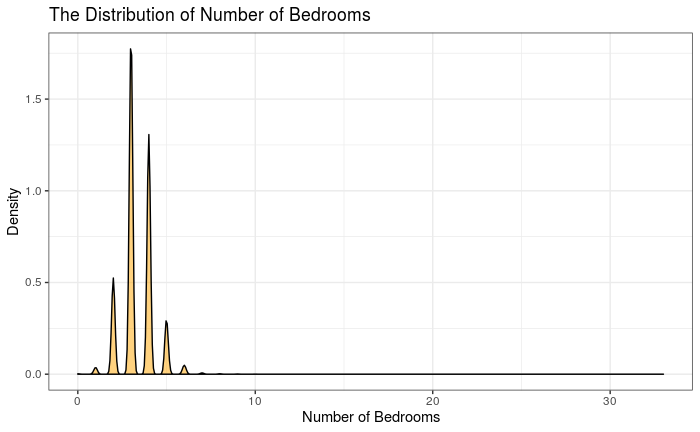
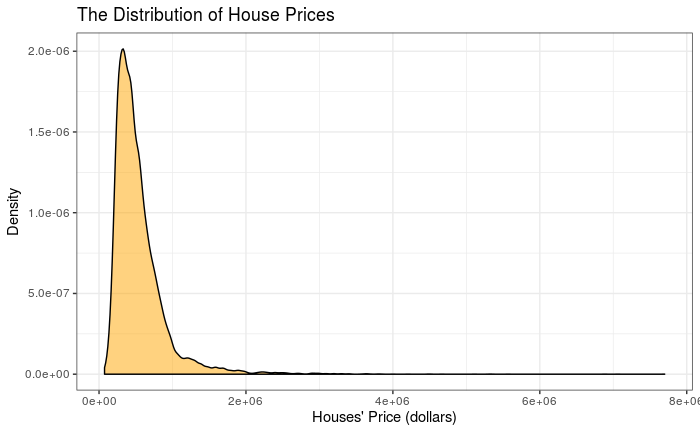
Due to our interest in analyzing and predicting the house price, we examine the relationship between house prices and other explanatory variables. The descriptions of the variables are as follows:

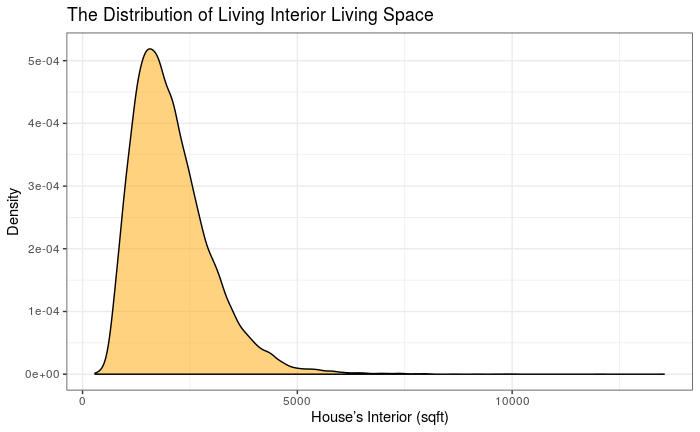
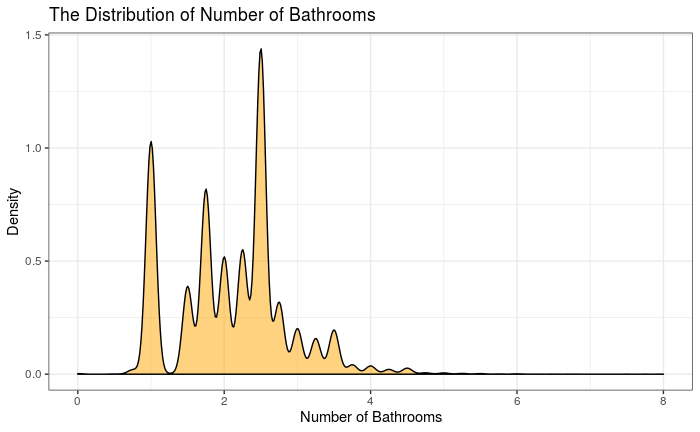
* `id` : The unique numeric number assigned to each house being sold.
* `date` : The date on which the house was sold.
* `price`: The sale price of each house.
* `bedrooms` : The number of bedrooms in a house.
* `bathrooms` : The number of bathrooms in a house (.5 refers to a bathroom with no shower).
* `sqft\_living` : The measurement in square feet of the house’s interior living space.
* `sqft\_lot` : The measurement in square feet of the house’s land space.
* `floors`: The number of floors of the house.
* `waterfront` : Variable describing whether a house has a view of the waterfront with 0 meaning no and 1 meaning yes.
* `view` : Variable describing how good of a view the house has with values ranging from 0 to 4.
* `condition` : The overall condition of a house on a scale of 1 to 5.
* `grade` : The overall grade given to the housing unit, based on the King County grading system on a scale of 1 to 13.
* ` sqft\_above` : The measurement in square feet of the house that is above ground level.
* `sqft\_basement` : The measurement in square feet of the basement of the house.
* `yr\_built` : The date the house was built.
* `yr\_renovated` : The year the house was last renovated.
* `zipcode` : The zipcode of the location of the house.
* `lat` : The latitude of the location of the house.
* `long` : The longitude of the location of the house.
* `sqft\_living15` : The measurement in square feet of the house’s interior of the nearest 15 neighbors.
* `sqft\_lot15` : The measurement in square feet of the house’s land space of the nearest 15 neighbors.

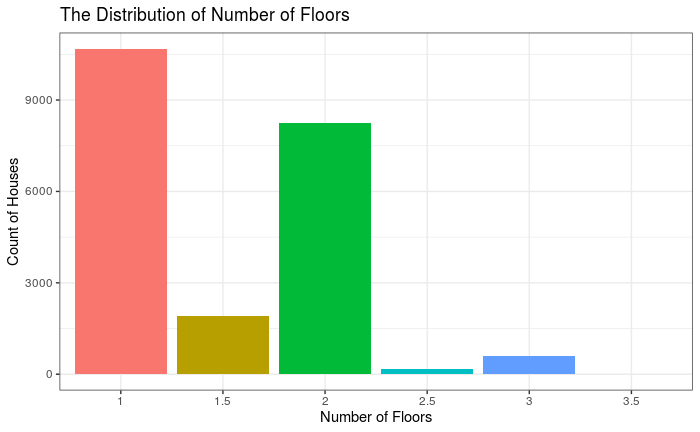
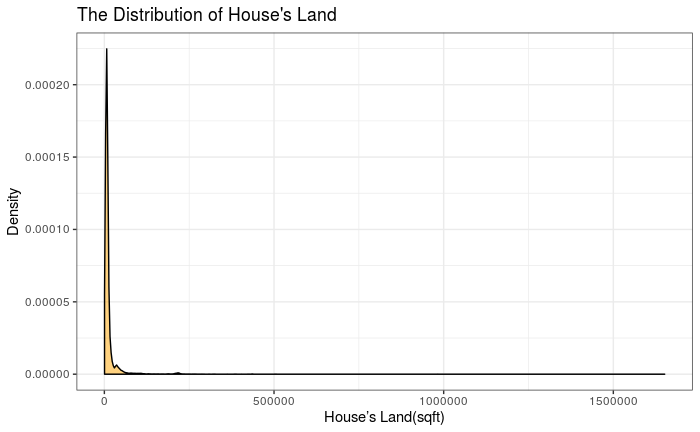
**Exploratory Data Analysis**

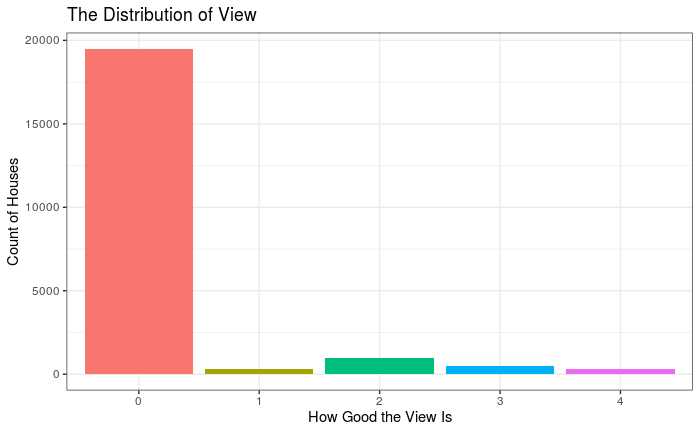
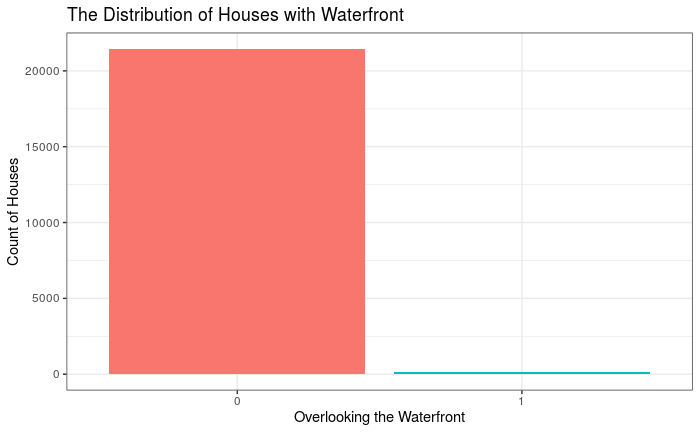
Univariate Exploration

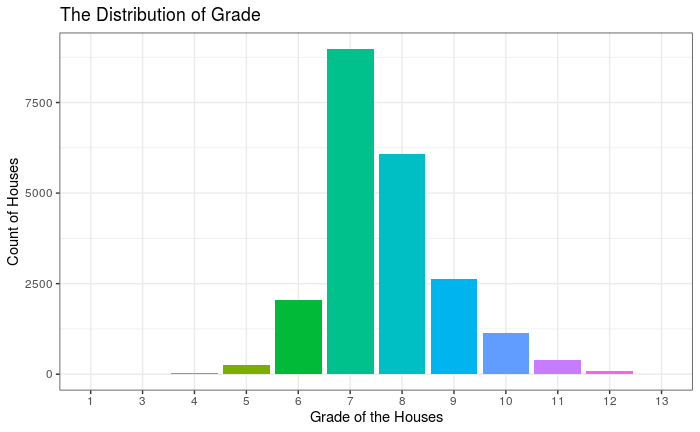
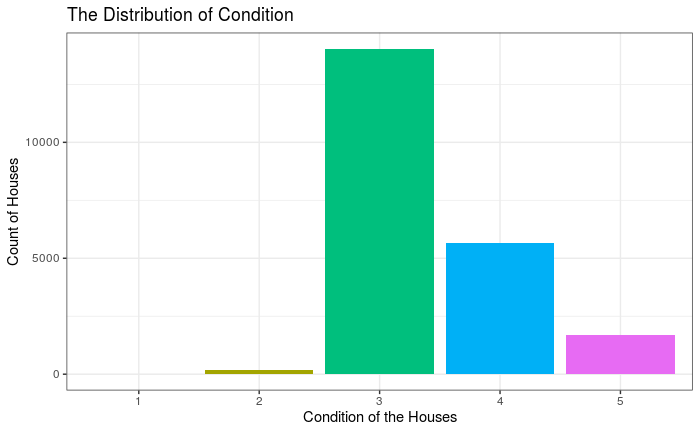
We first explore each variable individually, in order to determine what variables to include in our models. We will use density plots for numeric variables and bar plots for categorical variables.

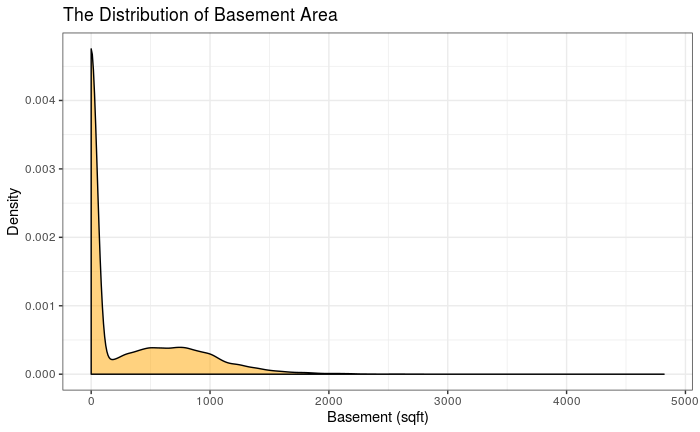
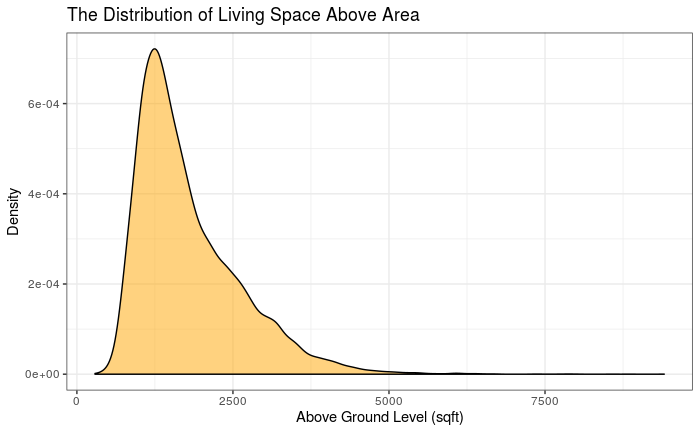


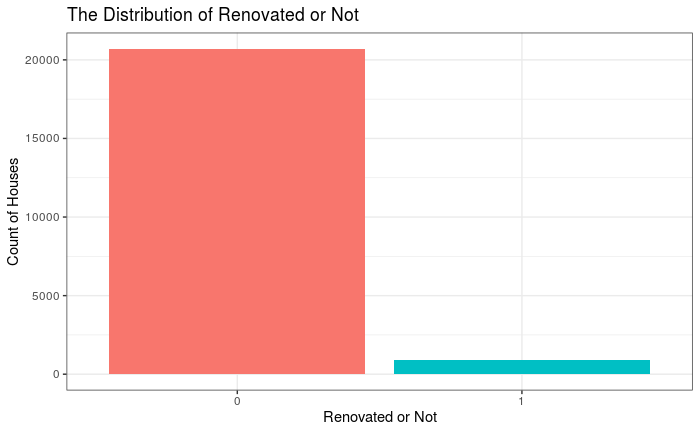
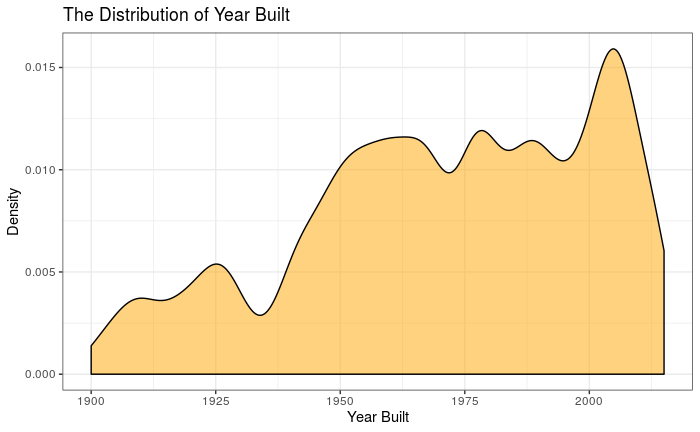


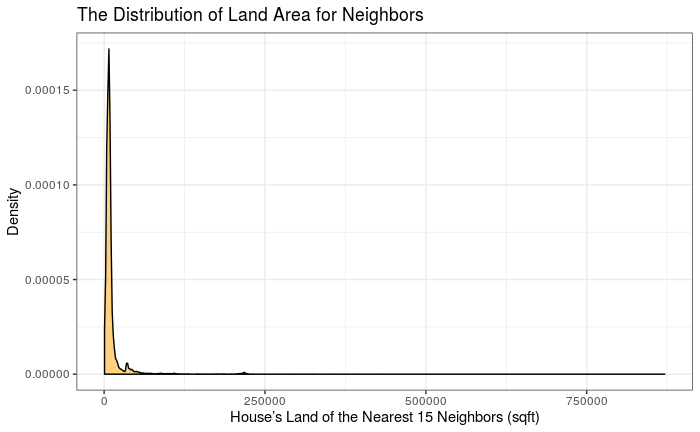
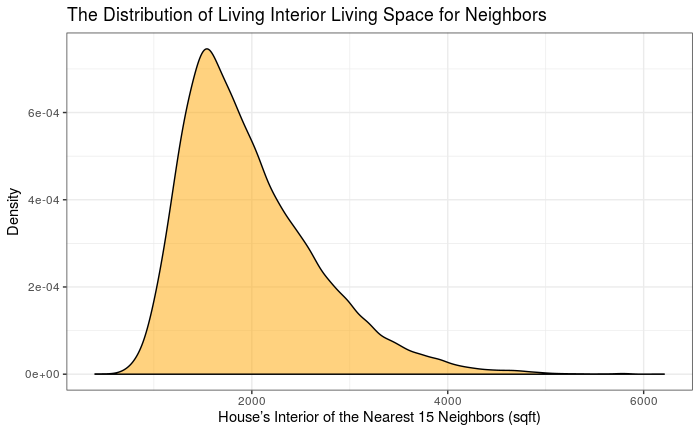






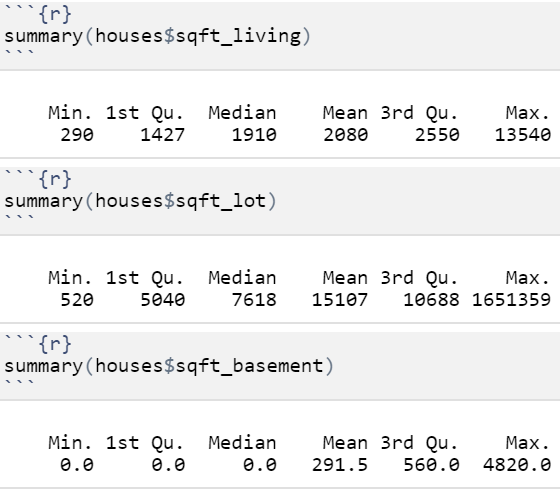
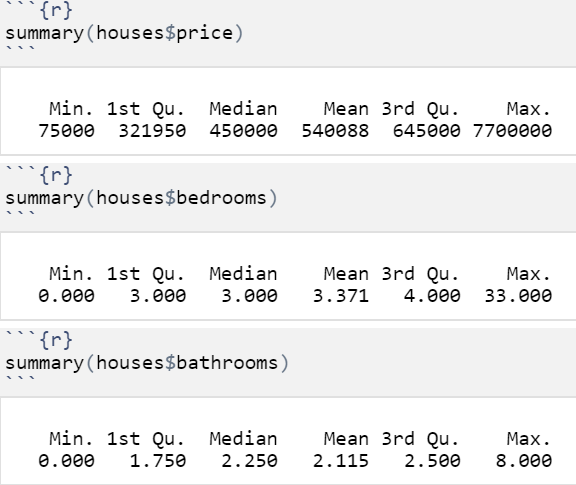


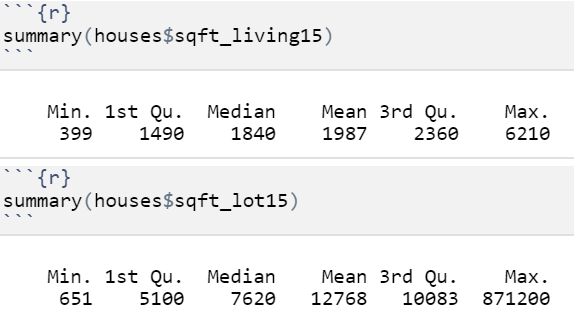




Figures: Distribution of Variables

For numeric variables, we supplement the graphical summary with numerical summaries.

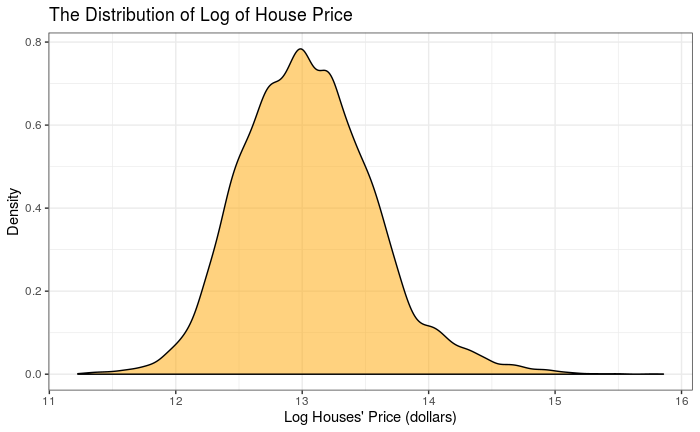




Tables: Numeric Summaries of variables

After looking at both the graphs and the summary statistics of our variables, we make the following observations: the distributions of most of our numeric variables are skewed to the right, the mean of these distributions being higher than their respective medians.

We notice that our response variable price is extremely skewed. Thus, a log transformation may normalize it. This transformation will be useful when we examine both price and log(price) in our linear regression models.



After doing a log transformation on house price, we can see that the distribution is almost normal.

Bivariate Exploration

First, we create a correlation plot between the numeric variables in the data set. In the graph below, red circles denote a positive correlation and blue circles denote a negative correlation. We can see that price and day, month, year, almost have no correlation. It has some strong correlation with sqft\_living15, sqft\_living, and sqft\_above. It also has some weak positive correlation with the rest of the numeric variables.

We also see that there seems to be some multicollinearity between some of the explanatory variables, such as sqft\_living and bathrooms, sqft\_living15 and sqft\_above, sqft\_above and bathrooms, sqft\_living15, and sqft\_living.

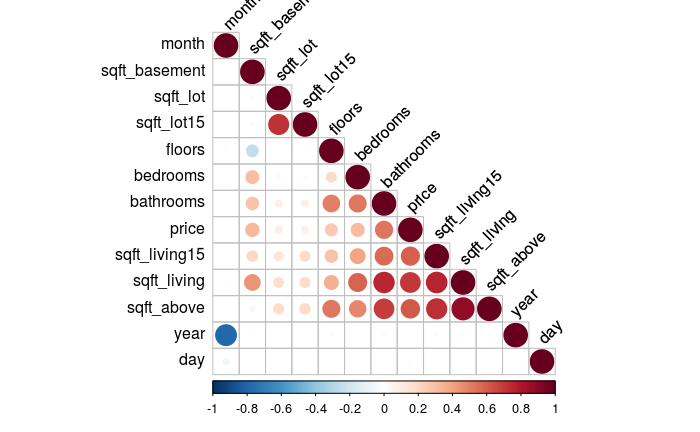
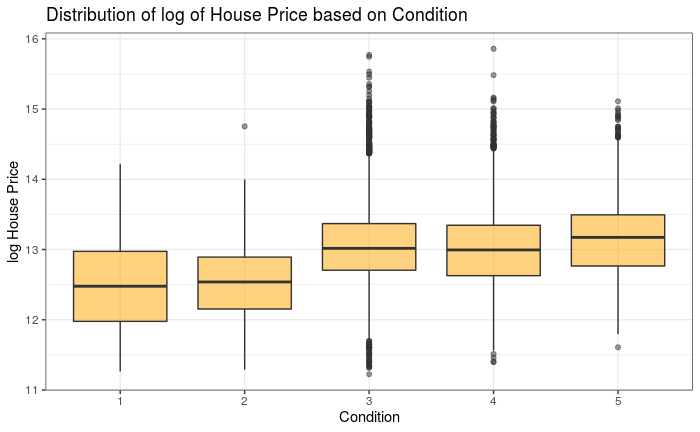
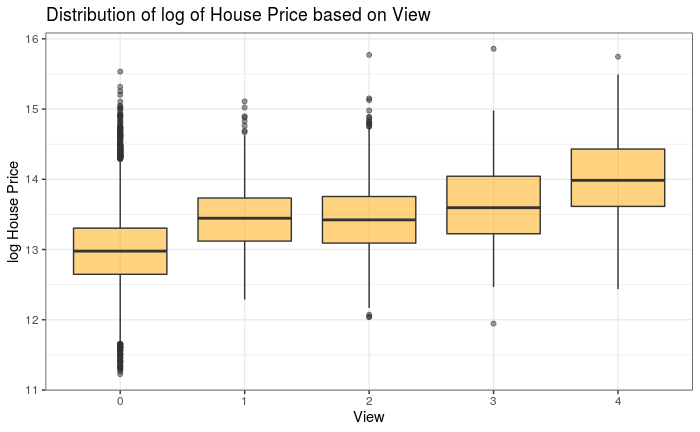
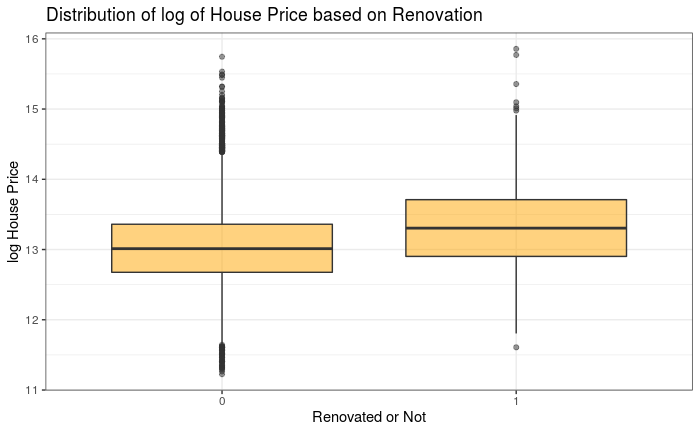
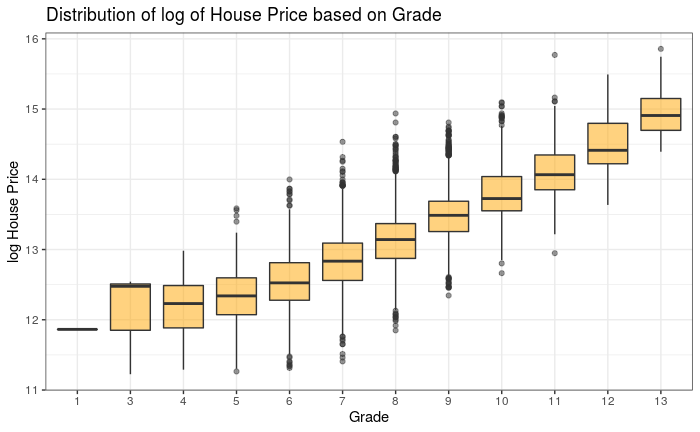


Figure: Correlation Plot for Quantitative Variables

In the next step, we observe how each categorical explanatory variable is associated with our response variable, the price of the house.





Figures: Boxplots for Categorical Variables to Log House Price

We can see that the distribution of the log(price) of the house is strongly related to waterfront, view, and grade and weakly related to renovation and condition. In addition, there are also many outliers for renovated houses and houses with a good condition.

Finally, for variables that indicates the location of the houses, we create a map to see the distribution of the houses using their longitude and latitude. We noticed that most of the houses are gathered together, while some of the houses are far away from most of the houses. Those far away house may have either higher prices or lower prices.

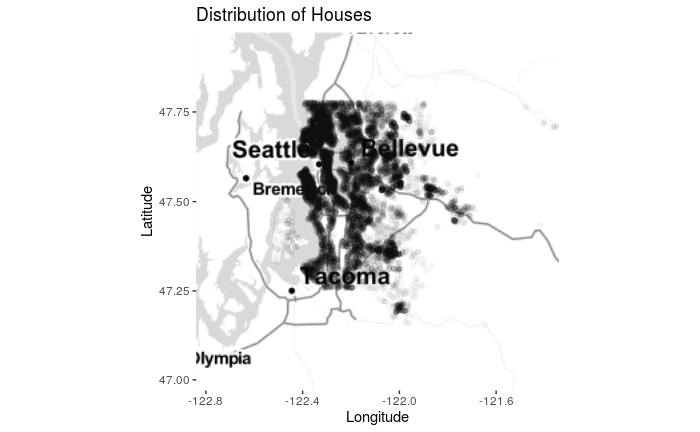


Figure: Distribution of Houses

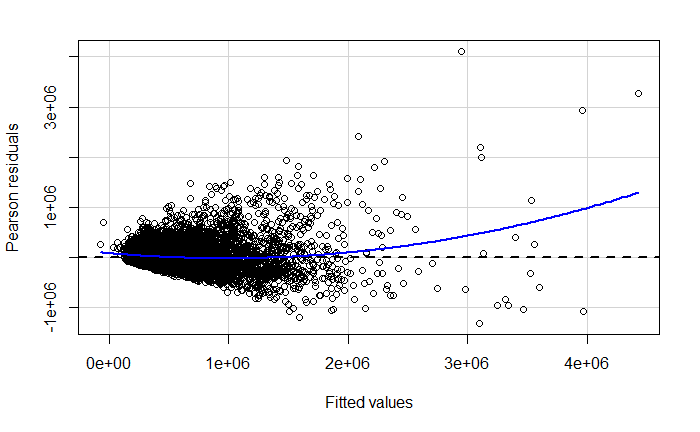
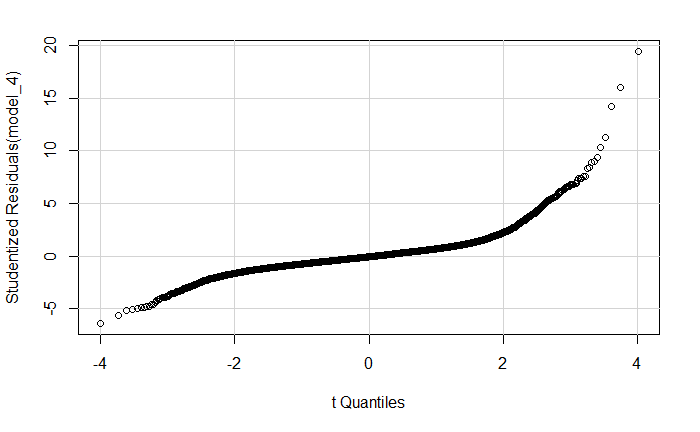
Through exploratory analysis, we choose to drop yr\_built, year, month, day because they are not suitable predictive variable for prices. Also, we drop zip code, lat, long because we cannot encode those geographic variables. Furthermore, including spatial information will make verifying the independence of our residuals in a linear regression more difficult.

**Modelling**

Linear Regression Models

Based on our EDA, we construct our first model by regressing price on bedrooms, bathrooms, living and lot size, number of floors, grade, condition, view rating, whether or not there is a waterfront, and the living and lot size of the 15 closest houses. We also thought about potential interactions in our model, but were unable to justify adding any meaningful interaction terms.

We first need to confirm whether our first model satisfies the conditions for a linear regression:

Figure: Residual Plot and QQ Plot for Model 1

Based on the residual plot and the quantile-quantile plot, we conclude our residuals are heteroskedastic and are not normally distributed, meaning that a linear regression will be inappropriate.

Therefore, we instead consider a second linear regression model on log(price) using the same explanatory variables. It is observable that the residuals are more homoskedastic (there is still some heteroskedasticity) and normally distributed.

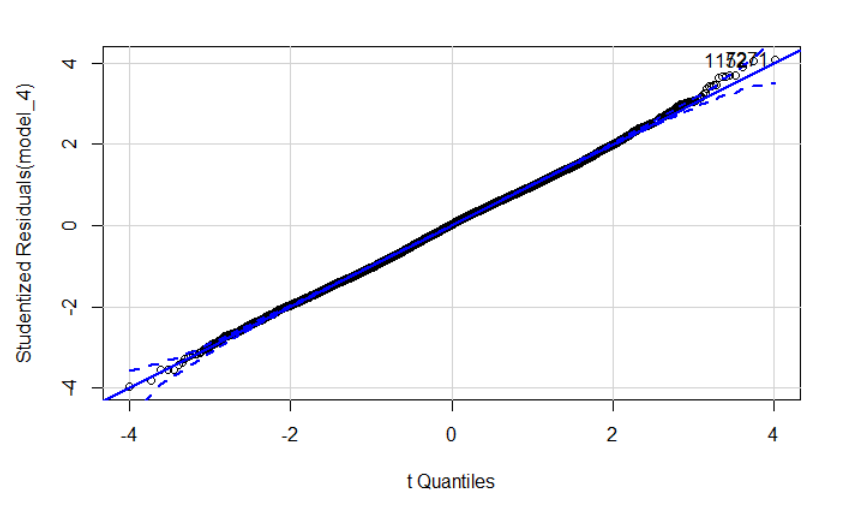
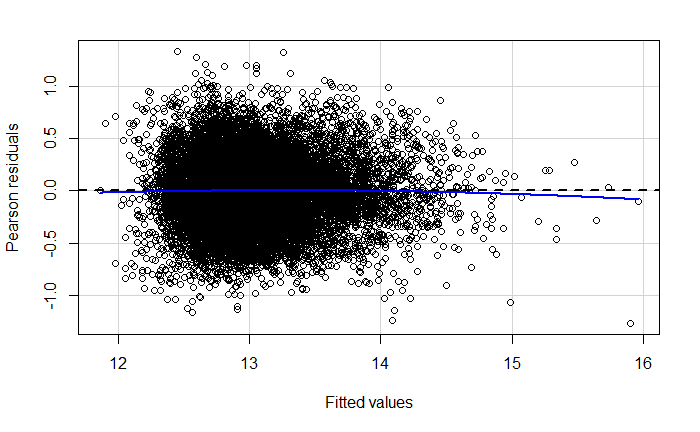


Figure: Residual Plot and QQ Plot For model 2

We proceed with our linear regression and present our results below:

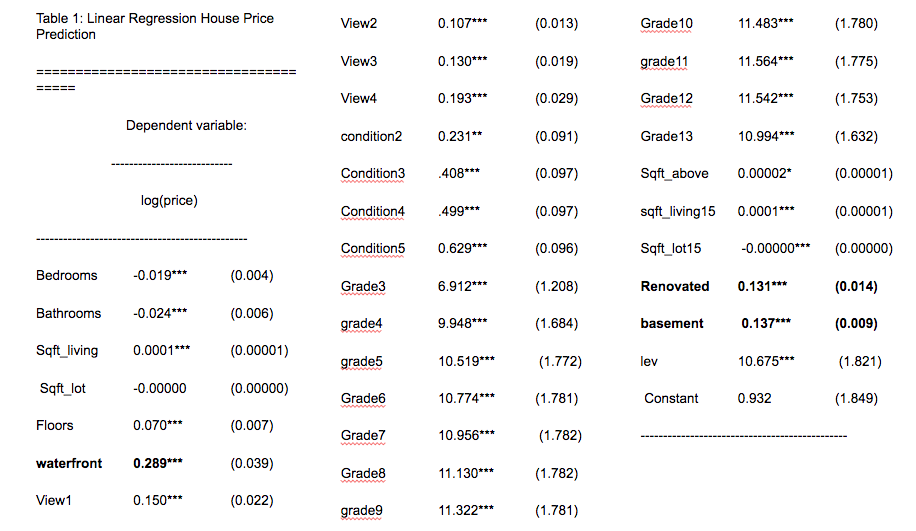


Table: Coefficients of Model 2

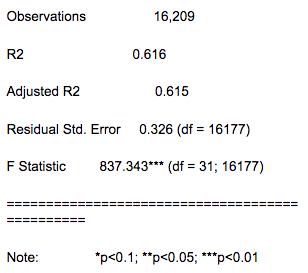


Table: Summary Statistics of Model2

When interpreting these coefficients, we need to keep in mind that we are predicting log(price), so each coefficient tells us a one unit increase for a given explanatory variable results in a 100% change in price. For example, our coefficient on whether the house has a waterfront is .289, which means a typical house, holding all other explanatory variables constant, would have a 28.9% increase in price if it had a waterfront. We also need to verify that there are not any extrapolations in our model; since it is possible to have a one unit increase in all of our explanatory variables and all factor levels are represented in the data, none of our interpretations are an extrapolation. Furthermore, most of our coefficients are statistically significant given a .05 p-value threshold. Lastly, our value is .616 which means our model explains 61.6% of the variation in the log(price) and our F-statistic is 837.343 which corresponds with a very low p-value which implies that our model is statistically significant.

After conducting some diagnostics on this model, we find that there are 1890 high leverage values out of 16209 observations in the dataset, which is approximately ⅛ of our observations. These observations have the potential to be problematic in our model.

This consideration motivates us to remove observations with high leverage values to generate a third linear regression model on log(price) using the same explanatory variables. We considered modifying this model slightly by removing observations with high Cook’s Distances; however, since we only found two such observations, we decided that removing these observations would not make a significant difference in our model. According to the residual plot and quantile-quantile plot, we notice that the residuals are reasonably homoscedastic (there is still some heteroskedasticity) and normally distributed.

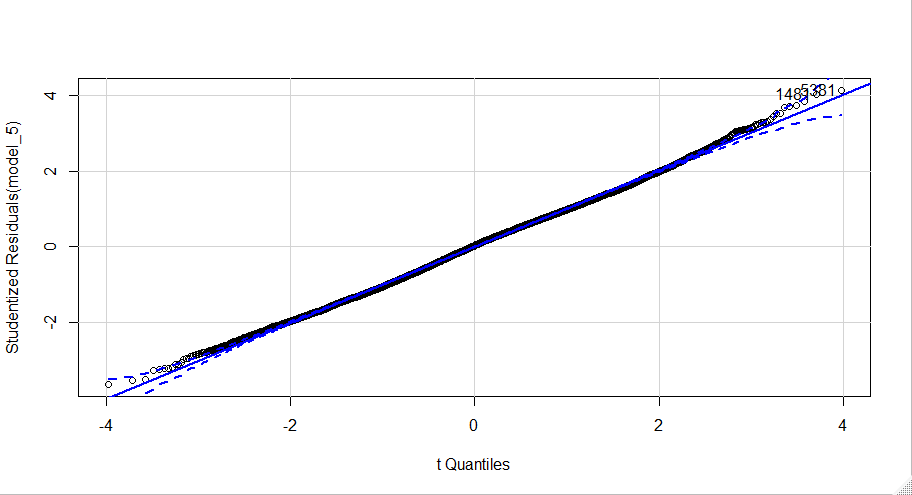
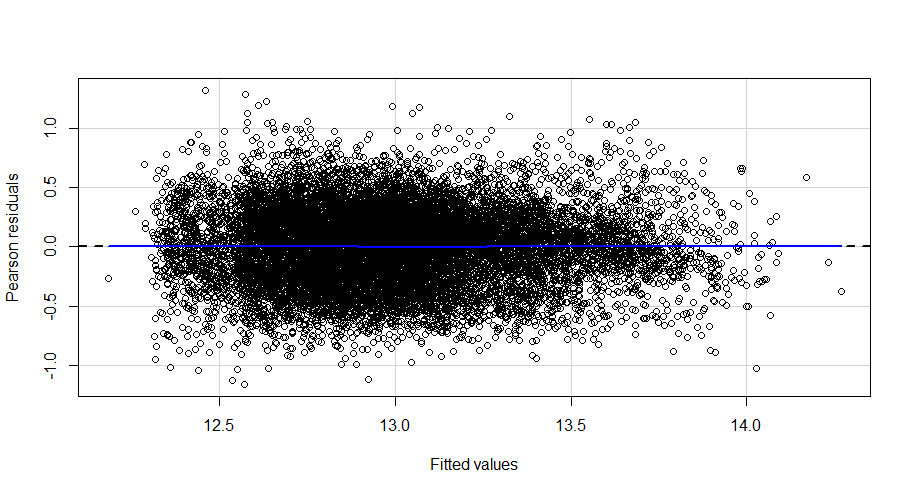


Figure: Residual Plot and QQ Plot for Model 3

In our third model, most of our coefficients are statistically significant given a .05 p-value threshold. In addition, our value is .502 and adjustedvalue is .501, meaning model 3 explains about half of the variation in the log(price). Although it is not fair to compare for model 2 and model 3, removing high leverage points may not have improved model 2. Regardless, the model 3 has an F-statistic of 848.070 which corresponds with a very low p-value, implying that our model is statistically significant.

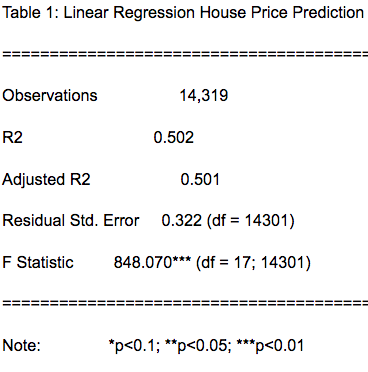


Table: Summary Statistics of Model3

XGboost and Random Forest

We use the Caret and Randomforest packages to build two machine learning models: XGboost and Random Forest. Both of the methods involve building many decision trees. XGboost builds one decision tree at a time. The new tree build will improve previous tree regarding errors. Thus, as more trees build, XGboost’s performance increases. Random forest builds a lot of trees at a time. However, the final prediction will be a weighted average of all the trees. Thus, random forest’s performance increase as more trees are created.

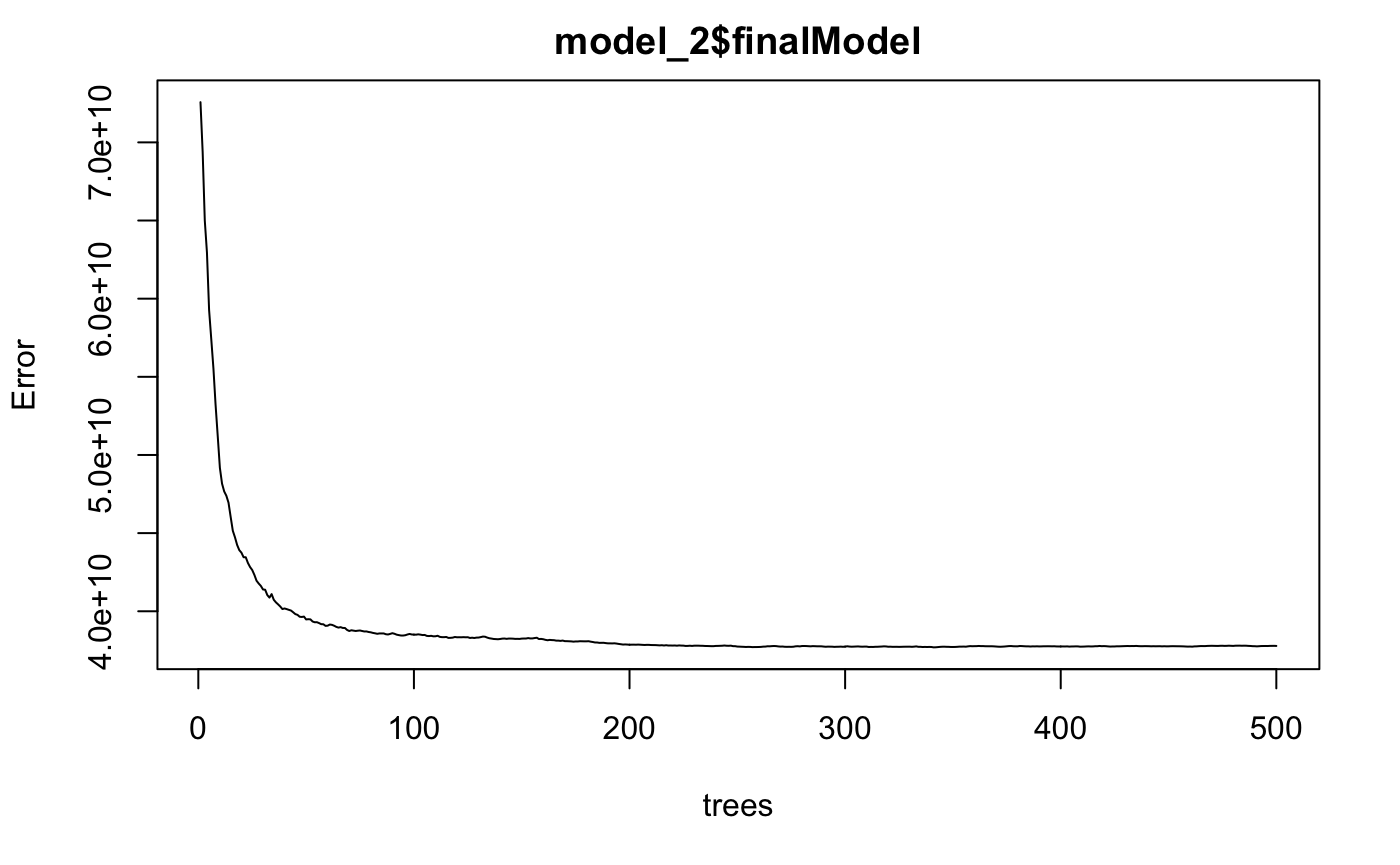
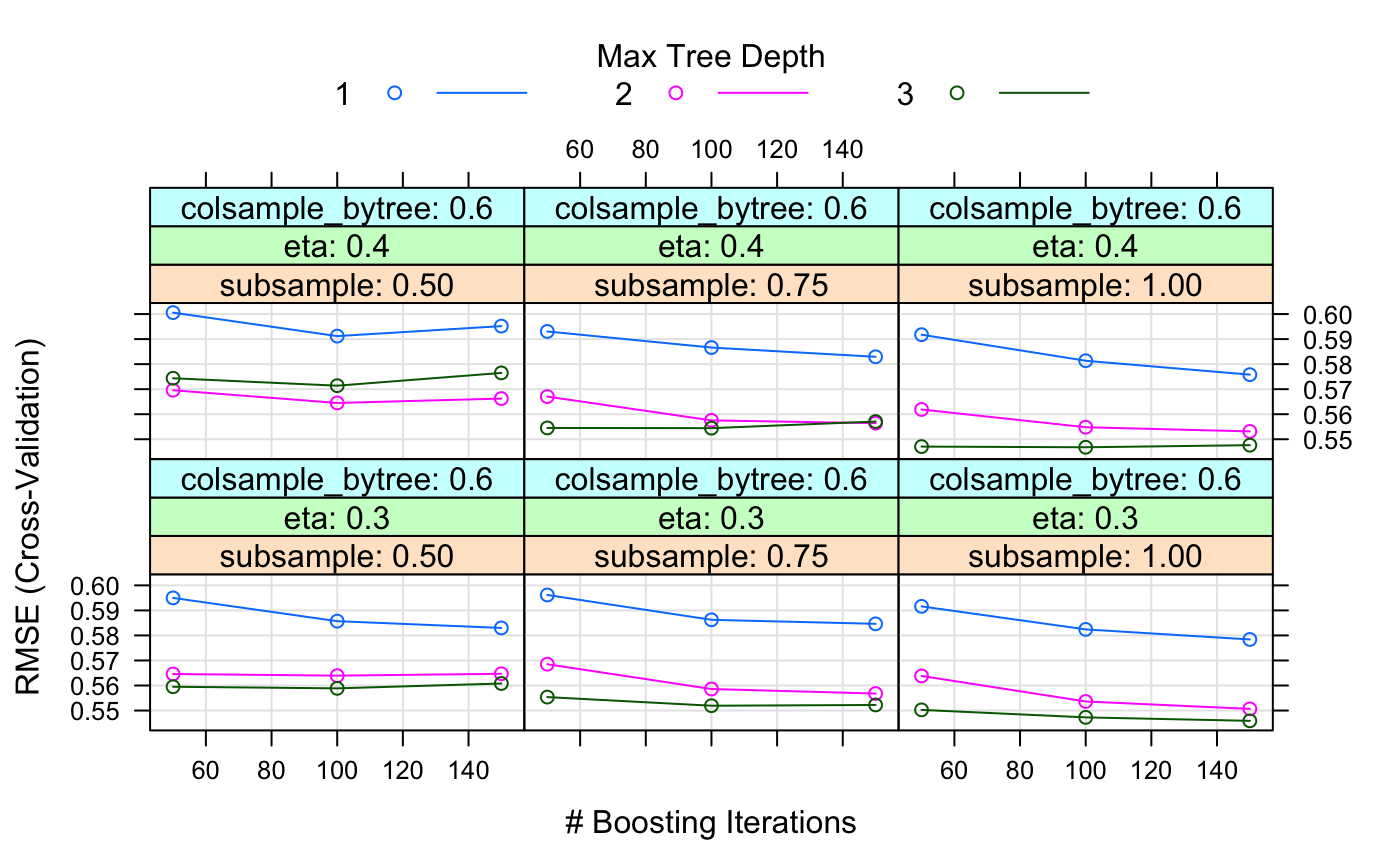
****

Figure: Change of Errors in XGboost(left) and Random Forest (right)

Neural Network

We used the neuralnet package for building a neural network with two hidden layers. To produce linear output, we did not employ activation function on the output layer. We chose resilient backpropagation with backtracking as the algorithm because it is fast and with high accuracy. However, R only allowed us the train the model with CPU. Thus, we decided to train the neural network one time.

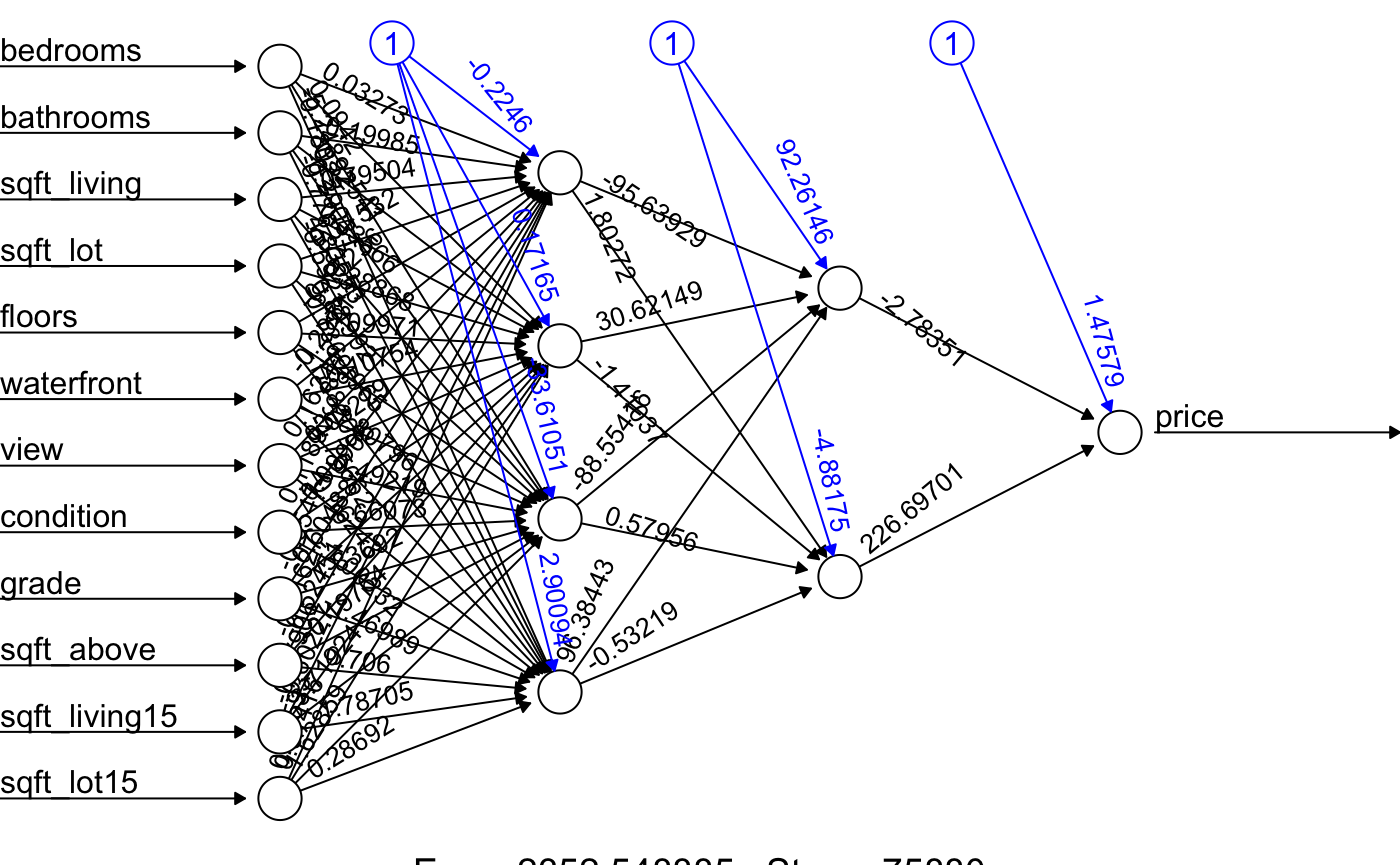
****

Figure: Neural Network Structure

**Model Evaluation**

We randomly split 75% of observations into a training set and another 25% in testing set. We also set seed so that we can get the same 75% of the training data for all the models.We use MAE and RMSE to evaluate the performance of our models. Mean Absolute Error (MAE) measures the average magnitude of the errors in a set of predictions, without considering their direction. It’s the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight. Root mean squared error (RMSE) is a quadratic scoring rule that also measures the average magnitude of the error. It’s the square root of the average of squared differences between prediction and actual observation. Compared to MAE, RMSE punishes larger errors.

Before we normalized the data, XG boost and random forest both have RMSE of about 190000 and MAE about 120000. Linear Regression Model with log price has RMSE of about 210000 and MAE about 140000, which are slightly worse than the accuracy of machine learning models. Unfortunately, the neural network does not converge with unscaled data. We conclude that, with unscaled data, machine learning models perform slightly better result that linear regression models.

After we normalized the data and do one hot encoding, we get different results.We find that XGboost and neural network perform the best. Random forest seem to be perform slightly worse. Linear Regression also perform worse. The change of result may due to one-hot encoding produce more variables, increase the likelihood of over fitting.

Table: Result of predictions with normalized data

|  |  |  |
| --- | --- | --- |
| Model | RMSE | MAE |
| XGboost | 0.52 | 0.34 |
| Random Forest | 1.29 | 0.832 |
| Linear Regression (log price) | 1.67 | 1.53 |
| Neural Network | 0.52 | 0.32 |

**Potential Weakness**

There are several potential weaknesses in our analysis. Firstly, we dropped the geographic variables, including zip code, latitude and longitude, which may be important to house prices because a house’s neighborhood quality directly affects the price of that house, Although the geographic variables we dropped might give us more insight into the determination of house prices, we were unable to find an appropriate way to encode them.

There were also some multicollinearity issues with our models, especially between the land and living space measurements of the house and the levels of the house. View may also be correlated with the size of the house as well as grade. In future models, it may be necessary to discard these variables.

We considered if any of our explanatory variables were endogenous i.e. they may be correlated with our error term. However, we struggled to find a convincing argument for why a particular variable would be endogenous. It’s possible that there is a degree of endogeneity in our model, and if so we might want to consider a TSLS model to decompose the exogenous and endogenous effects.

Lastly, another limitation, specifically for our machine learning models, was computational power. Machine learning models, especially our neural network, consume lots of power and take time. Unfortunately, our machines did not have the computational power and thus our models were not ideal.

**Conclusion**

Based on our linear regression models, we unsurprisingly find that having a basement, renovations, good condition, good structure and design, good view, and having a waterfront, are positively correlated to house prices. Surprisingly, the result shows that the size of the house, number of bedrooms and the number of bathrooms have little impact on house prices.

We also noticed that more complicated models, like machine learning models and neural networks, do not always outperform the simpler linear regression models. Machine learning models require large amounts of time with many iterations in order to reach their full potential which we were unable to exploit given our computational limitations.

Our group is sufficiently satisfied with the final linear model. We believe that this model is able to explain a reasonable amount of the variation in prices of houses in King County. We should remember that all our observations are about houses that were **sold**, so for a realtor interested in selling a house, this model allows for a better understanding of the characteristics of a high value house that sells.

**Future Research**

In the future, it may help to include more variables like the pricing of nearby houses. The housing market is also correlated with the general trends of the economy so King County economic data would be interesting to include. Additional information about the characteristics of the buyer would help our hypothetical realtor out too.

Lastly, our project only focuses on correlation rather than causality. Instrumental variables may be added to test the causality between other predictive variables price.