**ALY 6015 INTERMEDIATE ANALYTICS**

House Prices Prediction in King County, WA, USA

Final Project Report

ALY 6015, Sec 02 Fall 2018

(October 27, 2018)

Submitted By:

Venkata Kaushik Burra, Divya Gupta, Zeyu Hu, Sijiang Liu

Submitted To:

Ani Aghababyan

Northeastern University, Boston, MA

Master of Professional Studies in Analytics

### **Table of Contents**

Introduction

Research Questions

Data Source

Data Cleaning

Methods

Analysis Results

Conclusion

Appendices

References

### 

### 

**Introduction**

House prices are a big deal in today’s life. House prices dropped in King County for the first time since the recession, as competition among buyers weakened and the number of homes for sale surged. In this project we have used a real-world dataset to understand the practical inferences. To carry out these tasks we have chosen a dataset for predicting house prices in King County. The dataset focuses on how the price affects the features. To understand the house prices and how they vary with various features we are using predictive analytics. Predictive Analytics is widely used in real life. When the outputs are continuous numerical, we can use regression model. In our project, we are trying to use multiple regression models to predict the house price based on the features.

### **Research Questions**

1. Which are the house features that affect the prices the most?
2. Is it reliable to use regression models to predict the house prices?

### **Data Source**

This dataset is retrieved from Kaggle, [House Sales](https://www.kaggle.com/harlfoxem/housesalesprediction) . The dataset covers the house sales in King County, Washington State, USA.

It consists of 19 house features which are the independent variables, the dependent variable “Price “and the ID columns, along with 21613 observations.

### **Data Cleaning**

This dataset required slight cleaning. We excluded null values during our analyses.

Additionally, we filtered the ID and the date columns to work with our analysis.

### **Methods**

We began by exploring the dataset, then we performed Linear Regression and Random Forest Regression models between the Price and the significant features, and finally plotted graphs to understand the distribution of the dataset. After performing the Regression models, we began to explore further to understand the relationship between the features.

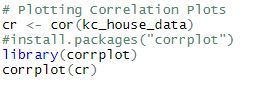
All the analysis in our project has been performed using R Language.

**Analysis Results**

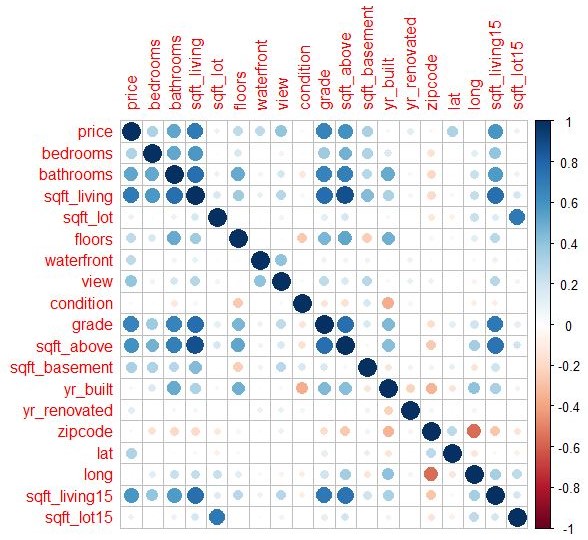
From our analysis, we discovered that the Price and the Sqft\_Living are highly correlated.

We also determined that there exists a positive correlation between the independent and the dependent variables.

Input code:



Graph 1:



As we can see from the above graph, the independent variables sqft\_living, grade, bedrooms and bathrooms have higher correlation to the dependent variable which is the price.

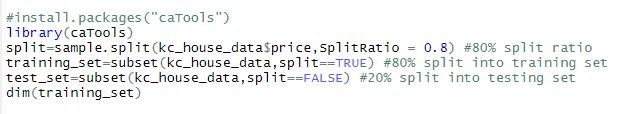
Our analyses result clearly showed that the house sales in King County depends on the sqft\_living, grade, bedrooms and bathrooms.

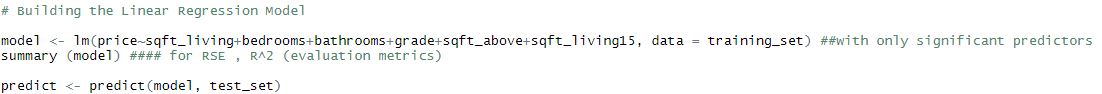
Visit Appendix 1 for more graphs. (heat map)

Further, we performed Linear Regression model analysis to predict how the prices vary with the significant features like sqft\_living, grade, bedrooms and bathrooms.

To begin with our analysis, first we fit the model by splitting the dataset into Training set and Test set respectively where we cover 80% of Training set and 20% of Test set. Then we build the model taking the significant features and predict the prices.

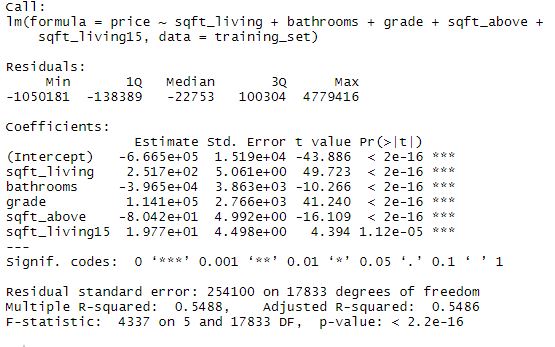
Input code:







Output:

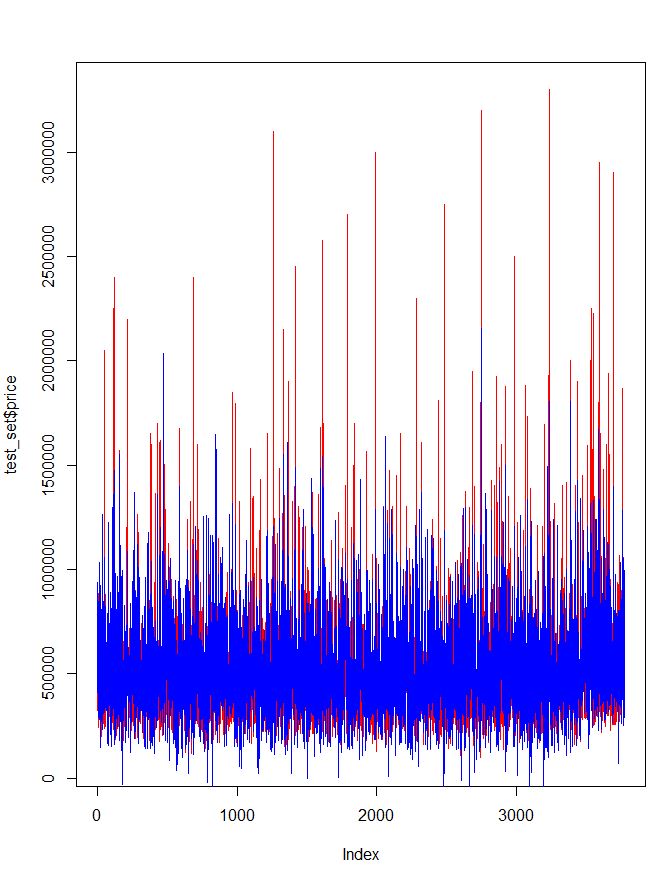


From the above output, the coefficients, Multiple R squared and Adjusted R score values evaluate the performance of the model.

We can see that the sqft\_living and the grad are having positive coefficients with the intercept which is a good sign and also explains why sqft\_living has the highest positive correlation with the price.

The Multiple R- squared and Adjusted R squared values are almost similar and have a decent value of 0.54.

Graph 2:



From the above graph, the Y-axis represents the price for the test set and the X-axis represents all the significant features index values.

We can see that; the red lines are the actual values of the price and the blue lines are the predicted values after modeling’s clearly the predictions decrease with the increase in price that shows most of the house sales (price) are between 50,000 and 100,000 with respect to all the significant features.

Additionally, we have shown useful visualizations while working with linear regression. The ‘lmplot’ is used for the four by four visualizations to check the correlation between the actual values and the predicted values. (below are the graph output).

Upper left plot: This plot shows how the actual values and predicted values are correlated.

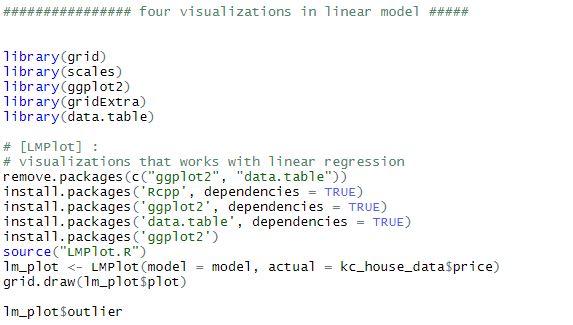
Upper right plot: This plot shows how the cook’s distance measures the accuracy for the predicted values.

Lower left plot: This plot shows how the residuals are symmetrical with the linear model predicted values.

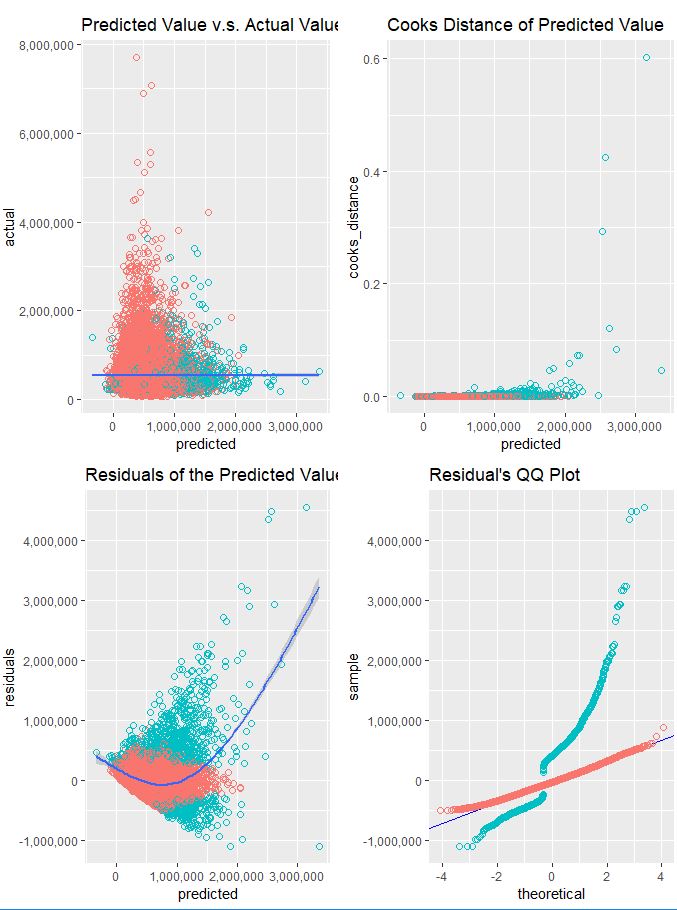
Lower right plot: This is the qq plot which shows how the residuals are aligned towards the predicted values.

From our analyses, the plots show significant amount of correlation for the linear model regression.

Input Code:

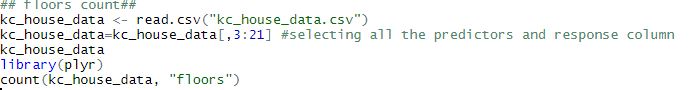


Graph 3:

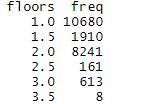


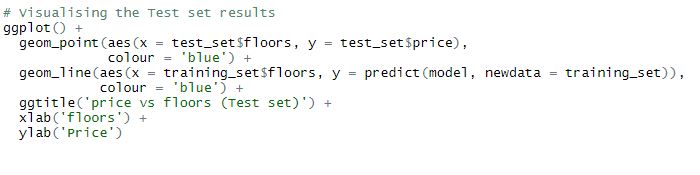
Further, we also analyzed how the number of floors for the house affect the house price with respect to the sqft\_living.

Input code:

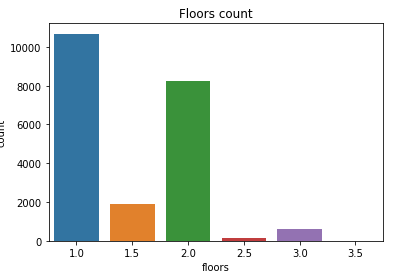


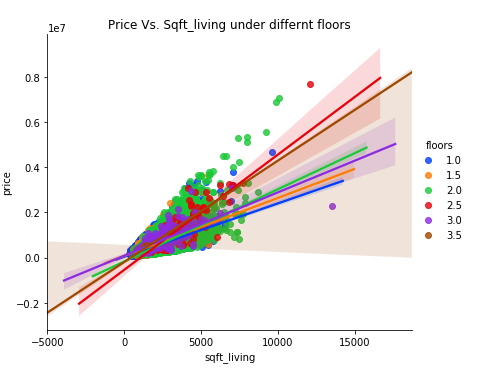
Output:





Graph 4:





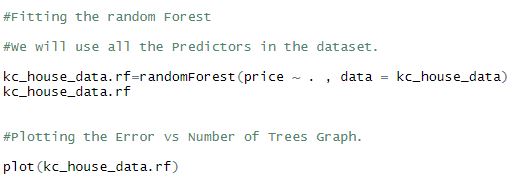
As we found out from the correlation graph, there is positive relationship between price and sqft\_living. The above graphs also present the floors which are the trend lines with different colors, where each color represents the count of floors the house consists.

The trend line is the price divided by sqft\_living, which, in other words, is the unit price of the house. We can see that house with 3 and 3.5 floors has the largest trend which means they are having the highest unit price.

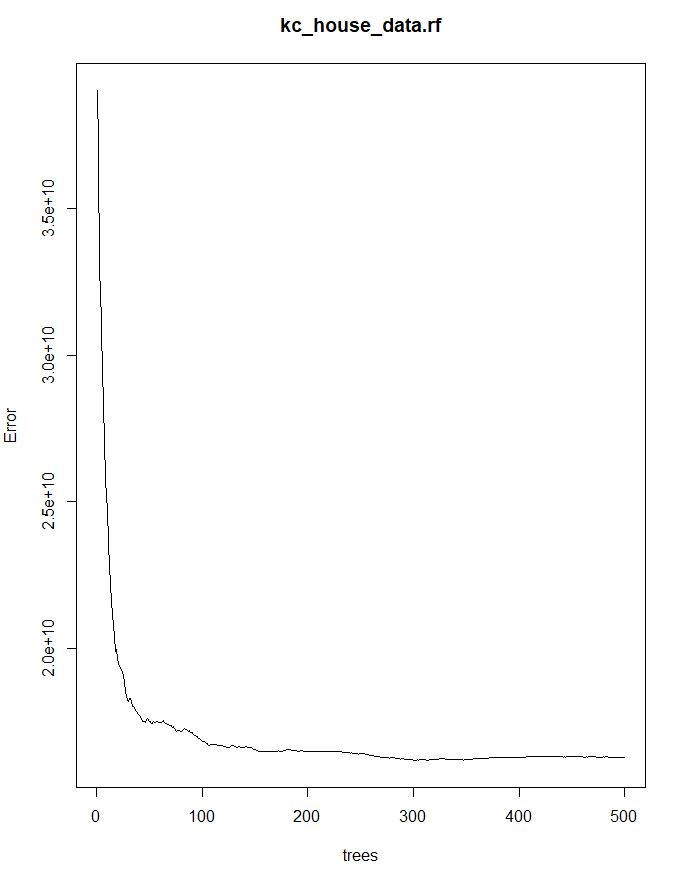
Random forest Regression:

Random Forest Regression is used to validate the variance and overfitting of the independent variables.





Graph 5:



In Random forest regression, the number of trees and the error are inversely related.

From the above graph, we see that the error decreases as the number of trees are increased.

The number of trees with a value of 500 is the ideal fit for our model.

Additional Random Forest analysis – Feature Importance (using Python) in Appendix 3 .

### 

### **Conclusion**

In this report, we demonstrated the correlation between the dependent variable “Price” and the independent variables “sqft\_living, grade, bedrooms, bathrooms, etc.”.

In fact, there exists a high positive correlation between the price and the sqft\_living. Furthermore, we performed the Linear Regression model and the Random Forest Model to predict the house price with respect to the significant features and found out the correlation with the linear graphs and scatter plot respectively.

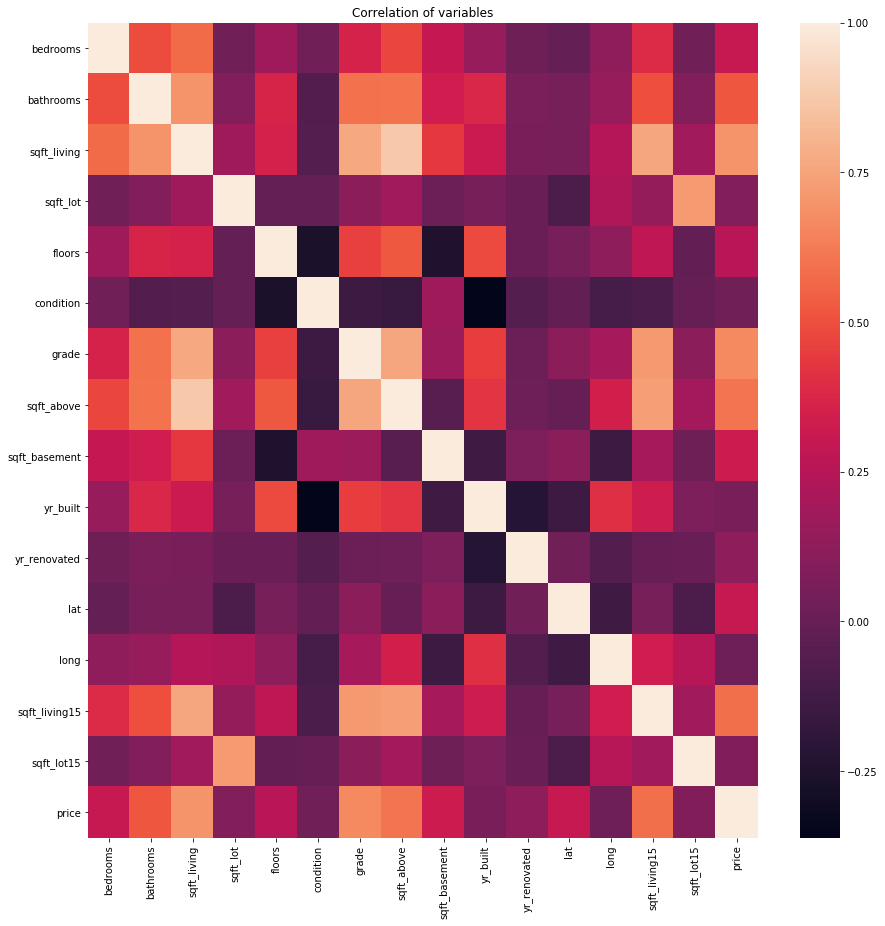
In addition, we also found out the correlation between the floors, price and the sqft\_living.

These set of analyses result in variation of house prices in Kings County, Washington State, USA.

**Appendices**

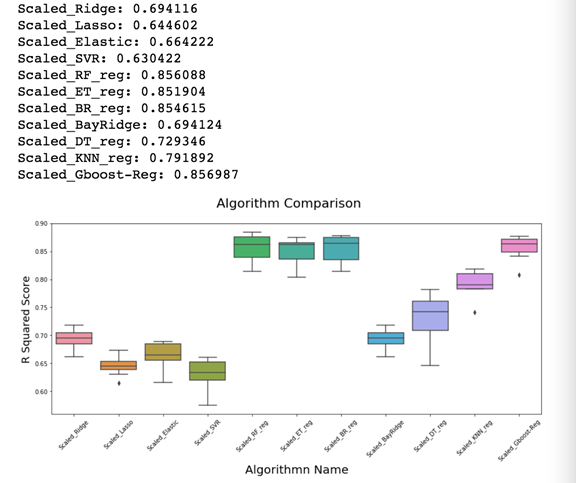
Appendix 1

HEAT MAP (using Python) for the correlation between the Dependent and the Independent variables.



Appendix 2

Python code: R Square Regression Score



Before we choose the model, we need to test which algorithm is the best for predicting the house price.

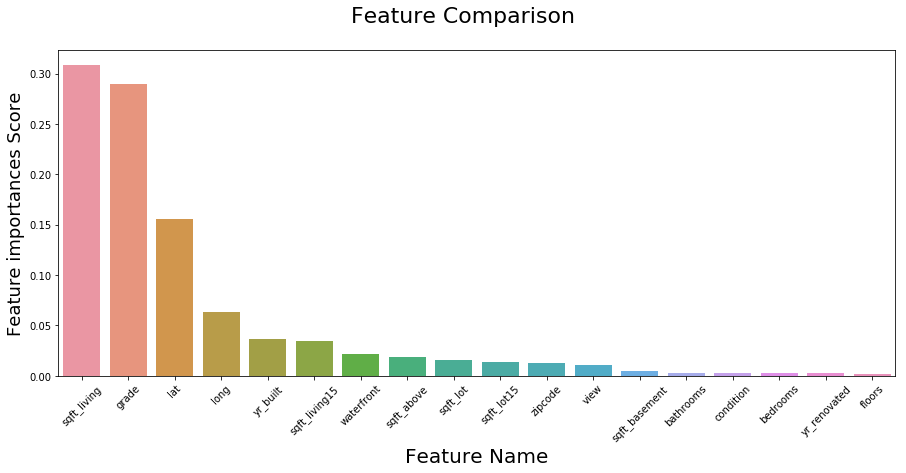
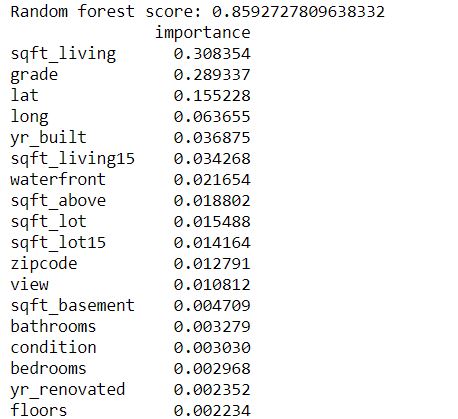
We know that a prediction model which uses the same data set to train and test is a mistake. If the model repeat using same data set will have a perfect score but would fail to predict useful on the future unknown data. We cannot let the model too overfitting. So, we used **train\_test\_split** function to hold out part of data as the test set.

To test which algorithm is best, we will use the k-fold cross validation method to test the accuracy of each algorithm. The test set will still be held out for the final evaluation. But the training set will split into 10 smaller sets. Each algorithm will use 9 of folds as training data. And each algorithm will validate on the remaining part of the training set. The performance score by k-fold cross validation is the average score of each algorithm computed in the loop. Form the chart, we can see that Random Forest, Extra Tree, Bagging, and GBDT regression are outstanding algorithms. The r squared score of them are higher than 80% which means those are useful algorithms for the model to predict the house price. We can also see that GBDT has more stable performance from the box plot, since it has smaller variance.

Appendix 3

Feature Importance from Random Forest Regressor (Python code)





#### **References**

1. Kaggle, Shiva chandel: <https://www.kaggle.com/shivachandel/kc-house-data>
2. “Random Forests in R.” *R-Bloggers*, 24 July 2017, [www.r-bloggers.com/random-forests-in-r/](http://www.r-bloggers.com/random-forests-in-r/).
3. Shekhar, Prashant. “How to Apply Linear Regression in R.” *DataScience+*, 21 Dec. 2017, <https://datascienceplus.com/how-to-apply-linear-regression-in-r/>