**PREDICTING HOUSE SALE PRICES IN KING COUNTY, USA**

**Jimit Shah (**[**jimshah@iu.edu**](mailto:jimshah@iu.edu)**) and Ritesh Tawde (**[**rtawde@iu.edu**](mailto:rtawde@iu.edu)**)**

**ABSTRACT**

Our project analyzes the dataset containing house sale prices for King County, USA. We have implemented prediction techniques like regression models- viz. simple linear regression and multiple regression- and Random Forest ensemble method. To apply these techniques, we have done a systematic analysis of the dataset through data exploration and visualization techniques.

**INTRODUCTION**

The data set we are using contains information on the house sale prices of King County, USA from May 2014 to May 2015. It is a great dataset for applying simple regression models. It is primarily used to mine a pattern and predict the prices of the houses sold, by analyzing various features provided in the data set. However, upon carefully analyzing and studying the dataset, we observed some features in the dataset are uncorrelated with the price of the house. Next section on data exploration and visualization provides an insight of how we analyzed the dataset

**DATA EXPLORATION AND VISUALIZATION**

We performed data exploration to find out if the dataset is clean i.e. whether it contains values like NA/NaN/Null, which features are categorical / continuous, which features are highly correlated / uncorrelated with price, etc. Through the experiments performed on the dataset, we observed the following:

* We didn’t find any dirty data i.e. the data doesn’t contain any Null/ NA/ NaN values. Therefore, we can conclude that there were no missing values.
* The “date” column’s scatter plot against house prices is shown below. We can understand that the price is independent of the timeline given in the dataset i.e. from May, 2014 to May, 2015. Therefore, for further analysis we did not take date feature into consideration.

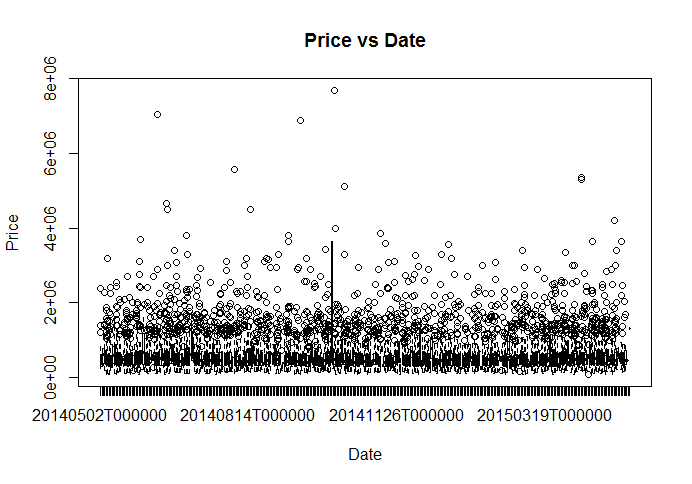
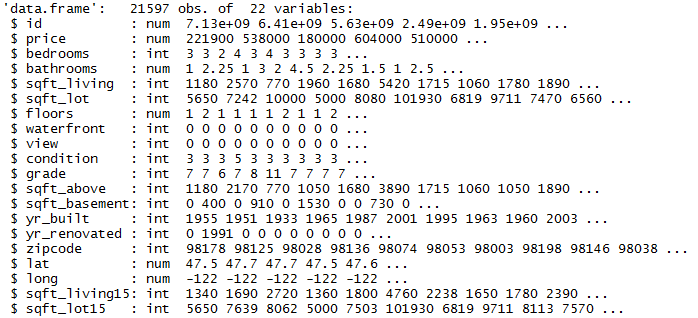


Figure : Date vs. Price plot

* Following table provides the description of each column of the data set in terms of its data types and values it stores.



On viewing the table, we can deduce the type of each feature. The following features are categorical: bedrooms, bathrooms, floors, waterfront, view, condition, grade, zip code, and id. The rest of the features are continuous.

* Let us look at the plots of the continuous variables against price

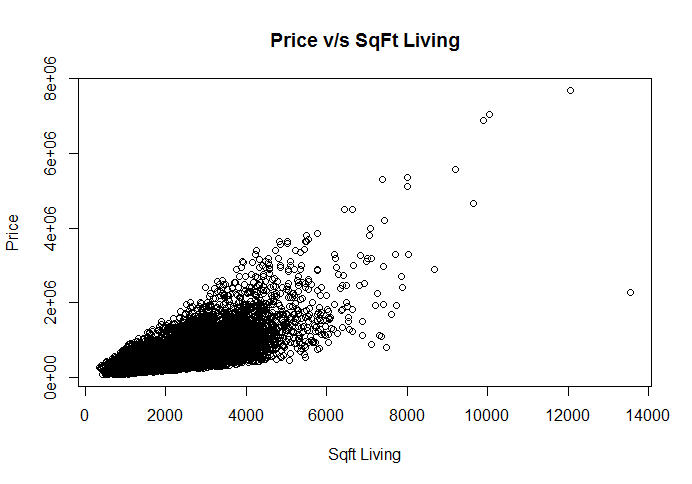
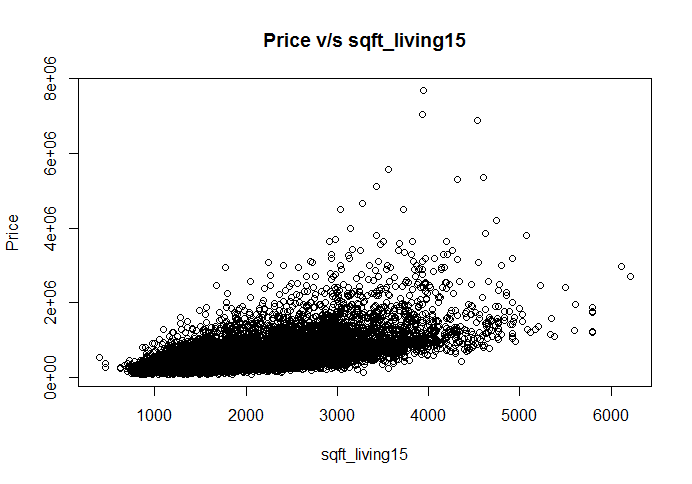


Figure : Price vs. Sqft\_Living



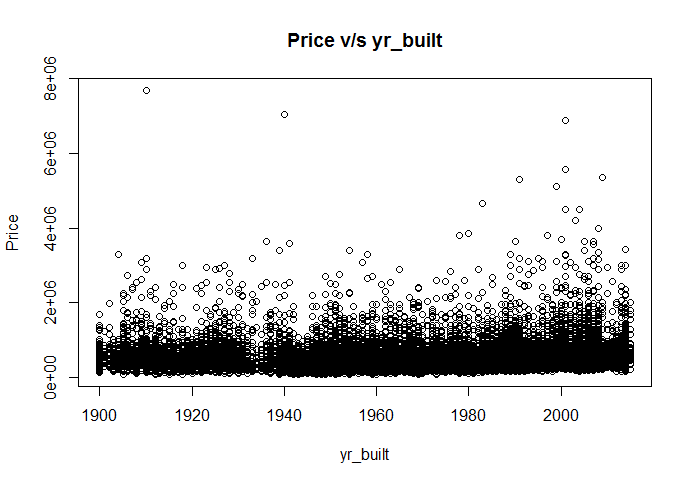


Figure 4: Price vs. Yr\_built

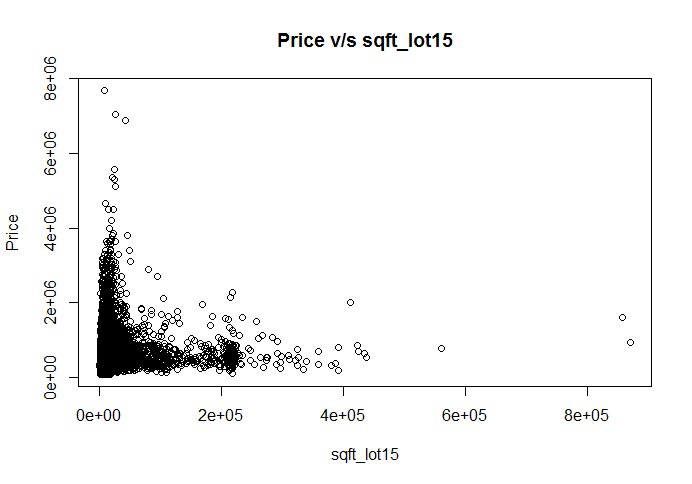


Figure 3: Price vs. Sqft\_Lot15

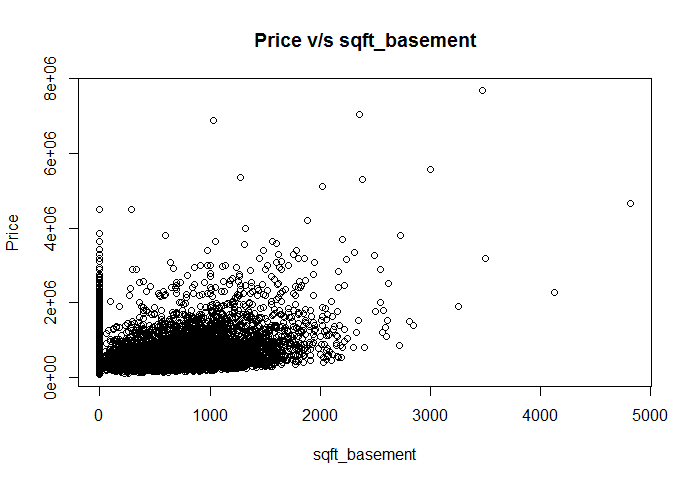


Figure 6 : Price vs. Sqft\_Basement

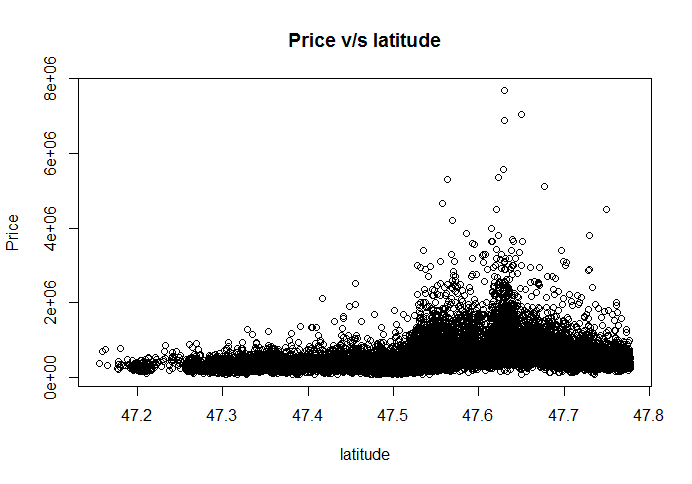


Figure 5: Price vs. Latitude

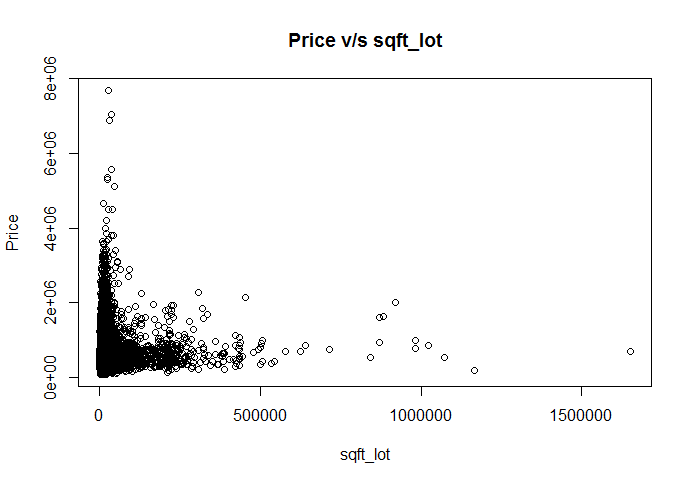


Figure 8: Price vs. Sqft\_Lot

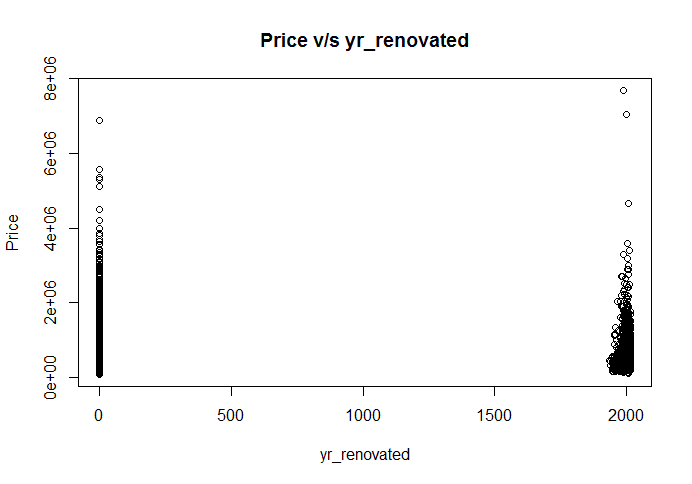


Figure 7: Price vs. Yr\_renovated

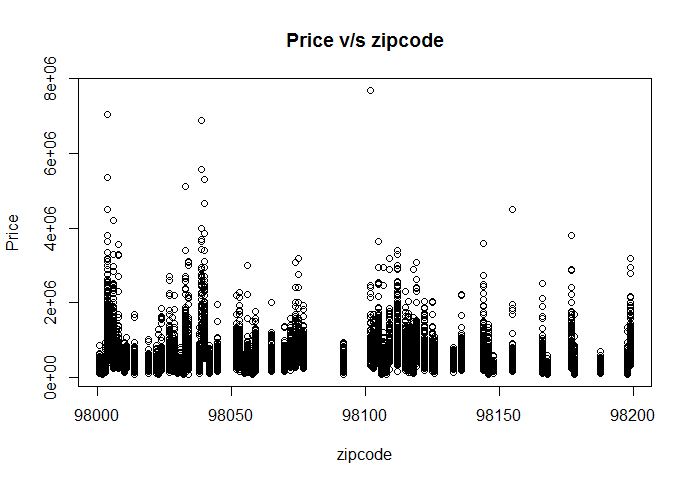


Figure 10: Price vs. Zipcode

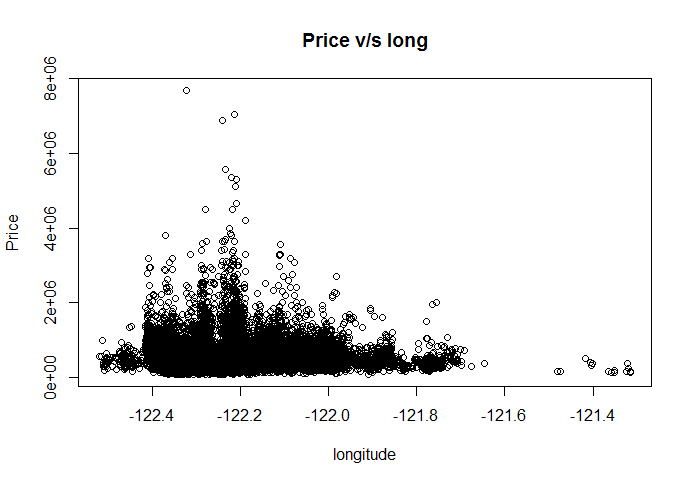


Figure 9: Price vs. Longitude

As we can observce from the plots, sqft\_living feature is highly correlated with price. Other variables like sqft\_above, sqft\_basement and sqft\_living15 are also nicely correlated to price. Features like zipcode, long, sqft\_lot15 and yr\_built are poorly correlated to price. The following plot depicts correlation between each feature.

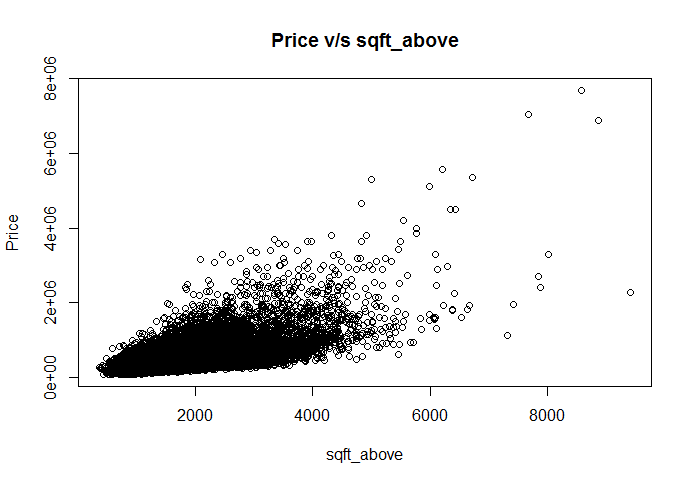


Figure : Price vs. Sqft\_above

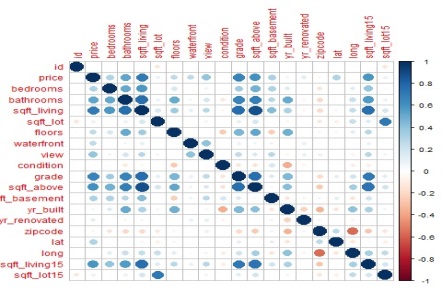
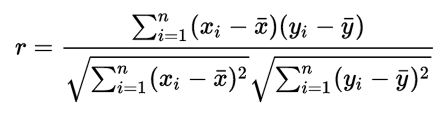


Figure : Correlation plot

To find correlation using a statistical tool, we need to apply the following formula:

* To calculate correlation between two continuous attributes, we use Pearson’s correlation formula. It is given as,



where n is total sample space, x and y are continuous attributes.

Using the above formula, we have calculated correlation of each continuous attribute with price to confirm our analysis we did using data visualization

Correlation between price and sqft\_living = 0.7019173

Correlation between price and sqft\_lot = 0.08987622

Correlation between price and sqft\_above = 0.6053679

Correlation between price and sqft\_basement = 0.3237989

Correlation between price and yr\_built = 0.05395333

Correlation between price and yr\_renovated = 0.1264236

Correlation between price and zipcode = -0.05340243

Correlation between price and lat = 0.3066923

Correlation between price and long = 0.02203632

Correlation between price and sqft\_living15 = 0.5852412

Correlation between price and sqft\_lot15 = 0.08284493

* Observe that features sqft\_living15 and sqft\_living are also highly correlated. We perform partial correlation test to find out the correlation of sqft\_living15 with price controlling the variable sqft\_living. On computing the partial correlation, we observed the partial correlation coefficient between sqft\_living15 and price controlling sqft\_living is 0.116. Thus, we can conclude that sqft\_living15 is not correlated to price given sqft\_living.
* Also, we observed that yr\_renovated and sqft\_basement features contain zeros for houses that either do not contain a basement or has not been renovated. Thus, it will be interesting to find out whether existence of basement or whether the house has been renovated affects the price. For this reason, we created two new binary features namely has\_basement and is\_renovated that denotes whether a house has a basement or not or whether it has been renovated or not respectively.
* Let us now look at the boxplots of categorical variables against price

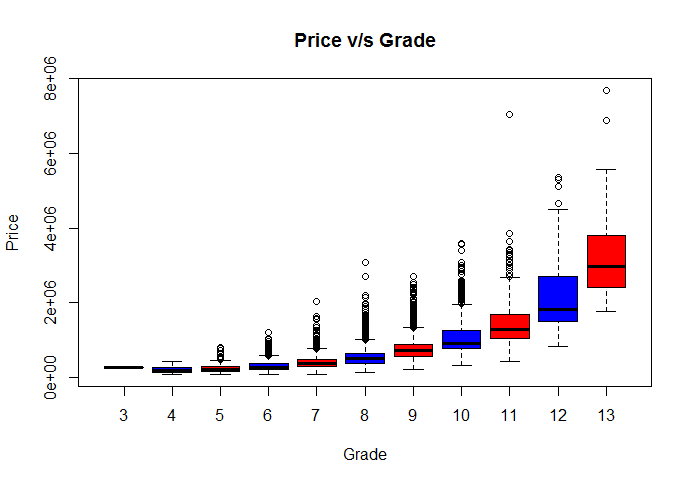


Figure : Boxplot - Price vs. Grade

``

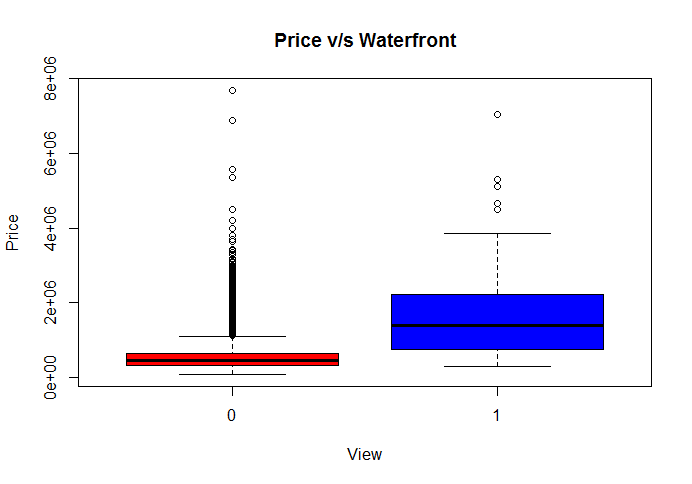


Figure : Boxplot - Price vs. Waterfront

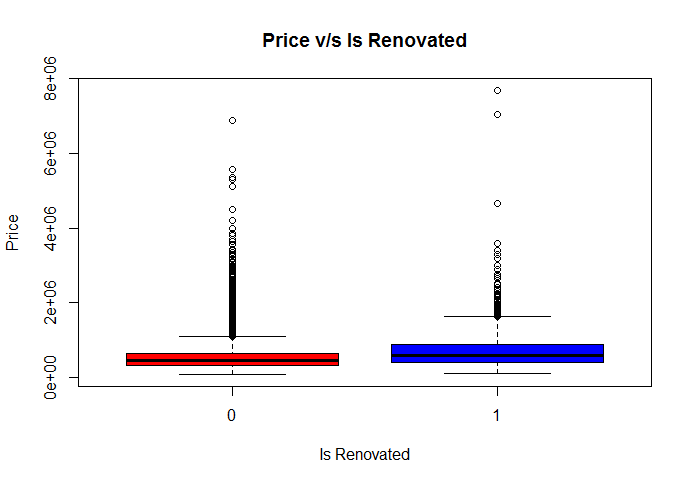


Figure 16: Boxplot - Price vs. Is\_Renovated

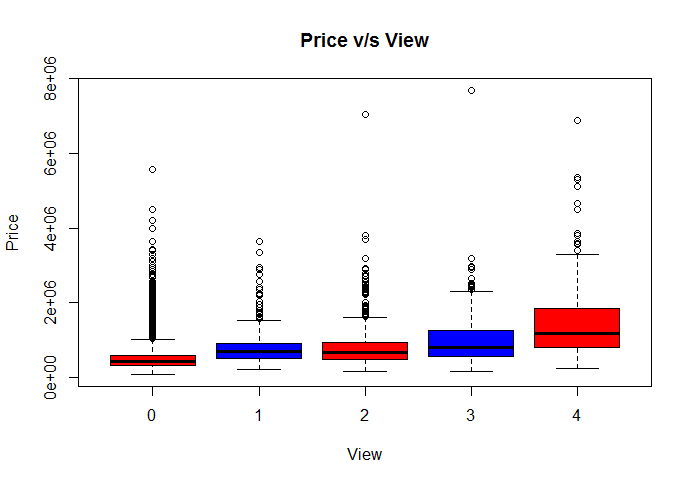


Figure 15: Boxplot - Price vs. View

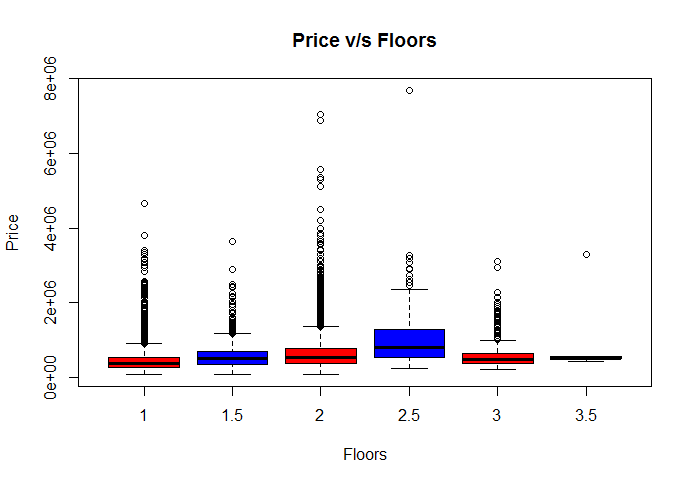


Figure 18: Boxplot - Price vs. Floors

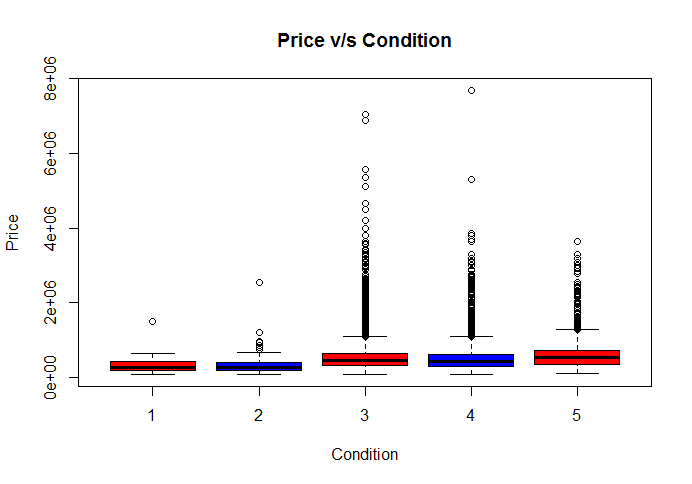


Figure 17: Boxplot - Price vs. Condition

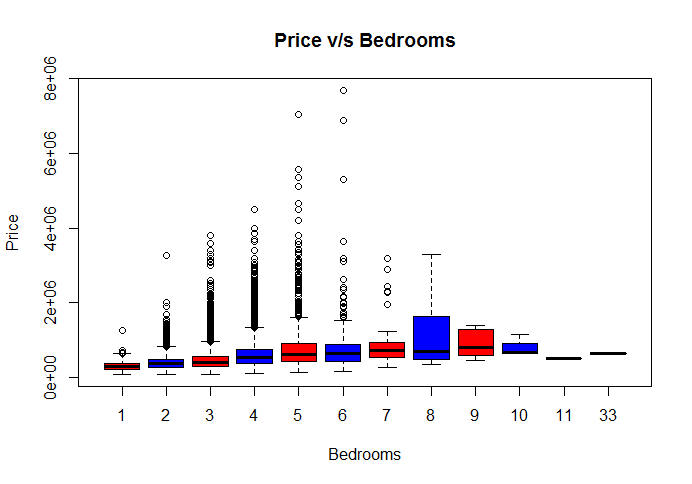


Figure 20: Boxplot - Price vs. Bedrooms

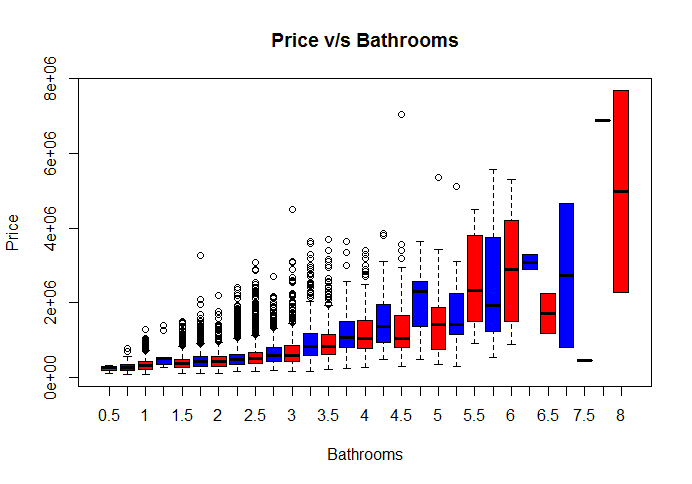


Figure 19: Boxplot - Price vs. Bathrooms

As we can observe, waterfront, bedrooms, bathrooms, floors, view and grade are correlated to price. To calculate correlation using some statistical tool, look at the following methods:

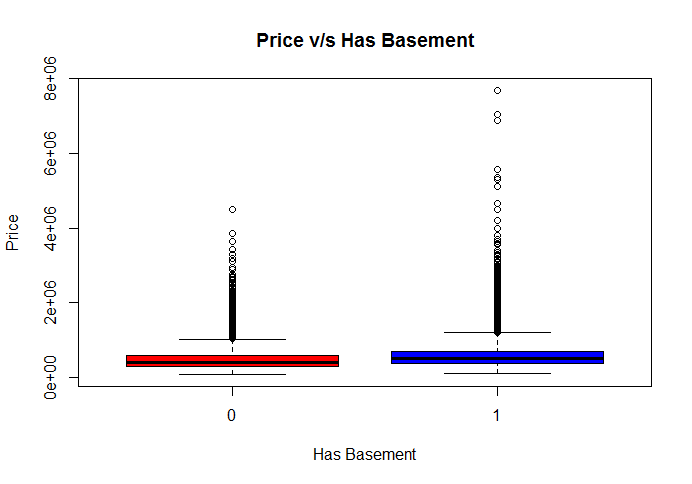
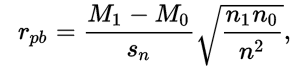
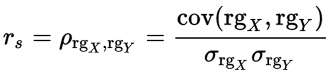


Figure : Boxplot - Price vs. Basement

* To calculate correlation between a binary attribute and a continuous attribute, we use point-biserial correlation formula. It is given as,

where sn is standard deviation of the sample, M1 and M0 are means of continuous attribute on all points that belong to group 1 and group 0 of the binary attribute respectively. Similarly, n1 and n0 are the number of data a point of continuous attribute that belong to group 1 and 0 of binary attribute respectively, and n is the total sample space.

* To calculate correlation between a discrete attribute and continuous attribute, we use Spearman’s correlation formula, as it does not make any assumption about the distribution of the underlying data. It is given as,



where denotes Pearson’s correlation coefficient applied to rank variables, cov(rgx, rgy) is the covariance of the rank variables and and are standard deviations of the rank variables.

Using the above formulae, we computed the correlation between binary features such as waterfront, has\_basement and is\_renovated using point-biserial correlation coefficient and compute correlation between other discrete features and price using spearman’s correlation coefficient:

Correlation between price and waterfront = -0.2663923

Correlation between price and is\_renovated = -0.126079

Correlation between price and has\_basement = -0.1800778

Correlation between price and bedrooms = 0.344245

Correlation between price and bathrooms = 0.4972984

Correlation between price and floors = 0.3224824

Correlation between price and view = 0.2939065

Correlation between price and condition = 0.01799459

Correlation between price and grade = 0.6581517

This confirms our observation about the correlation of categorical features with price.

Through the above experiments, we can draw following conclusions:

* sqft\_living, sqft\_above and sqft\_basement were moderately/strongly correlated to price.
* sqft\_living15, which is the average house square footage of the 15 closest neighbors, was strongly related to price. But when controlling sqft\_living, the relationship disappeared.
* sqft\_lot, sqft\_lot15 and yr\_built were poorly correlated to house.
* Three binary features namely waterfront, has\_basement and is\_renovated were associated with price. However, the associations were small.
* Five ordinal features such as bedrooms, bathrooms, floors, view and grade were moderately/strongly correlated to price.

We used these observations while performing regression.

**SIMPLE LINEAR REGRESSION AND MULTIPLE REGRESSIONS**

We performed simple linear regression i.e. linear regression with one input and then proceeded towards performing regression using multiple features/ inputs. We Split the data into Training set and Testing set for validation. We observed that linear regression with one input has large bias and the model it fits does not explain the variability of the response data around its mean well. However, when we performed multiple regression, we observed that the model has low bias and the model can explain the variability of the response data around its mean in a better way. We deduced the measure of how close the data is to the fitted regression using R squared statistical measure. It is given by:

where is sum of squares of residuals and is given by:

where are the data points and are the predicted values.

is the total sum of squares proportional to variance of the data. It is given by:

Also, we measured the error in prediction using Root Mean Squared Error (RMSE) measure. It is given by:

where N is the sample size.

* **Simple Linear Regression**

We performed simple linear regression on top four highly correlated data i.e. sqft\_living, grade, sqft\_above and bathrooms. Sqft\_living feature provided with the **lowest RMSE (268855.9) and highest R squared (0.4843) values**. The coefficients of linear model are:

intercept = -35201.273

slope = 275.334

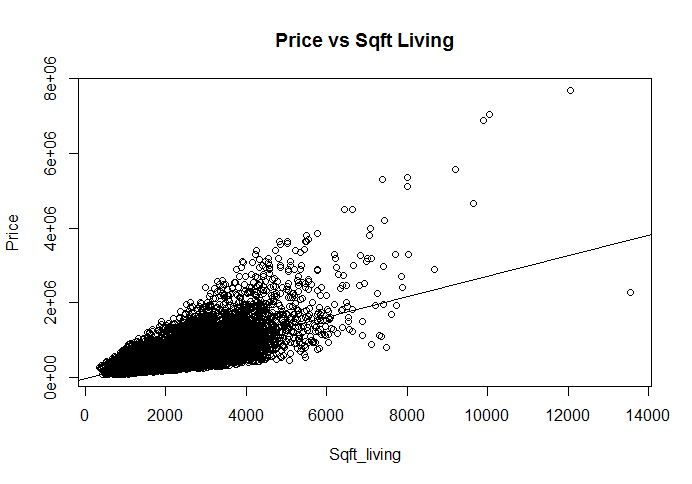


Figure 22: Price vs Square Feet Living Area

* **Multiple Linear Regression**

The plot of House price vs House Sqft\_living (Square footage living area) is observed to be too crowded. The plot of aggregated vectors below provides a better visualization to understand the relationship between price and sqft\_living

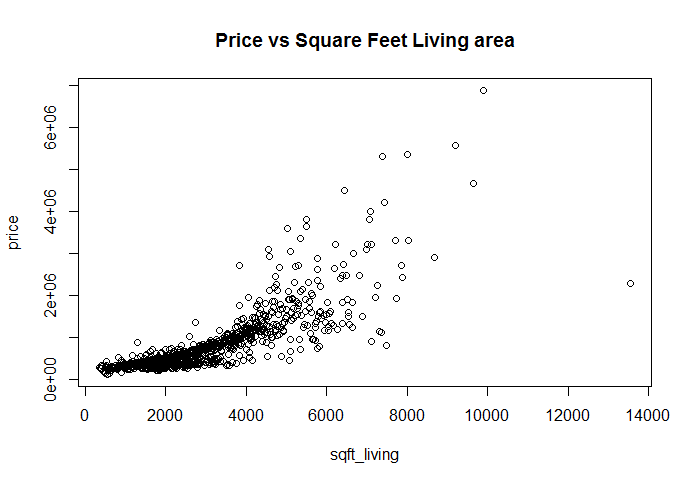


Figure 23: Price vs Square Feet Living Area

As we can observe, the relation between price and sqft\_living is not linear but exponential. Therefore, instead of fitting a linear regression line, we take exponential regression between price and sqft\_living. However, on further inspection, it is interesting to find out that we can observe a linear relationship between log(price) and log(sqft\_living).

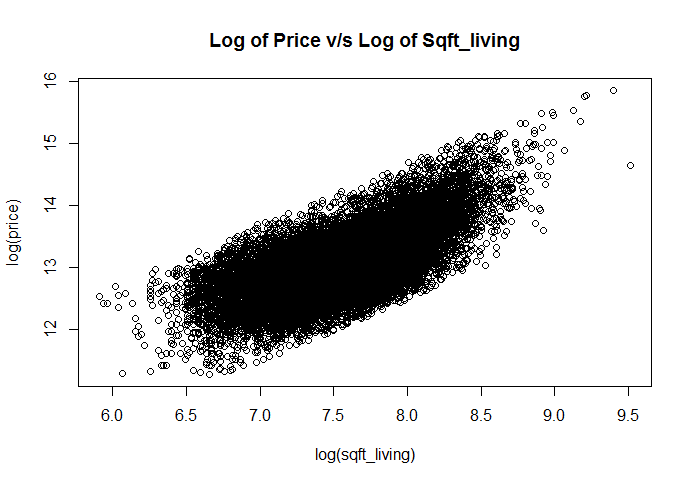


Figure 24: Log of Price vs. Log of Sqft\_Living

Thus, we perform input transformation of price as well as sqft\_living to fit a higher order polynomial regression model. Similarly, we observed the effect of input transformation on the relationship between different features with price. Following plots provides with a way to do the same:

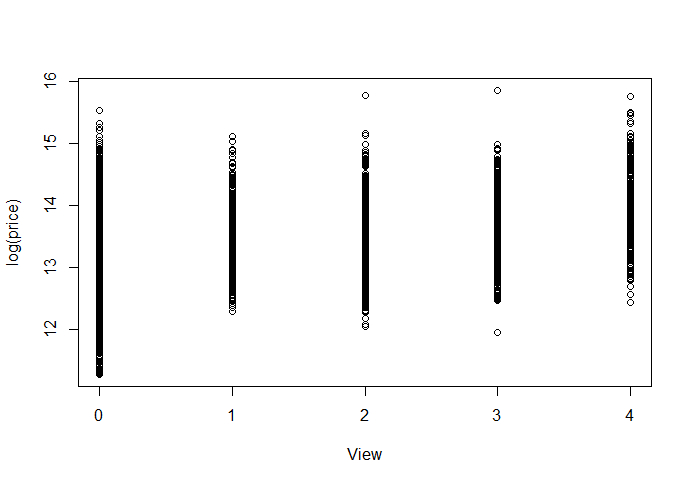
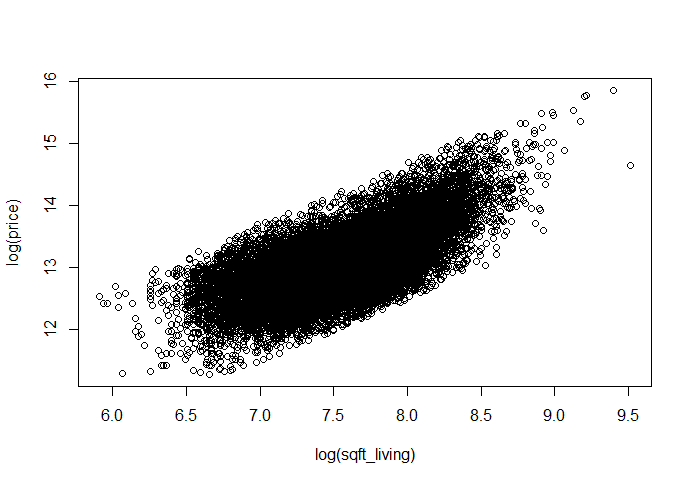
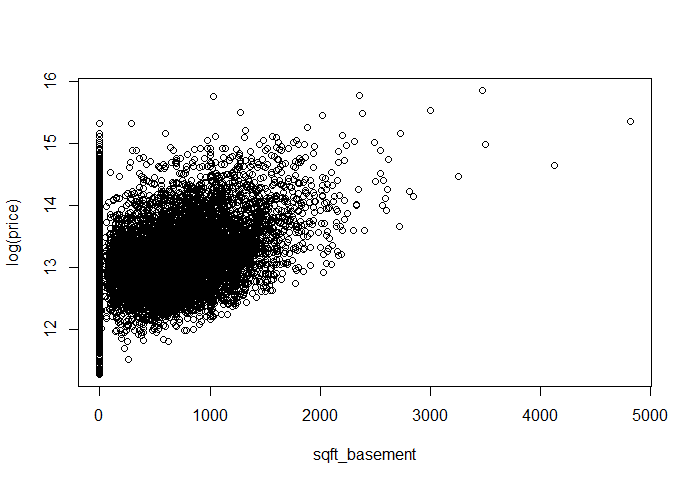
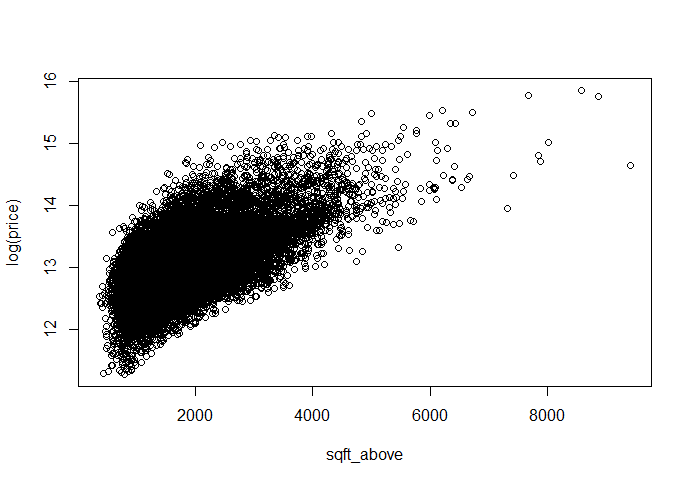
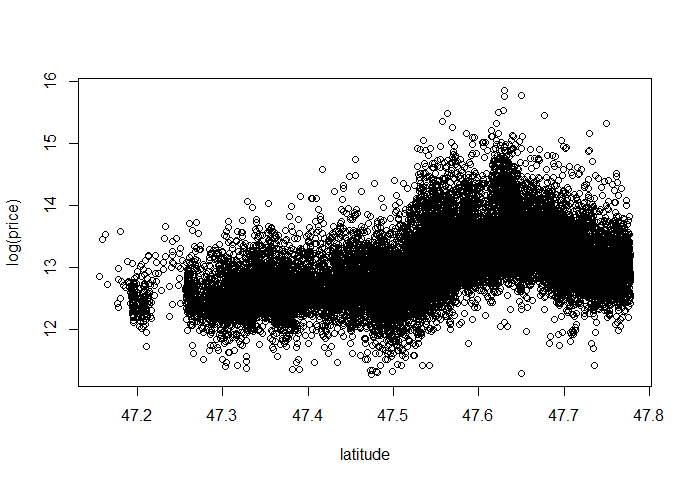
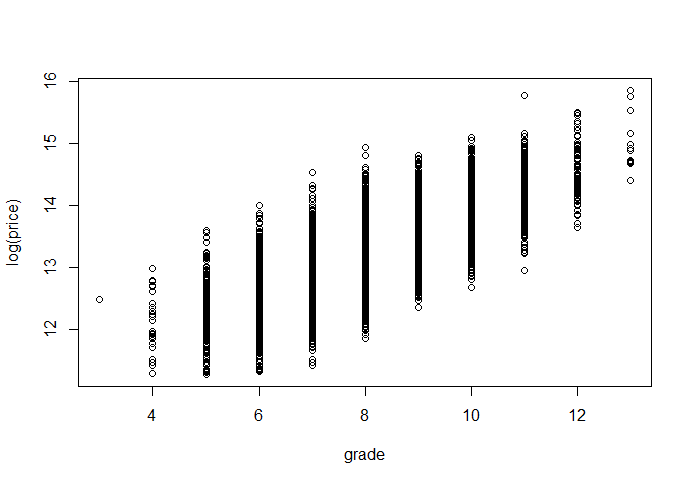
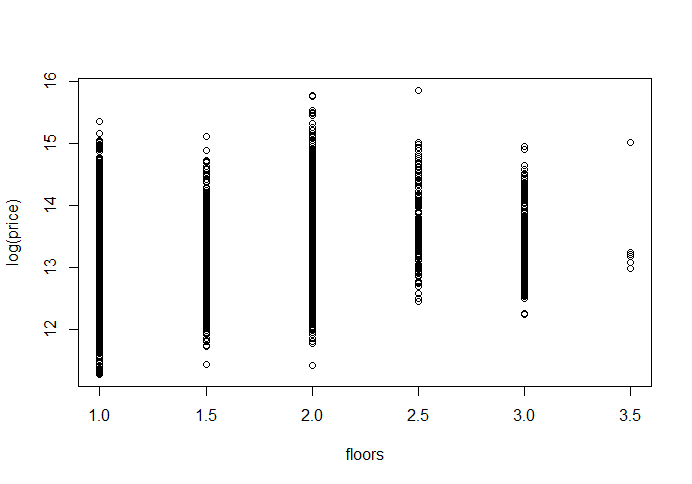
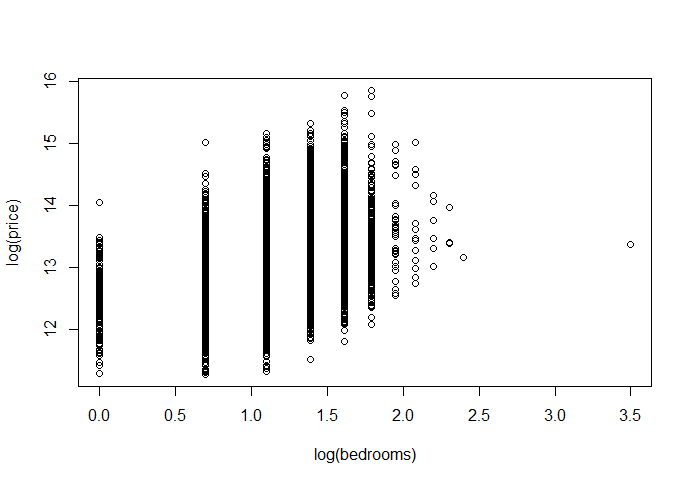
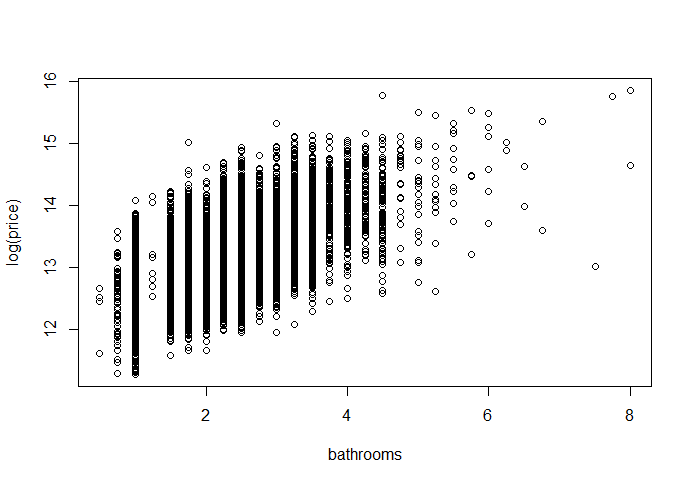
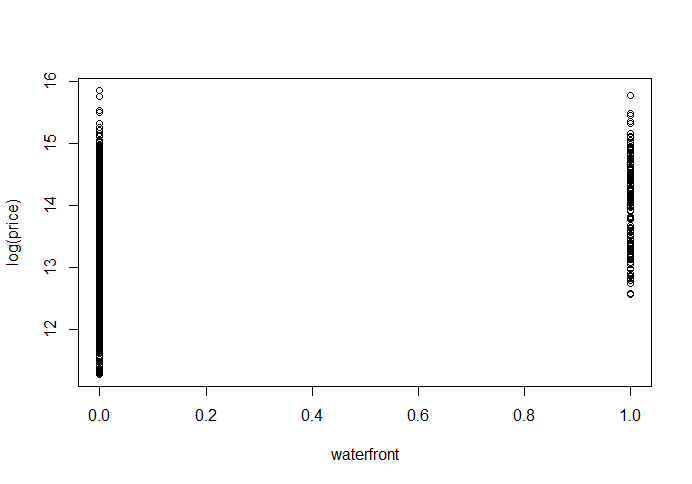


Figure 25: Plots of log(price) vs. other features

After performing multiple regression on this model, we achieved **R squared value of 0.73 and RMSE value of 664040.1**

**RANDOM FOREST**

We applied one of the ensemble methods called “random forest” to best visualize and finding the models of regression on the data. We observed the model using different tree sizes and have found out the Root Mean Squared Error (RMSE) for each model with different tree sizes. Comparison is provided for different tree sizes to conclude the best fit tree size model for regression. Plots are provided to better visualization.

Based on the correlation plotted for price against all the columns in figure12, the most correlated features with price were:

**View Bathrooms**

**Sqft\_Living15 Sqft\_Above**

**Grade Sqft\_Living**

**Bedrooms**

Plot obtained are as follows:

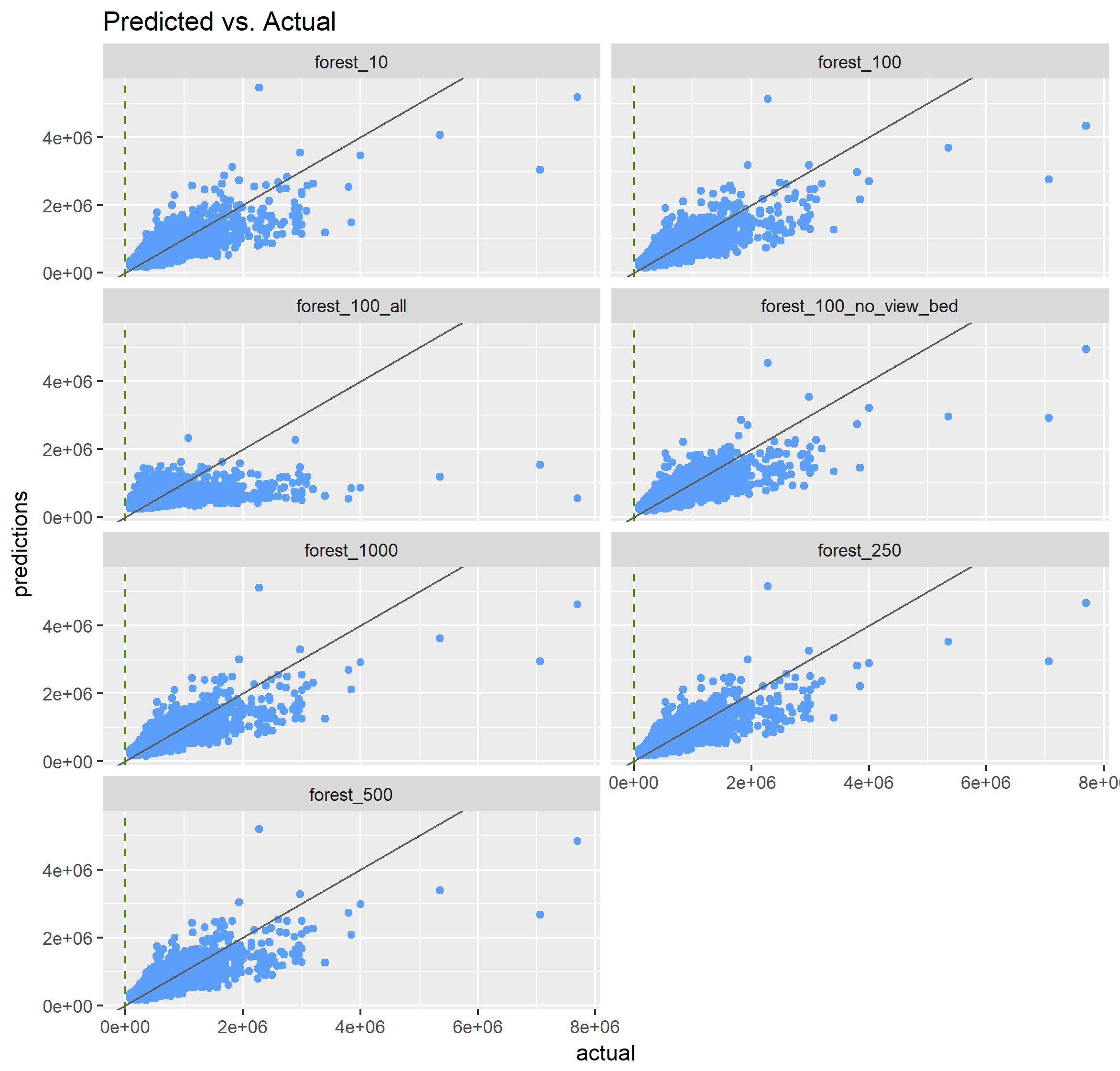


Figure: Random Forest output with different tree sizes

Figure 26: Random Forest plots with different ntree parameters

* Post obtaining the predictors based on random forest, we calculated the root mean squared error (RMSE) for each of the category we selected for prediction
* Based on such observation, following results were obtained for RMSE:

|  |  |
| --- | --- |
| **Ntree** | **RMSE** |
| 10 | 228453.742788293 |
| 100 | 224615.119381526 |
| 100\_all | 342945.78938366 |
| 100\_no\_view\_bed | 229440.210357277 |
| 250 | 222837.728176309 |
| 500 | 223819.312397516 |
| 1000 | 222887.504825556 |

Based on the above table, we can see that the best prediction obtained is from *insert name here* where ntree = 250

*Note: The above result is based on the sample training and testing data obtained while running the program. Above observations will change based on the data sampled while running.*

**CONCLUSION**

For this dataset, we performed 3 regression techniques, viz. Simple linear regression, multiple regression and Random Forest ensemble method based on the observations through data visualization techniques. Based on the RMSE calculated for each of these techniques, we conclude that Random Forest provides the model which best fits the data thus giving the best possible prediction of house sales prices.

**REFERENCES**

[**https://en.wikipedia.org/wiki/Root-mean-square\_deviation**](https://en.wikipedia.org/wiki/Root-mean-square_deviation)

[**https://www.linkedin.com/pulse/regression-analysis-how-do-i-interpret-r-squared-assess-gaurhari-dass**](https://www.linkedin.com/pulse/regression-analysis-how-do-i-interpret-r-squared-assess-gaurhari-dass)

[**https://www.kaggle.com/vmalyi/d/harlfoxem/housesalesprediction/finding-the-most-correlating-variables**](https://www.kaggle.com/vmalyi/d/harlfoxem/housesalesprediction/finding-the-most-correlating-variables)

[**https://en.wikipedia.org/wiki/Coefficient\_of\_determination**](https://en.wikipedia.org/wiki/Coefficient_of_determination)

[**https://www.kaggle.com/piyushauj/d/harlfoxem/housesalesprediction/predicting-the-price-using-multi-linear-regression/code**](https://www.kaggle.com/piyushauj/d/harlfoxem/housesalesprediction/predicting-the-price-using-multi-linear-regression/code)

[**http://www.statisticssolutions.com/correlation-pearson-kendall-spearman/**](http://www.statisticssolutions.com/correlation-pearson-kendall-spearman/)

[**https://www.youtube.com/watch?v=G\_obrpV70QQ&index=6&list=PLqzoL9-eJTNBJrvFcN-ohc5G13E7Big0e#t=0.671305**](https://www.youtube.com/watch?v=G_obrpV70QQ&index=6&list=PLqzoL9-eJTNBJrvFcN-ohc5G13E7Big0e#t=0.671305)

[**https://www.kaggle.com/vicky83/d/harlfoxem/housesalesprediction/price-prediction-comparison/code#L20**](https://www.kaggle.com/vicky83/d/harlfoxem/housesalesprediction/price-prediction-comparison/code#L20)

[**https://www.kaggle.com/auygur/d/harlfoxem/housesalesprediction/step-by-step-house-price-prediction-r-2-0-77**](https://www.kaggle.com/auygur/d/harlfoxem/housesalesprediction/step-by-step-house-price-prediction-r-2-0-77)

[**https://www.kaggle.com/harlfoxem/d/harlfoxem/housesalesprediction/house-price-prediction-part-1**](https://www.kaggle.com/harlfoxem/d/harlfoxem/housesalesprediction/house-price-prediction-part-1)

[**https://www.kaggle.com/harlfoxem/d/harlfoxem/housesalesprediction/house-price-prediction-part-2**](https://www.kaggle.com/harlfoxem/d/harlfoxem/housesalesprediction/house-price-prediction-part-2)