Stat-515 Final Project


# Final Project

# On

# Prediction of House Sales in King County, USA

**Submitted**

**Sowndarya Lahari Tadepalli**

## Introduction:

King County is a county located in the United States. State of Washington. This is the most populous county in the Washington and 13$^{th}$ most populous in the United States [1]. This dataset contains house prices for sale for the king County. It includes the factors concerning the house sale prices in king county sold between May 2014 and May 2015.

## Dataset Description:

The county comprises houses with varied features. The features include bedrooms, bathrooms, lot, presence of waterfront, condition of the house, built year, renovated year etc. The dataset (input_dataset.csv)[2] has 21 columns in it. The description of each of the column is as follows:

Id: a unique identifier for a house

Date: A date on which a house is sold

Price: Price of the house which is target prediction

Bedrooms: Number of Bedrooms in the House

Bathrooms: Number of bathrooms/bedrooms

Sqft_living: square footage of the home

Sqft_lot: square footage of the lot

Floors: Total floors (levels) in house

Waterfront: House which has a view to a waterfront

View: Has been viewed

Condition: How good the condition is (Overall)

Grade: overall grade given to the housing unit, based on King County grading system

Sqft_above: square footage of house apart from basement

Sqft_basement: square footage of the basement

Yr_built: The year in which the house was built

Yr_renovated: The year in which the house was renovated

Zip code: Zip-code of the location where the house is located

Lat: Latitude coordinate

Long: Longitude coordinate

Sqft_living15: Living room area in 2015 (implies-- some renovations)

Sqft_lot15: lot Size area in 2015

## Research Questions:

The research questions are as follows:

1. What is the Best statistical model for our dataset?
2. How to understand the deviation between the Predicted and Test value?
3. Are there any outliers in the dataset which may distort our statistical analysis?
4. How to find the best dependent variables that are highly associated with the response variables?

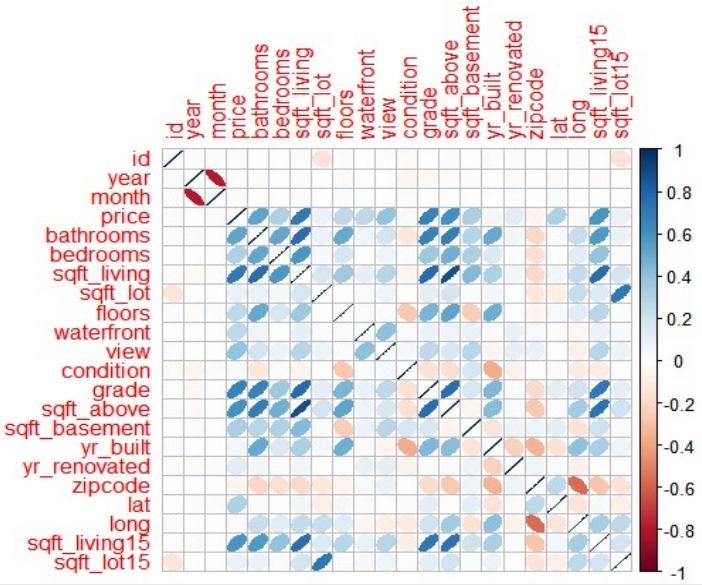
## Data cleaning and Preprocessing:

Data preprocessing and cleaning is one of the crucial parts of our statistical analysis. As part of the data cleaning, we have checked for the inconsistencies in our dataset. We checked for the missing values and found that there are no missing values. We identified some columns namely id, lat, long which would not be useful for our regression analysis

## Exploratory Analysis:

As part of the Exploratory Analysis, we have used summary statistics to get the summaries of all the columns of our dataset, correlation matrix to identify the correlation between variables and some important visualizations based on the correlation matrix.

**Correlation Matrix**



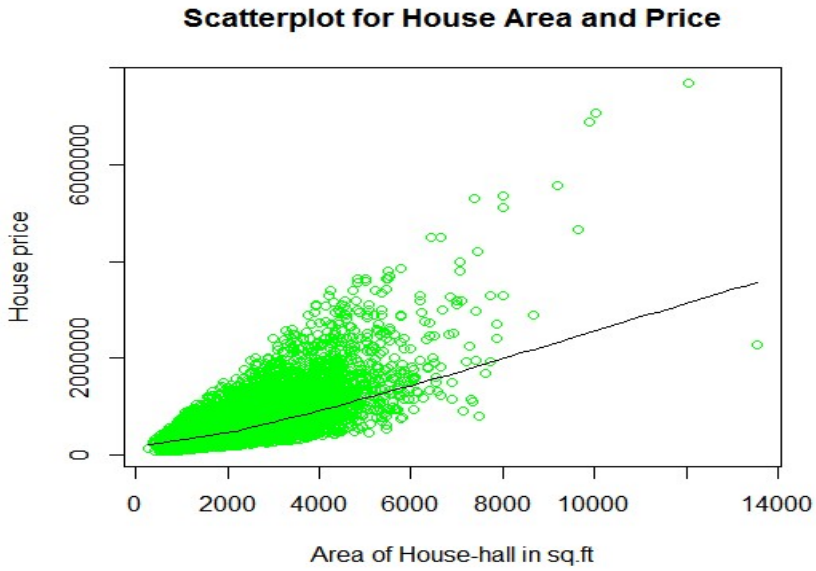

From the correlation plot, we can say that month and year are negatively correlated with each other. Also, the price is positively correlated with the sqft_living, grade, sqft_above, bathrooms, bedrooms, view, sqft_basement, waterfront, floors.

## Summary Statistics:

```
> summary(input)
       id                 year          month            price          bathrooms
 Min.   :1.000e+06   Min.   :2014   Min.   : 1.000   Min.   :  75000   Min.   :0.000
 1st Qu.:2.123e+09   1st Qu.:2014   1st Qu.: 4.000   1st Qu.: 321950   1st Qu.:1.750
 Median :3.905e+09   Median :2014   Median : 6.000   Median : 450000   Median :2.250
 Mean   :4.580e+09   Mean   :2014   Mean   : 6.574   Mean   : 540182   Mean   :2.115
 3rd Qu.:7.309e+09   3rd Qu.:2015   3rd Qu.: 9.000   3rd Qu.: 645000   3rd Qu.:2.500
 Max.   :9.900e+09   Max.   :2015   Max.   :12.000   Max.   :7700000   Max.   :8.000
    bedrooms       sqft_living       sqft_lot           floors        waterfront
 Min.   : 0.000   Min.   :  290   Min.   :    520   Min.   :1.000   Min.   :0.000000
 1st Qu.: 3.000   1st Qu.: 1427   1st Qu.:   5040   1st Qu.:1.000   1st Qu.:0.000000
 Median : 3.000   Median : 1910   Median :   7618   Median :1.500   Median :0.000000
 Mean   : 3.371   Mean   : 2080   Mean   :  15107   Mean   :1.494   Mean   :0.007542
 3rd Qu.: 4.000   3rd Qu.: 2550   3rd Qu.:  10688   3rd Qu.:2.000   3rd Qu.:0.000000
 Max.   :33.000   Max.   :13540   Max.   :1651359   Max.   :3.500   Max.   :1.000000
      view           condition        grade          sqft_above   sqft_basement
 Min.   :0.0000   Min.   :1.000   Min.   : 1.000   Min.   : 290   Min.   :   0.0
 1st Qu.:0.0000   1st Qu.:3.000   1st Qu.: 7.000   1st Qu.:1190   1st Qu.:   0.0
 Median :0.0000   Median :3.000   Median : 7.000   Median :1560   Median :   0.0
 Mean   :0.2343   Mean   :3.409   Mean   : 7.657   Mean   :1788   Mean   : 291.5
 3rd Qu.:0.0000   3rd Qu.:4.000   3rd Qu.: 8.000   3rd Qu.:2210   3rd Qu.: 560.0
 Max.   :4.0000   Max.   :5.000   Max.   :13.000   Max.   :9410   Max.   :4820.0
    yr_built     yr_renovated       zipcode           lat             long
 Min.   :1900   Min.   :   0.0   Min.   :98001   Min.   :47.16   Min.   :-122.5
 1st Qu.:1951   1st Qu.:   0.0   1st Qu.:98033   1st Qu.:47.47   1st Qu.:-122.3
 Median :1975   Median :   0.0   Median :98065   Median :47.57   Median :-122.2
 Mean   :1971   Mean   :  84.4   Mean   :98078   Mean   :47.56   Mean   :-122.2
 3rd Qu.:1997   3rd Qu.:   0.0   3rd Qu.:98118   3rd Qu.:47.68   3rd Qu.:-122.1
 Max.   :2015   Max.   :2015.0   Max.   :98199   Max.   :47.78   Max.   :-121.3
 sqft_living15    sqft_lot15
 Min.   : 399   Min.   :   651
 1st Qu.:1490   1st Qu.:  5100
 Median :1840   Median :  7620
 Mean   :1987   Mean   : 12768
 3rd Qu.:2360   3rd Qu.: 10083
 Max.   :6210   Max.   :871200
>
```

From the Summary Statistics we can say that Minimum House Price in the King County is 75000 whereas the Maximum House price is 7700000.

## Visualizations:



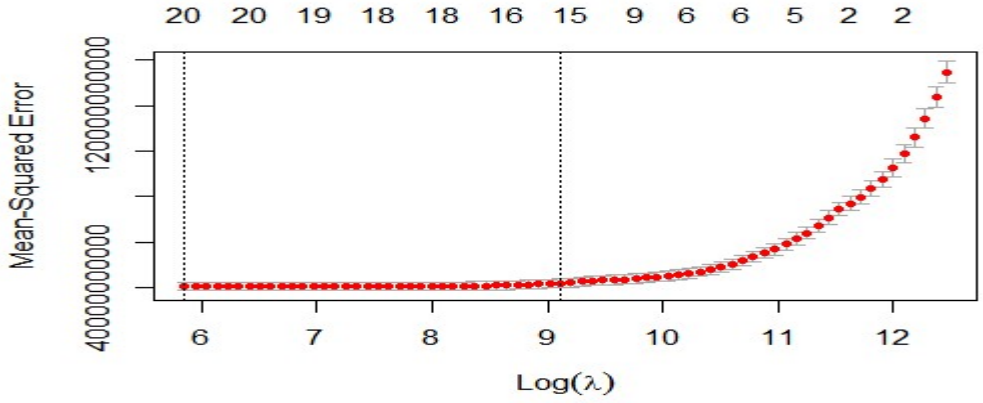

Price and Sqft_living is positively correlated with each other. As the Sqft_living increases the House sale price also increases.

## Statistical Models:

### 1. Lasso Regression:

Lasso regression is the variable selection method for linear regression models. The main goal of lasso regression is to get the subset of predictors that minimizes the prediction error for a quantitative response variable. This can be done by imposing a constraint on the parameters which results in the shrinkage of regression coefficients to zero. So, the variables with regression coefficient zero are excluded whereas the variables with the nonzero regression coefficient can be considered as they are strongly associated with the response variable.

**Lasso Regression**

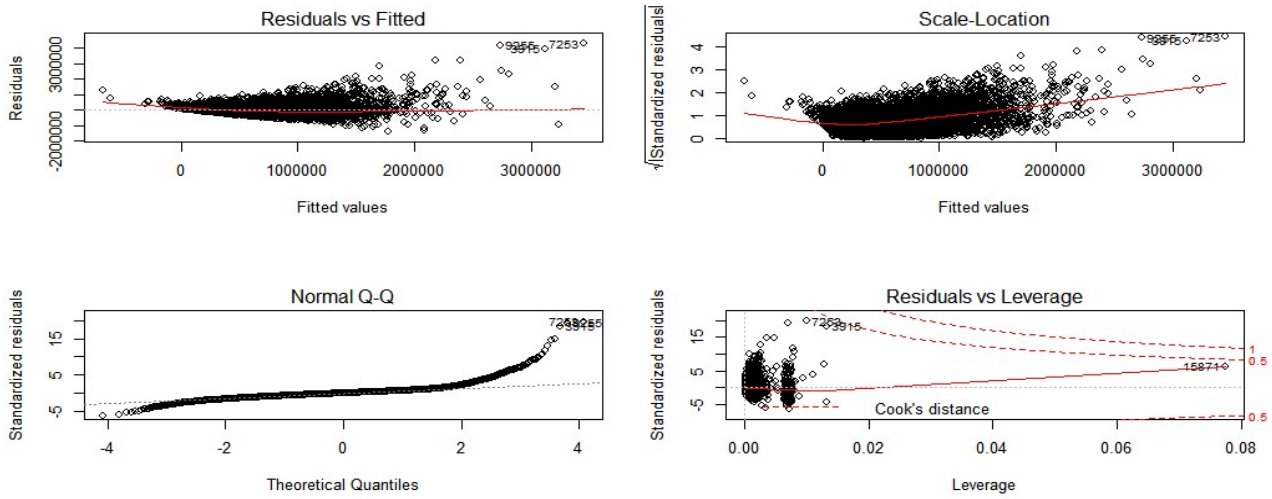



The best lambda value obtained is 348.9488. The log of lambda is taken in x-axis and Mean Squared Error is taken in the y-axis. Also, the columns named id, sqft_basement and sqft_lot has the regression coefficients approximated to zero. This answers our 4$^{th}$ research question.

## 2. Multiple Linear Regression

Multiple Linear Regression is a statistical technique which uses several explanatory variables to predict the outcome of a response variable. The main goal of Multiple Linear Regression is to model the linear relationship between the independent and dependent variables. The $R^2$ is the statistical measure that is used to measure how much of variation in the outcome can be explained by variation in the independent variables.

### Diagnostic Plots




**Residuals vs Fitted plot:**

In the Residuals vs Fitted plot the data distribution is in a funnel shaped which is pointed towards only center. So it is not linearly distributed.

**Scale-Location plot:**

There is no equal spread of residuals in the range of predictors. So, the data doesn't have the variance.

**Normal Q-Q plot:**

The curve depicts deviation of residuals from the diagonal line in both lower and upper bound, it follows characteristics of heavier head at right end. The residuals normality is violated because there is deviation near the ends.

**Residuals vs Leverage plot:**

All the residuals are less than the cooks' distance which indicates that there are no influential outliers from the dataset. This answers our third research question.

From the diagnostic plots, we can say that this model is not the best fit for our dataset.

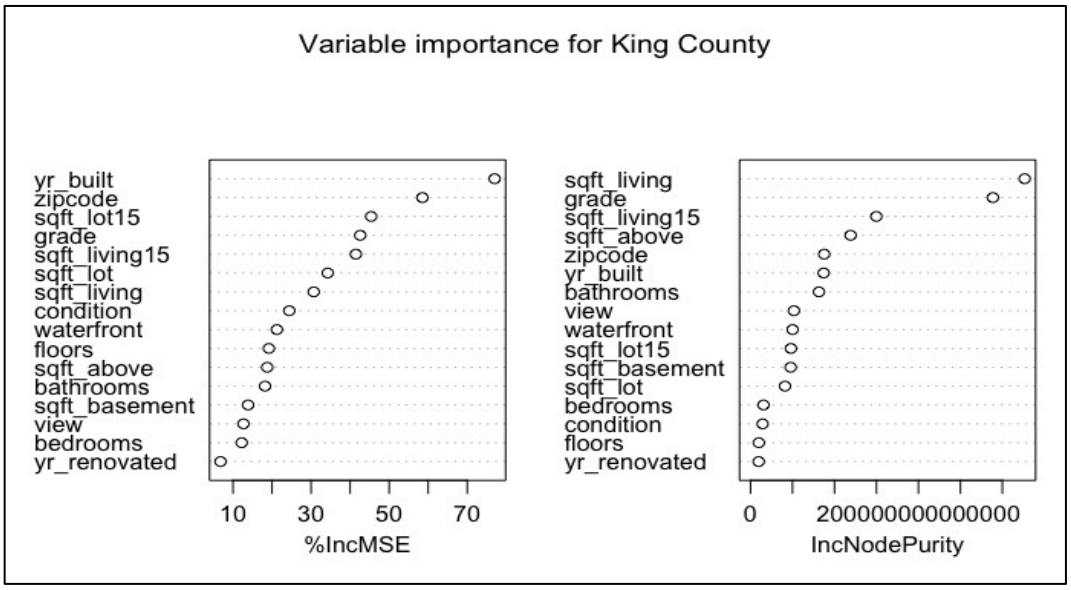
## 3. Random Forest

### 3.1 Regression using Random Forest

By means of a small tweak that decorrelates the trees, random forests provide improvement over bagged trees. We build several decision trees on bootstrapped training samples, as in bagging. But each time a split is considered in a tree, a random sample of m predictors is chosen as split candidates from the full set of p predictors when building those decision trees. Only one of those predictors m can use the split. In other words, the algorithm is not even permitted to consider most of the available predictors when constructing a random forest at each split in the tree. The bagged trees, most or all the trees in the top split will use the strong predictor. Many of the bagged trees would also look very close to each other. Therefore, the forecasts from the bagged trees will be highly correlated. By forcing each split to consider only a subset of predictors, random forests overcome this problem.

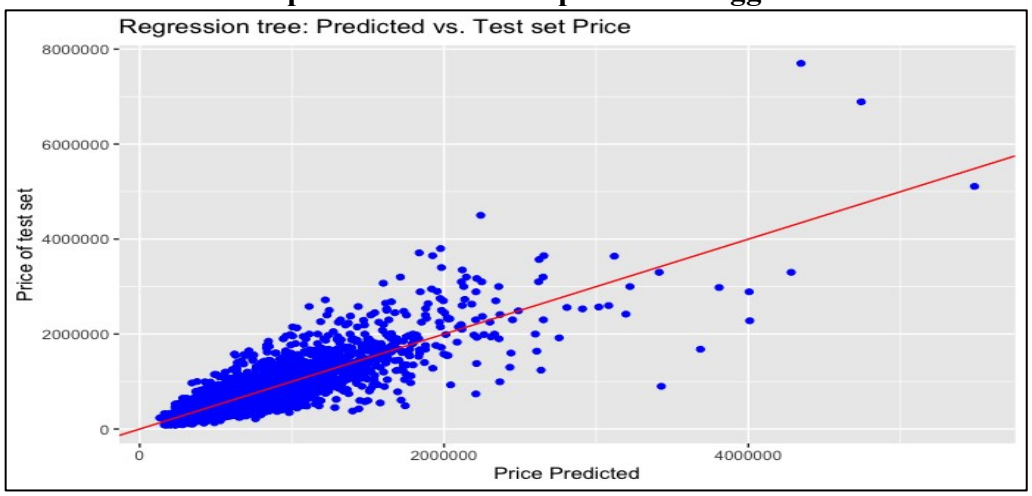
### 3.2 Model Implementation:

After the cleaning and preprocessing of the data, training dataset is created and test the Mean squared error (MSE) of the data. Then, bagging is performed to with an attribute split of 15 (mtry=15). Firstly, Random Forest function is called to compute the test MSE with a default value of 500 trees and a graph is plotted against Predicted value and test value with the 'Price' variable. Then, we find variable importance measures and create a regression model for the most important variables. We got the variance as 81%




Variable importance or Gini importance both can be performed but The variable importance's are critical. Hence, we decided to go with variable importance. The variable importance graph represents the percentage increase in Mean squared error and the important variable listed highest to lowest with node impurity. It can be noted the attributes year built and zip code have the highest percentage of MSE whereas, bedrooms and year renovated has the lowest percentage of MSE. It can also be noticed that square foot living and grade are the most important variable whereas, floors and year renovated are the least important variables.

**Scatter plot for test set and predicted bagged tree**

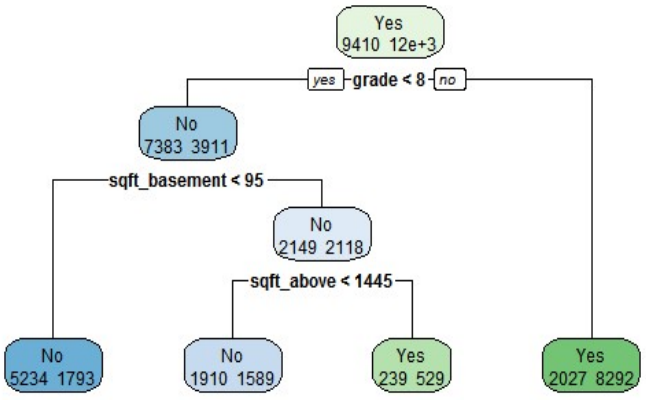



From the plot, it is noticed that there are three outliers. Both the predicted and test set are directly proportional as they both increase linearly. This answers our second research question.

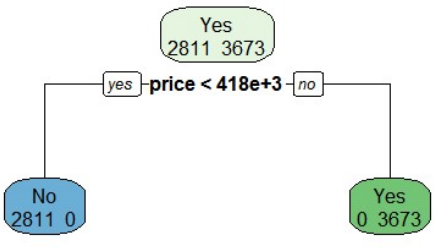
## 4. Decision Tree Regression

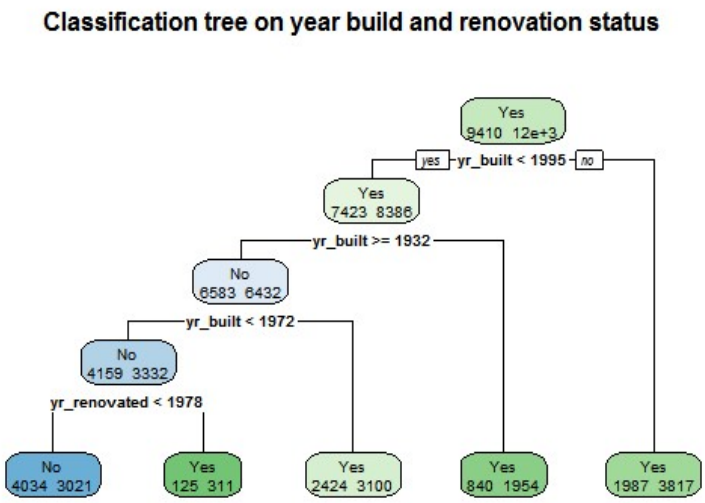
Decision Tree divides the dataset into smaller groups while at the same time builds the regression model incrementally in the form of a tree structure. The result is a tree with decision nodes and leaf nodes. The decision node has two or more branches which represents the values for the attribute tested. Leaf nodes represent the decision on the target variable. The topmost node is the root node which is the best predictor.

**Classification tree on House structure**




**Price wise Classification tree**






Here, we have constructed three decision trees based on yr_built, price and grade. The overall accuracy for the model is 97%.

The best model which fits our data is Decision tree when compared to the random forest. This answers our first research question.

## Results:

We can say the following from our statistical analysis

- There are no outliers which would distort the statistical analysis and multiple linear regression is not the best model.
- We found that both predicted and test values are directly proportional and thus model fits our data.
- Decision tree is the best model for our data when compared to the Random forest

We can conclude from our summary statistics and visualizations that

- House sales price is directly proportional to the Sqft_living.
- The most preferred house has 2.25 bathrooms and 3 Bedrooms.
- The idea grade for most of the houses are 7.
- The Min and Max House Sales price in king's county is 75000$ and 7700000$ respectively.

## References:

[1] "King County, Washington." Wikipedia. Wikimedia Foundation, April 19, 2020. https://en.wikipedia.org/wiki/King_County,_Washington.

[2] Harlfoxem. “House Sales in King County, USA.” Kaggle, August 25, 2016. https://www.kaggle.com/harlfoxem/housesalesprediction.