**Analysis of House Price**

  For many Americans homeownership is part of achieving their personal 'American Dream'. Most Millennial renters aged 18-34 (89%) plan to buy a home one day - more than any other generation. But what does this dream home look like? And where is located? What amenities do people dream of most?

For most people, a house is the biggest purchase they will make in their lives, one they will pay off for years, even decades, to come. Based on the type of homes people prefer and their priorities in amenities available in those homes answer the questions of new home buyers, sellers, remodelers and builders to make better decisions.

**Data**

The data set is Kaggle’s, ‘Iowa House Prices: Regression Techniques’. Dataset contains 81 variables and 1461 observations. Of the 81 variables, 23 are nominal, 23 are ordinal, 14 are discrete, and 21 are continuous. The variables included are basic characteristics that anyone wanting to buy a house would be interested in.

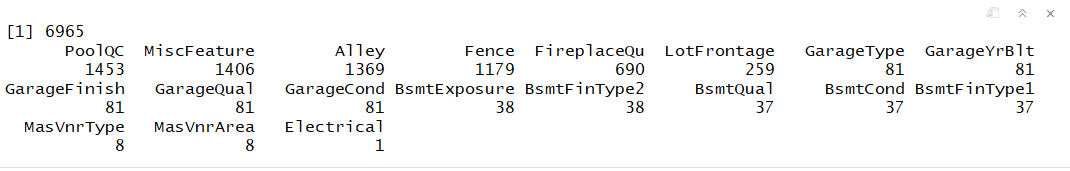
For the most part, the different variables may be split up into specific groups. In general, the 20 continuous variables relate measurements of area dimensions for each observation. These include, among others, the sizes of lots, rooms, porches, and garages. The 14 discrete variables mostly have to do with the number of bedrooms, bathrooms, kitchens, etc. that a given property has.The rest of the nominal variables identify characteristics of the property and dwelling type/structure. Most of the ordinal variables are rankings of the quality/condition of rooms and lot characteristics.

**Exploration Analysis**

Loading the data to r and converting it to a local data frame the exploration analysis begins.

**Step1: Cleaning the data**

Finding NA’s. Below shows the columns and the number of observations contains NA.



Checking with correlated variables helps to identify the NA’s as missing values or the house doesn’t have that feature. Such as the PoolQc missing more values, a quick look at pool area shows the house doesn’t have a pool. Alley have NA encoded as a level to specify No Alley Access. It's not "missing" values. There are 1369 missing values in Alley. Coding the NAs to No Alley, fixes all these missing values.

Data wrangling in this dataset includes changing the categorical value NA’s to respective feature name with no (e.g. “No Fireplace”, “No Fence” ...). Continuous variables NA’s need to change to 0. Lot Frontage we populate the missing values with the mean of lot frontage.5 Garage variables are missing 81 values. Checking the variables 'Garage Cars' and 'Garage Area' gives the idea there is no garage built with the home. So, the missing values are 'none'. Later created new variables total living area sqft, age of home and total number of bathrooms for the exploration analysis.

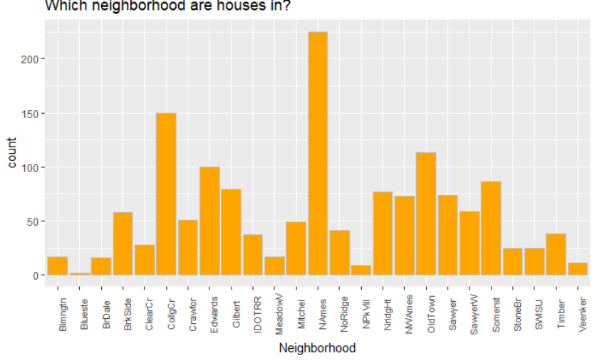
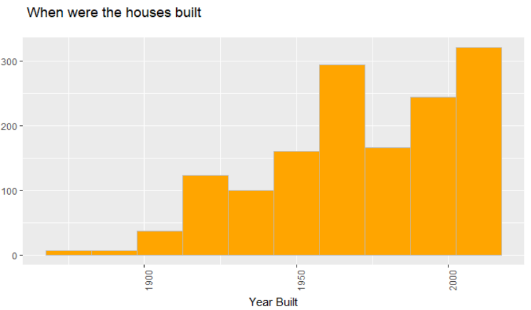
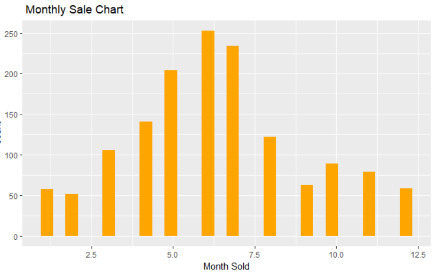
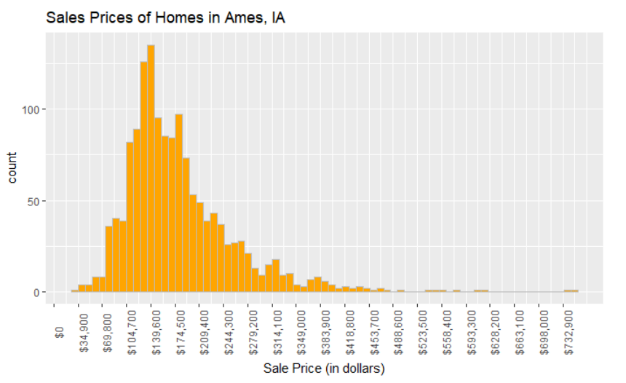
**Step 2: check for zero variance predictors**

We can reduce the dimension of the data by removing those columns for which variance is zero, because zero variance columns have unique values. So those columns don’t impact the output at all and can be safely removed from the dataset.

There are 20 near zero variance variables in the dataset. After removing the near zero variance predictors from dataset, the new data set contains 62 variables and 1460 observations.

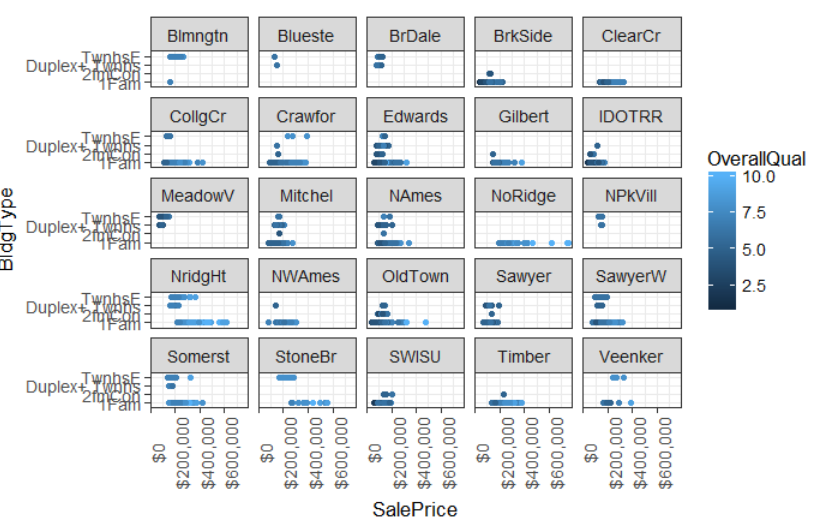
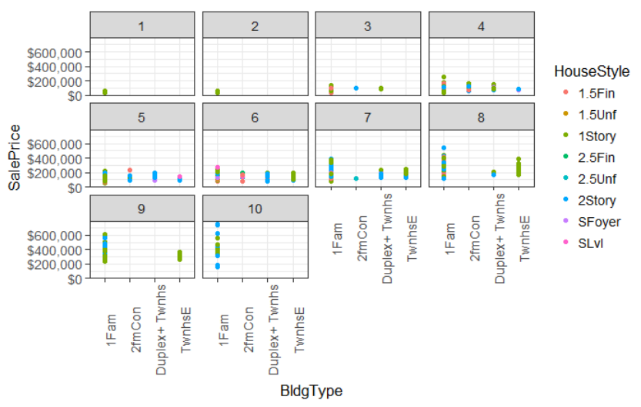
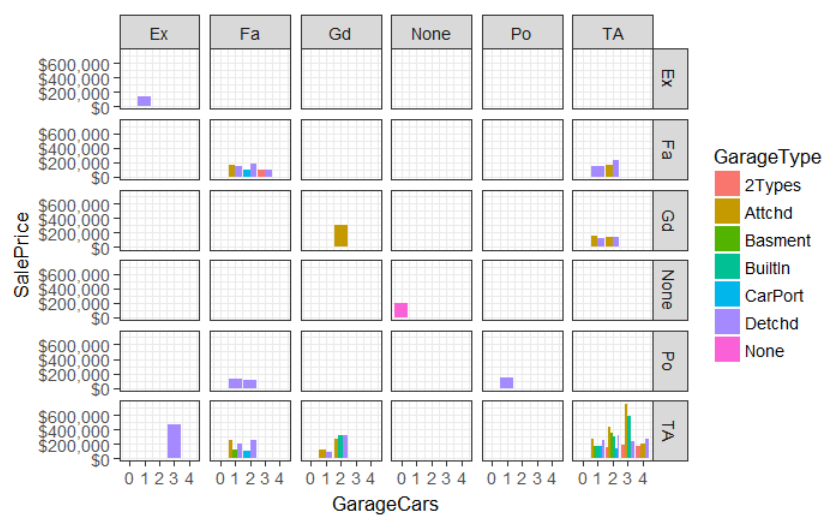
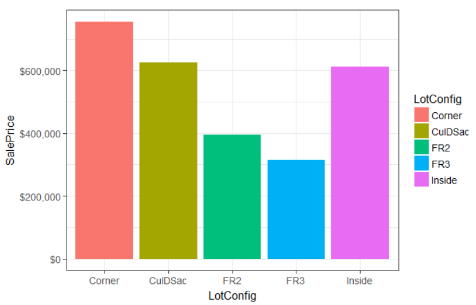
**Step 3: Preliminary exploration of data**

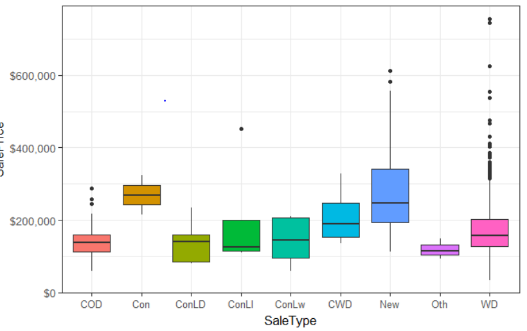
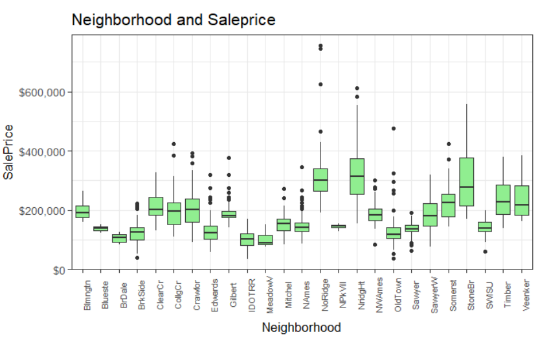
Using histograms graphs and plots examine the distribution of the variables.

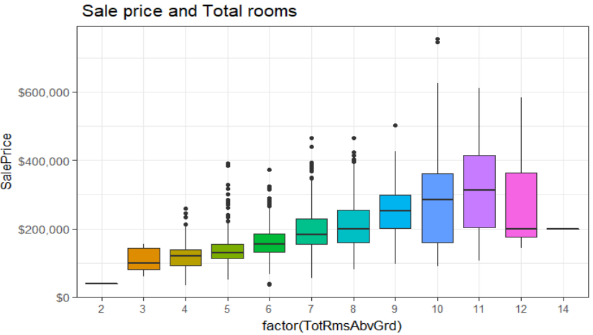
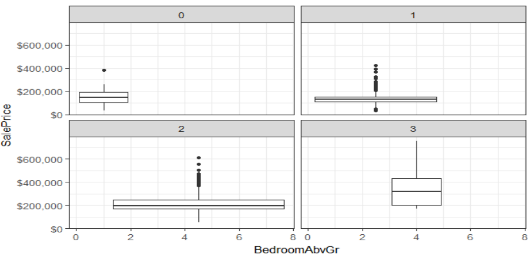


Used different type of bar graphs, plots and box plots to visualize and compare the variables with sale price to get a better idea about correlation of variables and price in our cleaned dataset.

**Step 4: Correlation and comparison analysis**

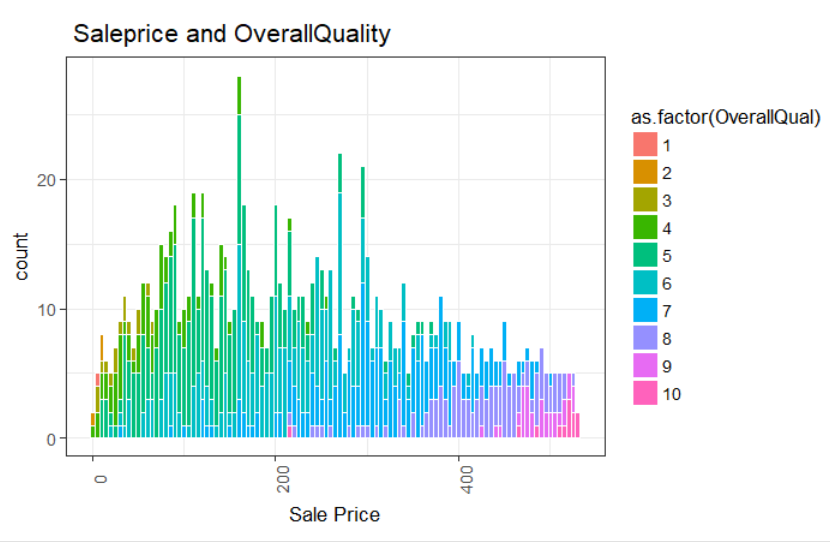


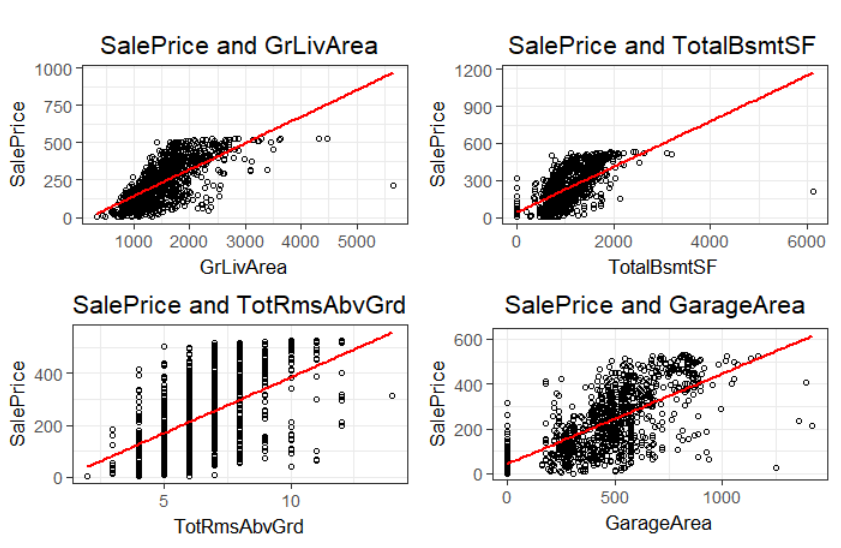


The exploration and visualizations reveal some interesting facts about the data. The minimum and maximum sale price is 34900 and 755000. The price of homes is more in some neighborhoods than others.  The oldest house builds in 1872 and the newest 2010.Home sales peaks in May, June and July. The lot size and home size are correlated. Single family homes sell more.

 The exploration analysis shows the strong correlation of sale price and overall quality of house, living area, basement area and garage area square foot.





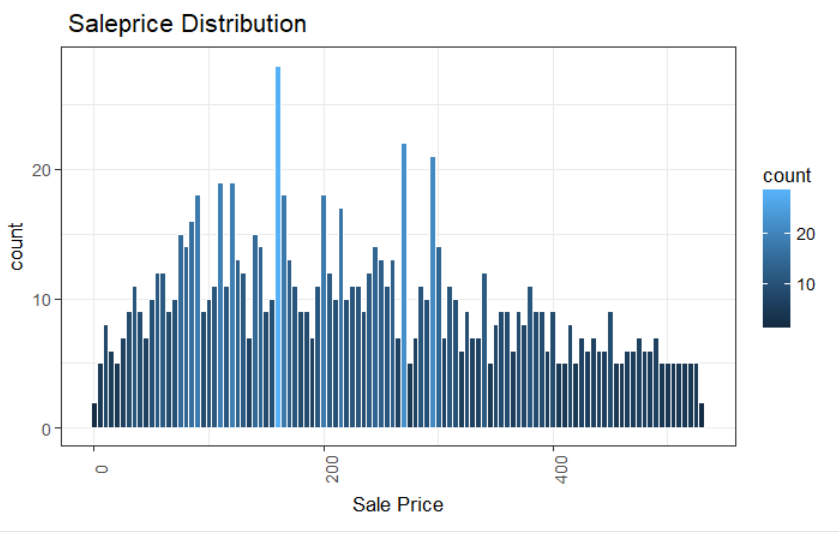
**Step 5: Train / Test split**

After cleaning and performing exploration analysis data set need to split to train and test data for prediction. We split the data set in to a 0.7/0.3.

Train - 1024 observations and 62 variables

Test - 436 observations and  62 variables.

We will use the train set to create our models and test set to predict the house price. Below is the histogram of sale price in our split train dataset.



**Step 6: Analysis**

Creating a correlation heat map shows the positive and negative correlations of variables impacts the price. We select the variables and create model\_var for the heat map and later to create our models.

model\_var <- c ('SalePrice',

               'OverallQual','OverallCond','YearBuilt','ExterCond',

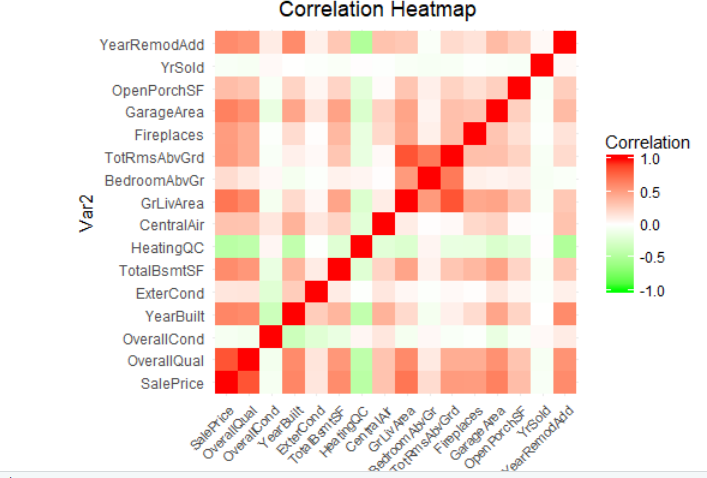
               'TotalBsmtSF','HeatingQC',

               'CentralAir','GrLivArea','BedroomAbvGr',

               'TotRmsAbvGrd','Fireplaces',

               'GarageArea','OpenPorchSF',

                'YrSold','YearRemodAdd')



In this graph, Red indicates perfect positive correlation and Green indicates perfect negative correlation. As we can see, there are several variables should be paid attention to: Garage Area, Fireplaces, TotRmsAbvGrd, GrLivArea, HeatingQC, TotalBsmtSF, YearBuild and YearRemodAdd.

**Step 7: Model fitting**

After Descriptive Analysis we are moving into Predictive Analysis section. We will try three different models.

Linear Regression Model

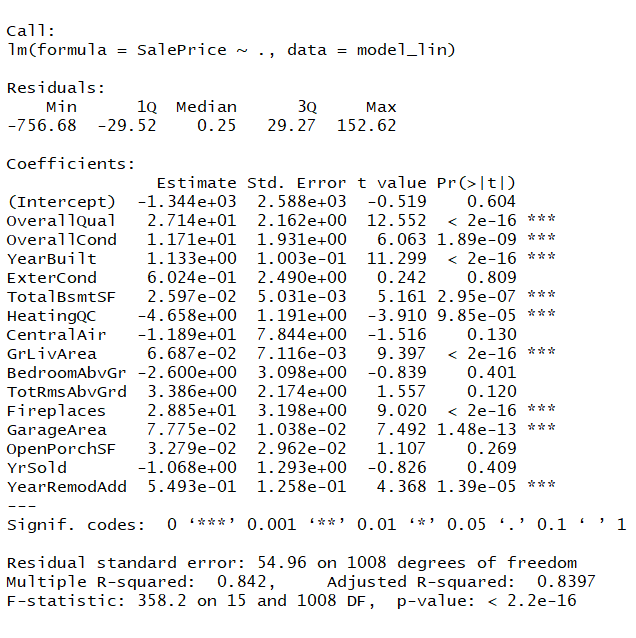
Classification & Regression Trees (CART) Model

Random Forest Model

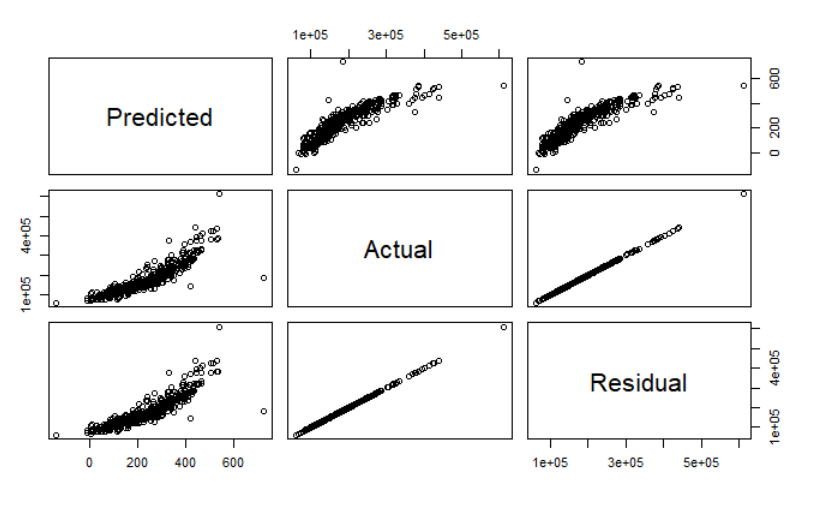
Linear Regression Model - In Linear Regression Model, the relationships between Dependent and Independent Variables is expressed by equation with coefficients. The aim of this model is to minimize the sum of the squared residuals.

\*\*Variables in this model\*\*:

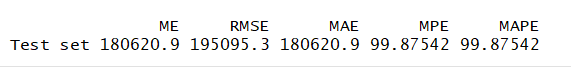
SalePrice, OverallQual, OverallCond, YearBuilt, ExterQual, ExterCond, TotalBsmtSF, HeatingQC, CentralAir, GrLivArea, BedroomAbvGr, TotRmsAbvGrd, Fireplaces, GarageArea, OpenPorchSF, YrSold, YearRemodAdd.



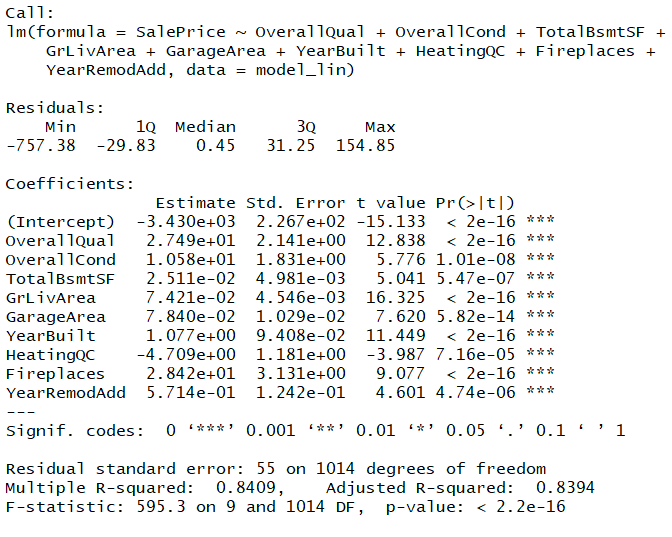
Plot of linear regression model

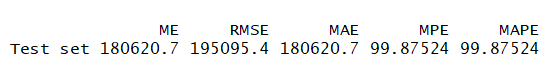


Forecast and model accuracy

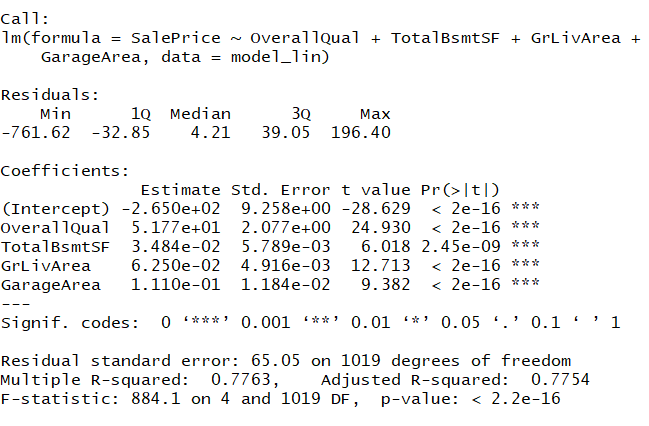


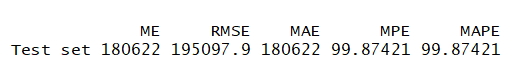
Linear regression Model 2





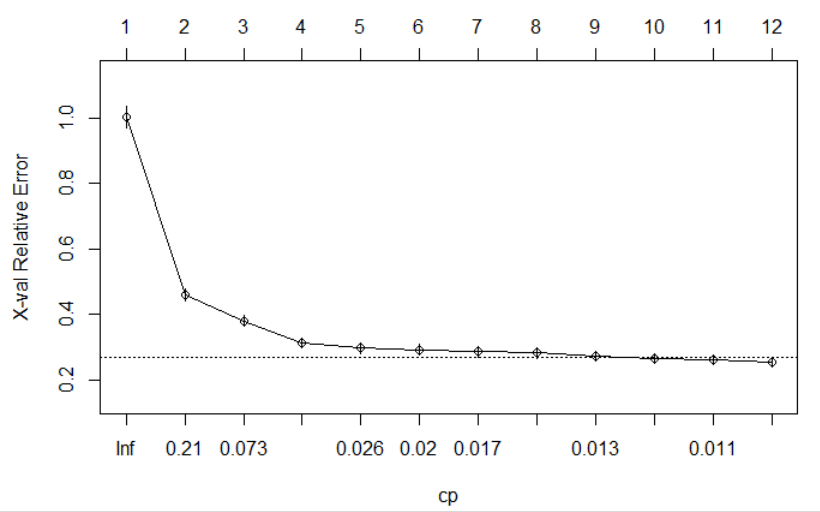
Linear Regression model 3

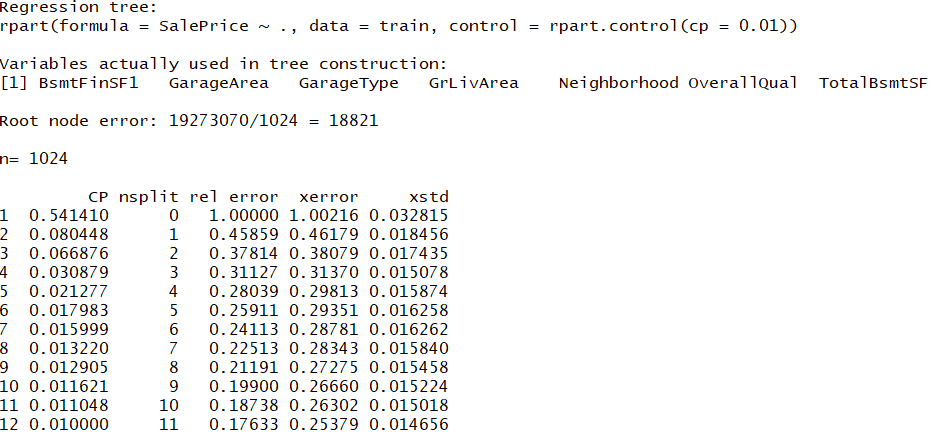




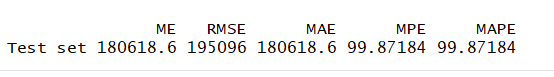
**Step 8: Tree Model**

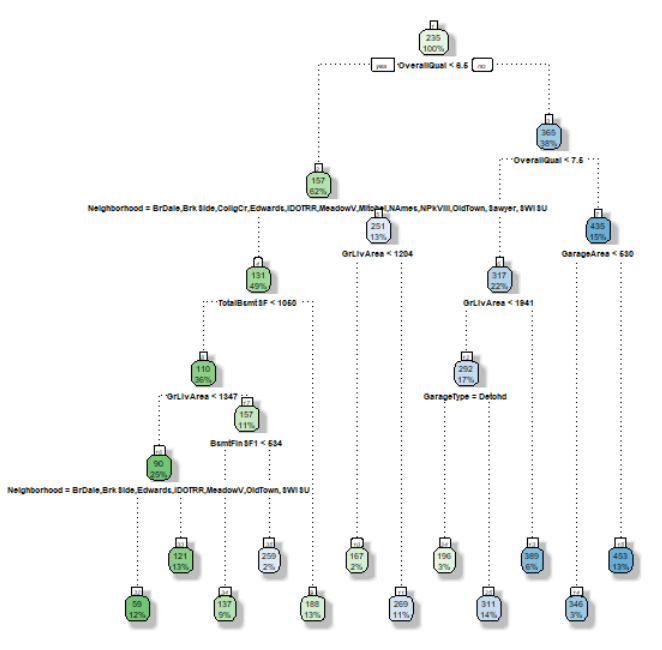
Size of tree



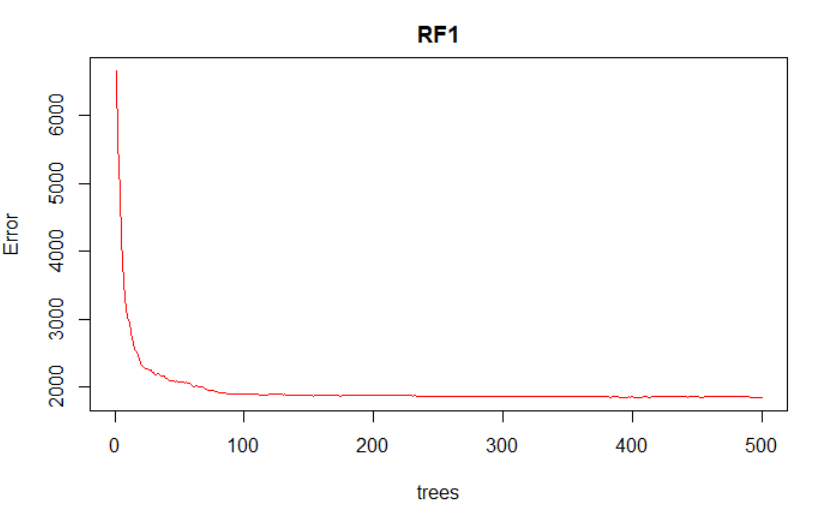
****

Forecast and model accuracy



****

**Step 9: Random Forest**

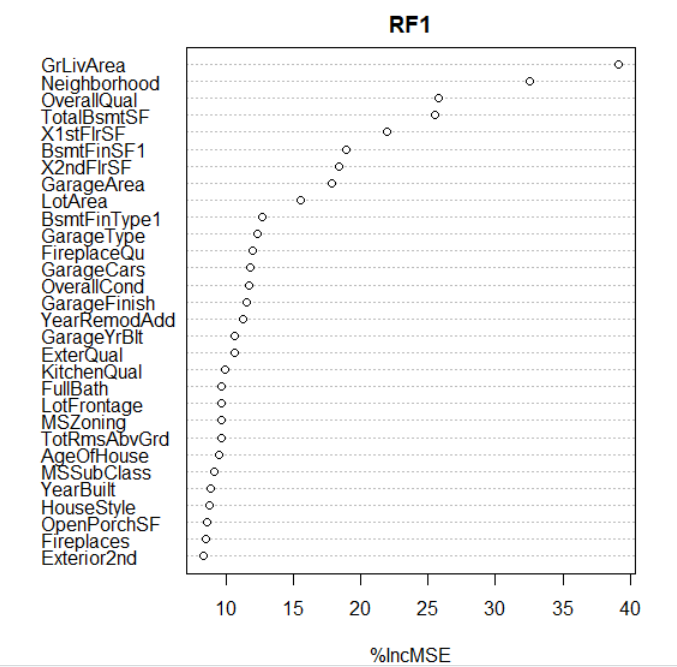
****

This plot shows the Error and the Number of Trees. We can easily notice that how the Error is dropping as we keep on adding more and more trees and average them.

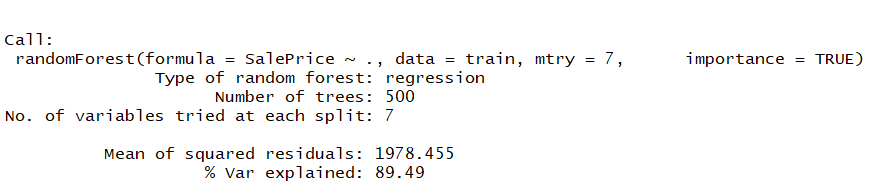
Forecast and model accuracy

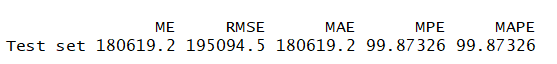
https://lh5.googleusercontent.com/bFWgXFaSlK5QveNDgUzT3RA9sd93T3uE1Jd6Gp0i9xFmZLW7hMSzjYqJein8bP9XFJH12p7AN1jz047aAOcmdoxI2_eigp61H0yXsNXxq-QY06oQBILAwVI-S53Oe5aISVZs0LSv

Variable importance plot

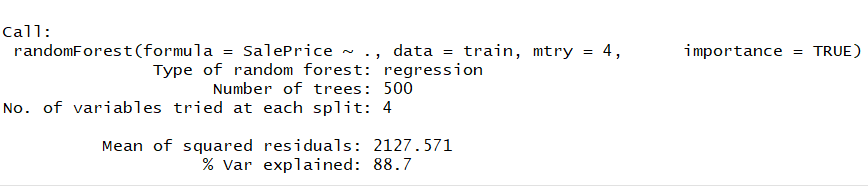


Random Forest Model 2



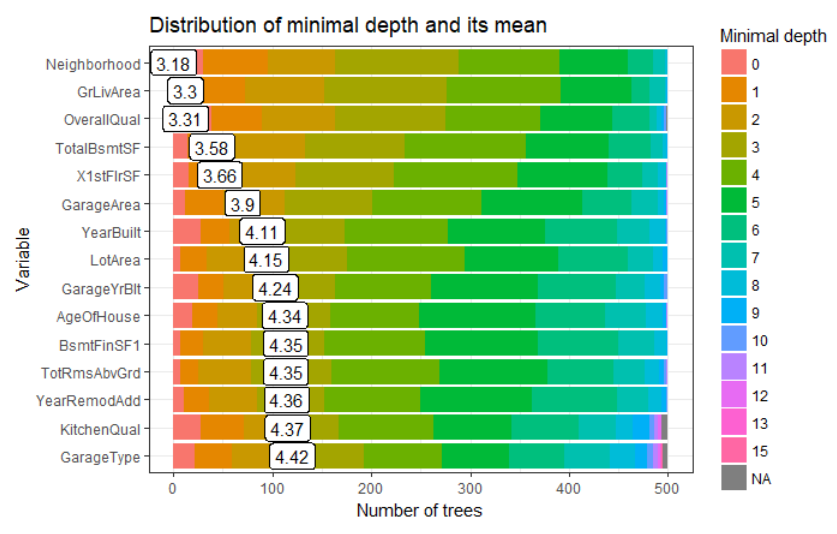


Random Forest Model 3

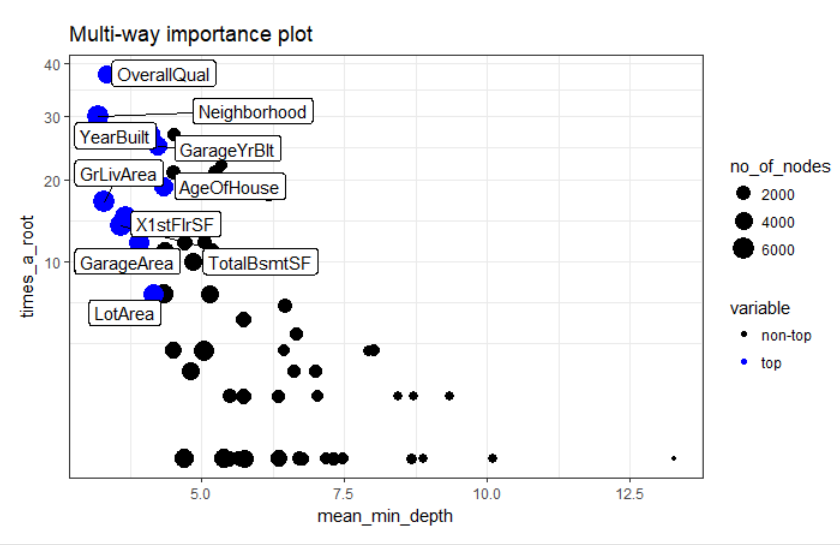


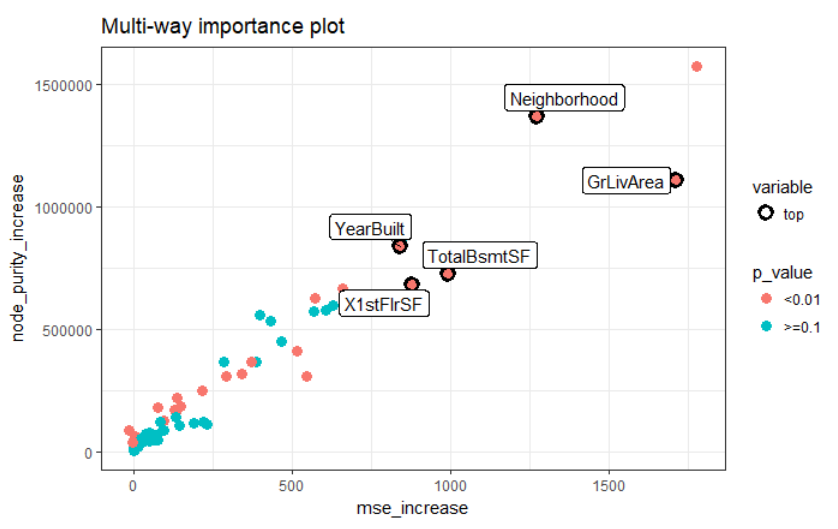
https://lh4.googleusercontent.com/KKQe3sRVbY0jSmmno4Ij-UoAakDBP1PNSCbdIjo_xanI9I6tHg_Xnm2fVIAjDirGAJzVtEpWkHe-1-2PlnIeAxZSidQMhB2Co9_1xy4g788xPAsYnKcc1iBC6krUQLPJgDiZHHVI

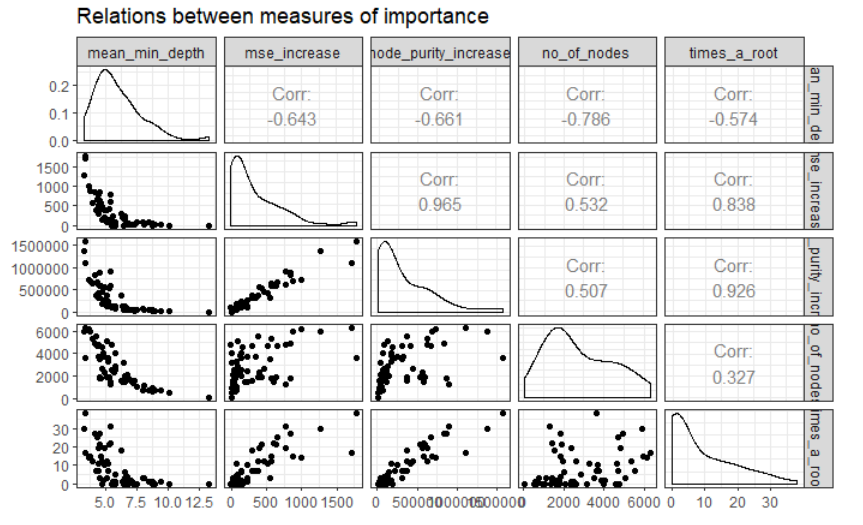
**Step 10 : Importance of Random Forest Model**

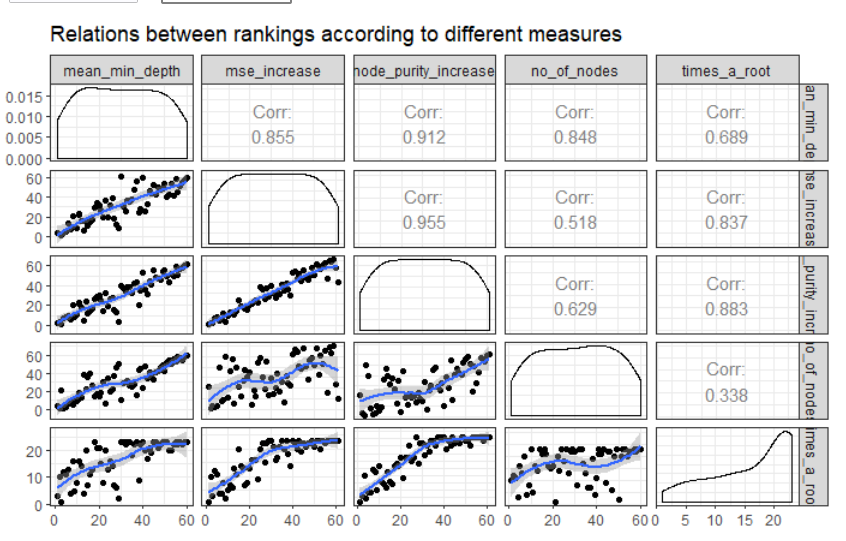


Plots created using Random forest explainer package further analyses the importance variable.



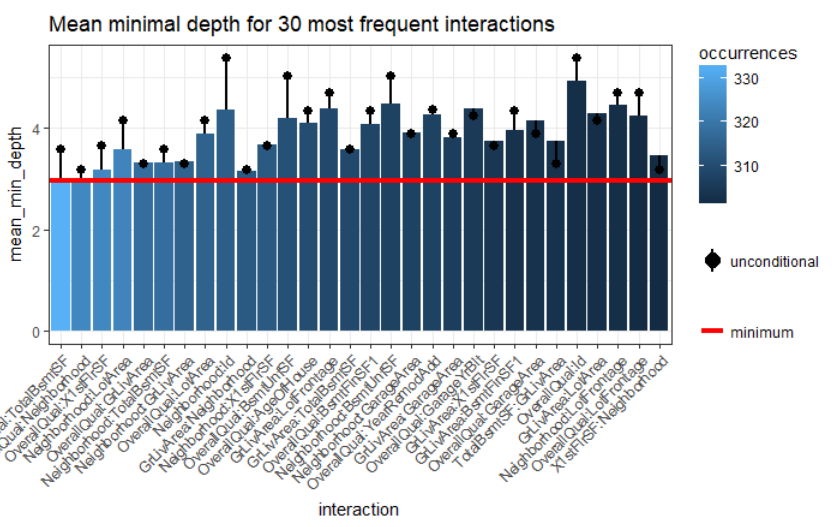


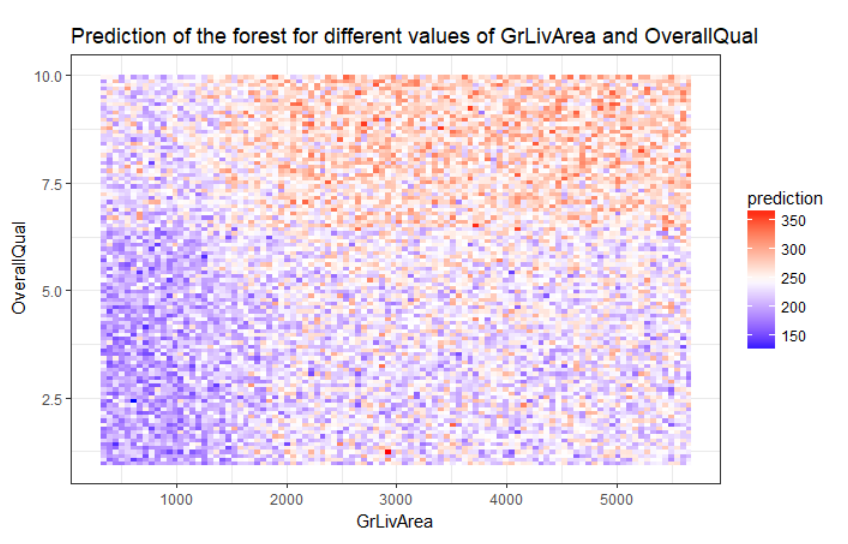




Important Variables

https://lh6.googleusercontent.com/yn-CAs1hsoBRU-8-mOs7KdQC-_Wb6XJuP7bdIDtdysjA3KezPUUcsCsa7ICFVrCDR6QJDEPWreDWab5FUPd3jfHEB80okcIEAnSi8MhJQDCOshahsFyTFJYl8fqYaCB8m42hwR_O





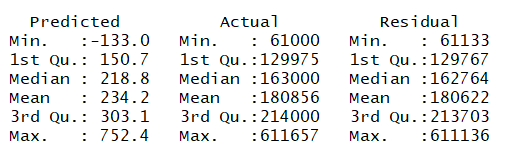
Sale price increases with Overall Quality and living area sqft increase.

**Step 11: Prediction Summary**

summary(train$SalePrice)

https://lh6.googleusercontent.com/YPjRjU_xmR0e9MKTkkKgyHCluIKgtrtdFldakkw7LpJK1D5UkvwAINgI4pRqKmwys4pFSrT6QMWp0iNgWWNEskJFgN1w6AQVMPDtEeUWNjUh_6qXYtXrbL1oXf-KqV24Mi9pWgG5

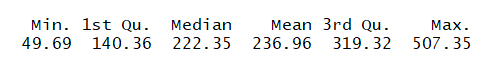
summary(linreg2\_pred)



summary(pred\_tree)

https://lh3.googleusercontent.com/ome7SvEY99GP5kBvRYGRJFjV1vNyV_pc67DzpPFKDX2zcP6ytEvppbDjTjSWzSya2j3WAYKgb-DxcO6qEKF5TgHRH32gLjrQq9ZYKSJ88PLMicJ9TBFThlPfcbJ179dX4CNX_BtM

summary (rf3.pred)



Summary comparison the random forest model predicts the price better than other models. Linear regression model prediction is the worst.

**Conclusion**

The dataset doesn’t include data like school rating, crime rate, asking price of house and how long it was in the market. Further analysis of data is required with more of these information’s in it to improve the predictability. The analysis proves the home’s sell in high price depends on overall quality of home, location and the square footage of the home.