STAT 515- Final Project Report

Instructor: Professor Daniel B. Carr

Name: Kacham Nikitha

G#: G01007820

Topic: House Price Predictions

1. Introduction

The Goal of the project titled “House price predictions” is to predict the price of a house based on various factors. The necessity of humans is to live in a house, an individual must buy or rent a house to survive in the society. One can choose the locality of the house by estimating the price whether he or she can afford or not. This major point has led me to work on this topic and answer an important question that is predicting the sale price of a house based on top most predictors or variables among the dataset. The purpose of this project is to visualize variety of graphics and models using the statistical analysis tool R.

2. Dataset Description

The dataset for the house price prediction has been extracted from ‘kaggle’ which is the famous platform for the predictive modeling and analytics competition. The dataset can be obtained from the link provided in the reference. The dataset consists of 1460 observations and 81 variables, which has really a huge number of predictors. The data description for the available dataset had the explanations for all the abbreviations of each column and its types. For instance, the column named ‘MSZoning’ identifies the general zoning classification of the sale which comprises of 8 levels such as A for Agriculture, C for commercial, FV represents Floating Village Residential, I for Industrial, RH for Residential High Density, RL for Residential Low Density, RP for Residential Low Density Park, RM for Residential Medium Density.

3. Data Preprocessing

This was the major task during the process, initially the dataset had 81 variables and 1460 observations. To check if there are any n/a values for each column I have used the function sum to know the count of n/a values out of which 17 variables had missing values. For example, the n/a count for the variable named ‘LotFrontage’ was 259. After getting the count of n/a values for each column I have saved all of them into the data frame and eliminated the columns having the missing values. The outcome after preprocessing the data had 64 predictors and 1460 observations. To be noted there were many outliers in the dataset, due to the time I could

not work more on eliminating the outliers. Through this process, I gained a very good knowledge on how to process the data.

Figure 1: After preprocessing the dataset

MSSubCla	MSZoning	LotArea	Street	Alley	LotShape	LandConto
60	RL	8450	Pave	NA	Reg	Lvl
20	RL	9600	Pave	NA	Reg	Lvl
60	RL	11250	Pave	NA	IR1	Lvl
70	RL	9550	Pave	NA	IR1	Lvl
60	RL	14260	Pave	NA	IR1	Lvl
50	RL	14115	Pave	NA	IR1	Lvl
20	RL	10084	Pave	NA	Reg	Lvl
60	RL	10382	Pave	NA	IR1	Lvl
50	RM	6120	Pave	NA	Reg	Lvl
190	RL	7420	Pave	NA	Reg	Lvl
20	RL	11200	Pave	NA	Reg	Lvl
60	RL	11924	Pave	NA	IR1	Lvl
20	RL	12968	Pave	NA	IR2	Lvl

In the above screenshot the variable named Alley has ‘NA’ values that doesn’t mean it contains n/a values it basically stands for ‘No Alley Access’. Each column has its own features, Street defines the type of road access to property it has two levels paved and gravel. The LotShape defines the general shape of property it has four levels ‘Reg’ stands for Regular, ‘IR1’ stands for Slightly irregular, ‘IR2’ represents Moderately Irregular and ‘IR3’ represents Irregular lot shape. The predictor Land contour defines the Flatness of the property it has four levels ‘Lvl’ stands for Near Flat/Level, ‘Bnk’ stands for Banked - Quick and significant rise from street grade to building, ‘HLS’ stands for Hillside - Significant slope from side to side, ‘Low’ stands for Depression.

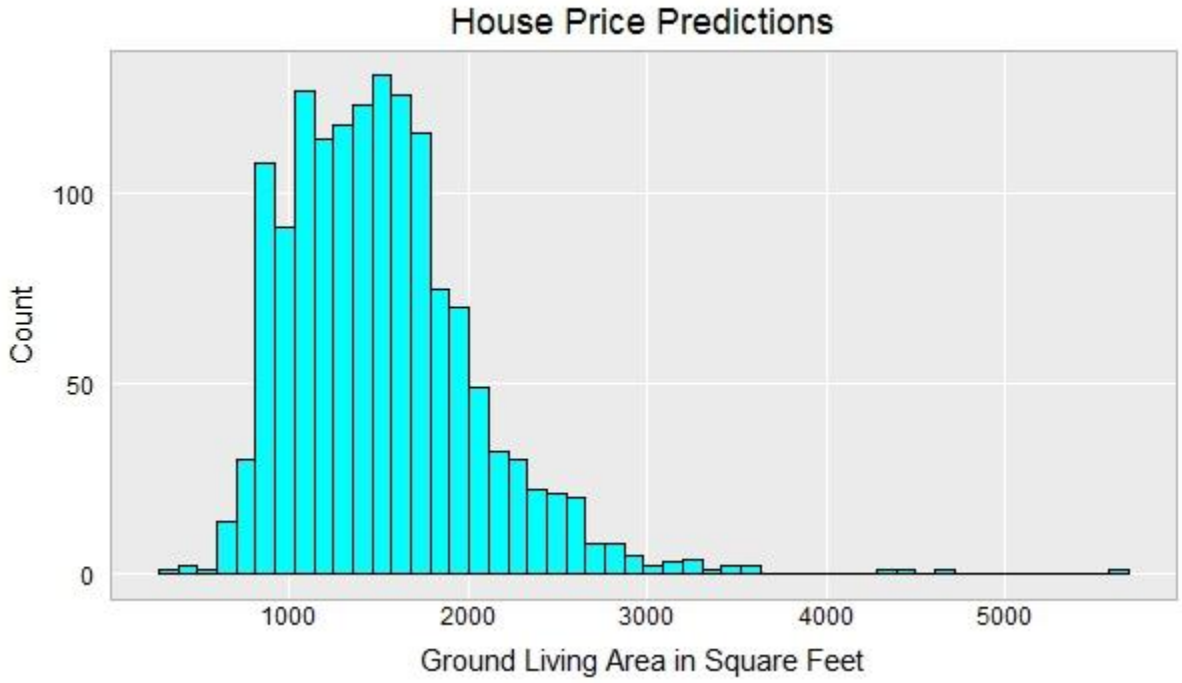
4. Visualizations

The data preprocessing was done successfully, to ensure the relationship between each variable I have analyzed the dataset and tried to implement graphs using the concepts thought in class. For the visualizations section I have used variety of simple graphs to complex graphs which includes histogram, juxtaposed scatter plot, kernel density plot, superposed density plot and two GGplots using the facet wrap.

A) Histogram

I started off with a basic plot, the very first graph in the visualizations is histogram to predict the sale price for a house based on the Ground living area in square feet. This is one observation among 64 other predictors. The histogram consists of rectangles, in the below graph the area is proportional to the Ground living area, we can interpret that the most preferable living area lies between 1000 to 2000 square feet.

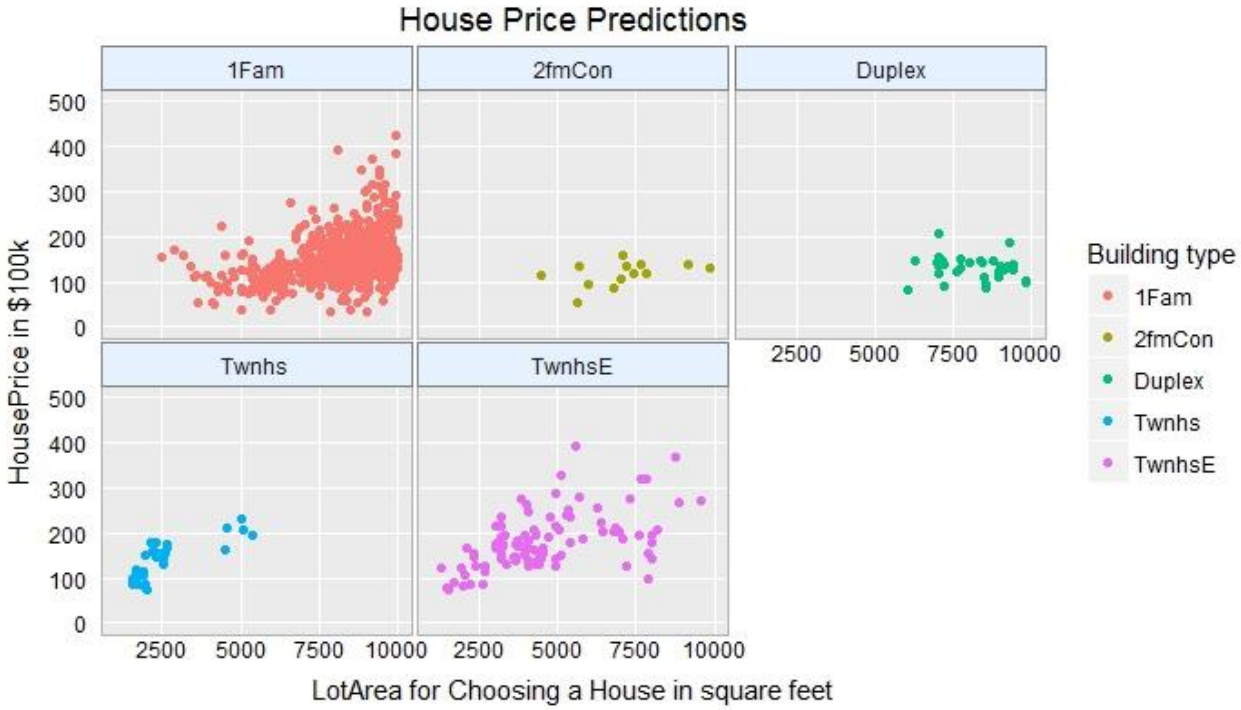
Figure 2: Histogram on ground living area



B) Juxtaposed scatter plot

The second variety of plot is juxtaposed scatter plot, where I have plotted the graph to show the relationship between the Lot Area and the sale price with respective to the building type. The abbreviations for the types of building are '1Fam' belongs to the single family, '2FamCon' represents two family conversion, 'Duplex' represents the duplex building, 'Twnhs' represents the townhouse end unit, 'TwnhsE' represents the town house inside unit. Each color dot identifies a type of building, we can observe from the plot that people are most likely to prefer the '1Fam' building type among the others. I have used the ggplot2 library and the facet_wrap to build the plot. I felt very interesting in finding unknown patterns by plotting various types og graphs.

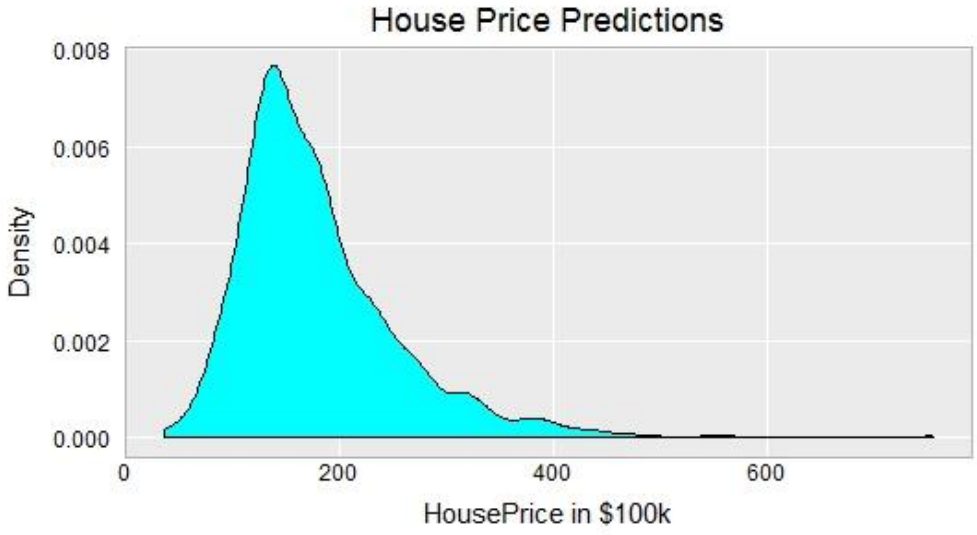
Figure 3: Juxtaposed scatter plot- Lot Area vs sale price with respective to Building type



C) Kernel Density plot

The third plot in the visualizations section is kernel density which is the other form of histogram, the density plot below shows us the distribution of saleprice in $100k over a continuous interval. I have plotted the graph by considering sale price on x axis, it gives a smooth curve which clearly interprets that the people are most likely to own a house which costs between 100k to 200k.

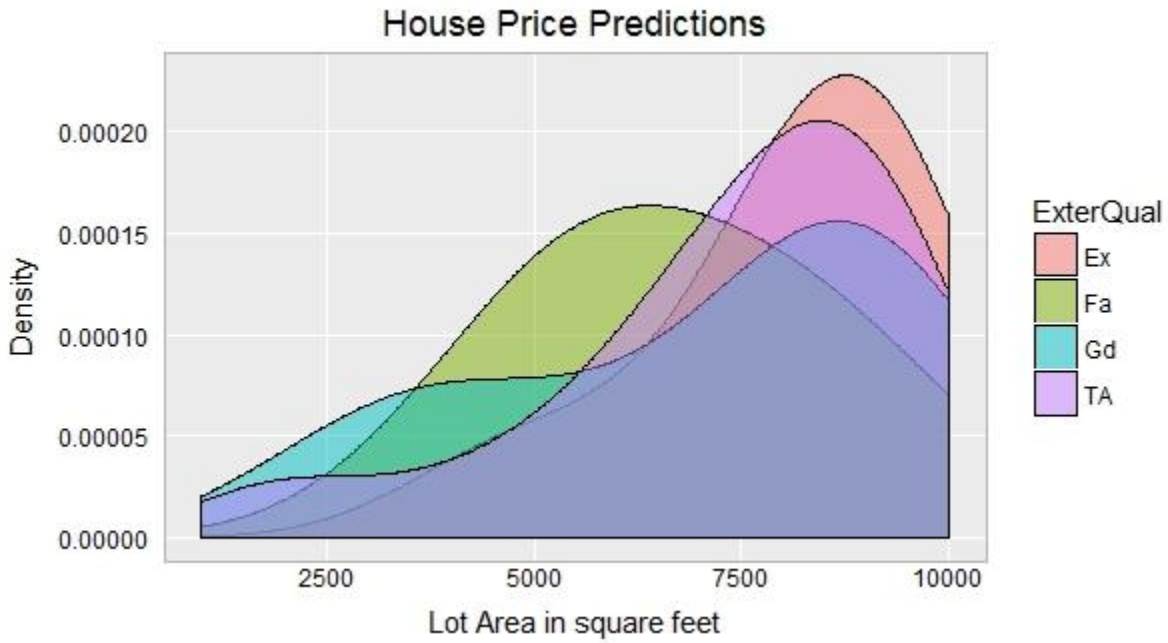
Figure 4: Kernel density plot – sale price predictions



D) Superposed Density plot

The fourth graph is the superposed density plot, after plotting a kernel density plot I got an idea to plot more complex density plot which depicts more information and clear understanding. To know the exact property of the lot area, I have decided to consider Lot area on the x axis and observation as external quality of the lot area. The legends indicate the quality levels of the lot area, Ex stands for excellent, Fa stands for fair, Gd for good and TA for average. From the graph, we can say that the excellent quality lot area lies between 7500 to 10000 square feet.

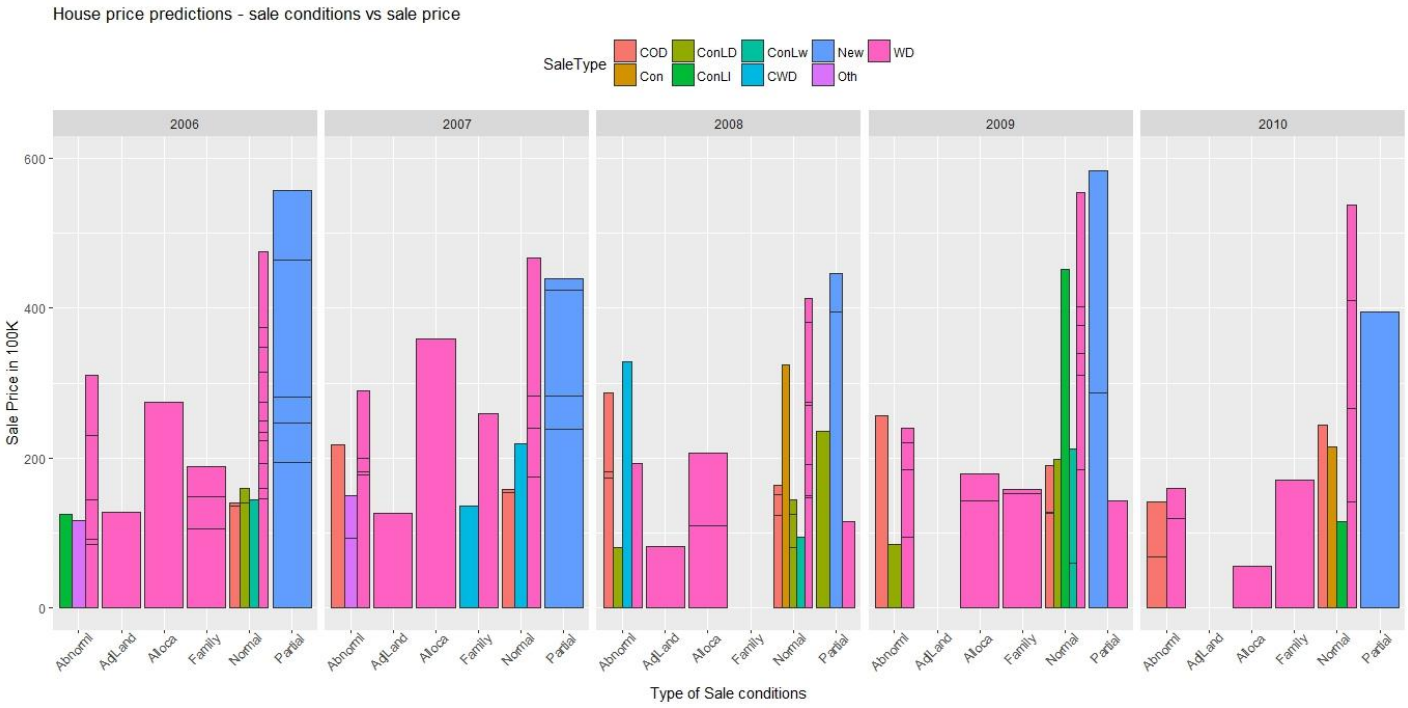
Figure 5: Superposed density plot- Lot Area vs External quality



E) First GGplot using Facet grid

After plotting the above graphs, I wanted to improve my observations to one level high by considering two or more variables. Later I got an idea to use the facet wrap option to combine the results of various graphs using the ggplot package and features in it. I had really a tough time in plotting this graph, the graph shows all the effort. I have considered the types of sale conditions from the year 2006 to year 2010 which defines the year built, one of the predictor from the dataset on the x-axis, sale price on the y axis and I used fill feature to add the fourth variable to be the sale type. I was really satisfied with the interesting outcome for this plot. By considering the observation for the year 2007 we can depict that the people who choose abnormal type of sale condition are likely to consider three types of sale which are COD- Court officer deed/estate, WD – warranty deed and other additional factors. Similarly, with the other produced interesting patterns.

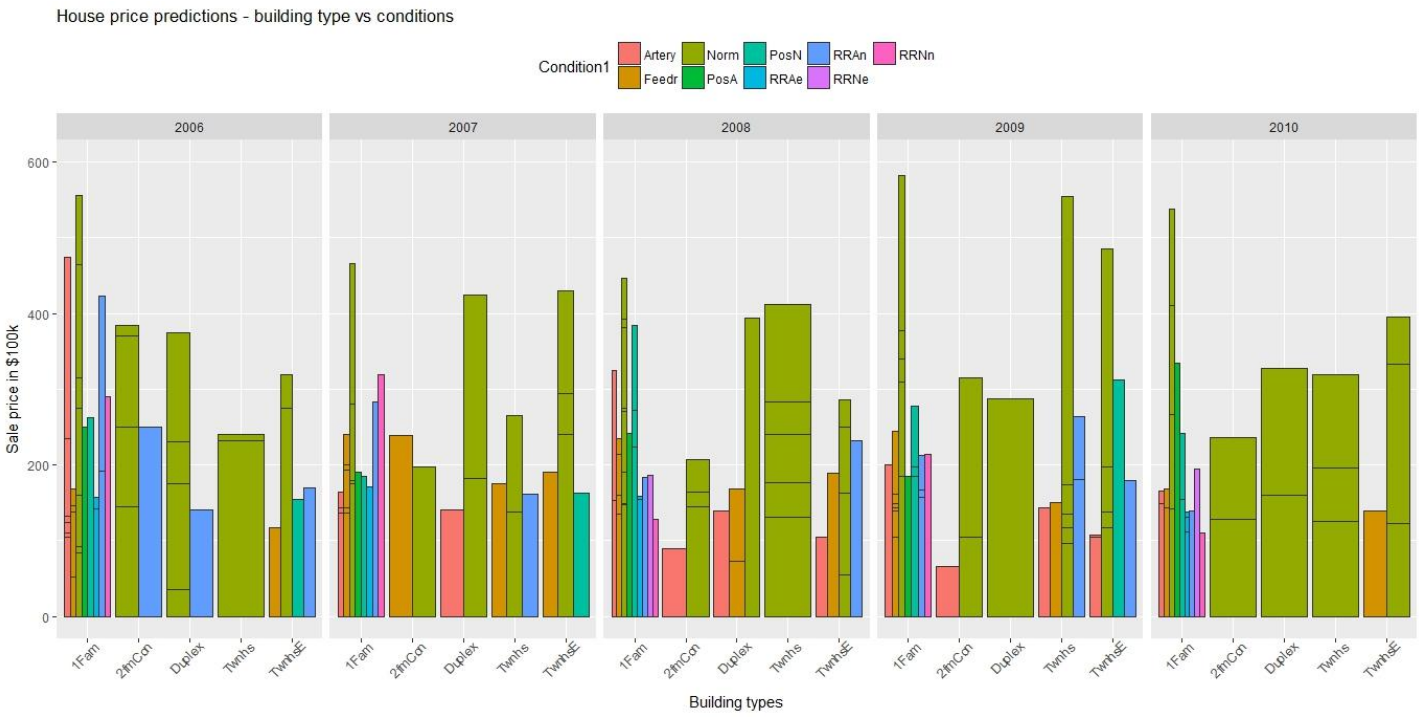
Figure 6: GGplot sale conditions vs sale price with respective to sale type



F) Second GGplot using Facet Grid

Like the figure 6, I have plotted one more ggplot by considering building type, condition1, lot area and sale price for finding interesting observations among the four predictors. From the figure 7 by considering the year 2007 the building type 2FmCon which stands for two family conversion are most likely to consider their house to satisfy the condition 1 Feedr which stands for adjacent to feeder street and the other condition to be normal.

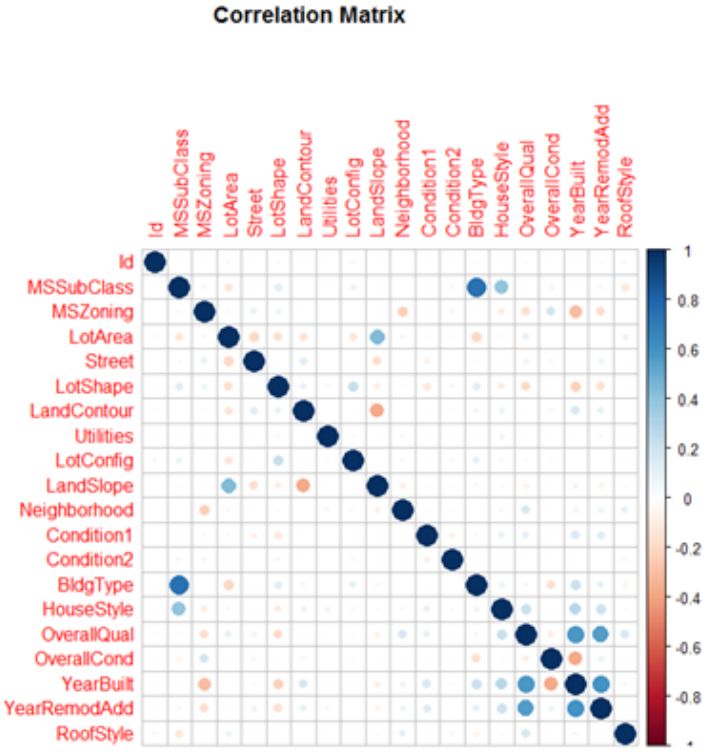
Figure 7: ggplot- building type vs sale price with respective to condition 1



G) Correlation plot

To know the relationship between various predictors I have opted to plot a correlation matrix using the corrplot package which clearly shows the correlation between one another. The package is very flexible it has many inbuilt features including choosing color, text labels, color labels, layout, etc. for the below plot I have used circle visualization method to represent the correlation between two variables.

Figure 8: Correlation Matrix



5. Modeling

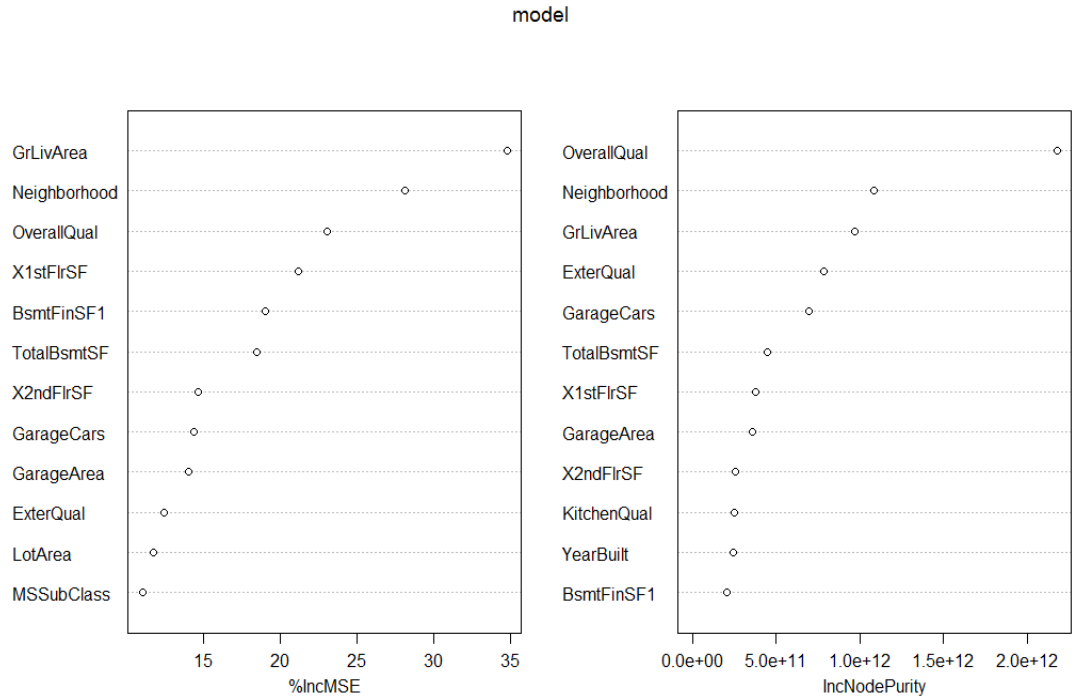
A) Random Forest Regression

To do some modeling on the data I have chosen to perform random forest regression first which I have produced the variable importance plot and error rate plot.

Variable importance plot

The variable importance plot is done using the VarImpPlot function. The plot is produced using the full dataset, I have given the number of predictors to be 12. The top variable is ground living area which gives us 35% accurate results, node purity is measured by Gini Index which is the difference between RSS before and after the split on that variable.

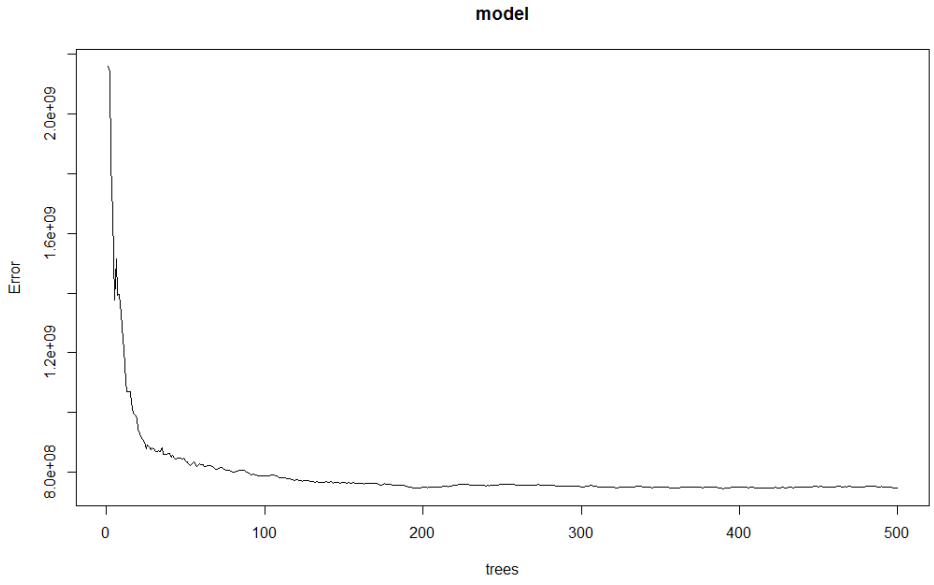
Figure 9: Variable importance matrix for 12 variables



Error rate plot

The error rate plot is nothing but the random forest model for the full dataset, I have considered number of tress to be 500 from the graph we can say that the more number of trees the more accurate is the model.

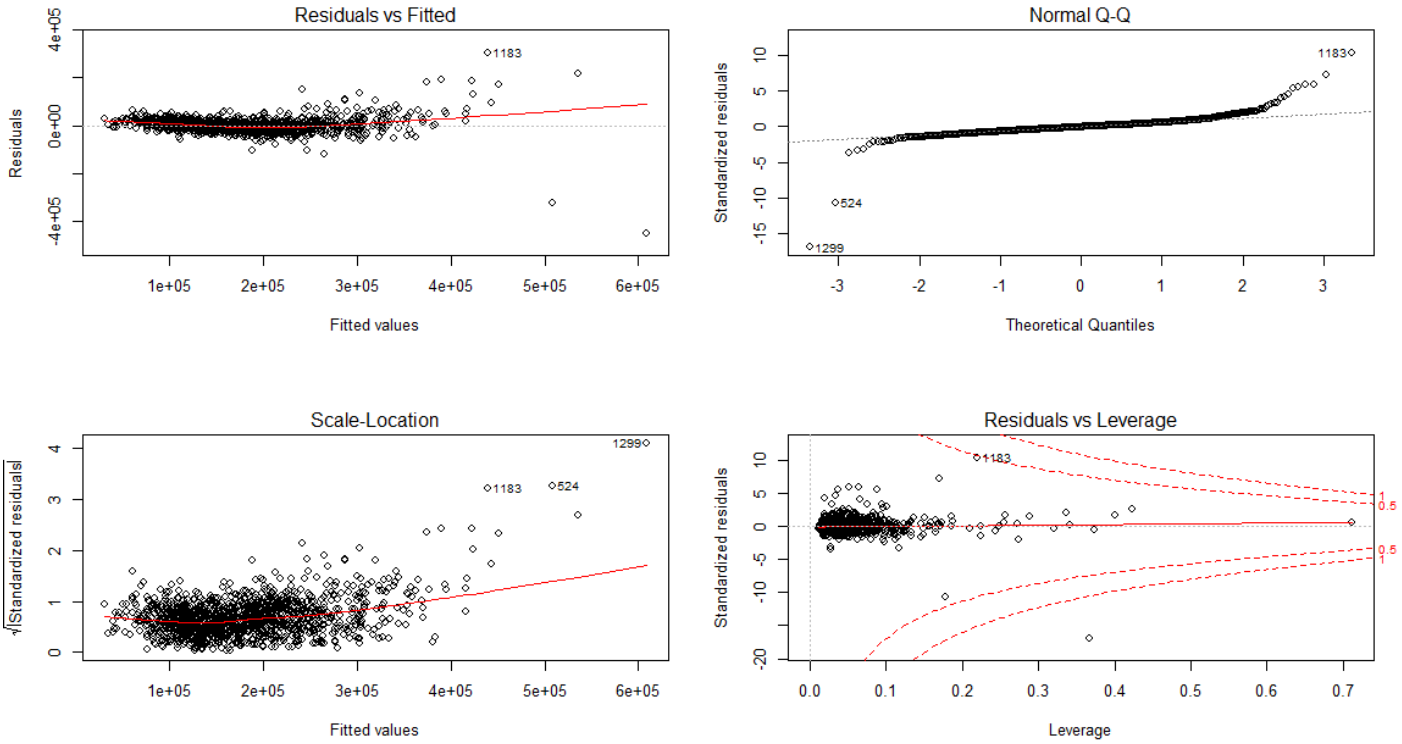
Figure 10: Error rate plot



B) Linear Regression

The second method of modeling was using the linear regression which is a statistical method to build a linear model and describes a continuous response variable as a function for one or more predictor variables with understanding the behavior of complex systems. I have divided the dataset into testing and training, first 200 observations as testing data and rest for the training dataset and applied linear model regression on the training dataset. The below four plots are produced for the linear regression. The residual vs fitted is the most frequent plot it is a scatter plot of residuals on the y axis and fitted values on the x axis, shows if the residuals have non-linearity. If we find equally spread residuals around a horizontal line without any patterns that indicates we don't have any non-linear relationships. 'due to the outliers in the dataset' the graph produced is not completely spread. The normal q-q plot shows that the residuals are normally distributed, if the residuals are on the straight like that means it's a good indication for the model, we can see a thick and tail forming for the q-q- plot. The scale location plot is like the residual vs fitted plot, the more the spread the more the model fits to the data. The last plot residual vs leverage finds the influential cases if any.

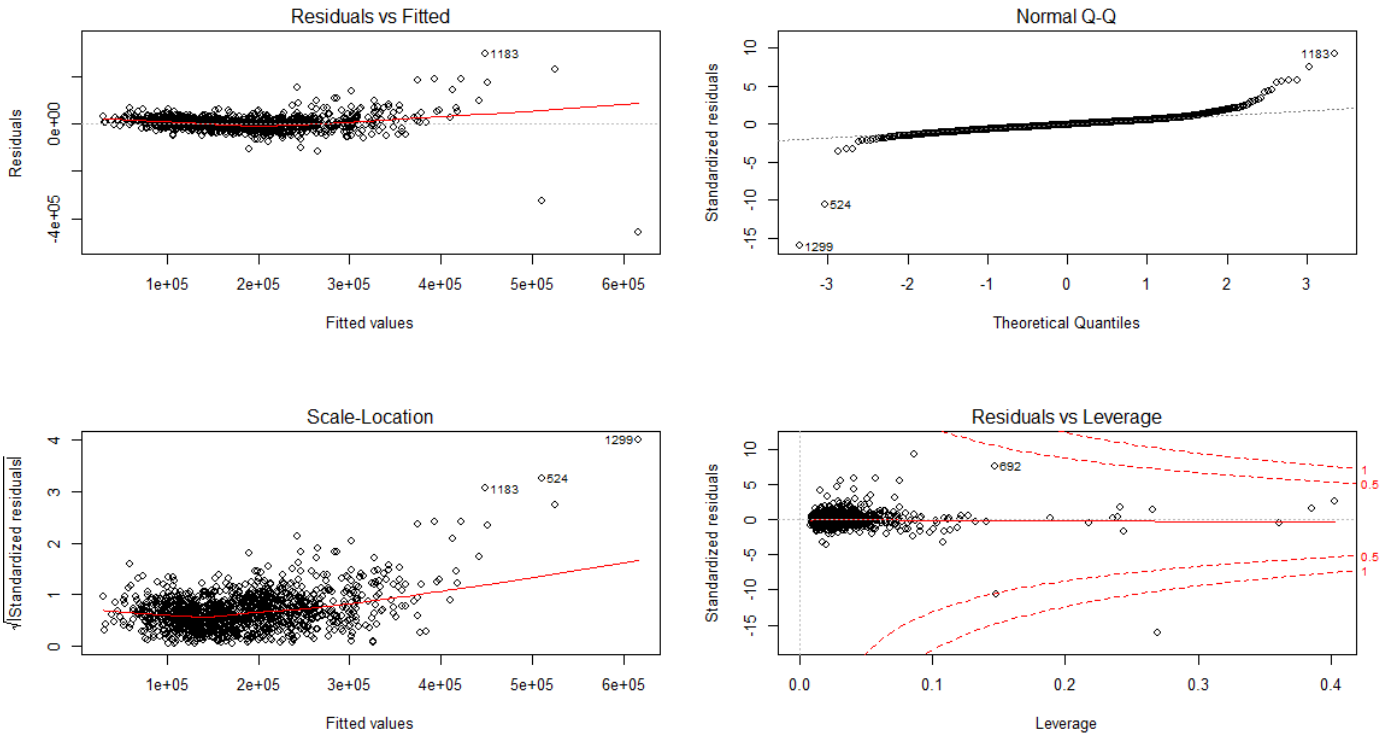
Figure 11: Linear regressions plots for Fit model



C)Step wise Regression

To check the how best the model fits to the data I have used another method which is step wise regression with the backward direction including 100 steps. The below are the four plots produced after modeling. I faced a lot of trouble in figuring out the outliers, I would definitely work on the outliers and improve my model and make the plots more appropriate with more widely spread residuals.

Figure 12: Step wise regression plots for Upfit model



Comparison between the Fit and Upfit model

We must pay a great attention to regression results such as slope coefficients, p value, multiple R-squared value, adjusted R-squared value to depict that how well a model can represent the data. below table represents the summary of both the models, by comparing the adjusted R-squared value we can say that Step regression model would best suit to the given data.

Models	Multiple R- squared value	Adjusted R-squared value
Regression model – FIT	0.8331	0.8245
Step regression model - UPFIT	0.8311	0.826

6. Observations

Performing ANOVA to find the best model

After observing the results produced by both the models I have decided to perform an Analysis of variance (ANOVA) test to confirm which model would be the best among the two. “The base case is the one-way ANOVA which is an extension of two-sample t test for independent groups covering situations where there are more than two groups being compared.” (Ralph, 2010).

By comparing the results of Anova test on both the models, we can interpret that the model 2 which is the step wise regression would best suit for the data.

Figure 13: Results for Anova

```
> anova(fit,upfit)
Analysis of Variance Table

Model 1: SalePrice ~ Id + MSSubClass + MSZoning + LotArea + Street + LotShape +
    LandContour + Utilities + LotConfig + LandSlope + Neighborhood +
    Condition1 + Condition2 + BldgType + HouseStyle + OverallQual +
    OverallCond + YearBuilt + YearRemodAdd + RoofStyle + RoofMatl +
    Exterior1st + Exterior2nd + MasVnrType + MasVnrArea + ExterQual +
    ExterCond + Foundation + BsmtFinSF1 + BsmtFinSF2 + BsmtUnfSF +
    TotalBsmtSF + Heating + HeatingQC + CentralAir + X1stFlrSF +
    X2ndFlrSF + LowQualFinSF + GrLivArea + BsmtFullBath + BsmtHalfBath +
    FullBath + HalfBath + BedroomAbvGr + KitchenAbvGr + KitchenQual +
    TotRmsAbvGrd + Functional + Fireplaces + GarageCars + GarageArea +
    PavedDrive + WoodDeckSF + OpenPorchSF + EnclosedPorch + X3SsnPorch +
    ScreenPorch + PoolArea + MiscVal + MoSold + YrSold + SaleType +
    SaleCondition
Model 2: SalePrice ~ MSSubClass + LotArea + Street + LotShape + LandContour +
    Utilities + LandSlope + Neighborhood + Condition2 + BldgType +
    HouseStyle + OverallQual + OverallCond + YearBuilt + RoofStyle +
    RoofMatl + Exterior1st + MasVnrType + MasVnrArea + ExterQual +
    BsmtFinSF1 + HeatingQC + X1stFlrSF + X2ndFlrSF + BsmtFullBath +
    FullBath + BedroomAbvGr + KitchenAbvGr + KitchenQual + TotRmsAbvGrd +
    Functional + Fireplaces + GarageCars + WoodDeckSF + ScreenPorch +
    YrSold + SaleCondition
  Res.Df        RSS  Df   Sum of Sq      F Pr(>F)
1   1190 1.3272e+12
2   1214 1.3428e+12 -24 -1.5603e+10 0.5829 0.9458
```

Performing Root mean square deviation

For further analysis, I have used the Caret library to produce RMSE values for the three model that is linear model, step wise regression model and random forest model. Based on the non-missing data fed model which the linear regression model has 2.57 RMSE and 88% R-squared value, the step wise regression has 2.52 RMSE and 89% R-squared value this is best because of the backward regression. The random forest has 1.02 RMSE and 98% R-squared value, this is low because the random forest is done on full dataset. Finally we can conclude by saying that among the three models random forest would be the best suitable model for the data.

Figure 14: Results for RMSE

```
> postResample(pred1,actual) # Based on Non Missing Data fed Model
        RMSE     Rsquared
2.578642e+04 8.864554e-01
> postResample(pred2,actual) # Best Because of Backward Regression
        RMSE     Rsquared
2.523627e+04 8.913083e-01
> postResample(pred3,actual)  # low because Random forest is done on full dataset
        RMSE     Rsquared
1.023402e+04 9.860591e-01
```

7. Conclusion

The selection, exploration, visualization and modeling for the data on the house price predictions was successful. Starting from the data description to finding the appropriate model for the data using various methods was accomplished. During the process, I have learnt a lot of new things, I achieved a good knowledge on how to deal with the data preprocessing, data exploration, data modeling. Among all the steps I spent lot of time on eliminating the missing values, initially the dataset had 81 variables after removing the n/a values 17 columns were deleted and the final dataset had 64 variables and 1460 observations. The aim of the project was observed based on various factors affecting the sale price to own a house by visualizing with variety of graphs. Modeling the dataset with different methods produced interesting patterns. Anova testing was done to choose the best model among the linear and step wise regression models. Random forest had the highest R- Squared value and lowest RMSE value.

References

1. Kaggle: House price advanced regression techniques data

Accessible link: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

2. R- bloggers: one way analysis of variance - anova

Accessible link: https://www.r-bloggers.com/one-way-analysis-of-variance-anova/

3. Visualizations on ggplot2

Accessible link: http://www.datacarpentry.org/R-ecology-lesson/04-visualization-ggplot2.html

4. simple linear regression

Accessible link: https://www.r-bloggers.com/simple-linear-regression-2/