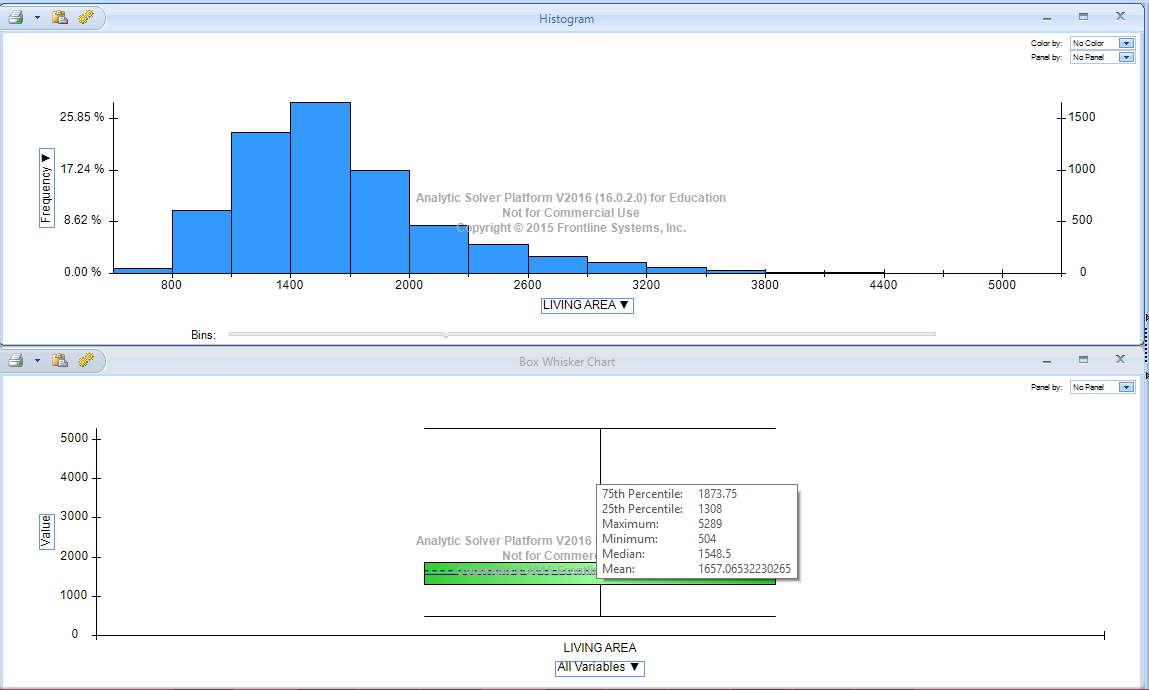
1. **West Roxbury**

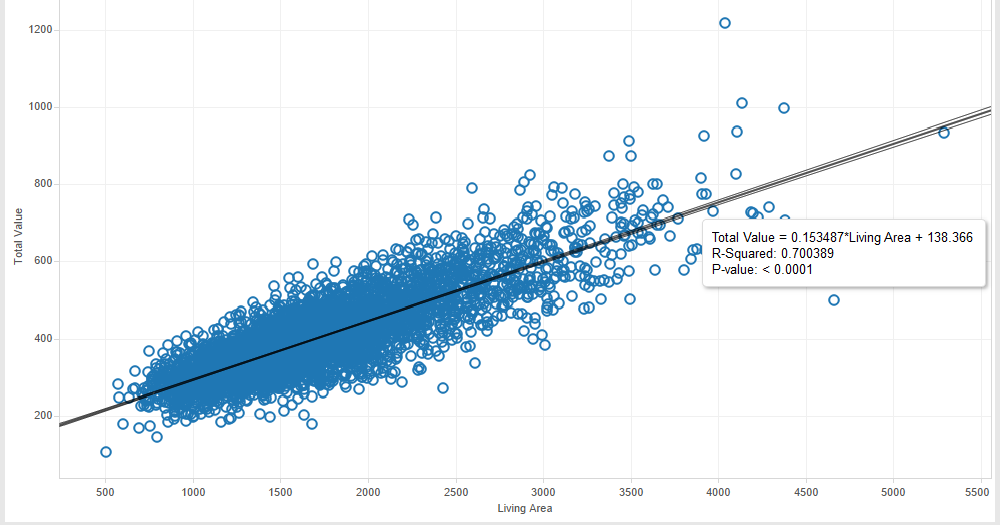
Problem: West Roxbury is a dataset about the houses in the West Roxbury neighborhood and we need to compute the expected prices of a home in future from this given dataset.

**1.1 Exploratory data analysis using TABLEAU and XLMiner**

Tool Used: Tableau and Excel

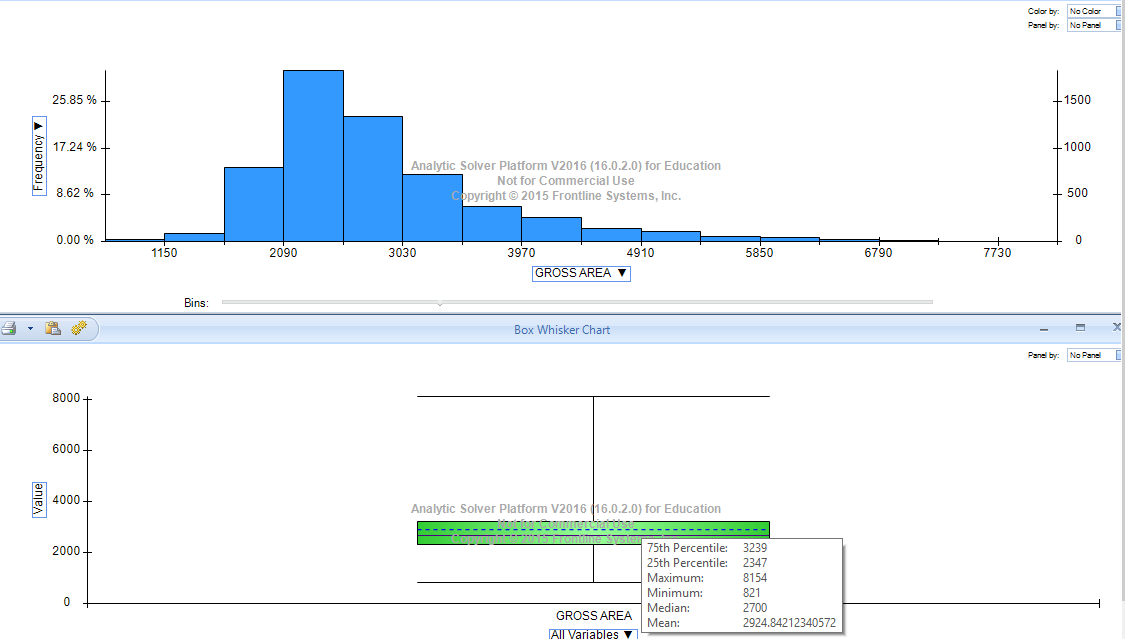
Living Area:



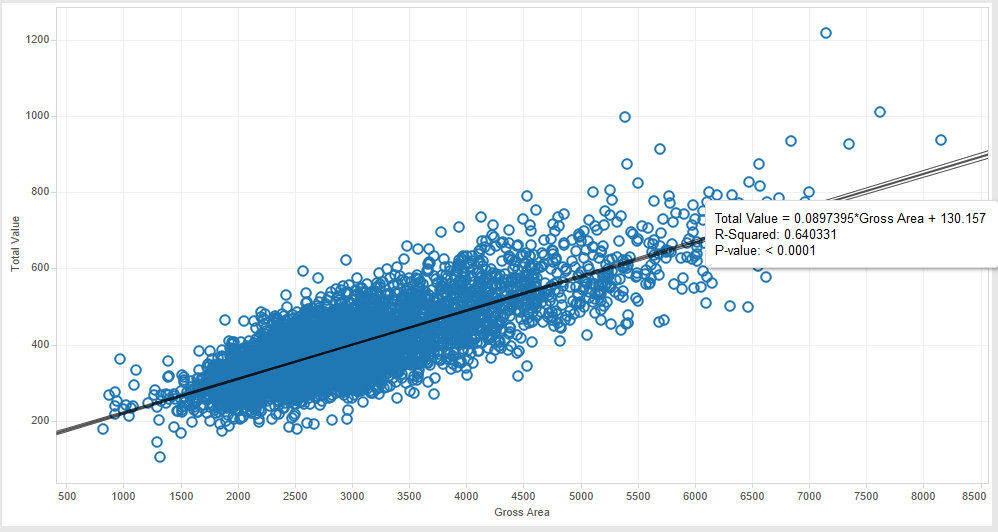
Scatter Plot in Tableau for Living Area

This Chart shows how the total price of the house depends on the Living area. The more the Living area of the house the price is higher.

Gross Area:

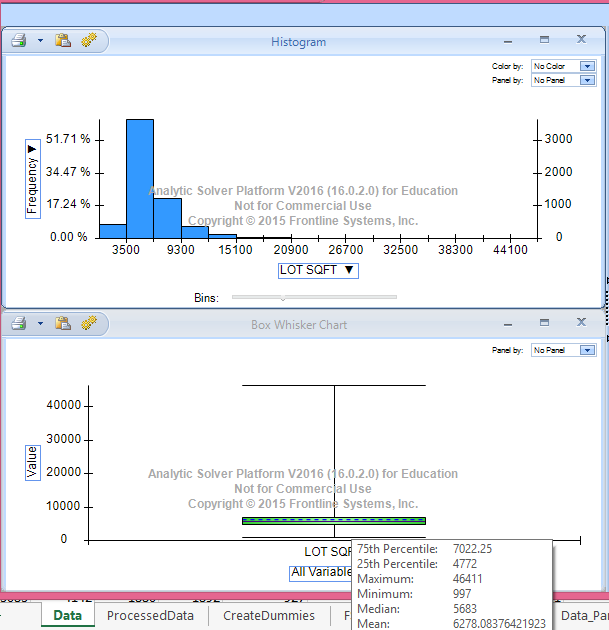


Scatter Plot in Tableau for Gross Area:

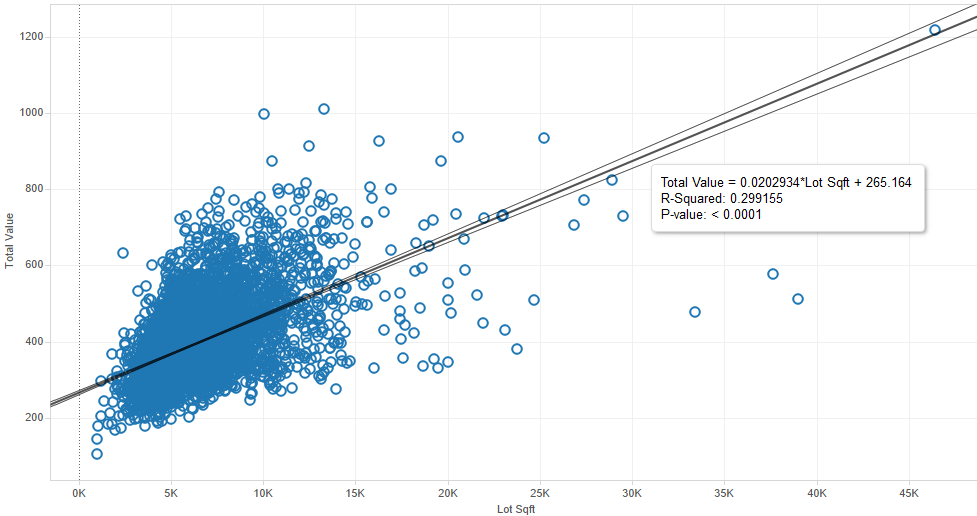


This Chart shows how the total price of the house depends on the Gross area. The more the gross area of the house the price is higher.

Lot Sqft:



Scatter Plot in Tableau for Lot Sqft

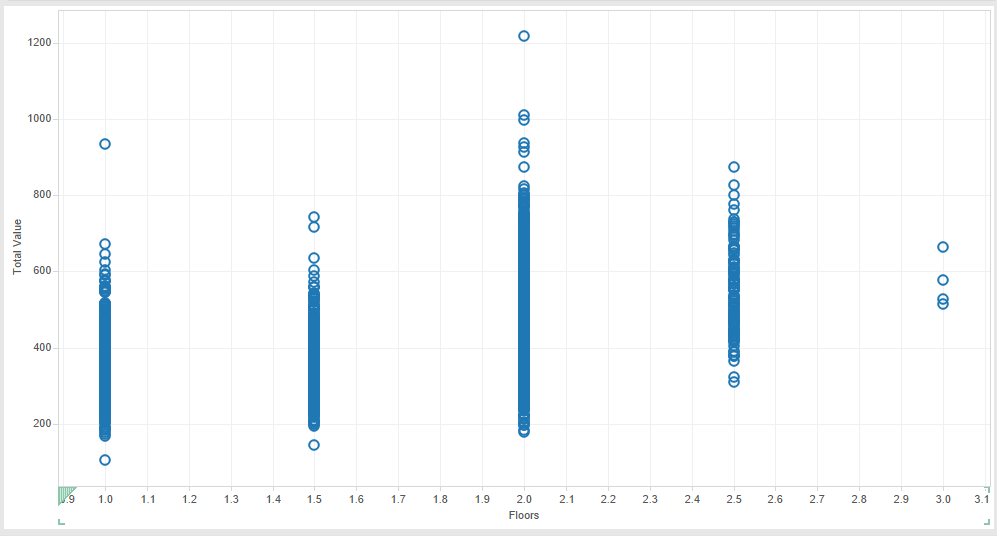


This chart shows how the Lot Sqft effect the pricing of the house. There are a lot of houses of 1000-15K sqft and the are priced between 100$ to 800$. As the size of the Lot increases the price increases too.

Floors:

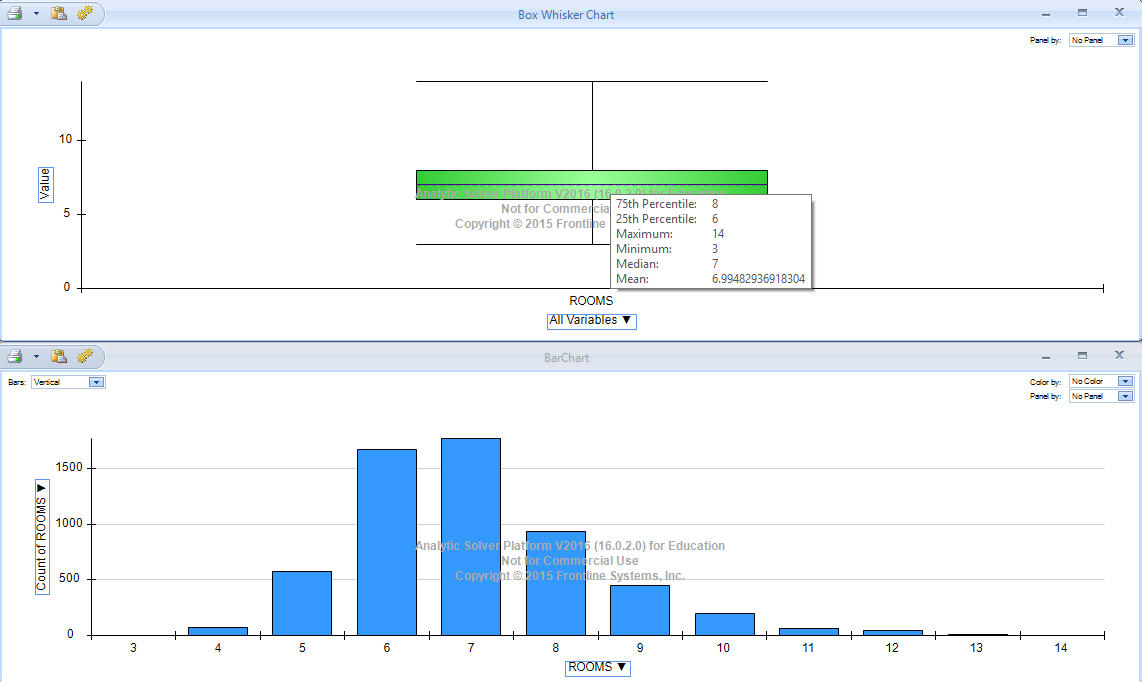


Scatter Plot on Tableau for Floor:

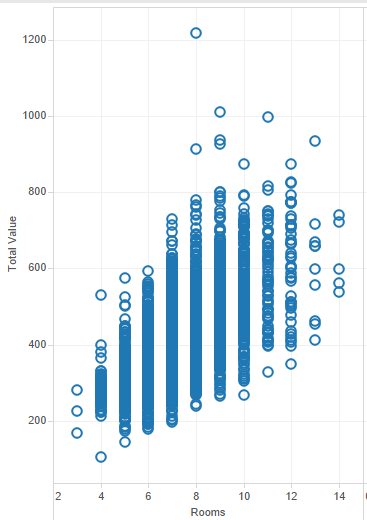


A very few houses in this dataset have 3 floors less than 1%. So this is an outlier.

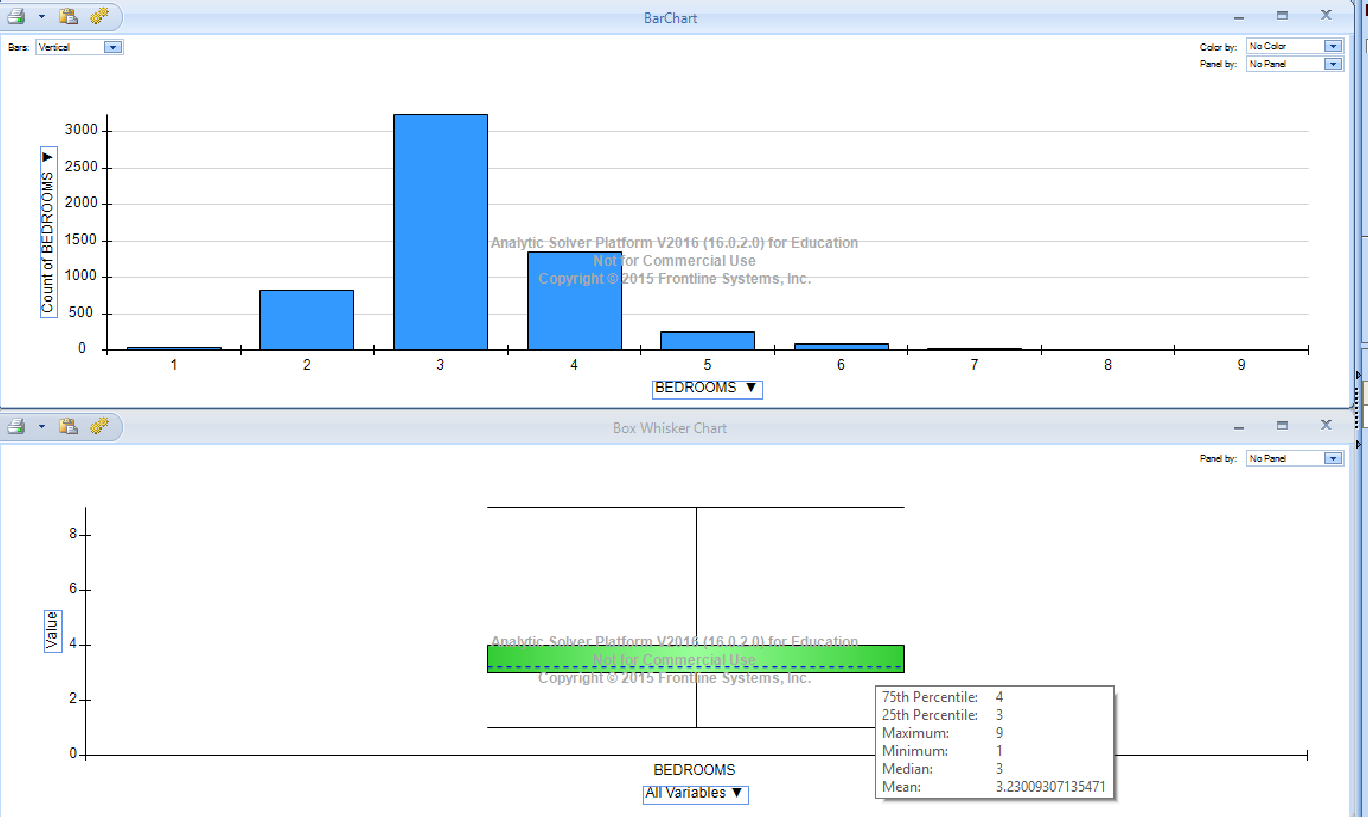
Rooms:



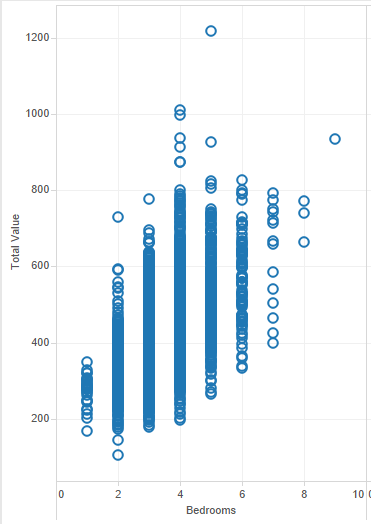
Scatter Plot on tableau for Room:



We can observe here that very few houses have 14 rooms. Most of the houses in this neighborhood has around 3-8 rooms in total.

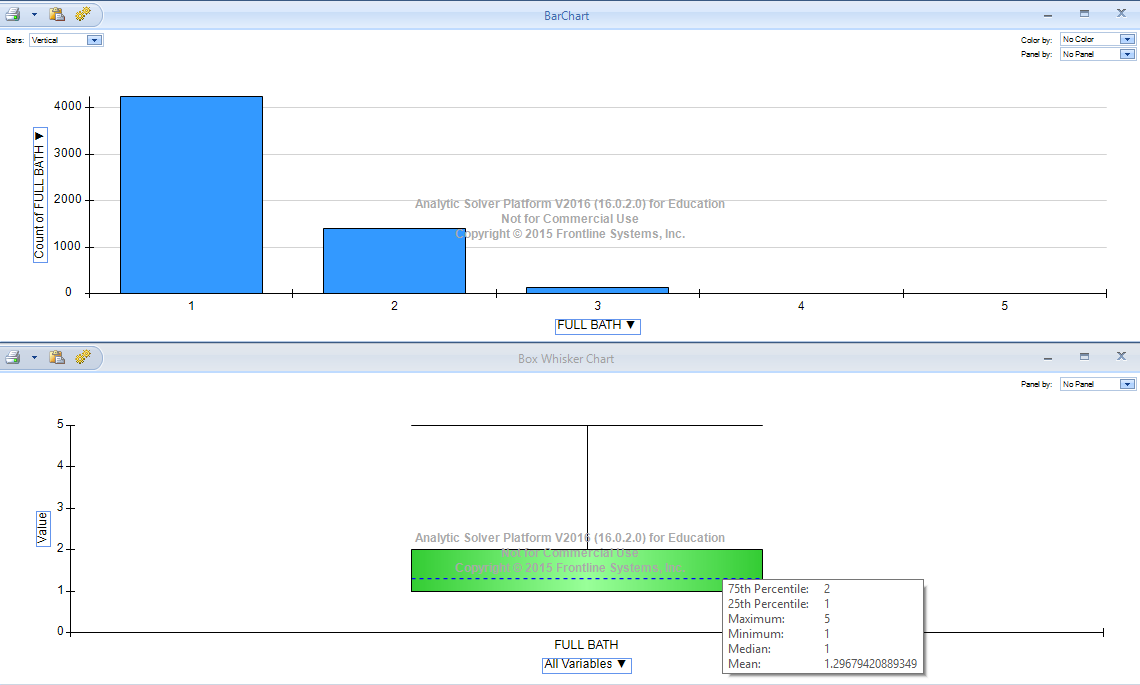
Bedrooms: 

Scatter Plot in Tableau for Bedrooms:

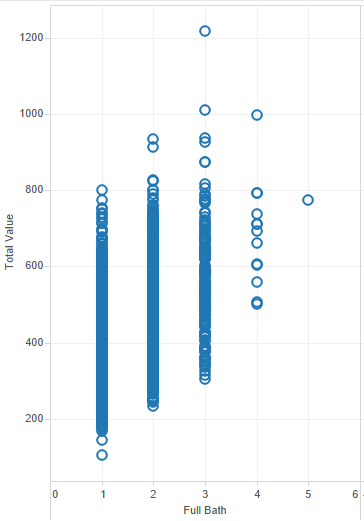


There is just one house with 9 bedrooms in this dataset. Also houses with 8 bedrooms are very less relatively.

Full Bath

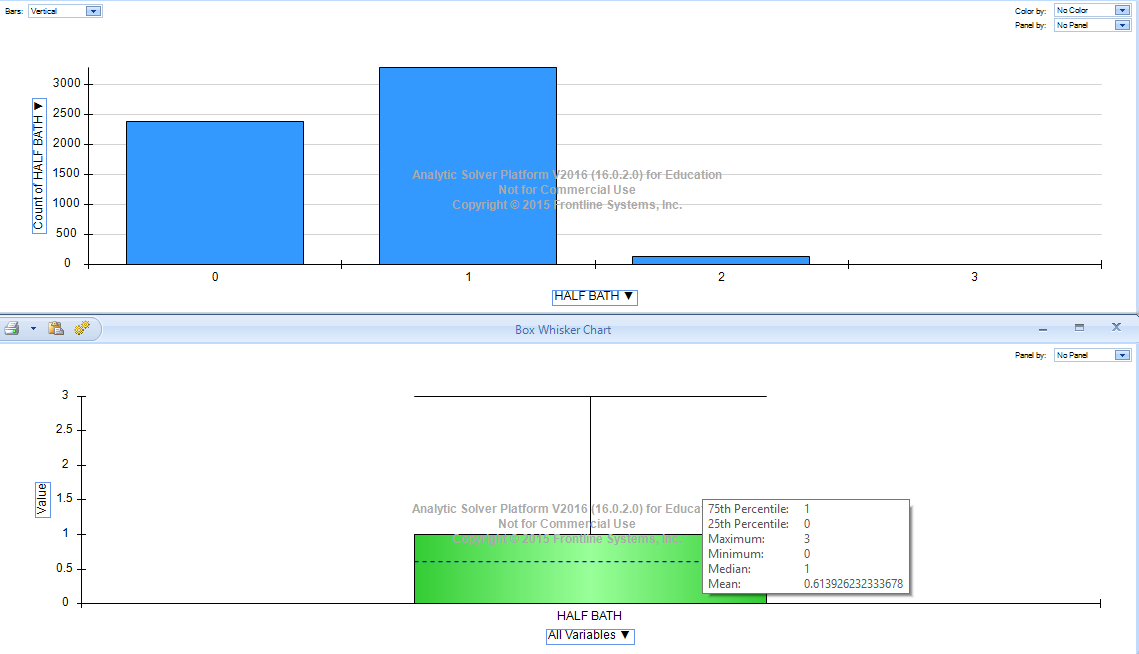


Scatter Plot in Tableau for Full Bath:

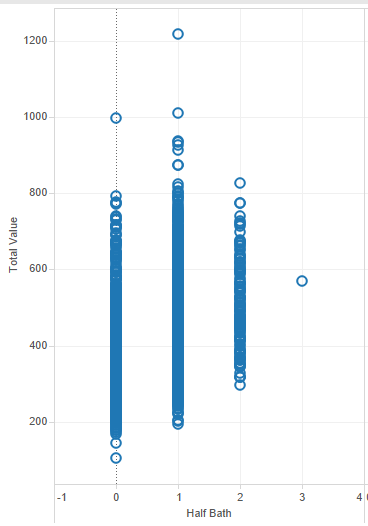


Here we can see very few houses have 4 Full Bath rooms and only 1 of them has 5 full baths. Also there are many houses with 1, 2 and 3 full baths. Houses with 3 full baths are priced a little higher.

Half bath

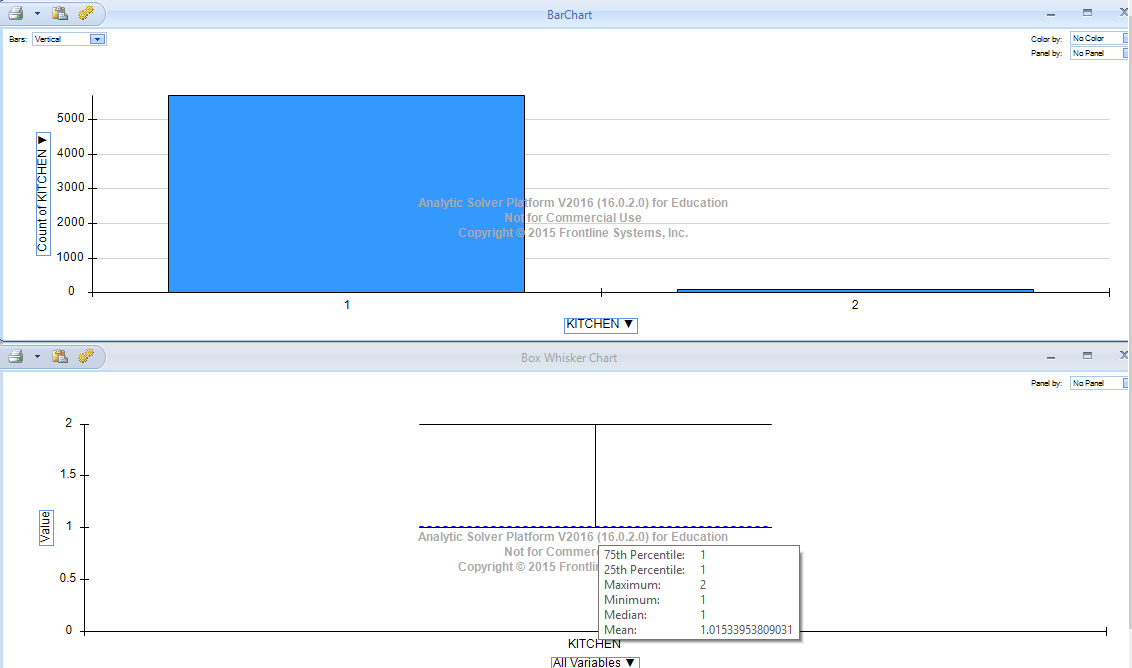


Scatter Plot in Tableau for Half bath

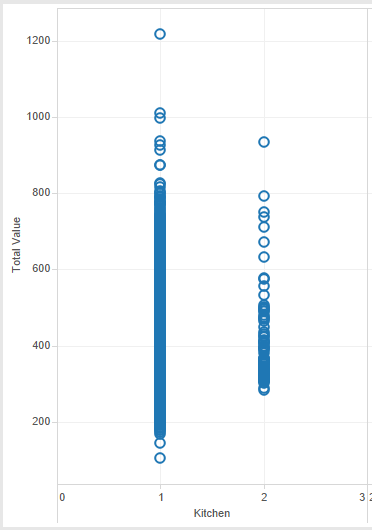


From looking at this we can find that there is only 1 data point which has 3 half baths, this may be an outlier.

Kitchen:

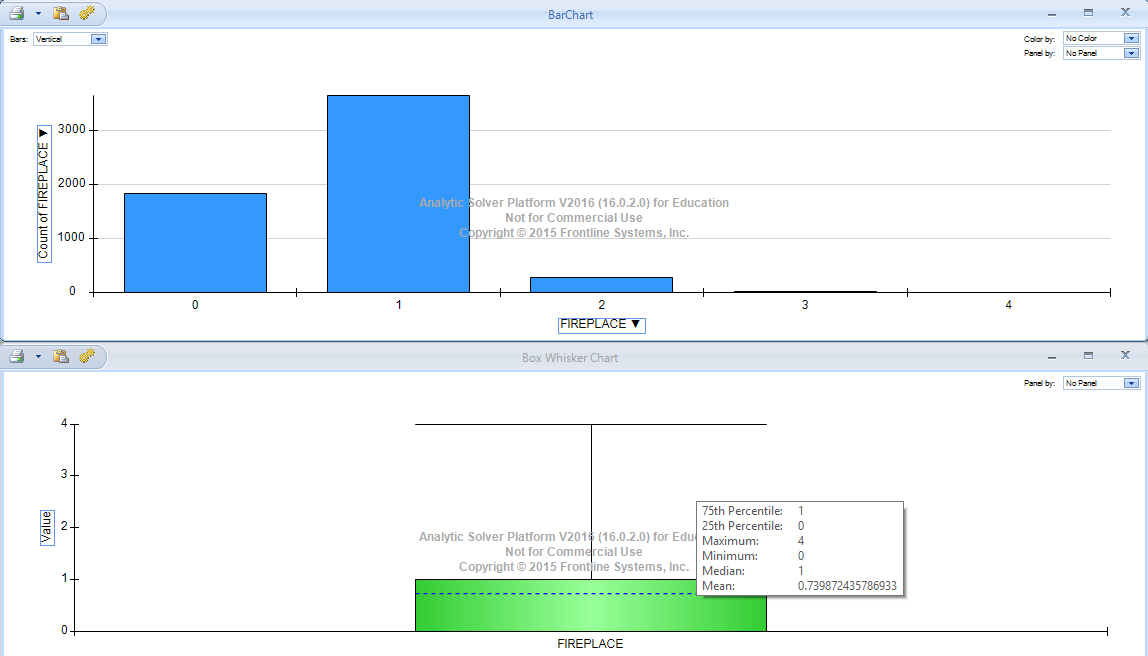


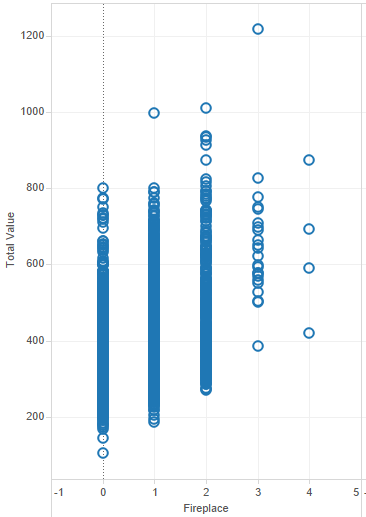
Scatter Plot in Tableau for Kitchen



Here there are very few houses with 2 kitchens which can be seen using the Barchart. Only 90 observations have 2 kitchen out of 5890 total observation.

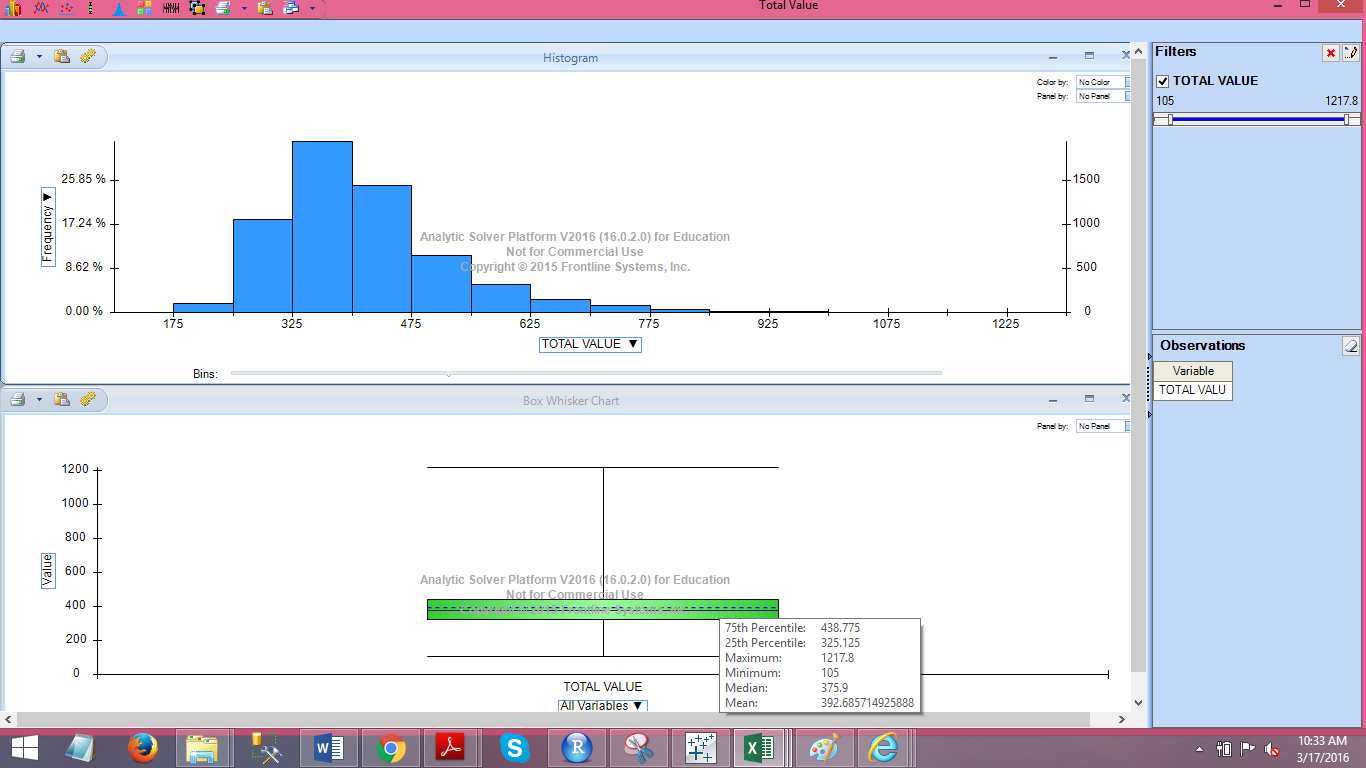
Fireplace:

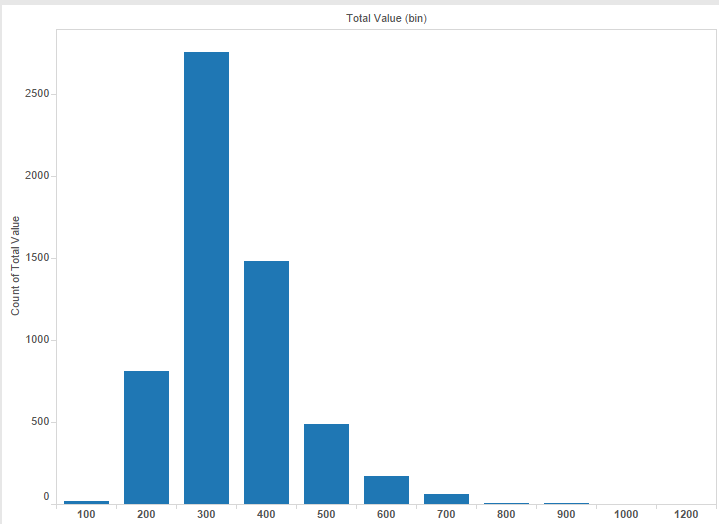




Here we can see there are very few (4) data points at fireplace=4 which means these are outliers.

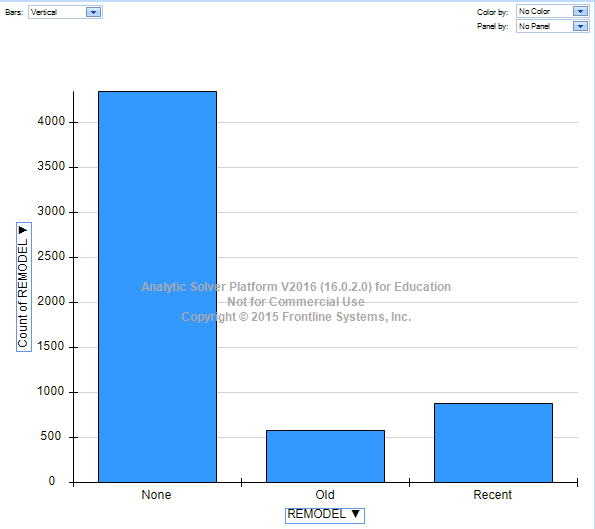
Summary of Total Value



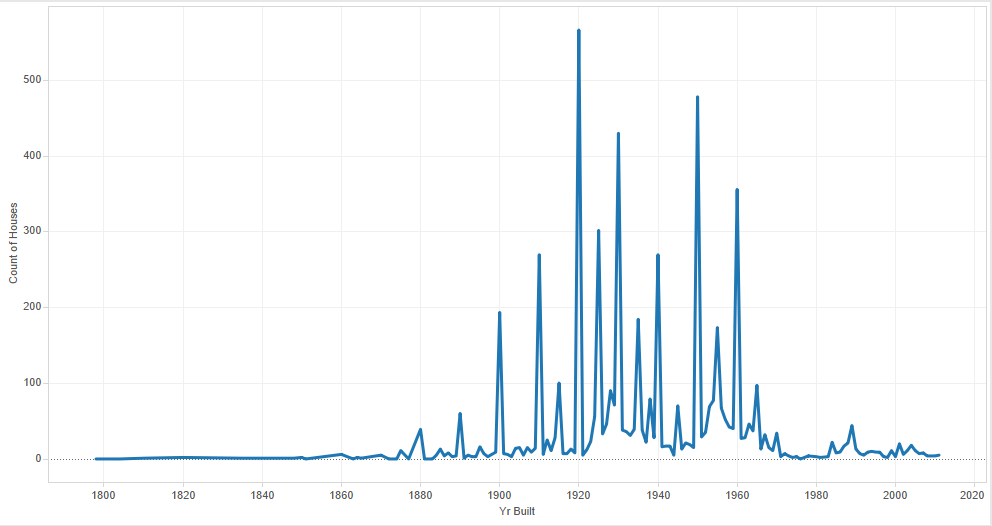


Most of the houses range from 200 to 600$ there are very few houses priced between 700-1000$ and only 1 house priced around 1200$.

Number of remodeled houses

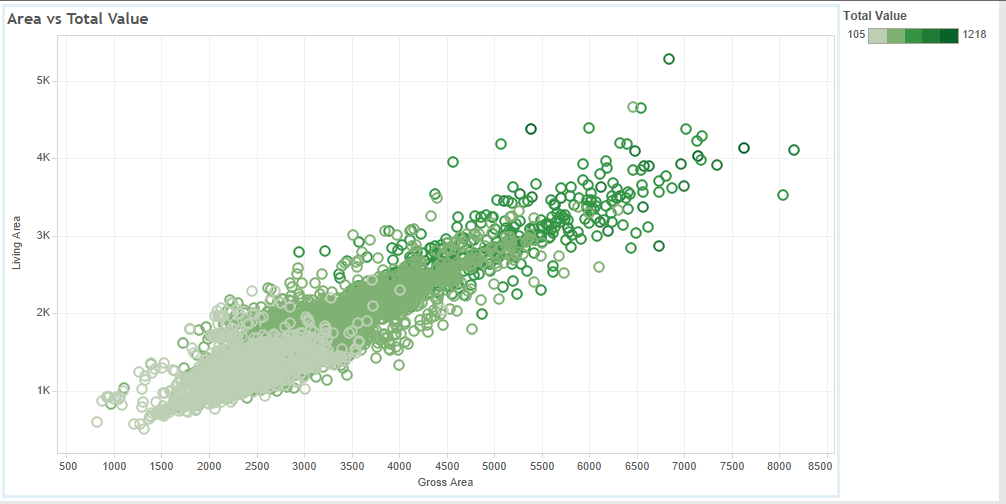


Trend of Year in which houses were built



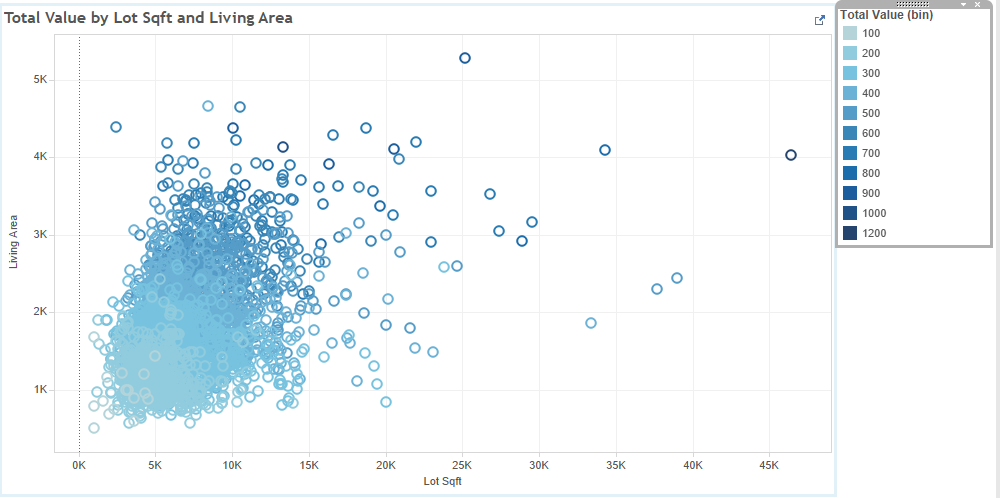
Below chart shows trend when most no. of houses were build. We can see here that there was increase in the number of houses build starting from 1880 and it increased a lot in and around the mid 90’s and it went down again the early 2000’s

Effect of Area on Total Value:



The chart shows when the Living Area and Gross area increases the total value increases too. The gradation in color shows the range of Total Value of the houses.

Effect of Lot Sqft and Living Area on Total Price



More Charts in the Problem.twb show detailed Exploration of Data.

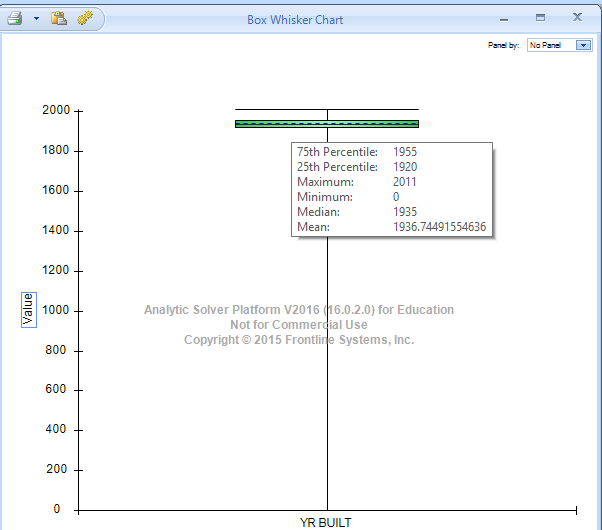
* 1. **Building Prediction model using Multiple Linear Regression, CART and Random Forest**

Tool: XLMiner

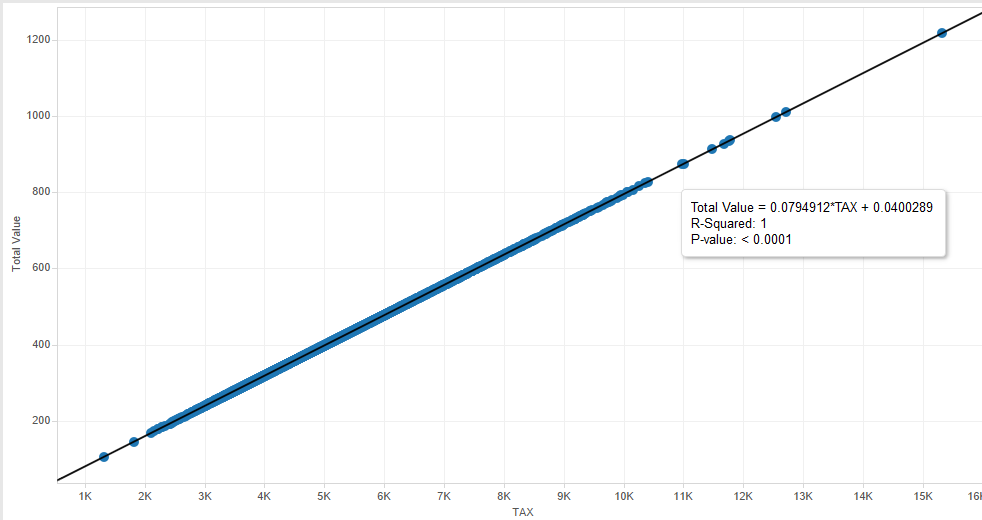
Steps:

1. **Data Pre-Processing:**

After carefully examining the dataset we find a row where the Year built is 0. Since we cannot randomly select a year and replace 0 with it, since many other predictors may be related to year. Also there is only one such row with year 0 so we can simply ignore this row and delete it.



Also the column Tax is calculated from the Total Value of the house. Tax is 12.5 times the Total Value of the house. Hence it is not an independent variable, it is directly dependent on Total Value. So we delete this column as we do not need it for the prediction.



The above chart shows the relation between Tax and Total amount.

1. **Creating Dummies for Categorical Variables:**

The Predictor Remodel (categorical feature) has 3 category which are NONE, OLD and RECENT. We create Dummies for this predictor since otherwise we cannot use them directly for Regression.

Similarly there are other variables like

ROOM (14 Categories)

BEDROOM (9 Categories)

KITCHEN (2 Categories)

FLOORS (5 Categories)

FULL BATH (5 Categories)

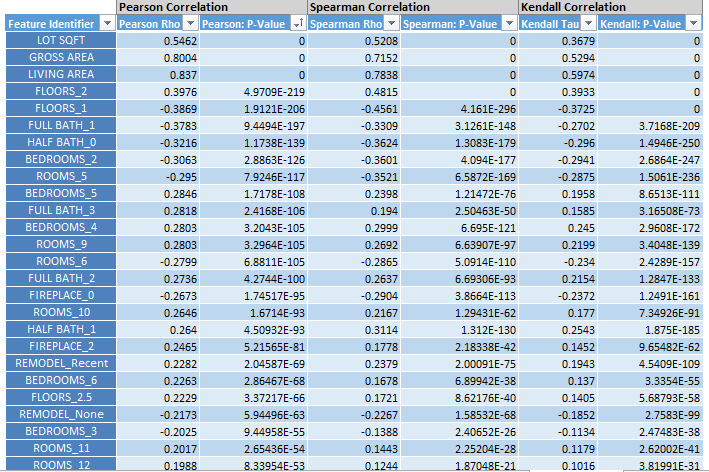
HALF BATH (4 Categories)

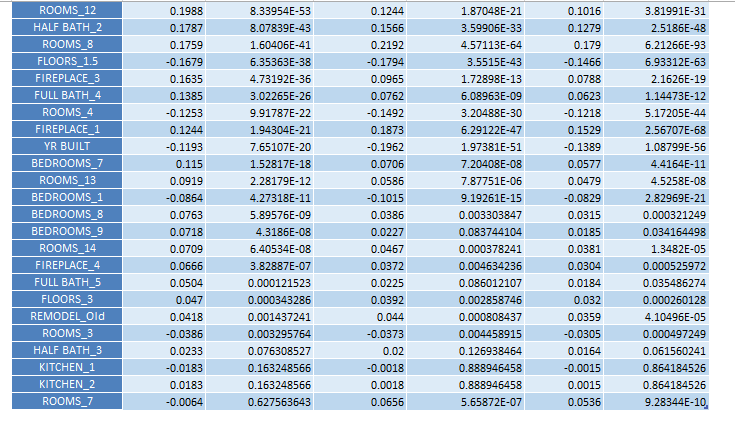
FIREPLACE (5 Categories)

Now we will have 49 variable in total now of which 8 of them would not be useful since if there are n dummy variables n-1 are only useful. So from these 41 variables we need to determine which of them matter how much.

We can determine this by Feature Selection feature of XLMiner. By using Feature selection we can see the Co-relation and the P value of the independent variables on Total Value. This Feature Selection uses Pearson, Spearman and Kendall Correlation for determining the Correlation between the variables. We will consider the values of Pearson Correlation for this problem. Higher Correlation and lower P values are considered to build a good model.

1. **Feature Selection**

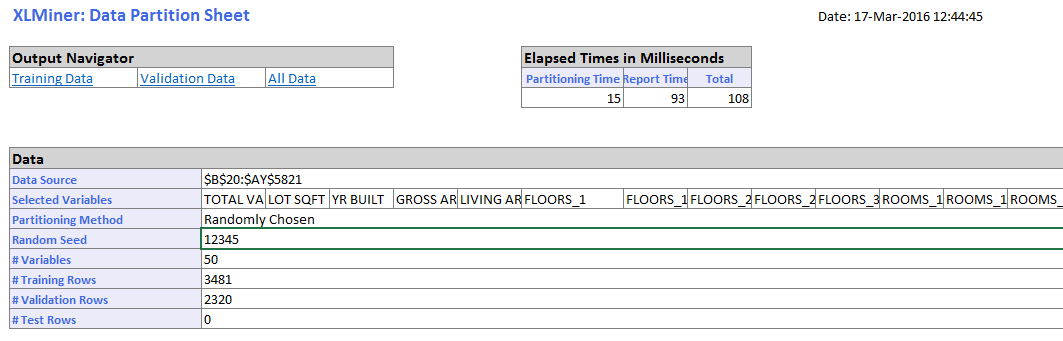




1. **Partition the Data**

Next Step would be to partition the data

We create a random standard partition of 60% Training data and 40% for Validation



1. **Build Models**
2. We build a model using Multiple Linear Regression, CART and Random Forest. For the Multiple Linear Regression we use the process of Backward Variable Selection to find out the best subsets of variables which would give us a good model.

The goal of the model is to have the RMSE value as low as possible. The root-mean-square deviation (RMSD) or root-mean-square error (RMSE) is a frequently used measure of the differences between values (sample and population values) predicted by a model or an estimator and the values actually observed.

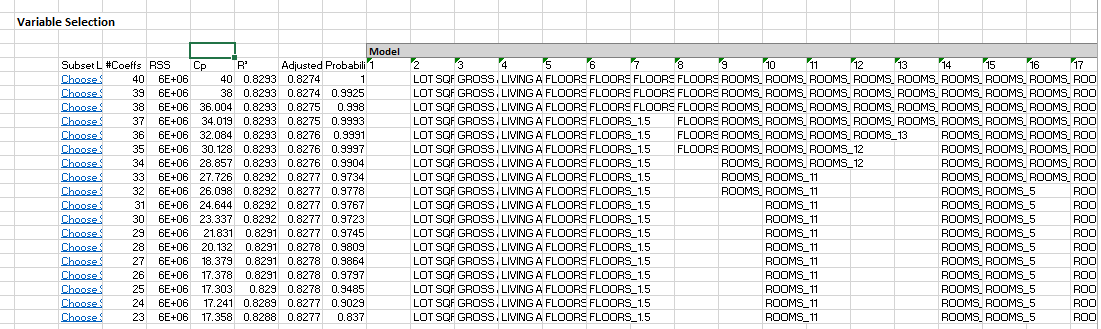
So a good model would be the one with low RMSE value and high R-squared and adjusted R squared value. Also for a good model the RMSE of Training and Validation should be close to each other.

1. The most important factor in predicting the total value in any prediction problem is to select the variables correctly and find the best regression equation for prediction using these features.

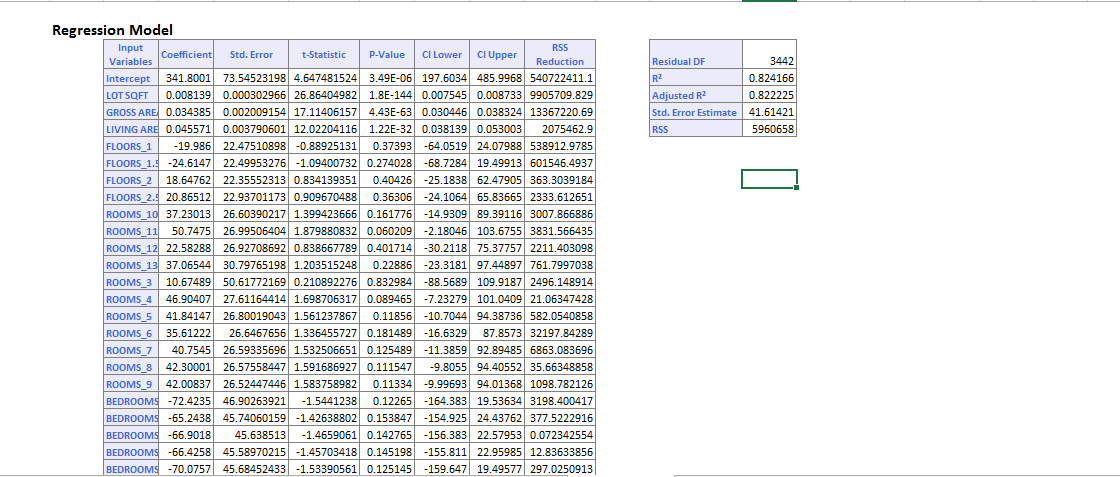
**1.3. The Model and its evaluation**

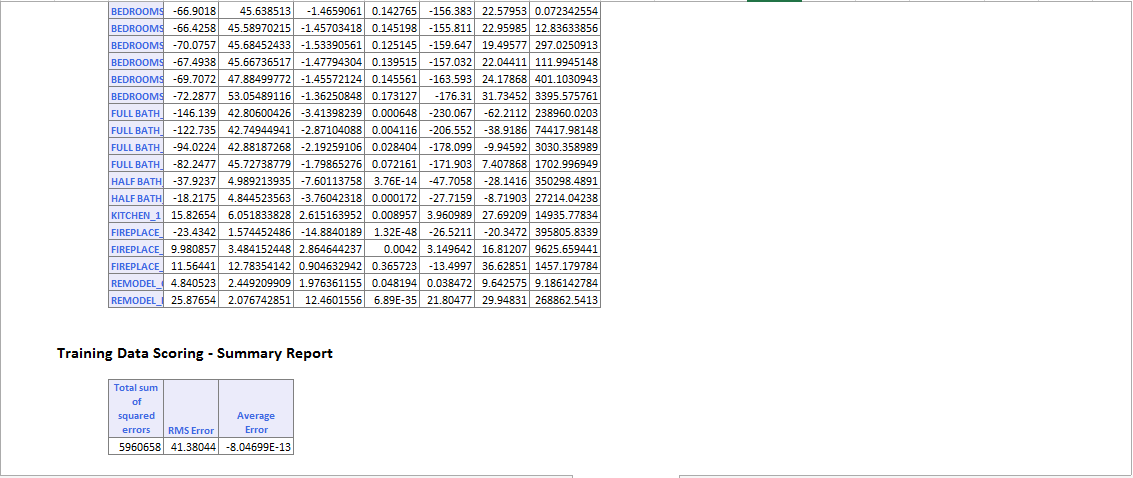
**Regression Model:**

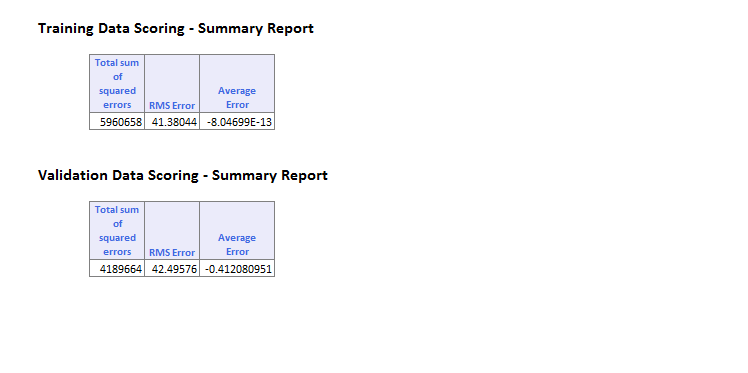
Using feature selection we can select the top variables and perform the regression. But how would we know how many variables will create best suitable model. For this we can use Variable Selection. We have used Backward Variable Selection as there are many variables to select from.



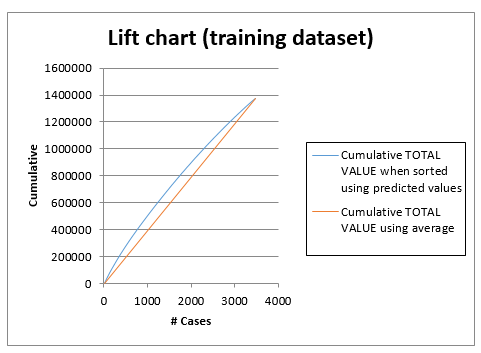
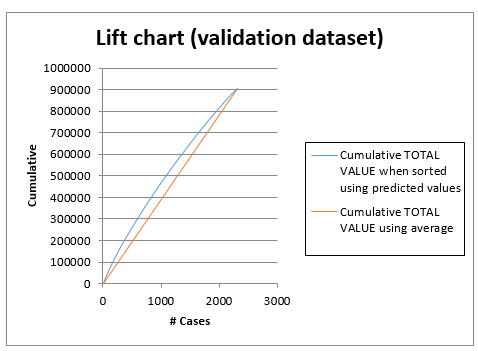
We select the Subset with 38 variables because the Cp value is close to no. of variables and also the R square and adjusted R squared values are pretty good and close to each other.



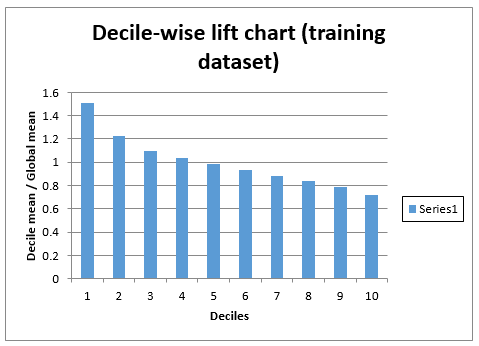
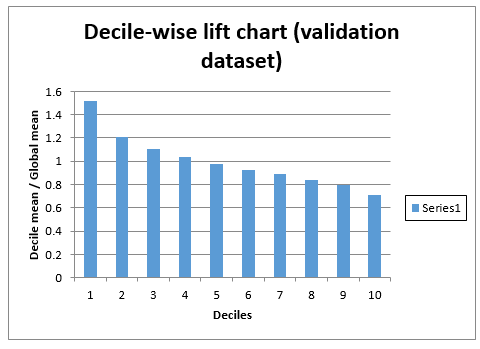


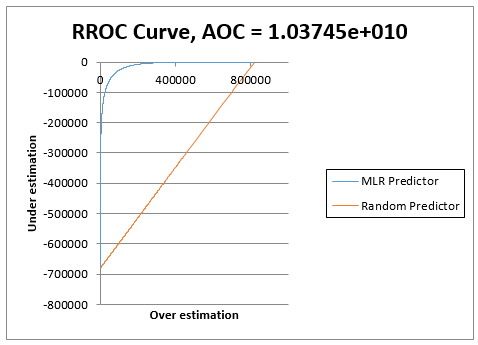
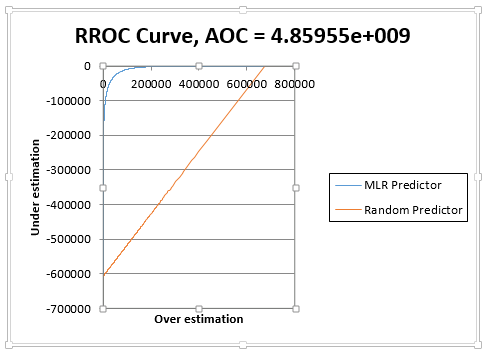


The RMSE we obtain for the Training is 41.38 and Validation dataset is 42.49 which are very much similar and also the average error is low. Also if we check the R-squared and adjusted R squared they are as high as 0.822 and are close to each other indicating this is a good model.

**Lift Charts for Regression:**

**Decile wise lift chart for Regression**



**ROC Curve**

The above charts show the Lift Chart, Decile Wise Lift Chart and the RROC charts for training and validation dataset of Multiple Linear Regression.

**Lift Chart**

Lift is a measure of the effectiveness of a predictive model calculated as the ratio between the results obtained with and without the predictive model.

The red line in the lift chart shows the baseline, which indicates the measure of effectiveness of any random model. The blue line shows the measure of effectiveness of our model which is above the baseline indicating our model is better than the baseline.

**Decile Chart**

After building a statistical model, a decile analysis is created to test the model’s ability to predict the intended outcome. Each column in the decile analysis chart represents a collection of records that have been scored using the model. The height of each column represents the average of those records’ actual behavior.

**Ideal Situation: The Staircase Effect**  
When you’re looking at a decile analysis, you want to see a staircase effect that is, you’ll want the bars to descend in order from left to right.

**Not-So-Ideal Situations**

In contrast, if the bars seem to be out of order, the decile analysis is telling you that the model is not doing a very good job of predicting actual responses.

If the bars seem to be the same height, or the decile analysis looks “flat”, the decile analysis is telling you that the model isn’t performing any better than randomly binning people into deciles would. In both cases, your model should be improved before moving forward with it.

Our linear regression model follows the staircase effect to some extent which means it is a good model.

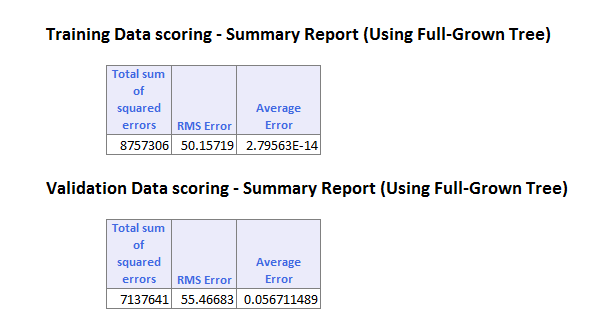
**RROC Curve**

RROC curves plot the performance of regressors by graphing over estimations (or predicted values that are too high) versus under estimations (or predicted values that are too low.) The closer the curve is to the top left corner of the graph (the smaller the area above the curve), the better the performance of the model. Area Over the Curve (AOC) is the space in the graph that appears above the ROC curve and is calculated using the formula: sigma2 \* n2/2 where n is the number of records. The smaller the AOC, the better the performance of the model.

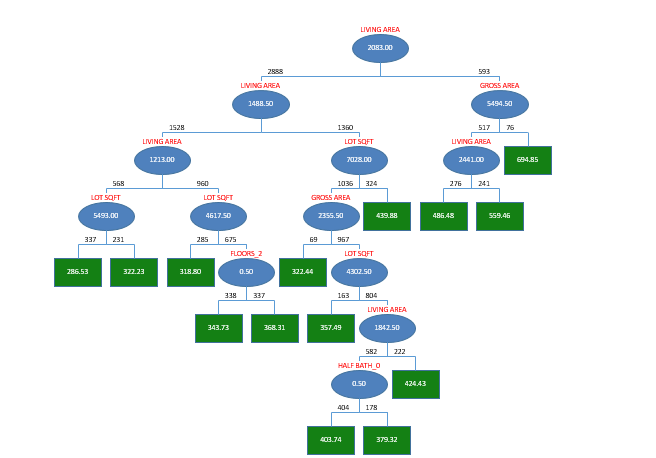
From above charts we can see that the AOC is very less which means that the model is a good model.

**CART**

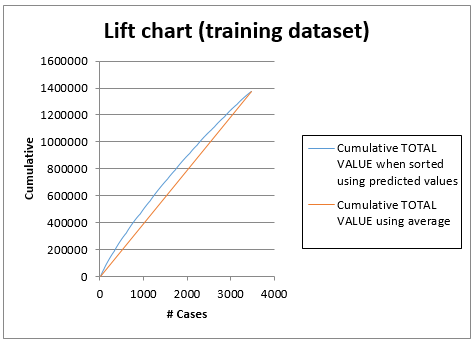
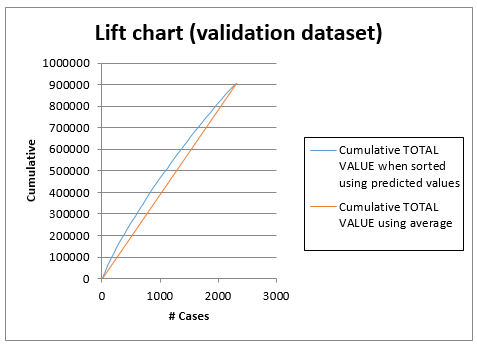
Results for CART are as follows:



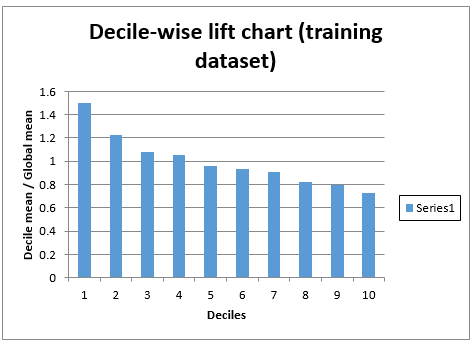
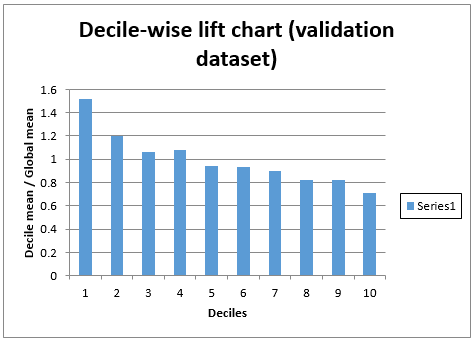
Here the RMSE is 50.15 for training dataset and 55.46 for Validation dataset. Also, only 5 variables are used by the Regression tree which can be seen below. The difference between RMSE for training and validation is greater in this model than in Multiple Linear Regression.

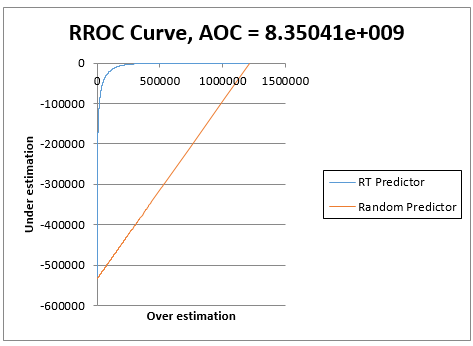
Full Grown Regression Tree

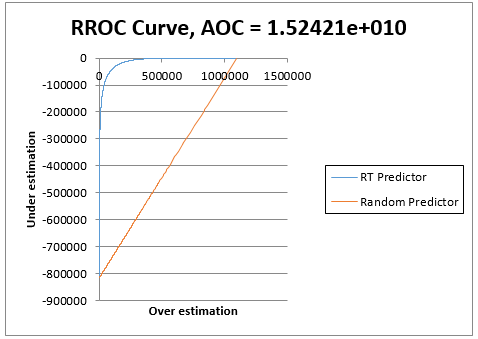
**Lift Charts for CART**

****

**Decile wise Lift Chart for CART**

****

**RROC Curve**



**Lift Chart**

The lift chart for Regression Tree is similar to the one of Multi Linear Regression so it is better than the baseline model. We cannot compare the Linear Regression with CART based on this lift charts as it is not giving us any significant difference.

**Decile Chart**

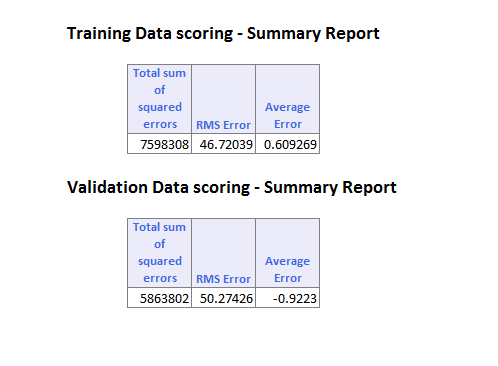
The decile chart for CART is little unstable which shows that the model is not very good.

**RROC Curve**

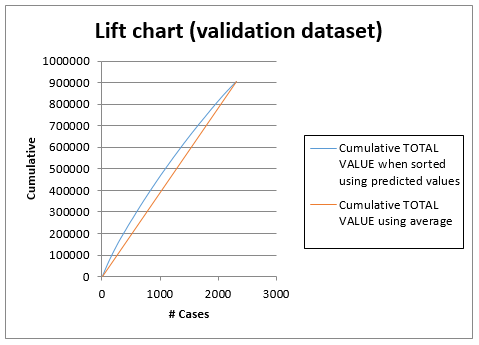
The value of the AOC is less in this model, but the AOC for multiple linear regression is the least amongst the 3 models.

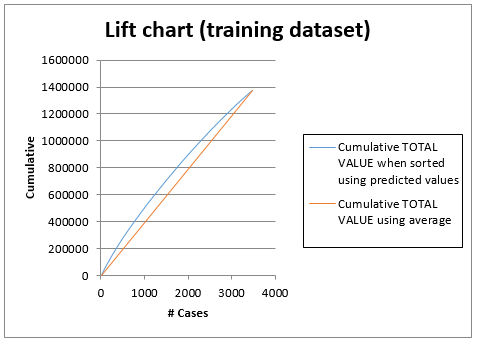
**Random Forest**

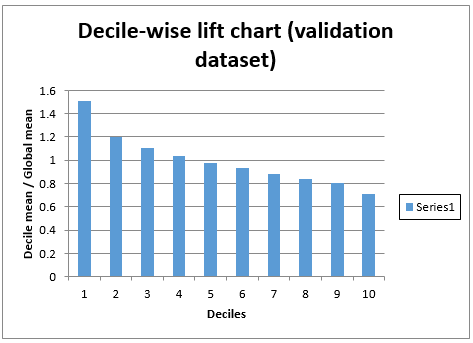
Result for Random Forest are as follow:

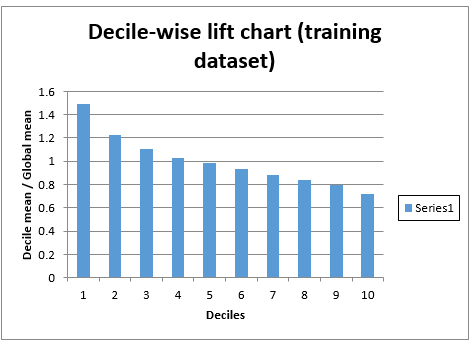


Here the RMSE for Training and Validation is better than CART since in CART only single tree is used and Random Forest creates many such trees and gives best output. The RMSE for Training and validation still differ more than that in Linear Regression. Also the RMSE for Multiple Linear Regression was lower than that in Random Forest.

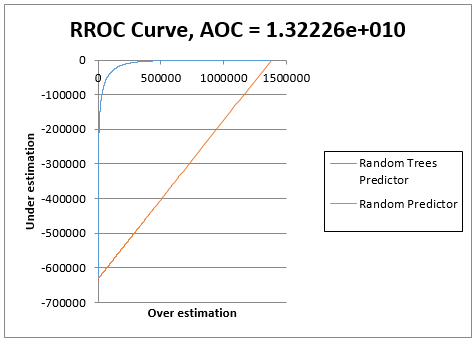
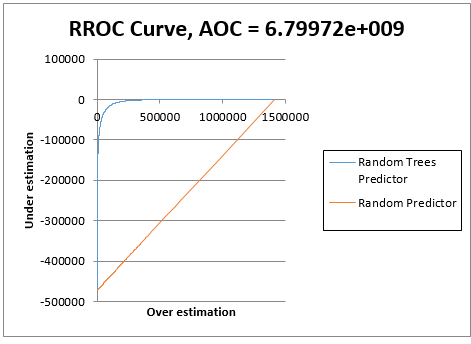
Lift Charts for Random Forest:



Decile Wise Lift Charts:



RROC Curve



**Lift Chart**

The lift chart for Random Forest is similar to the one of Multi Linear Regression so it is better than the baseline model. We cannot compare the Linear Regression with CART based on this lift charts as it is not giving us any significant difference.

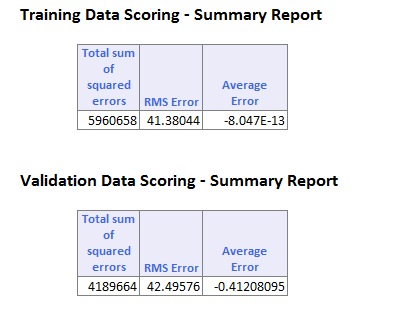
**Decile Chart**

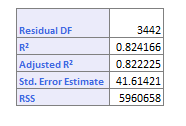
The decile chart for Random Forest is better in this case as compared to CART.

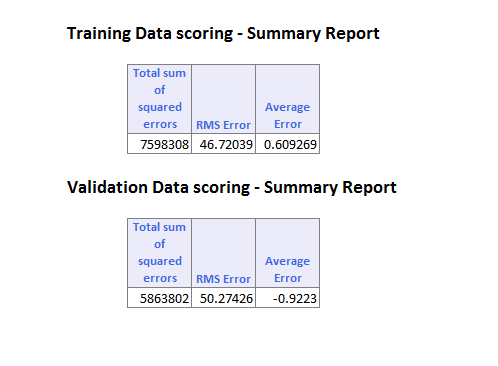
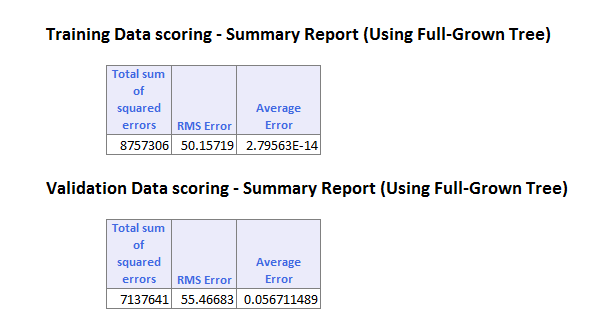
**RROC Curve**

The value of the AOC is less in this model. But the AOC for multiple linear regression is the least amongst the 3 models.

**1.4. Model Recommendation**

**Multiple Linear Regression:**



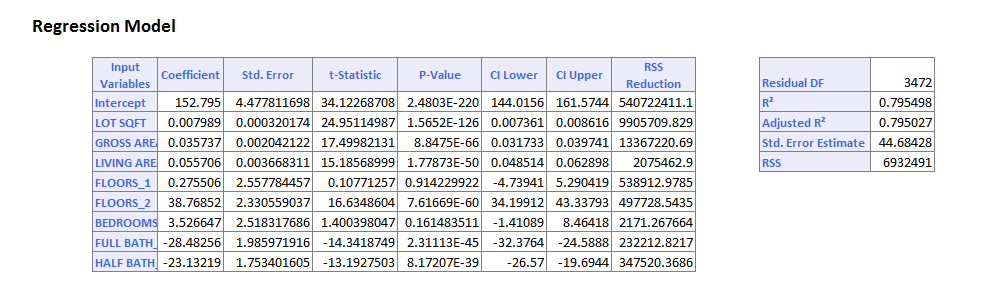
**Random Forest: CART Results:**

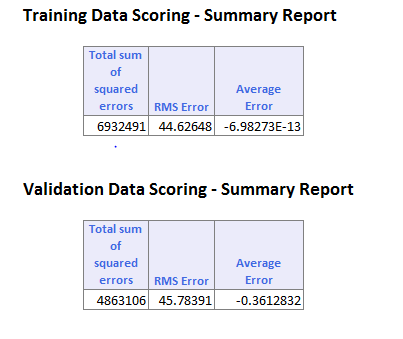
Comparing these three we can see the best (lowest) RMSE value of 41.38 for training and 42 for validation is derived by multiple linear Regression and the R-square as well as Adjusted R-squared is also around 0.82.

Also the Lift charts the Decile wise lift chart for the Multiple Linear regression are better in the case of Multiple Linear Regression.

It can be argued that since we have included a large no. of features for Multiple Linear Regression we are getting RMSE smaller than the other two model. But this is not the case. We have selected the variables by performing variable selection and feature selection. Also if we reduce the number of features that we input into the Multiple Linear Regression model by selecting only top 5-6 features even then the RMSE would be lesser than what we get by Random Forest and CART.

Proof that Multi Linear Model works better in comparison to CART and Random Forest even with lower number of features:





So the model derived by using Multi Linear regression would be the best model for this case of determining the prices of the houses in West Roxbury. The second best based on performance is Random Forest and CART is the last based on performance that we would recommend.