**Analysis on housing data: if sales prices can be accurately predicted using disparate but relevant data**

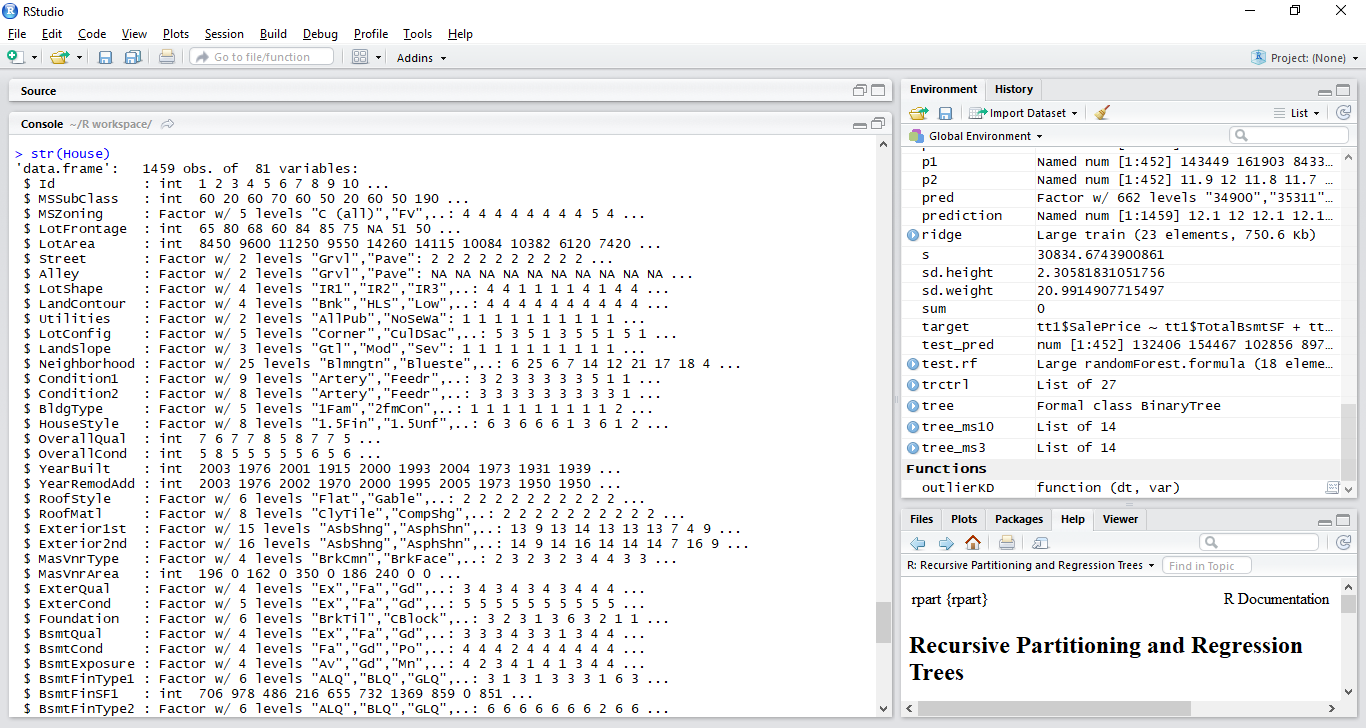
**Overview**

Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

**Objective**

With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, find the most influential set of predictors that influence predicting the final price of each home.

* **Data set Description**



Important set of predictors are :

TotalBsmtSF – total basement square footage

FFlrSF – total first floor square footage

SFlrSF – second floor square footage

YearBuilt – year the home was built

GarageArea – total Garage area

Yrremod – year in which the house was remodeled

PoolArea – total Pool area

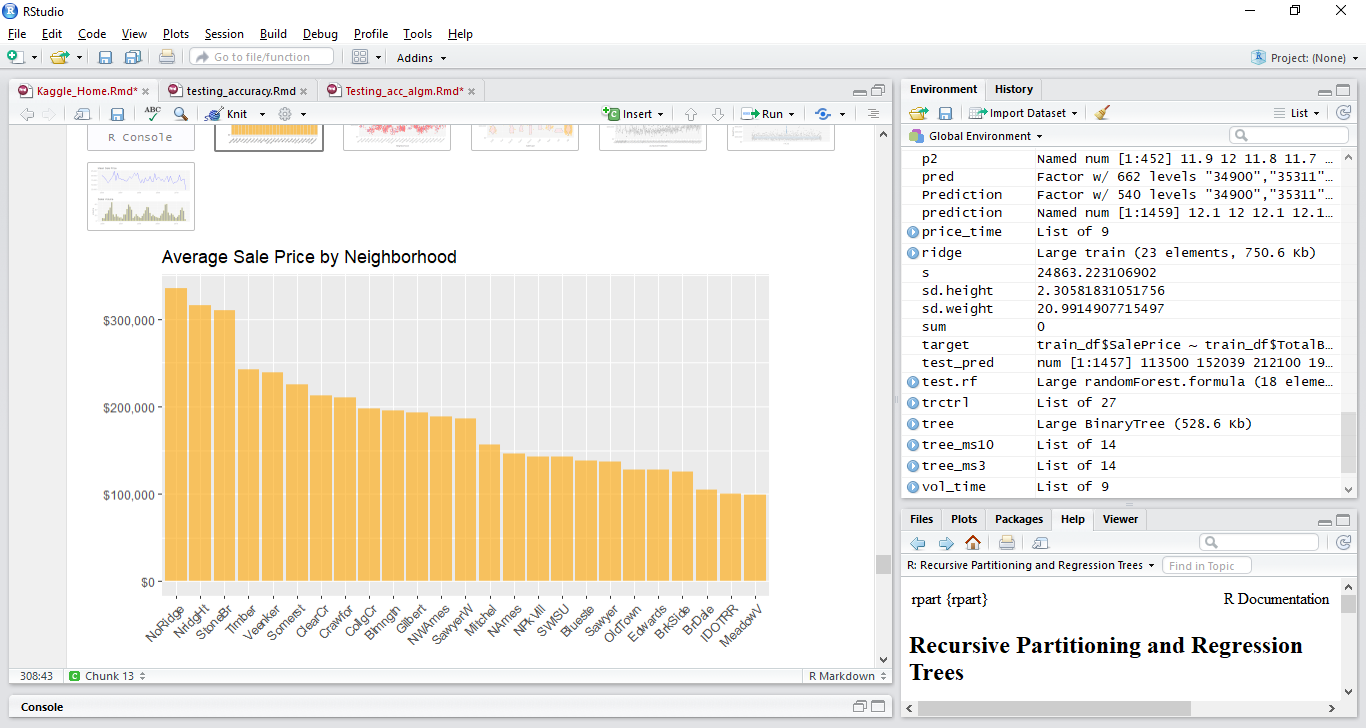
SalePrice – saleprice of each home

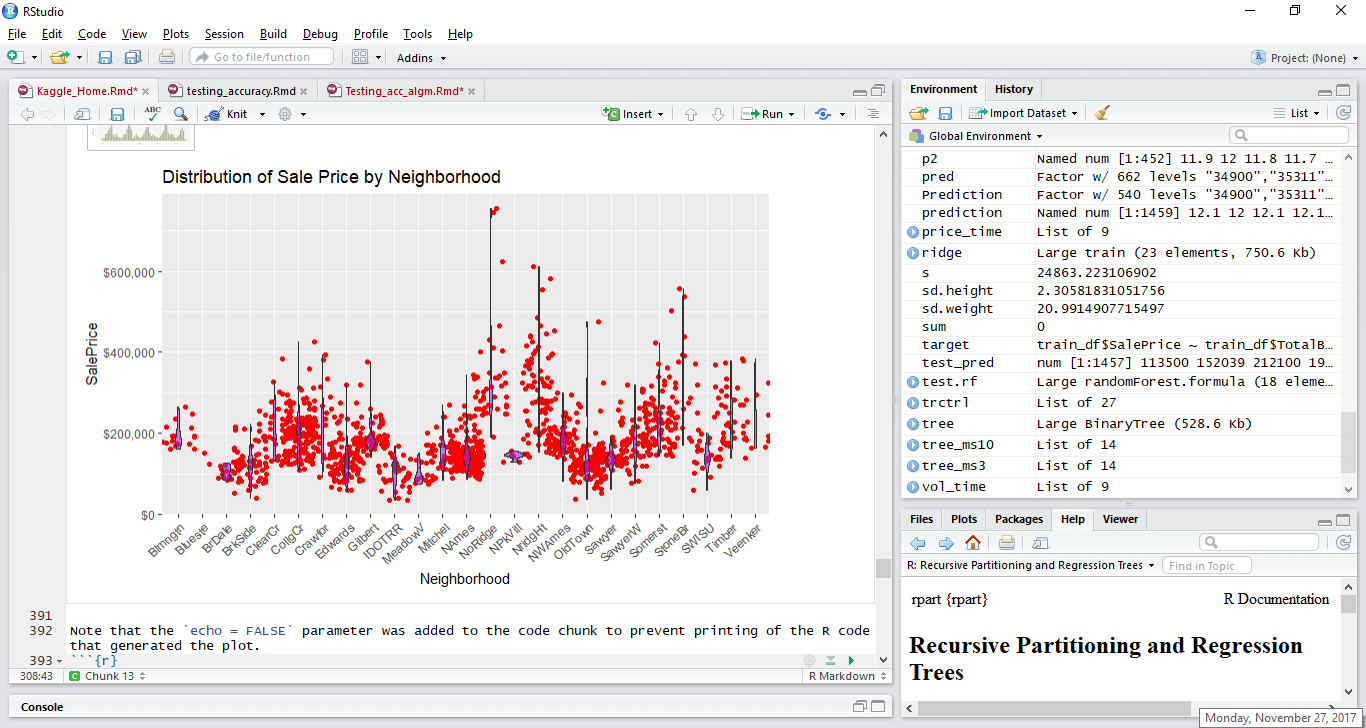
**Initial Exploratory Analysis**

Distribution of Sale Price

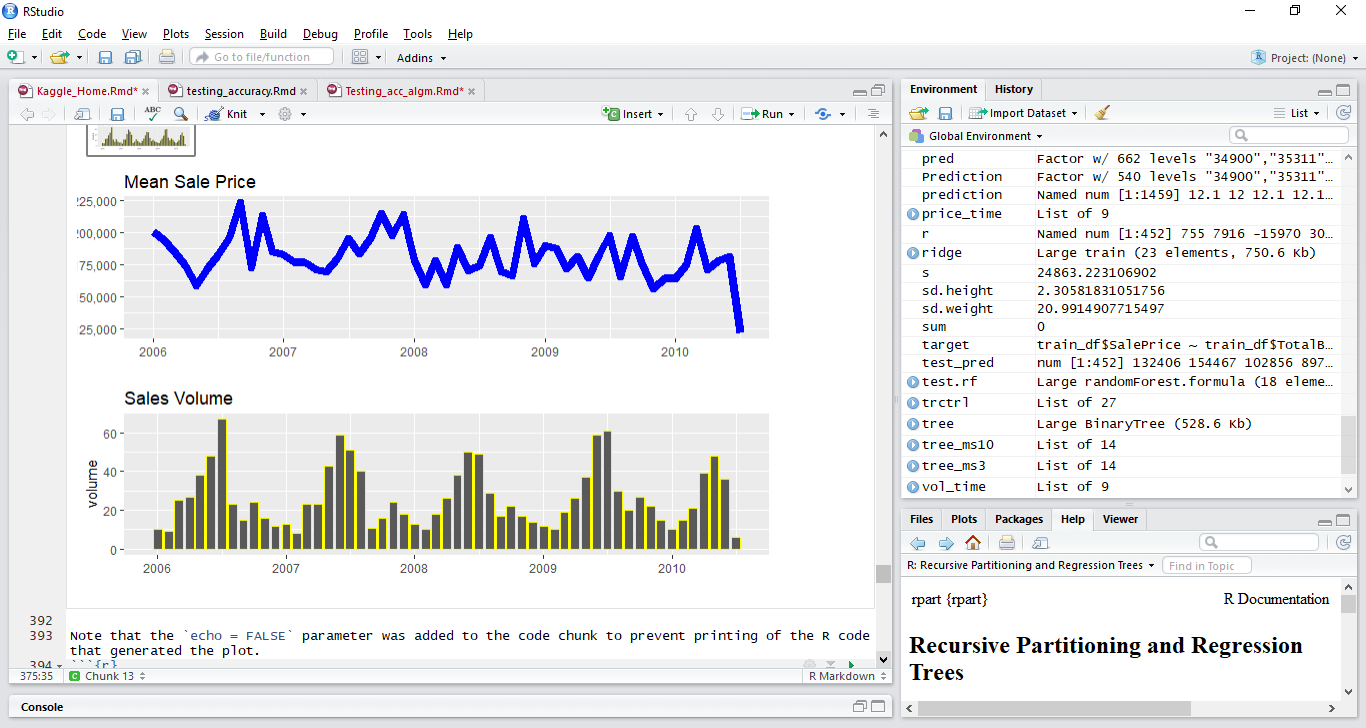


Distribution of Sale Price by Neighborhood



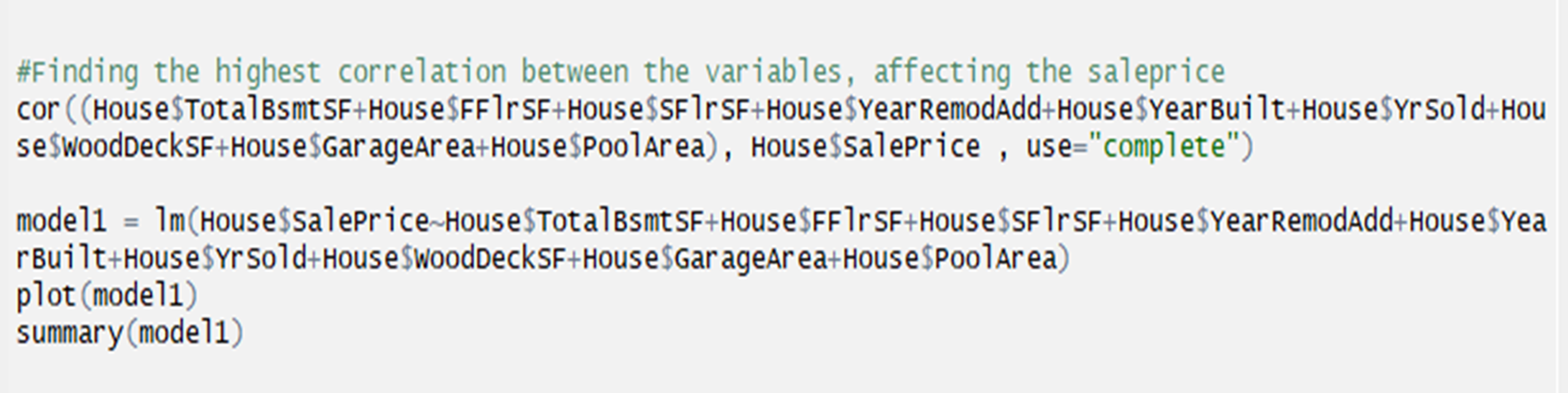


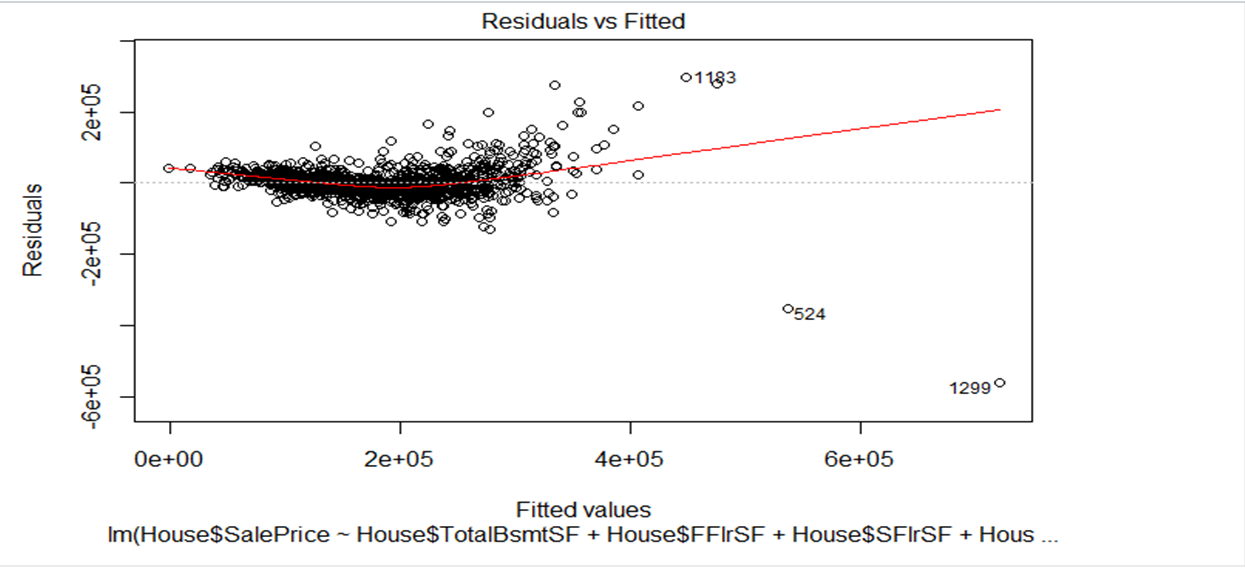
Sale Price and sales volume over a period.

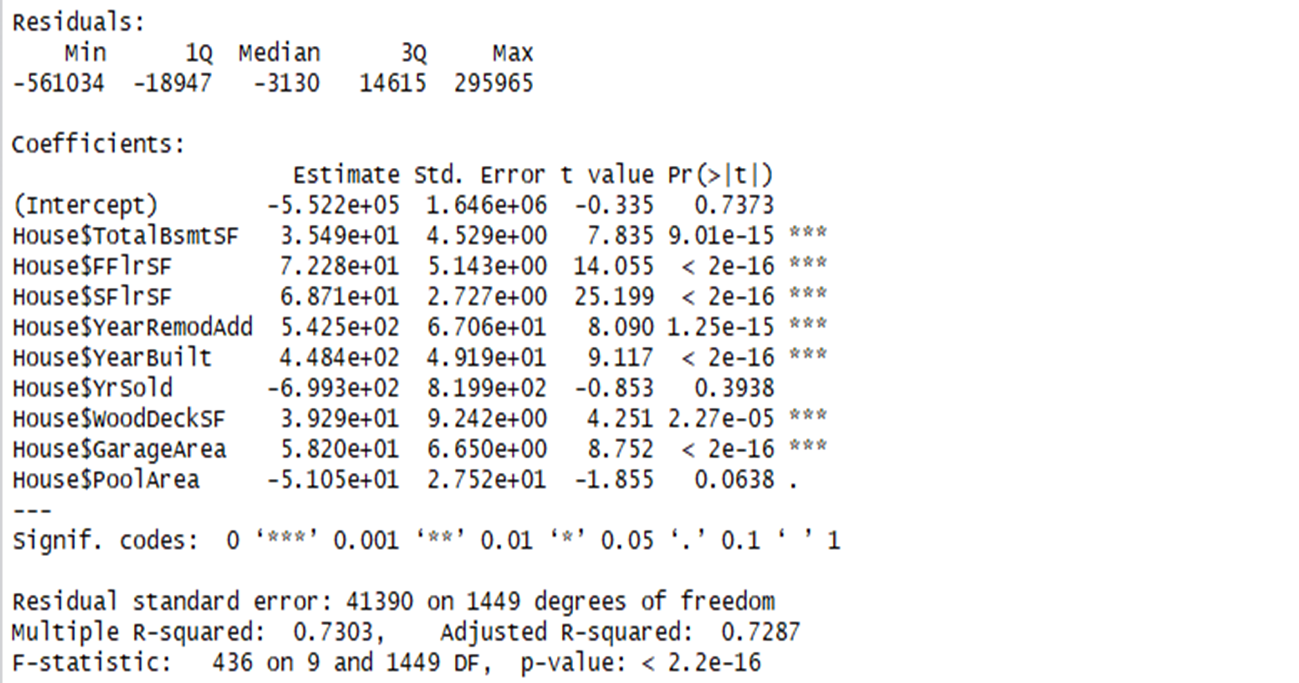


**Correlation Factor and Analysis between the data points Regression**

* Analysis – Relationship between data points Predictor variables vs response
* Finding the variables that are correlated to the sale price
* The correlation coefficient between **sale price and predictors is 0.83**, which shows a strong relationship







72.87% variance in sale price can be explained by the variance in the set of predictor variables.

**Analysis – Confidence values and Z scores**

|  |
| --- |
| C.I. = 97.5 % |
| Variables | ZSCORE | Standard error | Lower limit | Upper limit |
| Sale price | 1.9599 | 9284.6 | 171659.5 | 190228.7 |
| Total Basement SF | 1.9599 | 22.51 | 1034.7 | 1079.8 |
| First Floor SF | 1.9599 | 22.40 | 1142 | 1182 |
| Second Floor SF | 1.9599 | 19.8 | 324.8 | 369.7 |

**Strategy Overview**

* Divide the training data set into subset of train and test (70 percent train and 30 percent test)
* Try KNN, Decision tree and random forest, to predict the sale price, using train dataset as reference.
* Once the Sale price values are predicted, compare it with the values, present in the test set
* Find the accuracy.
* Used R- squared as an estimate of accuracy
* Depending on accuracy, use the method with highest adjusted R square value, to predict the final
* Sale Price on testing data set.

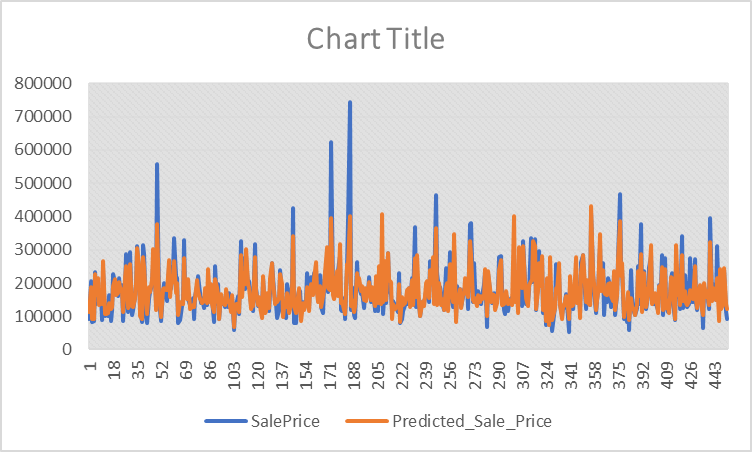
Results

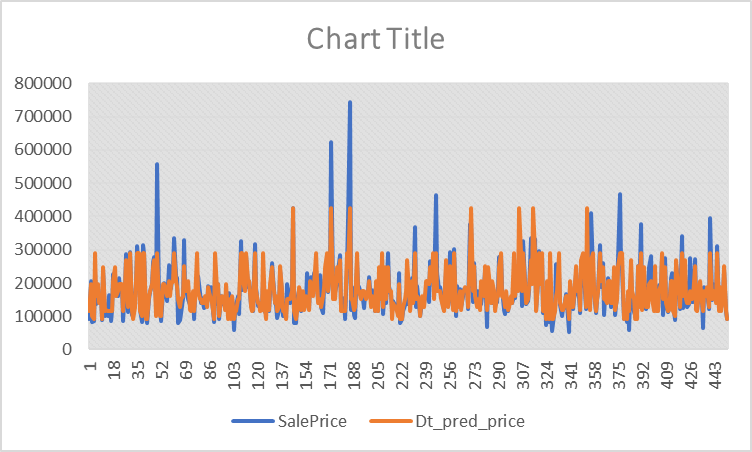
Accuracy (Adjusted R squared)

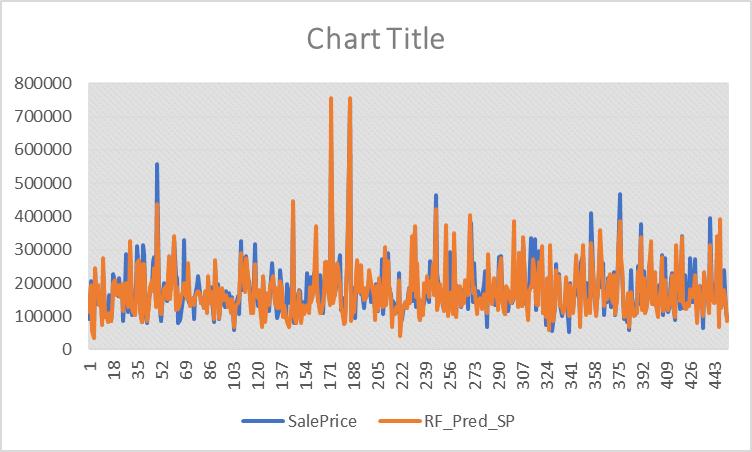
KNN on the sub test data = 86.9 %

Decision trees on the sub test data = 75.9%

Random Forest on the sub test data = 65.7%

****

****

****

**Conclusion.**

We have predicted the values using all the above described methods. So, we found KNN to be the best fit and hence we proceed with KNN method for our final prediction.

**Balineni Priyanka**

**Hanus Andrew**

**Waingankar Varsha**