**PROJECT REPORT**

**REAL ESTATE HOUSE PRICE PREDICTION**

**ADTA 5130**

***DR. YIXUN XING***

**GROUP -- 16**

**Team Members:**

**Harini Chaturvedula (11522208):** Data Understanding

**Uday Raj Suthapalli (11518664):** Parameter Estimation and Model Building

**Venkat Akshay Reddy Jaggavarapu (11514134):** Parameter Estimation and Model Building

**Prakash Manne (11515064):** Model Evaluation and Conclusion

1. **BUSINESS UNDERSTANDING/ANALYTICS QUESTIONS:**

In housing market, real estate companies mostly rely on Comparative Market Analysis apart from their own experience and understanding to decide the proposed price. Comparative Market Analysis can be used to estimate or measure the value of properties by examining and comparing to those properties that are similar. Our analysis aims to help the real estate company to optimize its sales by determining the prices based on factors such as square feet, number of bedrooms, baths, number of stories., etc. It is important for a real estate company to understand how each attribute can have an impact on house prices. This understanding will help the real estate company to better assess the sale and be confident about their proposed price. Hence, our analysis includes how each attribute correlates with the dependent variable which is price. We also apply a regression technique to build an optimal pricing model to better estimate the prices of houses closer to their original value so that this can be utilized by the real estate company to optimize their sales.

The problem statement for our project includes:

* Identifying which factors have highest impact on house prices I.e., square-feet, number of bedrooms, number of stories, age of the property etc.,
* To create a multi linear model that quantitatively relates house price with variables such as number of bedrooms, square-feet, number of stories, age of the property etc.
* To create a model to understand the accuracy of how well these factors can predict house prices.

**2. DATA SPECIFICATION:**

Our dataset contains 7 variables recorded for 550 observations. The seven attributes are described below.

* price - Price of the house in thousands of dollars.
* Sqft - Total square-feet of living area.
* Beds - Total number of bedrooms in the house.
* Baths - Total number of full baths in the house.
* Age - Age of the property in years.
* Stories - Number of stories of the house
* Vacant - If the house is empty or occupied.
* ‘1’ Specifies that the house is empty.
* ‘0’ Specifies that the house is occupied.

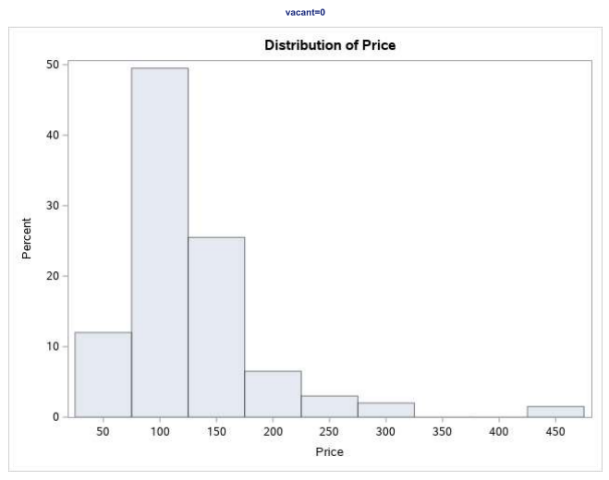
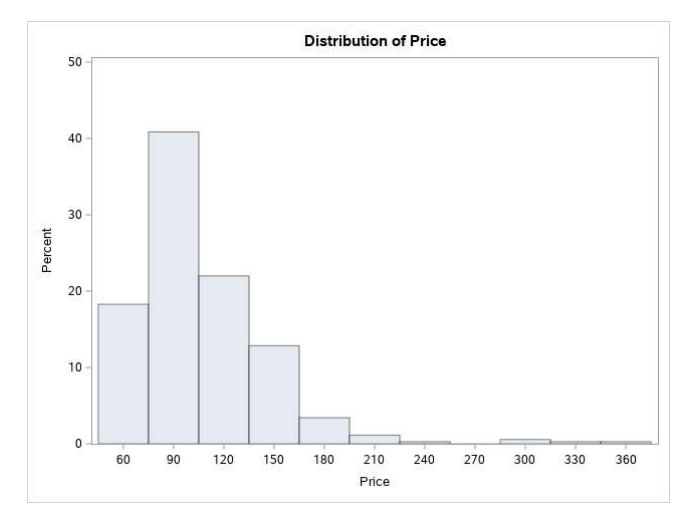
**3. EDA/ DATA UNDERSTANDING & DATA PREPARATION:**

Traditionally, the first step in data preparation involves data cleaning where null values are removed and columns with more than 85% missing values are also removed. But our dataset doesn’t have any null values or missing values. Hence, we will be jumping onto Exploratory data analysis to understand and broaden our perspective about our dataset. To see which variable has highest correlation and that are likely to affect the price, we perform Exploratory data analysis to showcase the correlation between the attributes. Here, a distribution analysis is done to depict the relationship between price and stories.

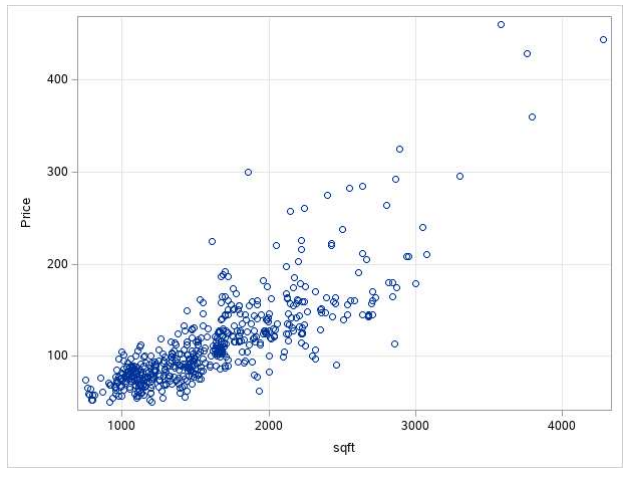
 **Houses with Stories = 1 vs Price Houses with Stories = 2 vs Price**

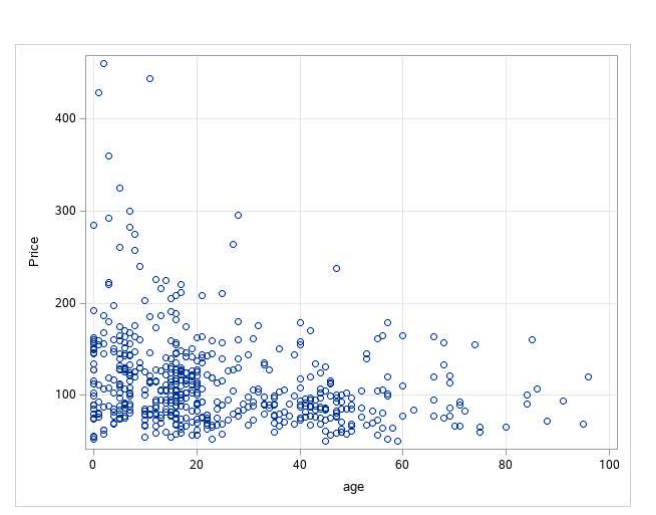


**Houses that are occupied vs Price Houses that are Vacant vs Price**



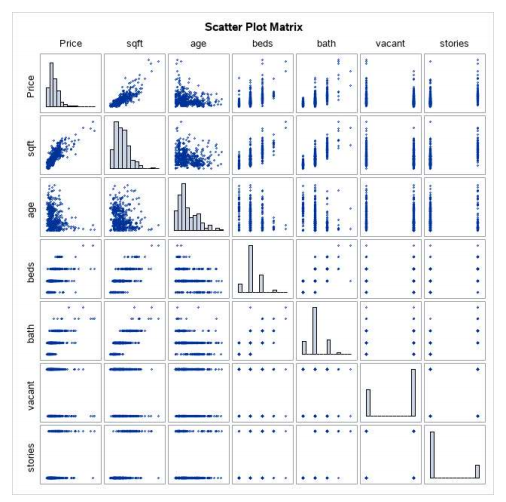
Now using a scatter plot, let’s depict the relationship between price and age, price and sqft.

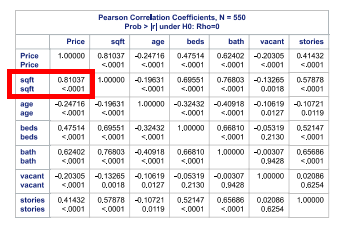
**Scatter Plot for Price Vs Age** **Scatter Plot for Price Vs Sqft**



**3.1 Correlation analysis:**

A correlation analysis is applied on this data set to understand how independent variables such as square-feet, number of bedrooms, number of beds, age of property etc., affect our dependent variable which is price. A scatter plot Matrix depicts the relationship between all the variables vs Price.





From the above correlation analysis, it is clear that sqft shares highest correlation with the dependent variable which is price. The correlation coefficient for sqft with price is 0.81037 with p-value <0.0001. The next highly correlated variable with price is bath with a correlation coefficient of 0.62402 with p-value of <0.0001. The least correlated variable with the dependent variable, price is age of the property with a correlation coefficient of -0.24716 with p-value <0.0001. Another simple interesting fact from this dataset is the price of house is more dependent on the number of bathrooms than the number of bedrooms in a house.

Data transformation was required since our data is not normally distributed. Hence, we have done data transformation for our dependent variable price by square root of price which gives more normally distributed values which is better for our analysis.

**4.MODELING**

**4.1 Model - 1: Simple linear Regression Model**

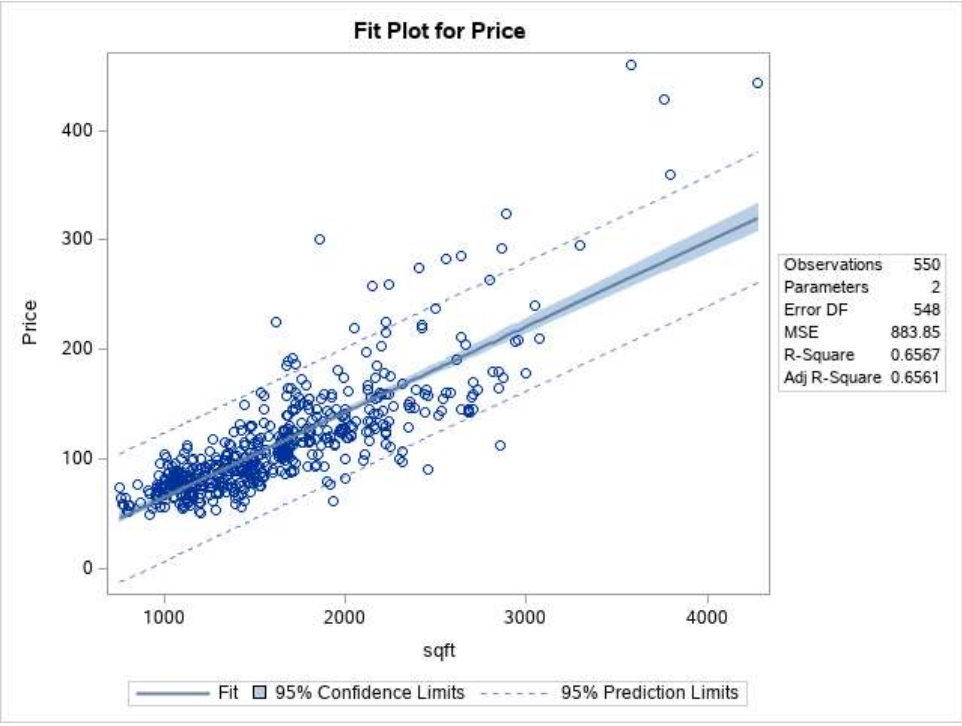
From the correlation analysis we have observed the 0.81037 positive correlation between the price and sqft, which means there is 81% of price of a house depends on the sqft of the house. So, we have applied the simple linear regression model with price as the dependent variable and sqft as the independent variable, then we got the regression model that has the R-squared value of 0.6567 and Adjusted R-squared value of 0.6561. The formula to predict the price of the house using this model is

**Price = -12.40880 + 0.07783\*sqft.**

The fit regression plot for the price and sqft looks like below.

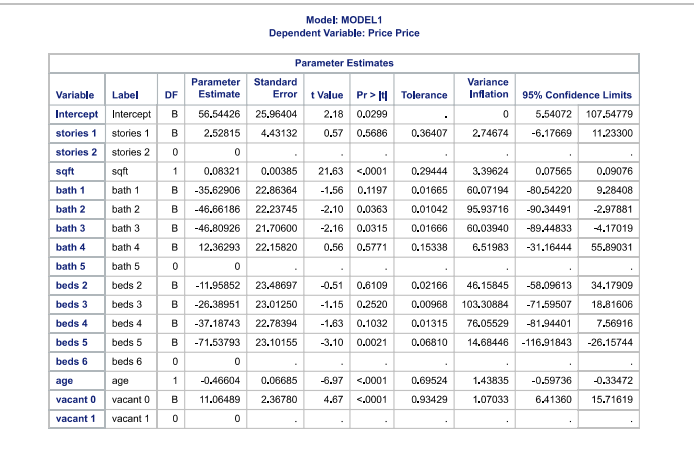
As we got R-squared as 0.6561 we wanted to improve the performance of the model by taking advantage of remaining attributes in the dataset which are age, beds, baths, vacant, stories.

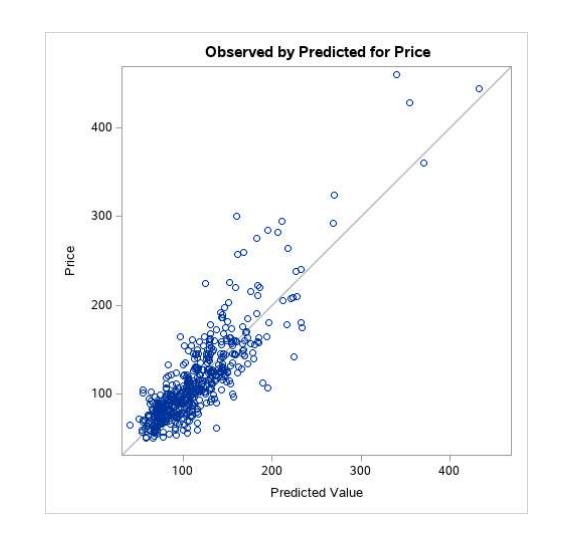
We haven’t used the squared rooted price and squared rooted sqft values in building a model as it doesn’t give much results than we expected and also it feels like the model is overfitting with the square rooted values based on the values of the VI and tolerance.



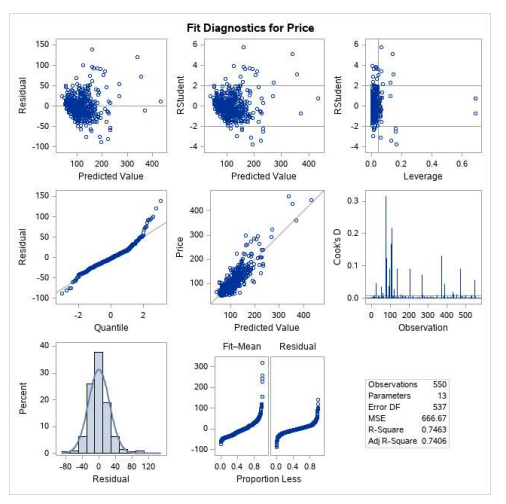
**4.2 Model - 2: Multi-linear Regression Model**

Now, we will apply a multi linear regression to build an optimal pricing model where our variables are logarithmically transformed and our categorical variables vacancy and stories are label encoded. We have performed multi linear regression with no interactions and no selection method. From the below estimates we can observe the Variance Inflation have a higher value greater than 10 which gives us the collinearity in the model we built. But we are not going to remove the attributes that give the high VI values as we will not consider the categorical duplicates VI’s in determining the collinearity of the model.



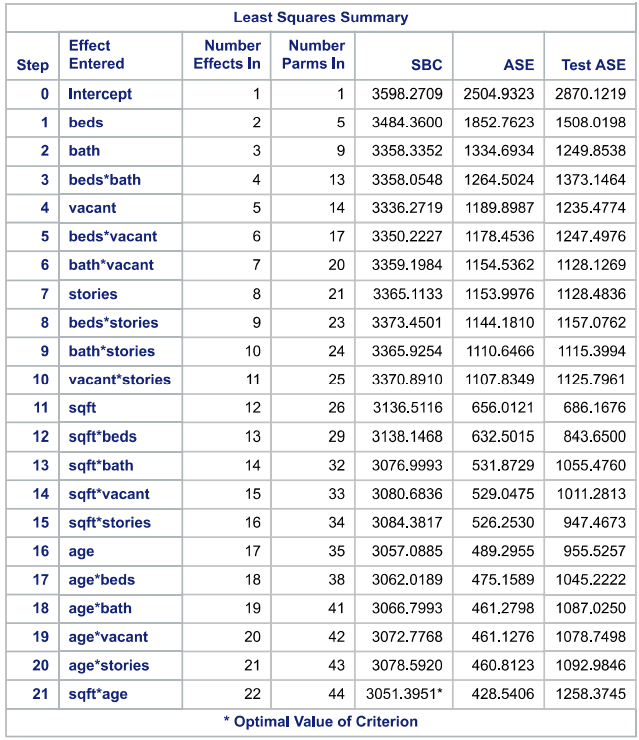


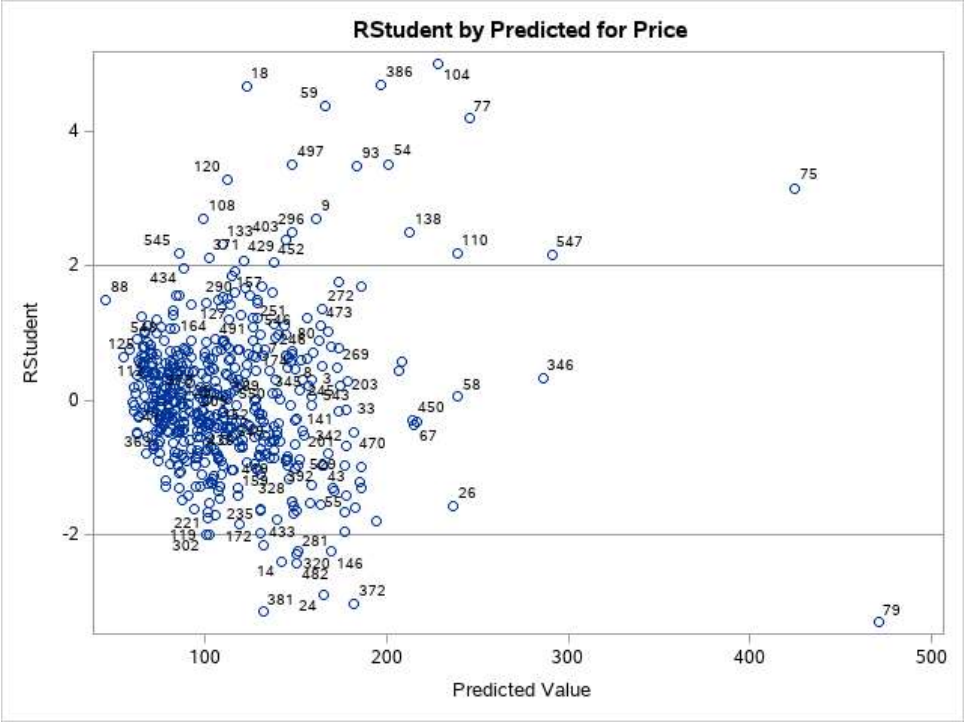
There has been a diagonal plot when we do a scatter plot for the predicted price and the real price value. We have got p values less than 0.05 for almost all of the variables except some which are derived as duplicates for the categorical variables, which we can ignore.

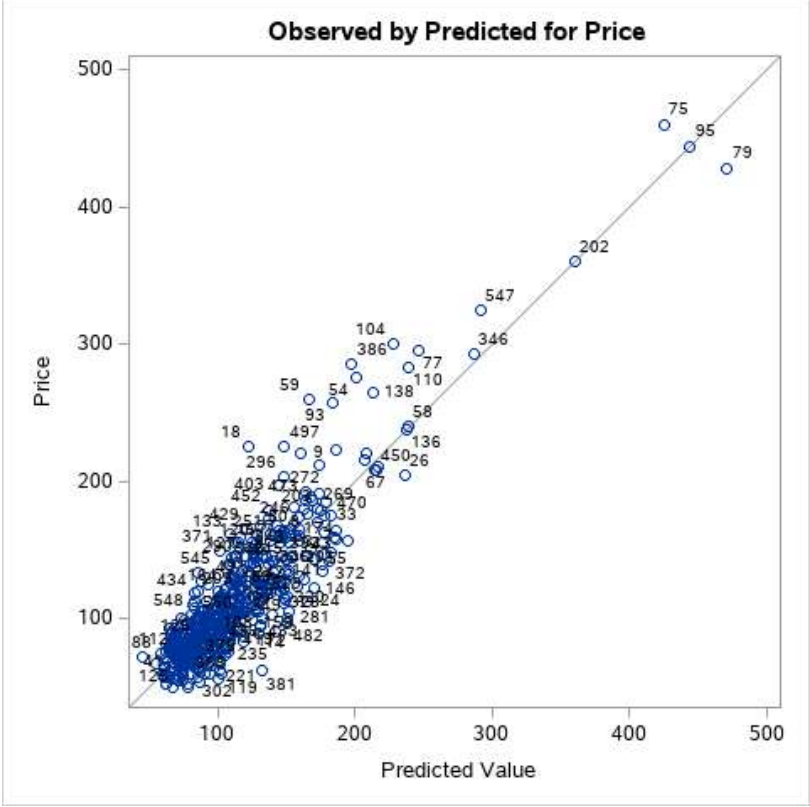


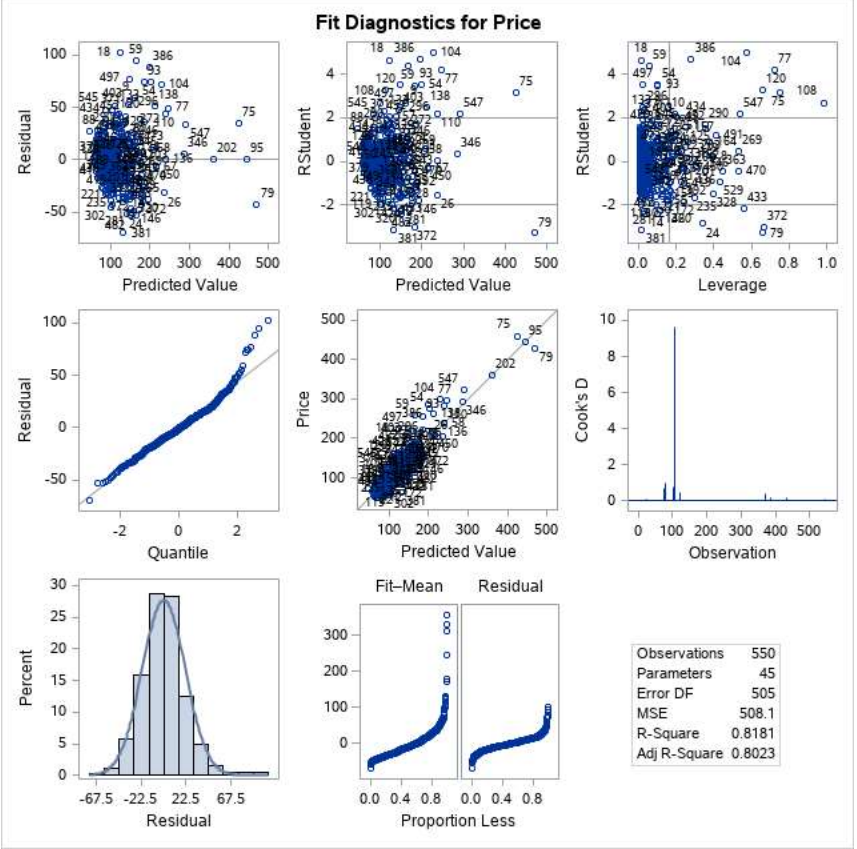
**4.3 Model - 3: Predictive Regression Model**

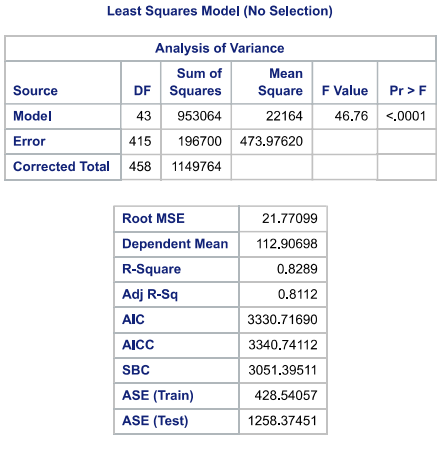
Now, we will apply a Regression model using the Predictive regression model in the SAS to build an optimal pricing model where our variables are logarithmically transformed and our categorical variables vacancy, beds and baths are label encoded. For this model we have used main effects and interactions on the with all variables for an N value of 2. We have used 80% of the dataset for training the model and 20% for the testing the model we have built which is 459 data for training and 91 data for testing. After applying the model with all the variables and all interactions we have got a R-square of 0.8289 and Adjusted R-square of 0.8112.





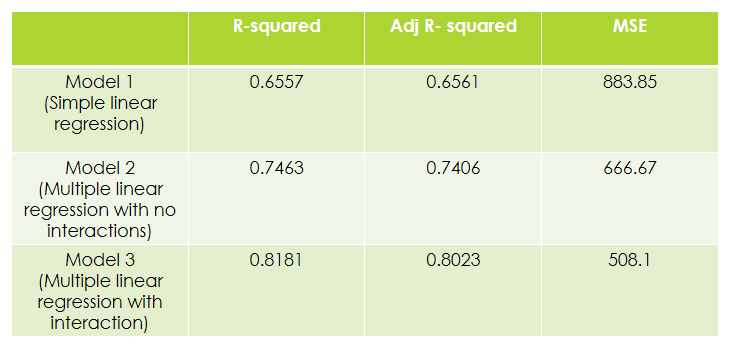
S





**5. Evaluation / Conclusion**

After building our models, we will evaluate the performance and accuracy of our models by considering R-square and MSE as our evaluation metrics. We can conclude that model -3 performs better when compared to model -1 and model 2 since the R-square value of model - 3 is 82% whereas R-square value of model -2 is 74% and model -1 is 65%. According to the mean squared error, model -3 has a lowest error rate when compared to model - 2 and model - 1. Hence, we can say that model - 3 is more optimized model which gives predictions of prices of houses with highest accuracy and low error rate. We can finally conclude that model -3 satisfies our business objectives by rightly predicting the prices.

The attributes that most contributes in predicting the price of a house for the dataset we have taken is the sqft of the house. From the predictions of multilinear model that we have created, we can conclude that the price of the house can be decided based on the attributes of the house like age, no. of bedrooms, no. of bathrooms and other attributes in the dataset we worked on. We cannot conclude the accuracy of our model based on the R-squared as it might be overfitted for our dataset. So, we have divided the dataset into training and testing set of 80 and 20 percentages and predicted the results on the testing data which is not involved in the created a model. Then we compared the predicted results of test data with its original data and then decided on the R-squared and the accuracy of the model.