**Housing Price Prediction**

**Data Description:**

The dataset used for this study is a CSV file named “Housing.csv” that contain information of House Data

The data include the following 13 features and 545 Sample:

* Price
* Area
* Bedrooms
* Bathrooms
* Stories
* Mainroad
* Guestroom
* Basement
* Hot Water Heating
* Air Conditioning
* Parking
* Prefarea
* Furnishingstatus

**Objective:**

* Understand the Dataset & cleanup (if required).
* Build Regression models to predict the sales w.r.t a single & multiple feature.
* Also evaluate the models & compare thier respective scores like R2, RMSE, etc.

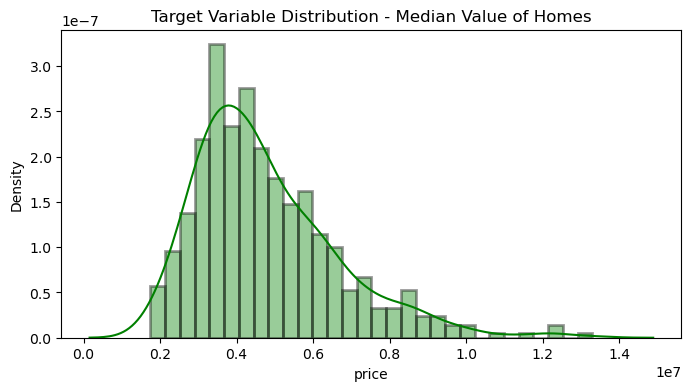
**Here are some of the necessary steps:**

1. Data Exploration
2. Exploratory Data Analysis (EDA)
3. Data Pre-processing
4. Data Manipulation
5. Feature Selection/Extraction
6. Predictive Modelling
7. Project Outcomes & Conclusion

**Result**

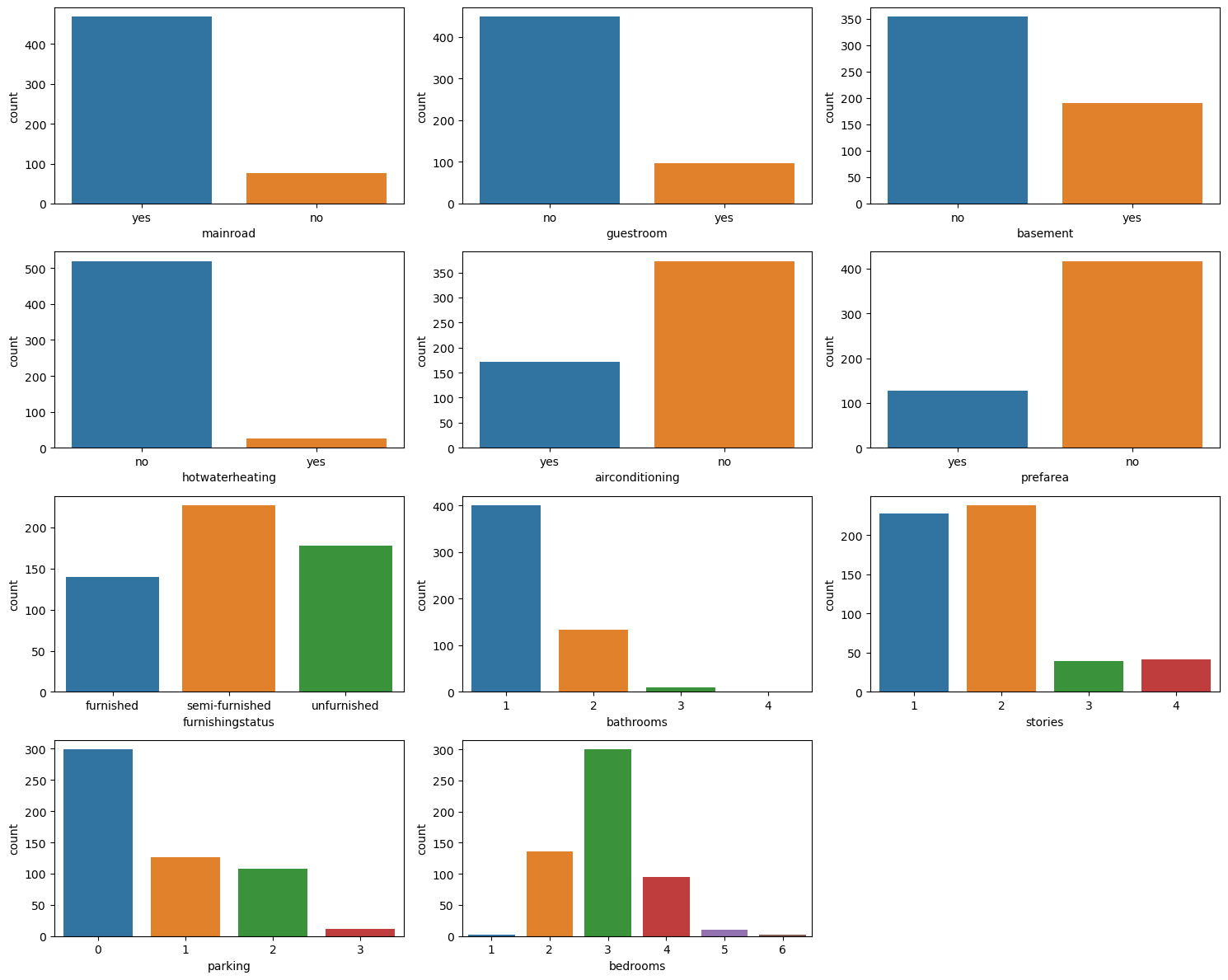
Exploratory Data Analysis (EDA)

* Analyze the distribution of the target Variable

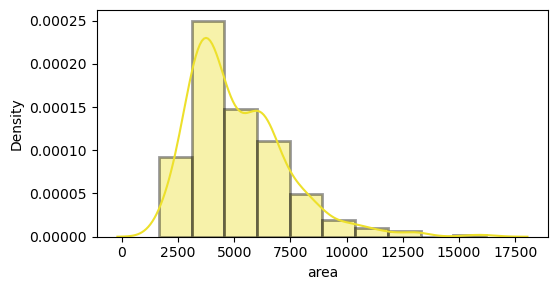


* The Target Variable seems to be normally distributed, averaging around 20 units.

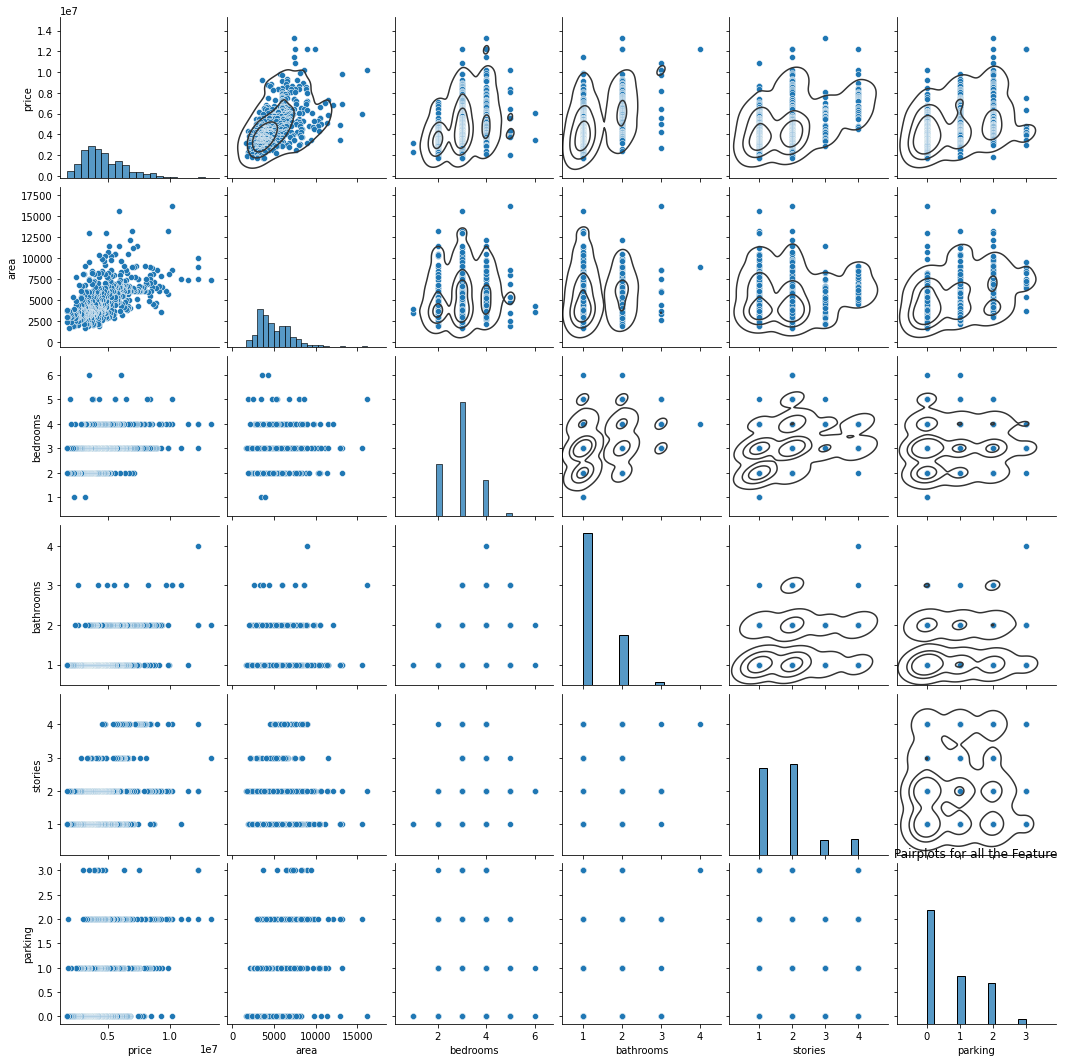
**Visualising Categorical Features**



**Visualising the Numeric features**

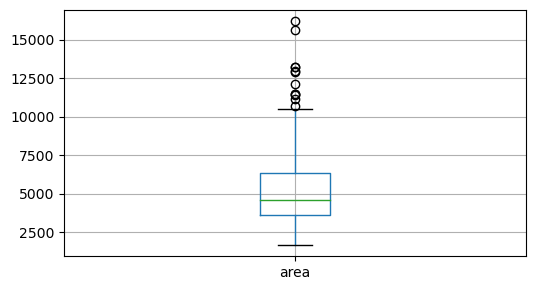
****

**Relationship between all features**



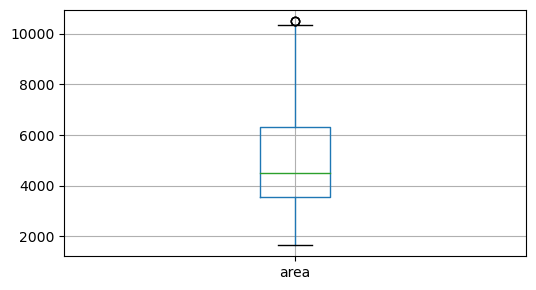
* We can notice that some features have linear relationship, let us futher analyze the detect multicollinearity.

After removal of outlier

****

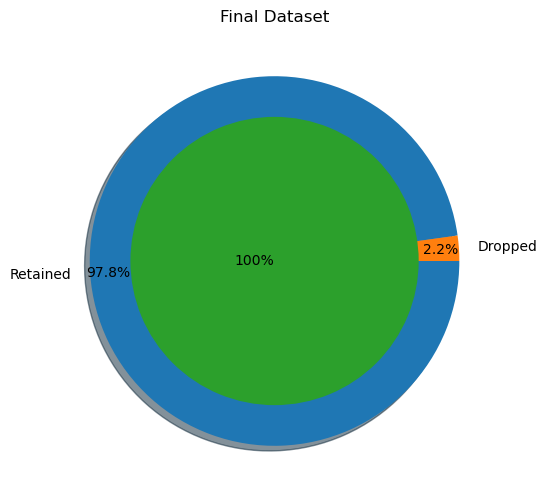
* The data set had 545 sample

Before removal of outlier



* The data set had 533 Sample

**Final Dataset size after performing preprocessing**

****

After the cleanup process, 12 samples were dropped

While retaining 2.2 % of the data.

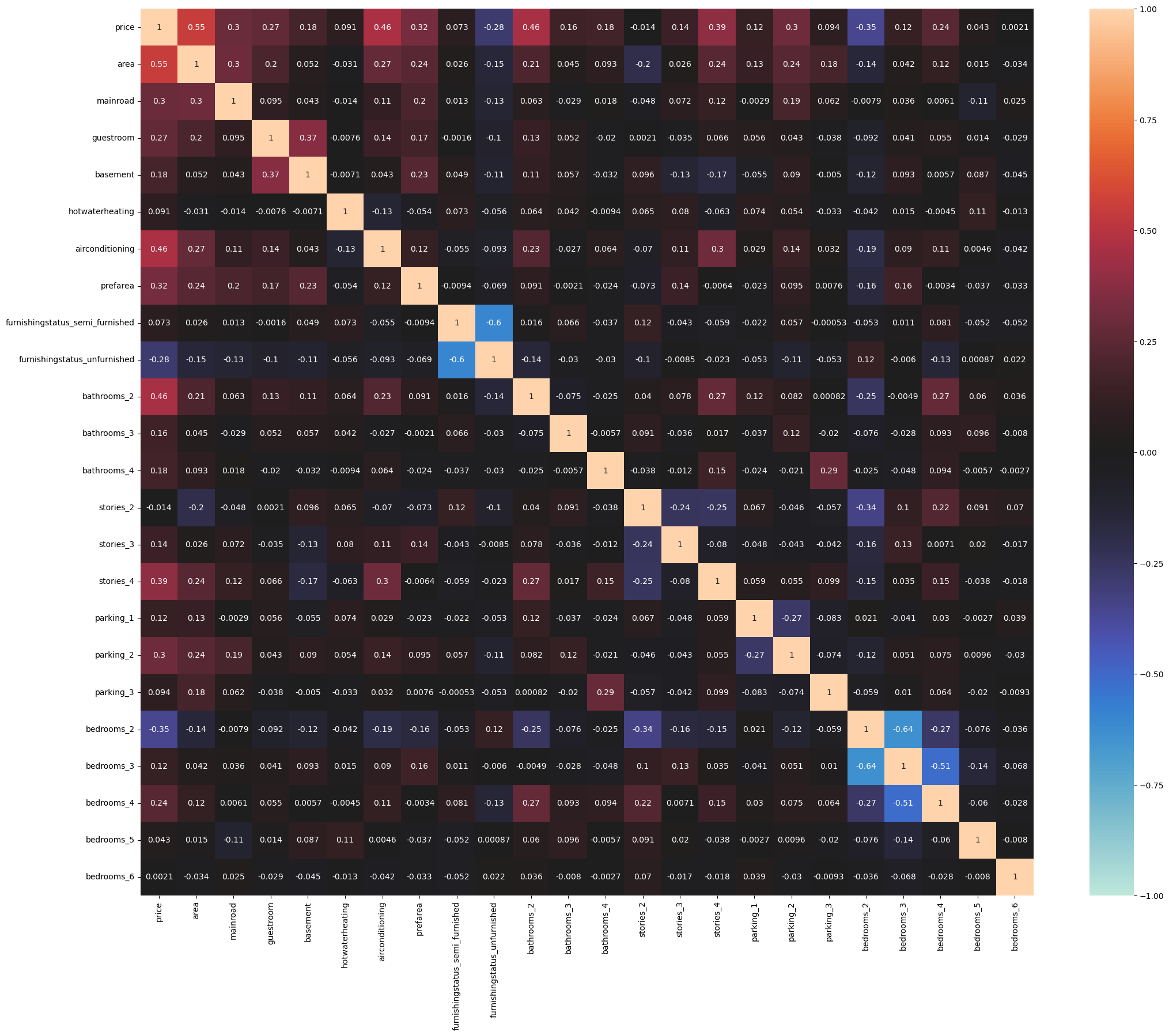
**Training & Testing Sets Data**

Original set : (533, 23) (533)

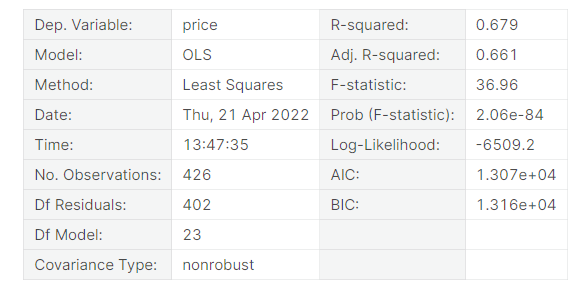
Training set : (426, 23) (426)

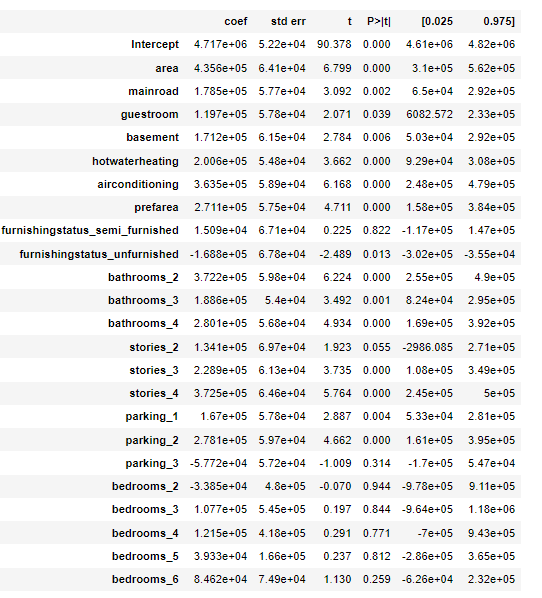
Testing set : (107, 23) (107)

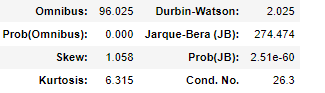
**Correlation Matrix**

****

**OLS Regression Results**

****

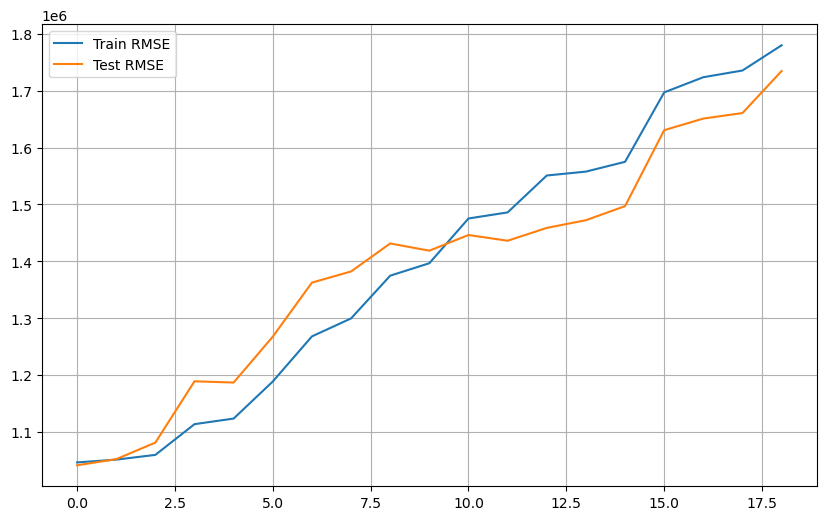




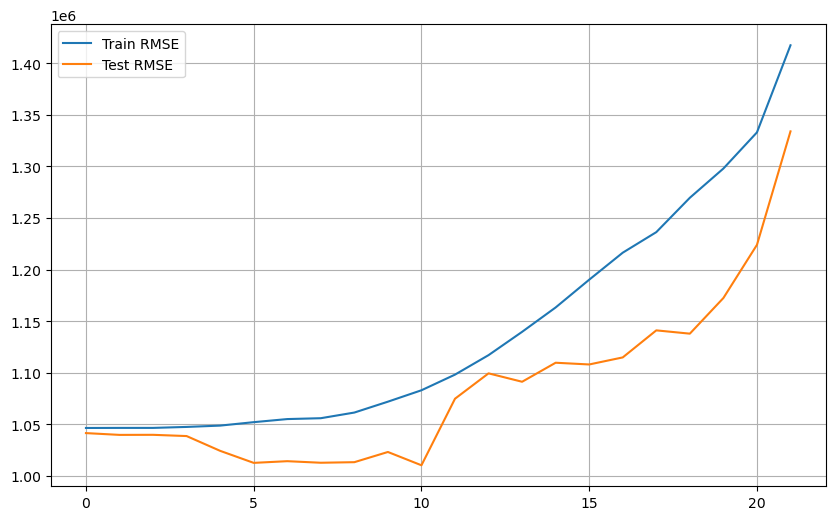
**Approach:**

1. Manual Method - Variance Inflation Factor (VIF)
2. Automatic Method – Recursive Feature Elimination (RFE)
3. Feature Elmination using PCA Decomposition

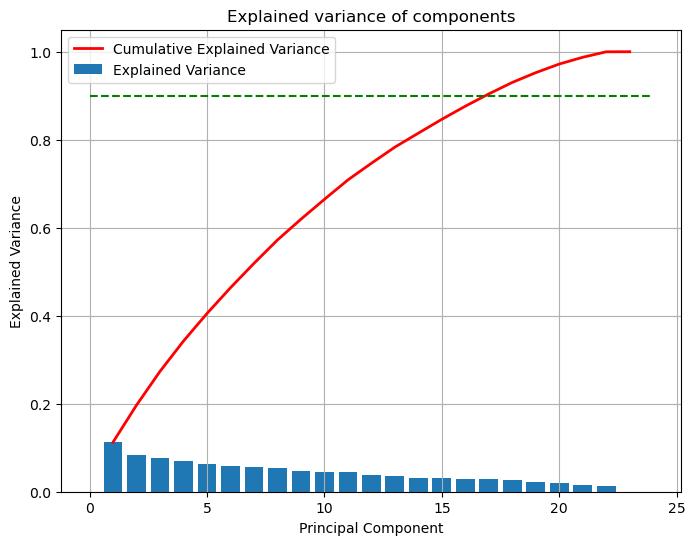
Manual Method - VIF

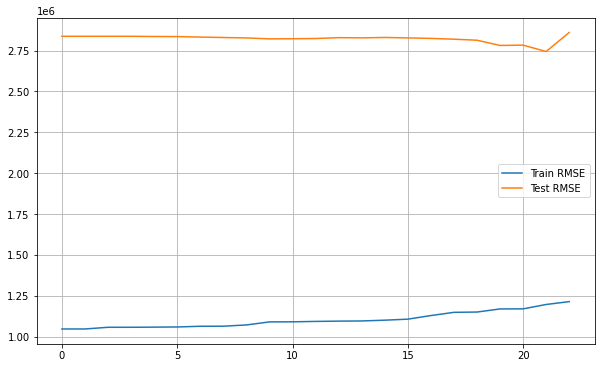


Automatic Method - RFE



PCA Decomposition





It can be seen that the performance of the modelsis quiet comparable unpon dropping features using VIF, RFE & PCA

Techniques. Comparing the RMSE plots, the optimal values were found for dropping most features using manual RFE

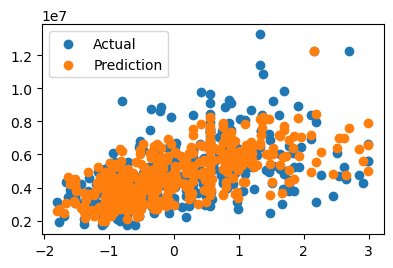
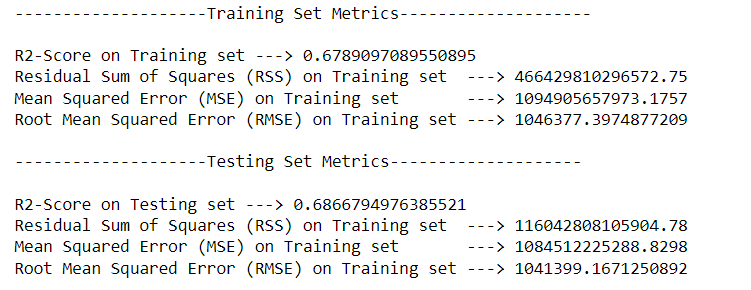
Technique. But let us skip these for now, as the advanced ML Algorithms take care of multicollinearity.

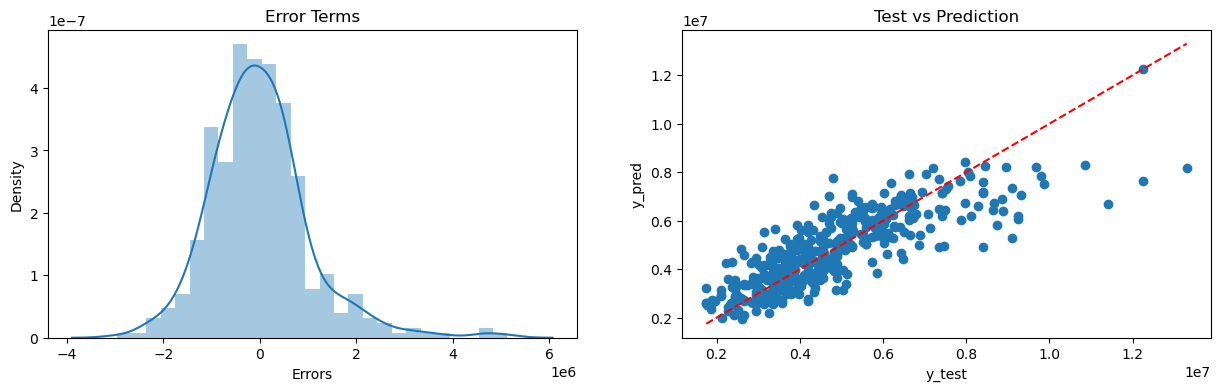
Modelling

**Objective** :

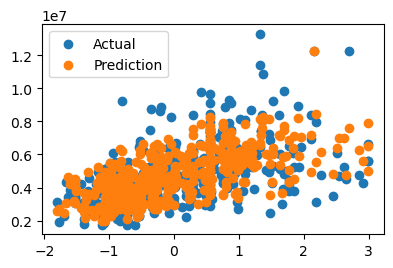
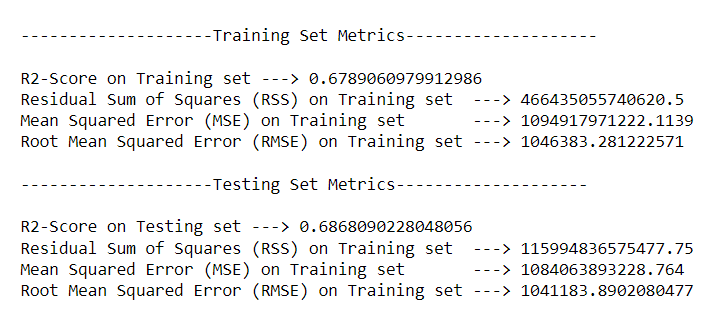
Let us now try building multiple regression models & compare their evaluation metrics to choose the best fit model both training and testing sets.

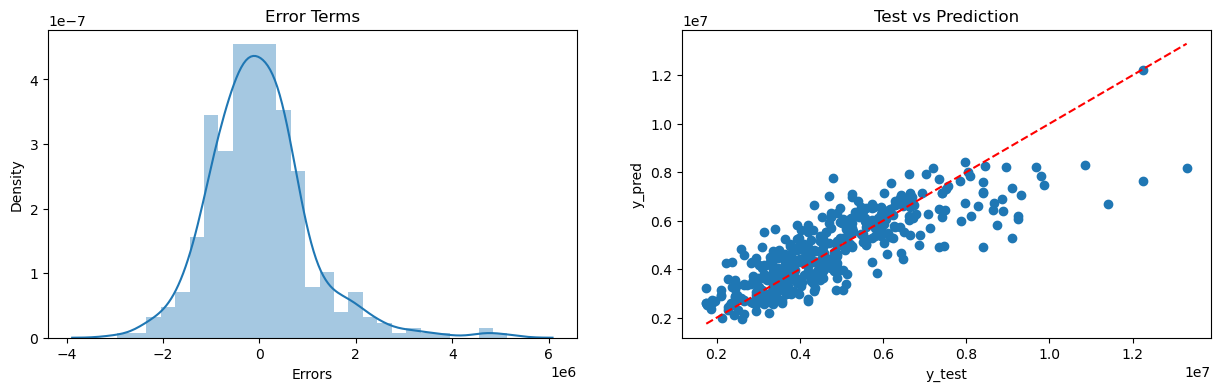
Multiple Linear Regression (MLR)



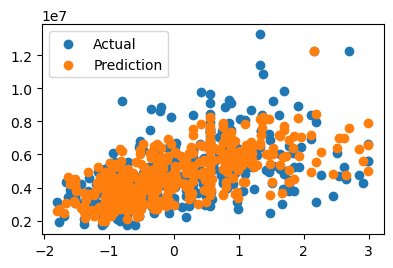
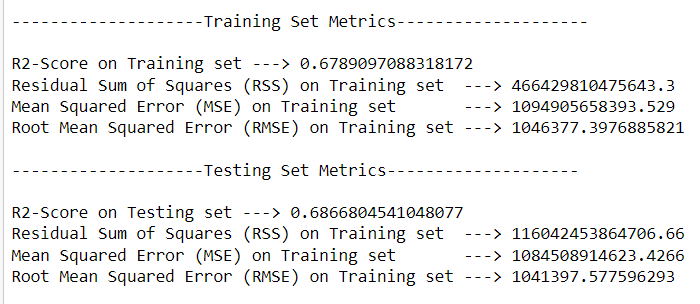


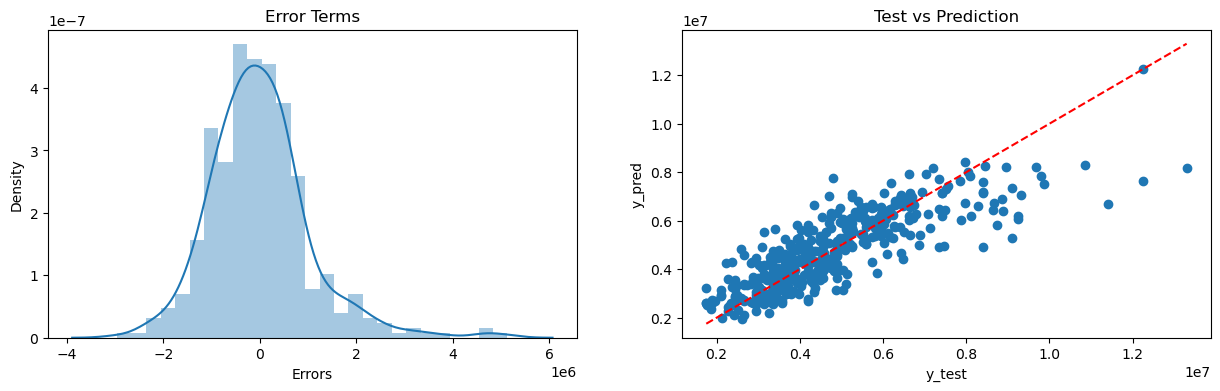
Ridge Rgeression Model



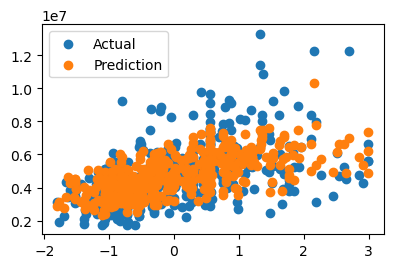
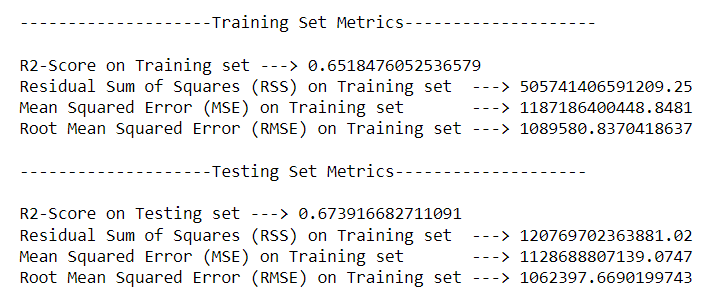


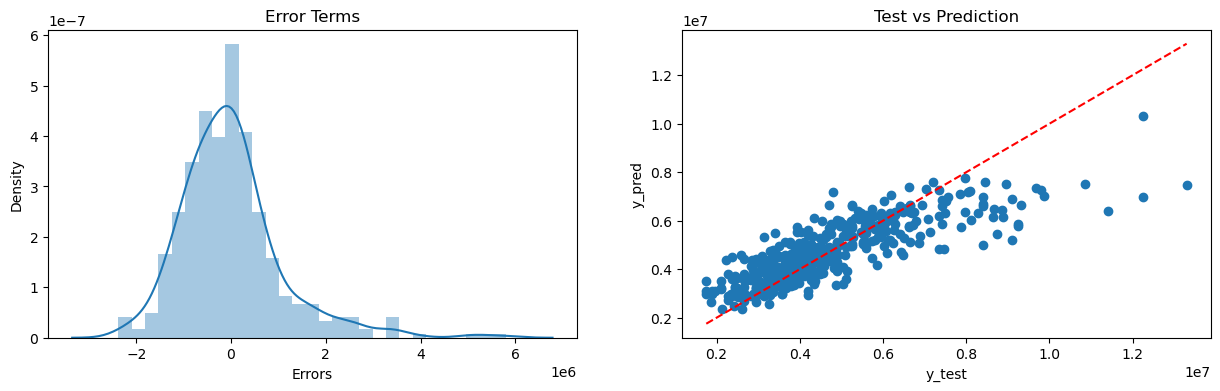
Lasso Regression Model



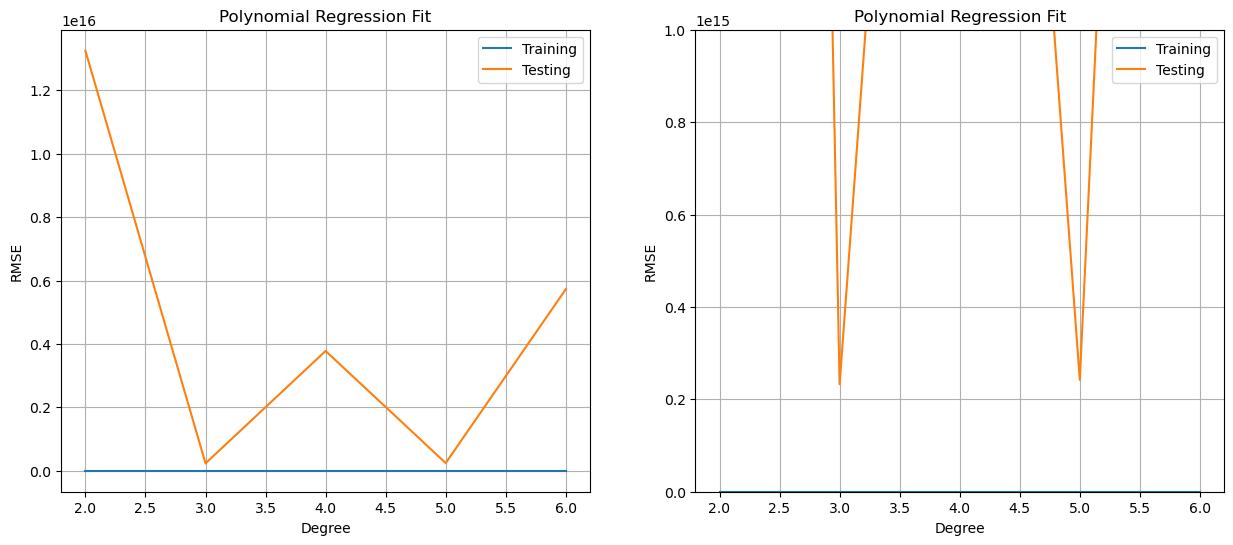


Elastic-Net Regression

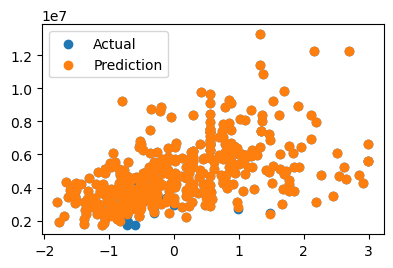
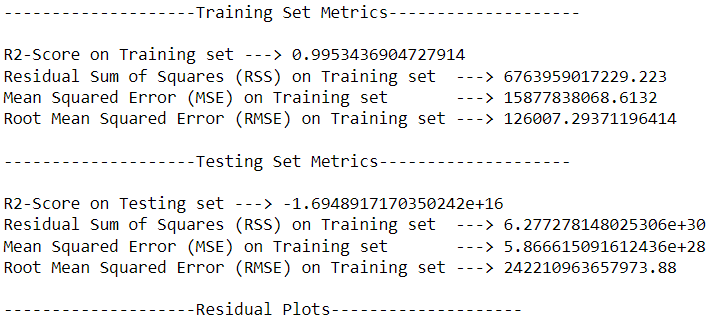


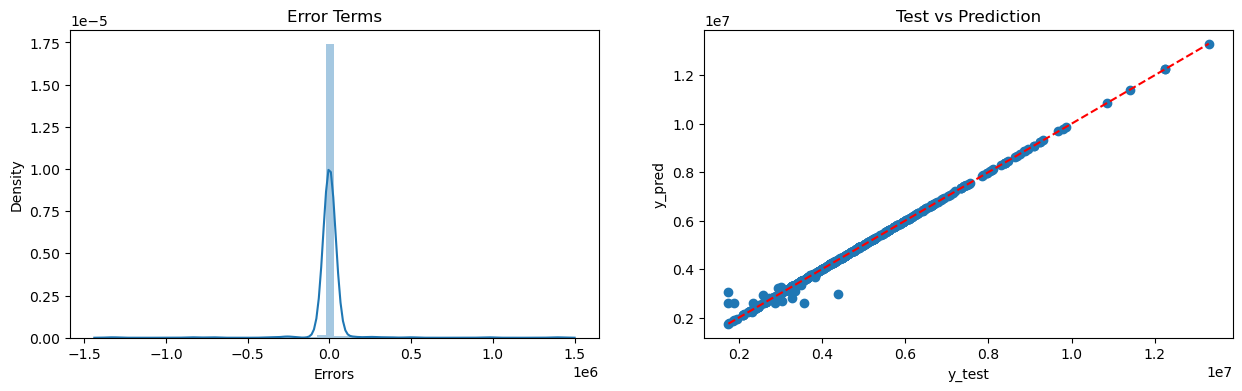


Polynomial Regression Model

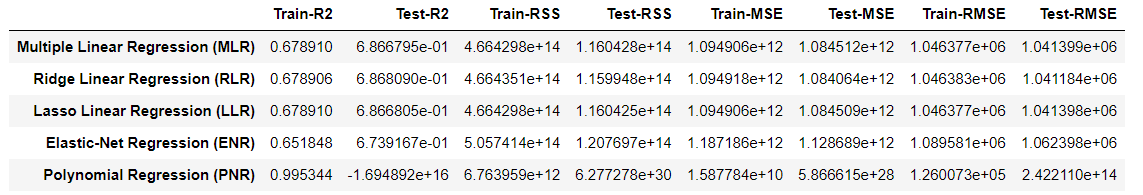


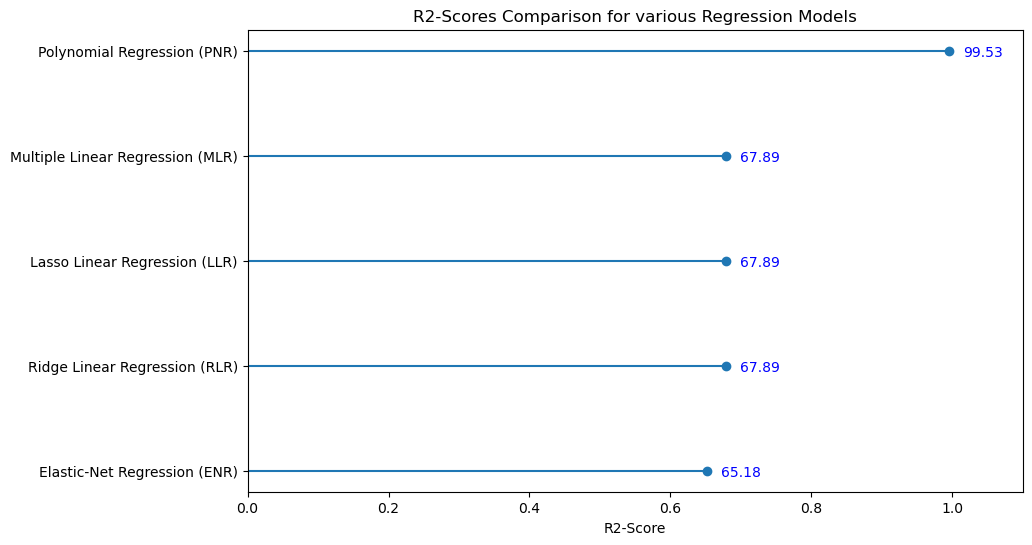
We can choose 5th order polynomial regression .



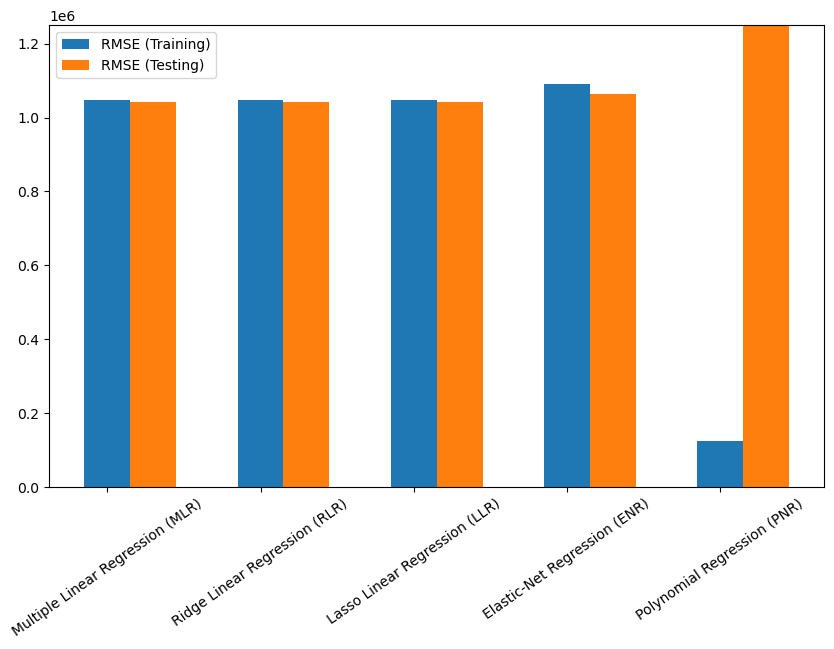


**Comparing the Evaluation Metics of the Models**

****

****

Root Mean square Error Comparison for different Regression Models



Provided the model should have close proximity with the training & testing scores.

It can be said that Polynomial regressions clearly overffitting the current problem. Hence, Simple MLR Model gave best results.

**Conclusions**

* The dataset was quiet small with just 545 samples & after preprocessing 2.2 % of the data samples were dropped.
* The features had high multicollinearity, hence in feature extraction step, we shortlisted the appropriate features with VIF Technique.
* Testing multiple algorithms with default hyper-paramters gave us some understanding for various models performance on this specific dataset.
* Polynomial Regression was the over-fitting, yet it is safe to use multiple regression algorithm, as their scores were quiet comparable & also they are more generalizable.

**Team Details** (Group-2)

|  |  |  |
| --- | --- | --- |
| **Name** | **Branch** | **Semester/Year** |
| Vikas Kumar Yadav | Information Technology | 5Th Sem / 3rd Year |
| Shalini Kumari | Information Technology | 5Th Sem / 3rd Year |
| Awnish Kumar | Information Technology | 5Th Sem / 3rd Year |

