**A PROJECT REPORT**

**ON**

##### “HOUSE PRICE PREDITION USING PYTHON”

**BY**

**Aditya Kedar**

.

**Index**

|  |  |  |
| --- | --- | --- |
| **SR.NO.** | **CONTENTS** | **PAGE NO.** |
| **1** | **Abstract** | **1** |
| **2** | **Objectives** | **2** |
| **3** | **Introduction** | **3** |
| **4** | **Statistical Software & Statistical Tools** | **5** |
| **5** | **Graphical Representation** | **7** |
| **6** | **Methodology** | **12** |
| **7** | **Model Building** | **15** |
| **8** | **Conclusion** | 25 |

# ABSTRACT

House price forecasting is an important topic of real estate. The literature attempts to derive useful knowledge from historical data of property markets. Machine learning techniques are applied to analyze historical property transactions in India to discover useful models for house buyers and sellers. Revealed is the high discrepancy between house prices in the most expensive and most affordable suburbs in the city of Pune.

Moreover, experiments demonstrate that the Multiple Linear Regression that is based on mean squared error measurement is a competitive approach.

## Objective

1. Create an effective price prediction model
2. Validate the model’s prediction accuracy
3. Identify the important home price attributes which feed the model’s predictive power
4. Understand the Dataset & cleanup (if required).
5. Build Regression models to predict the sales w.r.t a single & multiple features.
6. Also evaluate the models & compare their respective scores like R2, RMSE, etc.

# INTRODUCTION

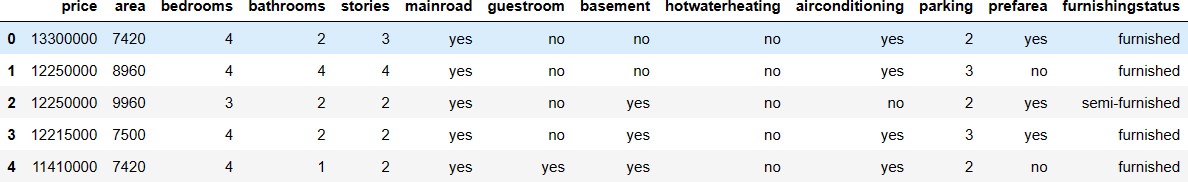
### Need and Motivation

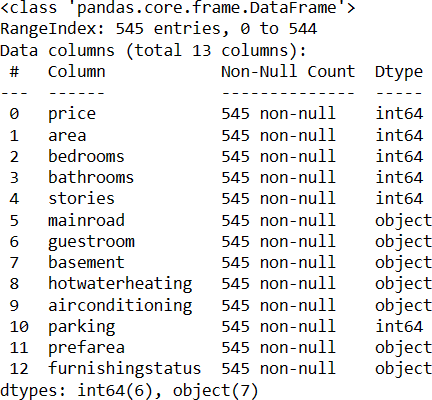
Having lived in India for so many years if there is one thing that I had been taking for granted, it’s that housing and rental prices continue to rise. Since the housing crisis of 2008, housing prices have recovered remarkably well, especially in major housing markets.

However, in the 4th quarter of 2016, I was surprised to read that Pune housing prices had fallen the most in the last 4 years. In fact, median resale prices for condos and coops fell 6.3%, marking the first time there was a decline since Q1 of 2017. The decline has been partly attributed to political uncertainty domestically and abroad and the 2014 election. So, to maintain the transparency among customers and also the comparison can be made easy through this model. If customer finds the price of house at some given website higher than the price predicted by the model, so he can reject that house.

### DATASET

Here we have the Data is taken from Broker in Pune. He is popular Real estate broker in Pune city. He gives the data in csv file format.





### Data Exploration

Data exploration is the first step in data analysis and typically involves summarizing the main characteristics of a data set, including its size, accuracy, initial patterns in the data and other attributes. It is commonly conducted by data analysts using visual analytics tools, but it can also be done in more advanced statistical software, Python. Before it can conduct analysis on data collected by multiple data sources and stored in data warehouses, an organization must know how many cases are in a data set, what variables are included, how many missing values there are and what general hypotheses the data is likely to support. An initial exploration of the data set can help answer these questions by familiarizing analysts with the

data with which they are working. We divided the data 9:1 for Training and Testing purpose respectively.

### Statistical software and statistical tool

* Advanced Excel
* Python
* Power Bi
* MS-Word

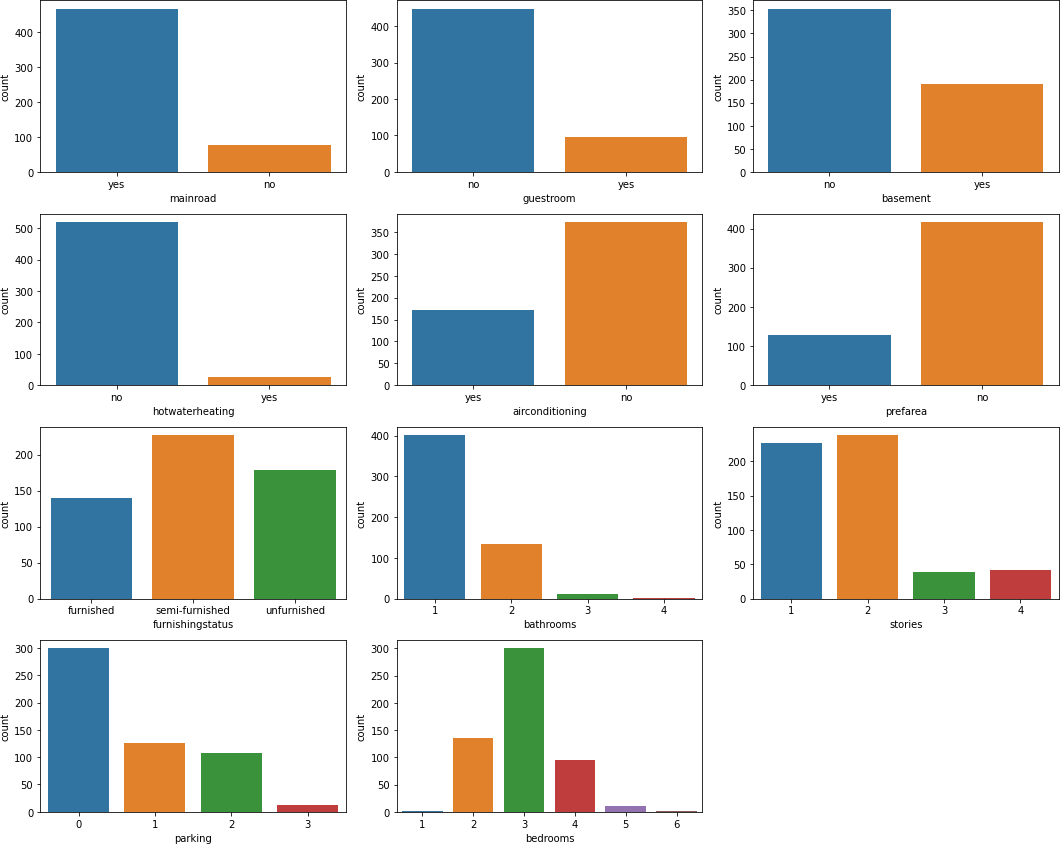
### Data Selection

Data selection is defined as the process of determining the appropriate data type and source, as well as suitable instruments to collect data. Data selection precedes the actual practice of data collection. This definition distinguishes data selection from selective data reporting (selectively excluding data that is not supportive of a research hypothesis) and interactive/active data selection (using collected data for monitoring activities/events, or conducting secondary data analyses). The process of selecting suitable data for a research project can impact data integrity. The primary objective of data selection is the determination of appropriate data type, source, and instrument(s) that allow investigators to adequately answer research questions. This determination is often discipline-specific and is primarily driven by the nature of the investigation, existing literature, and accessibility to necessary data sources.

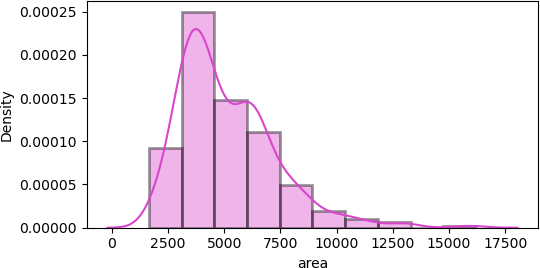
## Graphical Representation

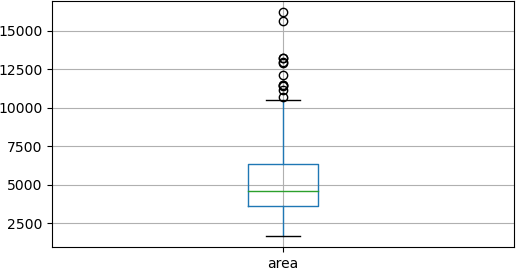
### Data Visualization

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. In the world of Big Data, data visualization tools and technologies are essential to analyses massive amounts of information and make data-driven decisions.

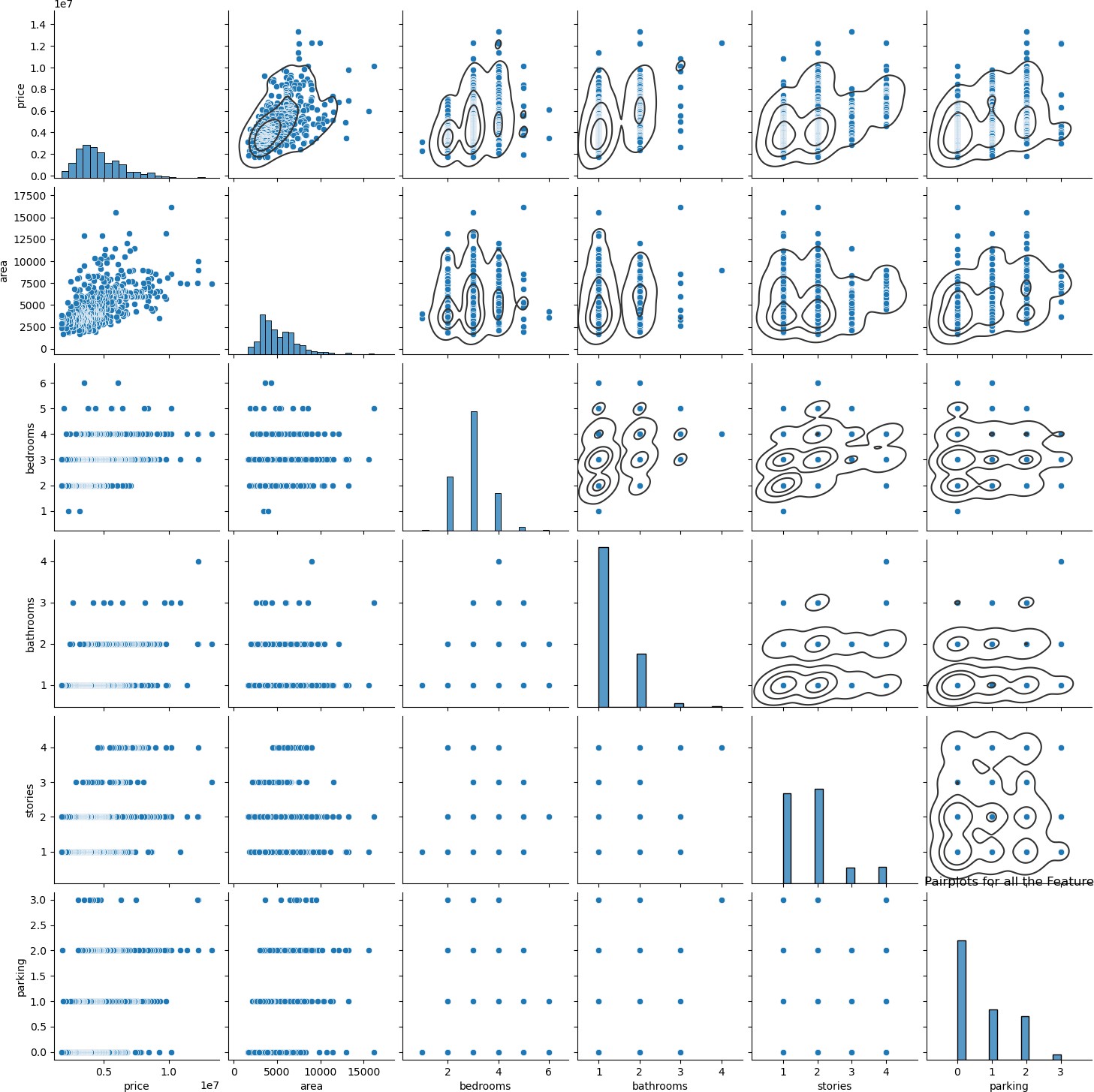


**Numeric Features Distribution**

****

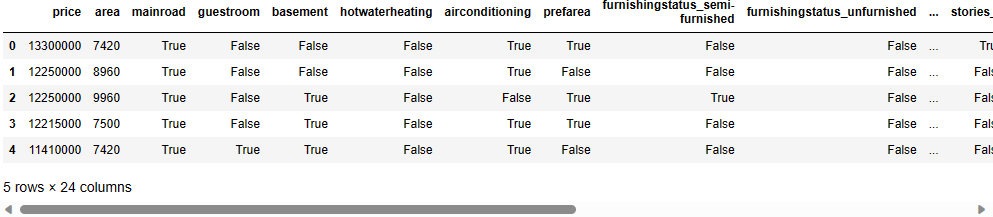
****

**Conclusion: -** According to the data there seen to be some outliers. Let us fix these in the upcoming section.



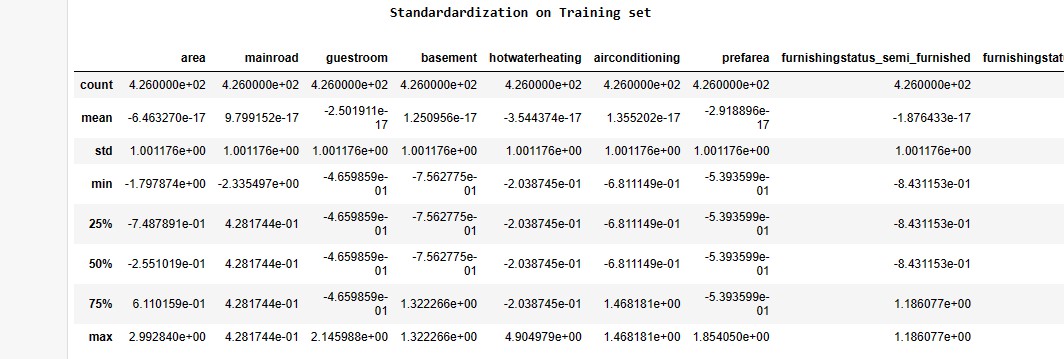
**Conclusion: -** According to the data set, We can notice that some features have linear relationship, let us further analyse the to detect the multicollinearity.

### Removal of outlier

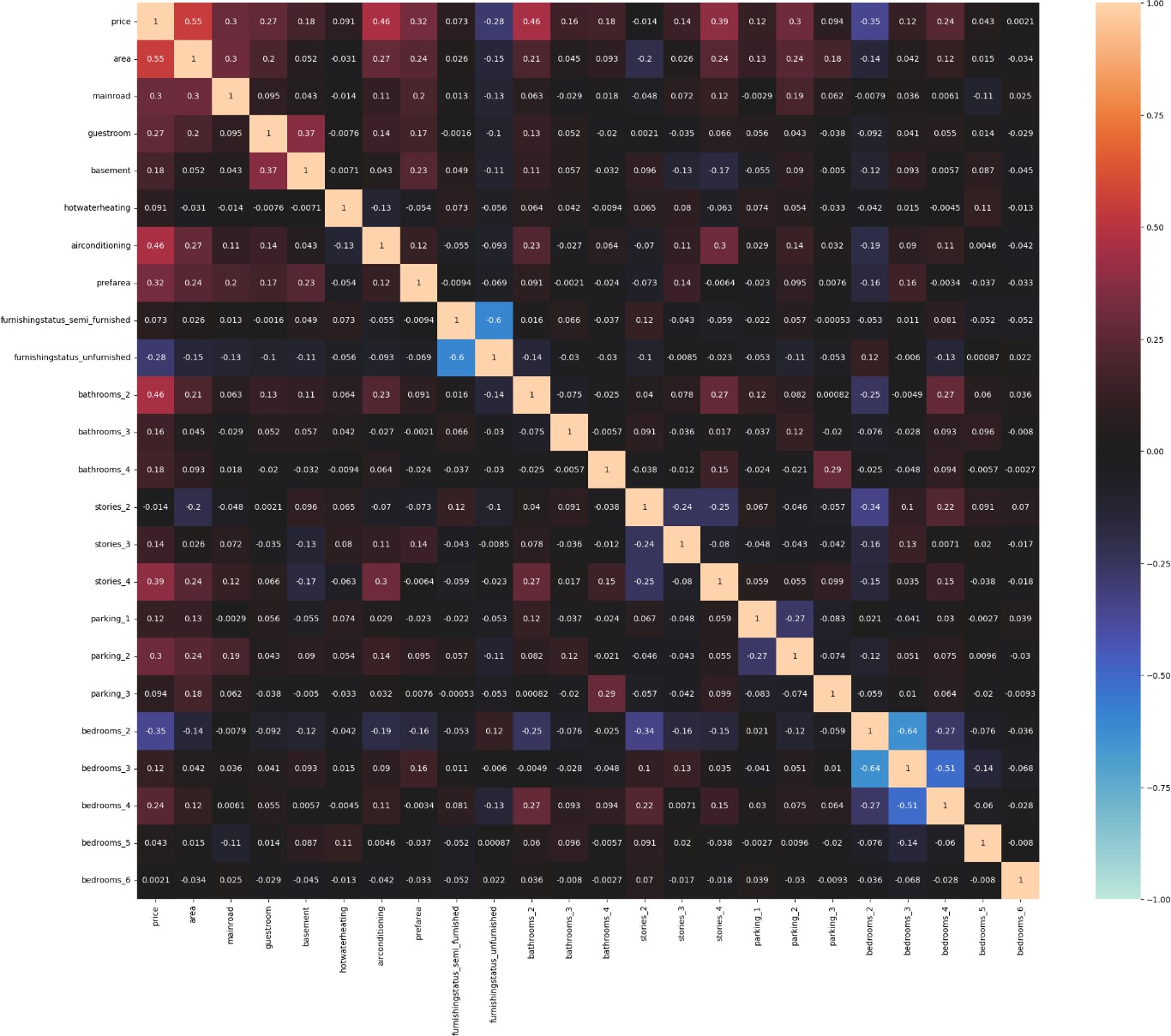
****

**Conclusion: -** Before removal of outliers, the dataset had 545 samples. After removal of outliers, the dataset now has 533 samples.

### Standardization the data

****

**Correlation Matrix**

****

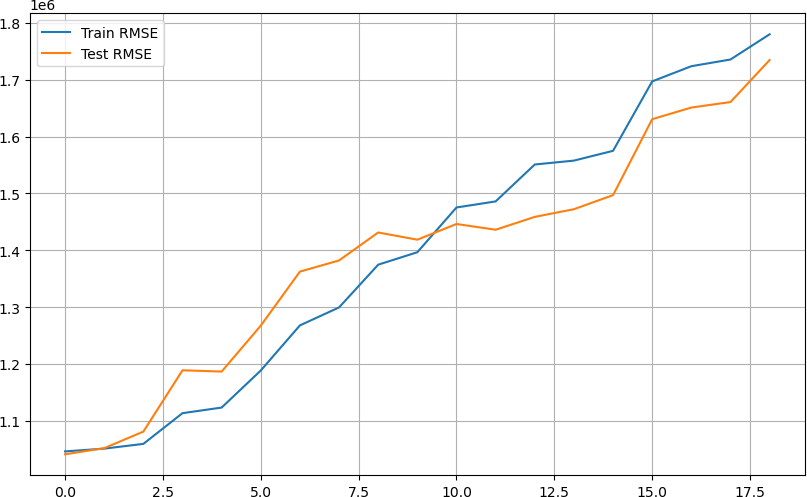
**Conclusion: -** According to data set correlation concept better understand using Heatmap. There seen to be multi-correlation between the features. Let us try to fix these. Using following methods.

### Approach: -

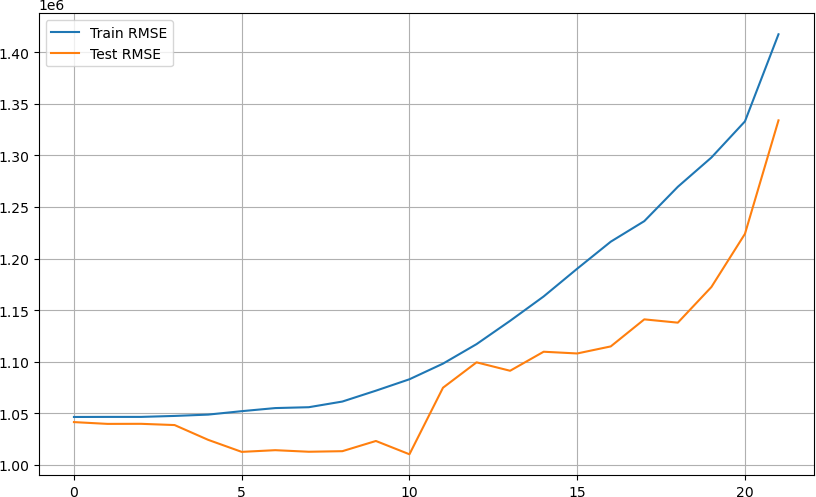
We can fix these multicollinearity with two techniques:

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

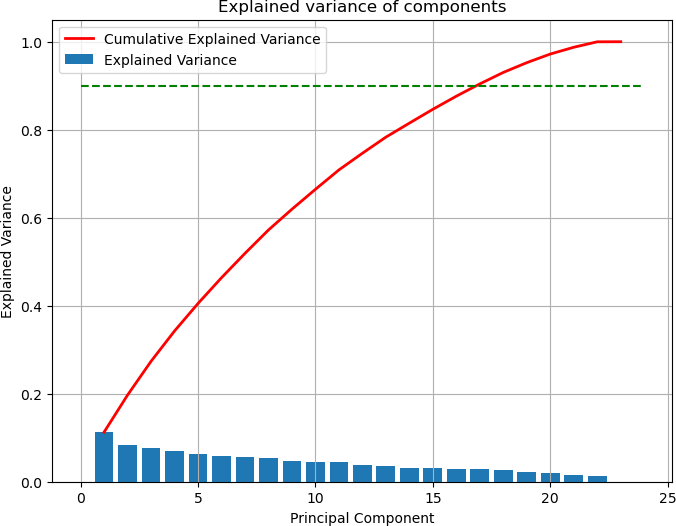
**VIF (**Variance Inflation Factor)

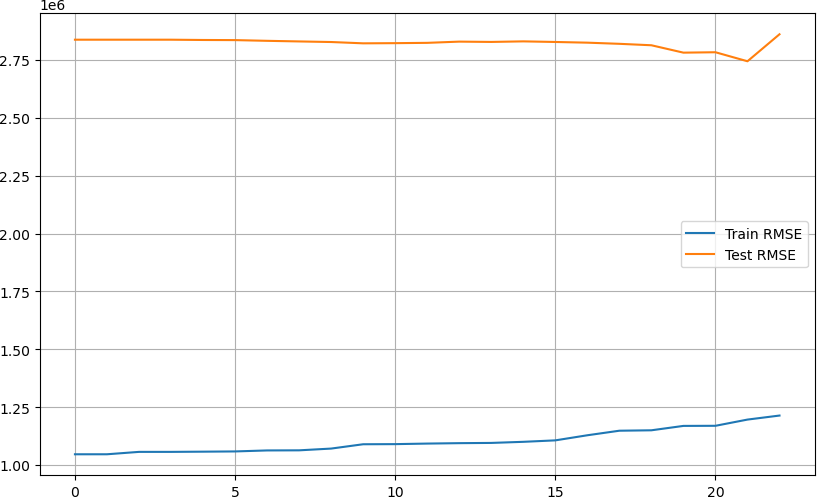


**RFE** (Recursive Feature Elimination)



**PCA** (Principle component analysis)

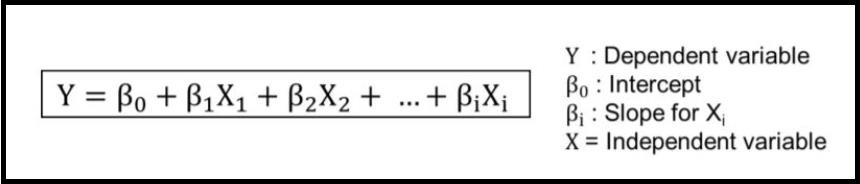


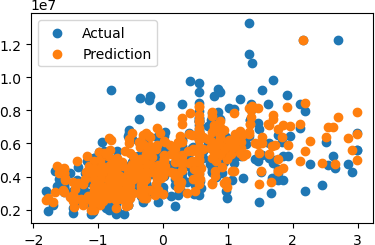


**Over all Conclusion: -** It can be seen that the performance of the models is quiet comparable upon dropping features using VIF, RFE & PCA Techniques. Comparing the RMSE plots, the optimal values were found for dropping most features using manual RFE Technique. But let us skip these for now, as the advanced ML Algorithms take care of multicollinearity.

### Predictive Modelling

#### Multiple Linear Regression (MLR):-





**Training Set Metrics**

R2-Score on Training set ---> 0.6789060979912986

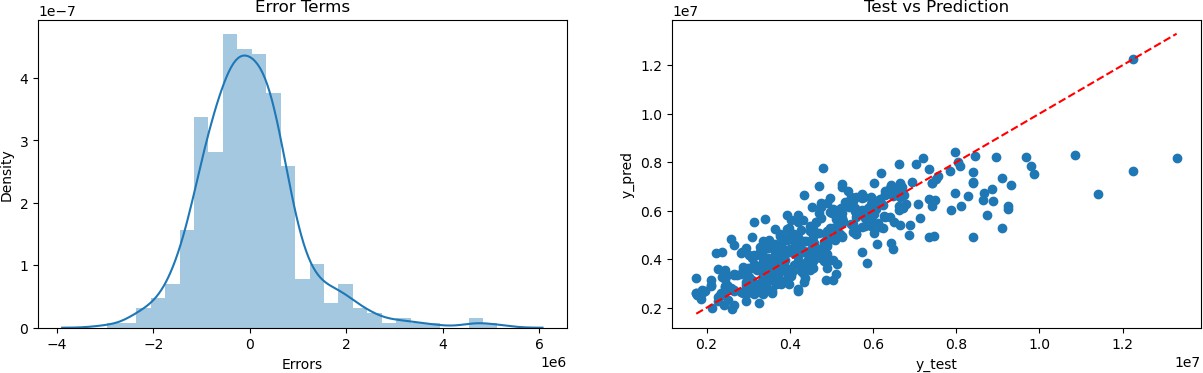
Residual Sum of Squares (RSS) on Training set ---> 466435055740620.5 Mean Squared Error (MSE) on Training set ---> 1094917971222.1139 Root Mean Squared Error (RMSE) on Training set ---> 1046383.281222571

**Testing Set Metrics**

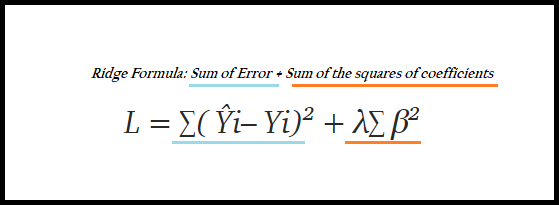
R2-Score on testing set ---> 0.6868090228048056

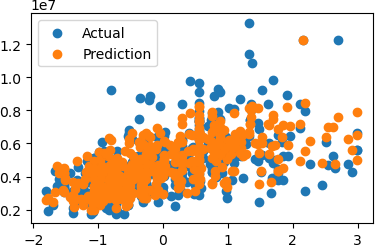
Residual Sum of Squares (RSS) on Training set ---> 115994836575477.75 Mean Squared Error (MSE) on Training set ---> 1084063893228.764 Root Mean Squared Error (RMSE) on Training set ---> 1041183.8902080477

**Residual Plots**

****

#### Ridge Regression Model:-





**Training Set Metrics**

R2-Score on Training set ---> 0.6789060979912986

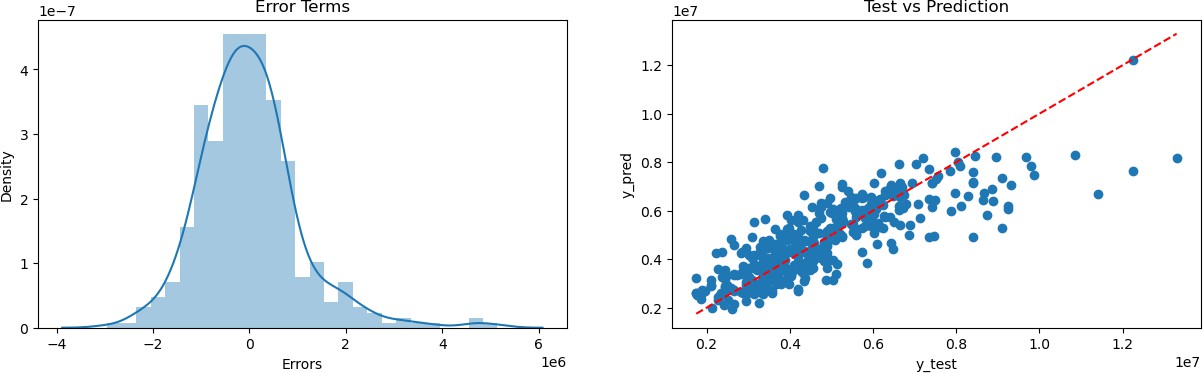
Residual Sum of Squares (RSS) on Training set ---> 466435055740620.5 Mean Squared Error (MSE) on Training set ---> 1094917971222.1139 Root Mean Squared Error (RMSE) on Training set ---> 1046383.281222571

**Testing Set Metrics**

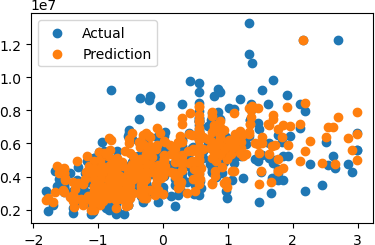
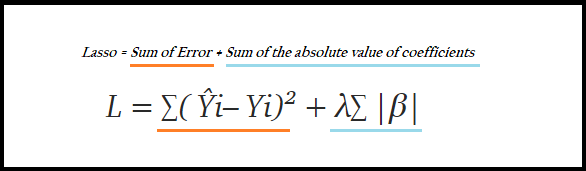
R2-Score on testing set ---> 0.6868090228048056

Residual Sum of Squares (RSS) on Training set ---> 115994836575477.75 Mean Squared Error (MSE) on Training set ---> 1084063893228.764 Root Mean Squared Error (RMSE) on Training set ---> 1041183.8902080477

**Residual Plots**

****

#### Lasso Regression Model:-



**Training Set Metrics**

R2-Score on Training set ---> 0.6789097088318172

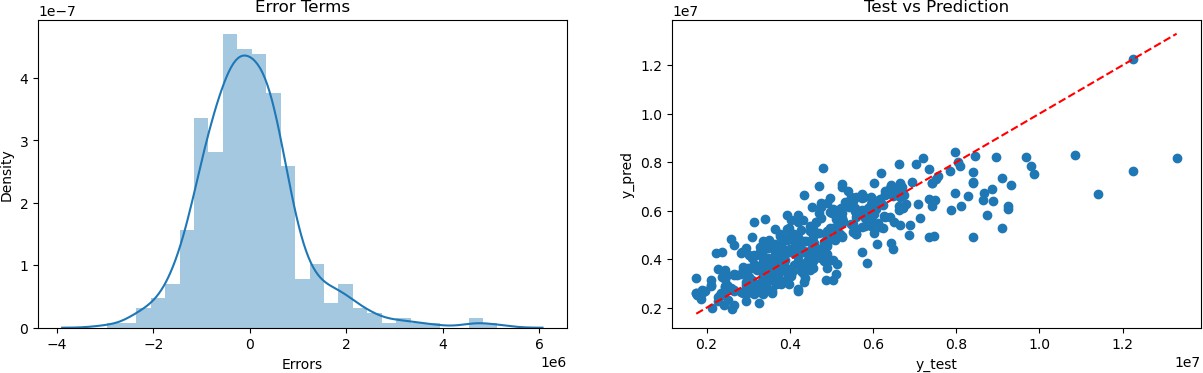
Residual Sum of Squares (RSS) on Training set ---> 466429810475643.3 Mean Squared Error (MSE) on Training set ---> 1094905658393.529 Root Mean Squared Error (RMSE) on Training set ---> 1046377.3976885821

**Testing Set Metrics**

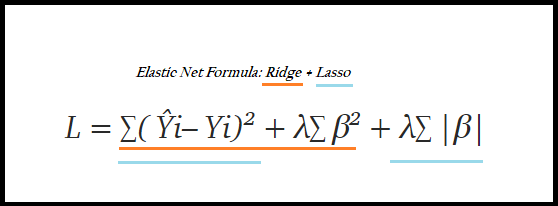
R2-Score on testing set ---> 0.6866804541048077

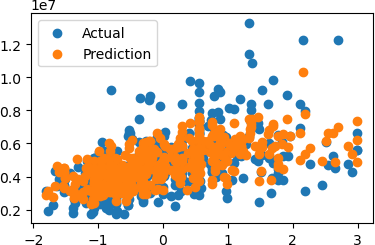
Residual Sum of Squares (RSS) on Training set ---> 116042453864706.66 Mean Squared Error (MSE) on Training set ---> 1084508914623.4266 Root Mean Squared Error (RMSE) on Training set ---> 1041397.577596293

**Residual Plots**

****

#### Elastic-Net Regression:-





**Training Set Metrics**

R2-Score on Training set ---> 0.6518476052536579

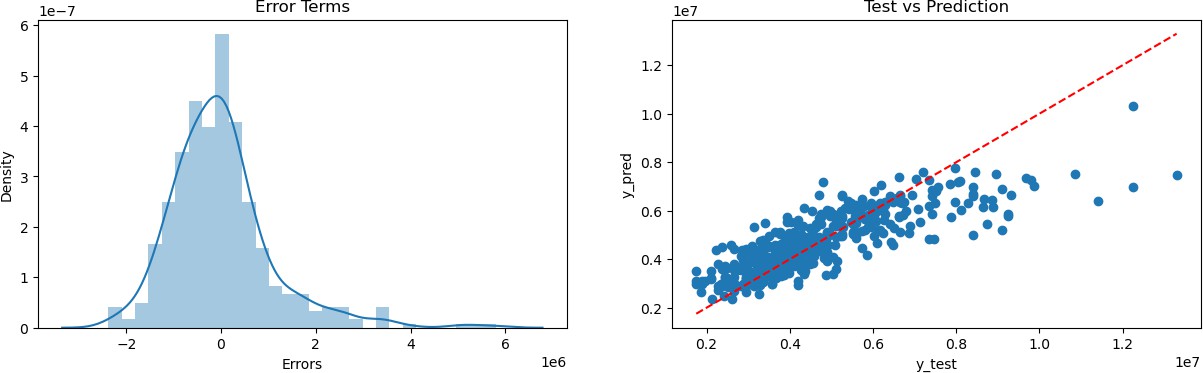
Residual Sum of Squares (RSS) on Training set ---> 505741406591209.25 Mean Squared Error (MSE) on Training set ---> 1187186400448.8481 Root Mean Squared Error (RMSE) on Training set ---> 1089580.8370418637

**Testing Set Metrics**

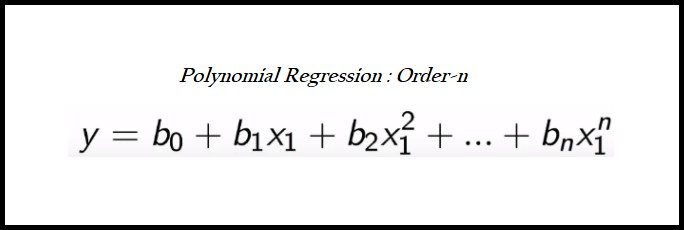
R2-Score on testing set ---> 0.673916682711091

Residual Sum of Squares (RSS) on Training set ---> 120769702363881.02 Mean Squared Error (MSE) on Training set ---> 1128688807139.0747 Root Mean Squared Error (RMSE) on Training set ---> 1062397.6690199743

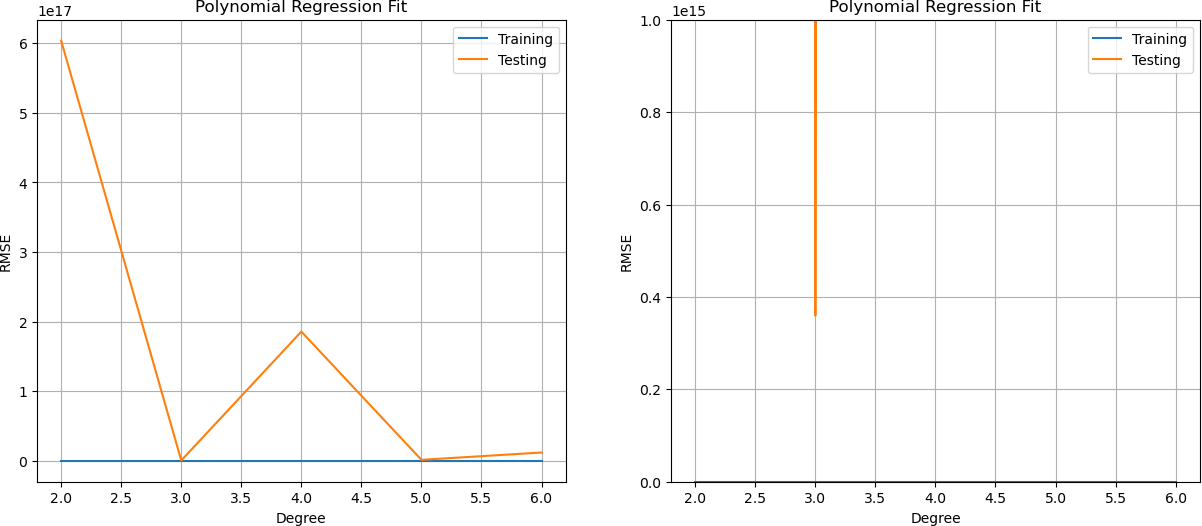
**Residual Plots**

****

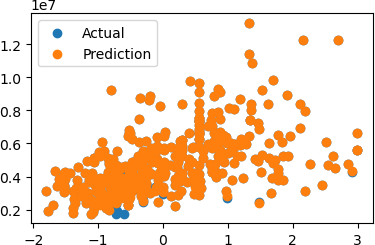
#### Polynomial Regression Model:-



##### Checking polynomial regression performance on various degrees

****

**Conclusion: -** We can choose 5th order polynomial regression as it gives the optimal training & testing scores.



**Training Set Metrics**

R2-Score on Training set ---> 0.9953436876838865

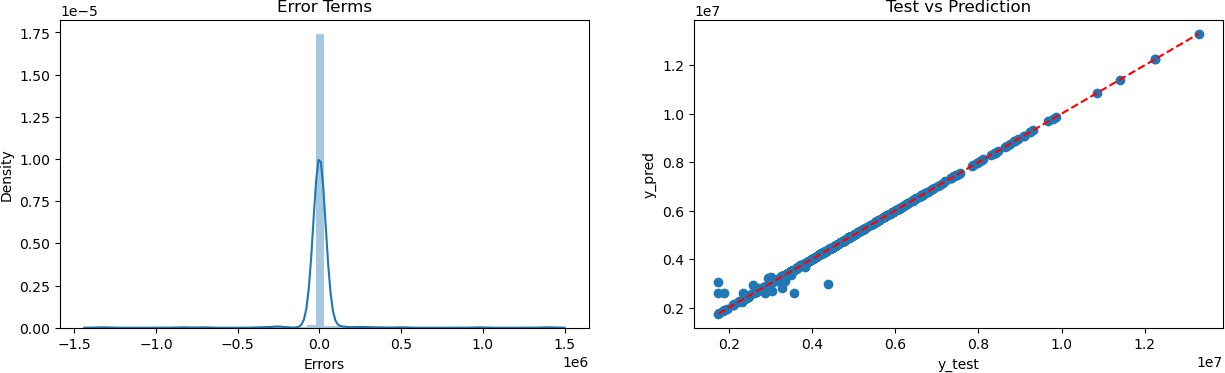
Residual Sum of Squares (RSS) on Training set ---> 6763963068514.785 Mean Squared Error (MSE) on Training set ---> 15877847578.673206 Root Mean Squared Error (RMSE) on Training set ---> 126007.33144810742

**Testing Set Metrics**

R2-Score on testing set ---> -4.6732570382202554e+17

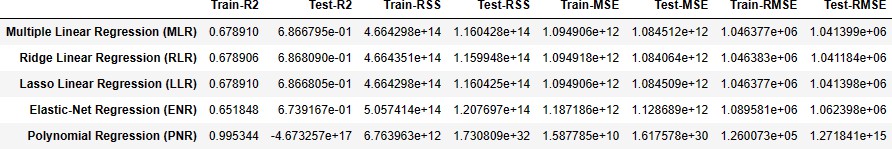
Residual Sum of Squares (RSS) on Training set ---> 1.7308087585349424e+32 Mean Squared Error (MSE) on Training set ---> 1.617578279004619e+30 Root Mean Squared Error (RMSE) on Training set ---> 1271840508477623.2

**Residual Plots**

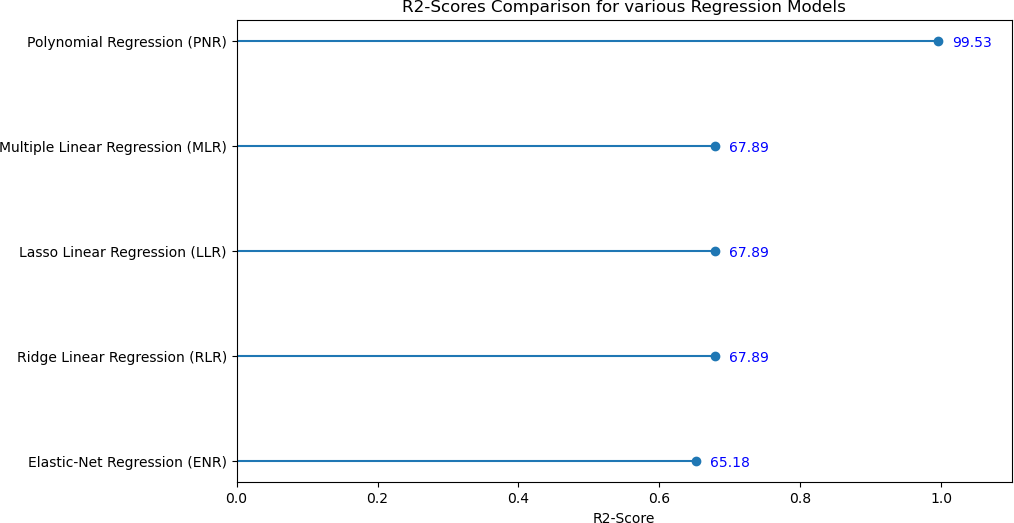
****

### Comparing the Evaluation Metrics of the Models

**Regression Models Results Evaluation**

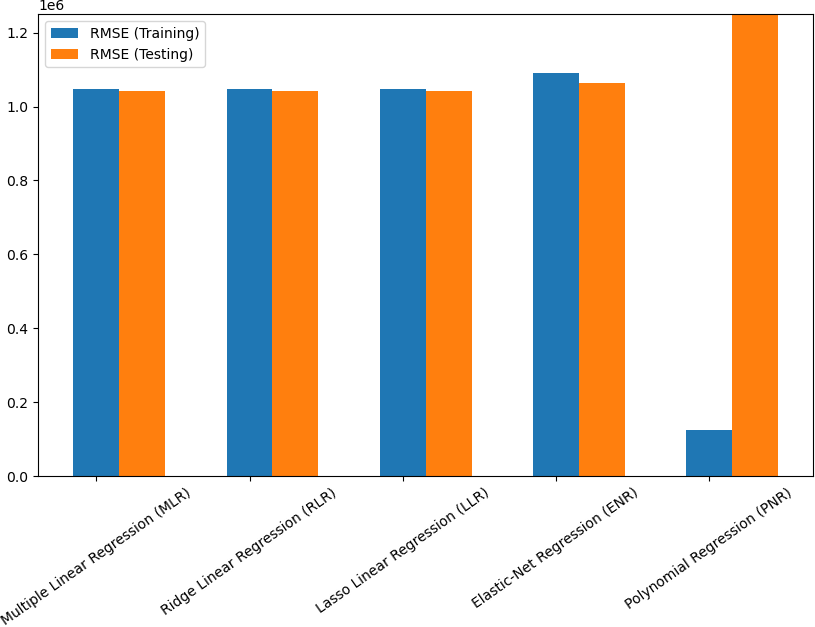
****

**R2-Scores Comparison for different Regression Models**

****

**Conclusion: -** From the above plot, it is clear that the polynomial regression models have the highest explain ability power to understand the dataset.

**Root Mean Squared Error Comparison for different Regression Models**

****

**Conclusion: -** Lesser the RMSE, better the model! Also, provided the model should have close proximity with the training & testing scores. For this problem, it is can be said that polynomial regressions clearly over fitting the current problem. Surprisingly simple MLR Model gave the best results.

### Project Outcomes & Conclusions

* The Dataset was quiet small with just 545 samples & after pre-processing 2.2% of the data samples were dropped.
* Visualising the distribution of data & their relationships, helped us to get some insights on the feature-set.
* The features had high multicollinearity, hence in Feature Extraction step, we shortlisted the appropriate features with VIF Technique.
* Testing multiple algorithms with default hyperparamters gave us some understanding for various models performance on this specific dataset.
* While, Polynomial Regression (Order-2) was the over fitting, yet it is safe to use multiple regression algorithm, as their scores were quiet comparable & also they're more generalizable.