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.

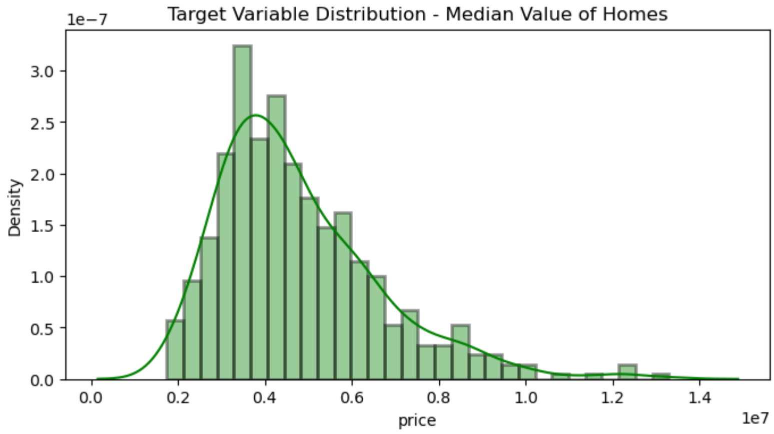
Here are some of the necessary steps:

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

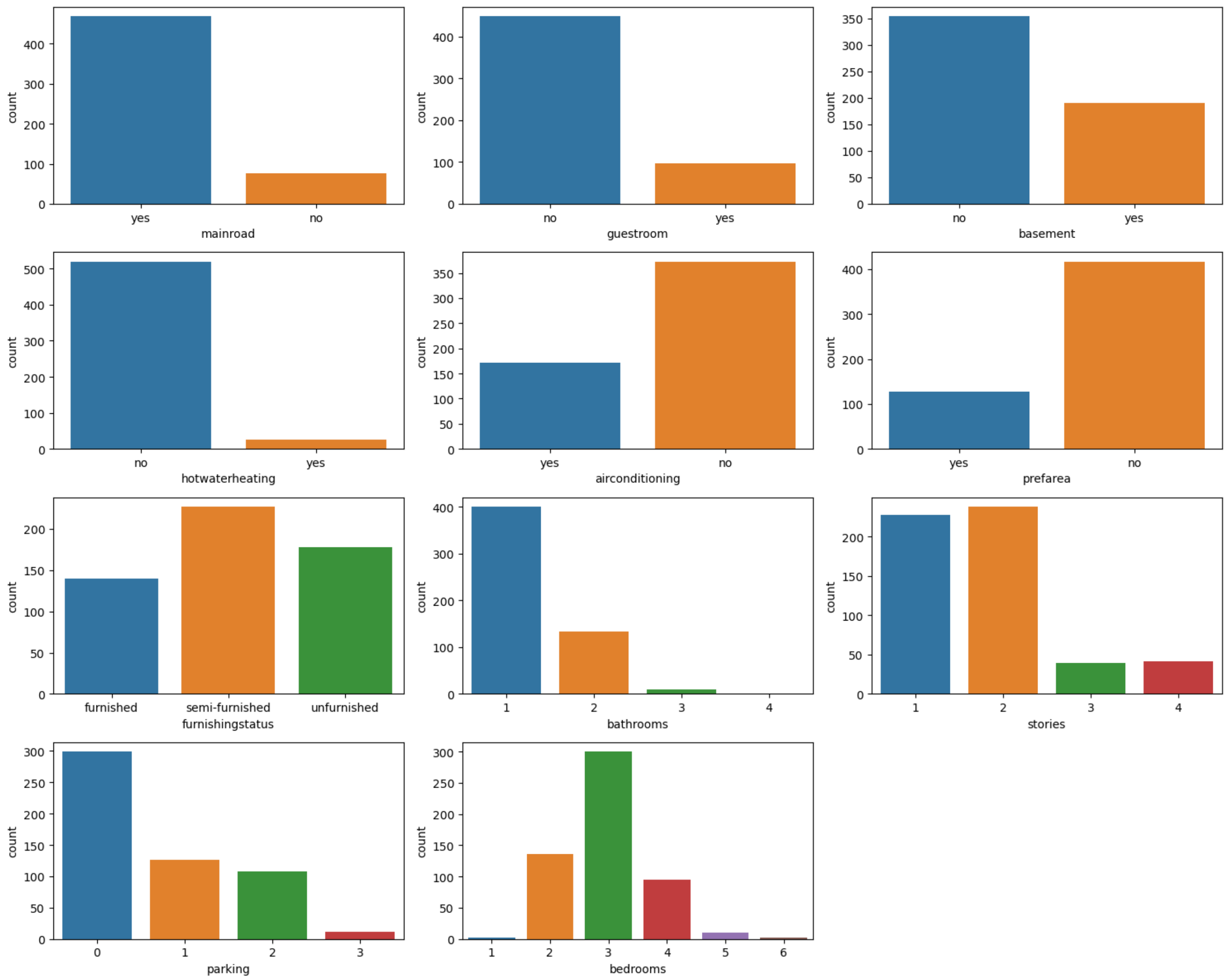
Exploratory Data Analysis (EDA)

❖ Analyze the distribution of the target Variable

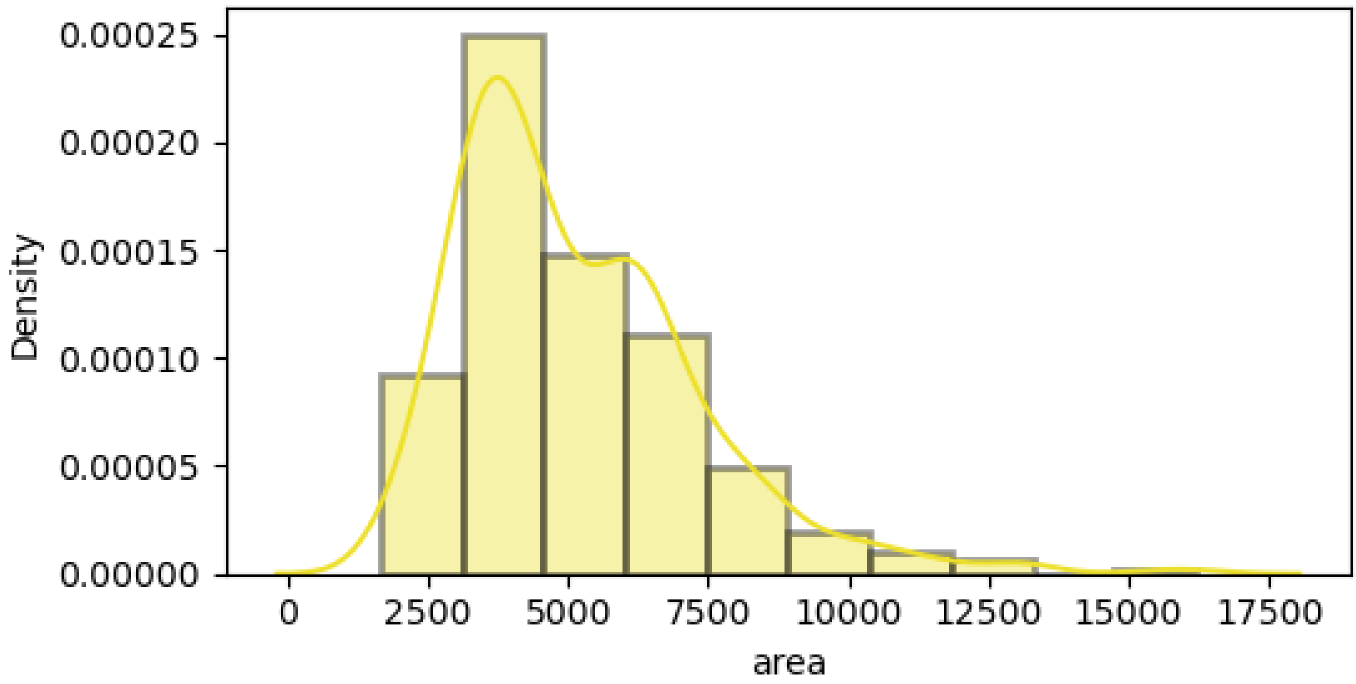


✓ The Target Variable seems to be normally distributed, averaging around 20 units.

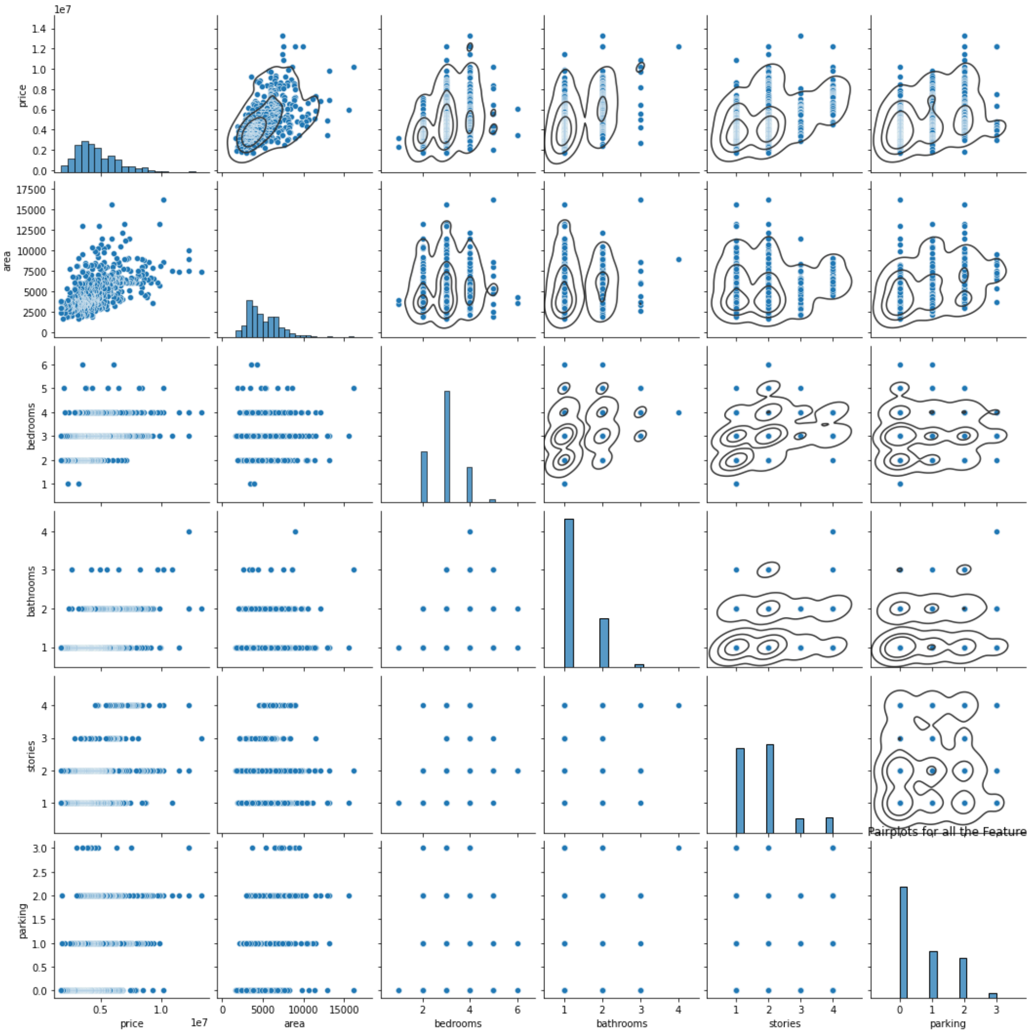
Visualising Categorical Features



Visualising the Numeric features

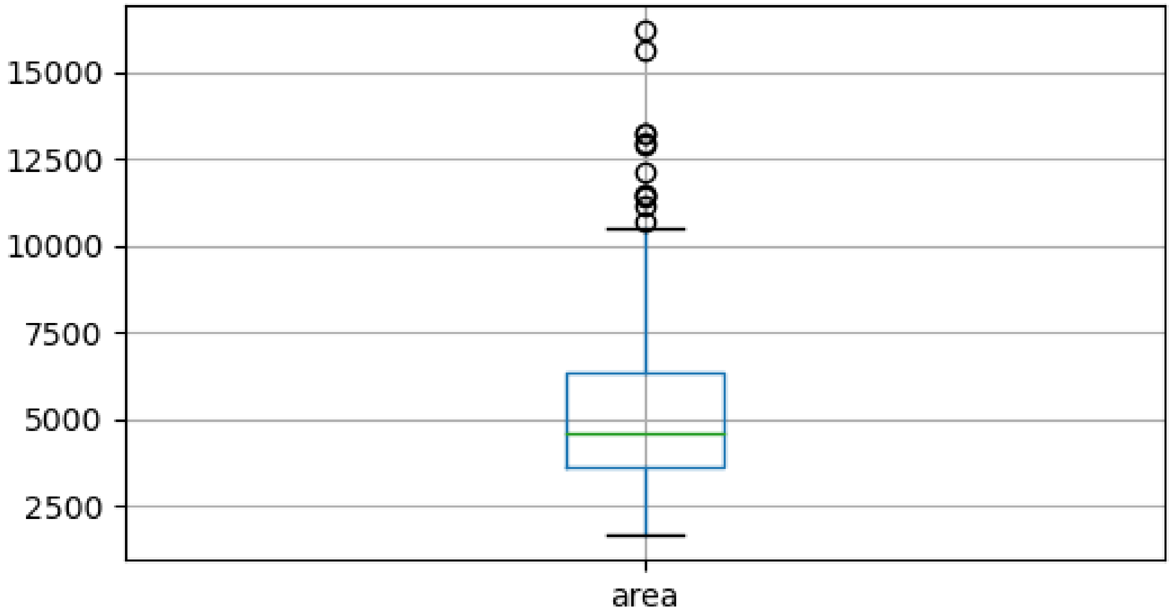


Relationship between all features



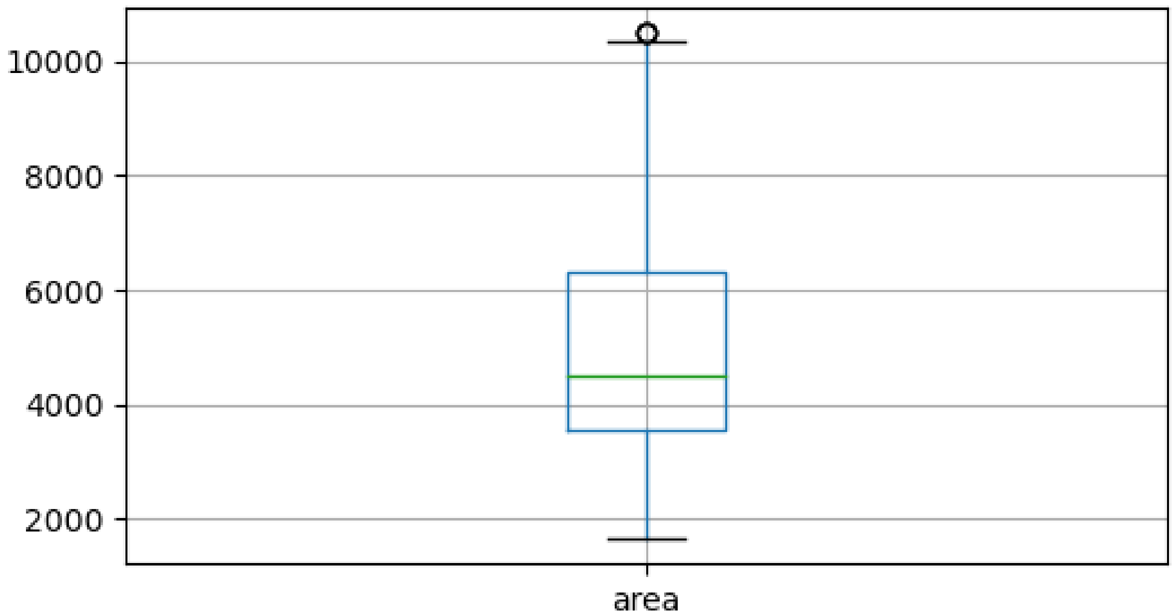
✓ We can notice that some features have linear relationship, let us futher analyze the detect multicollinearity.

After removal of outlier



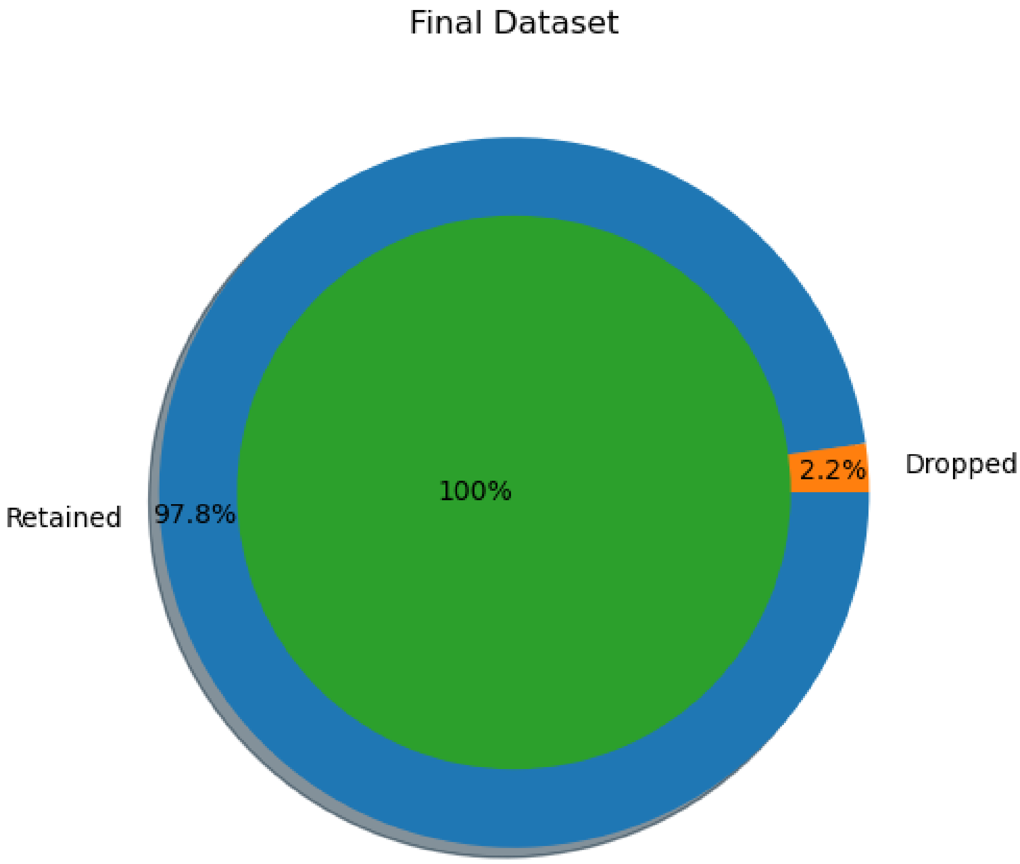
- ✓ The data set had 545 sample

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.

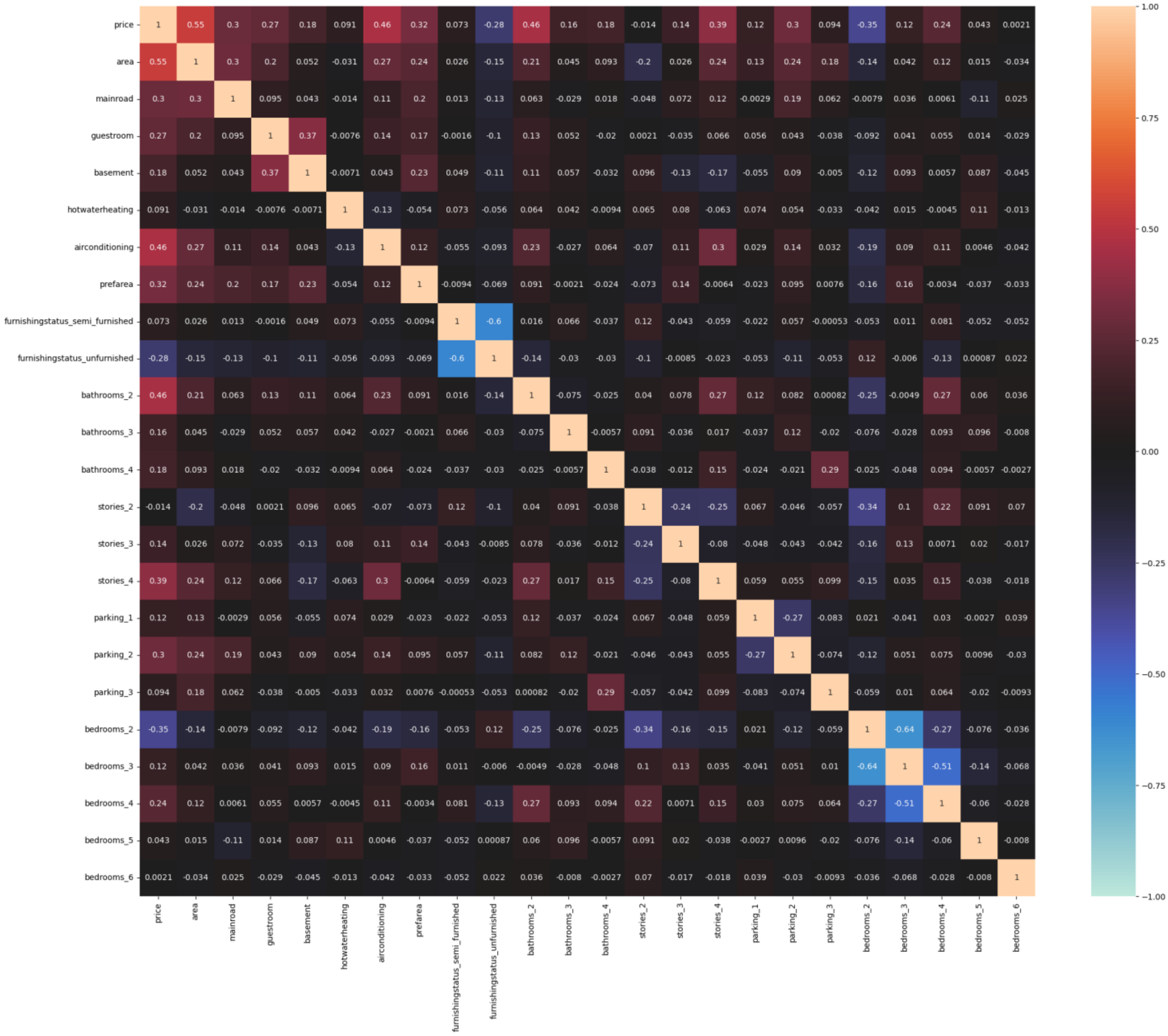
Training & Testing Sets Data

Original set : (533, 23) (533)

Training set : (426, 23) (426)

Testing set : (107, 23) (107)

Correlation Matrix



OLS Regression Results

Dep. Variable:	price	R-squared:	0.679
Model:	OLS	Adj. R-squared:	0.661
Method:	Least Squares	F-statistic:	36.96
Date:	Thu, 21 Apr 2022	Prob (F-statistic):	2.06e-84
Time:	13:47:35	Log-Likelihood:	-6509.2
No. Observations:	426	AIC:	1.307e+04
Df Residuals:	402	BIC:	1.316e+04
Df Model:	23		
Covariance Type:	nonrobust		

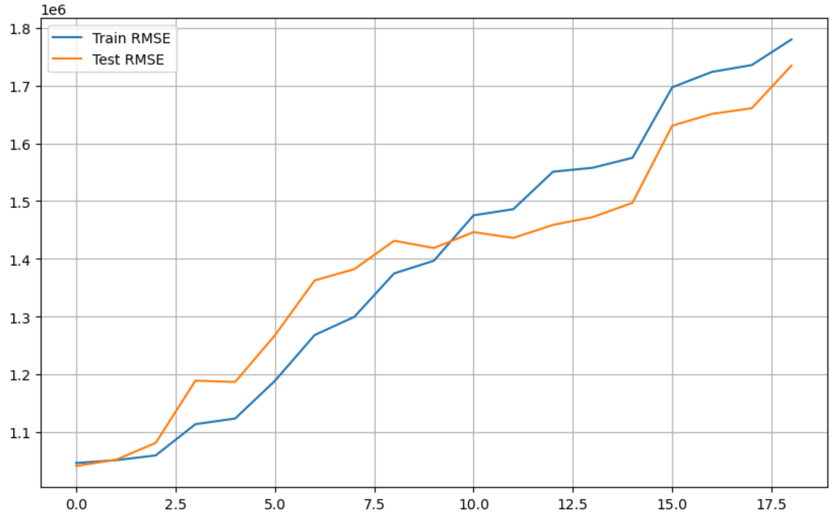
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.717e+06	5.22e+04	90.378	0.000	4.61e+06	4.82e+06
area	4.356e+05	6.41e+04	6.799	0.000	3.1e+05	5.62e+05
mainroad	1.785e+05	5.77e+04	3.092	0.002	6.5e+04	2.92e+05
guestroom	1.197e+05	5.78e+04	2.071	0.039	6082.572	2.33e+05
basement	1.712e+05	6.15e+04	2.784	0.006	5.03e+04	2.92e+05
hotwaterheating	2.006e+05	5.48e+04	3.662	0.000	9.29e+04	3.08e+05
airconditioning	3.635e+05	5.89e+04	6.168	0.000	2.48e+05	4.79e+05
prefarea	2.711e+05	5.75e+04	4.711	0.000	1.58e+05	3.84e+05
furnishingstatus_semi_furnished	1.509e+04	6.71e+04	0.225	0.822	-1.17e+05	1.47e+05
furnishingstatus_unfurnished	-1.688e+05	6.78e+04	-2.489	0.013	-3.02e+05	-3.55e+04
bathrooms_2	3.722e+05	5.98e+04	6.224	0.000	2.55e+05	4.9e+05
bathrooms_3	1.886e+05	5.4e+04	3.492	0.001	8.24e+04	2.95e+05
bathrooms_4	2.801e+05	5.68e+04	4.934	0.000	1.69e+05	3.92e+05
stories_2	1.341e+05	6.97e+04	1.923	0.055	-2986.085	2.71e+05
stories_3	2.289e+05	6.13e+04	3.735	0.000	1.08e+05	3.49e+05
stories_4	3.725e+05	6.46e+04	5.764	0.000	2.45e+05	5e+05
parking_1	1.67e+05	5.78e+04	2.887	0.004	5.33e+04	2.81e+05
parking_2	2.781e+05	5.97e+04	4.662	0.000	1.61e+05	3.95e+05
parking_3	-5.772e+04	5.72e+04	-1.009	0.314	-1.7e+05	5.47e+04
bedrooms_2	-3.385e+04	4.8e+05	-0.070	0.944	-9.78e+05	9.11e+05
bedrooms_3	1.077e+05	5.45e+05	0.197	0.844	-9.64e+05	1.18e+06
bedrooms_4	1.215e+05	4.18e+05	0.291	0.771	-7e+05	9.43e+05
bedrooms_5	3.933e+04	1.66e+05	0.237	0.812	-2.86e+05	3.65e+05
bedrooms_6	8.462e+04	7.49e+04	1.130	0.259	-6.26e+04	2.32e+05

Omnibus:	96.025	Durbin-Watson:	2.025
Prob(Omnibus):	0.000	Jarque-Bera (JB):	274.474
Skew:	1.058	Prob(JB):	2.51e-60
Kurtosis:	6.315	Cond. No.	26.3

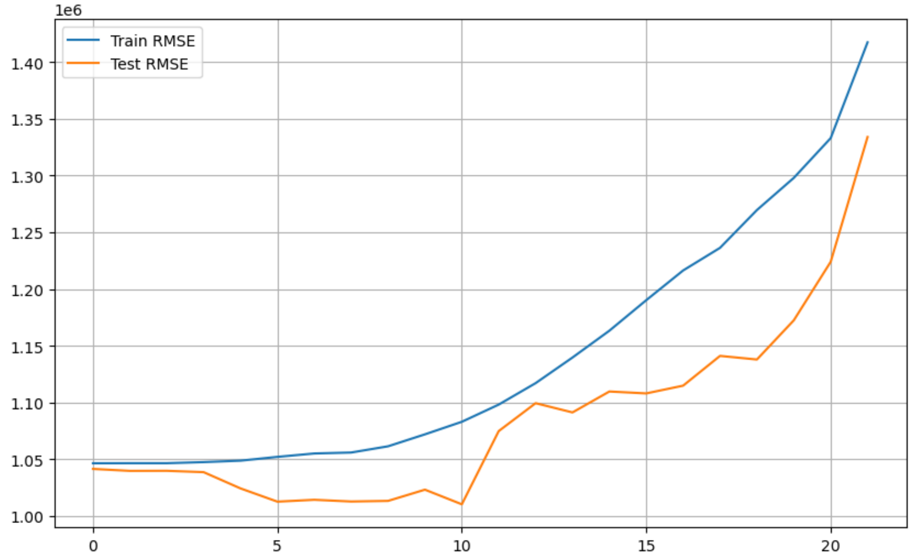
Approach:

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

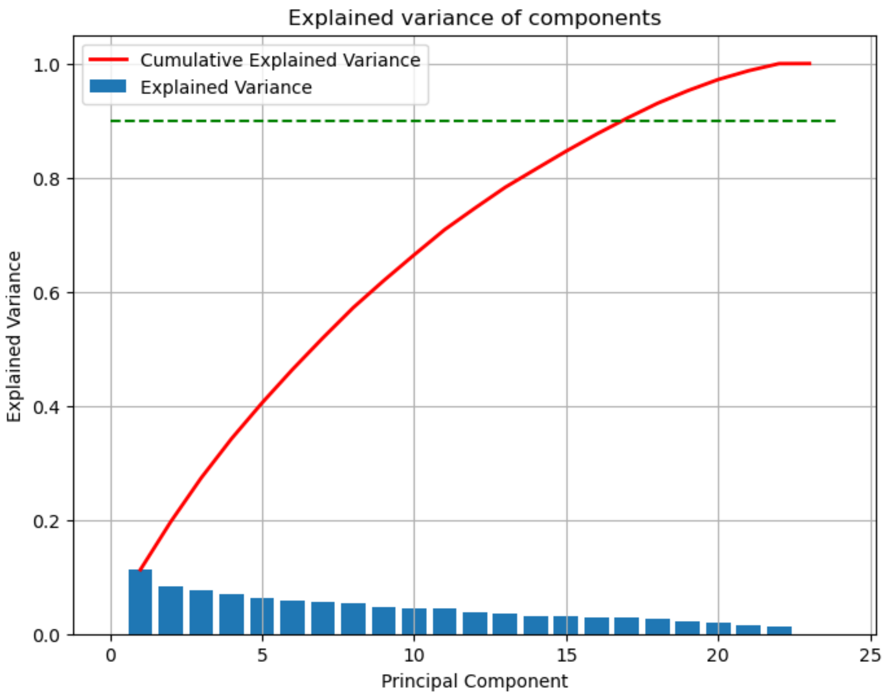
Manual Method - VIF

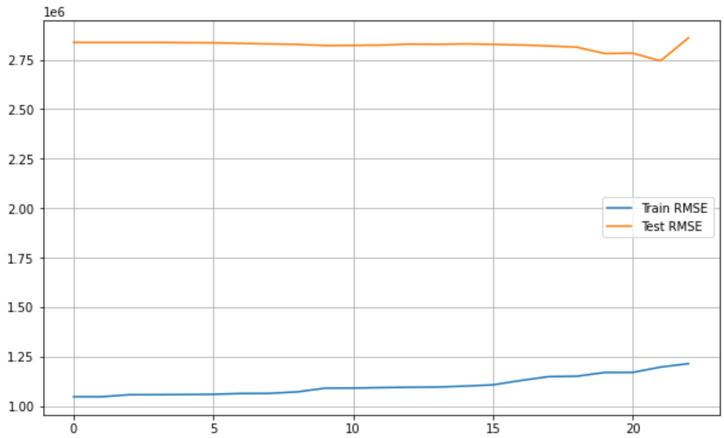


Automatic Method - RFE



PCA Decomposition





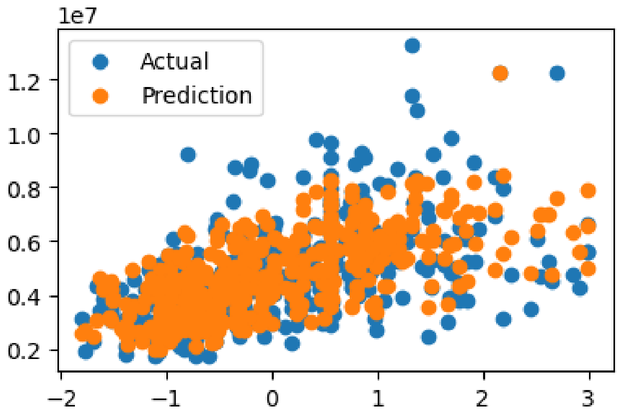
It can be seen that the performance of the modelsis quiet comparable unpon dropping features using VIF, RFE & PCA Techniques. Comparing the RMSE plots, the optimal values were found for dropping most features using manual RFE Technique. But let us skip these for now, as the advanced ML Algorithms take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

Modelling

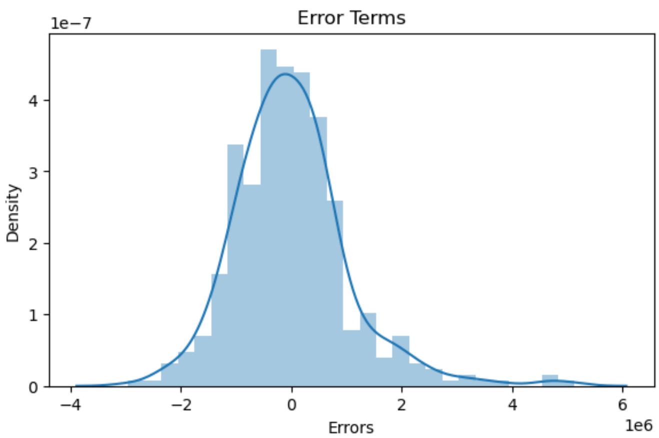
Objective :

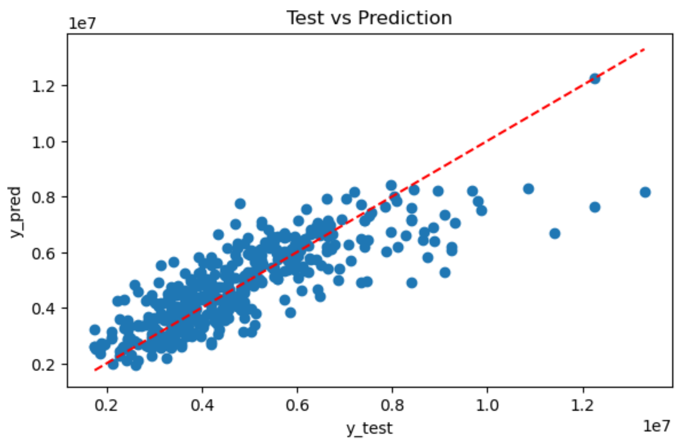
Let us now try building multiple regression models & compare their evaluation metrics to choose the best fit model both training and testing sets.

<u>Multiple Linear Regression (MLR)</u>

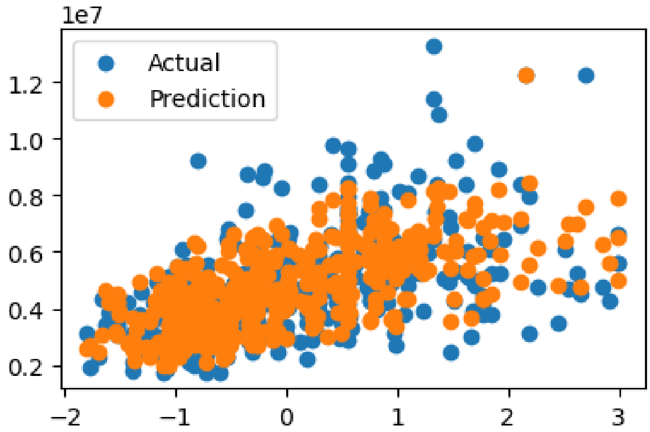



```
-------------------Training Set Metrics-------------------

R2-Score on Training set ---> 0.6789097089550895
Residual Sum of Squares (RSS) on Training set  ---> 466429810296572.75
Mean Squared Error (MSE) on Training set       ---> 1094905657973.1757
Root Mean Squared Error (RMSE) on Training set ---> 1046377.3974877209

-------------------Testing Set Metrics-------------------

R2-Score on Testing set ---> 0.6866794976385521
Residual Sum of Squares (RSS) on Training set  ---> 116042808105904.78
Mean Squared Error (MSE) on Training set       ---> 1084512225288.8298
Root Mean Squared Error (RMSE) on Training set ---> 1041399.1671250892
```

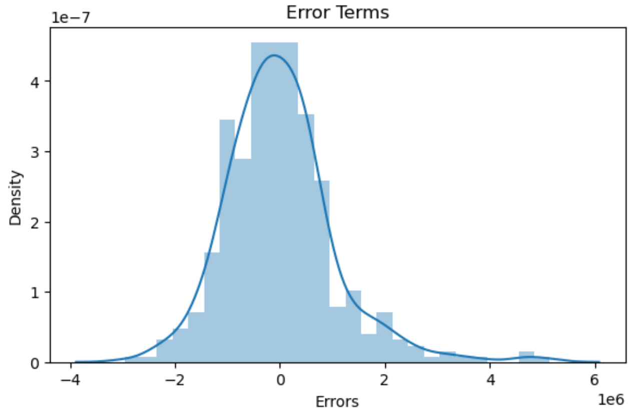


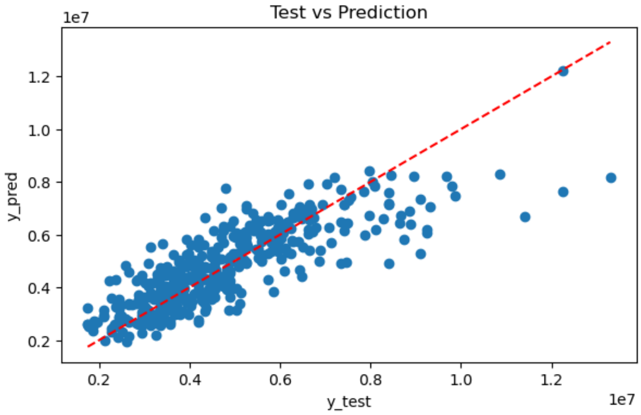


Ridge Rgeression Model

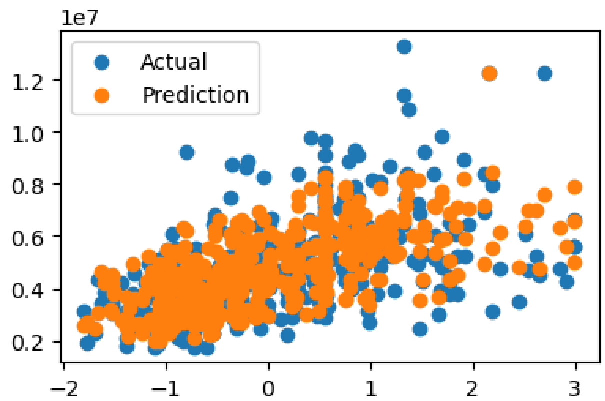



```
-------------------Training Set Metrics-------------------

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
```

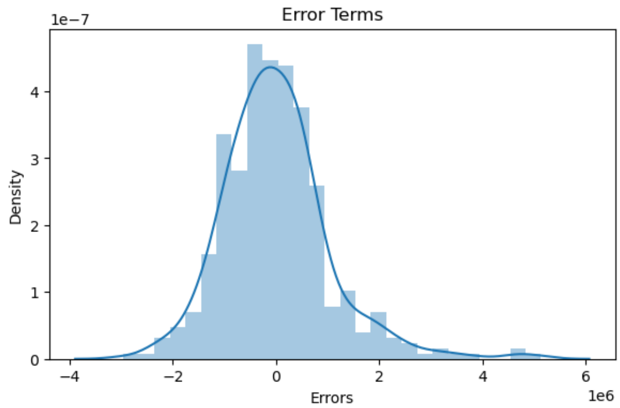


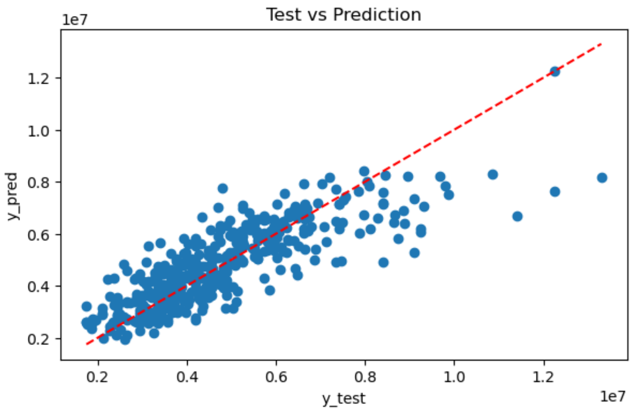


Lasso Regression Model

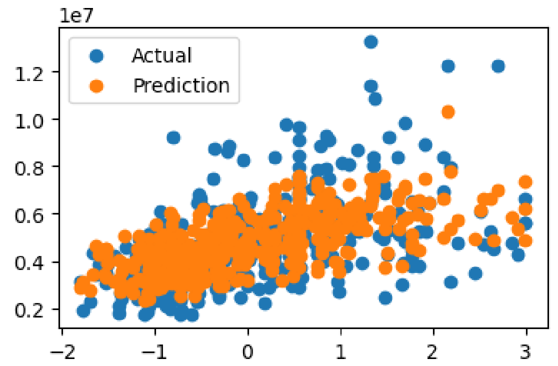



```
-------------------Training Set Metrics-------------------

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
```

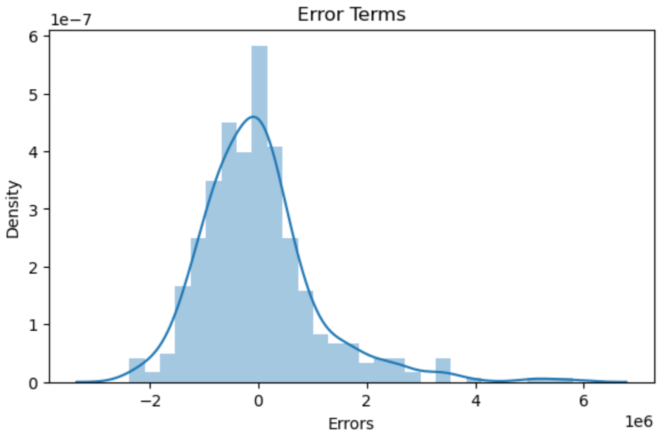


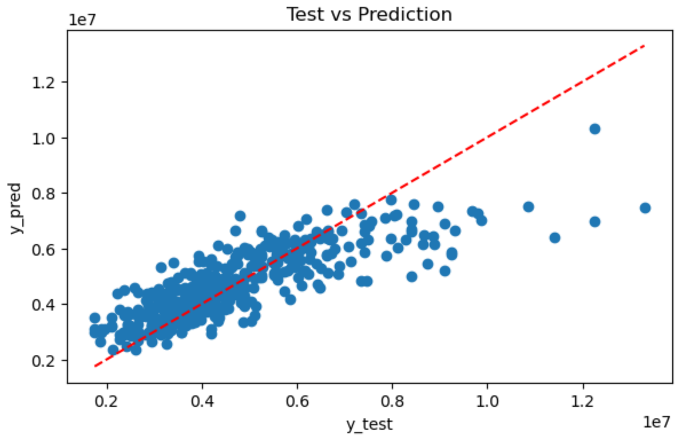


Elastic-Net Regression

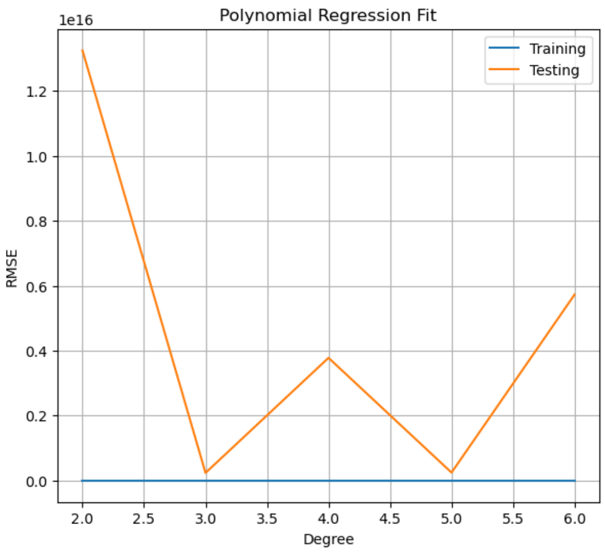


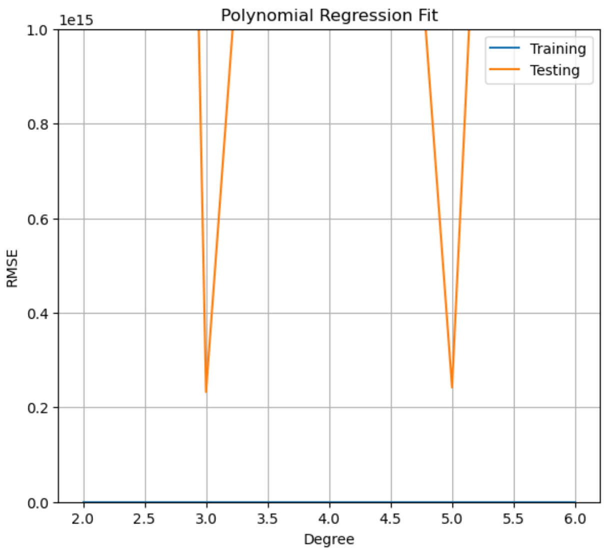

```
-------------------Training Set Metrics-------------------

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
```



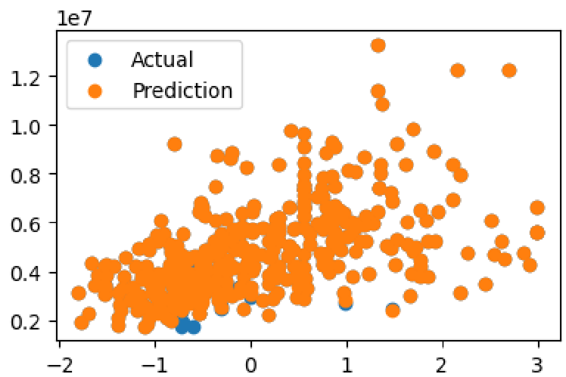


Polynomial Regression Model

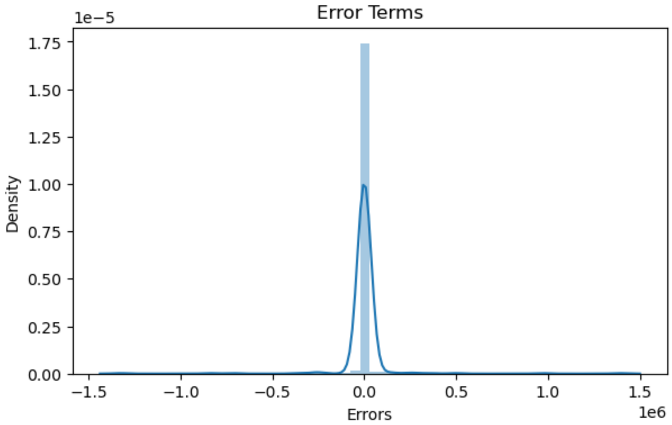


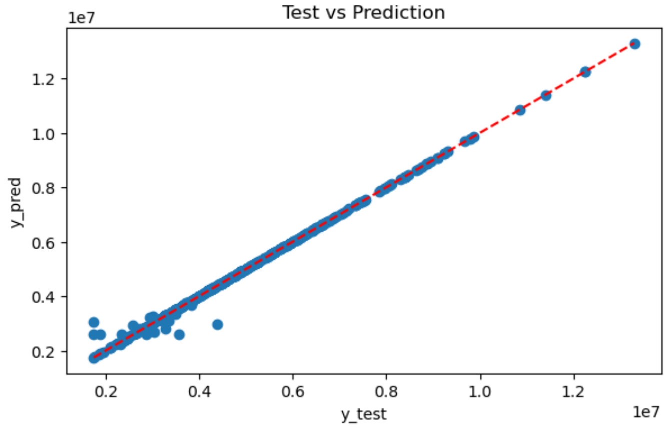


We can choose 5th order polynomial regression .



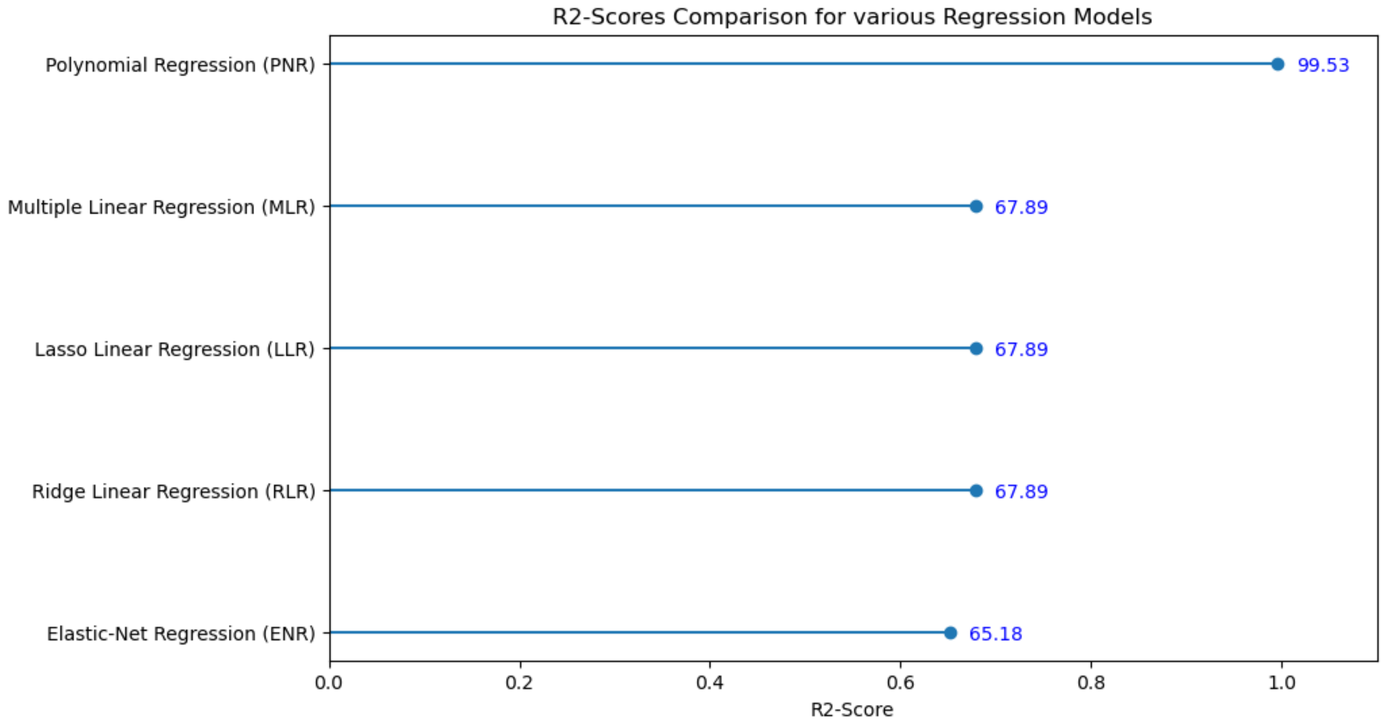

```
-------------------Training Set Metrics-------------------

R2-Score on Training set ---> 0.9953436904727914
Residual Sum of Squares (RSS) on Training set  ---> 6763959017229.223
Mean Squared Error (MSE) on Training set       ---> 15877838068.6132
Root Mean Squared Error (RMSE) on Training set ---> 126007.29371196414

-------------------Testing Set Metrics-------------------

R2-Score on Testing set ---> -1.6948917170350242e+16
Residual Sum of Squares (RSS) on Training set  ---> 6.277278148025306e+30
Mean Squared Error (MSE) on Training set       ---> 5.866615091612436e+28
Root Mean Squared Error (RMSE) on Training set ---> 242210963657973.88

-------------------Residual Plots-------------------
```



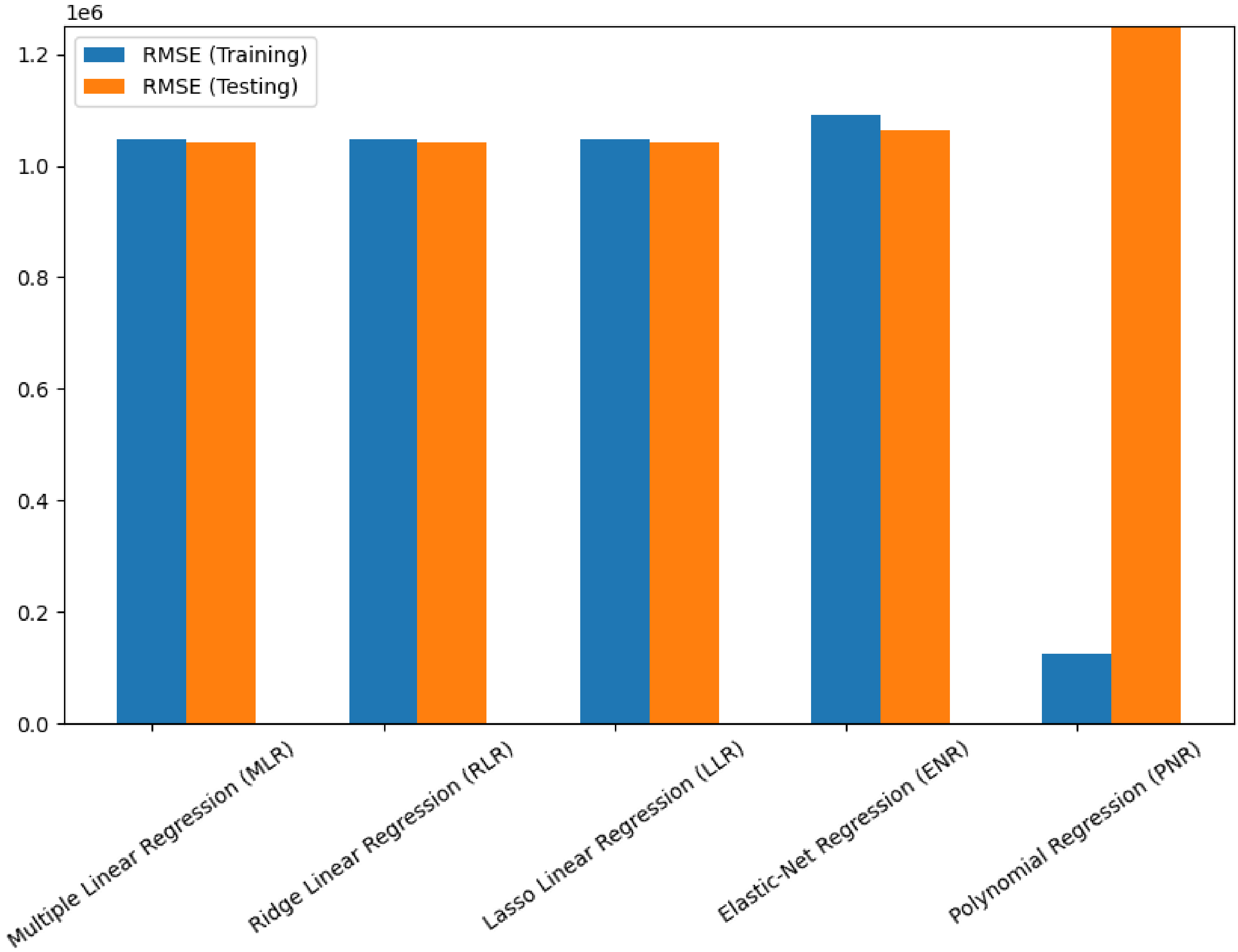


Comparing the Evaluation Metics of the Models

	Train-R2	Test-R2	Train-RSS	Test-RSS	Train-MSE	Test-MSE	Train-RMSE	Test-RMSE
Multiple Linear Regression (MLR)	0.678910	6.866795e-01	4.664298e+14	1.160428e+14	1.094906e+12	1.084512e+12	1.046377e+06	1.041399e+06
Ridge Linear Regression (RLR)	0.678906	6.868090e-01	4.664351e+14	1.159948e+14	1.094918e+12	1.084064e+12	1.046383e+06	1.041184e+06
Lasso Linear Regression (LLR)	0.678910	6.866805e-01	4.664298e+14	1.160425e+14	1.094906e+12	1.084509e+12	1.046377e+06	1.041398e+06
Elastic-Net Regression (ENR)	0.651848	6.739167e-01	5.057414e+14	1.207697e+14	1.187186e+12	1.128689e+12	1.089581e+06	1.062398e+06
Polynomial Regression (PNR)	0.995344	-1.694892e+16	6.763959e+12	6.277278e+30	1.587784e+10	5.866615e+28	1.260073e+05	2.422110e+14



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	5^{Th} Sem / 3^{rd} Year
Shalini Kumari	Information Technology	5^{Th} Sem / 3^{rd} Year
Awnish Kumar	Information Technology	5^{Th} Sem / 3^{rd} Year

Thank You Sir