Housing Case Study

Problem Statement:

Consider a real estate company that has a dataset containing the prices of properties in the Delhi region. It wishes to use the data to optimise the sale prices of the properties based on important factors such as area, bedrooms, parking, etc.

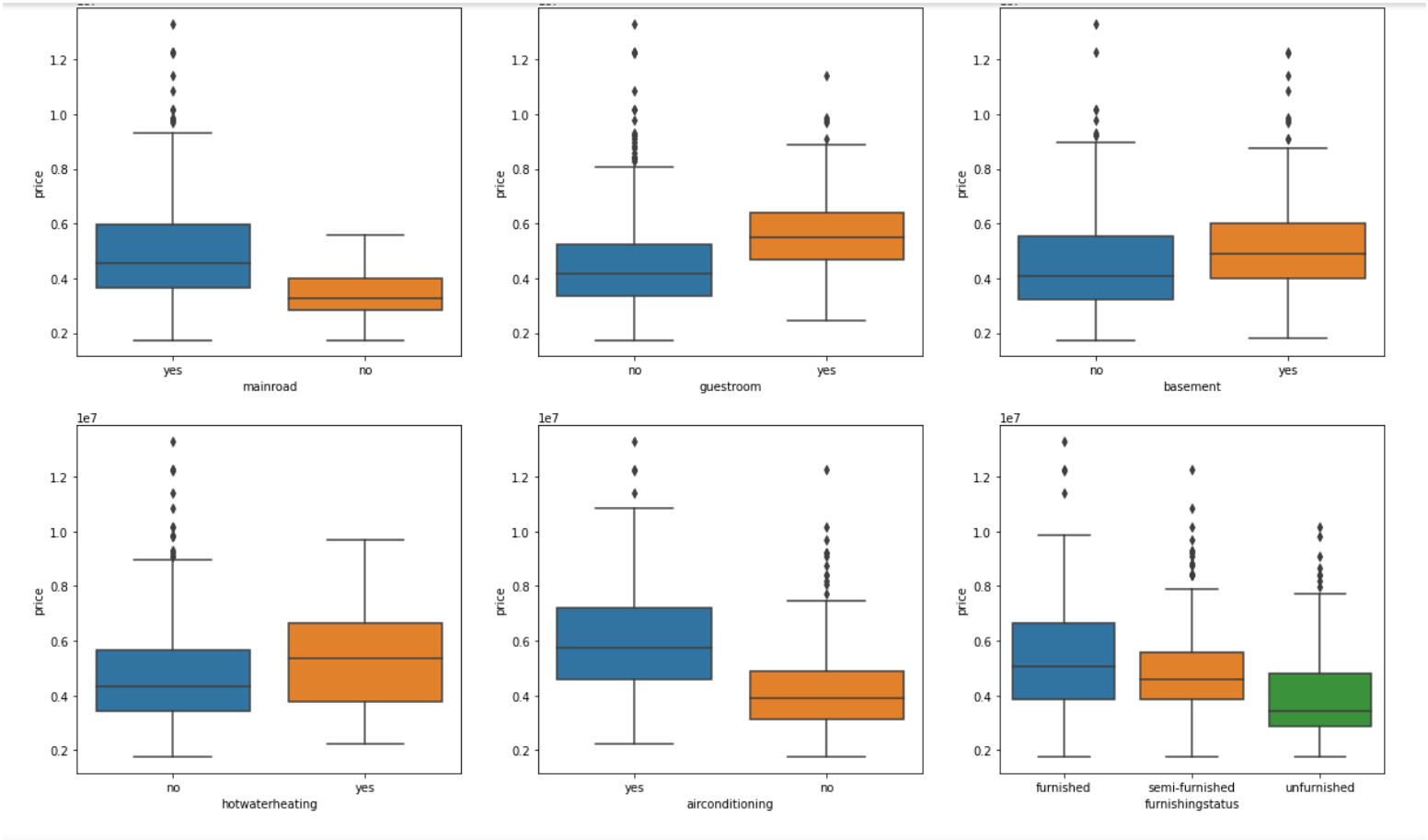
Essentially, the company wants —

1. To identify the variables affecting house prices, e.g. area, number of rooms, bathrooms, etc.
2. To create a linear model that quantitatively relates house prices with variables such as number of rooms, area, number of bathrooms, etc.
3. To know the accuracy of the model, i.e. how well these variables can predict house prices.

Steps:

1. Reading , understanding and visualizing the data
2. Preparing the data for modelling (train-test split, rescaling etc)
3. Training the data
4. Residual analysis
5. Predictions and evaluation on the test set

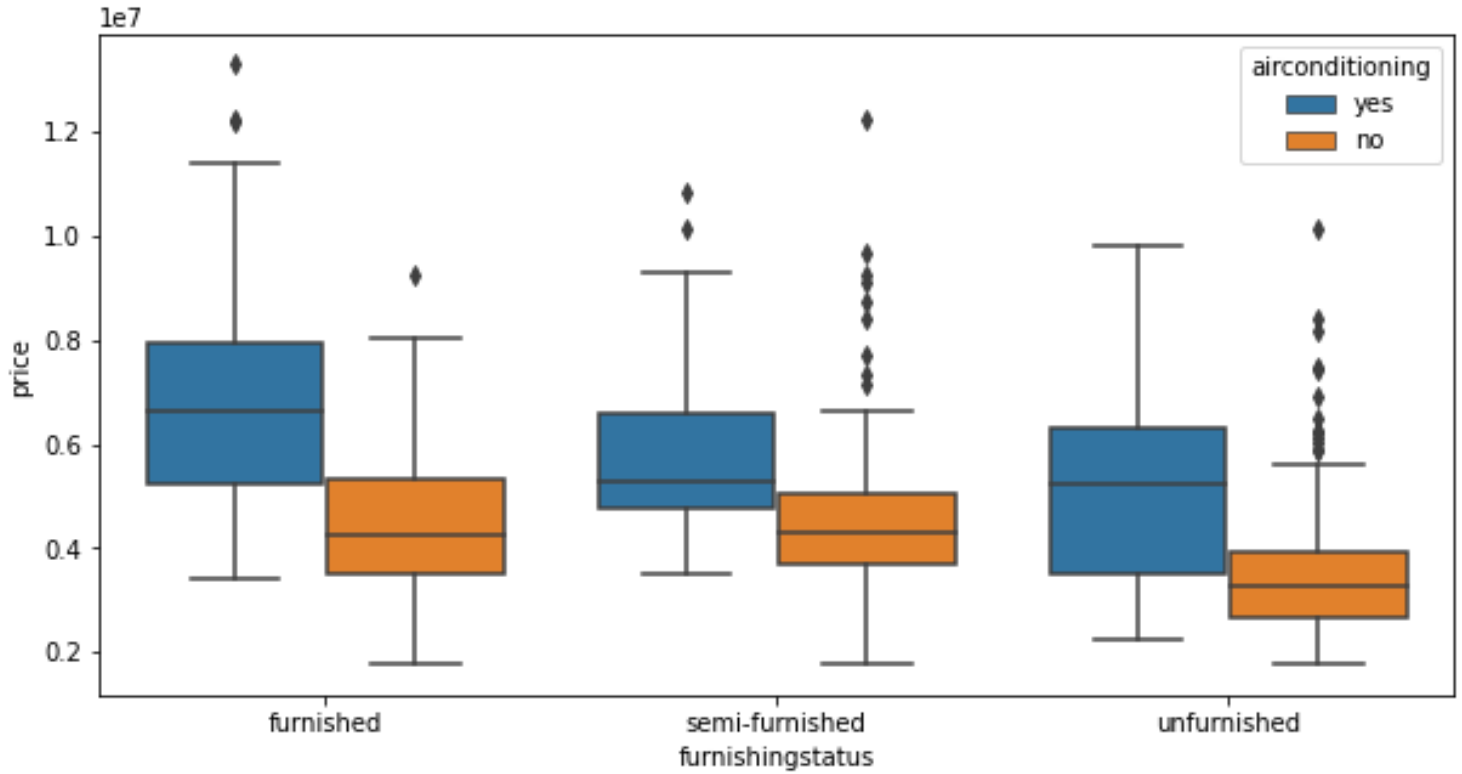
Visualising Categorical Variables

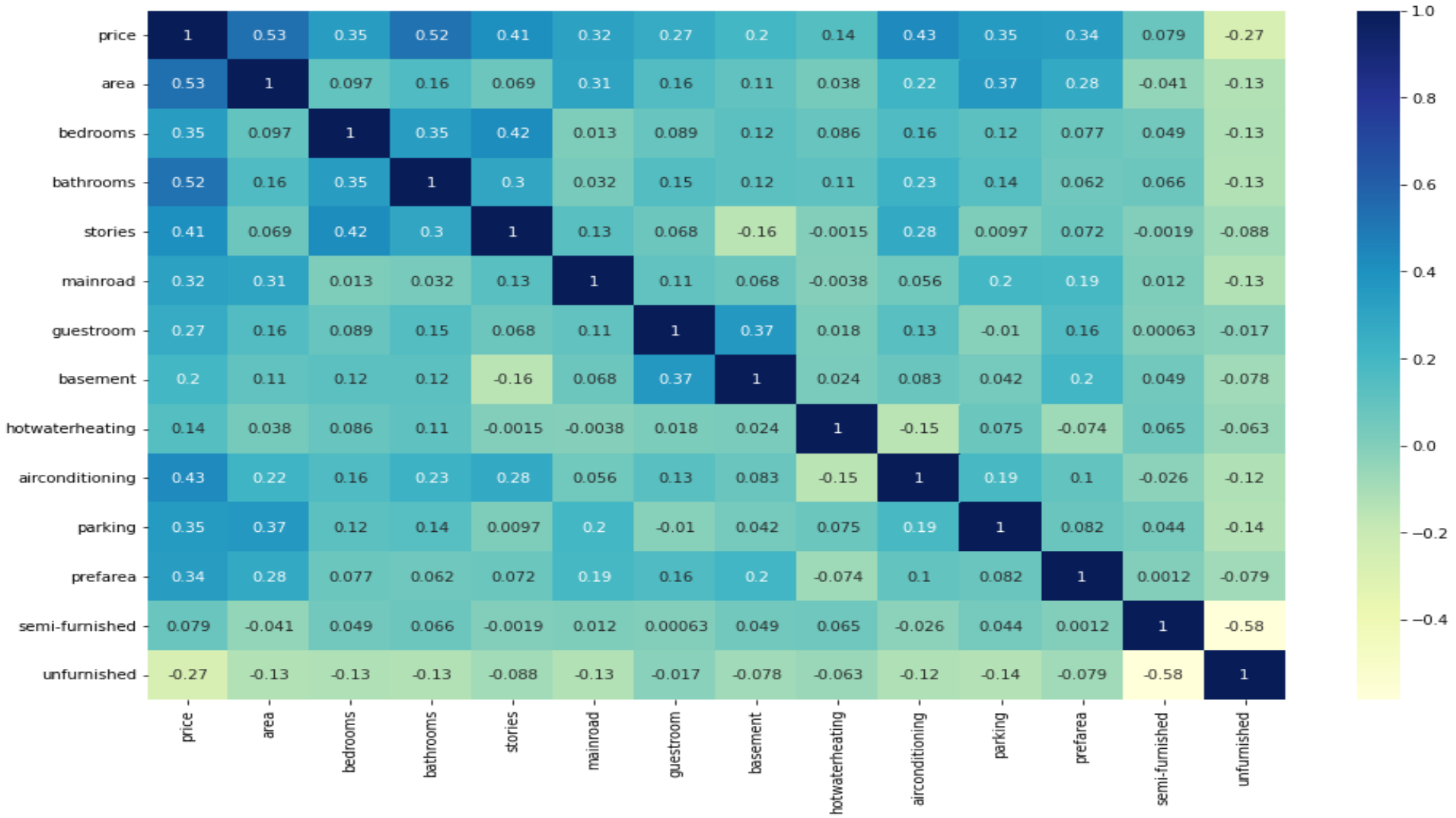


1. The location of the house on a main road is an important variable that affects the price, with houses on **main roads generally having higher prices than those that are not.**

2. There is not significant difference between furnished and unfurnished houses, the prompt does not provide enough information to make a clear summary. Additional information is needed to understand how furnishing affects the price of a house, such as the location of the house, the quality and quantity of furnishings, and local market conditions.

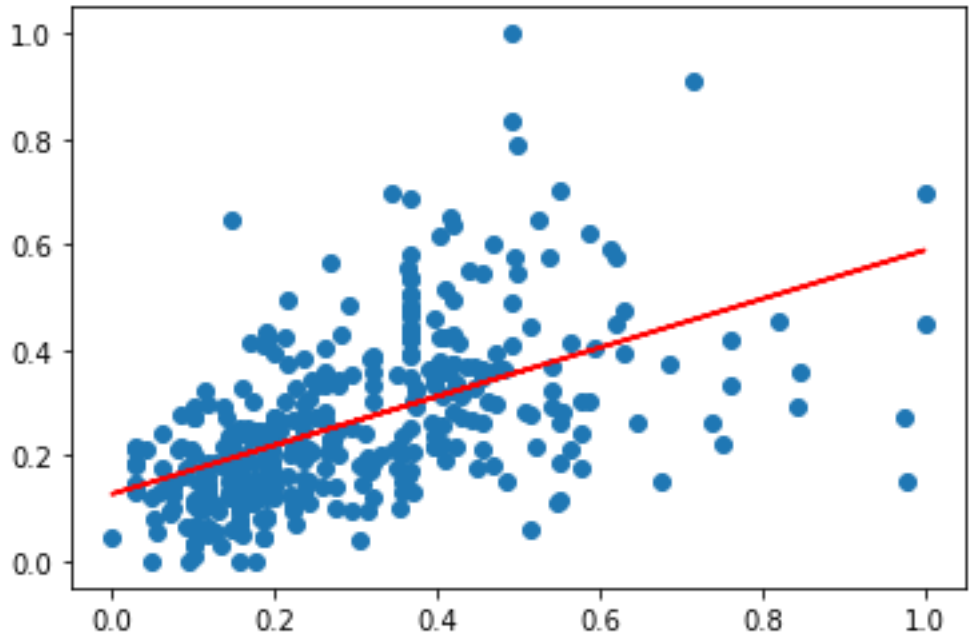
Conclusion: Air conditioning has impact on the pricing for every type of furnishing status





Conclusion: There is a significant correlation of **Price** on **Area, Stories, Bathroom and Air Conditioning**. Meaning the change in these variables will have affect on pricing

Visualize the data with a scatter plot and the fitted regression lin
e



Parameters of linear Regression

Const= 0.126894
Area= 0.462192

Summary

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.283
Model:	OLS	**Adj. R-squared:**	0.281
Method:	Least Squares	**F-statistic:**	149.6
Date:	Mon, 02 Jan 2023	**Prob (F-statistic):**	3.15e-29
Time:	18:19:27	**Log-Likelihood:**	227.23
No. Observations:	381	**AIC:**	-450.5
Df Residuals:	379	**BIC:**	-442.6
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	0.1269	0.013	9.853	0.000	0.102	0.152
area	0.4622	0.038	12.232	0.000	0.388	0.536

Omnibus:	67.313	**Durbin-Watson:**	2.018
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	143.063
Skew:	0.925	**Prob(JB):**	8.59e-32
Kurtosis:	5.365	**Cond. No.**	5.99

The coefficient pvalue is 0 the result is statistically significance R square is 0.28 means about 28% of significance in price is explained by the variable area

Final Model- Parameters

OLS Regression Results

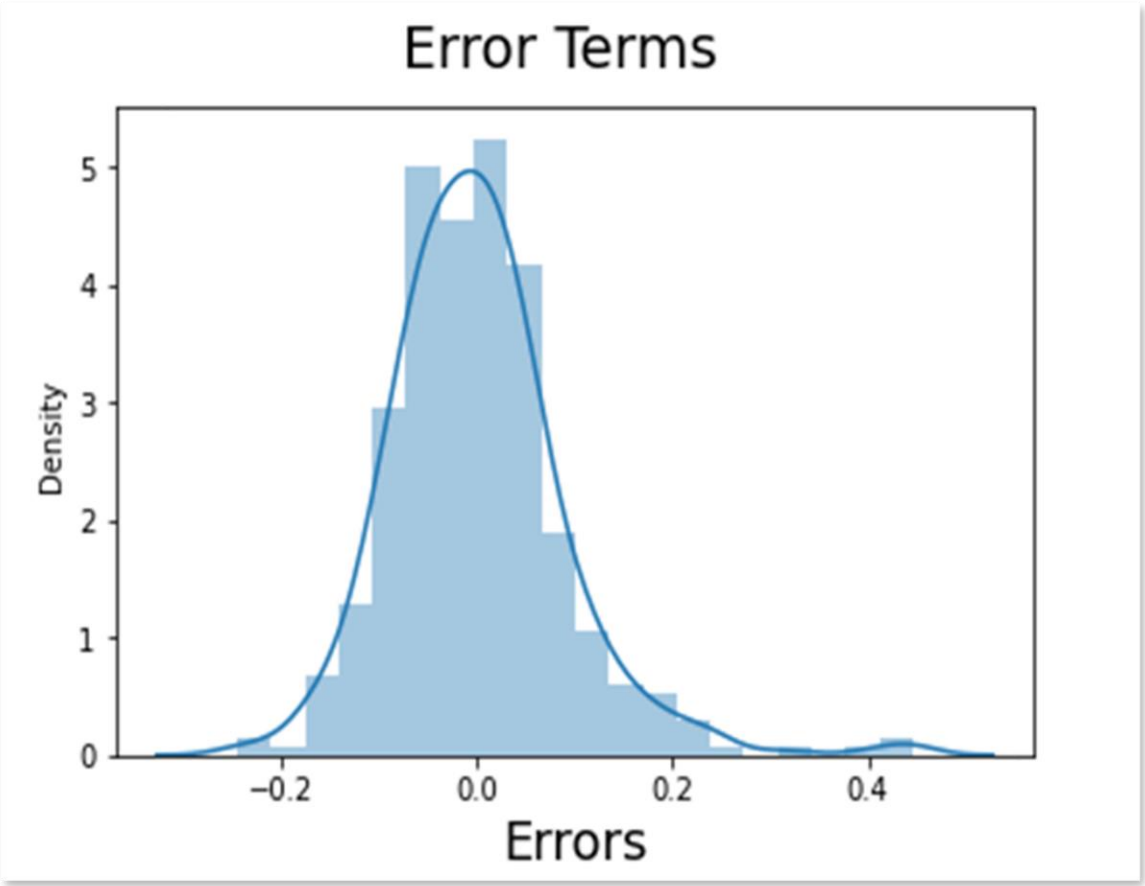
Dep. Variable:	price	R-squared:	0.676
Model:	OLS	Adj. R-squared:	0.667
Method:	Least Squares	F-statistic:	77.18
Date:	Mon, 02 Jan 2023	Prob (F-statistic):	3.13e-84
Time:	18:19:28	Log-Likelihood:	378.51
No. Observations:	381	AIC:	-735.0
Df Residuals:	370	BIC:	-691.7
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0428	0.014	2.958	0.003	0.014	0.071
area	0.2335	0.030	7.772	0.000	0.174	0.293
bathrooms	0.2019	0.021	9.397	0.000	0.160	0.244
stories	0.1081	0.017	6.277	0.000	0.074	0.142
mainroad	0.0497	0.014	3.468	0.001	0.022	0.078
guestroom	0.0402	0.013	3.124	0.002	0.015	0.065
hotwaterheating	0.0876	0.022	4.051	0.000	0.045	0.130
airconditioning	0.0682	0.011	6.028	0.000	0.046	0.090
parking	0.0629	0.018	3.482	0.001	0.027	0.098
prefarea	0.0637	0.012	5.452	0.000	0.041	0.087
unfurnished	-0.0337	0.010	-3.295	0.001	-0.054	-0.014

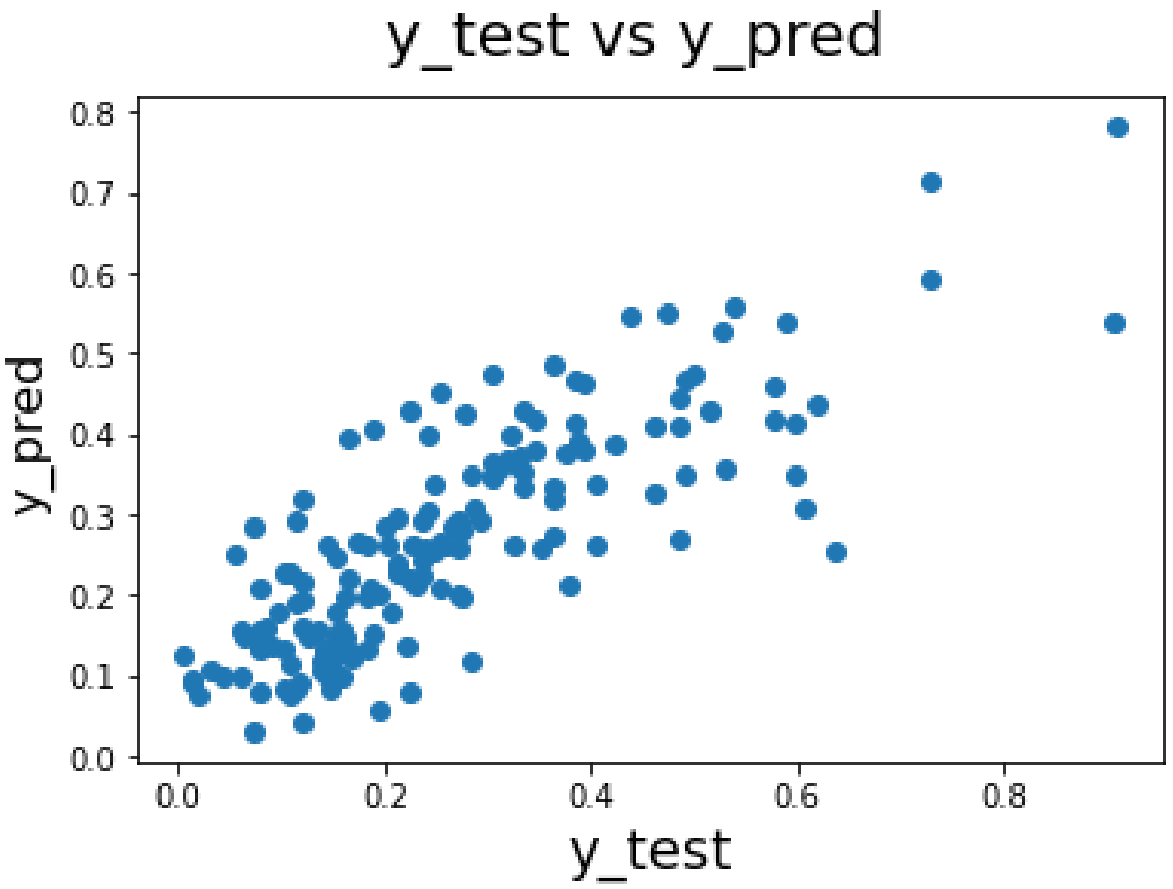
Omnibus:	97.054	Durbin-Watson:	2.099
Prob(Omnibus):	0.000	Jarque-Bera (JB):	322.034
Skew:	1.124	Prob(JB):	1.18e-70
Kurtosis:	6.902	Cond. No.	10.3

Error term is centered around zero and normally distributed.

The VIFs and p-values both are within an acceptable range. So we go ahead and make our predictions using this model only.



	Features	VIF
3	mainroad	4.55
0	area	4.54
2	stories	2.12
7	parking	2.10
6	airconditioning	1.75
1	bathrooms	1.58
8	prefarea	1.47
9	unfurnished	1.33
4	guestroom	1.30
5	hotwaterheating	1.12



We can see that the equation of our best fitted line is:

price=0.236×**area**+0.202×**bathrooms**+0.11×**stories**+0.05×**mainroad**+0.04×**guestroom**+0.0876×**hotwaterheating**+0.0682×**airconditioning**+0.0629×**parking**+0.0637×**prefarea**−0.0337×**unfurnished**