Predictive Modeling
REPORT

Linear Regression

Problem Statement:

Predicting the price of a house based on several independent and dependent features from the dataset.

Data Shape:

- ***The number of rows (observations) is 545***
- ***The number of columns (variables) is 13***

Data information:

RangeIndex: 545 entries, 0 to 544

Data columns (total 13 columns):

#	Column	Non-Null Count	Dtype
0	price	545 non-null	int64
1	area	545 non-null	int64
2	bedrooms	545 non-null	int64
3	bathrooms	545 non-null	int64
4	stories	545 non-null	int64
5	mainroad	545 non-null	object
6	guestroom	545 non-null	object
7	basement	545 non-null	object
8	hotwaterheating	545 non-null	object
9	airconditioning	545 non-null	object
10	parking	545 non-null	int64
11	prefarea	545 non-null	object
12	furnishingstatus	545 non-null	object

dtypes: int64(6), object(7)

memory usage: 55.5+ KB

Describe the data: Numeric data:

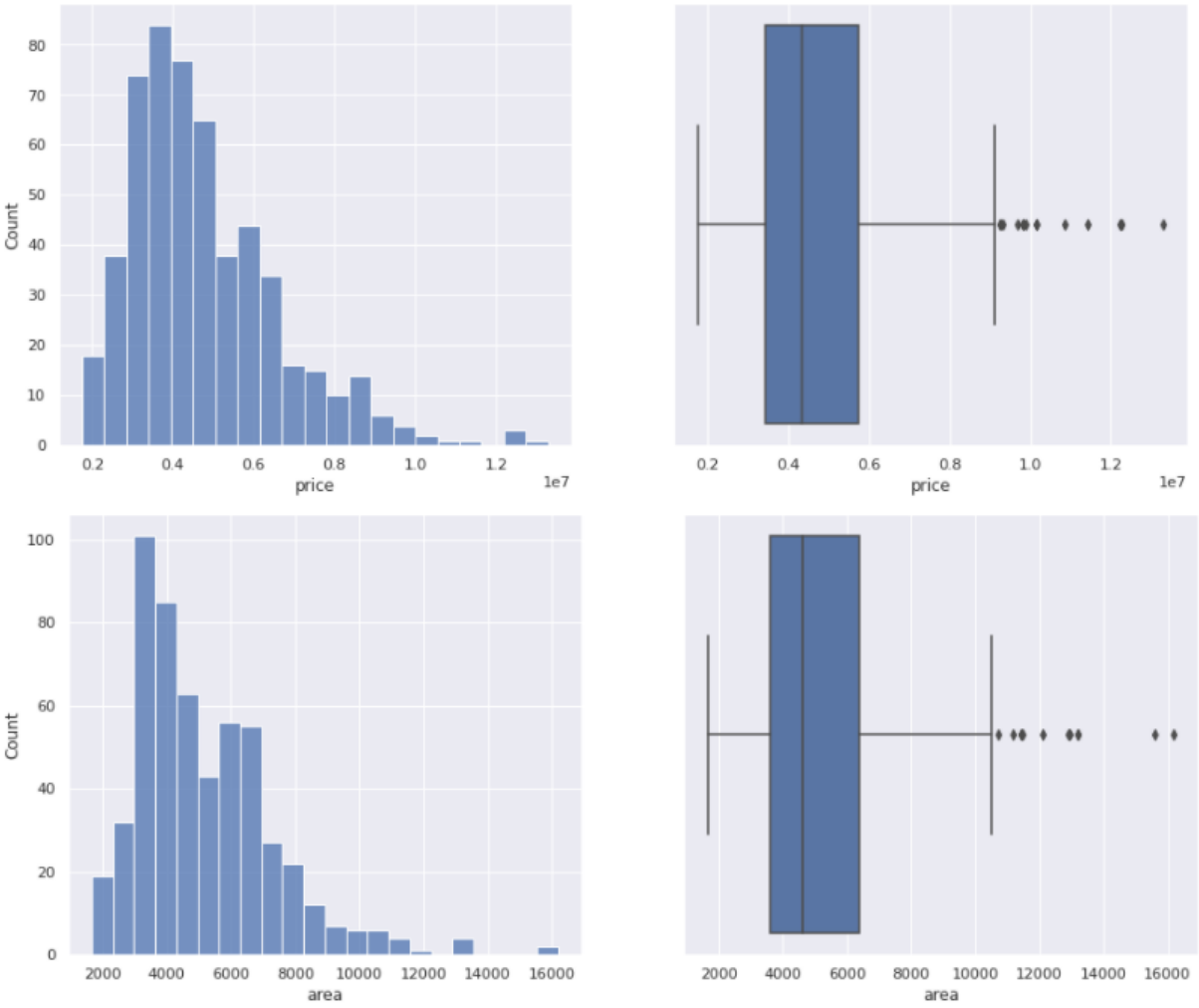
	Count	Mean	STD	MIN	25.00%	50.00%	75.00%	MAX
price	545	4766729.25	1870439.62	1750000	3430000	4340000	5740000	13300000
area	545	5150.54	2170.14	1650	3600	4600	6360	16200
bedrooms	545	2.97	0.74	1	2	3	3	6
bathrooms	545	1.29	0.5	1	1	1	2	4
stories	545	1.81	0.87	1	1	2	2	4
parking	545	0.69	0.86	0	0	0	1	3

Describe the data: Categorical data:

	Count	Unique	Top	Freq
mainroad	545	2	yes	468
guestroom	545	2	no	448
basement	545	2	no	354
hotwaterheating	545	2	no	520
airconditioning	545	2	no	373
prefarea	545	2	no	417
furnishingstatus	545	3	semi-furnished	227

- ***We have 6 numerical values, 7 categorical values and in which 6 are binary variable.***
- ***There are 0 null variables in our dataset.***
- ***Price is our Target variable.***
- ***Price ranges from 1750000/- to 13300000/-***
- ***Average Price is approximate 4766729.25/-***
- ***The average House Area is approximately 5150.54.***

Outliers: For House Price and Area:



- ***The Boxplot tells us there are a fair amount of outliers in both Area and Price distribution.***
- ***The distplot distribution can be said to be highly left-skewed in both cases.***
- ***We can see that there are outliers for each numeric variable.***
- ***As there is few outliers in each case, treating them would help the model predict the outcome correctly.***
- ***Value counts of `furnishingstatus` Variables:***

```
semi-furnished     227
unfurnished        178
furnished          140
Name: furnishingstatus, dtype: int64
```

- *Done one hard encoding for it. Values given:*
  - *semi-furnished- 2*
  - *Unfurnished- 1*
  - *furnished - 3*
- **Binary variables are Encoded using dummy variable encoding.**
- **Ordinal variables Encoded using dummy variable encoding:**

	count	mean	std	min	25.00%	50.00%	75.00%	max
price	545	5102.25	2005.8	1650	3600	4600	6360	10500
area	545	2.97	0.74	1	2	3	3	6
bedrooms	545	1.29	0.5	1	1	1	2	4
bathrooms	545	1.81	0.87	1	1	2	2	4
stories	545	0.69	0.86	0	0	0	1	3
parking	545	0.93	0.76	0	0	1	2	2
furnishingstatus	545	0.86	0.35	0	1	1	1	1
mainroad_yes	545	0.18	0.38	0	0	0	0	1
guestroom_yes	545	0.35	0.48	0	0	0	1	1
basement_yes	545	0.05	0.21	0	0	0	0	1
hotwaterheating_yes	545	0.32	0.47	0	0	0	1	1
airconditioning_yes	545	0.23	0.42	0	0	0	0	1
prefarea_yes	545	5102.25	2005.8	1650	3600	4600	6360	10500

- *All the variables are now encoded and ready for model implementation.*
- ***Splitting data into Train and Test at the default radio of 30:70% with random state as 123.***
- *After splitting the data into test and train, distribution is:*
  - `X_train (381, 12)`
  - `X_test (164, 12)`
  - `y_train (381,)`

- y_test (164,)

Model Iteration:1

- **Linear Regression Model score for R Square:**
  - The model score on traing set: 0.706
  - The model score on test set: 0.631
- 70% of the variation in the price is explained by the predictors in the model for train set.
- 63% of the variation in the price is explained by the predictors in the model for train set.
- **Coefficients with variables:**
  - The coefficient for area is 280.16828141892
  - The coefficient for bedrooms is 98528.01971678424
  - The coefficient for bathrooms is 905411.1470633473
  - The coefficient for stories is 390766.8163013215
  - The coefficient for parking is 150906.6983386266
  - The coefficient for furnishingstatus is 194137.67902612602
  - The coefficient for mainroad_yes is 369831.7825696926
  - The coefficient for guestroom_yes is 403971.37500637537
  - The coefficient for basement_yes is 284402.0494674524
  - The coefficient for hotwaterheating_yes is 471298.5442953159
  - The coefficient for airconditioning_yes is 887693.0839815252
  - The coefficient for prefarea_yes is 653178.4367847284
- **Linear Regression Model score for Adj R Square:**
  - **Combining X and Y to get one dataset for Adj R Square, implemented using statsmodel.**
- **Intercept and coefficients associated with variables:**
  - Intercept -122008.937969
  - Variables:

area	280.168281
bedrooms	98528.019717
bathrooms	905411.147063
stories	390766.816301
parking	150906.698339

furnishingstatus	*194137.679026*
mainroad_yes	*369831.782570*
guestroom_yes	*403971.375006*
basement_yes	*284402.049467*
hotwaterheating_yes	*471298.544295*
airconditioning_yes	*887693.083982*
prefarea_yes	*653178.4367*

Regression model summary:

OLS Regression Results

Dep. Variable:	price	R-squared:	0.706
Model:	OLS	Adj. R-squared:	0.697
Method:	Least Squares	F-statistic:	73.74
Date:	Sun, 19 Jun 2022	Prob (F-statistic):	4.29e-90
Time:	20:59:18	Log-Likelihood:	-5781.0
No. Observations:	381	AIC:	1.159e+04
Df Residuals:	368	BIC:	1.164e+04
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.22e+05	2.51e+05	-0.487	0.627	-6.15e+05	3.71e+05
area	280.1683	29.464	9.509	0.000	222.229	338.107
bedrooms	9.853e+04	7.61e+04	1.295	0.196	-5.1e+04	2.48e+05
bathrooms	9.054e+05	1.12e+05	8.055	0.000	6.84e+05	1.13e+06
stories	3.908e+05	6.76e+04	5.785	0.000	2.58e+05	5.24e+05
parking	1.509e+05	6.12e+04	2.464	0.014	3.05e+04	2.71e+05
furnishingstatus	1.941e+05	6.7e+04	2.896	0.004	6.23e+04	3.26e+05
mainroad_yes	3.698e+05	1.56e+05	2.374	0.018	6.35e+04	6.76e+05
guestroom_yes	4.04e+05	1.43e+05	2.828	0.005	1.23e+05	6.85e+05
basement_yes	2.844e+05	1.18e+05	2.406	0.017	5.2e+04	5.17e+05
hotwaterheating_yes	4.713e+05	2.57e+05	1.830	0.068	-3.51e+04	9.78e+05
airconditioning_yes	8.877e+05	1.16e+05	7.682	0.000	6.6e+05	1.11e+06
prefarea_yes	6.532e+05	1.25e+05	5.220	0.000	4.07e+05	8.99e+05

Omnibus:	19.086	Durbin-Watson:	2.079
Prob(Omnibus):	0.000	Jarque-Bera (JB):	23.215
Skew:	0.453	Prob(JB):	9.10e-06
Kurtosis:	3.801	Cond. No.	3.00e+04

- ***According to our result above, p-value is greater than alpha, hence null hypothesis is accepted, we will have to further tuning our formula to get better result.***
- ***For this model, we are removing the variables with high p values than alpha. Starting with bedrooms and hotwaterheating_yes.***

Iteration:2

- ***For this model, we are removing the variables with high p values. Starting with bedrooms and hotwaterheating_yes. Followed by: mainroad_yes and basement_yes.***

OLS Regression Results

Dep. Variable:	price	R-squared:	0.693
Model:	OLS	Adj. R-squared:	0.687
Method:	Least Squares	F-statistic:	105.2
Date:	Mon, 20 Jun 2022	Prob (F-statistic):	1.19e-90
Time:	15:09:29	Log-Likelihood:	-5789.2
No. Observations:	381	AIC:	1.160e+04
Df Residuals:	372	BIC:	1.163e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.875e+05	1.84e+05	2.100	0.036	2.47e+04	7.5e+05
area	289.6672	29.015	9.983	0.000	232.613	346.721
bathrooms	9.615e+05	1.09e+05	8.847	0.000	7.48e+05	1.18e+06
stories	4.021e+05	6.23e+04	6.457	0.000	2.8e+05	5.25e+05
parking	1.774e+05	6.17e+04	2.874	0.004	5.6e+04	2.99e+05
furnishingstatus	2.17e+05	6.77e+04	3.205	0.001	8.39e+04	3.5e+05
guestroom_yes	5.322e+05	1.37e+05	3.878	0.000	2.62e+05	8.02e+05
airconditioning_yes	8.767e+05	1.16e+05	7.550	0.000	6.48e+05	1.11e+06
prefarea_yes	7.443e+05	1.23e+05	6.059	0.000	5.03e+05	9.86e+05

Omnibus:	21.787	Durbin-Watson:	2.070
Prob(Omnibus):	0.000	Jarque-Bera (JB):	26.474
Skew:	0.503	Prob(JB):	1.78e-06
Kurtosis:	3.811	Cond. No.	2.12e+04

- ***The overall P-value is less than alpha, so rejecting H0 and accepting Ha that at least 1 regression coefficient is not 0. Here all regression coefficients are not 0.***
  - ***The model RMSE on training set: 961109.876***
  - ***The model RMSE on test set: 1143904.04***
- ***The final Linear Regression equation is:***

  price = (387492.41) * Intercept + (289.67) * area + (961515.25) * bathrooms + (402104.66) * stories + (177373.28) * parking + (216980.26) * furnishingstatus + (532247.15) * guestroom_yes + (876699.66) * airconditioning_yes + (744317.05) * prefarea_yes

Inference :

- ***With Adj R Square of 0.687.***
- ***When area increases by 1 unit, house price increases by 961515.25 units, keeping all other predictors constant.***
- ***Similarly, when no. of bathrooms increases by 1 unit, house price increases by 402104.66 units, keeping all other predictors constant and so on.***
- ***The feature that has the most effect on the housing price is Area.***

END !