BUSINESS REPORT

Terro's real estate agency

SBMITTED BY

R.ANU

Problem Statement

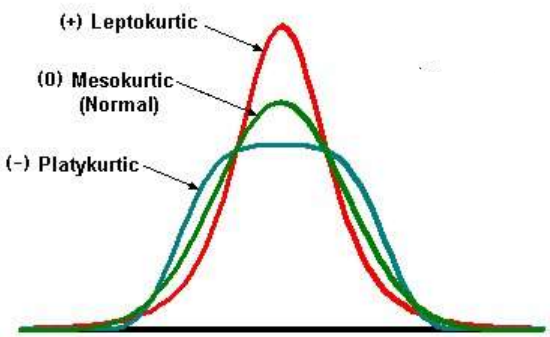
Terro's real-estate is an agency that estimates the pricing of houses in a certain locality. The pricing is concluded based on different features / factors of a property. This also helps them in identifying the business value of a property. To do this activity the company employs an "Auditor", who studies various geographic features of a property like pollution level (NOX), crime rate, education facilities (pupil to teacher ratio), connectivity (distance from highway), etc. This helps in determining the price of a property.

1) Generate the summary statistics for each variable in the table. (Use Data analysis tool pack). Write down your observation.

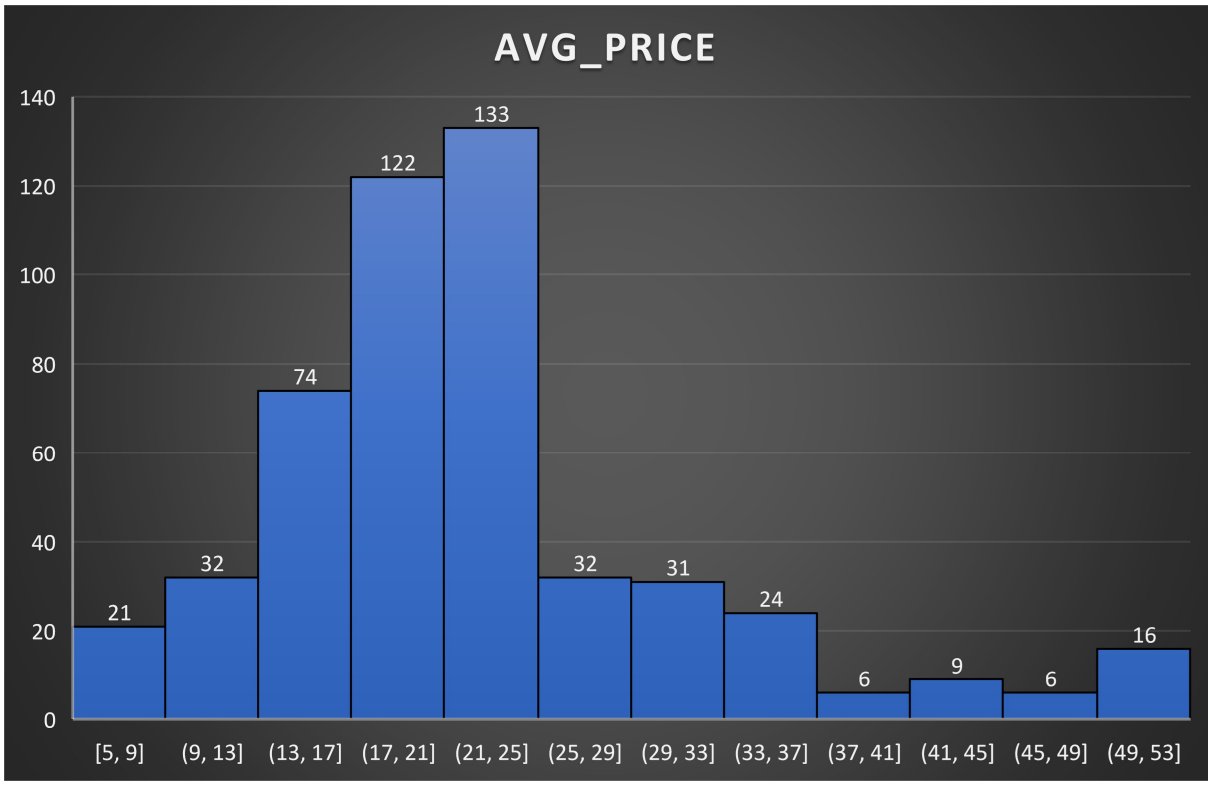
Tax	
Mean	408.2371542
Standard Error	7.492388692
Median	330
Mode	666
Standard Deviation	168.5371161
Sample Variance	28404.75949
Kurtosis	-1.142407992
Skewness	0.669955942
Range	524
Minimum	187
Maximum	711
Sum	206568
Count	506

I generate descriptive statistics for each variable in the given table. From above the table am showing the statistical table of tax.

Here the kurtosis is negative so it is platykurtic.



2) Plot a histogram of the Avg_Price variable. What do you infer?



Above histogram shows the average price of houses.

It represents the average price of houses in a particular area. The X- axis represent the number of houses and the Y-axis represent the average price of houses.

The above histogram might show the highest bar around (21,25] range, it means there are many houses in that price range. This suggests that houses in that range are popular in a particular area.

The above histogram might show the lowest bar around (37, 41] and (45, 49] same range, it means there are few houses in that price range. this suggests that houses highly paid in a particular area.

3) Compute the covariance matrix. Share your observations.

Covariance helps us understand the direction and strength of the relationship between two variables.

If the covariance is positive, it means that when one variable goes up, the other tends to go up as well. Conversely, if the covariance is negative, it means that when one variable goes up, the other tends to go down.

	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RATE	8.516147873									
AGE	0.562915215	790.7924728								
INDUS	-0.110215175	124.2678282	46.97142974							
NOX	0.000625308	2.381211931	0.605873943	0.013401099						
DISTANCE	-0.229860488	111.5499555	35.47971449	0.615710224	75.66653127					
TAX	-8.229322439	2397.941723	831.7133331	13.02050236	1333.116741	28348.6236				
PTRATIO	0.068168906	15.90542545	5.680854782	0.047303654	8.74340249	167.8208221	4.677726296			
AVG_ROOM	0.056117778	-4.74253803	-1.884225427	-0.024554826	-1.281277391	-34.51510104	-0.539694518	0.492695216		
LSTAT	-0.882680362	120.8384405	29.52181125	0.487979871	30.32539213	653.4206174	5.771300243	-3.073654967	50.89397935	
AVG_PRICE	1.16201224	-97.39615288	-30.46050499	-0.454512407	-30.50083035	-724.8204284	-10.09067561	4.484565552	-48.35179219	84.41955616

From the above table, the covariance is negative for Avg.Price and Distance so, the Avg.Price is inversely proportional to the Distance.

The covariance is positive for LSTAT, Indus, NOX, Distance, Tax, and Ptratio so, the LSTAT is Directly proportional to the Indus, NOX, Distance, Tax.

4) Create a correlation matrix of all the variables (Use Data analysis tool pack).

a) Which are the top 3 positively correlated pairs

Positive correlated pairs

- 0.731470104 between NOX and Age
- 0.763651447 between Indus and NOX
- 0.910228189 between Tax and Distance

Negative correlated pairs

- -0.613808272 between LSTAT and Avg.Room
- -0.507786686 between Avg.Price and Ptratio
- -0.737662726 between Avg.Price and LSTAT

5) Build an initial regression model with AVG_PRICE as 'y' (Dependent variable) and LSTAT variable as Independent Variable. Generate the residual plot.

a) What do you infer from the Regression Summary output in terms of variance explained, coefficient value, Intercept, and the Residual plot?

	Coefficients	Standard error	t stat	p- value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	34.0553	0.56262	61.4151	3.7431E-236	33.4484	35.6592	33.4484	35.6592
LSTAT	-0.9500	0.03873	24.5278	5.0811E-88	1.0261	0.8739	1.0261	0.8739

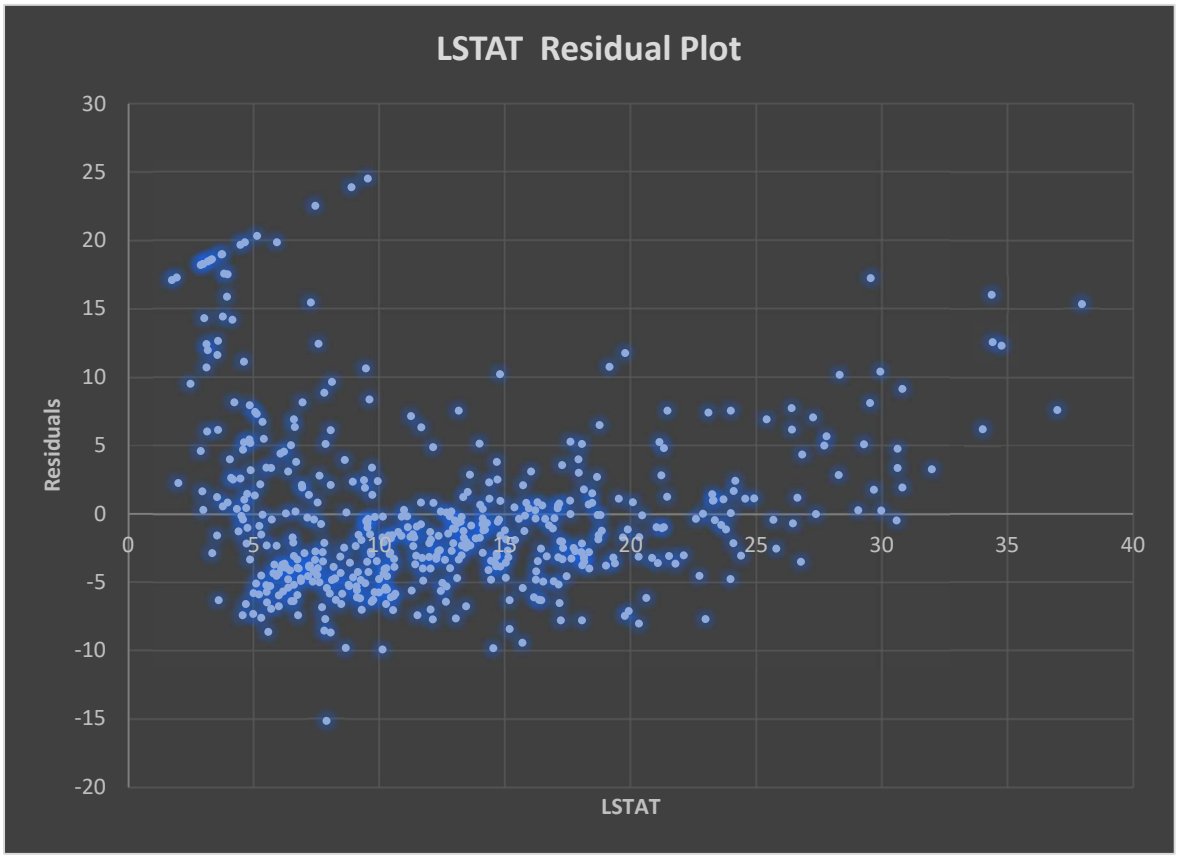
- **Coefficient value**

From the above table the LSTAT coefficient value is negative -0.9500 so, if LSTAT is increases the average price is decrease.

- **Intercept**

From the above table the intercept value is 34.0553.

- **Residual plot**



Observation:

From the above residual chart,

- Below 5 it has more residual error so it is lower biased.
- Greater than 25 it has more residual errors so it is upper biased.
- Between 5 to 25 the residuals errors are biased.

b) Is LSTAT variable significant for the analysis based on your model?

	Coefficients	Standard error	t stat	p- value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	34.0553	0.56262	61.4151	3.7431E-236	33.4484	35.6592	33.4484	35.6592
LSTAT	0.9500	0.03873	24.5278	5.0811E-88	1.0261	0.8739	1.0261	0.8739

The significance value of LSTAT is 5.0811E-88 is lesser than P value 0.05 hence we reject null hypothesis. So, the LSTAT is significant.

6) Build a new Regression model including LSTAT and AVG_ROOM together as independent variables and AVG_PRICE as dependent variable.

a) Write the Regression equation. If a new house in this locality has 7 rooms (on an average) and has a value of 20 for L-STAT, then what will be the value of AVG_PRICE? How does it compare to the company quoting a value of 30000 USD for this locality? Is the company Overcharging/ Undercharging?

Regression => $Y = m_1x_1 + m_2x_2 + b$

m_1 = slope = 5.094787984

m_2 = slope = -0.642358334

$x_1 = 7$

$x_2 = 20$

b = intercept = -1.358272812

Here my regression value is 21.4580764 like 21,000 which is lesser than company price so the company charge more price than what we expect.

b) Is the performance of this model better than the previous model you built in Question 5? Compare in terms of adjusted R-square and explain.

- **Previous model**

Regression Statistics	
Multiple R	0.737662726
R Square	0.544146298
Adjusted R Square	0.543241826
Standard Error	6.215760405
Observations	506

- **This model**

Regression Statistics	
Multiple R	0.799100498
R Square	0.638561606
Adjusted R Square	0.637124475
Standard Error	5.540257367
Observations	506

Adjusted R-square of Previous model -- 0.543241826

Adjusted R-square of This model -- 0.637124475

This model has higher value of adjusted R-square so this model is better than previous model.

7) Build another Regression model with all variables where AVG_PRICE alone be the Dependent Variable and all the other variables are independent.

a) Interpret the output in terms of adjusted R-square, coefficient and Intercept values.

Regression Statistics	
Multiple R	0.832978824
R Square	0.69385372
Adjusted R Square	0.688298647
Standard Error	5.1347635
Observations	506

Since the above regression table shows the Adjusted R Square is 0.688298647 (68.82%) it is above 50% so, the higher Adjusted R-Square is better.

- **Coefficient**

	Coefficients
Intercept	29.24131526
CRIME_RATE	0.048725141
AGE	0.032770689
INDUS	0.130551399
NOX	-10.3211828
DISTANCE	0.261093575
TAX	-0.01440119
PTRATIO	-1.074305348
AVG_ROOM	4.125409152
LSTAT	-0.603486589

From the above table,

The coefficient variables of Crime rate, Age, Indus, Distance, and Avg-Room are positive. So, the variables are Directly proportional.

The Other coefficient variables of NOX, Tax, Ptratio, and LSTAT are negative. So, the variables are Inversely proportional.

- **Intercept**

The intercept value 29.24131526 is positive.

b) Explain the significance of each independent variable with respect to AVG_PRICE.

	Coefficients	**P-value**
CRIME_RATE	0.048725141	0.534657
AGE	0.032770689	0.01267
INDUS	0.130551399	0.039121
NOX	-10.3211828	0.008294
DISTANCE	0.261093575	0.000138
TAX	-0.01440119	0.000251
PTRATIO	-1.074305348	6.59E-15
AVG_ROOM	4.125409152	3.89E-19
LSTAT	-0.603486589	8.91E-27

- **CRIME_RATE**

The significant value of Crime Rate is 0.534657 which is greater than P value 0.05 the level of significance. Hence, we accept the null hypothesis and we conclude that there is no significant.

- **AGE**

The significant value of Age is 0.01267 which is lesser than P value 0.05 the level of significance. Hence, we do not accept the null hypothesis and we conclude that there is significant.

- **INDUS**

The significant value of Indus is 0.039121 which is lesser than P value 0.05 the level of significance. Hence, we do not accept the null hypothesis and we conclude that there is significant.

- **NOX**

The significant value of NOX is 0.008294 which is lesser than P value 0.05 the level of significance. Hence, we do not accept the null hypothesis and we conclude that there is significant.

- **DISTANCE**

The significant value of Distance is 0.000138 which is lesser than P value 0.05 the level of significance. Hence, we do not accept the null hypothesis and we conclude that there is significant.

- **TAX**

The significant value of Tax is 0.000251 which is lesser than P value 0.05 the level of significance. Hence, we do not accept the null hypothesis and we conclude that there is significant.

- **PTRATIO**

The significant value of Tax is 6.59E-15 which is lesser than P value 0.05 the level of significance. Hence, we do not accept the null hypothesis and we conclude that there is significant.

- **AVG_ROOM**

The significant value of Avg-Room is 3.89E-19 which is lesser than P value 0.05 the level of significance. Hence, we do not accept the null hypothesis and we conclude that there is significant.

- **LSTAT**

The significant value of LSTAT is 8.91E-27 which is lesser than P value 0.05 the level of significance. Hence, we do not accept the null hypothesis and we conclude that there is significant.

8) Pick out only the significant variables from the previous question. Make another instance of the Regression model using only the significant variables you just picked and answer the questions below.

a) Interpret the output of this model.

Regression Statistics	
Multiple R	0.832835773
R Square	0.693615426
Adjusted R Square	0.688683682
Standard Error	5.131591113
Observations	506

Since the above regression table shows the Adjusted R Square is 0.688683682 (68.86%) it is above 50% so, the higher Adjusted R-Square is better.

b) Compare the adjusted R-square value of this model with the model in the previous question, which model performs better according to the value of adjusted R-square?

- **Previous model**

Regression Statistics	
Multiple R	0.832978824
R Square	0.69385372
Adjusted R Square	0.688298647
Standard Error	5.1347635
Observations	506

- **This model**

Regression Statistics	
Multiple R	0.832835773
R Square	0.693615426
Adjusted R Square	0.688683682
Standard Error	5.131591113
Observations	506

Adjusted R-square of Previous model -- 0.688298647

Adjusted R-square of This model -- 0.688683682

This model has higher value of adjusted R-square so this model is better than previous model.

c) Sort the values of the Coefficients in ascending order. What will happen to the average price if the value of NOX is more in a locality in this town?

	Coefficients
NOX	-10.27270508
PTRATIO	-1.071702473
LSTAT	-0.605159282
TAX	-0.014452345
AGE	0.03293496
INDUS	0.130710007
DISTANCE	0.261506423
AVG_ROOM	4.125468959
Intercept	29.42847349

The coefficient value of NOX -10.27270508 is negative. So, the variable of NOX is Inversely proportional. It means the NOX decreases the average price is Increases.

d) Write the regression equation from this model.

Regression => $Y = m_1x_1 + m_2x_2 + \ldots + m_8x_8 + b$

Where,

b = intercept,

$m_1, m_2, m_3, \ldots, m_8$ = slopes

Using the coefficients values in the table, the regression equation for this model is:

y = 29.42847349 + 0.03293496 * AGE + 0.130710007 * INDUS - 10.27270508 * NOX + 0.261506423 * DISTANCE - 0.014452345 * TAX - 1.071702473 * PTRATIO + 4.125468959 * AVG_ROOM - 0.605159282 * LSTAT

Here, the NOX, INDUS, AGE, LSTAT, PTRATIO, DISTANCE, AVG-ROOM, TAX are predicting variables.