TERRO'S REAL ESTATE AGENCY

D. Harish Reddy

DA-MAR-2023-GLCA

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.

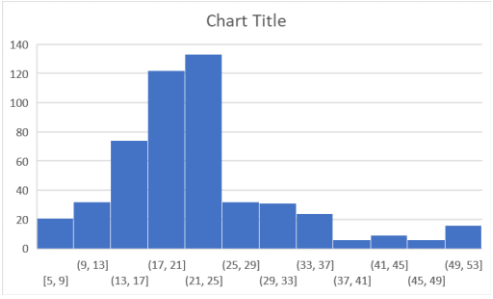
1.Generate the summary statistics for each variable in the table. (Use Data analysis tool pack). Write down your observation

CRIME_RATE		*AGE*		*INDUS*		*NOX*		*DISTANCE*		*TAX*		*PTRATIO*		*AVG_ROOM*		*LSTAT*		*AVG_PRICE*	
Mean	4.871976	Mean	68.5749	Mean	11.13678	Mean	0.554695	Mean	9.549407	Mean	408.2372	Mean	18.45553	Mean	6.284634	Mean	12.65306	Mean	22.53281
Standard Error	0.12986	Standard E	1.25137	Standard E	0.30498	Standard E	0.005151	Standard E	0.387085	Standard E	7.492389	Standard E	0.096244	Standard E	0.031235	Standard E	0.317459	Standard E	0.408861
Median	4.82	Median	77.5	Median	9.69	Median	0.538	Median	5	Median	330	Median	19.05	Median	6.2085	Median	11.36	Median	21.2
Mode	3.43	Mode	100	Mode	18.1	Mode	0.538	Mode	24	Mode	666	Mode	20.2	Mode	5.713	Mode	8.05	Mode	50
Standard Devia	2.921132	Standard D	28.14886	Standard D	6.860353	Standard D	0.115878	Standard D	8.707259	Standard D	168.5371	Standard D	2.164946	Standard D	0.702617	Standard D	7.141062	Standard D	9.197104
Sample Varian	8.533012	Sample Va	792.3584	Sample Va	47.06444	Sample Va	0.013428	Sample Va	75.81637	Sample Va	28404.76	Sample Va	4.686989	Sample Va	0.493671	Sample Va	50.99476	Sample Va	84.58672
Kurtosis	-1.18912	Kurtosis	-0.96772	Kurtosis	-1.23354	Kurtosis	-0.06467	Kurtosis	-0.86723	Kurtosis	-1.14241	Kurtosis	-0.28509	Kurtosis	1.8915	Kurtosis	0.49324	Kurtosis	1.495197
Skewness	0.021728	Skewness	-0.59896	Skewness	0.295022	Skewness	0.729308	Skewness	1.004815	Skewness	0.669956	Skewness	-0.80232	Skewness	0.403612	Skewness	0.90646	Skewness	1.108098
Range	9.95	Range	97.1	Range	27.28	Range	0.486	Range	23	Range	524	Range	9.4	Range	5.219	Range	36.24	Range	45
Minimum	0.04	Minimum	2.9	Minimum	0.46	Minimum	0.385	Minimum	1	Minimum	187	Minimum	12.6	Minimum	3.561	Minimum	1.73	Minimum	5
Maximum	9.99	Maximum	100	Maximum	27.74	Maximum	0.871	Maximum	24	Maximum	711	Maximum	22	Maximum	8.78	Maximum	37.97	Maximum	50
Sum	2465.22	Sum	34698.9	Sum	5635.21	Sum	280.6757	Sum	4832	Sum	206568	Sum	9338.5	Sum	3180.025	Sum	6402.45	Sum	11401.6
Count	506	Count	506	Count	506	Count	506	Count	506	Count	506	Count	506	Count	506	Count	506	Count	506

- By observing above summary statistics, the average price of a flat is about 22.53(amount). The positive reason to purchase a flat is Pupil and Teacher ratio which is in good range (9.40 range of PTRATIO) that can attract a greater number of people to buy flats in that region.
- One more positive reason is that the average no of rooms 6.28 which is almost 6 rooms in a flat this can also attract the buyers. Some people want to buy flats near highway and the average distance from highway is 9.55 which is approximately 10 miles. But there are some negative reasons.

  Like the average crime rate 4.87, average tax 408.24 and average age of buildings 68.57.

2.Plot a histogram of the Average Price variable. What do you infer?



- The price of each house starts from $5000 to $50000 and the average price was $22000 and by the histogram there are more houses in a price range of $20000 to $25000.
- The histogram is a "right skewed histogram." The average price is affected by the other variables since the average price is the dependent variable for all the other variables in the table.
- The other variables like tax, crime rate, NOX, average room etc., will affect the average price.

Example: - If crime rate and NOX is high the price will be low and if the rooms are more the

price will be high.

3.Compute the covariance matrix. Share your observations.

- Covariance is a measure of the relationship between two random variables where it describes up to what extent they change together.

- In simple words covariance describes about the direction, and if the value is positive integer, then the variables move in the same direction, or if the value is negative integer then the variables move in inverse direction.
- By analysing above covariance matrix Average price and tax have a negative relationship where as Average _price and Average_rooms have a positive relationship.

	CRIME_RATI	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RATE	8.516148									
AGE	0.562915	790.7925								
INDUS	-0.11022	124.2678	46.97143							
NOX	0.000625	2.381212	0.605874	0.013401						
DISTANCE	-0.22986	111.55	35.47971	0.61571	75.66653					
TAX	-8.22932	2397.942	831.7133	13.0205	1333.117	28348.62				
PTRATIO	0.068169	15.90543	5.680855	0.047304	8.743402	167.8208	4.677726			
AVG_ROOM	0.056118	-4.74254	-1.88423	-0.02455	-1.28128	-34.5151	-0.53969	0.492695		
LSTAT	-0.88268	120.8384	29.52181	0.48798	30.32539	653.4206	5.7713	-3.07365	50.89398	
AVG_PRICE	1.162012	-97.3962	-30.4605	-0.45451	-30.5008	-724.82	-10.0907	4.484566	-48.3518	84.41956

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

b) Which are the top 3 negatively correlated pairs.

	CRIME_RATI	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RATE	1									
AGE	0.006859	1								
INDUS	-0.00551	0.644779	1							
NOX	0.001851	0.73147	0.763651	1						
DISTANCE	-0.00906	0.456022	0.595129	0.611441	1					
TAX	-0.01675	0.506456	0.72076	0.668023	0.910228	1				
PTRATIO	0.010801	0.261515	0.383248	0.188933	0.464741	0.460853	1			
AVG_ROOM	0.027396	-0.24026	-0.39168	-0.30219	-0.20985	-0.29205	-0.3555	1		
LSTAT	-0.0424	0.602339	0.6038	0.590879	0.488676	0.543993	0.374044	-0.61381	1	
AVG_PRICE	0.043338	-0.37695	-0.48373	-0.42732	-0.38163	-0.46854	-0.50779	0.69536	-0.73766	1

- A correlation matrix is a table showing relation of coefficients between variables. Each cell in the table shows the correlation between two variables.
- A correlation matrix is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses.

a) Which are the top 3 positively correlated pairs

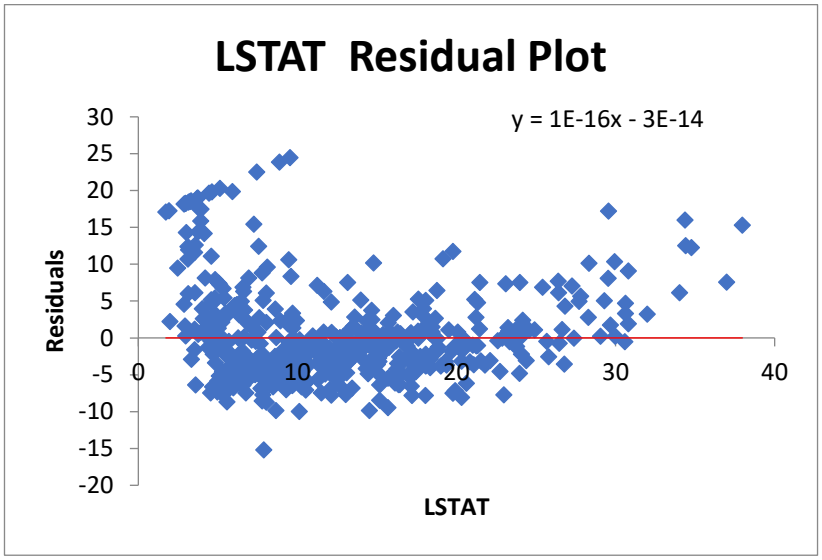
- Top three positively corelated pairs are 0.9102(Tax & Distance), 0.7636(NOX &Indus) and 0.7314(NOX& Age).

b) Which are the top 3 negatively correlated pairs.

- Top three negatively corelated pairs are -0.7376(AVG-Price& LSTAT), -0.6138(LSTAT & AVG-Room) and -0.5077(AVG-Price & PTRATIO).

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

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.737662726							
R Square	0.544146298							
Adjusted R Square	0.543241826							
Standard Error	6.215760405							
Observations	506							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	23243.914	23243.91	601.6178711	5.0811E-88			
Residual	504	19472.38142	38.63568					
Total	505	42716.29542						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	34.55384088	0.562627355	61.41515	3.7431E-236	33.44845704	35.65922	33.44846	35.65922
LSTAT	-0.950049354	0.038733416	-24.5279	5.0811E-88	-1.0261482	-0.87395	-1.02615	-0.87395



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

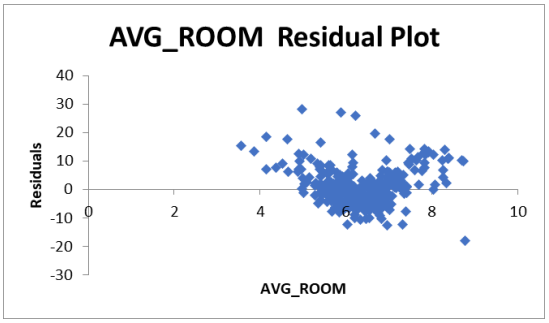
- By observing the regression summary output, we know that if the coefficient value is positive and also it is increasing the variance will also increases but if coefficient value is negative and it is increasing the variance will decreases, there is some pattern in the trendline where it is a straight line. By residuals, we know that if the value is less than 0.05 then it is significant.

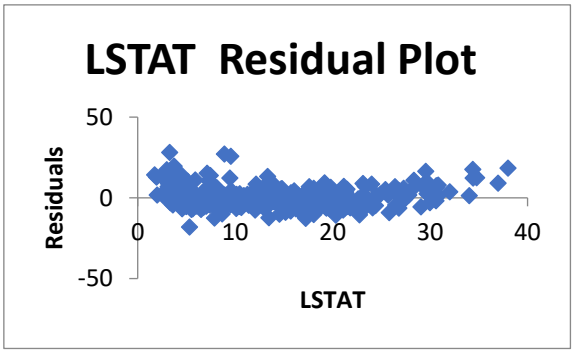
b) Is LSTAT variable significant for the analysis based on your model?

- LSTAT is significant because the p-value is less than 0.05.

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

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.7991							
R Square	0.638562							
Adjusted R	0.637124							
Standard E	5.540257							
Observatic	506							
ANOVA								
	df	*SS*	*MS*	*F*	*ignificance F*			
Regressior	2	27276.99	13638.49	444.3309	7E-112			
Residual	503	15439.31	30.69445					
Total	505	42716.3						
	Coefficients	*andard Erı*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-1.35827	3.172828	-0.4281	0.668765	-7.5919	4.875355	-7.5919	4.875355
AVG_ROO	5.094788	0.444466	11.46273	3.47E-27	4.22155	5.968026	4.22155	5.968026
LSTAT	-0.64236	0.043731	-14.6887	6.67E-41	-0.72828	-0.55644	-0.72828	-0.55644





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?

- Y= a+B1*X1+B2*X2...+E Equation defines multiple linear regression so, if we have 7 Rooms and 20 for LSTAT then Y= -1.35 + (5.09*7) + (-0.64*20) = 21.48. 21.48 is equivalent to 21480USD. Therefore, the company is undercharging because it is less than the 30000USD.

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 has an adjusted R square of 0.543 and for this model we have an adjusted R square of 0.637, By these values we can state that this regression has high performances rather than the previous model. Because we know that if adjusted R square is higher the model works better, and if the adjusted R square is lower the model does not work better.

7. Build another Regression model with all variables where AVG_PRICE alone be the Dependent Variable and all the other variables are independent. Interpret the output in terms of adjusted R-square, coefficient and Intercept values. Explain the significance of each independent variable with respect to AVG_PRICE.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.832979
R Square	0.693854
Adjusted R	0.688299
Standard E	5.134764
Observatic	506

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regressior	9	29638.86	3293.207	124.9045	1.9E-121
Residual	496	13077.43	26.3658		
Total	505	42716.3			

	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	29.24132	4.817126	6.070283	2.54E-09	19.77683	38.7058	19.77683	38.7058
CRIME_RA	0.048725	0.078419	0.621346	0.5346572	-0.10535	0.202799	-0.10535	0.202799
AGE	0.032771	0.013098	2.501997	0.0126704	0.007037	0.058505	0.007037	0.058505
INDUS	0.130551	0.063117	2.068392	0.0391209	0.006541	0.254562	0.006541	0.254562
NOX	-10.3212	3.894036	-2.65051	0.0082939	-17.972	-2.67034	-17.972	-2.67034
DISTANCE	0.261094	0.067947	3.842603	0.0001375	0.127594	0.394593	0.127594	0.394593
TAX	-0.0144	0.003905	-3.68774	0.0002512	-0.02207	-0.00673	-0.02207	-0.00673
PTRATIO	-1.07431	0.133602	-8.0411	6.586E-15	-1.3368	-0.81181	-1.3368	-0.81181
AVG_ROO	4.125409	0.442759	9.317505	3.893E-19	3.255495	4.995324	3.255495	4.995324
LSTAT	-0.60349	0.053081	-11.3691	8.911E-27	-0.70778	-0.49919	-0.70778	-0.49919

a. **Intercept the output in terms of adjusted R-Square**
   - we know that the value of R square and adjusted R square indicates the performances of the model i.e., "69.3% and here the regression co efficient is used to describe the relation
     between an independent variable and dependent variable**.**

b. **Coefficient and Intercept values.**
   - Most of the variables have perfectly positive linear relationship with AVG_PRICE. Variables like NOX, tax, PTRATIO, LSTAT have a perfectly negative linear relationship with AVG_PRICE.

c. **Significance of each independent variable with respect to AVG_PRICE.**
   - P-Values we can state except the CRIME_RATE (0.53) is less than P-value, remaining all variables are significant because P-value is Less than 0.05 i.e., said to be 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.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.832835773							
R Square	0.693615426							
Adjusted R	0.688683682							
Standard E	5.131591113							
Observatic	506							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regressior	8	29628.68142	3703.585178	140.6430411	1.911E-122			
Residual	497	13087.61399	26.33322735					
Total	505	42716.29542						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	29.42847349	4.804728624	6.124898157	1.84597E-09	19.98838959	38.8685574	19.98838959	38.8685574
AGE	0.03293496	0.013087055	2.516605952	0.012162875	0.007222187	0.058647734	0.007222187	0.058647734
INDUS	0.130710007	0.063077823	2.072202264	0.038761669	0.006777942	0.254642071	0.006777942	0.254642071
NOX	-10.27270508	3.890849222	-2.640221837	0.008545718	-17.9172457	-2.628164466	-17.9172457	-2.628164466
DISTANCE	0.261506423	0.067901841	3.851242024	0.000132887	0.128096375	0.394916471	0.128096375	0.394916471
TAX	-0.014452345	0.003901877	-3.703946406	0.000236072	-0.022118553	-0.006786137	-0.022118553	-0.006786137
PTRATIO	-1.071702473	0.133453529	-8.030529271	7.08251E-15	-1.333905109	-0.809499836	-1.333905109	-0.809499836
AVG_ROOI	4.125468959	0.44248544	9.323400461	3.68969E-19	3.256096304	4.994841615	3.256096304	4.994841615
LSTAT	-0.605159282	0.0529801	-11.42238841	5.41844E-27	-0.70925186	-0.501066704	-0.70925186	-0.501066704

a. Interpret the output of this model.

- It represents the regression statistics of the significant variables if the P-value variables is less than 0.05.

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?

- We can observe that the current model of adjusted R-Square is 0.6886 gives a slight better performance compare with the previous model (0.6882) because it is slightly greater than the previous model and that the higher adjusted R square gives the great performances.

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?

- A positive coefficient means when the value of independent variables decreases the mean of the dependent variables increases and negative coefficient means as the values of the independent variable increase the mean of the dependent variables decreases, after sorting the value of NOX increases AVG_PRICE decreases. In other word NOX (pollution) increase the AVG_PRICE decreases because of the pollution.

d. Write the regression equation from this model.

- The regression equation is AVG_PRICE = Intercept + (NOX*X1) + (PTRATIO*X2) + (LSTAT*X3) + (TAX*X4) + (AGE*X5) + (INDUS*X6) + (DISTANCE*X7) + (AVG_ROOM*X8)
  Were, AVG_PRICE is dependent with other variables.