TERRO'S REAL ESTATE

Assignment –Terro's real estate agency

JANUARY 15, 2023

Problem Statement (Situation): "Finding out the most relevant features for pricing of a house" 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.

The agency has provided a dataset of 506 houses in Boston. Following are the details of the dataset:

Data Dictionary:

Attribute	Description
CRIME RATE	CRIME RATE per capita crime rate by town
INDUSTRY	proportion of non-retail business acres per town (in percentage terms)
NOX	nitric oxides concentration (parts per 10 million)
AVG_ROOM	average number of rooms per house
AGE	AGE proportion of houses built prior to 1940 (in percentage terms)
DISTANCE	distance from highway (in miles)
TAX	full-value property-tax rate per $10,000
PTRATIO	pupil-teacher ratio by town
LSTAT %	lower status of the population
AVG_PRICE	Average value of houses in $1000's

To do the analysis, you are expected to solve these questions:

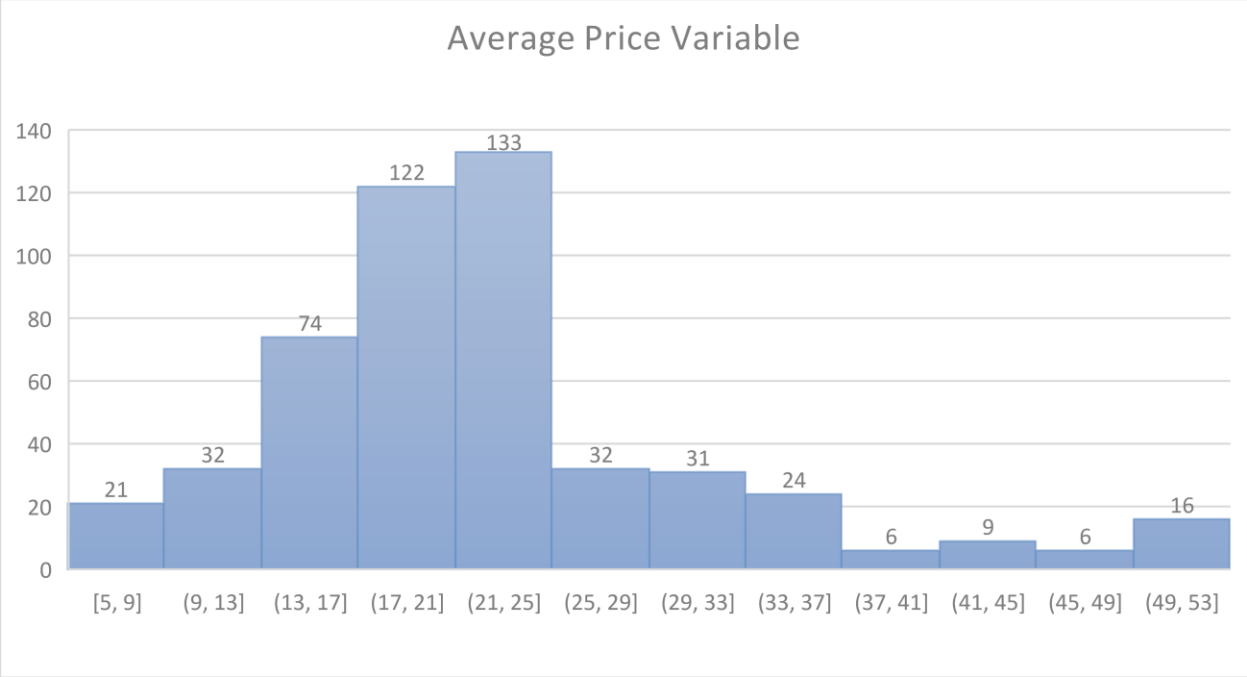
1.Generate the summary statistics for each variable in the table. (Use Data analysis tool pack). Write down your observation.

- ★ With respect to average distance of 10km with respect to crime rate 5 per capita crime rate in town.
- ★ Average value of house is 23$ with respect to average distance of 10km
- ★ The full value property tax rate is 408 with average house of 23$
- ★ Within 10km there is an average of 69% of house build prior to 1910.
- ★ With respect to distance the lower status population is 13%.
- ★ For the NOX the range is 0.48 with respect to average price is 22

Observation :

Based on the variable that effect the pricing is based on the different variable if the crime rate per capita by town then the pricing is increased Some other factor like Distance from highway (in miles), Industry and average age of the house build Prior to 1940**.**

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



A histogram shows how frequently a value falls into a particular bin. The height of each bar represents the number of values in the data set that fall within a particular bin. When the y-axis is labeled as "count" or "number", the numbers along the y-axis tend to be discrete positive integers

As the highest number comes under (21,25) that is 133 and lowest price is (37,41) is 6.

As I can infer from the graph that variable Represent Right-Skewed (Positive Skewness)

3.Compute the covariance matrix. Share your observations.

Column1	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RATE	8.516147873									
AGE	0.562915215	790.7925								
INDUS	-0.110215175	124.2678	46.97143							
NOX	0.000625308	2.381212	0.605874	0.013401						
DISTANCE	-0.229860488	111.55	35.47971	0.61571	75.6665313					
TAX	-8.229322439	2397.942	831.7133	13.0205	1333.11674	28348.62				
PTRATIO	0.068168906	15.90543	5.680855	0.047304	8.74340249	167.8208	4.6777263			
AVG_ROOM	0.056117778	-4.74254	-1.88423	-0.02455	-1.28127739	-34.5151	-0.5396945	0.492695216		
LSTAT	-0.882680362	120.8384	29.52181	0.48798	30.3253921	653.4206	5.7713002	-3.073654967	50.89398	
AVG_PRICE	1.16201224	-97.3962	-30.4605	-0.45451	-30.5008304	-724.82	-10.090676	4.484565552	-48.3518	84.41955616

Observation: By using the conditional formatting for this covariance matrix table I come to this infer that. In this table all the positive (+) element form in a diagonal. For all the negative (-) number I had mention red color and for the zero I had mention as no color.

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

Column1	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RATE	1									
AGE	0.006859463	1								
INDUS	-0.005510651	0.644778511	1							
NOX	0.001850982	0.731470104	0.763651447	1						
DISTANCE	-0.009055049	0.456022452	0.595129275	0.611440563	1					
TAX	-0.016748522	0.506455594	0.72076018	0.6680232	0.910228189	1				
PTRATIO	0.010800586	0.261515012	0.383247556	0.188932677	0.464741179	0.460853035	1			
AVG_ROOM	0.02739616	-0.240264931	-0.391675853	-0.302188188	-0.209846668	-0.292047833	-0.355501495	1		
LSTAT	-0.042398321	0.602338529	0.603799716	0.590878921	0.488676335	0.543993412	0.374044317	-0.613808272	1	
AVG_PRICE	0.043337871	-0.376954565	-0.48372516	-0.427320772	-0.381626231	-0.468535934	-0.507786686	0.695359947	-0.73766	1

a) Which are the top 3 positively correlated pairs

positive correlated pairs	Value
Tax and Distance	0.910228189
Nox and indus	0.763651447
Nox and age	0.731470104

b) Which are the top 3 negatively correlated pairs.

Negative correlated pairs	Value
INDUS AND CRIME RATE	-0.005510651
DISTANCE AND CRIME RATE	-0.009055049
TAX AND CRIME_RATE	-0.016748522

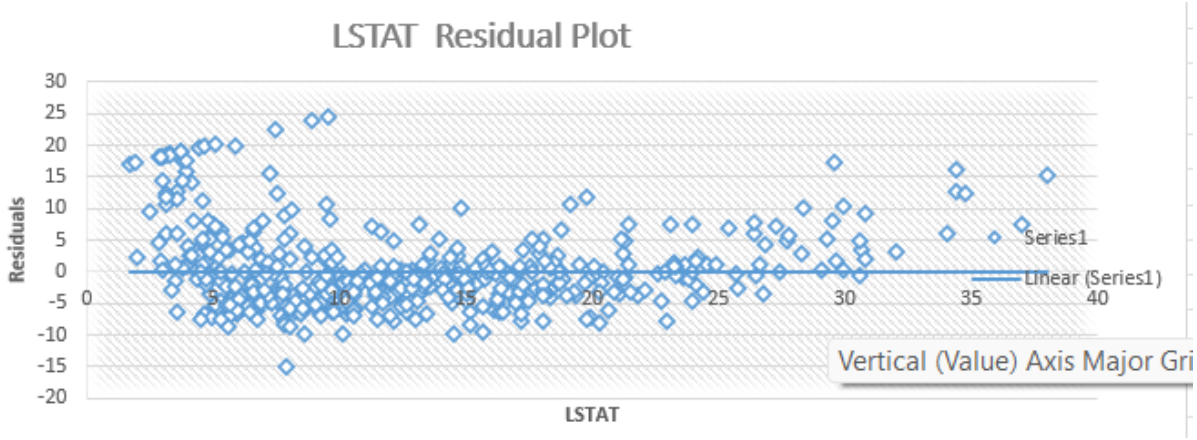
5. Build an initial regression model with AVG_PRICE as 'y' (Dependent variable) and LSTAT variable as Independent Variable. Generate the residual plot

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*	*gnificance F*
Regression	1	23243.914	23243.91	601.6178711	5.08E-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	33.44846	35.65922	33.44846	35.65922
LSTAT	-0.95004935	0.038733416	-24.5279	5.0811E-88	-1.02615	-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 the Residual plot?

The value of R square percentage variance is 54% then the value is significant for 54% with respect variance Y

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

Yes this model is significant to analysis its P-value is 5.0811E-88 and t Stat -24.5279

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	0.799100498
R Square	0.638561606
Adjusted R Square	0.637124475
Standard Error	5.540257367
Observations	506

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	27276.98621	13638.49311	444.3309	7.0085E-112
Residual	503	15439.3092	30.69445169		
Total	505	42716.29542			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-1.358272812	3.17282778	-0.428095348	0.668765	-7.591900282	4.875354658	-7.591900282	4.875354658
AVG_ROOM	5.094787984	0.4444655	11.46272991	3.47E-27	4.221550436	5.968025533	4.221550436	5.968025533
LSTAT	-0.642358334	0.043731465	-14.68869925	6.67E-41	-0.728277167	-0.556439501	-0.728277167	-0.556439501

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 EQUATION

Y= B0 + B1 * X1 + B2 * X2

Y = -1.358+(5.094*7) + (-0.642*20) = 21.46

As the company is quoting for a value of 30000 USD for this locality by regression equation we get to know that the company is overcharging.

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

Yes,this model is better than the previous model because in the previous model the adjusted R-square is 54% and this model the adjusted R-square value is 0.63 which is independent variable that explain 63% of the variation in the dependent variable Ideally this model performance well compare to this 5question.model.

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.

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

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	9	29638.8605	3293.206722	124.9045049	1.9328E-121
Residual	496	13077.43492	26.3657962		
Total	505	42716.29542			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	29.24131526	4.817125596	6.070282926	2.53978E-09	19.77682784	38.70580267	19.77682784	38.7058027
CRIME_RATE	0.048725141	0.078418647	0.621346369	0.534657201	-0.105348544	0.202798827	-0.10534854	0.20279883
AGE	0.032770689	0.013097814	2.501996817	0.012670437	0.00703665	0.058504728	0.00703665	0.05850473
INDUS	0.130551399	0.063117334	2.068392165	0.03912086	0.006541094	0.254561704	0.006541094	0.2545617
NOX	-10.3211828	3.894036256	-2.650510195	0.008293859	-17.97202279	-2.670342809	-17.9720228	-2.67034281
DISTANCE	0.261093575	0.067947067	3.842602576	0.000137546	0.127594012	0.394593138	0.127594012	0.39459314
TAX	-0.01440119	0.003905158	-3.687736063	0.000251247	-0.022073881	-0.0067285	-0.02207388	-0.0067285
PTRATIO	-1.074305348	0.133601722	-8.041104061	6.58642E-15	-1.336800438	-0.811810259	-1.33680044	-0.81181026
AVG_ROOM	4.125409152	0.442758999	9.317504929	3.89287E-19	3.255494742	4.995323561	3.255494742	4.99532356
LSTAT	-0.603486589	0.053081161	-11.36912937	8.91071E-27	-0.70777824	-0.499194938	-0.70777824	-0.49919494

As we know, if the p-value is less than 0.05 then the variable is significance and if the p-value is greater than 0.05 then the variable is insignificance.

So the significance variables are AGE, INDUS, NOX, DISTANCE and TAX and the insignificance variables are CRIME_RATE, PTRATIO, AVG_ROOM, LSTAT.

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:

SUMMARY OUTPUT

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

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	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_ROOM	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

From the model we can interpret the regression statistics output as we can observe from the table Multiple

Multiple R	0.8
R square	0.69
Adjusted R square	0.68
Standard Error	5.13

From the ANOVA regression significant is negative valued so all the category value are 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?

Adjusted R square	0.688683682
Adjusted R square	0.688298647

As we can infer from this table that the adjusted R square from the previous table negative

And the for this model the adjusted R Square is positive .

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	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
NOX	-10.27270508	3.890849222	-2.640221837	0.008545718	-17.9172457	-2.628164466	-17.9172457	-2.628164
PTRATIO	-1.071702473	0.133453529	-8.030529271	7.08251E-15	-1.333905109	-0.809499836	-1.333905109	-0.8095
LSTAT	-0.605159282	0.0529801	-11.42238841	5.41844E-27	-0.70925186	-0.501066704	-0.70925186	-0.501067
TAX	-0.014452345	0.003901877	-3.703946406	0.000236072	-0.022118553	-0.006786137	-0.022118553	-0.006786
AGE	0.03293496	0.013087055	2.516605952	0.012162875	0.007222187	0.058647734	0.007222187	0.0586477
INDUS	0.130710007	0.063077823	2.072202264	0.038761669	0.006777942	0.254642071	0.006777942	0.2546421
DISTANCE	0.261506423	0.067901841	3.851242024	0.000132887	0.128096375	0.394916471	0.128096375	0.3949165
AVG_ROOM	4.125468959	0.44248544	9.323400461	3.68969E-19	3.256096304	4.994841615	3.256096304	4.9948416
Intercept	29.42847349	4.804728624	6.124898157	1.84597E-09	19.98838959	38.8685574	19.98838959	38.868557

After we sort the coefficient in ascending order the coefficients respect to average price of NOX decrease so if the nitric oxide decrease then the locality in town will increase

d)Write the regression equation from this model.

Regression Equation

B0+(B1*X1)+(B2*X2)+(B3*X3)+(B4*X4)+(B5*X5)+(B6*X6)+(B7*X7)+(B8*X8)