BUSINESS REPORT

SHIKHAR TYAGI

DA ONLINE SEPT 23

30/09/23

Great Learning

POWER AHEAD

PROBLEM STATEMENT

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.

Objective (Task):

Your job, as an auditor, is to analyze the magnitude of each variable to which it can affect the price of a house in a particular locality.

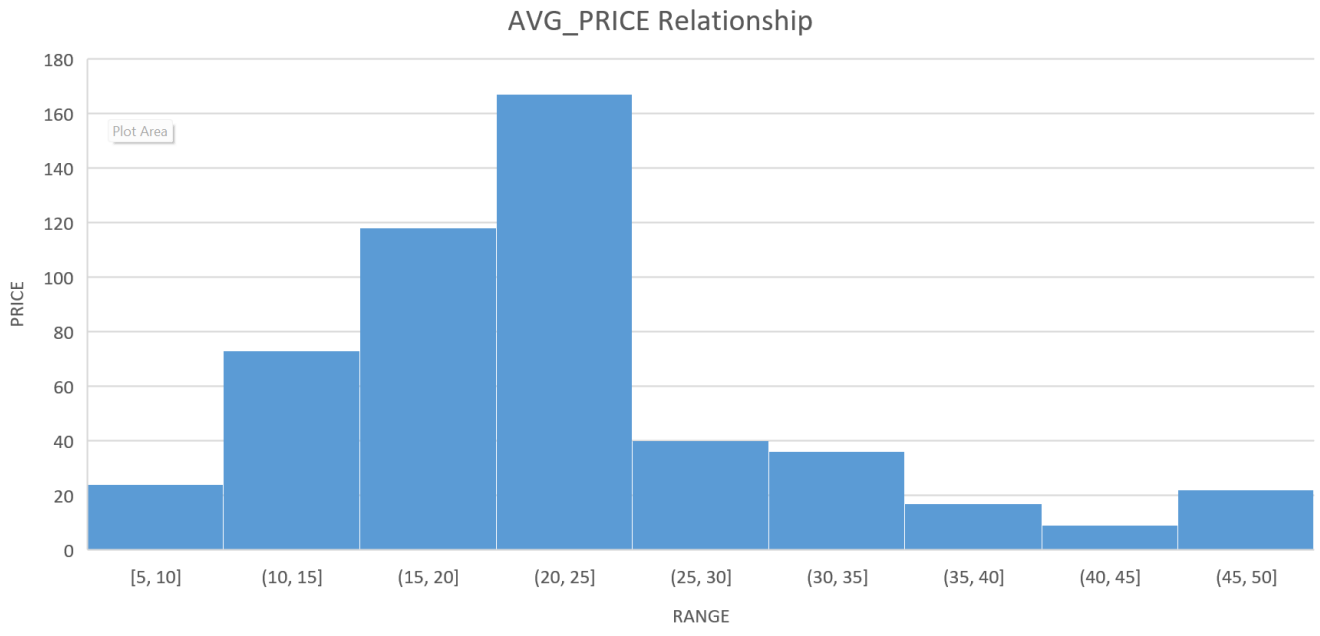
1) Generate the summary statistics for each variable in the table. (Use Data analysis tool pack). Write down your observation.

	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
Mean	4.871976285	68.57490119	11.13677866	0.554695059	9.549407115	408.2371542	18.4555336	6.284634387	12.65306324	22.53280632
Standard Error	0.129860152	1.251369525	0.304979888	0.005151391	0.387084894	7.492388692	0.096243568	0.031235142	0.317458906	0.408861147
Median	4.82	77.5	9.69	0.538	5	330	19.05	6.2085	11.36	21.2
Mode	3.43	100	18.1	0.538	24	666	20.2	5.713	8.05	50
Standard Deviation	2.921131892	28.14886141	6.860352941	0.115877676	8.707259384	168.5371161	2.164945524	0.702617143	7.141061511	9.197104087
Sample Variance	8.533011532	792.3583985	47.06444247	0.013427636	75.81636598	28404.75949	4.686989121	0.49367085	50.99475951	84.58672359
Kurtosis	-1.189122464	-0.967715594	-1.233539601	-0.064667133	-0.867231994	-1.142407992	-0.285091383	1.891500366	0.493239517	1.495196944
Skewness	0.021728079	-0.59896264	0.295021568	0.729307923	1.004814648	0.669955942	-0.802324927	0.403612133	0.906460094	1.108098408
Range	9.95	97.1	27.28	0.486	23	524	9.4	5.219	36.24	45
Minimum	0.04	2.9	0.46	0.385	1	187	12.6	3.561	1.73	5
Maximum	9.99	100	27.74	0.871	24	711	22	8.78	37.97	50
Sum	2465.22	34698.9	5635.21	280.6757	4832	206568	9338.5	3180.025	6402.45	11401.6
Count	506	506	506	506	506	506	506	506	506	506
CV	1.667838517	2.436151864	1.623353602	4.786901844	1.096717887	2.422238874	8.524710389	8.944607239	1.771874282	2.449989269

Figure 1 Summary Statistics

- We observe, AVG_PRICE is highest positively skewed.(Right skewed)
- PTRATIO has the highest negative skewness.(Left skewed)
- We calculate the Coefficient of Variance to be highest for AVG_ROOM hence it is the least consistent and Distance is the most consistent.
- Looking at the Min and Max, There could be some outliers in AGE and CRIME_RATE columns.

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



The Distribution shows a Positive skewed or Right skewed Relationship.

3) Compute the covariance matrix. Share your observations.

	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RATE	8.516147873									
AGE	0.562915215	790.7924728								
INDUS	-0.11021518	124.2678282	46.97142974							
NOX	0.000625308	2.381211931	0.605873943	0.013401099						
DISTANCE	-0.22986049	111.5499555	35.47971449	0.615710224	75.66653127					
TAX	-8.22932244	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	-1.88422543	-0.02455483	-1.28127739	-34.515101	-0.53969452	0.492695216		
LSTAT	-0.88268036	120.8384405	29.52181125	0.487979871	30.32539213	653.4206174	5.771300243	-3.07365497	50.89397935	
AVG_PRICE	1.16201224	-97.3961529	-30.460505	-0.45451241	-30.5008304	-724.820428	-10.0906756	4.484565552	-48.3517922	84.4195562

TAX and AGE have the highest Covariance, meanwhile TAX and AVG_PRICE have the least Covariance.

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

a) Which are the top 3 positively correlated pairs and
b) Which are the top 3 negatively correlated pairs.

	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	-0.737662726	1

We observe, TAX & DISTANCE, NOX & INDUS, NOX & AGE are the top positively correlated pairs.
While, AVG_PRICE& LSTAT, AVG_ROOM& LSTAT and AVG_PRICE& PTRATIO
Are the least correlated pairs.

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?

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

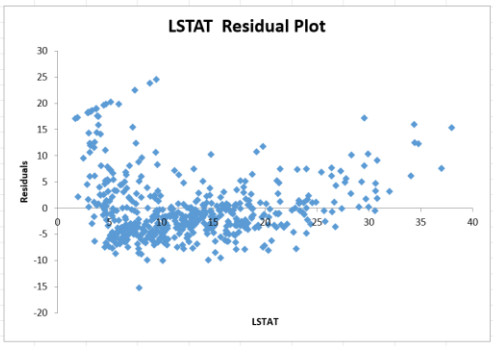
SUMMARY OUTPUT

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	23243.914	601.6178711	5.0811E-88
Residual	504	19472.38142	38.63567742		
Total	505	42716.29542			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	34.55384088	0.562627355	61.41514552	3.7431E-236	33.44845704	35.65922472	33.44845704	35.65922472
LSTAT	-0.950049354	0.038733416	-24.52789985	5.0811E-88	-1.0261482	-0.873950508	-1.0261482	-0.873950508



RESIDUAL OUTPUT

Observation	Predicted AVG_PRICE	Residuals
1	29.8225951	-5.822595098
2	25.87038979	-4.270389786
3	30.72514198	3.974858016
4	31.76069578	1.639304221
5	29.49007782	6.709922176
6	29.60408375	-0.904083746
7	22.74472741	0.155272588
8	16.36039575	10.73960425
9	6.118863721	10.38113628
10	18.30799693	0.59200307
11	15.1253316	-0.125331595
12	21.94668596	-3.046685955
13	19.62856553	2.071434468

a) Since the value of R Square is less than 60%, the model is not fit to describe the variation in Price. Negative Coefficient represents that that LSTAT and price are inversely related. The trendline is flat, representing Normal distribution.

b) P value is less than 0.05 hence the variables are significant according to this model.

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?

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.

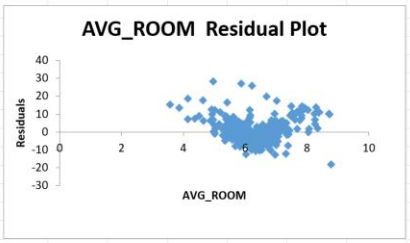
SUMMARY OUTPUT

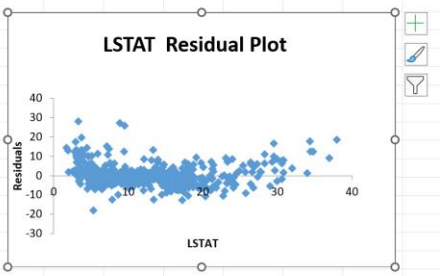
Regression Statistics	
Multiple R	0.799100498
R Square	0.638561606
Adjusted R Square	0.637124475
Standard Error	5.540257367
Observations	506

ANOVA

	df	SS	MS	F	Significance F
Regression	2	27276.98621	13638.49311	444.3308922	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	0.668764941	-7.591900282	4.875354658	-7.591900282	4.875354658
AVG_ROOM	5.094787984	0.4444655	11.46272991	3.47226E-27	4.221550436	5.968025533	4.221550436	5.968025533
LSTAT	-0.642358334	0.043731465	-14.68869925	6.66937E-41	-0.728277167	-0.556439501	-0.728277167	-0.556439501





RESIDUAL OUTPUT

Observation	Predicted AVG_PRICE	Residuals
1	28.94101368	-4.941013681
2	25.48420566	-3.884205661
3	32.65907477	2.040925231
4	32.40652	0.99348
5	31.63040699	4.569593009
6	28.05452701	0.645472994
7	21.28707846	1.612921545

a) Regression Equation can be written as: y=mx+c

Putting values in equation, We get Predicted value to be 21.46K, Hence the company is overcharging.

b) Since, the value of Rsquare and adjusted R square is greater than previous model, this model is 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 Rsquare, coefficient and Intercept values. Explain the significance of each independent variable with respect to AVG_PRICE.

SUMMARY OUTPUT

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

Met

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	9	29638.8605	3293.206722	124.904505	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	2.5398E-09	19.77682784	38.70580267	19.77682784	38.70580267
CRIME_RATE	0.048725141	0.078418647	0.621346369	0.5346572	-0.10534854	0.202798827	-0.10534854	0.202798827
AGE	0.032770689	0.013097814	2.501996817	0.01267044	0.00703665	0.058504728	0.00703665	0.058504728
INDUS	0.130551399	0.063117334	2.068392165	0.03912086	0.006541094	0.254561704	0.006541094	0.254561704
NOX	-10.3211828	3.894036256	-2.6505102	0.00829386	-17.9720228	-2.67034281	-17.9720228	-2.67034281
DISTANCE	0.261093575	0.067947067	3.842602576	0.00013755	0.127594012	0.394593138	0.127594012	0.394593138
TAX	-0.01440119	0.003905158	-3.68773606	0.00025125	-0.02207388	-0.0067285	-0.02207388	-0.0067285
PTRATIO	-1.074305348	0.133601722	-8.04110406	6.5864E-15	-1.33680044	-0.81181026	-1.33680044	-0.81181026
AVG_ROOM	4.125409152	0.442758999	9.317504929	3.8929E-19	3.255494742	4.995323561	3.255494742	4.995323561
LSTAT	-0.603486589	0.053081161	-11.3691294	8.9107E-27	-0.70777824	-0.49919494	-0.70777824	-0.49919494

RESIDUAL OUTPUT

Observation	*Predicted AVG_PRICE*	*Residuals*
1	30.1153558	-6.115355802
2	27.00714024	-5.407140244
3	32.83291255	1.867087455
4	31.20703392	2.192966083

- Value of R square is significant for a good model.
- Difference in R square and Adjusted Rsquare is less than 1% hence this condition is met
- Coefficient is negative for NOX, TAX,PTRATIO, and LSTAT which means these variables are inversely related to the AVG_Price.
- Other coefficients – CRIME_RATE, AGE,INDUS,DISTANCE,AVG_ROOM are directly related to AVG_PRICE which means AVG Price increases with increase in these values.
- CRIME_RATE is the only quantity with p value greater than 0.05 which makes it insignificant.

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.

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?

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?

d) Write the regression equation from this model

SUMMARY OUTPUT

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

Met

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

RMSE

RESIDUAL OUTPUT

RMSE

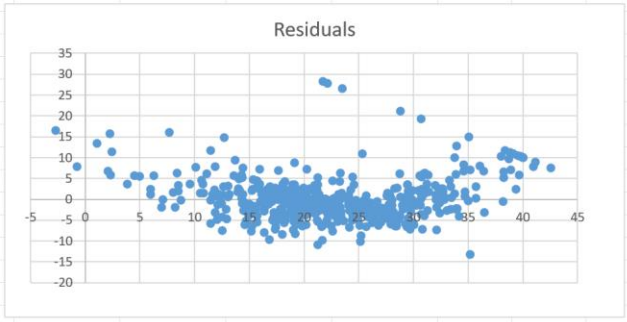
RESIDUAL OUTPUT

Observation	*Predicted AVG_PRICE*	*Residuals*	*Square*	*Mean*	*RMSE*	*Max Possible*
1	30.04888734	-6.048887337	36.58903801	25.86484979	5.085749678	22.53280632
2	27.04098462	-5.440984617	29.60431361			
3	32.69896454	2.001035462	4.004142921			
4	31.14306949	2.256930513	5.093735341			
5	30.58808735	5.611912655	31.49356364			
6	27.85095254	0.849047463	0.720881594			
7	25.07089688	-2.170896878	4.712793257			
8	22.63588287	4.464117131	19.92834176			
9	14.00883345	2.491166552	6.205910791			
10	22.84744402	-3.947444016	15.58231426			
11	22.63561401	-7.63561401	58.30260132			
12	25.08702653	-6.18702653	38.27929728			
13	21.66953684	0.030463156	0.000928004			
14	20.64832118	-0.248321176	0.061663407			
15	20.79207015	-2.592070151	6.718827667			
16	19.87225351	0.027746494	0.000769868			
17	20.53684599	2.563154009	6.569758476			
18	17.59380012	-0.093800118	0.008798462			
19	15.70880764	4.491192361	20.17080882			
20	18.15848523	0.041514769	0.001723476			
21	12.55847507	1.041524935	1.084774189			
22	18.24600939	1.353990606	1.83329056			
23	16.09932591	-0.899325912	0.808787096			
24	14.31342203	0.186577971	0.034811339			

23% High Prediction Error

Assumptions of Residuals

1, Mean	-1.03948E-14	Met
2. Distribution	1.643869514	
3. Scatterplot	Parabolic/Normally distributed.	



- Value of R square is similar but adjusted Rsquare is marginally higher so model is slightly accurate and better than the previous one.
- P values are all under 0.05 hence all the values are significant now.
- Though this model has high prediction error(23%), rest of the conditions for residuals are met (refer to fig).
- If the value of NOX increases, average price decreases. Since there is an inverse relation.
- Y=mx+c: NOX*-10.27+PTRATIO*-1.0717+LSTAT*-0.605+TAX*-0.0144+AGE*0.1307+DISTANCE*0.26+AVG_ROOM*4.125

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

Coefficients Ascending