Business Report

Title:

Terro's Real Estate Agency

Submitted By,

Arulson.Y

Q1) Generate the summary statistics for each variable in the table. (Use Data analysis tool pack). Write down your observation.

CRIME_RATE		*AGE*		*INDUS*	
Mean	4.871976285	Mean	68.57490119	Mean	11.13677866
Standard Error	0.129860152	Standard Error	1.251369525	Standard Error	0.304979888
Median	4.82	Median	77.5	Median	9.69
Mode	3.43	Mode	100	Mode	18.1
Standard Deviation	2.921131892	Standard Deviation	28.14886141	Standard Deviation	6.860352941
Sample Variance	8.533011532	Sample Variance	792.3583985	Sample Variance	47.06444247
Kurtosis	-1.189122464	Kurtosis	-0.967715594	Kurtosis	-1.233539601
Skewness	0.021728079	Skewness	-0.59896264	Skewness	0.295021568
Range	9.95	Range	97.1	Range	27.28
Minimum	0.04	Minimum	2.9	Minimum	0.46
Maximum	9.99	Maximum	100	Maximum	27.74
Sum	2465.22	Sum	34698.9	Sum	5635.21
Count	506	Count	506	Count	506

NOX		*DISTANCE*		*TAX*	
Mean	0.554695059	Mean	9.549407115	Mean	408.2371542
Standard Error	0.005151391	Standard Error	0.387084894	Standard Error	7.492388692
Median	0.538	Median	5	Median	330
Mode	0.538	Mode	24	Mode	666
Standard Deviation	0.115877676	Standard Deviation	8.707259384	Standard Deviation	168.5371161
Sample Variance	0.013427636	Sample Variance	75.81636598	Sample Variance	28404.75949
Kurtosis	-0.064667133	Kurtosis	-0.867231994	Kurtosis	-1.142407992
Skewness	0.729307923	Skewness	1.004814648	Skewness	0.669955942
Range	0.486	Range	23	Range	524
Minimum	0.385	Minimum	1	Minimum	187
Maximum	0.871	Maximum	24	Maximum	711
Sum	280.6757	Sum	4832	Sum	206568
Count	506	Count	506	Count	506

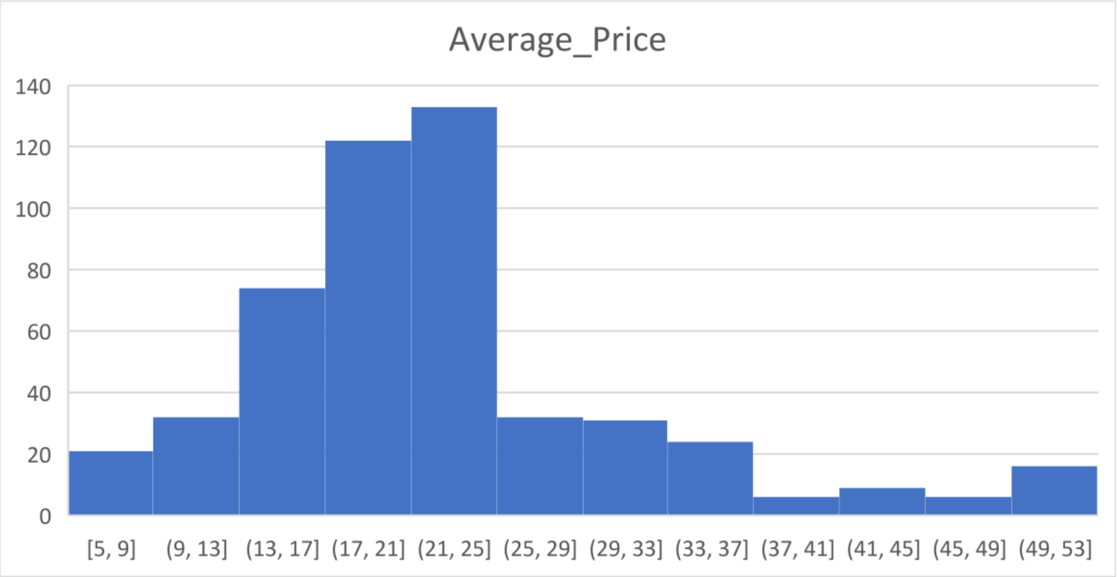
PTRATIO		*AVG_ROOM*		*LSTAT*	
Mean	18.4555336	Mean	6.284634387	Mean	12.65306324
Standard Error	0.096243568	Standard Error	0.031235142	Standard Error	0.317458906
Median	19.05	Median	6.2085	Median	11.36
Mode	20.2	Mode	5.713	Mode	8.05
Standard Deviation	2.164945524	Standard Deviation	0.702617143	Standard Deviation	7.141061511
Sample Variance	4.686989121	Sample Variance	0.49367085	Sample Variance	50.99475951
Kurtosis	-0.285091383	Kurtosis	1.891500366	Kurtosis	0.493239517
Skewness	-0.802324927	Skewness	0.403612133	Skewness	0.906460094
Range	9.4	Range	5.219	Range	36.24
Minimum	12.6	Minimum	3.561	Minimum	1.73
Maximum	22	Maximum	8.78	Maximum	37.97
Sum	9338.5	Sum	3180.025	Sum	6402.45
Count	506	Count	506	Count	506

AVG_PRICE	
Mean	22.53280632
Standard Error	0.408861147
Median	21.2
Mode	50
Standard Deviation	9.197104087
Sample Variance	84.58672359
Kurtosis	1.495196944
Skewness	1.108098408
Range	45
Minimum	5
Maximum	50
Sum	11401.6
Count	506

Observation:

- The mean provides the average value of each variable.
- Standard deviation measures the dispersion or spread of the data.
- Median is the middle value in a data set.
- Mode is the most frequently occurring value.
- Skewness measures the asymmetry of the data distribution.
- Kurtosis measures the "tailedness" of the data distribution.
- Range is the difference between the maximum and minimum values.

Q2) Plot a histogram of the Avg_Price variable. What do you infer?



A histogram is a visual depiction of the distribution of AVG_PRICE values in a dataset. The x-axis displays intervals or bins of average prices, while the y-axis indicates the frequency or count of instances falling within each AVG_PRICE range. In this histogram plot (21,25) have the most AVG_PRICEs and (37,41) have the least AVG_PRICEs .

Q3) Compute the covariance matrix. Share your observations.

	CRIME_RATE	*AGE*	*INDUS*	*NOX*	*DISTANCE*	*TAX*	*PTRATIO*	*AVG_ROOM*	*LSTAT*	*AVG_PRICE*
CRIME_RATE	8.5161									
AGE	0.5629	790.7925								
INDUS	-0.1102	124.2678	46.9714							
NOX	0.0006	2.3812	0.6059	0.0134						
DISTANCE	-0.2299	111.5500	35.4797	0.6157	75.6665					
TAX	-8.2293	2397.9417	831.7133	13.0205	1333.1167	28348.6236				
PTRATIO	0.0682	15.9054	5.6809	0.0473	8.7434	167.8208	4.6777			
AVG_ROOM	0.0561	-4.7425	-1.8842	-0.0246	-1.2813	-34.5151	-0.5397	0.4927		
LSTAT	-0.8827	120.8384	29.5218	0.4880	30.3254	653.4206	5.7713	-3.0737	50.8940	
AVG_PRICE	1.1620	-97.3962	-30.4605	-0.4545	-30.5008	-724.8204	-10.0907	4.4846	-48.3518	84.4196

Observation:

Positive Relationships:

- Variables with positive covariances tend to increase or decrease together.
- **Examples:** AGE and CRIME_RATE, DISTANCE and NOX.

Negative Relationship:

- Variables with negative covariances tend to move in opposite directions.
- **Examples:** TAX and CRIME_RATE, AVG_PRICE and NOX.

Q4) Create a correlation matrix of all the variables.

(Use Data analysis tool pack).

	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

a) Which are the top 3 positively correlated pairs.

Top 3 positively correlated:

1.DISTANCE and TAX:	0.910228189
2.INDUS and NOX:	0.763651447
3.AGE and NOX:	0.731470104

b) Which are the top 3 negatively correlated pairs.

Top 3 negatively correlated:

1.AVG_PRICE and LSTAT:	-0.737662726
2.LSTAT and AVG_ROOM:	-0.613808272
3.AVG_PRICE and PTRATIO:	-0.507786686

Q5) Build an initial regression model with AVG_PRICE as 'y' (Dependent variable) and LSTAT 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	23243.914	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

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

Variance Explained:

R-Square value is (0.544146298)

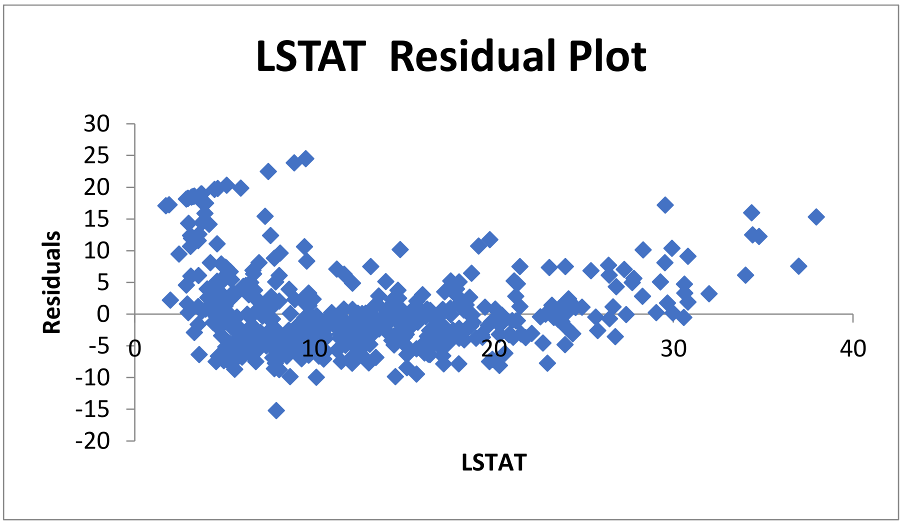
Coefficient value:

The Coefficient for the variable LSTAT is (-0.95004)

Intercept:

The Intercept is (34.5538)

Residual Plot:



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

Yes, the LSTAT variable is significant for the analysis as indicates by it's very low P-value (5.0811E-88) in the regression output.

Q6) 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	0.973885353							
R Square	0.948452681							
Adjusted R Square	0.946366278							
Standard Error	5.53576654							
Observations	506							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	2	284181.4056	142090.7028	4636.712087	0			
Residual	504	15444.93444	30.64471119					
Total	506	299626.34						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
AVG_ROOM	4.906906071	0.070193339	69.90557997	1.6137E-261	4.768998482	5.044813661	4.768998482	5.044813661
LSTAT	-0.655739993	0.030558561	-21.45847115	4.81185E-73	-0.715777847	-0.595702138	-0.715777847	-0.595702138

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?

Ans:

Regression Equation:

AVG_PRICE=4.9069*AVG_ROOM+(-0.6557) *LSTAT

AVG_ROOM=7
LSTAT=20

AVG_PRICE is 21.23354265

Predicted value is	$	21,233.54
Company quoting value is	$	30,000.00

Yes, 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.

Ans:

Adjusted R Square is (0.946366278)

Question 5 Adjusted R Square is (0.543241826)

Based on the adjusted R-Square values, it can be inferred that the performance of the current model is better than the previous model. The higher adjusted R-Square indicates a most effective and accurate model in explaining the variance in the dependent variable.

Q7) 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	0.832978824							
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	38.70580267
CRIME_RATE	0.048725141	0.078418647	0.621346369	0.534657201	-0.105348544	0.202798827	-0.105348544	0.202798827
AGE	0.032770689	0.013097814	2.501996817	0.012670437	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	-2.650510195	0.008293859	-17.97202279	-2.670342809	-17.97202279	-2.670342809
DISTANCE	0.261093575	0.067947067	3.842602576	0.000137546	0.127594012	0.394593138	0.127594012	0.394593138
TAX	-0.01440119	0.003905158	-3.687736063	0.000251247	-0.022073881	-0.0067285	-0.022073881	-0.0067285
PTRATIO	-1.074305348	0.133601722	-8.041104061	6.58642E-15	-1.336800438	-0.811810259	-1.336800438	-0.811810259
AVG_ROOM	4.125409152	0.442758999	9.317504929	3.89287E-19	3.255494742	4.995323561	3.255494742	4.995323561
LSTAT	-0.603486589	0.053081161	-11.36912937	8.91071E-27	-0.70777824	-0.499194938	-0.70777824	-0.499194938

Adjusted R-Square:

The adjusted R-Square is 68.83%.

Coefficients:

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)

Intercept:

The Intercept value is (29.2413).

Explanation:

The model suggests that AGE, INDUS, NOX, DISTANCE, TAX, PTRATIO, AVG_ROOM, LSTAT are statistically significant predictors of AVG_PRICE, while CRIME_RATE is not statistically significant.

Q8) 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%*
NOX		3.890849222	-2.640221837	0.008545718	-17.9172457	-2.628164466	-17.9172457	-2.628164466
PTRATIO	-1.071702473	0.133453529	-8.030529271	7.08251E-15	-1.333905109	-0.809499836	-1.333905109	-0.809499836
LSTAT	-0.605159282	0.0529801	-11.42238841	5.41844E-27	-0.70925186	-0.501066704	-0.70925186	-0.501066704
TAX	-0.014452345	0.003901877	-3.703946406	0.000236072	-0.022118553	-0.006786137	-0.022118553	-0.006786137
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
DISTANCE	0.261506423	0.067901841	3.851242024	0.000132887	0.128096375	0.394916471	0.128096375	0.394916471
AVG_ROOM	4.125468959	0.44248544	9.323400461	3.68969E-19	3.256096304	4.994841615	3.256096304	4.994841615
Intercept	29.42847349	4.804728624	6.124898157	1.84597E-09	19.98838959	38.8685574	19.98838959	38.8685574

a) Interpret the output of this model.

Adjusted R-Square:

The adjusted R-Square is 68.86%.

Coefficients:

AVG_ROOM:	(4.125468959)
DISTANCE:	(0.261506423)
INDUS:	(0.130710007)
AGE:	(0.03293496)
TAX:	(-0.014452345)
LSTAT:	(-0.605159282)
PTRATIO:	(-1.071702473)
NOX:	(-10.27270508)

Intercept:

The Intercept value is (29.4284).

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?

Ans:

The adjusted R-square for this model is 0.6887, while the adjusted R-square for the model in the previous question was 0.6883.

The difference is minimal, and both models perform similarly in explaining the variability in AVG_PRICE.

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 in Ascending order:

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)

If the value of NOX is more in a locality according to the model, the average price tends to decrease.

d) Write the regression equation from this model.

AVG_PRICE=(4.125*AVG_ROOM)+(0.261*DISTANCE)+(0.1307*INDUS)+

(0.032*AGE)+(-0.014*TAX)+(-0.6051*LSTAT)

+(-1.0717*PTRATIO)+(-10.2727*NOX)+(29.4284)