TERRO'S REAL ESTATE

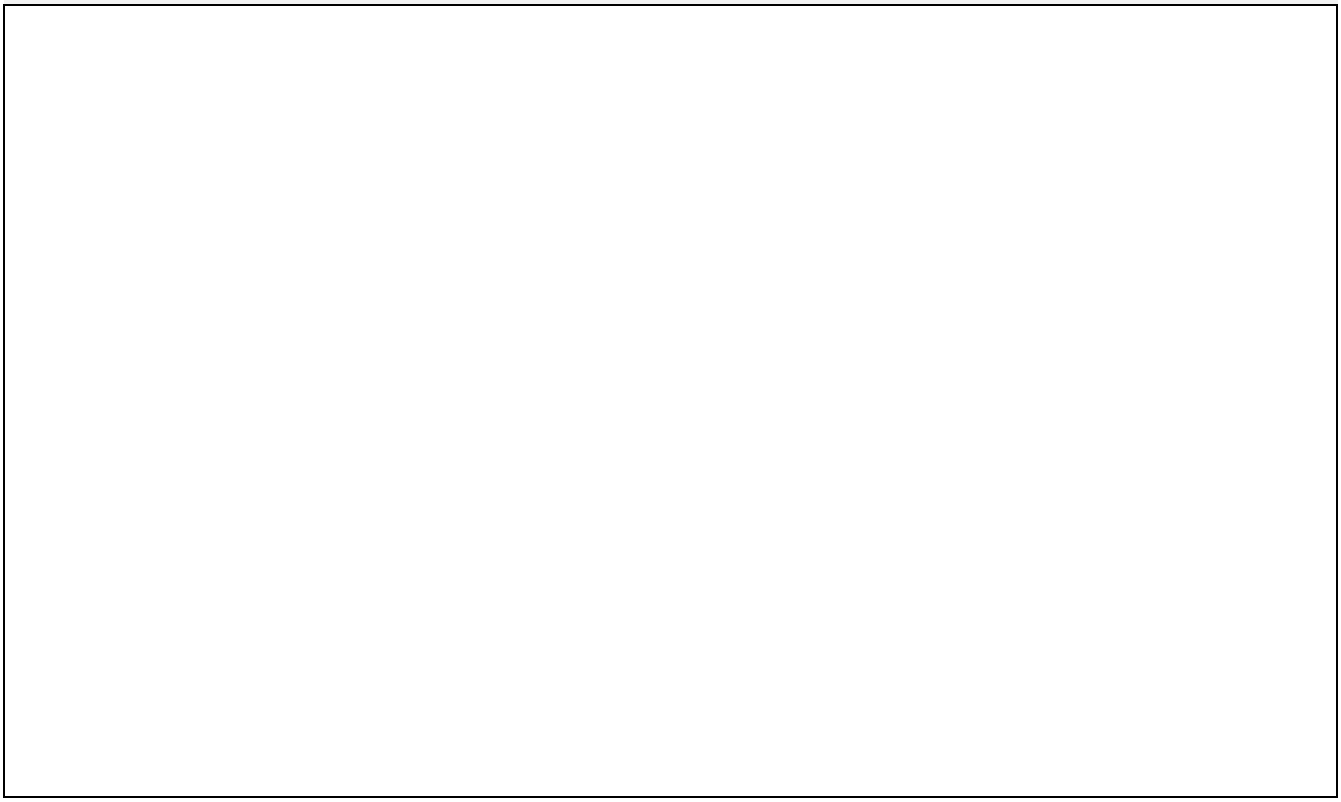
2023

JULY 21

TERRO'S REAL ESTATE AGENCY

Authored by: Bhavani A

Logo Name



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

CRIME_RATE		*AGE*	
Mean	4.871976285	Mean	68.57490119
Standard Error	0.129860152	Standard Error	1.251369525
Median	4.82	Median	77.5
Mode	3.43	Mode	100
Standard Deviation	2.921131892	Standard Deviation	28.14886141
Sample Variance	8.533011532	Sample Variance	792.3583985
Kurtosis	-1.189122464	Kurtosis	-0.967715594
Skewness	0.021728079	Skewness	-0.59896264
Range	9.95	Range	97.1
Minimum	0.04	Minimum	2.9
Maximum	9.99	Maximum	100
Sum	2465.22	Sum	34698.9
Count	506	Count	506
Confidence Level(95.0%)	0.255132688	Confidence Level(95.0%)	2.458531467

INDUS		*NOX*	
Mean	11.13677866	Mean	0.554695059
Standard Error	0.304979888	Standard Error	0.005151391
Median	9.69	Median	0.538
Mode	18.1	Mode	0.538
Standard Deviation	6.860352941	Standard Deviation	0.115877676
Sample Variance	47.06444247	Sample Variance	0.013427636
Kurtosis	-1.233539601	Kurtosis	-0.064667133
Skewness	0.295021568	Skewness	0.729307923
Range	27.28	Range	0.486
Minimum	0.46	Minimum	0.385
Maximum	27.74	Maximum	0.871
Sum	5635.21	Sum	280.6757
Count	506	Count	506
Confidence Level(95.0%)	0.599185642	Confidence Level(95.0%)	0.010120797

DISTANCE		*TAX*	
Mean	9.549407115	Mean	408.2371542
Standard Error	0.387084894	Standard Error	7.492388692
Median	5	Median	330
Mode	24	Mode	666
Standard Deviation	8.707259384	Standard Deviation	168.5371161
Sample Variance	75.81636598	Sample Variance	28404.75949
Kurtosis	-0.867231994	Kurtosis	-1.142407992
Skewness	1.004814648	Skewness	0.669955942
Range	23	Range	524
Minimum	1	Minimum	187
Maximum	24	Maximum	711
Sum	4832	Sum	206568
Count	506	Count	506
Confidence Level(95.0%)	0.760495101	Confidence Level(95.0%)	14.72009106

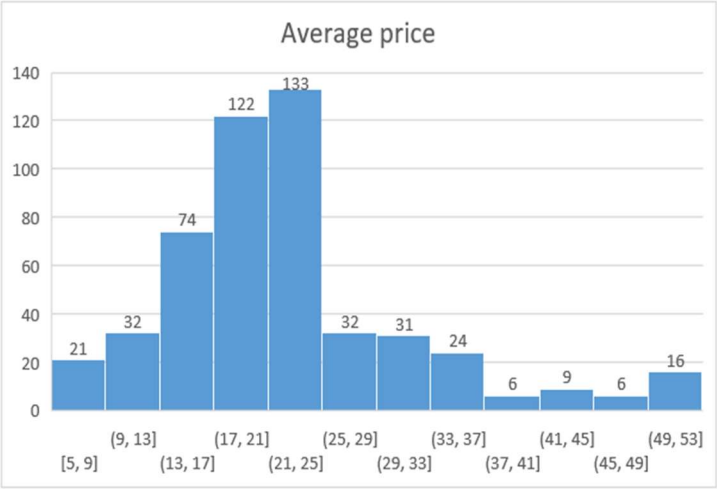
PTRATIO		*AVG_ROOM*	
Mean	18.4555336	Mean	6.284634387
Standard Error	0.096243568	Standard Error	0.031235142
Median	19.05	Median	6.2085
Mode	20.2	Mode	5.713
Standard Deviation	2.164945524	Standard Deviation	0.702617143
Sample Variance	4.686989121	Sample Variance	0.49367085
Kurtosis	-0.285091383	Kurtosis	1.891500366
Skewness	-0.802324927	Skewness	0.403612133
Range	9.4	Range	5.219
Minimum	12.6	Minimum	3.561
Maximum	22	Maximum	8.78
Sum	9338.5	Sum	3180.025
Count	506	Count	506
Confidence Level(95.0%)	0.189087104	Confidence Level(95.0%)	0.061366829

	AVG_PRICE		
Mean	12.65306324	Mean	22.53280632
Standard Error	0.317458906	Standard Error	0.408861147
Median	11.36	Median	21.2
Mode	8.05	Mode	50
Standard Deviation	7.141061511	Standard Deviation	9.197104087
Sample Variance	50.99475951	Sample Variance	84.58672359
Kurtosis	0.493239517	Kurtosis	1.495196944
Skewness	0.906460094	Skewness	1.108098408
Range	36.24	Range	45
Minimum	1.73	Minimum	5
Maximum	37.97	Maximum	50
Sum	6402.45	Sum	11401.6
Count	506	Count	506
Confidence Level(95.0%)	0.623702827	Confidence Level(95.0%)	0.80327831

OBSERVATION:

- Skewness measures the Dispersion of data's from the Mean or Average.
- If the mean >median, then the curve is Positively Skewed.
- If the mean <median, then the curve is Negatively Skewed.
- Kurtosis measures the Sharpness or Peakedness or Tailedness of the curve.
- If the Kurtosis value is Below 3 it will results in flat curve called Platykurtic.
- If the Kurtosis value is Above 3 it will results in Sharp curve called Leptokurtic.
- In this Model PTRATIO's Skewness is (-0.803325) which is more negatively Skewed And Average price is more positively skewed(1.108099)
- The Kurtosis value is Ranging between -2 to 2. So the curve is platykurtic.

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



OBSERVATION:

- Here the Bin width of the Average value of the houses is Splited in the interval of 4.
- With this histogram we can infer 133 count of people has average value of houses between $21000 to $25000 which is highest.
- And 6 people has average value of houses between $37000 to $41000 and another 6 people has average value of houses between $45000 to $49000.

3. Compute the covariance matrix. Share your observations.

B21

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		*CRIME_RATE*	*AGE*	*INDUS*	*NOX*	*DISTANCE*	*TAX*	*PTRATIO*	*AVG_ROOM*	*LSTAT*	*AVG_PRICE*		
2	CRIME_RATE	8.516147873											
3	AGE	0.562915215	790.7924728										
4	INDUS	-0.110215175	124.2678282	46.97142974									
5	NOX	0.000625308	2.381211931	0.605873943	0.013401099								
6	DISTANCE	-0.229860488	111.5499555	35.47971449	0.615710224	75.66653127							
7	TAX	-8.229322439	2397.941723	831.7133331	13.02050236	1333.116741	28348.6236						
8	PTRATIO	0.068168906	15.90542545	5.680854782	0.047303654	8.74340249	167.8208221	4.677726296					
9	AVG_ROOM	0.056117778	-4.74253803	-1.884225427	-0.024554826	-1.281277391	-34.515101	-0.539694518	0.492695216				
10	LSTAT	-0.882680362	120.8384405	29.52181125	0.487979871	30.32539213	653.4206174	5.771300243	-3.073654967	50.89397935			
11	AVG_PRICE	1.16201224	-97.39615288	-30.46050499	-0.454512407	-30.50083035	-724.820428	-10.09067561	4.484565552	-48.35179219	84.41955616		

OBSERVATION:

- Covariance matrix tells how variance in Independent variable(X) affects the variance of Dependent variable(Y).
- For example if we take the covariance of Distance and Average prices it has negative covariance. Because the distance increases the Average price decreases .

- So the Distance and Average Prices are Inversely Proportional.

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

	A	B	C	D	E	F	G	H	I	J	K	L
1		*CRIME_RATE*	*AGE*	*INDUS*	*NOX*	*DISTANCE*	*TAX*		*CRIME_RATE*	*AGE*	*INDUS*	
2	CRIME_RATE											
3	AGE	0.006859463										
4	INDUS	-0.005510651	0.644778511									
5	NOX	0.001850982	0.731470104	0.763651447								
6	DISTANCE	-0.009055049	0.456022452	0.595129275	0.611440563							
7	TAX	-0.016748522	0.506455594	0.72076018	0.6680232	0.910228189						
8	PTRATIO	0.010800586	0.261515012	0.383247556	0.188932677	0.464741179	0.460853035					
9	AVG_ROOM	0.02739616	-0.240264931	-0.391675853	-0.302188188	-0.209846668	-0.292047833	-0.355501495				
10	LSTAT	-0.042398321	0.602338529	0.603799716	0.590878921	0.488676335	0.543993412	0.374044317	-0.613808272			
11	AVG_PRICE	0.043337871	-0.376954565	-0.48372516	-0.427320772	-0.381626231	-0.468535934	-0.507786686	0.695359947	-0.737662726		
12												

OBSERVTIONS:

- Correlation -The coeffiecients of independent variables(X) predicting the Dependent Variables(Y) Result.
- For Example:

X1	X2	X3	Y
Calorie intake	Stress	Sedentary Lifestyle	Diabetes

a) Which are the top 3 positively correlated pairs?

1 0.910228189

2 0.763651447

3 0.731470104

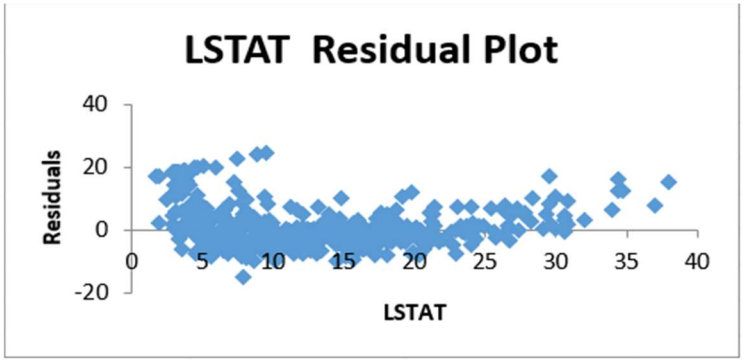
b) Which are the top 3 negatively correlated pairs?

1 -0.737662726

2 -0.613808272

3 -0.507786686

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 Residual plot?

R square value	0.5441463	54%
Coefficient of LSTAT	-0.95004935	
Coefficient of Intercept	34.5538409	

- 1.R square value is the coefficient of determination.
- Here the R square value is 54% ,if we get more percentage It will gives more nearest answer.
- 2. The coefficient of LSTAT is negative . So the LSTAT and Average price are inversely Proportional to each other.
- 3. The LSTAT line fit plot gives negative slope .Then LSTAT and Average price are negatively Correlated.
- 4. In Residual plot We are getting More Biased in all the intervals except 10-20 range.
- So if the residual falls Before 10 or above 20 we will reject this Model.
- 5. Intercept- 34.55384088 .If Y=X the is no intercept point. Then the intercept is Zero. That means aal the X's can predict the 100% of Y.

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

- If p value <0.05 the Model is Significant.
- If p value >0.05 the Model is Not Significant.
- Here the Confidence level is not given .So we assume that it as 95%. So the Level of significance is 1-0.95=0.05.
- Here P value of LSTAT is 5.0811E-88 . Which is <0.05.
- Here we will Reject the Null Hypothesis

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?

1. REGRESSION EQUATION

Y=(M1*X1+M2*X2)+B

Slope 1	M1	5.09478798
Slope 2	M2	-0.64235833
Intercept	B	1.35827281
Independent 1	X1	7
Independent 2	X2	20
	Y	21.4580764

Regression Equation:

Y=21.458 USD

The company is Overcharging .Because the predicted value is $21,458.08 only.

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

Adjusted R Square of 5 th Question 0.543241826

Adjusted R Square of 6 th Question 0.637124475

The highest R square will cause Better results. The current model is more preferable.

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&#65533;square, coefficient and Intercept values. Explain the significance of each independent variable with respect to AVG_PRICE.

1. Adjusted R Square 0.688298647 = 68.8%

Based on the Number of Samples and Number of Independent Variable Adjusted R Square will varies. The Highest adjusted R Square results in better outcome.

2. Intercept value is 29.24131526.

	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

Significance of Each Independent Variable:

	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

1. The P value of Crime rate is greater than 0.05. So it is not significant.
2. AGE, INDUS, NOX, DISTANCE, TAX, PTRATIO ,AVG_ROOM and LSTAT are 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**.**

- Coefficients of AGE,INDUS,DISTANCE,AVG_ROOM is Positive.
- Coefficients of NOX,TAX,PTRATIO,LSTAT is Negative.

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?

1 Adjusted R Square 0.688298647 68.8%

2 Adjusted R Square 0.688683682 68.9%

Based on the Number of Samples and Number of Independent Variable Adjusted R Square will varies. The Highest adjusted R Square results in better outcome.

Comparing the previous model and the current model the adjusted R Square is slightly increasing .

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 of NOX is negative, Nox is inversely Proportional to AVG_PRICE. 2So If the NOX increases ,then the AVG_PRICE will decreases .

d) Write the regression equation from this model.

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

Regression Equation:

Y=(M1*X1+M2*X2+M3*X3+M4*X4+M5*X5+M6*X6+M7*X7+M8*X8)+B

Y=(0.0329349604286303*AGE(X1)+0.130710006682182*INDUS(X2)
+(-10.2727050815094)*NOX(X3)+0.261506423001819*DISTANCE(X4)+
(-0.0144523450364819)*TAX(X5)+(-1.07170247269449)*PTRATIO(X6)+
4.12546895908474*AVG_ROOM(X7)+(-0.605159282035406)*LSTAT(X8)+29.4284734939458