GRADED WEEK 2 PROJECT REPORT

TERRO'S REAL ESTATE AGENCY

BY

GOUTHAMEN.S

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

By using the Data analytics tool pack, we generate summary statistics for each variable in the table.

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

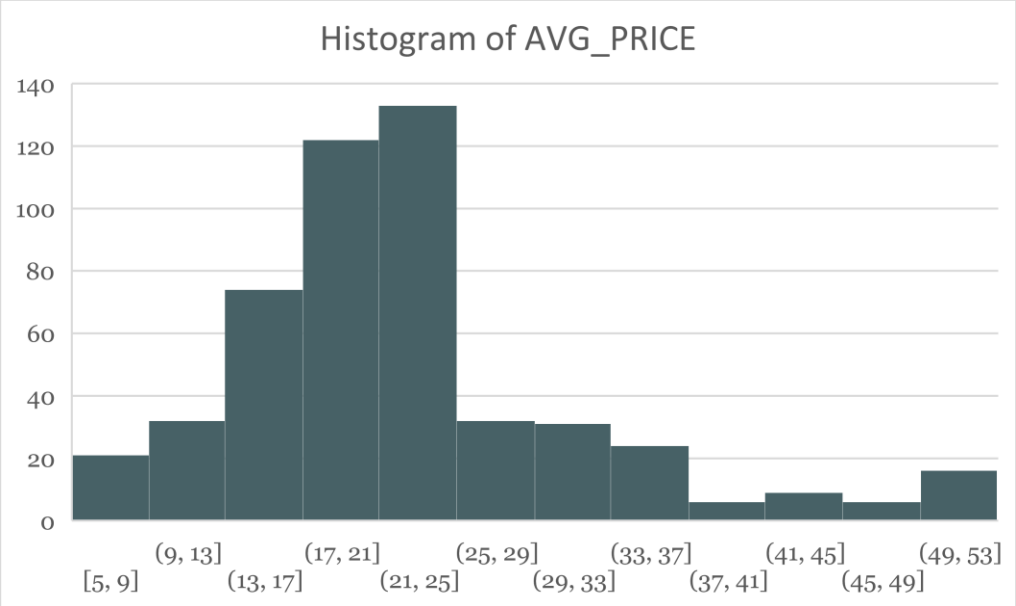
NOX		DISTANCE		TAX	
Mean	0.554695059	Mean	9.549407115	Mean	408.2371542
Standard Error	0.005151391	Standard Error	0.387084894	Standard Error	7.492388692
Median	0.538	Median	5	Median	330
Mode	0.538	Mode	24	Mode	666
Standard Deviation	0.115877676	Standard Deviation	8.707259384	Standard Deviation	168.5371161
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

From the descriptive statistics we can infer the number of observations, minimum, maximum and sum of the values. It computes the measure of central tendency like mean, median and mode. It computes the measure of dispersion like range, variance and standard deviation. It computes the measure of symmetery as skewness, which measure how tailed it is. It computes the measure of peakedness as kurtosis, which measure how peak it is. Positive kurtosis is called leptokurtic and negative kurtosis is called platykurtic.

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



We can plot histogram of Avg_price by navigating to insert The histogram chart of Average price tells the count of occurrence of the average price. The x-axis is average price and y-axis is the count of average price. The average prices are grouped into 12 groups from $5,000 to $53,000, each bin having range of $6,000. From the histogram chart of AVG_PRICE, we can infer that most of the average price are within the range of $21,000 to $25,000.

3) 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

By using the data analytics tool pack, we can compute covariance matrix. While looking at the covariance matrix, we can know about the relationship of the variables.
The covariance matrix tell about the relationship with positive or negative sign but not the magnitude. The AVG_PRICE has negative relation with AGE, INDUS, NOX, DISTANCE, TAX, PTRATIO and LSTAT. AVG_PRICE has positive relation with CRIME_RATE, AVG_ROOM, AVG_PRICE. Negative sign indicates when there is increase in one variable, the other variable decreases. Wherease positive sign indicates increase in one variable also increases other variable.

4) 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	1.0000									
AGE	0.0069	1.0000								
INDUS	-0.0055	0.6448	1.0000							
NOX	0.0019	0.7315	0.7637	1.0000						
DISTANCE	-0.0091	0.4560	0.5951	0.6114	1.0000					
TAX	-0.0167	0.5065	0.7208	0.6680	0.9102	1.0000				
PTRATIO	0.0108	0.2615	0.3832	0.1889	0.4647	0.4609	1.0000			
AVG_ROOM	0.0274	-0.2403	-0.3917	-0.3022	-0.2098	-0.2920	-0.3555	1.0000		
LSTAT	-0.0424	0.6023	0.6038	0.5909	0.4887	0.5440	0.3740	-0.6138	1.0000	
AVG_PRICE	0.0433	-0.3770	-0.4837	-0.4273	-0.3816	-0.4685	-0.5078	0.6954	-0.7377	1.0000

By using the data analytics pack, we can compute correlation matrix. It is between -1 to +1. Unlike covariance matrix, correlation matrix can also tell the magnitude. Positive correlation means increase in one variable also increases other variable, wherease negative correlation means increase in one variable decreases other variable and vice-versa. We can use conditional formatting to find the top 3 positively correlated pairs and top 3 negatively correlated pairs. The green shows top 3 positively correlated pairs and red shows top 3 negatively correlated pairs.

a) Which are the top 3 positively correlated pairs.

The top 3 postively correlated pairs are TAX and DISTANCE (0.9102), NOX and INDUS (0.7637), NOX and AGE (0.7315).

b) Which are the top 3 negatively correlated pairs.

The top 3 negatively correlated pairs are AVG_PRICE and LSTAT (-0.7377), LSTAT and AVG_ROOM (-0.6138), AVG_PRICE and PTRATIO (-0.5078).

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

By using data analytics pack, we can create a regression model with AVG_PRICE as Dependent variable in y-axis and LSTAT as Independent Variable in x-axis. We set the confidence level of 95%.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.737662726
R Square	0.544146298
Adjusted R Square	0.543241826
Standard Error	6.215760405
Observations	506

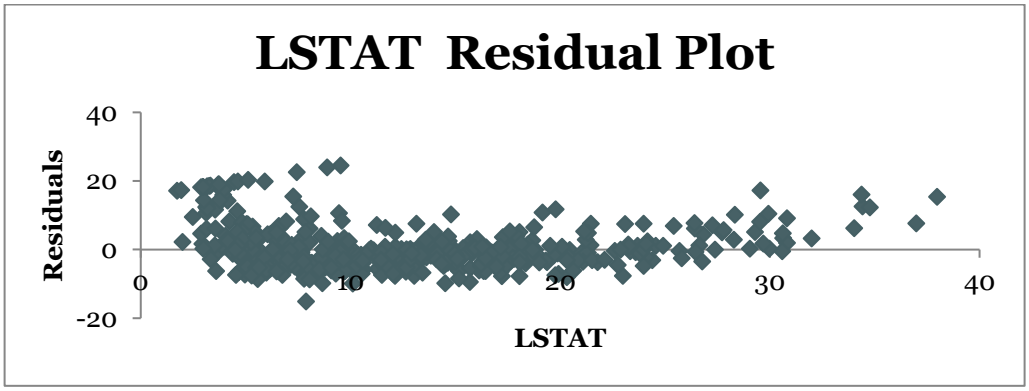
ANOVA

	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	35.6592247
LSTAT	-0.950049354	0.038733416	-24.52789985	5.0811E-88	-1.0261482	-0.873950508	-1.0261482	-0.87395050

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

In a single linear regression R^2 = (covar)^2, R^2 tells how much of dependent variable can be predicted from independent variable with a certain confidence level.



For this model R^2 = 54.41% where LSTAT can predict AVG_PRICE with 54.41% of accuracy with a confidence level of 95%

AVG_PRICE = (-0.95) * LSTAT + 34.5538
R^2 = 54.41%
Coefficient = -0.95
Intercept = 34.5538

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

As the p value is less than 0.05, we reject the null hypothesis.
LSTAT can predict AVG_PRICE 54.41% accurately with confidence level of 95%.
So LSTAT is statistically significant for this model.

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

By using data analytics pack, we can create a regression model with AVG_PRICE as Dependent variable in y-axis and LSTAT and AVG_ROOM as Independent Variable in x-axis. We set the confidence level of 95%. As the p value of intercept is greater than 0.05, we eliminate intercept by keeping the constant as 0.

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	4636.71208	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	1.613E-261	4.768998482	5.044813661	4.768998482	5.04481366
LSTAT	-0.655739993	0.030558561	-21.45847115	4.8115E-73	-0.715777847	-0.595702138	-0.715777847	-0.59570213

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?

Regression equation is AVG_PRICE = 4.9069*AVG_ROOM +(-0.6557) * LSTAT

AVG_ROOM = 7
LSTAT = 20

AVG_PRICE = 4.9069*7 + (-0.6557) * 20
= 21.23354265
As the values are in $1000s, we can multiply
= 21.23354265 * $1000
= $21,233

The predicted average price is $21,234
The company quoted average price is $30,000
Difference is $30,000-$21,234 = $8,766

The company is overcharging amount of $8,766 than the predicted price

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 for this model is 94.64%
Adjusted R-square for question 5 is 54.32%

This model has high adjusted R-square, so it has better performance than previous 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 Rsquare, coefficient and Intercept values. Explain the significance of each independent variable with respect to AVG_PRICE.

By using data analytics pack, we can create a regression model with AVG_PRICE as Dependent variable in y-axis and other variables as Independent Variable in x-axis. We set the confidence level of 95%.

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	29.241	4.817	6.070	0.000	19.777	38.706	19.777	38.706
CRIME_RATE	0.049	0.078	0.621	0.535	-0.105	0.203	-0.105	0.203
AGE	0.033	0.013	2.502	0.013	0.007	0.059	0.007	0.059
INDUS	0.131	0.063	2.068	0.039	0.007	0.255	0.007	0.255
NOX	-10.321	3.894	-2.651	0.008	-17.972	-2.670	-17.972	-2.670
DISTANCE	0.261	0.068	3.843	0.000	0.128	0.395	0.128	0.395
TAX	-0.014	0.004	-3.688	0.000	-0.022	-0.007	-0.022	-0.007
PTRATIO	-1.074	0.134	-8.041	0.000	-1.337	-0.812	-1.337	-0.812
AVG_ROOM	4.125	0.443	9.318	0.000	3.255	4.995	3.255	4.995
LSTAT	-0.603	0.053	-11.369	0.000	-0.708	-0.499	-0.708	-0.499

Adjusted R square is 68.83%
Intercept is 29.24

Coefficients of

AGE = 0.033
INDUS = 0.131
NOX = -10.321
DISTANCE = 0.261
TAX = -0.014
PTRATIO = -1.074
AVG_ROOM = 4.125
LSTAT = -0.603

As the p value of crime rate is greater than 0.05, it fails to reject null hypothesis.
As the p value of other variables is less than 0.05, it rejects null hypothesis.

CRIME_RATE is statistically insignificant with respect to AVG_PRICE.

AGE, INDUS, NOX, DISTANCE, TAX, PTRATIO, AVG_ROOM, LSTAT are statistically significant with respect to AVG_PRICE.

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:

By using data analytics pack, we can create a regression model with AVG_PRICE as Dependent variable in y-axis and other significant variables as Independent Variable in x-axis. We set the confidence level of 95%.

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	-10.27270508	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.

The R-square is 69.36%
The adjusted R-Square is 68.86%.

The Intercept value is 29.4284.

Coefficients of

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

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 of this model is 68.87%
Adjusted R Square of previous question is 68.83%

So, this model performs better as it has higher 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?

Coefficient of NOX = -10.27270508

As the coefficient of NOX is negative, if the value of NOX is more, then Average price will decrease.

d) Write the regression equation from this model.

Regression equation for this model is AVG_PRICE=-10.27*NOX +(-1.07) *PTRATIO + (-0.61)*LSTAT + (-0.01)*TAX + 0.03*AGE + 0.13*INDUS + 0.26*DISTANCE + 4.13*AVG_ROOM + 29.4284