TERROS REAL ESTATE

DATA ANALYSIS

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

AVG_PRICE	
Mean	22.53281
Standard Error	0.408861
Median	21.2
Mode	50
Standard Deviation	9.197104
Sample Variance	84.58672
Kurtosis	1.495197
Skewness	1.108098
Range	45
Minimum	5
Maximum	50
Sum	11401.6
Count	506

CRIME_RATE	
Mean	4.871976
Standard Error	0.12986
Median	4.82
Mode	3.43
Standard Deviation	2.921132
Sample Variance	8.533012
Kurtosis	-1.18912
Skewness	0.021728
Range	9.95
Minimum	0.04
Maximum	9.99
Sum	2465.22
Count	506

PTRATIO	
Mean	18.45553
Standard Error	0.096244
Median	19.05
Mode	20.2
Standard Deviation	2.164946
Sample Variance	4.686989
Kurtosis	-0.28509
Skewness	-0.80232
Range	9.4
Minimum	12.6
Maximum	22
Sum	9338.5
Count	506

LSTAT	
Mean	12.65306
Standard Error	0.317459
Median	11.36
Mode	8.05
Standard Deviation	7.141062
Sample Variance	50.99476
Kurtosis	0.49324
Skewness	0.90646
Range	36.24
Minimum	1.73
Maximum	37.97
Sum	6402.45
Count	506

AVG_ROOM	
Mean	6.284634
Standard Error	0.031235
Median	6.2085
Mode	5.713
Standard Deviation	0.702617
Sample Variance	0.493671
Kurtosis	1.8915
Skewness	0.403612
Range	5.219
Minimum	3.561
Maximum	8.78
Sum	3180.025
Count	506

TAX	
Mean	408.2372
Standard Error	7.492389
Median	330
Mode	666
Standard Deviation	168.5371
Sample Variance	28404.76
Kurtosis	-1.14241
Skewness	0.669956
Range	524
Minimum	187
Maximum	711
Sum	206568
Count	506

DISTANCE	
Mean	9.549407
Standard Error	0.387085
Median	5
Mode	24
Standard Deviation	8.707259
Sample Variance	75.81637
Kurtosis	-0.86723
Skewness	1.004815
Range	23
Minimum	1
Maximum	24
Sum	4832
Count	506

NOX	
Mean	0.554695
Standard Error	0.005151
Median	0.538
Mode	0.538
Standard Deviation	0.115878
Sample Variance	0.013428
Kurtosis	-0.06467
Skewness	0.729308
Range	0.486
Minimum	0.385
Maximum	0.871
Sum	280.6757
Count	506

INDUS	
Mean	11.13678
Standard Error	0.30498
Median	9.69
Mode	18.1
Standard Deviation	6.860353
Sample Variance	47.06444
Kurtosis	-1.23354
Skewness	0.295022
Range	27.28
Minimum	0.46
Maximum	27.74
Sum	5635.21
Count	506

AGE	
Mean	68.5749
Standard Error	1.25137
Median	77.5
Mode	100
Standard Deviation	28.14886
Sample Variance	792.3584
Kurtosis	-0.96772
Skewness	-0.59896
Range	97.1
Minimum	2.9
Maximum	100
Sum	34698.9
Count	506

INFERENCE

#1. **Crime Rate could be normally distributed** the mean is **4.87** and median is **4.82** and skewness is **0.02**

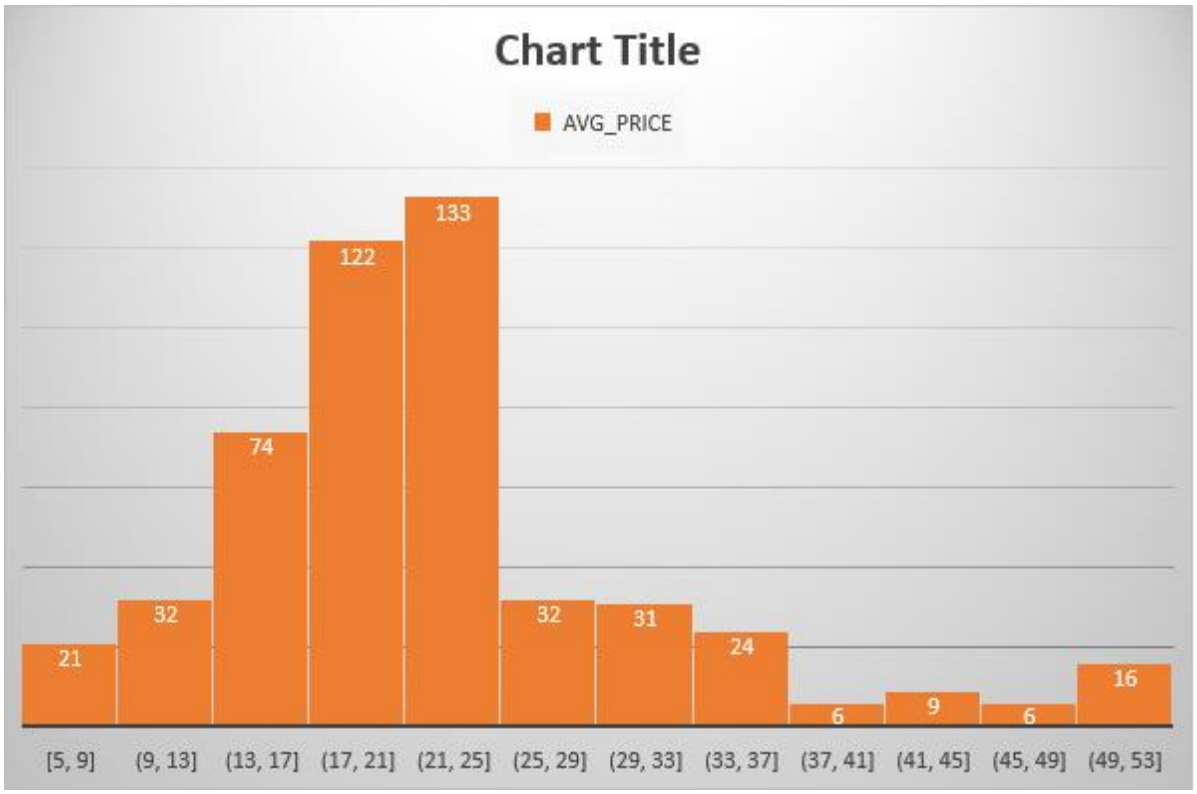
#2.**Average price** of the owned houses **22.5** with positive skewnees that is much of its prices are towards the right and the range is 45 with min 5, max45

#3. The **mean** of **AVG ROOM** is **6.28** and **median** is **6.20** which shows that it could be normally distributed

#4.The **Tax** varies from the range of **187 to 711** , the **most common tax** rate being **666**

#5.The **range** of age varies from **2.9 to 100**,with the **average** being **68.5**

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



INFERENCE

This histogram Indicates that AvgPrice is positively skewed and the values are tail towards the right.
Around 50 % of the houses are priced in the range of 17.1 to 24.9(17100 to 24900 dollars).

3.Compute the covariance matrix. Share your observations.

	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RATE	8.52									
AGE	0.56	790.79								
INDUS	-0.11	124.27	46.97							
NOX	0.00	2.38	0.61	0.01						
DISTANCE	-0.23	111.55	35.48	0.62	75.67					
TAX	-8.23	2397.94	831.71	13.02	1333.12	28348.62				
PTRATIO	0.07	15.91	5.68	0.05	8.74	167.82	4.68			
AVG_ROOM	0.06	-4.74	-1.88	-0.02	-1.28	-34.52	-0.54	0.49		
LSTAT	-0.88	120.84	29.52	0.49	30.33	653.42	5.77	-3.07	50.89	
AVG_PRICE	1.16	-97.40	-30.46	-0.45	-30.50	-724.82	-10.09	4.48	-48.35	84.42

Considering Independent variable as AVG_PRICE , CRIME RATE and AVG_ROOM are positively related WHITE CELLS represents negatively related variables and yellow cells represents positively related cells

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

a) Which are the top 3 positively correlated pairs

b) Which are the top 3 negatively correlated pairs.

	CRIME_RATE	*AGE*	*INDUS*	*NOX*	*DISTANCE*	*TAX*	*PTRATIO*	*AVG_ROOM*	*LSTAT*	*AVG_PRICE*
CRIME_RATE	0									
AGE	0.007	0								
INDUS	-0.006	0.645	0							
NOX	0.002	0.731	0.764	0						
DISTANCE	-0.009	0.456	0.595	0.611	0					
TAX	-0.017	0.506	0.721	0.668	0.910	0				
PTRATIO	0.011	0.262	0.383	0.189	0.465	0.461	0			
AVG_ROOM	0.027	-0.240	-0.392	-0.302	-0.210	-0.292	-0.356	0		
LSTAT	-0.042	0.602	0.604	0.591	0.489	0.544	0.374	-0.614	0	
AVG_PRICE	0.043	-0.377	-0.484	-0.427	-0.382	-0.469	-0.508	0.695	-0.738	0

Top 3 Positively correlated variables is indiacted in red coloured cells, They are as follows

1.NOX and AGE = 0.731

2.NOX and INDUS =0.764

3.TAX and DISTANCE = 0.910

Top 3 Negatively correlated variables is indicated green coloured cells, They are as follows

1. AVG PRICE and PTRATIO = -0.508

2.LSTAT and AVG ROOM = -0.614

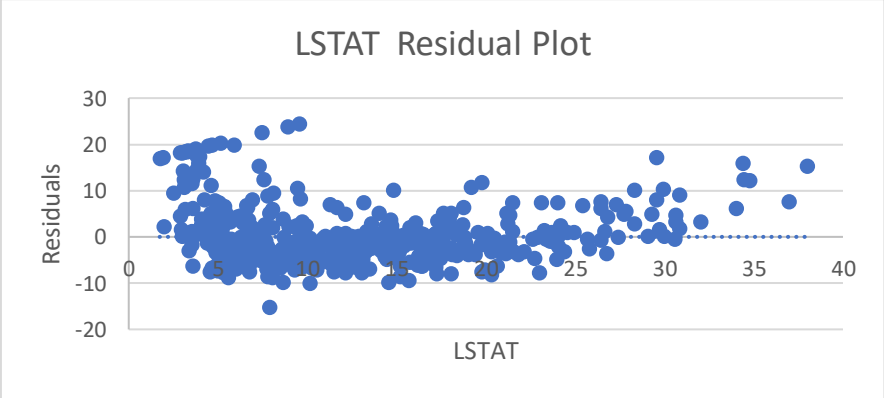
3.LSTAT and AVG PRICE = -0.738

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

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.737663
R Square	0.544146
Adjusted R Square	0.543242
Standard Error	6.21576
Observations	506



ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	23243.91	23243.91	601.6179	5.08E-88
Residual	504	19472.38	38.63568		
Total	505	42716.3			

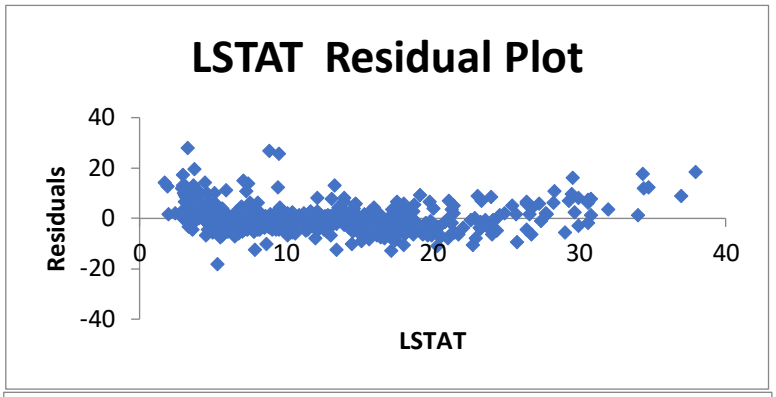
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	34.55384	0.562627	61.41515	3.7E-236	33.44846	35.65922	33.44846	35.65922
LSTAT	-0.95005	0.038733	-24.5279	5.08E-88	-1.02615	-0.87395	-1.02615	-0.87395

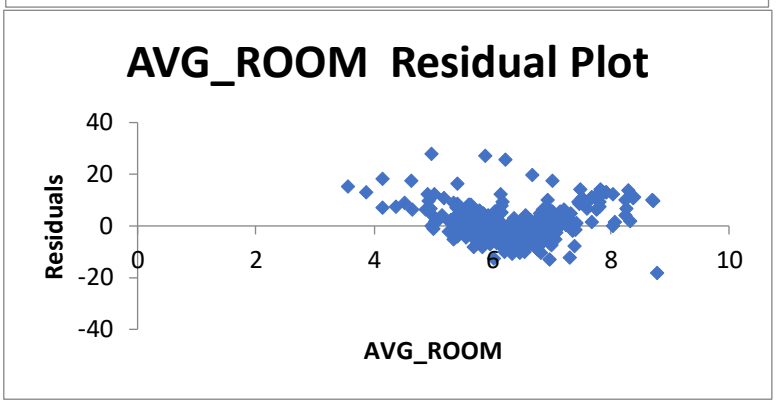
INFERENCE

The built model has a **R SQUARE value of 0.544** and **P value is 5.08E-88**
This indicates that this model can predict or explain 54.4% of the variance in the AVG PRICE
By Looking at the Residual plot we can say that there might be a pattern or yet we have to build a model by including other Variables as well
The Intercept is 34.55, which indiactes that Even if AVG PRICE is 0 , Value of LSAT is 34.55
The Equation is AVG PRICE = **34.553 - 0.95(LSTAT)**

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	0.7991
R Square	0.638562
Adjusted R Square	0.637124
Standard Error	5.540257
Observations	506

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	27276.99	13638.49	444.3309	7E-112
Residual	503	15439.31	30.69445		
Total	505	42716.3			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-1.35827	3.172828	-0.4281	0.668765	-7.5919	4.875355	-7.5919	4.875355
AVG_ROOM	5.094788	0.444466	11.46273	3.47E-27	4.22155	5.968026	4.22155	5.968026
LSTAT	-0.64236	0.043731	-14.6887	6.67E-41	-0.72828	-0.55644	-0.72828	-0.55644

6A. The Equation is **AVG PRICE = -1.36 + 5.1*AVG ROOM - 0.64*LSTAT**
AVG ROOM = 7
LSTAT = 20
Substituting in the above eqaution
AVG PRICE = -1.36 + 5.1*7-0.64*20
AVG PRICE = 21.54
The company is quoting the value of **30** , where the value supposed to be **21.54**
Hence The company is overcharging

6B. The **R squared** value of **this model is 63.8** and for **previous model** it was **54.4**. hence this model can explain the variance of AVG PRIVE which is nearly **10%(9.4%) is better than previous one**

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	0.832979
R Square	0.693854
Adjusted R Square	0.688299
Standard Error	5.134764
Observations	506

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	9	29638.8605	3293.2067	124.9045	1.933E-121
Residual	496	13077.4349	26.365796		
Total	505	42716.2954			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	29.24132	4.8171256	6.0702829	2.54E-09	19.7768278	38.705803	19.7768278	38.705803
CRIME_RATE	0.048725	0.07841865	0.6213464	0.5346572	-0.1053485	0.2027988	-0.1053485	0.2027988
AGE	0.032771	0.01309781	2.5019968	0.0126704	0.00703665	0.0585047	0.00703665	0.0585047
INDUS	0.130551	0.06311733	2.0683922	0.0391209	0.00654109	0.2545617	0.00654109	0.2545617
NOX	-10.3212	3.89403626	-2.6505102	0.0082939	-17.972023	-2.670343	-17.972023	-2.670343
DISTANCE	0.261094	0.06794707	3.8426026	0.0001375	0.12759401	0.3945931	0.12759401	0.3945931
TAX	-0.0144	0.00390516	-3.6877361	0.0002512	-0.0220739	-0.006728	-0.0220739	-0.006728
PTRATIO	-1.07431	0.13360172	-8.0411041	6.586E-15	-1.3368004	-0.81181	-1.3368004	-0.81181
AVG_ROOM	4.125409	0.442759	9.3175049	3.893E-19	3.25549474	4.9953236	3.25549474	4.9953236
LSTAT	-0.60349	0.05308116	-11.369129	8.911E-27	-0.7077782	-0.499195	-0.7077782	-0.499195

INFERENCE

This model has a R sqaured value of .69 which is higher than the previous model whose r sqaured value is 0.64 hence this model explain 5% of variance better than the previous model

The intercept value is 29.24 ,which indicates that even if all the variable are zero , The AVG PRICE Would be 29.24

By looking at the P-Values only CRIME RATE is insignificant which is higher than 0.05 , and indicated by grey cell in the table

Adjusted R Square value is 0.688, which indicates it explains 68.8 % variance of the dependent variable

By looking at coefficients NOX,TAX,LSTAT are negatively correlated

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	0.832836
R Square	0.693615
Adjusted R Square	0.688684
Standard Error	5.131591
Observations	506

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	8	29628.68	3703.585	140.643	1.9E-122
Residual	497	13087.61	26.33323		
Total	505	42716.3			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	29.42847	4.804729	6.124898	1.85E-09	19.98839	38.86856	19.98839	38.86856
NOX	-10.2727	0.013087	2.516606	0.012163	0.007222	0.058648	0.007222	0.058648
PTRATIO	-1.0717	0.063078	2.072202	0.038762	0.006778	0.254642	0.006778	0.254642
LSTAT	-0.60516	3.890849	-2.64022	0.008546	-17.9172	-2.62816	-17.9172	-2.62816
TAX	-0.01445	0.067902	3.851242	0.000133	0.128096	0.394916	0.128096	0.394916
AGE	0.032935	0.003902	-3.70395	0.000236	-0.02212	-0.00679	-0.02212	-0.00679
INDUS	0.13071	0.133454	-8.03053	7.08E-15	-1.33391	-0.8095	-1.33391	-0.8095
DISTANCE	0.261506	0.442485	9.3234	3.69E-19	3.256096	4.994842	3.256096	4.994842
AVG_ROOM	4.125469	0.05298	-11.4224	5.42E-27	-0.70925	-0.50107	-0.70925	-0.50107

A. The R square Value of this model is 0.6936 which explains 69.36% of variance of AVG PRICE
The INTERCEPT value is 29.42 which indictes even if all the variables are zero the AVG PRICE would be 29.42

B. Adjusted R Sqare of previous model is 68.83
Adjusted R square of this model is 68.87
There is no significant change in R square value but we have all significant varaibles here which shows this is the better model

C. The Sorted cells are in Gray colour
NOX is negatively related to AVG price that is when nox increases AVG price decreases
for every one unit increase in NOX , The AVG PRICE decreases by 10.27 units

D. The Equation is as follows

AVG PRICE = 29.42 - (10.27*NOX) - (1.07*PTRATIO) - (0.6*LSTAT) - (0.01*TAX) + (0.033*AGE) + (0.13*INDUS) + (0.26*DISTANCE) + (4.12*AVG_ROOM)