PROBLEM STATEMENT:

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:

To analyse 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. **(5 marks)**

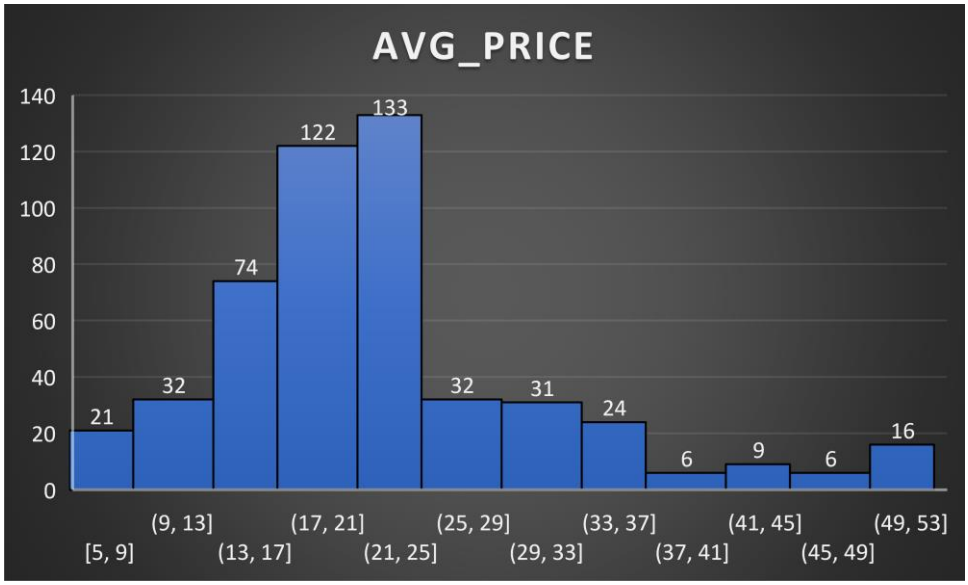
CRIME_RATE		*AGE*		*INDUS*		*NOX*		*DISTANCE*	
Mean	4.87197628	Mean	68.5749	Mean	11.1368	Mean	0.5547	Mean	9.54941
Standard Error	0.12986015	Standard Error	1.25137	Standard Error	0.30498	Standard Error	0.00515	Standard Error	0.38708
Median	4.82	Median	77.5	Median	9.69	Median	0.538	Median	5
Mode	3.43	Mode	100	Mode	18.1	Mode	0.538	Mode	24
Standard Deviation	2.92113189	Standard Deviation	28.1489	Standard Deviation	6.86035	Standard Deviation	0.11588	Standard Deviation	8.70726
Sample Variance	8.53301153	Sample Variance	792.358	Sample Variance	47.0644	Sample Variance	0.01343	Sample Variance	75.8164
Kurtosis	-1.1891225	Kurtosis	-0.96772	Kurtosis	-1.23354	Kurtosis	-0.06467	Kurtosis	-0.86723
Skewness	0.02172808	Skewness	-0.59896	Skewness	0.29502	Skewness	0.72931	Skewness	1.00481
Range	9.95	Range	97.1	Range	27.28	Range	0.486	Range	23
Minimum	0.04	Minimum	2.9	Minimum	0.46	Minimum	0.385	Minimum	1
Maximum	9.99	Maximum	100	Maximum	27.74	Maximum	0.871	Maximum	24
Sum	2465.22	Sum	34698.9	Sum	5635.21	Sum	280.676	Sum	4832
Count	506	Count	506	Count	506	Count	506	Count	506

TAX		*PTRATIO*		*AVG_ROOM*		*LSTAT*		*AVG_PRICE*	
Mean	408.237	Mean	18.4555	Mean	6.28463	Mean	12.6531	Mean	22.5328
Standard Error	7.49239	Standard Error	0.09624	Standard Error	0.03124	Standard Error	0.31746	Standard Error	0.40886
Median	330	Median	19.05	Median	6.2085	Median	11.36	Median	21.2
Mode	666	Mode	20.2	Mode	5.713	Mode	8.05	Mode	50
Standard Deviation	168.537	Standard Deviation	2.16495	Standard Deviation	0.70262	Standard Deviation	7.14106	Standard Deviation	9.1971
Sample Variance	28404.8	Sample Variance	4.68699	Sample Variance	0.49367	Sample Variance	50.9948	Sample Variance	84.5867
Kurtosis	-1.14241	Kurtosis	-0.28509	Kurtosis	1.8915	Kurtosis	0.49324	Kurtosis	1.4952
Skewness	0.66996	Skewness	-0.80232	Skewness	0.40361	Skewness	0.90646	Skewness	1.1081
Range	524	Range	9.4	Range	5.219	Range	36.24	Range	45
Minimum	187	Minimum	12.6	Minimum	3.561	Minimum	1.73	Minimum	5
Maximum	711	Maximum	22	Maximum	8.78	Maximum	37.97	Maximum	50
Sum	206568	Sum	9338.5	Sum	3180.03	Sum	6402.45	Sum	11401.6
Count	506	Count	506	Count	506	Count	506	Count	506

Attribute	Skewness
CRIME_RATE	Positive
AGE	Negative
INDUS	Positive
NOX	Positive
DISTANCE	Positive
TAX	Positive
PTRATIO	Negative
AVG_ROOM	Positive
LSTAT	Positive
AVG_PRICE	Positive

1. Looking into the data we can see the summary of the data set
2. The mean and median is being highlighted
3. The Skewness of each of the attributes are mentioned
4. We can see the presence of outliers
5. From just the summary we can't infer much, so we are going deep into each attribute with different methods

2) Plot a histogram of the Avg_Price variable. What do you infer? **(5 marks)**



1. The greatest number of houses (133) are sold in the price range of 21k to 25k.
2. The least number of houses (6) are sold in the price range of 37k to 41k and 45k to 49k.

3) Compute the covariance matrix. Share your observations. **(5 marks)**

	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RATE	8.516147873									
AGE	0.562915215	790.7924728								
INDUS	-0.110215175	124.2678282	46.97142974							
NOX	0.000625308	2.381211931	0.605873943	0.013401099						
DISTANCE	-0.229860488	111.5499555	35.47971449	0.615710224	75.66653127					
TAX	-8.229322439	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	-1.884225427	-0.024554826	-1.281277391	-34.51510104	-0.539694518	0.492695216		
LSTAT	-0.882680362	120.8384405	29.52181125	0.487979871	30.32539213	653.4206174	5.771300243	-3.073654967	50.89397935	
AVG_PRICE	1.16201224	-97.39615288	-30.46050499	-0.454512407	-30.50083035	-724.8204284	-10.09067561	4.484565552	-48.35179219	84.4195562

1. Higher covariance tells us they are highly correlated.
2. We can see that TAX has high variance value.

4) Create a correlation matrix of all the variables (Use Data analysis tool pack). **(5 marks)**

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	0.6447785	1							
NOX	0.001850982	0.7314701	0.76365	1						
DISTANCE	-0.009055049	0.4560225	0.59513	0.61144	1					
TAX	-0.016748522	0.5064556	0.72076	0.66802	0.91023	1				
PTRATIO	0.010800586	0.261515	0.38325	0.18893	0.46474	0.46085	1			
AVG_ROOM	0.02739616	-0.240265	-0.39168	-0.30219	-0.20985	-0.29205	-0.35550149	1		
LSTAT	-0.042398321	0.6023385	0.6038	0.59088	0.48868	0.54399	0.374044317	-0.61380827	1	
AVG_PRICE	0.043337871	-0.376955	-0.48373	-0.42732	-0.38163	-0.46854	-0.50778669	0.695359947	-0.73766273	1

POSITIVE

Sl.No	Attributes	Value
1	TAX vs DISTANCE	0.9102282
2	INDUX vs NOX	0.7636514
3	AGE vs NOX	0.7314701

NEGATIVE

Sl.No	Attributes	Value
1	LSAT vs AVG_PRICE	-0.737663
2	AVG_ROOM vs LSTAT	-0.613808
3	PTRATIO vs AVG_PRICE	-0.507787

5) Build an initial regression model with AVG_PRICE as **'y'** (Dependent variable) and LSTAT variable as Independent Variable. Generate the residual plot. **(8 marks)**

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?

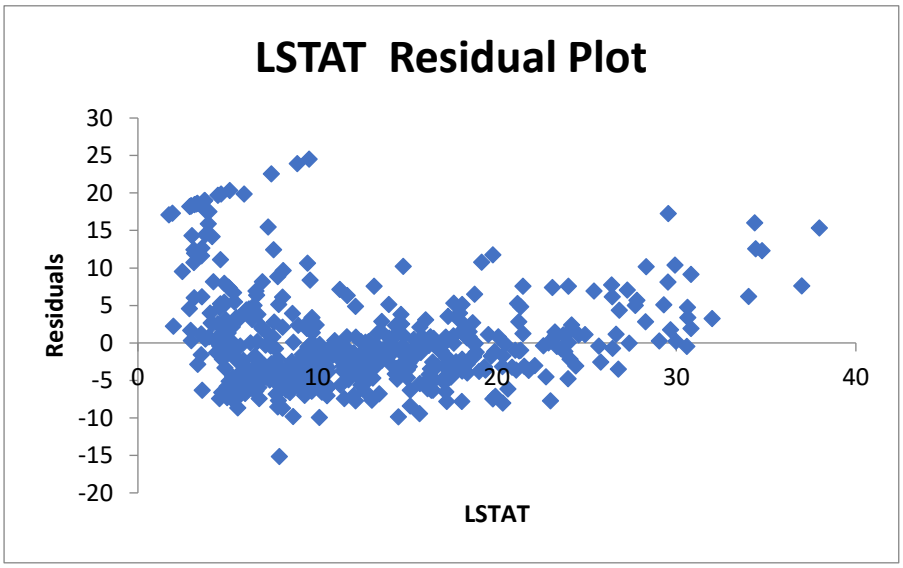
SUMMARY OUTPUT

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	601.617871	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	3.743E-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



1. R square value is highlighted showing that variance is 54%
2. Intercept value is highlighted, showing that, when LSTAT is 0 the AVG_PRICE will be 34.55384
3. The significance value is highlighted and is above 0.05, therefore LSTAT is significant

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

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.

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	444.330892	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	-0.42809535	0.66876494	-7.59190028	4.875354658	-7.59190028	4.875354658
AVG_ROOM	5.094787984	0.4444655	11.46272991	3.4723E-27	4.221550436	5.968025533	4.221550436	5.968025533
LSTAT	-0.642358334	0.043731465	-14.6886992	6.6694E-41	-0.72827717	-0.5564395	-0.72827717	-0.5564395

AVG_ROOMS (x1) =7

L-STAT(x2) =20

y= dependent variable (AVG_PRICE)

x1= Independent variable (AVG_ROOM)

x2= Independent variables (LSTAT)

Regression Equation:

y = a+bx1 + bx2

y= -1.358 +5.09 (x1) -0.642 (x2)

y= -1.358 + 5.09(7) -0.642(20) = 21.44

The price for the new house is $ 21440.

Comparing to the company quoting a value of 3000 USD is clearly overcharging.

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. **(8 marks)**

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

1. R square value is 0.6938 therefore the variance is 69%
2. Checking at the significance we can the that CRIME_RATE is not a significant variable as the p-value is greater that 0.05 all other variables are significant
3. NOX, TAX, PTRATIO, LSTAT have a negative coefficient, which means they are inversely proportional.

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: **(8 marks)**

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.

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%*
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

SUMMARY OUTPUT

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

SUMMARY OUTPUT

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

	Coefficients	*Standard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
NOX	-10.27270508	3.890849222	-2.640221837	0.00854572	-17.9172457	-2.628164466	-17.9172457	-2.628164466
PTRATIO	-1.071702473	0.133453529	-8.030529271	7.0825E-15	-1.333905109	-0.809499836	-1.333905109	-0.809499836
LSTAT	-0.605159282	0.0529801	-11.42238841	5.4184E-27	-0.70925186	-0.501066704	-0.70925186	-0.501066704
TAX	-0.014452345	0.003901877	-3.703946406	0.00023607	-0.022118553	-0.006786137	-0.022118553	-0.006786137
AGE	0.03293496	0.013087055	2.516605952	0.01216288	0.007222187	0.058647734	0.007222187	0.058647734
INDUS	0.130710007	0.063077823	2.072202264	0.03876167	0.006777942	0.254642071	0.006777942	0.254642071
DISTANCE	0.261506423	0.067901841	3.851242024	0.00013289	0.128096375	0.394916471	0.128096375	0.394916471
AVG_ROOM	4.125468959	0.44248544	9.323400461	3.6897E-19	3.256096304	4.994841615	3.256096304	4.994841615
Intercept	29.42847349	4.804728624	6.124898157	1.846E-09	19.98838959	38.8685574	19.98838959	38.8685574

1. Looking at the significance all the variable are significant as their p-value is less than 0.05
2. Both the models have similar Multiple R value and R square value, both of them perform well

Regression equation:

Y=0.03293496 x0 + 0.130710007 x1 -10.27270508 x3 +0.261506423 x4 - 0.014452345 x5 -1.071702473 x6 + 4.125468959 x7 -0.605159282 x8 +29.42847349