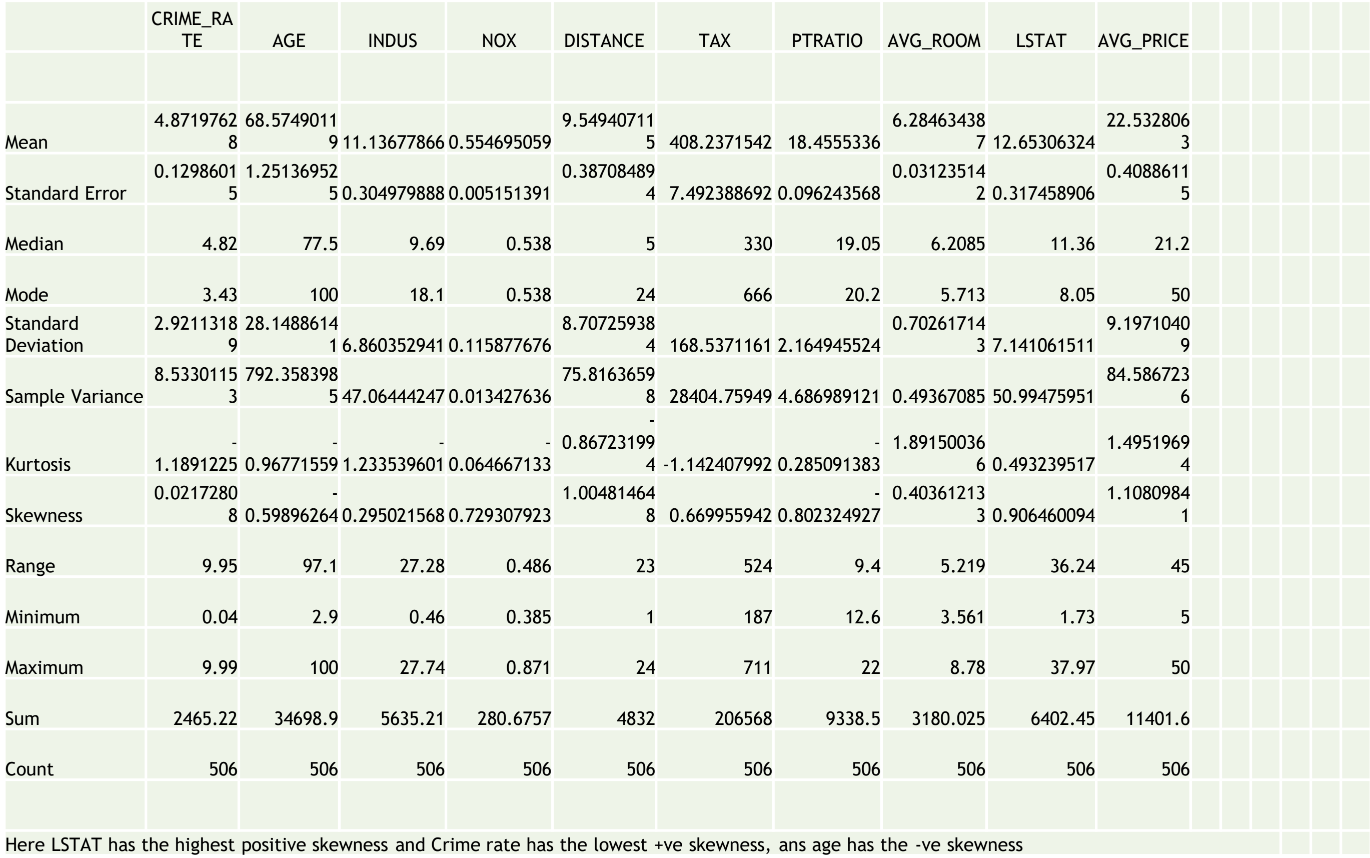
TERRO-REAL-ESTATE REPORT

~MOHAMMED SHAMIS KOLA

SO HERE WE HAD THE DATA GIVEN

- CRIME RATE: per capita crime rate by town
- INDUSTRY: the proportion of non-retail business acres per town (in percentage terms)
- NOX: nitric oxides concentration (parts per 10 million)
- AVG ROOM: average number of rooms per house
- AGE: the proportion of houses built prior to 1940 (in percentage terms)
- DISTANCE: distance from highway (in miles)
- TAX: full-value property-tax rate per $10,000
- PTRATIO: pupil-teacher ratio by town
- LSTAT:% lower status of the population
- AVG_PRICE: Average value of houses in $1000's

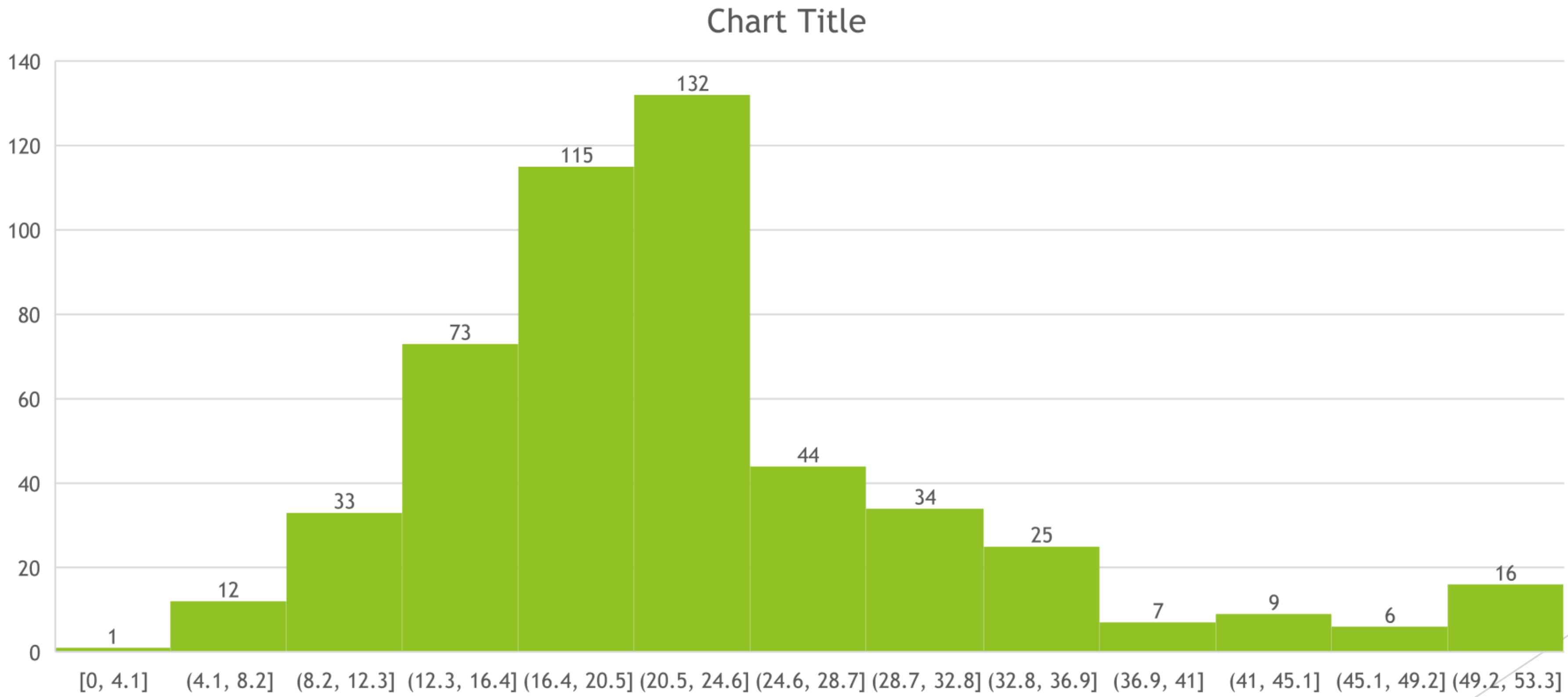
1) Generate the summary statistics for each variable in the table. (Use Data analysis tool pack). Write down your observation.



	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
Mean	4.87197628	68.57490119	11.13677866	0.554695059	9.549407115	408.2371542	18.4555336	6.284634387	12.65306324	22.5328063
Standard Error	0.12986015	1.251369525	0.304979888	0.005151391	0.387084894	7.492388692	0.096243568	0.031235142	0.317458906	0.40886115
Median	4.82	77.5	9.69	0.538	5	330	19.05	6.2085	11.36	21.2
Mode	3.43	100	18.1	0.538	24	666	20.2	5.713	8.05	50
Standard Deviation	2.92113189	28.14886141	6.860352941	0.115877676	8.707259384	168.5371161	2.164945524	0.702617143	7.141061511	9.19710409
Sample Variance	8.53301153	792.3583985	47.06444247	0.013427636	75.81636598	28404.75949	4.686989121	0.49367085	50.99475951	84.5867236
Kurtosis	-1.1891225	-0.96771559	-1.233539601	-0.064667133	-0.867231994	-1.142407992	-0.285091383	1.891500366	0.493239517	1.49519694
Skewness	0.02172808	-0.59896264	0.295021568	0.729307923	1.004814648	0.669955942	-0.802324927	0.403612133	0.906460094	1.10809841
Range	9.95	97.1	27.28	0.486	23	524	9.4	5.219	36.24	45
Minimum	0.04	2.9	0.46	0.385	1	187	12.6	3.561	1.73	5
Maximum	9.99	100	27.74	0.871	24	711	22	8.78	37.97	50
Sum	2465.22	34698.9	5635.21	280.6757	4832	206568	9338.5	3180.025	6402.45	11401.6
Count	506	506	506	506	506	506	506	506	506	506

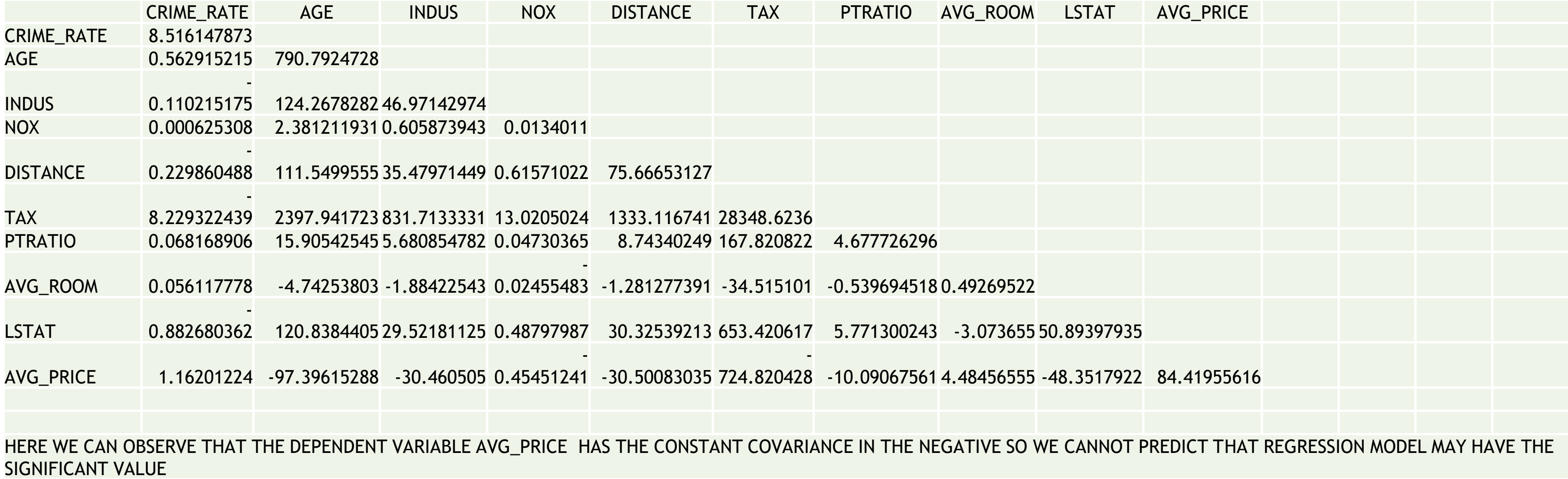
Here LSTAT has the highest positive skewness and Crime rate has the lowest +ve skewness, ans age has the -ve skewness

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



Here we can observe that there 132 persons which are having the average b/w the range 20.5-24.6

3. Compute the covariance matrix. Share your observations.

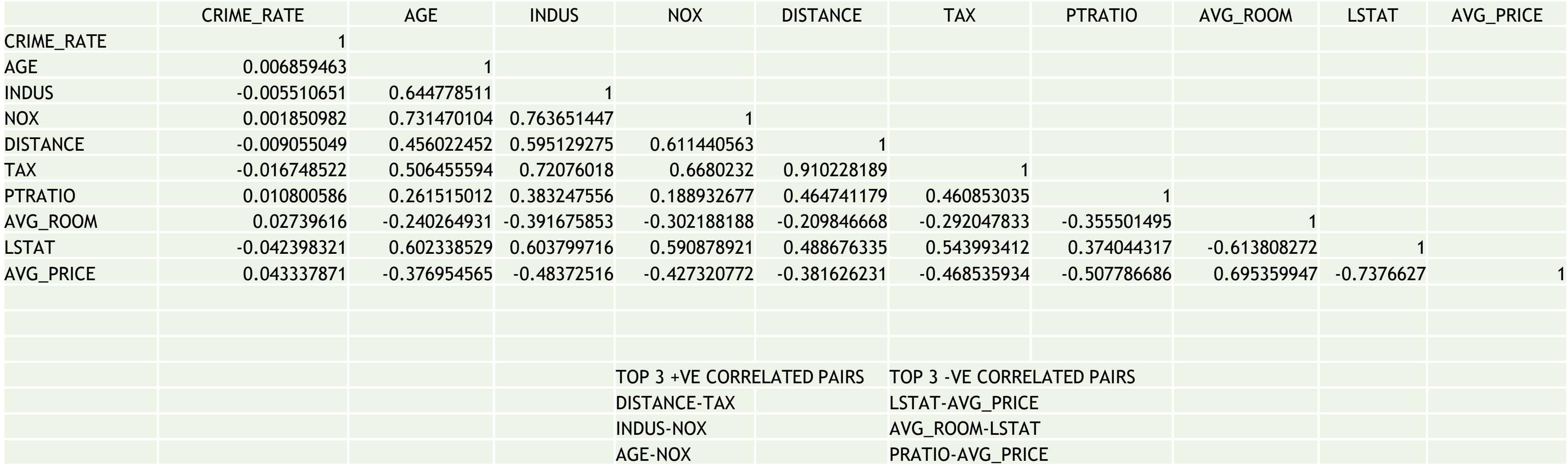


	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RATE	8.516147873									
AGE	0.562915215	790.7924728								
INDUS	-0.110215175	124.2678282	46.97142974							
NOX	0.000625308	2.381211931	0.605873943	0.0134011						
DISTANCE	-0.229860488	111.5499555	35.47971449	0.61571022	75.66653127					
TAX	-8.229322439	2397.941723	831.7133331	13.0205024	1333.116741	28348.6236				
PTRATIO	0.068168906	15.90542545	5.680854782	0.04730365	8.74340249	167.820822	4.677726296			
AVG_ROOM	0.056117778	-4.74253803	-1.88422543	-0.02455483	-1.281277391	-34.515101	-0.539694518	0.49269522		
LSTAT	-0.882680362	120.8384405	29.52181125	0.48797987	30.32539213	653.420617	5.771300243	-3.073655	50.89397935	
AVG_PRICE	1.16201224	-97.39615288	-30.460505	-0.45451241	-30.50083035	-724.820428	-10.09067561	4.48456555	-48.3517922	84.41955616

HERE WE CAN OBSERVE THAT THE DEPENDENT VARIABLE AVG_PRICE HAS THE CONSTANT COVARIANCE IN THE NEGATIVE SO WE CANNOT PREDICT THAT REGRESSION MODEL MAY HAVE THE SIGNIFICANT VALUE

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

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	0.644778511	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	-0.7376627	1
				TOP 3 +VE CORRELATED PAIRS		TOP 3 -VE CORRELATED PAIRS				
				DISTANCE-TAX		LSTAT-AVG_PRICE				
				INDUS-NOX		AVG_ROOM-LSTAT				
				AGE-NOX		PRATIO-AVG_PRICE				

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	35.65922472	33.448457	35.65922
LSTAT	-0.95005	0.038733	-24.5279	5.08E-88	-1.02615	-0.873950508	-1.0261482	-0.87395

RESIDUAL PLOT WE CAN SAY THAT THERE IS NO CONSTANT VARIATION

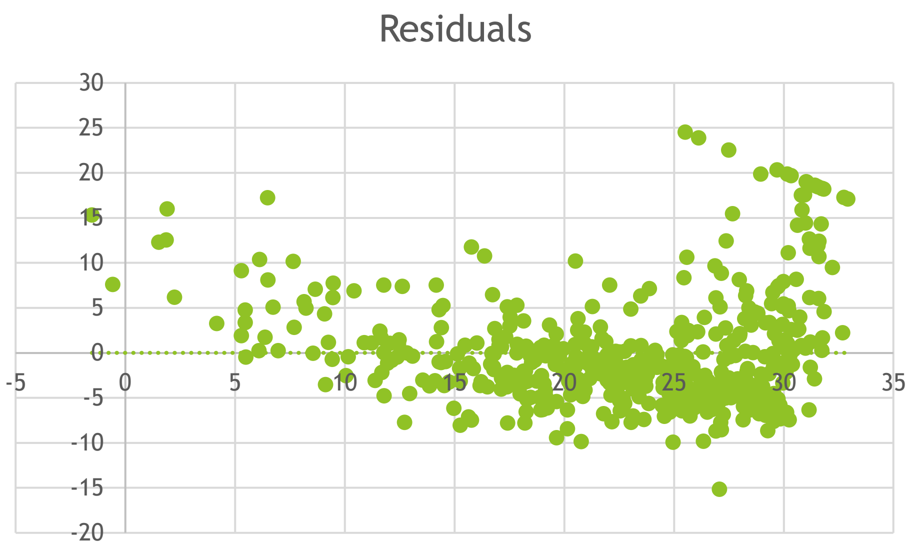
LSTAT HAS THE P-VALUE LESS THAN 0.05 SO WE CAN DO THE REGRESSION ANALYSIS

LSTAT IS NOT THE SIGNIFICANT VARIABLE

SO WE CANNOT PROCEED WITH THIS MODEL BECAUSE RSQUARE IS LESS THAN 60% AND MAX POSSIBLE ERROR IS GREATER THAN 10%

MEAN	ROOT	AVERAGE OF Y	%
38.48297	6.203464	22.53280632	0.2753081

ASSUMPTIONS		
MEAN	-2.7365E-14	MET
SKEWNESS	1.45706199	NOT MET
THERE IS NO CONSTANT VARIANCE		MET



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		<0.6 MET					
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	-7.59190028	4.875355
AVG_ROOM	5.094788	0.444466	11.46273	3.47E-27	4.22155	5.968026	4.221550436	5.968026
LSTAT	-0.64236	0.043731	-14.6887	6.67E-41	-0.72828	-0.55644	-0.72827717	-0.55644
				BOTH HAS THE P-VALUE GREATER THAN 5%				

MEAN	ROOT	AVERAGE Y	%
30.51246878	5.523809263	22.53281	0.245145198

ASSUMPTION		
MEAN	1.44741E-14	MET
SKEW	1.347227992	NOT MET
NO CONSTANT VARIANCE		MET

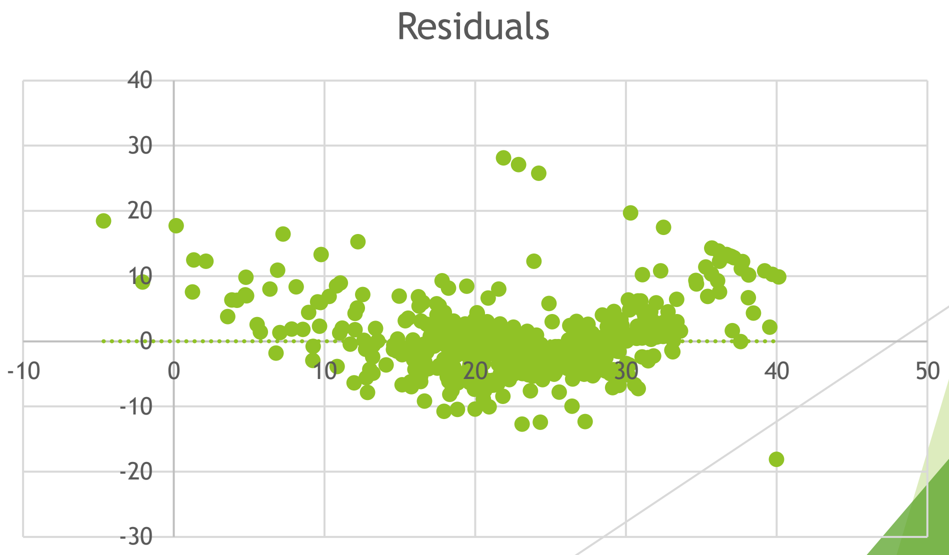
AVG ROOMS	LSTAT	AVG PRICE
7	20	21.45808 =B17+(B18*L20)+(B19*M20)

How does it compare to the company quoting a value of 30000 USD for this locality? Is the company Overcharging/ Undercharging?

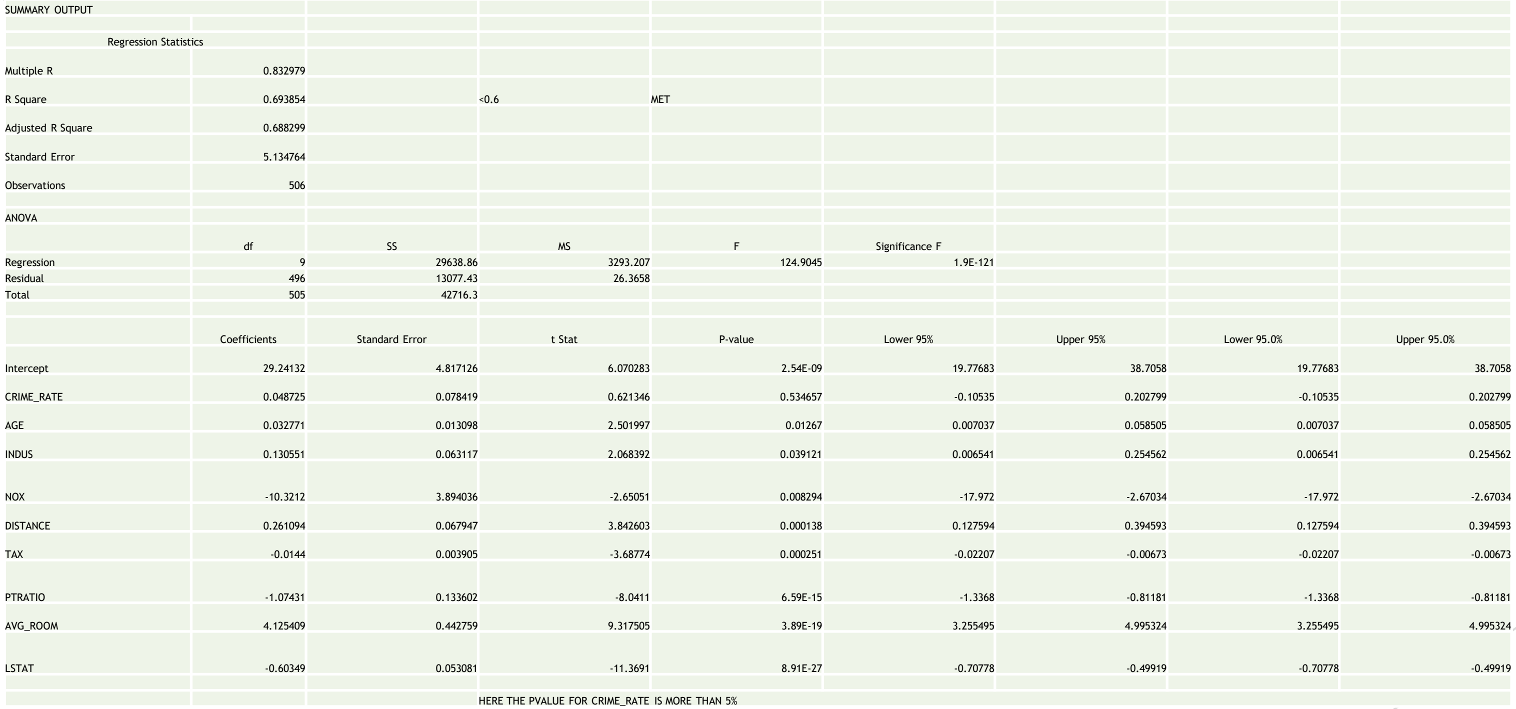
COMPANY IS OVERCHARGING

RSQUARE IS BETTER FOR THIS REGRESSSION MODEL THAN THE PREVIOUS MODEL SINCE THIS MODEL HAS R SQUARE GREATER THEN 60%

SO WE CANNOT PROCEED WITH THIS MODEL BECAUSE SKEWNESS IS NOT MET



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�square, 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		<0.6	MET				
Adjusted R Square	0.688299							
Standard Error	5.134764							
Observations	506							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	9	29638.86	3293.207	124.9045	1.9E-121			
Residual	496	13077.43	26.3658					
Total	505	42716.3						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	29.24132	4.817126	6.070283	2.54E-09	19.77683	38.7058	19.77683	38.7058
CRIME_RATE	0.048725	0.078419	0.621346	0.534657	-0.10535	0.202799	-0.10535	0.202799
AGE	0.032771	0.013098	2.501997	0.01267	0.007037	0.058505	0.007037	0.058505
INDUS	0.130551	0.063117	2.068392	0.039121	0.006541	0.254562	0.006541	0.254562
NOX	-10.3212	3.894036	-2.65051	0.008294	-17.972	-2.67034	-17.972	-2.67034
DISTANCE	0.261094	0.067947	3.842603	0.000138	0.127594	0.394593	0.127594	0.394593
TAX	-0.0144	0.003905	-3.68774	0.000251	-0.02207	-0.00673	-0.02207	-0.00673
PTRATIO	-1.07431	0.133602	-8.0411	6.59E-15	-1.3368	-0.81181	-1.3368	-0.81181
AVG_ROOM	4.125409	0.442759	9.317505	3.89E-19	3.255495	4.995324	3.255495	4.995324
LSTAT	-0.60349	0.053081	-11.3691	8.91E-27	-0.70778	-0.49919	-0.70778	-0.49919
			HERE THE PVALUE FOR CRIME_RATE IS MORE THAN 5%					

	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE		
CRIME_RATE	1											
AGE	0.006859	1										
INDUS	-0.00551	0.644779	1									
NOX	0.001851	0.73147	0.763651	1								
DISTANCE	-0.00906	0.456022	0.595129	0.611441	1							
TAX	-0.01675	0.506456	0.72076	0.668023	0.910228	1						
PTRATIO	0.010801	0.261515	0.383248	0.188933	0.464741	0.460853	1					
AVG_ROOM	0.027396	-0.24026	-0.39168	-0.30219	-0.20985	-0.29205	-0.3555	1				
LSTAT	-0.0424	0.602339	0.6038	0.590879	0.488676	0.543993	0.374044	-0.61381	1			
AVG_PRICE	0.043338	-0.37695	-0.48373	-0.42732	-0.38163	-0.46854	-0.50779	0.69536	-0.73766	1		

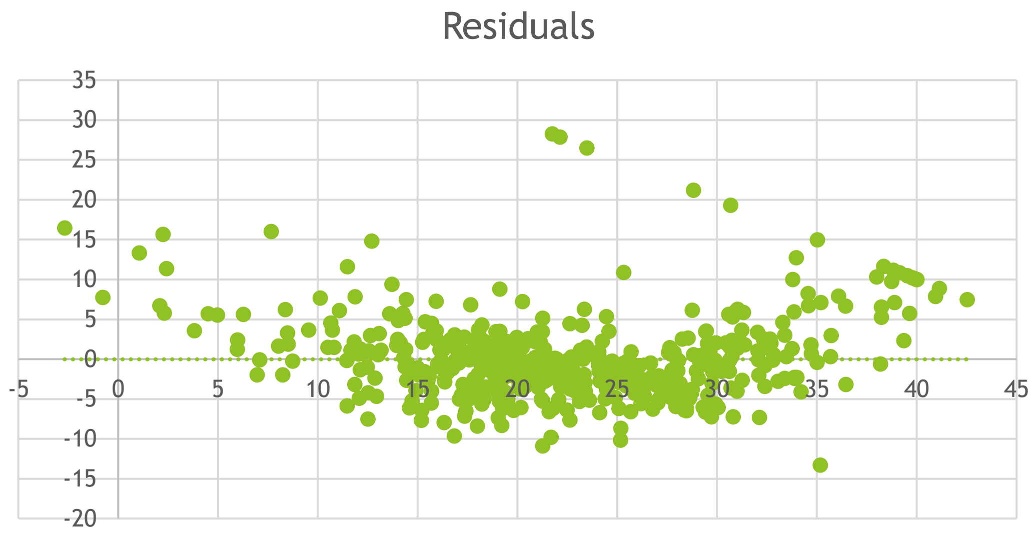
HERE WE CAN PREDICT THAT WE CAN GET THE GOOD MODEL SINCE WE THE GOOD RELATIONSHIP WITH THE DEPENDENT VARIABLE

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			THERE IS NO EFFECTIVE CHANGE IN THE VALUE OF RSQUARE				
R Square	0.693615		<0.6	EVEN AFTER REMOVING THE CRIME_RATE				
Adjusted R Square	0.688684							
Standard Error	5.131591							
Observations	506							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	8	29628.68	3703.585	140.6430411	1.911E-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	1.84597E-09	19.9883896	38.8685574	19.98839	38.8685574
AGE	0.032935	0.013087	2.516606	0.012162875	0.00722219	0.058647734	0.0072222	0.058647734
INDUS	0.13071	0.063078	2.072202	0.038761669	0.00677794	0.254642071	0.0067779	0.254642071
NOX	-10.2727	3.890849	-2.64022	0.008545718	-17.917246	-2.628164466	-17.917246	-2.628164466
DISTANCE	0.261506	0.067902	3.851242	0.000132887	0.12809638	0.394916471	0.1280964	0.394916471
TAX	-0.01445	0.003902	-3.70395	0.000236072	-0.0221186	-0.006786137	-0.0221186	-0.006786137
PTRATIO	-1.0717	0.133454	-8.03053	7.08251E-15	-1.3339051	-0.809499836	-1.3339051	-0.809499836
AVG_ROOM	4.125469	0.442485	9.3234	3.68969E-19	3.2560963	4.994841615	3.2560963	4.994841615
LSTAT	-0.60516	0.05298	-11.4224	5.41844E-27	-0.7092519	-0.501066704	-0.7092519	-0.501066704
				ALL THE VALUES ARE LESS THAN 5%				

MEAN	ROOT	AVG Y	%					
25.8648497 9	5.08574968	22.53280632	0.2257042					
			NOT MET GREATER THAN 5%					
	ASSUMPTION							
	MEAN	-1.03948E-14						
	SKEW	1.643869514	NOT MET					
	THERE IS NO CONSTANT VARIATION							WE CANNOT USE THIS MODEL FURTHER



Column1	Coefficients											
NOX	-10.27270508											
PTRATIO	-1.071702473		HERE WE CAN SEE THAT NOX HAS THE -VE VALUE SO THE AVG_PRICE WILL DECREASE IF THE NOX IS MORE IN A PARTICULAR AREA IN THIS TOWN									
LSTAT	-0.605159282											
TAX	-0.014452345											
AGE	0.03293496											
INDUS	0.130710007		AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE	
DISTANCE	0.261506423		50	9	0.9	4	300	15.1	6.567	30	12.47097	=I77+I73*K75+I74*L75+I69*M75+Table1[@Coefficients]*N75+I72*O75+I70*P75+I76*Q75+I71*R75
AVG_ROOM	4.125468959		50	9	0.6	4	300	15.1	6.567	30	15.55278	=I77+Table1[@Coefficients]*Q76+I75*N76+I74*L76+I73*K76+I72*O76+I71*R76+I70*P76+I69*M76
Intercept	29.42847349											
				HERE WE CAN SAY THAT IF VALUE OF NOX IS MORE THEN AVG_PRICE IS LESS AND VICE-VERSA								