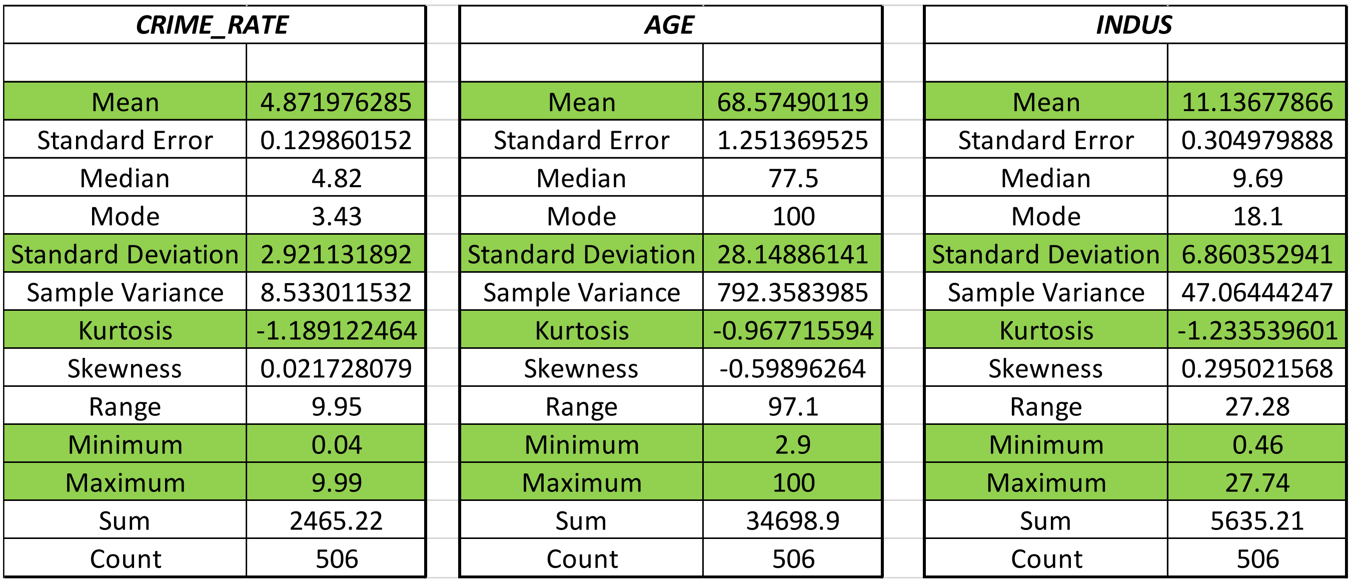
Dominic John

Terro’s Real Estate Agency - Assignment

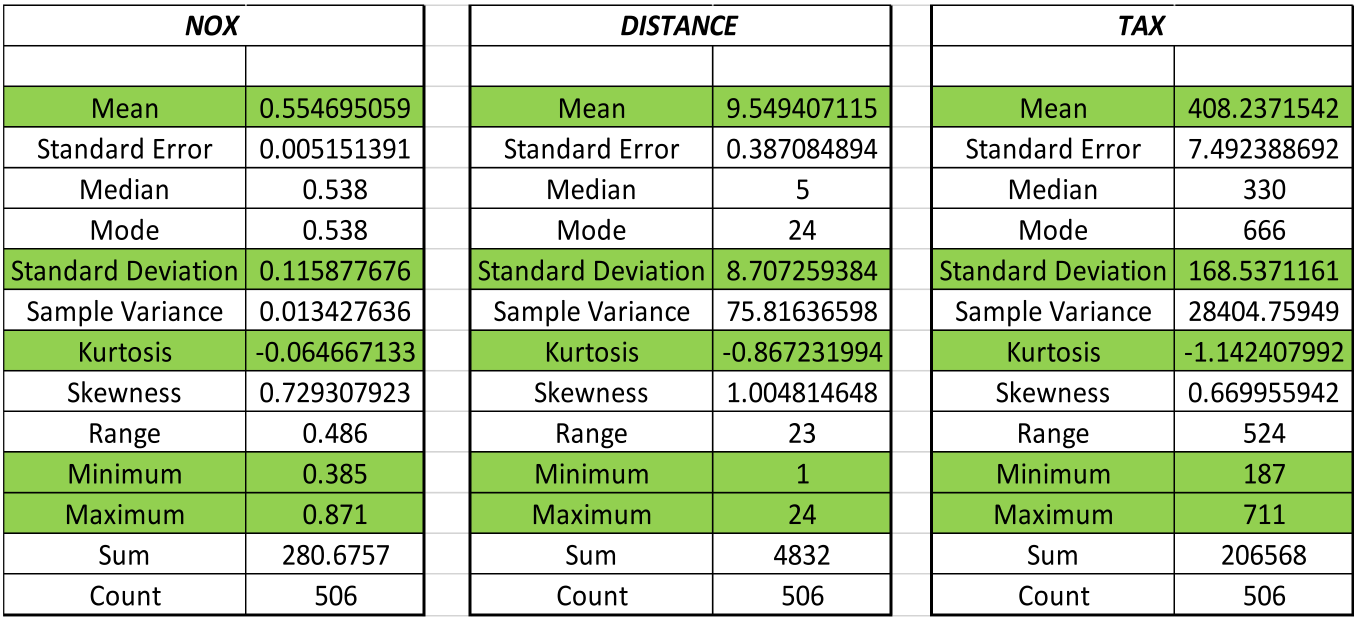
Q.1. Generate the summary statistics for each variable in the table. Write down your observation.



CRIME_RATE	
Mean	4.871976285
Standard Error	0.129860152
Median	4.82
Mode	3.43
Standard Deviation	2.921131892
Sample Variance	8.533011532
Kurtosis	-1.189122464
Skewness	0.021728079
Range	9.95
Minimum	0.04
Maximum	9.99
Sum	2465.22
Count	506

AGE	
Mean	68.57490119
Standard Error	1.251369525
Median	77.5
Mode	100
Standard Deviation	28.14886141
Sample Variance	792.3583985
Kurtosis	-0.967715594
Skewness	-0.59896264
Range	97.1
Minimum	2.9
Maximum	100
Sum	34698.9
Count	506

INDUS	
Mean	11.13677866
Standard Error	0.304979888
Median	9.69
Mode	18.1
Standard Deviation	6.860352941
Sample Variance	47.06444247
Kurtosis	-1.233539601
Skewness	0.295021568
Range	27.28
Minimum	0.46
Maximum	27.74
Sum	5635.21
Count	506



NOX	
Mean	0.554695059
Standard Error	0.005151391
Median	0.538
Mode	0.538
Standard Deviation	0.115877676
Sample Variance	0.013427636
Kurtosis	-0.064667133
Skewness	0.729307923
Range	0.486
Minimum	0.385
Maximum	0.871
Sum	280.6757
Count	506

DISTANCE	
Mean	9.549407115
Standard Error	0.387084894
Median	5
Mode	24
Standard Deviation	8.707259384
Sample Variance	75.81636598
Kurtosis	-0.867231994
Skewness	1.004814648
Range	23
Minimum	1
Maximum	24
Sum	4832
Count	506

TAX	
Mean	408.2371542
Standard Error	7.492388692
Median	330
Mode	666
Standard Deviation	168.5371161
Sample Variance	28404.75949
Kurtosis	-1.142407992
Skewness	0.669955942
Range	524
Minimum	187
Maximum	711
Sum	206568
Count	506

PTRATIO	
Mean	18.4555336
Standard Error	0.096243568
Median	19.05
Mode	20.2
Standard Deviation	2.164945524
Sample Variance	4.686989121
Kurtosis	-0.285091383
Skewness	-0.802324927
Range	9.4
Minimum	12.6
Maximum	22
Sum	9338.5
Count	506

AVG_ROOM	
Mean	6.284634387
Standard Error	0.031235142
Median	6.2085
Mode	5.713
Standard Deviation	0.702617143
Sample Variance	0.49367085
Kurtosis	1.891500366
Skewness	0.403612133
Range	5.219
Minimum	3.561
Maximum	8.78
Sum	3180.025
Count	506

LSTAT	
Mean	12.65306324
Standard Error	0.317458906
Median	11.36
Mode	8.05
Standard Deviation	7.141061511
Sample Variance	50.99475951
Kurtosis	0.493239517
Skewness	0.906460094
Range	36.24
Minimum	1.73
Maximum	37.97
Sum	6402.45
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

Observations

I. The mean value of the crime rate variable is relatively low at 4.87, which provides insight into the prevailing level of criminal activity within the localities. This suggests that, on average, the communities experience a moderate or subdued crime rate.

II. Reflecting the historical nature of properties in the localities, the average age is notably high, registering at 68.57. However, the substantial range between the maximum and minimum ages indicates a diverse mix of both newer and older properties. This range showcases the varied architectural timelines and development phases within the region.

III. The average extent of non-retail business acres per town, approximately 11.14, serves as a measure of commercial diversification. This statistic implies that a considerable number of areas display a substantial proportion of land dedicated to non-retail business activities, portraying a vibrant economic landscape with a focus on various industries.

IV. The localities' average pollution level hovers around 0.55, indicating a moderate level of environmental contamination. This insight highlights the region's overall efforts to maintain a balance between industrial and ecological considerations.

V. With an average distance from the highway of about 9.55, the data reveals an interesting spatial aspect of the localities. This average distance underscores a prevailing trend where properties tend to be situated at relatively greater distances from major transportation routes, possibly reflecting a deliberate preference for quieter, less congested residential areas.

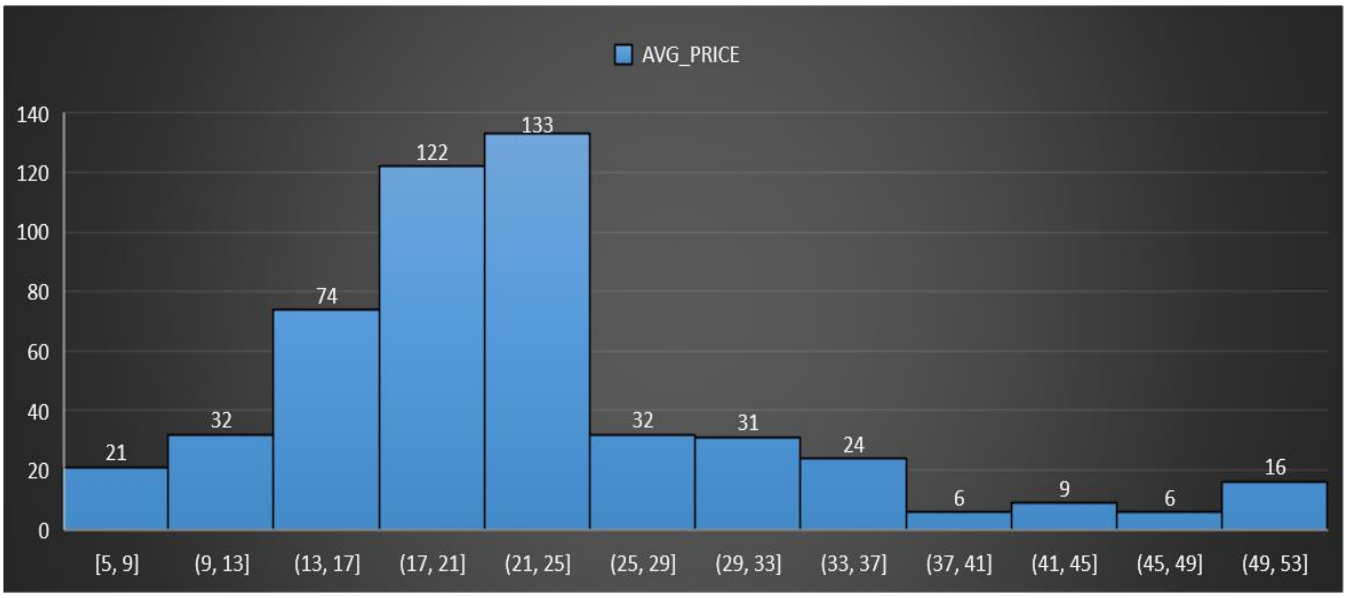
VI. The mean property tax rate, approximately 408.24, provides valuable context about the financial obligations of residents. This figure signifies a moderate taxation level, suggesting a balanced approach by local authorities to fund public services and infrastructure while not overly burdening property owners.

VII. The average pupil-teacher ratio, at around 18.46, sheds light on the educational dynamics within the localities. This ratio indicates a moderate balance between the number of students and teachers, implying an environment where educators can provide a reasonable level of attention to each pupil for effective learning.

VIII. An average of approximately 6.28 rooms in properties offers insights into the local housing landscape. This moderate room count suggests a standard expectation for dwelling sizes, potentially reflecting a common preference for a balanced living space that accommodates various needs.

IX. The approximate average house price of 22.53 hints at the economic spectrum across the localities. This statistic suggests that a considerable number of regions tend to experience slightly higher house prices, possibly due to factors such as location, amenities, or overall desirability.

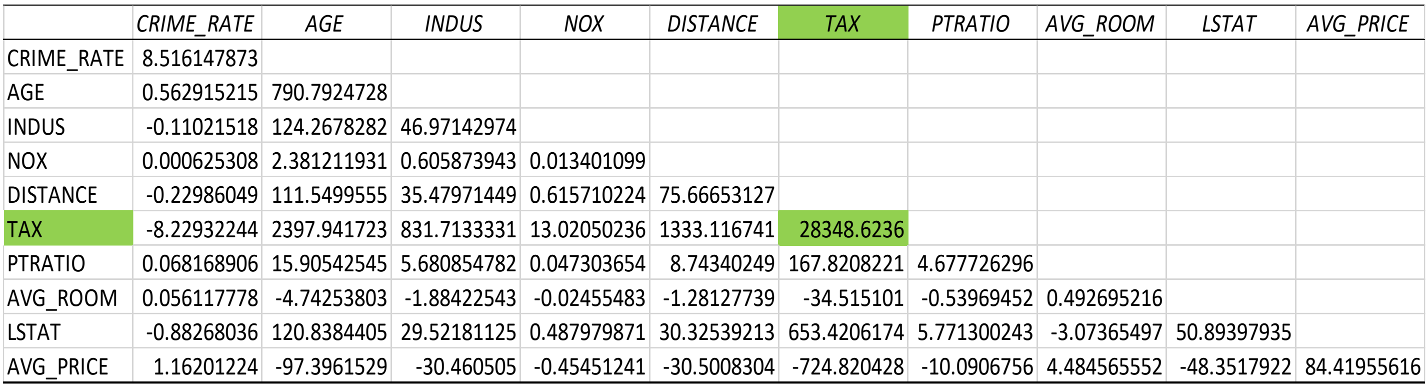
Q. 2. Plot a histogram of the average price variable. What do you infer?



Observations

The histogram is like a picture that helps us understand how much houses cost on average and how these prices are spread out in different groups. We can see where most prices are, which helps us know the normal price range for houses in the data. The shape of the histogram tells us something too. It shows that some houses have really high prices, more than most others. The highest part of the histogram tells us the prices most houses have. We can also see that in many places, house prices are a bit higher.

Q.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	-0.11021518	124.2678282	46.97142974							
NOX	0.000625308	2.381211931	0.605873943	0.013401099						
DISTANCE	-0.22986049	111.5499555	35.47971449	0.615710224	75.66653127					
TAX	-8.22932244	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	-1.88422543	-0.02455483	-1.28127739	-34.515101	-0.53969452	0.492695216		
LSTAT	-0.88268036	120.8384405	29.52181125	0.487979871	30.32539213	653.4206174	5.771300243	-3.07365497	50.89397935	
AVG_PRICE	1.16201224	-97.3961529	-30.460505	-0.45451241	-30.5008304	-724.820428	-10.0906756	4.484565552	-48.3517922	84.41955616

Observations

I. Crime rate's variance is about 8.516, which shows how much it changes by itself.

II. More "Indus" usually means less crime, with a number around -0.1102 showing this link.

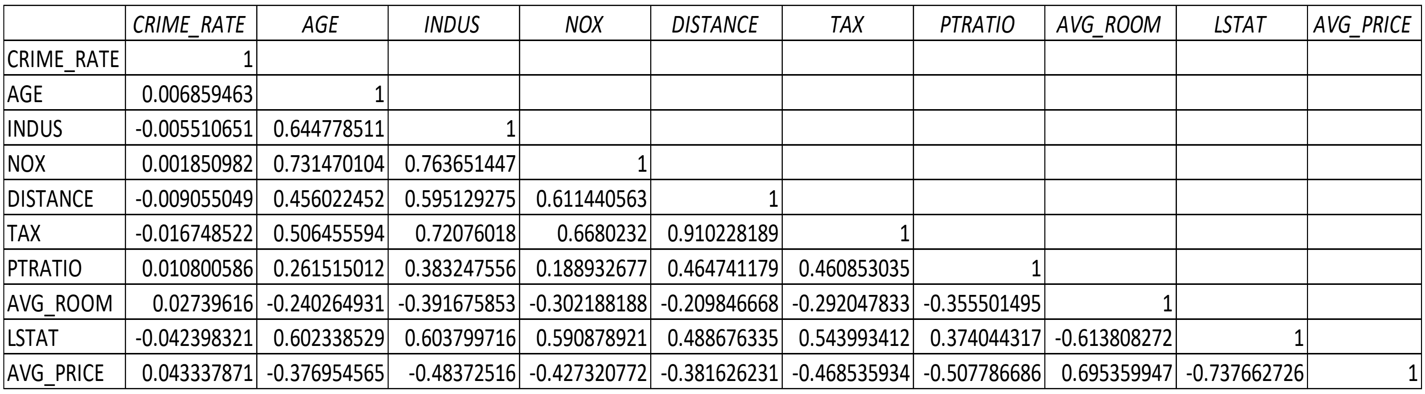
III. High "Tax" often goes with less crime, with a strong connection shown by a big negative number (-8.2293).

IV. As areas get older, crime tends to go up, shown by a positive number of about 0.5629 for "Age."

V. A bit of air pollution "Nox" has a very small link with a little more crime, about 0.0006.

VI. Being farther from things "Distance" often means less crime, as indicated by a negative number of around -0.2298.

Q.4. Create a correlation matrix of all the variables?



	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	-0.737662726	1

a) Which are the top 3 positively correlated pairs and

I. Indus – Nox
II. Age – Nox
III. Distance – Tax

b) Which are the top 3 negatively correlated pairs

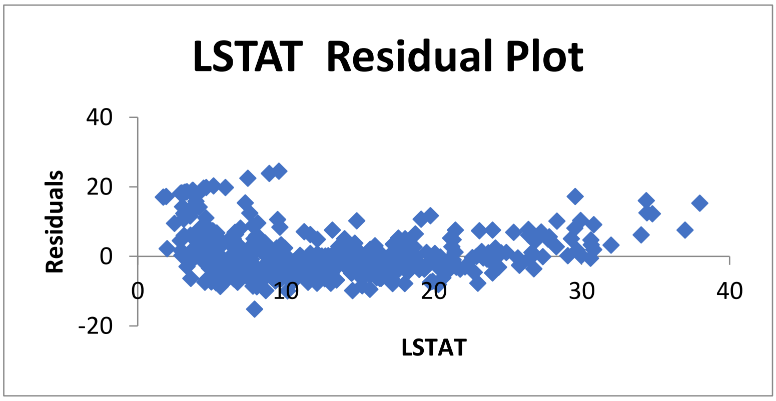
IV. Lstat – Avg price
V. Avg room – Lstat
VI. Ptratio – Avg Price

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

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	35.65922472
LSTAT	-0.950049354	0.038733416	-24.52789985	5.0811E-88	-1.0261482	-0.873950508	-1.0261482	-0.873950508



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

We're looking at the connection between the independent variable LSTAT and the dependent variable Average Price. In our analysis, we find a multiple R value of 0.7377. Checking the P Value, we see it's extremely low (0), indicating that the regression model is very meaningful statistically. The starting point for this analysis, called the intercept, is at 34.5538. When we examine the data, we notice that as the independent variable LSTAT increases, the dependent variable AVG PRICE tends to decrease.

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

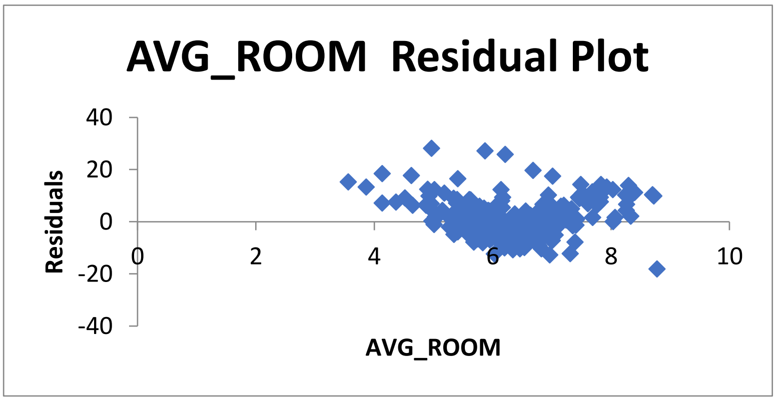
The variable LSTAT plays an important role in our analysis. Every time LSTAT changes by one unit, we see a corresponding change of 0.95 in Average Price. This pair of variables has a strong negative relationship, which is more noticeable compared to other variables in our data.

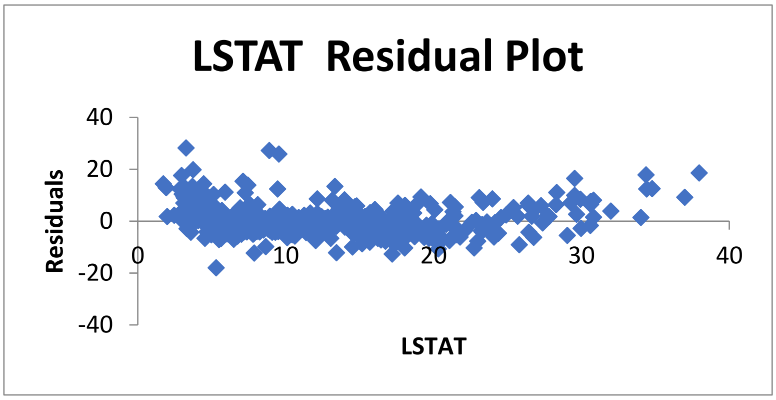
Q.6. Build a new Regression model including LSTAT and AVG_ROOM together as Independent variables and AVG_PRICE as dependent variables.

SUMMARY OUTPUT	
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	444.3308922	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	-0.428095348	0.668764941	-7.591900282	4.875354658	-7.591900282	4.875354658
AVG_ROOM	5.094787984	0.4444655	11.46272991	3.47226E-27	4.221550436	5.968025533	4.221550436	5.968025533
LSTAT	-0.642358334	0.043731465	-14.68869925	6.66937E-41	-0.728277167	-0.556439501	-0.728277167	-0.556439501





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

Y = -1.358 +5.09 X0 -0.642 X1

Where Y=AVG ROOM

X0 = AVG ROOM

X1 = LSTAT As per the model,

AVG PRICE for new house

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

Ie; the price for the new house is 21440

So, we can conclude that company is Overcharging

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.***

When we compared to previous model performance of this model is better

From the linear equation model

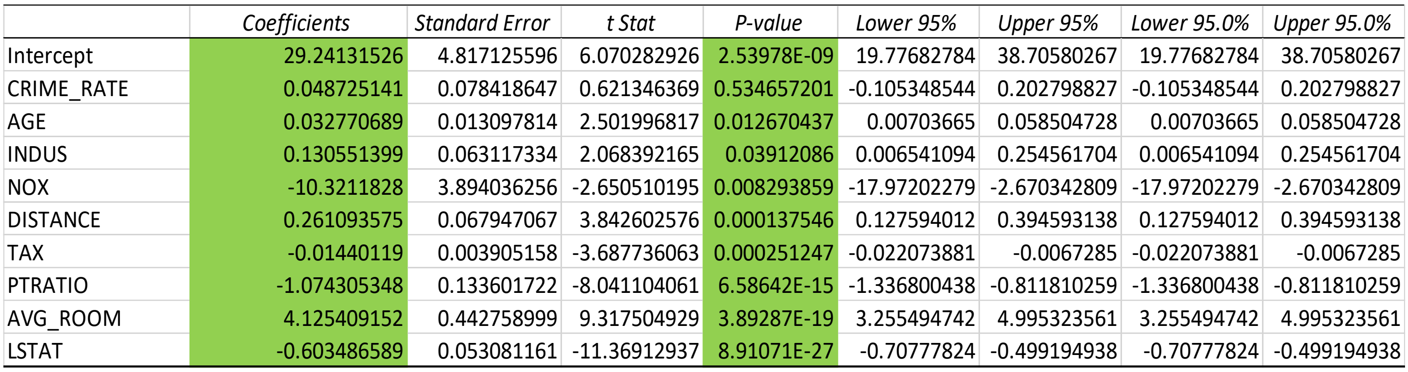
Y = -1.35 +5.09a -0.64b

A = Avg_room B = LSTAT

And Value of R square = 0.6385

We found that both Avg_room and LSTAT together make up 63% of the changes in the average price. This is supported by a strong connection shown by the multiple R value of 0.79. In the previous model, only LSTAT explained 54% of the changes in the average price.

Q.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.



	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

The crime rate doesn't have a significant effect on the average house price because its p-value is higher than 0.5. When we consider all the features, they together account for 69% of the differences in the average house price. Features such as NOX, TAX, PTRATIO, and LSTAT have negative coefficients, which means that higher values of these features result in lower house prices, and vice versa.

Q.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

We have examined important factors such as AGE, INDUS, NOX, DISTANCE, TAX, PTRATIO, AVG ROOM, and LSTAT in relation to the dependent variable AVG PRICE.

a) *Interpret the output of this model*

The multiple R Square value is 0.8328, which is quite close to 1. This indicates a strong positive connection.

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.*

When comparing the previous model's Adjusted R Square value (0.688298647) with the current model's R Squared value (0.688683682), they are very similar. This suggests that both models explain things similarly. The inclusion of the variable "crime rate" did not significantly change the model.

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?***

If the level of NOX increases, the average price value will decrease. In simpler terms, better air quality is linked to higher average prices.

d) ***Write the regression equation from this model.***

Equation:-

Y = 29.42847348+0.032934961 * AGE + 0.130710006 * INDUS - 10.27270514 * NOX+0.261506423 * DISTANCE - 0.014452345 * TAX - 1.071702472 * PTRATIO + 4.125468961*AVG ROOM - 0.605159282*LSTAT