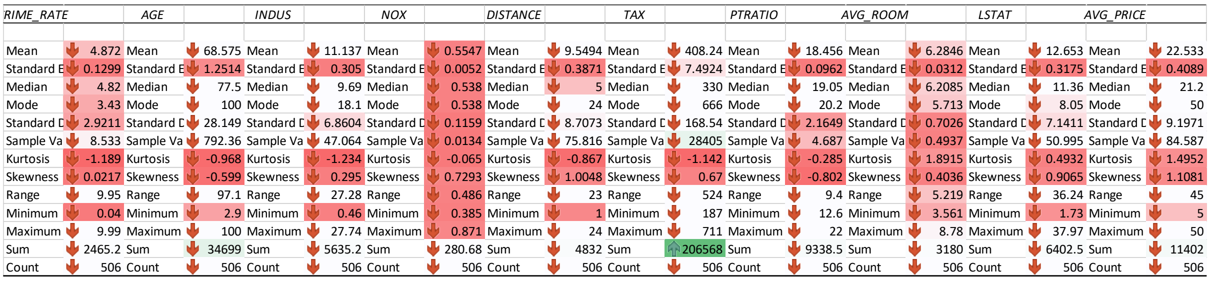
Project Report

Question 1: - The first step to any project is understanding the data. So, for this step, generate the summary statistics for each of the variables. What do you observe?

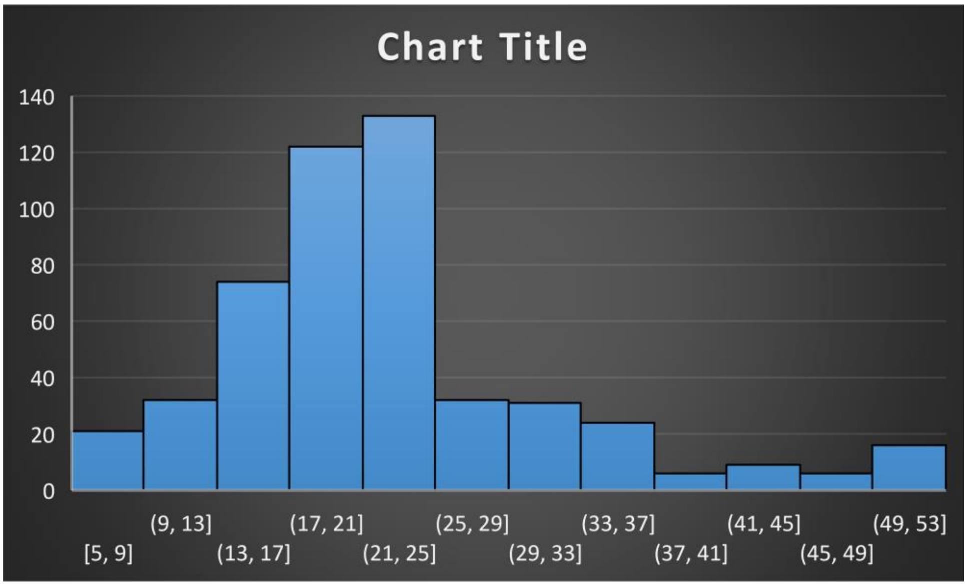


RIME_RATE		AGE		INDUS		NOX		DISTANCE		TAX		PTRATIO		AVG_ROOM		LSTAT		AVG_PRICE	
Mean	4.872	Mean	68.575	Mean	11.137	Mean	0.5547	Mean	9.5494	Mean	408.24	Mean	18.456	Mean	6.2846	Mean	12.653	Mean	22.533
Standard E	0.1299	Standard E	1.2514	Standard E	0.305	Standard E	0.0052	Standard E	0.3871	Standard E	7.4924	Standard E	0.0962	Standard E	0.0312	Standard E	0.3175	Standard E	0.4089
Median	4.82	Median	77.5	Median	9.69	Median	0.538	Median	5	Median	330	Median	19.05	Median	6.2085	Median	11.36	Median	21.2
Mode	3.43	Mode	100	Mode	18.1	Mode	0.538	Mode	24	Mode	666	Mode	20.2	Mode	5.713	Mode	8.05	Mode	50
Standard D	2.9211	Standard D	28.149	Standard D	6.8604	Standard D	0.1159	Standard D	8.7073	Standard D	168.54	Standard D	2.1649	Standard D	0.7026	Standard D	7.1411	Standard D	9.1971
Sample Va	8.533	Sample Va	792.36	Sample Va	47.064	Sample Va	0.0134	Sample Va	75.816	Sample Va	28405	Sample Va	4.687	Sample Va	0.4937	Sample Va	50.995	Sample Va	84.587
Kurtosis	-1.189	Kurtosis	-0.968	Kurtosis	-1.234	Kurtosis	-0.065	Kurtosis	-0.867	Kurtosis	-1.142	Kurtosis	-0.285	Kurtosis	1.8915	Kurtosis	0.4932	Kurtosis	1.4952
Skewness	0.0217	Skewness	-0.599	Skewness	0.295	Skewness	0.7293	Skewness	1.0048	Skewness	0.67	Skewness	-0.802	Skewness	0.4036	Skewness	0.9065	Skewness	1.1081
Range	9.95	Range	97.1	Range	27.28	Range	0.486	Range	23	Range	524	Range	9.4	Range	5.219	Range	36.24	Range	45
Minimum	0.04	Minimum	2.9	Minimum	0.46	Minimum	0.385	Minimum	1	Minimum	187	Minimum	12.6	Minimum	3.561	Minimum	1.73	Minimum	5
Maximum	9.99	Maximum	100	Maximum	27.74	Maximum	0.871	Maximum	24	Maximum	711	Maximum	22	Maximum	8.78	Maximum	37.97	Maximum	50
Sum	2465.2	Sum	34699	Sum	5635.2	Sum	280.68	Sum	4832	Sum	206568	Sum	9338.5	Sum	3180	Sum	6402.5	Sum	11402
Count	506	Count	506	Count	506	Count	506	Count	506	Count	506	Count	506	Count	506	Count	506	Count	506

According to the summary statistics, a flat typically costs around 22.53 (amount). Pupil and Teacher Ratio, which is in a good range (9.40 range of PTRATIO), is a good reason to buy a flat and can draw more people to buy apartments in that area. A further

The average number of rooms in a flat is 6.28, or nearly 6, which is a good reason and may also draw buyers. Some people are looking to purchase apartments close to the highway, which is 9.55 miles away on average. However, there are some drawbacks, such as the 4.87 average crime rate, 408.24 average tax rate, and 68.57 average age of buildings.

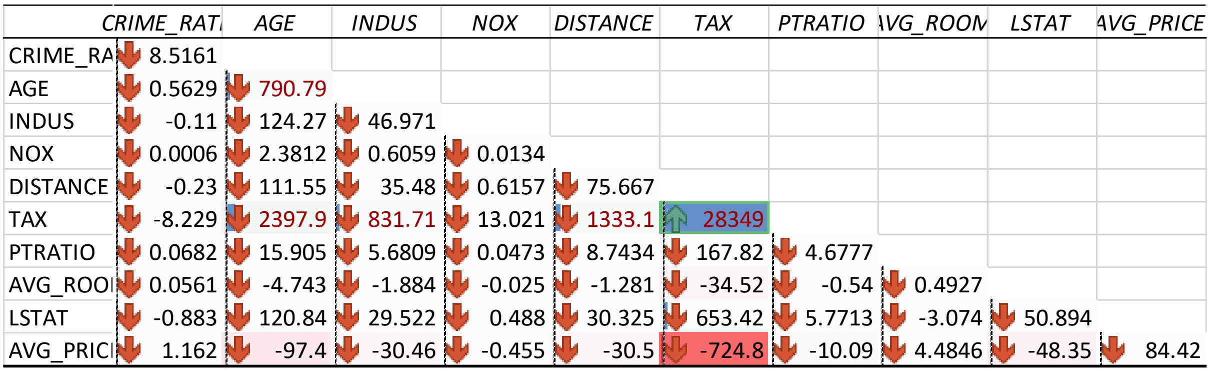
Question 2: - Plot the histogram of the Avg_Price Variable. What do you infer?



The prices of the houses range from $5000 to $50000, with the average price being $22000. Based on the histogram, we can see that there are more houses in the price range of $20000 to $50000. $25000. The histogram is described as "right skewed." Because the Avg_price is the dependent variable for all of the other variables in the table, it is influenced by them. The Avg_price will be affected by other variables such as tax, crime rate, nox, avg_room, and so on. For instance, if the crime rate and nox are high, the price will be low, whereas if the rooms are more, the price will be high.

Question 3: -

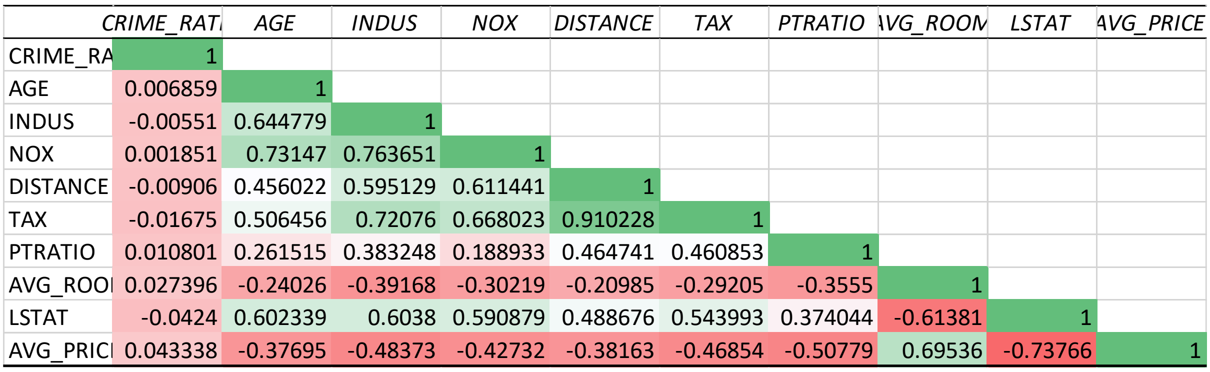
Compute the covariance matrix. Share your observations.



	CRIME_RATI	*AGE*	*INDUS*	*NOX*	*DISTANCE*	*TAX*	*PTRATIO*	*VG_ROOM*	*LSTAT*	*VG_PRICE*
CRIME_RA	⬇ 8.5161									
AGE	⬇ 0.5629	⬇ 790.79								
INDUS	⬇ -0.11	⬇ 124.27	⬇ 46.971							
NOX	⬇ 0.0006	⬇ 2.3812	⬇ 0.6059	⬇ 0.0134						
DISTANCE	⬇ -0.23	⬇ 111.55	⬇ 35.48	⬇ 0.6157	⬇ 75.667					
TAX	⬇ -8.229	⬇ 2397.9	⬇ 831.71	⬇ 13.021	⬇ 1333.1	⬆ 28349				
PTRATIO	⬇ 0.0682	⬇ 15.905	⬇ 5.6809	⬇ 0.0473	⬇ 8.7434	⬇ 167.82	⬇ 4.6777			
AVG_ROOI	⬇ 0.0561	⬇ -4.743	⬇ -1.884	⬇ -0.025	⬇ -1.281	⬇ -34.52	⬇ -0.54	⬇ 0.4927		
LSTAT	⬇ -0.883	⬇ 120.84	⬇ 29.522	⬇ 0.488	⬇ 30.325	⬇ 653.42	⬇ 5.7713	⬇ -3.074	⬇ 50.894	
AVG_PRICI	⬇ 1.162	⬇ -97.4	⬇ -30.46	⬇ -0.455	⬇ -30.5	⬇ -724.8	⬇ -10.09	⬇ 4.4846	⬇ -48.35	⬇ 84.42

Covariance is a measure of the relationship between two random variables that describes how much they change in tandem. In simple terms, covariance describes the direction; if the value is positive, the variables move in the same direction; if the value is negative, the variables move in the opposite direction. According to the above covariance matrix, average price and tax have a negative relationship, whereas average price and average rooms have a positive relationship.

Question 4: - Create a correlation matrix of all the variables as shown in the Videos and various case studies. State top 3 positively correlated pairs and top 3 negatively correlated pairs.



	CRIME_RATI	*AGE*	*INDUS*	*NOX*	*DISTANCE*	*TAX*	*PTRATIO*	*VG_ROOM*	*LSTAT*	*VG_PRICE*
CRIME_RA	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_ROOI	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_PRICI	0.043338	-0.37695	-0.48373	-0.42732	-0.38163	-0.46854	-0.50779	0.69536	-0.73766	1

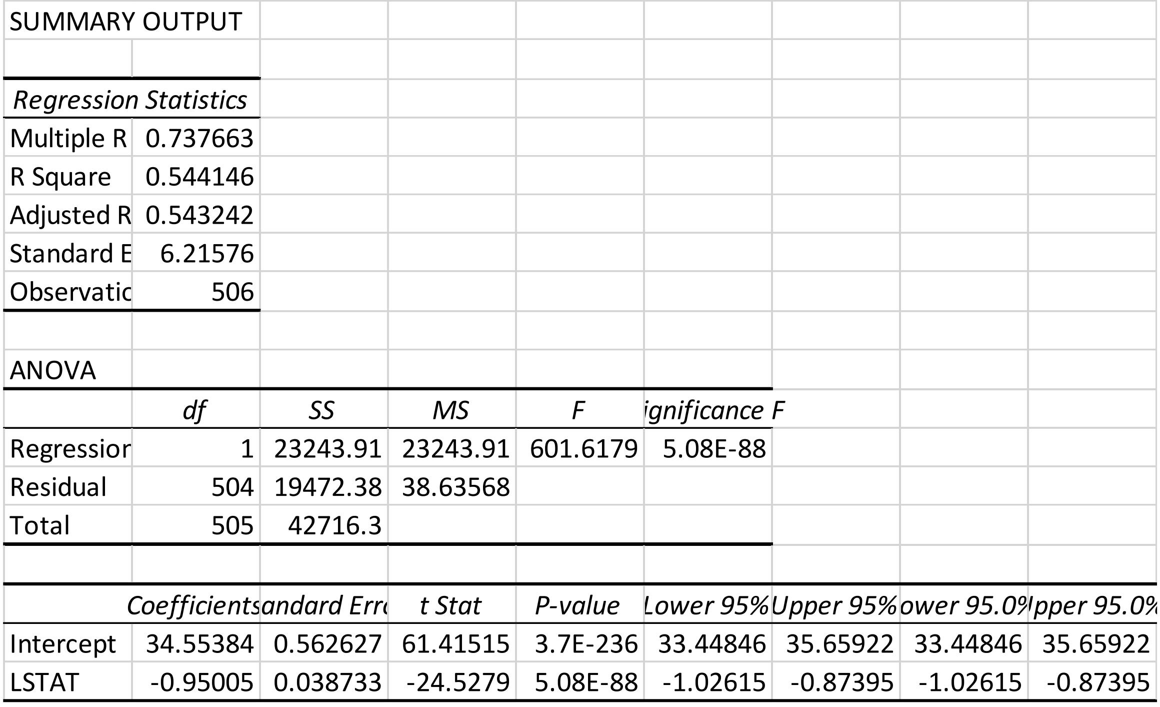
A correlation matrix is a table that displays the coefficient relationships between variables. Each cell in the table represents the relationship between two variables. A correlation matrix is used to summarize data, as an input into a more advanced analysis, and as a diagnostic tool for advanced analytics.

analyses. The top three positively related pairs are 0.9102 (Tax and Distance), 0.7636 (Nox and Indus), and 0.7314 (Nox and Age), while the top three negatively related pairs are -0.7376 (Avg_price & LSTAT), -0.6138 (LSTAT & Avg_room), and -0.5077 (Avg_Price & Ptratio).

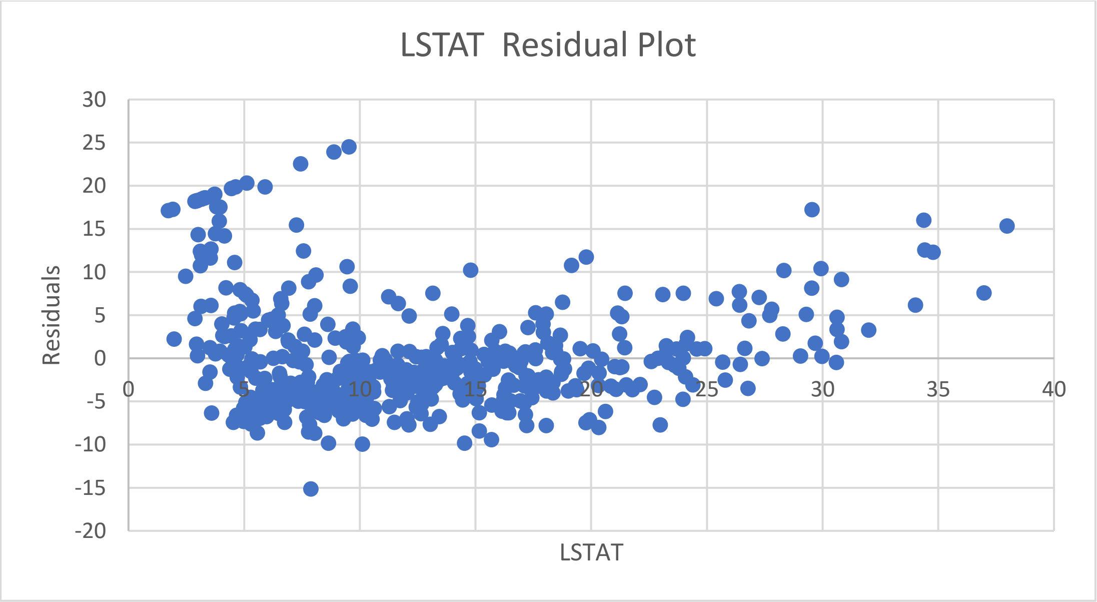
Question 5: - Build an initial regression model with AVG_PRICE as the y or the Dependent variable and LSTAT as the Independent variable. Generate the residual plot too.

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	0.737663							
R Square	0.544146							
Adjusted R	0.543242							
Standard E	6.21576							
Observatic	506							
ANOVA								
	df	*SS*	*MS*	*F*	*ignificance F*			
Regressior	1	23243.91	23243.91	601.6179	5.08E-88			
Residual	504	19472.38	38.63568					
Total	505	42716.3						
	Coefficients	*andard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 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



a.) From the regression summary output, we can see that if the coefficient value is positive and increasing, the variance will increase as well, but if the coefficient value is negative and increasing, the variance will decrease, and we can see if there is a pattern in the trendline where it is a straight line.

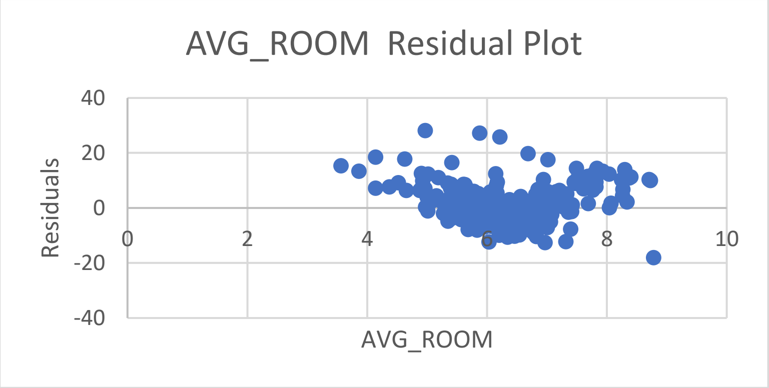
b.) We know from residuals that if the value is less than 0.05, it is significant. b.) Because the p-value is less than 0.05, LSTAT is significant.

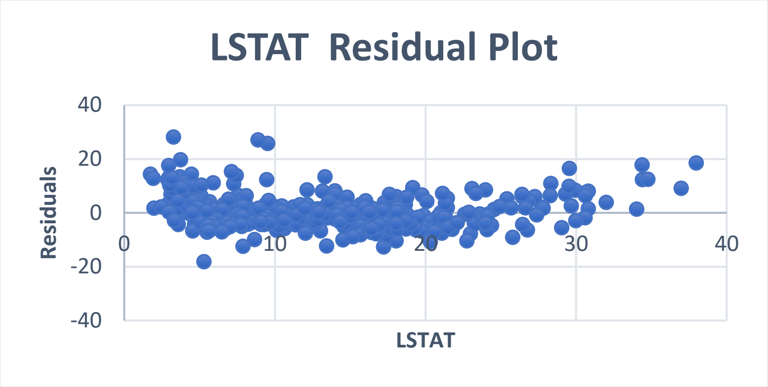
Question 6: - Build another instance of the Regression model but this time include LSTAT and AVG_ROOM together as independent variables and AVG_PRICE as the dependent variable.

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. Explain.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.7991							
R Square	0.638562							
Adjusted R	0.637124							
Standard E	5.540257							
Observatic	506							
ANOVA								
	df	*SS*	*MS*	*F*	*ignificance F*			
Regressior	2	27276.99	13638.49	444.3309	7E-112			
Residual	503	15439.31	30.69445					
Total	505	42716.3						
	Coefficien	Standard E	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.(	Upper 95.(
Intercept	-1.35827	3.172828	-0.4281	0.668765	-7.5919	4.875355	-7.5919	4.875355
AVG_ROOI	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

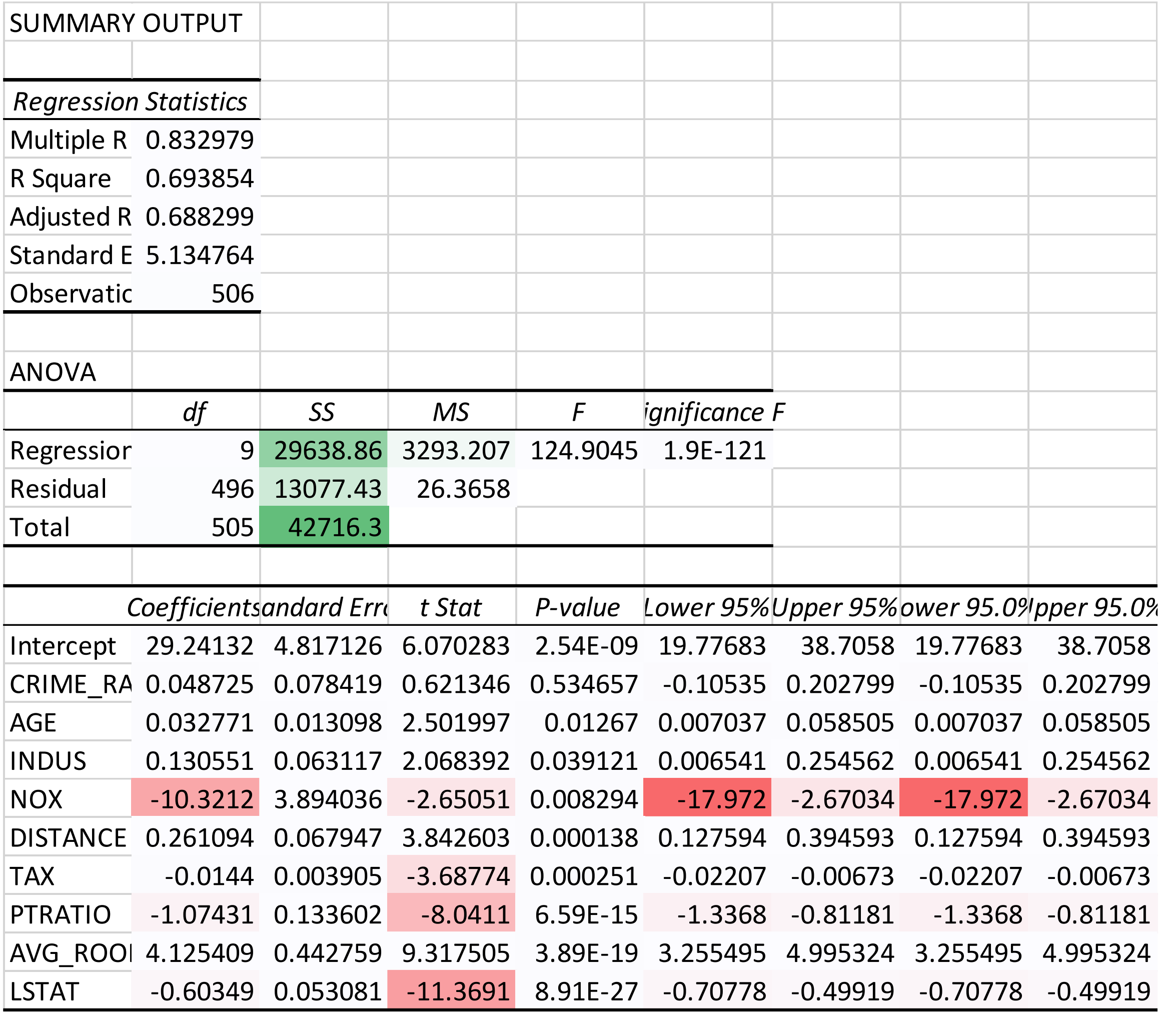




a.) Because the equation (Y= a+B1*X1+B2*X2...+E) defines multiple linear regression, if we have 7 Rooms and 20 for LSTAT, Y= -1.35 + (5.09*7) + (-0.64*20) = 21.48, which is equivalent to 21480USD. As a result, the company is undercharging because it is less than 30000 USD.

b.) The adjusted R square for the previous model was 0.543, while the adjusted R square for this model is 0.637. Based on these values, we can conclude that this regression outperforms the previous model. Because we know that if the adjusted R square is higher, the model performs better, and if the adjusted R square is lower, the model performs poorly.

Question 7: - Now, build a Regression model with all variables. AVG_PRICE shall be the Dependent Variable. Interpret the output in terms of adjusted R-square, coefficient and Intercept values, Significance of variables with respect to AVG_price. Explain.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.832979
R Square	0.693854
Adjusted R	0.688299
Standard E	5.134764
Observatic	506

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regressior	9	29638.86	3293.207	124.9045	1.9E-121
Residual	496	13077.43	26.3658		
Total	505	42716.3			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	29.24132	4.817126	6.070283	2.54E-09	19.77683	38.7058	19.77683	38.7058
CRIME_RA	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_ROOI	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, we know that the value of R square and adjusted R square indicates the model's performance, i.e., "69.3%" The regression coeffient is used to describe the relationship between an independent variable and a dependent variable. Furthermore, the majority of the variables have a perfectly positive linear relationship with Avg_price. Nox, tax, ptratio, and LSTAT all have a perfectly negative linear relationship with Avg_price. We can conclude that, with the exception of the Crime_Rate (0.53), all remaining variables are significant because we know that if the P-value is less than 0.05, the variable is said to be significant.

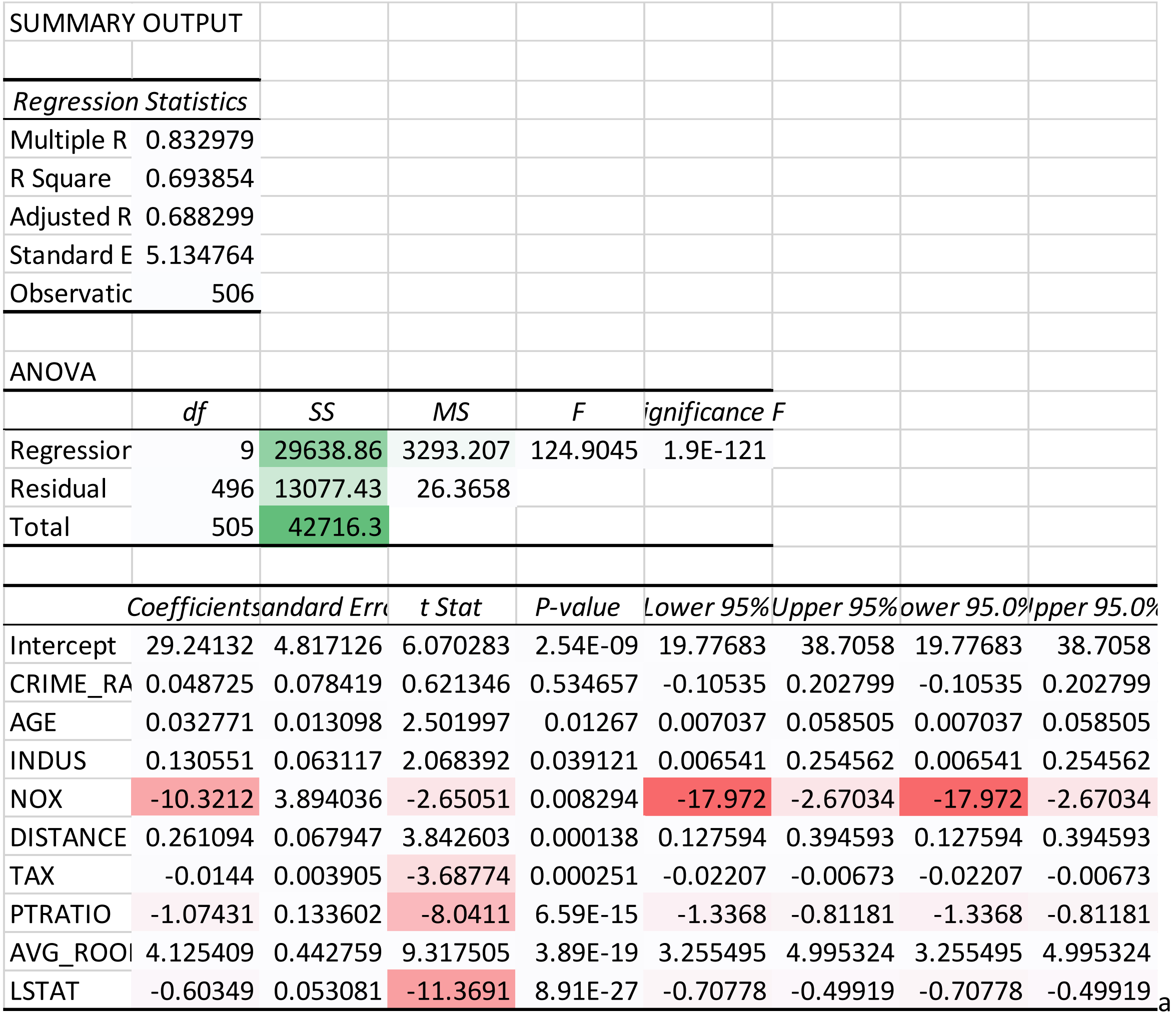
Question 8: - Pick out only the significant variables from the previous question. Make another instance of the Regression model using only the significant variables you just picked.

(HINT: Significant variables are those whose p-values are less than 0.05. If the p-value is greater than 0.05 then it is insignificant) 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 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	0.832979
R Square	0.693854
Adjusted R	0.688299
Standard E	5.134764
Observatic	506

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regressior	9	29638.86	3293.207	124.9045	1.9E-121
Residual	496	13077.43	26.3658		
Total	505	42716.3			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	29.24132	4.817126	6.070283	2.54E-09	19.77683	38.7058	19.77683	38.7058
CRIME_RA	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_ROOI	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

a.)

a.) The graph above depicts regression statistics for significant variables with p-values less than 0.05.

b.) We can see that the current model (0.6886) performs slightly better than the previous model (0.6882) because it is slightly greater than the previous model, and we know that the higher the adjusted R square, the better the performance.

c.) A positive coefficient means that as the value of the independent variables decreases, the mean of the dependent variables increases; a negative coefficient means that as the value of the independent variables increases, the mean of the dependent variables decreases; and after sorting, as the value of nox increases, the avg_price decreases. In other words, as NOX (pollution) levels rise, the average price falls.

d.) Theregression equation is AVG_PRICE = Intercept + (NOX*X1) + (PTRATIO*X2)

+ (LSTAT*X3) + (TAX*X4) + (AGE*X5) + (INDUS*X6) + (DISTANCE*X7) +

(AVG_ROOM*X8)

Were, AVG_PRICE is dependent with other variables.