Q1. 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?

CRIME_RATE	
Mean	4.8719763
Standard Error	0.1298602
Median	4.82
Mode	3.43
Standard Deviation	2.9211319
Sample Variance	8.5330115
Kurtosis	-1.1891225
Skewness	0.0217281
Range	9.95
Minimum	0.04
Maximum	9.99
Sum	2465.22
Count	506

AGE	
Mean	68.574901
Standard Error	1.2513695
Median	77.5
Mode	100
Standard Deviation	28.148861
Sample Variance	792.3584
Kurtosis	-0.9677156
Skewness	-0.5989626
Range	97.1
Minimum	2.9
Maximum	100
Sum	34698.9
Count	506

TAX	
Mean	408.23715
Standard Error	7.4923887
Median	330
Mode	666
Standard Deviation	168.53712
Sample Variance	28404.759
Kurtosis	-1.142408
Skewness	0.6699559
Range	524
Minimum	187
Maximum	711
Sum	206568
Count	506

PTRATIO	
Mean	18.455534
Standard Error	0.0962436
Median	19.05
Mode	20.2
Standard Deviation	2.1649455
Sample Variance	4.6869891
Kurtosis	-0.2850914
Skewness	-0.8023249
Range	9.4
Minimum	12.6
Maximum	22
Sum	9338.5
Count	506

INDUS	
Mean	11.136779
Standard Error	0.3049799

NOX	
Mean	0.5546951
Standard Error	0.0051514

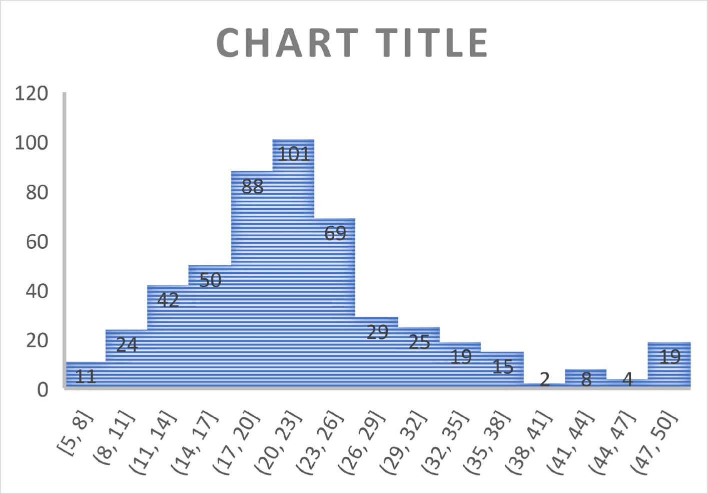
Median	9.69
Mode	18.1
Standard Deviation	6.8603529
Sample Variance	47.064442
Kurtosis	-1.2335396
Skewness	0.2950216
Range	27.28
Minimum	0.46
Maximum	27.74
Sum	5635.21
Count	506

Median	0.538
Mode	0.538
Standard Deviation	0.1158777
Sample Variance	0.0134276
Kurtosis	-0.0646671
Skewness	0.7293079
Range	0.486
Minimum	0.385
Maximum	0.871
Sum	280.6757
Count	506

AVG_ROOM	
Mean	6.2846344
Standard Error	0.0312351
Median	6.2085
Mode	5.713
Standard Deviation	0.7026171
Sample Variance	0.4936709
Kurtosis	1.8915004
Skewness	0.4036121
Range	5.219
Minimum	3.561
Maximum	8.78
Sum	3180.025
Count	506

LSTAT	
Mean	12.653063
Standard Error	0.3174589
Median	11.36
Mode	8.05
Standard Deviation	7.1410615
Sample Variance	50.99476
Kurtosis	0.4932395
Skewness	0.9064601
Range	36.24
Minimum	1.73
Maximum	37.97
Sum	6402.45
Count	506

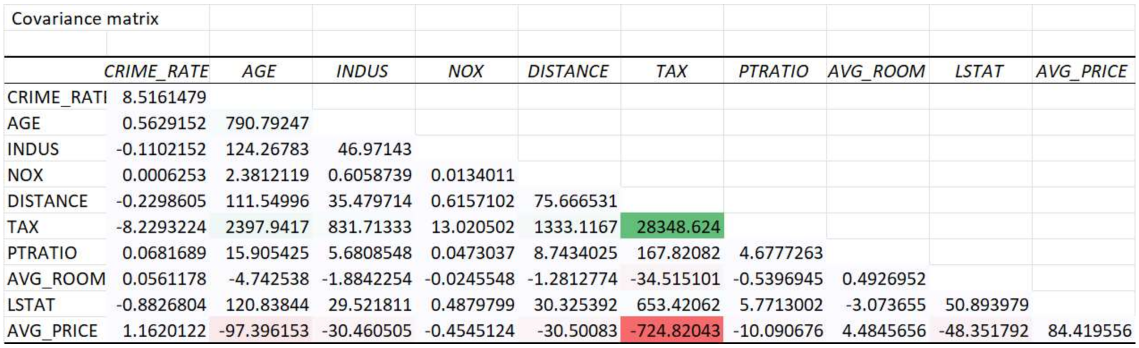
Q2 .Plot the histogram of the Avg_Price Variable. What do you infer?



As clearly we can observe from the figure that it is skewed towards right and have a positive kurtosis (that means it has values at extreme ends).

We can observe that Avg_price of the most properties are nearly towards mean but there are some extreme/large values of the properties.

Q3. Compute the covariance matrix. Share your observations.

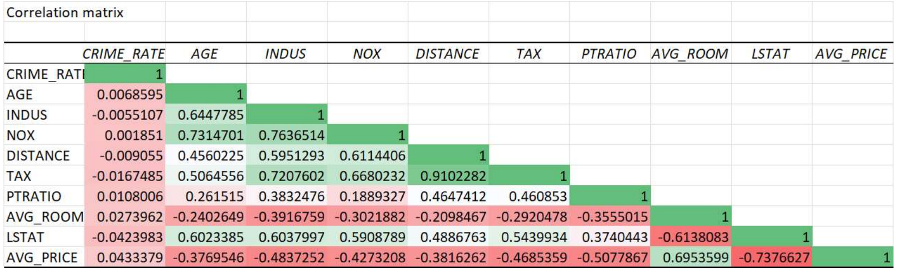


Covariance matrix

	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RATI	8.5161479									
AGE	0.5629152	790.79247								
INDUS	-0.1102152	124.26783	46.97143							
NOX	0.0006253	2.3812119	0.6058739	0.0134011						
DISTANCE	-0.2298605	111.54996	35.479714	0.6157102	75.666531					
TAX	-8.2293224	2397.9417	831.71333	13.020502	1333.1167	28348.624				
PTRATIO	0.0681689	15.905425	5.6808548	0.0473037	8.7434025	167.82082	4.6777263			
AVG_ROOM	0.0561178	-4.742538	-1.8842254	-0.0245548	-1.2812774	-34.515101	-0.5396945	0.4926952		
LSTAT	-0.8826804	120.83844	29.521811	0.4879799	30.325392	653.42062	5.7713002	-3.073655	50.893979	
AVG_PRICE	1.1620122	-97.396153	-30.460505	-0.4545124	-30.50083	-724.82043	-10.090676	4.4845656	-48.351792	84.419556

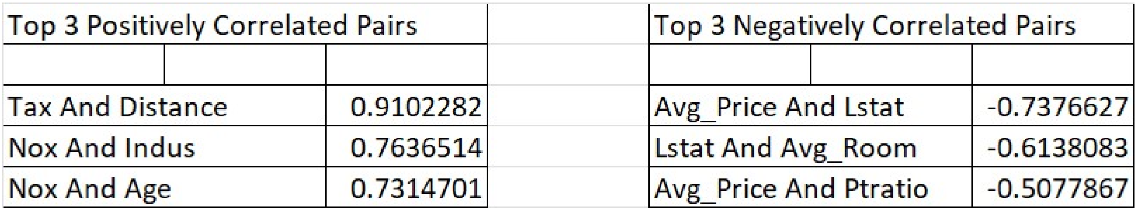
Covariance matrix is used to describe the relationship between the variables, either the variable's follow a positive relationship or negative. But problem with covariance matrix is that it gives random numbers between negative infinity to positive infinity, so we can't have a comparison between variable. So here correlation matrix comes into the picture, this matrix comprises of values between -1 to +1. So, it is ease to compare.

Q4. 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.



Correlation matrix

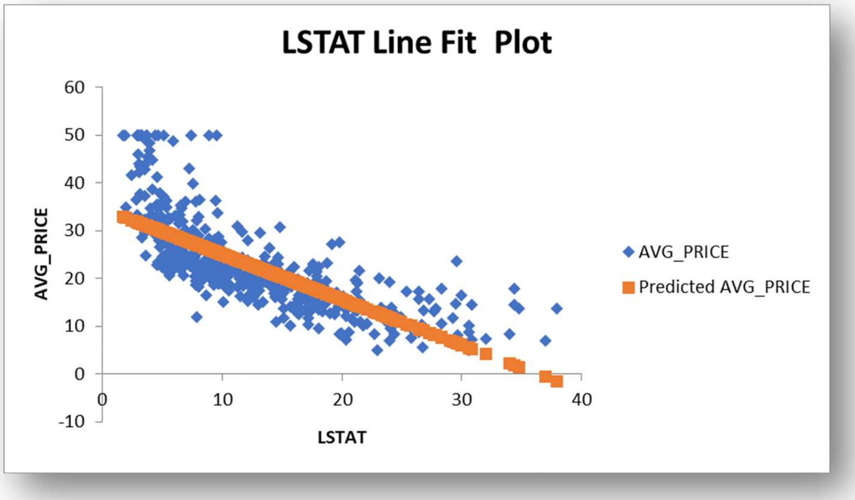
	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RATI	1									
AGE	0.0068595	1								
INDUS	-0.0055107	0.6447785	1							
NOX	0.001851	0.7314701	0.7636514	1						
DISTANCE	-0.009055	0.4560225	0.5951293	0.6114406	1					
TAX	-0.0167485	0.5064556	0.7207602	0.6680232	0.9102282	1				
PTRATIO	0.0108006	0.261515	0.3832476	0.1889327	0.4647412	0.460853	1			
AVG_ROOM	0.0273962	-0.2402649	-0.3916759	-0.3021882	-0.2098467	-0.2920478	-0.3555015	1		
LSTAT	-0.0423983	0.6023385	0.6037997	0.5908789	0.4886763	0.5439934	0.3740443	-0.6138083	1	
AVG_PRICE	0.0433379	-0.3769546	-0.4837252	-0.4273208	-0.3816262	-0.4685359	-0.5077867	0.6953599	-0.7376627	1

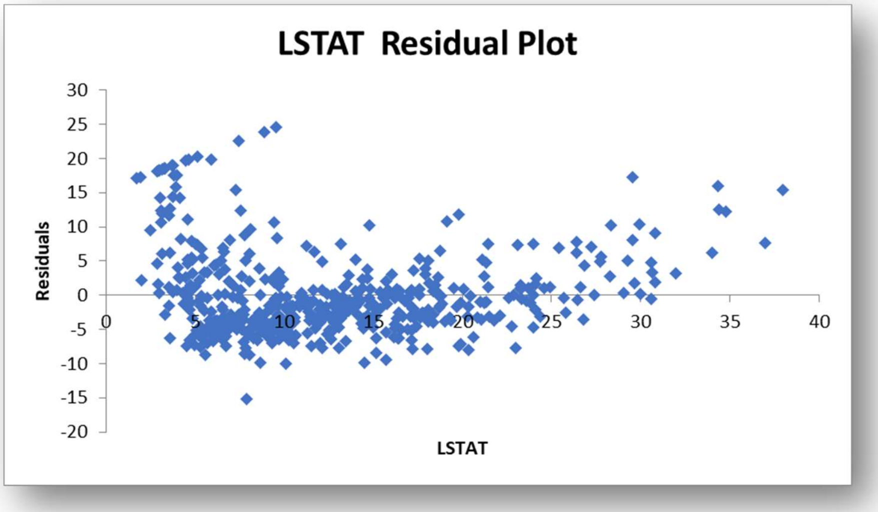


Top 3 Positively Correlated Pairs		
Tax And Distance		0.9102282
Nox And Indus		0.7636514
Nox And Age		0.7314701

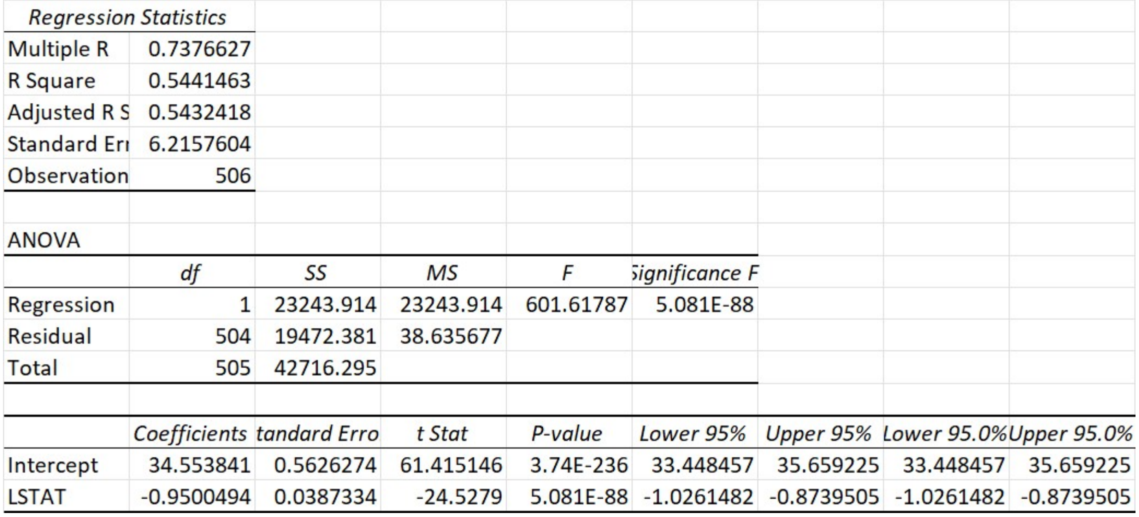
Top 3 Negatively Correlated Pairs		
Avg_Price And Lstat		-0.7376627
Lstat And Avg_Room		-0.6138083
Avg_Price And Ptratio		-0.5077867

5. Build an initial regression model with AVG_PRICE as the y or the Dependent variable and LSTAT variable 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?**

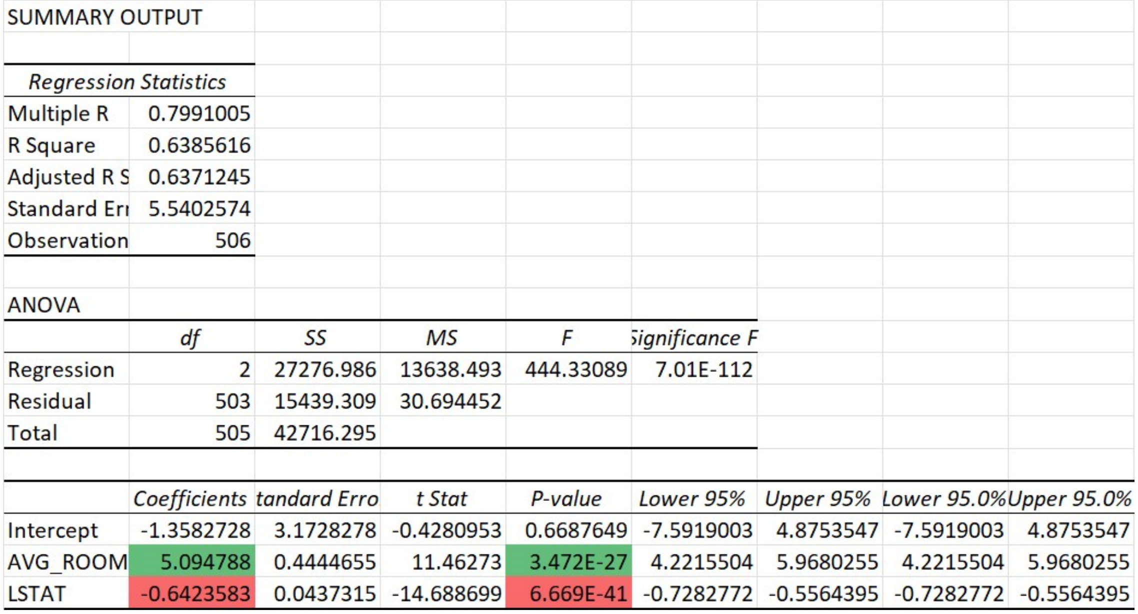


Regression Statistics								
Multiple R	0.7376627							
R Square	0.5441463							
Adjusted R S	0.5432418							
Standard Err	6.2157604							
Observation	506							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	23243.914	23243.914	601.61787	5.081E-88			
Residual	504	19472.381	38.635677					
Total	505	42716.295						
	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	34.553841	0.5626274	61.415146	3.74E-236	33.448457	35.659225	33.448457	35.659225
LSTAT	-0.9500494	0.0387334	-24.5279	5.081E-88	-1.0261482	-0.8739505	-1.0261482	-0.8739505

B. Is LSTAT variable significant for the analysis based on your model?

Lstat is for sure a important variable for analysis for building the model.

Q6. 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.



SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.7991005							
R Square	0.6385616							
Adjusted R S	0.6371245							
Standard Err	5.5402574							
Observation	506							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	2	27276.986	13638.493	444.33089	7.01E-112			
Residual	503	15439.309	30.694452					
Total	505	42716.295						
	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-1.3582728	3.1728278	-0.4280953	0.6687649	-7.5919003	4.8753547	-7.5919003	4.8753547
AVG_ROOM	5.094788	0.4444655	11.46273	3.472E-27	4.2215504	5.9680255	4.2215504	5.9680255
LSTAT	-0.6423583	0.0437315	-14.688699	6.669E-41	-0.7282772	-0.5564395	-0.7282772	-0.5564395

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?

Linear equation is -> $y=m_1x_1+m_2x_2+c$

Y= dependent variable (avg_price)

x1= independent variable 1 (avg_room)

x2= independent variable 2 (Lstat)

m1= coeff. independent variable 1 (avg_room)

m2= coeff. independent variable 2 (Lstat)

$y=m_1x_1+m_2x_2+c$

y=(5.095)*(7)+(-0.6423)*(20)+(-1.35)

y= 21.469

According to regression formula 21.469 is the predicted price and company is quoting 30.000 as the price for the locality. Company is quoting overcharging the price of the location.

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.

Adjusted R^2 is a corrected goodness-of-fit (model accuracy) measure for linear models. Previous model has Adjusted R^2 (0.5432418) and Current model has Adjusted R^2 (0.6371245). So defiantly the model accuracy is increased.

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.

	Coefficients
Intercept	29.241315
CRIME_RATE	0.0487251
AGE	0.0327707
INDUS	0.1305514
NOX	-10.321183
DISTANCE	0.2610936
TAX	-0.0144012
PTRATIO	-1.0743053
AVG_ROOM	4.1254092
LSTAT	-0.6034866

Regression Statistics	
Multiple R	0.8329788
R Square	0.6938537
Adjusted R Square	0.6882986
Standard Error	5.1347635
Observations	506

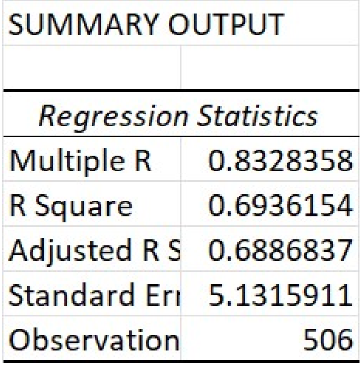
Adjusted R square of this model is 0.688 which is higher then all the previous model, this model definitely increases a little bit. But adding so many variables and accuracy of the model is not that much increase, and this will unnecessarily increase the complexity of the model.

And from seeing weights from coefficients table we can observe that AVG_ROOM and DISTANCE, these variables has more impact on the equation (so these variable are more important).

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)

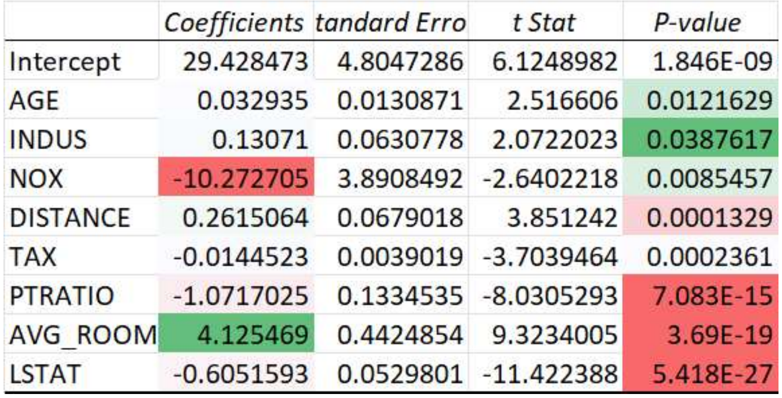
Answer the questions below:

a. Interpret the output of this model.



SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.8328358
R Square	0.6936154
Adjusted R S	0.6886837
Standard Err	5.1315911
Observation	506

Based on p-values below are more appropriate variables for making a regression equations.



	Coefficients	*tandard Erro*	*t Stat*	*P-value*
Intercept	29.428473	4.8047286	6.1248982	1.846E-09
AGE	0.032935	0.0130871	2.516606	0.0121629
INDUS	0.13071	0.0630778	2.0722023	0.0387617
NOX	-10.272705	3.8908492	-2.6402218	0.0085457
DISTANCE	0.2615064	0.0679018	3.851242	0.0001329
TAX	-0.0144523	0.0039019	-3.7039464	0.0002361
PTRATIO	-1.0717025	0.1334535	-8.0305293	7.083E-15
AVG_ROOM	4.125469	0.4424854	9.3234005	3.69E-19
LSTAT	-0.6051593	0.0529801	-11.422388	5.418E-27

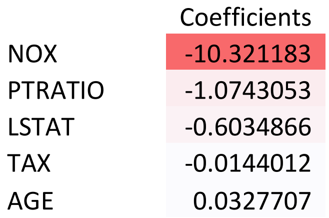
b. Compare the adjusted R-square value of this model with the model in the previous question, which model performs better according to the value of adjusted R-square?

The adjusted r-square is slightest high then previous model, the value is not that significant.

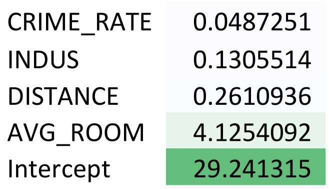
Current model r- square -> 0.6886837

Previous model r – square -> 0.6882986

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?



	Coefficients
NOX	-10.321183
PTRATIO	-1.0743053
LSTAT	-0.6034866
TAX	-0.0144012
AGE	0.0327707



CRIME_RATE	0.0487251
INDUS	0.1305514
DISTANCE	0.2610936
AVG_ROOM	4.1254092
Intercept	29.241315

The weight/coefficient of Nox is -10.3, so if Nox is in more in a locality then the price of the house is less than average because they have a negative relationship.

d. Write the regression equation from this model.

Y=m1x1+ m2x2+ m3x3+ m4x4+ m5x5+ m6x6+ m7x7+ m8x8+ m9x9+ Intercept

Y=(-10.3)*x1+(-1.07)*x2+(0.603)*x3+(0.014)*x4+(0.03)*x5+(0.048)*x6+(0.130)*x7+(0.261)*x8+(4.125)*x9+29.241

Y= dependent variable

X1-9 = independent variables