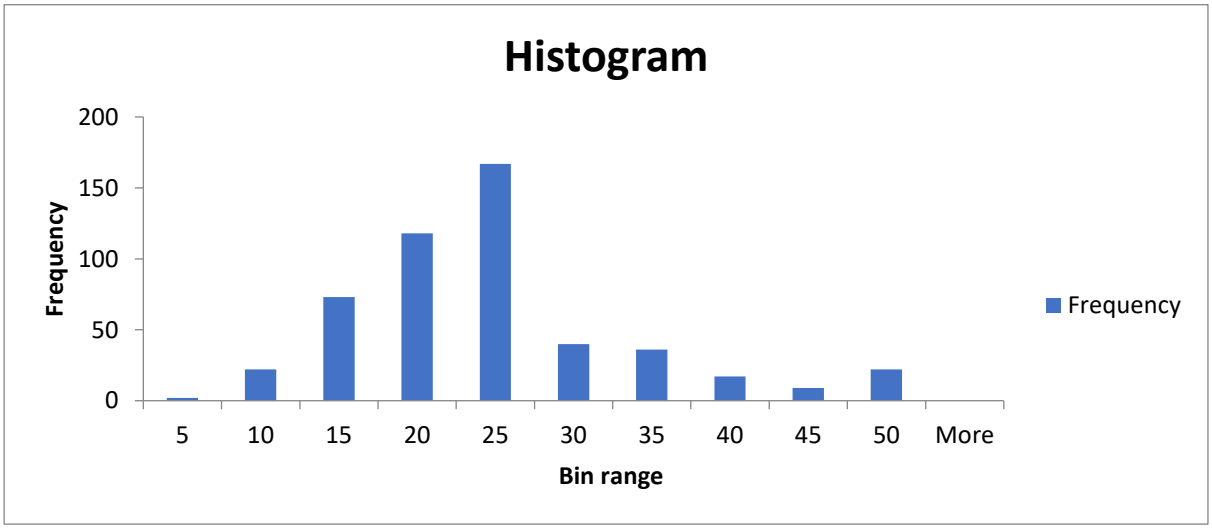
Business Report

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?

CRIME_RATE		AGE		INDUS		NOX		DISTANCE		TAX		PTRATIO		AVG_ROOM		LSTAT		AVG_PRICE	
Mean	4.87	Mean	68.57	Mean	11.14	Mean	0.55	Mean	9.55	Mean	408.24	Mean	18.46	Mean	6.28	Mean	12.65	Mean	22.53
Standard Error	0.13	Standard Error	1.25	Standard Error	0.30	Standard Error	0.01	Standard Error	0.39	Standard Error	7.49	Standard Error	0.10	Standard Error	0.03	Standard Error	0.32	Standard Error	0.41
Median	4.82	Median	77.50	Median	9.69	Median	0.54	Median	5.00	Median	330.00	Median	19.05	Median	6.21	Median	11.36	Median	21.20
Mode	3.43	Mode	100.00	Mode	18.10	Mode	0.54	Mode	24.00	Mode	666.00	Mode	20.20	Mode	5.71	Mode	8.05	Mode	50.00
Standard Deviation	2.92	Standard Deviation	28.15	Standard Deviation	6.86	Standard Deviation	0.12	Standard Deviation	8.71	Standard Deviation	168.54	Standard Deviation	2.16	Standard Deviation	0.70	Standard Deviation	7.14	Standard Deviation	9.20
Sample Variance	8.53	Sample Variance	792.36	Sample Variance	47.06	Sample Variance	0.01	Sample Variance	75.82	Sample Variance	28404.76	Sample Variance	4.69	Sample Variance	0.49	Sample Variance	50.99	Sample Variance	84.59
Kurtosis	-1.19	Kurtosis	-0.97	Kurtosis	-1.23	Kurtosis	-0.06	Kurtosis	-0.87	Kurtosis	-1.14	Kurtosis	-0.29	Kurtosis	1.89	Kurtosis	0.49	Kurtosis	1.50
Skewness	0.02	Skewness	-0.60	Skewness	0.30	Skewness	0.73	Skewness	1.00	Skewness	0.67	Skewness	-0.80	Skewness	0.40	Skewness	0.91	Skewness	1.11
Range	9.95	Range	97.10	Range	27.28	Range	0.49	Range	23.00	Range	524.00	Range	9.40	Range	5.22	Range	36.24	Range	45.00
Minimum	0.04	Minimum	2.90	Minimum	0.46	Minimum	0.39	Minimum	1.00	Minimum	187.00	Minimum	12.60	Minimum	3.56	Minimum	1.73	Minimum	5.00
Maximum	9.99	Maximum	100.00	Maximum	27.74	Maximum	0.87	Maximum	24.00	Maximum	711.00	Maximum	22.00	Maximum	8.78	Maximum	37.97	Maximum	50.00
Sum	2465.22	Sum	34698.90	Sum	5635.21	Sum	280.68	Sum	4832.00	Sum	206568.00	Sum	9338.50	Sum	3180.03	Sum	6402.45	Sum	11401.60
Count	506.00	Count	506.00	Count	506.00	Count	506.00	Count	506.00	Count	506.00	Count	506.00	Count	506.00	Count	506.00	Count	506.00

By observing above summary statistics, the average price of a flat is about 22.53(amount). The positive reason to purchase a flat is Pupil and Teacher ratio which is in good range (9.40 range of PTRATIO) that can attract a greater number of people to buy flats in that region. One more positive reason is that the average no of rooms 6.28 which is almost 6 rooms in a flat this can also attract the buyers. Some people want to buy flats near highway and the average distance from highway is 9.55 which is approximately 10 miles. But there are some negative reasons like the average crime rate 4.87, average tax 408.24 and average age of buildings 68.57.

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



The price of each house starts from $5000 to $50000 and the average price was $22000 and by the histogram we can analyze that there are more houses in a price range of $20000 to $25000. The histogram is a “right skewed histogram.” The Avg_price is affected by the other variables since the Avg_price is the dependent variable for all the other variables in the table. The other variables like tax, crime rate, nox, avg_room etc., will affect the Avg_price.

Example: - If crime rate and nox is high the price will be low and if the rooms are more the price will be high.

Question 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	790.7925								
INDUS	-0.110215175	124.2678	46.97143							
NOX	0.000625308	2.381212	0.605874	0.013401						
DISTANCE	-0.229860488	111.55	35.47971	0.61571	75.66653					
TAX	-8.229322439	2397.942	831.7133	13.0205	1333.117	28348.62				
PTRATIO	0.068168906	15.90543	5.680855	0.047304	8.743402	167.8208	4.677726296			
AVG_ROOM	0.056117778	-4.74254	-1.88423	-0.02455	-1.28128	-34.5151	-0.539694518	0.492695216		
LSTAT	-0.882680362	120.8384	29.52181	0.48798	30.32539	653.4206	5.771300243	-3.073654967	50.89397935	
AVG_PRICE	1.16201224	-97.3962	-30.4605	-0.45451	-30.5008	-724.82	-10.09067561	4.484565552	-48.35179219	84.41955616

Covariance is a measure of the relationship between two random variables where it describes up to what extent they change together. In simple words covariance describes about the direction, and if the value is positive integer, then the variables move in the same direction, or if the value is negative integer then the variables move in inverse direction. By analysing above covariance matrix Avg_price and tax have a negative relationship where as Avg_price and Avg_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_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RATE	1									
AGE	0.006859463	1								
INDUS	-0.005510651	0.644779	1							
NOX	0.001850982	0.73147	0.763651	1						
DISTANCE	-0.009055049	0.456022	0.595129	0.611441	1					
TAX	-0.016748522	0.506456	0.72076	0.668023	0.910228	1				
PTRATIO	0.010800586	0.261515	0.383248	0.188933	0.464741	0.460853	1			
AVG_ROOM	0.02739616	-0.24026	-0.39168	-0.30219	-0.20985	-0.29205	-0.3555	1		
LSTAT	-0.042398321	0.602339	0.6038	0.590879	0.488676	0.543993	0.374044	-0.613808272	1	
AVG_PRICE	0.043337871	-0.37695	-0.48373	-0.42732	-0.38163	-0.46854	-0.50779	0.695359947	-0.73766	1

A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. A correlation matrix is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses. Top three positively corelated pairs are 0.9102(Tax & Distance), 0.7636(Nox & Indus) and 0.7314(Nox & Age) and top three negatively corelated 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

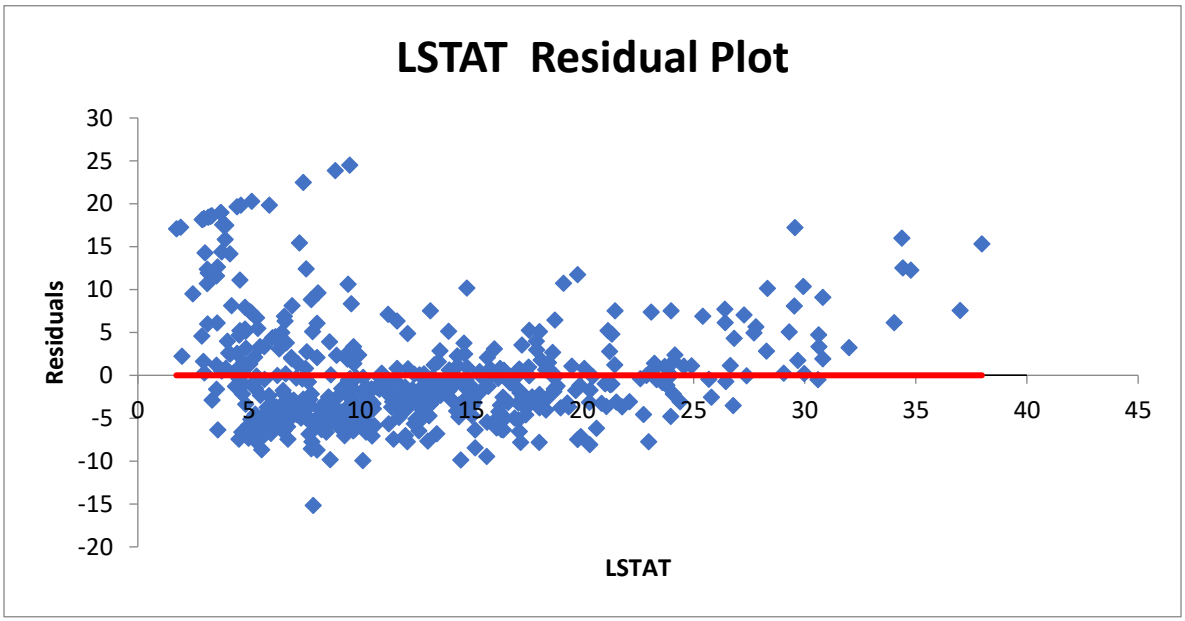
Regression Statistics	
Multiple R	0.737663
R Square	0.544146
Adjusted R	0.543242
Standard E	6.21576
Observatic	506

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 Erı*	*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

Regression model



a.) By observing the regression summary output, we know that if the coeffient value is positive and also it is increasing the variance will also increases but if coeffient value is negative and it is increasing the variance will decreases, and we can analyze there is some pattern in the trendline where it is a straight line. By residuals, we know that if the value is less than 0.05 then it is significant.

b.) LSTAT is significant because the p-value is less than 0.05.

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

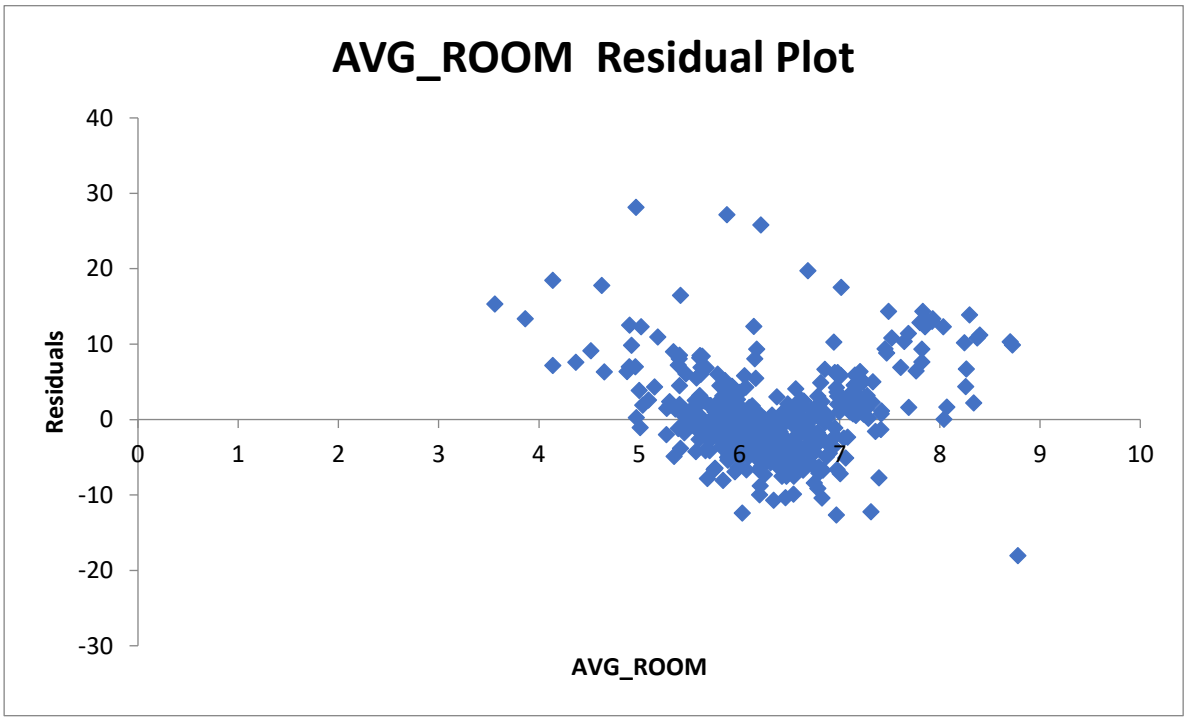
Regression Statistics	
Multiple R	0.7991
R Square	0.6386
Adjusted R	0.6371
Standard E	5.5403
Observatic	506

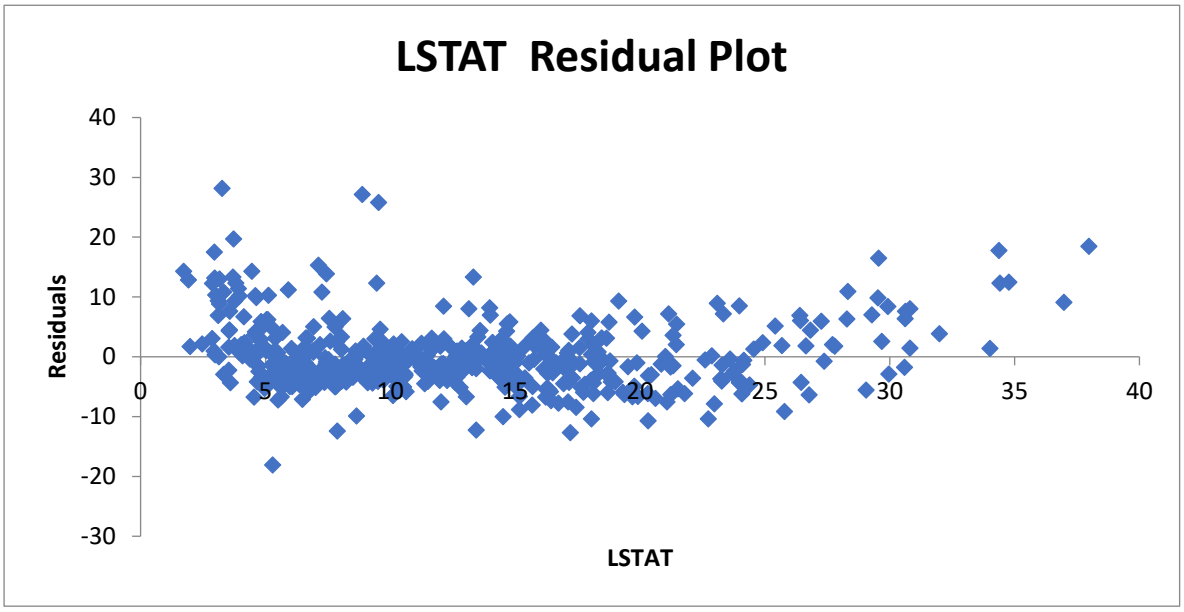
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regressior	2	27276.99	13638.49	444.3308922	7.0085E-112
Residual	503	15439.31	30.69445		
Total	505	42716.3			

	Coefficient	*tandard Erı*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-1.358	3.172828	-0.4281	0.668764941	-7.591900282	4.875355	-7.5919	4.875355
AVG_ROO	5.0948	0.444466	11.46273	3.47226E-27	4.221550436	5.968026	4.22155	5.968026
LSTAT	-0.642	0.043731	-14.6887	6.66937E-41	-0.728277167	-0.55644	-0.72828	-0.55644

Regression model with Avg_rooms and LSTAT





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

b.) Previous model has an adjusted R square of 0.543 and for this model we have an adjusted R square of 0.637, By these values we can state that this regression has high performances rather than the previous model. Because we know that if adjusted R square is higher the model works better, and if the adjusted R square is lower the model does not work better.

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	0.832978824
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

Here, we know that the value of R square and adjusted R square indicates the performances of the model i.e., "69.3%" here the regression coeffient is used to describe the relation between an independent variable and dependent variable. Also, most of the variables have perfectly positive linear relationship with Avg_price. Variables like Nox, tax, ptratio, LSTAT have a perfectly negative linear relationship with Avg_price. By considering the P-Values we can state except the Crime_Rate (0.53) remaining all variables are significant because we know that if P-value is Less than 0.05 i.e., 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	0.832835773
R Square	0.693615426
Adjusted R Square	0.688683682
Standard Error	5.131591113
Observations	506

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regression	8	29628.68	3703.585	140.643	1.9E-122
Residual	497	13087.61	26.33323		
Total	505	42716.3			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	29.42847349	4.804729	6.124898	1.85E-09	19.98839	38.86856	19.98839	38.86856
AGE	0.03293496	0.013087	2.516606	0.012163	0.007222	0.058648	0.007222	0.058648
INDUS	0.130710007	0.063078	2.072202	0.038762	0.006778	0.254642	0.006778	0.254642
NOX	-10.27270508	3.890849	-2.64022	0.008546	-17.9172	-2.62816	-17.9172	-2.62816
DISTANCE	0.261506423	0.067902	3.851242	0.000133	0.128096	0.394916	0.128096	0.394916
TAX	-0.014452345	0.003902	-3.70395	0.000236	-0.02212	-0.00679	-0.02212	-0.00679
PTRATIO	-1.071702473	0.133454	-8.03053	7.08E-15	-1.33391	-0.8095	-1.33391	-0.8095
AVG_ROOM	4.125468959	0.442485	9.3234	3.69E-19	3.256096	4.994842	3.256096	4.994842
LSTAT	-0.605159282	0.05298	-11.4224	5.42E-27	-0.70925	-0.50107	-0.70925	-0.50107

a.) The above picture represents regression statistics of the significant variables only where the p-value of variables is less than 0.05.

b.) We can observe that the current model (0.6886) gives a slight better performance compare with the previous model (0.6882) because it is slightly greater than the previous model and we know that the higher the adjusted R square gives the greater performances.

c.) A positive coefficient means when the value of independent variables decreases the mean of the dependent variables increases and negative coefficient means as the values of the independent variables increases the mean of the dependent variables decreases, after sorting the value of nox increases avg_price decreases. In other words NOX(pollution) increases the Avg_price decreases because of the pollution.

d.) The regression 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.