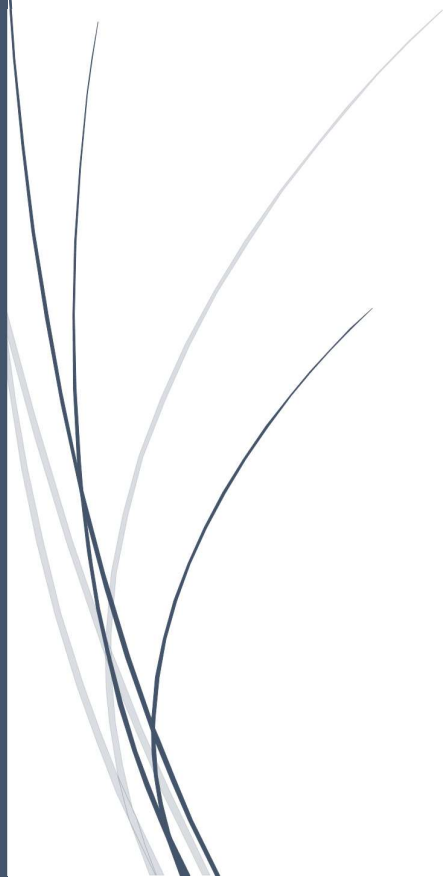
28/12/2023

TERRO'S REAL ESTATE AGENCY

PROJECT REPORT



PROJECT REPORT – TERRO'S REAL ESTATE AGENCY

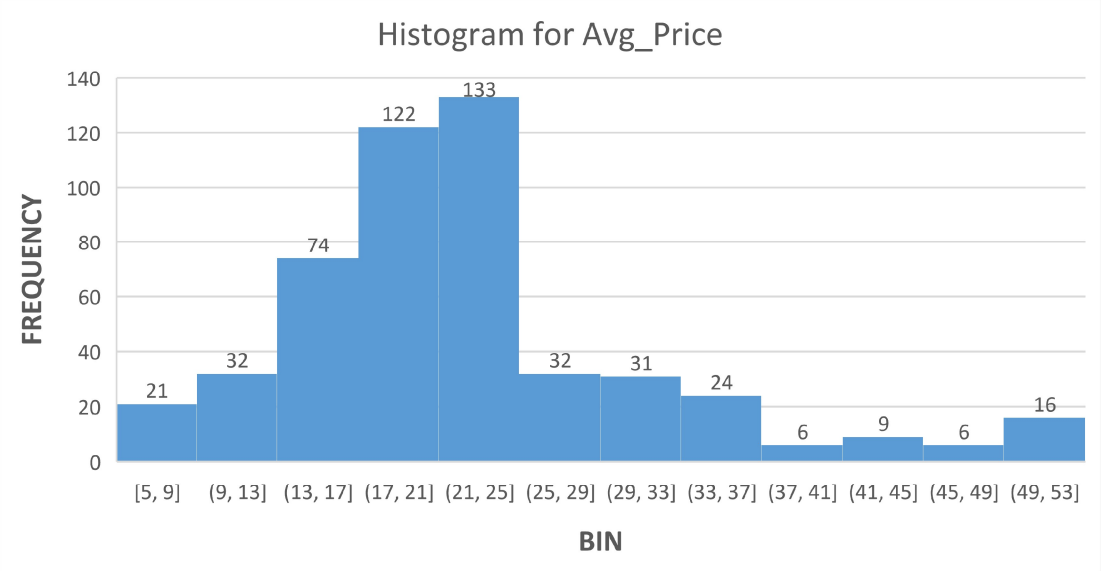
1) Generate the summary statistics for each variable in the table. (Use Data analysis tool pack). Write down your observation.

CRIME_RATE		AGE		INDUS		NOX		DISTANCE	
Mean	4.87197628	Mean	68.574901	Mean	11.1367787	Mean	0.55469506	Mean	9.5494071
Standard Error	0.12986015	Standard Error	1.2513695	Standard Error	0.30497989	Standard Error	0.00515139	Standard Error	0.3870849
Median	4.82	Median	77.5	Median	9.69	Median	0.538	Median	5
Mode	3.43	Mode	100	Mode	18.1	Mode	0.538	Mode	24
Standard Deviatior	2.92113189	Standard Deviatior	28.148861	Standard Deviatior	6.86035294	Standard Deviatio	0.11587768	Standard Deviatio	8.7072594
Sample Variance	8.53301153	Sample Variance	792.3584	Sample Variance	47.0644425	Sample Variance	0.01342764	Sample Variance	75.816366
Kurtosis	-1.1891225	Kurtosis	-0.967716	Kurtosis	-1.2335396	Kurtosis	-0.06466713	Kurtosis	-0.867232
Skewness	0.02172808	Skewness	-0.598963	Skewness	0.29502157	Skewness	0.72930792	Skewness	1.0048146
Range	9.95	Range	97.1	Range	27.28	Range	0.486	Range	23
Minimum	0.04	Minimum	2.9	Minimum	0.46	Minimum	0.385	Minimum	1
Maximum	9.99	Maximum	100	Maximum	27.74	Maximum	0.871	Maximum	24
Sum	2465.22	Sum	34698.9	Sum	5635.21	Sum	280.6757	Sum	4832
Count	506	Count	506	Count	506	Count	506	Count	506

PTRATIO		AVG_ROOM		LSTAT		AVG_PRICE		TAX	
Mean	18.4555336	Mean	6.2846344	Mean	12.6530632	Mean	22.5328063	Mean	408.23715
Standard Error	0.09624357	Standard Error	0.0312351	Standard Error	0.31745891	Standard Error	0.40886115	Standard Error	7.4923887
Median	19.05	Median	6.2085	Median	11.36	Median	21.2	Median	330
Mode	20.2	Mode	5.713	Mode	8.05	Mode	50	Mode	666
Standard Deviatior	2.16494552	Standard Deviatior	0.7026171	Standard Deviatior	7.14106151	Standard Deviatio	9.19710409	Standard Deviatio	168.53712
Sample Variance	4.68698912	Sample Variance	0.4936709	Sample Variance	50.9947595	Sample Variance	84.5867236	Sample Variance	28404.759
Kurtosis	-0.2850914	Kurtosis	1.8915004	Kurtosis	0.49323952	Kurtosis	1.49519694	Kurtosis	-1.142408
Skewness	-0.8023249	Skewness	0.4036121	Skewness	0.90646009	Skewness	1.10809841	Skewness	0.6699559
Range	9.4	Range	5.219	Range	36.24	Range	45	Range	524
Minimum	12.6	Minimum	3.561	Minimum	1.73	Minimum	5	Minimum	187
Maximum	22	Maximum	8.78	Maximum	37.97	Maximum	50	Maximum	711
Sum	9338.5	Sum	3180.025	Sum	6402.45	Sum	11401.6	Sum	206568
Count	506	Count	506	Count	506	Count	506	Count	506

- Mean has large amount of deviation in Tax.
- Mean has large amount of deviation from median in Age and Tax.
- Indus has a flat curve than other datasets such as Tax or Crime rate. And it is flat as per kurtosis measures.
- Positive and large skewed value in tax and distance is seen in the dataset.
- The range value is the maximum for tax.
- The range value is the maximum for tax.

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



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

From the Histogram Plot we can observe the shape of the distribution of AVG_PRICE variable.

- 20-25 has the maximum frequency in the Average Price variable.
- 17-21 has the second maximum frequency.
- Most houses have average price in the range of $17K-$25K.
- Shows that the distribution has a sharp peak.
- Positive tail skewed dataset is seen.
- The Average Price variable has positive Kurtosis.
- 37-41 and 45-49 has the minimum frequency in this dataset.

3) Compute the covariance matrix. Share your observations.

Column1	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	0.013401099						
DISTANCE	-0.229860488	111.5499555	35.47971449	0.615710224	75.66653127					
TAX	-8.229322439	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	-1.884225427	-0.024554826	-1.281277391	-34.51510104	-0.539694518	0.492695216		
LSTAT	-0.882680362	120.8384405	29.52181125	0.487979871	30.32539213	653.4206174	5.771300243	-3.073654967	50.89397935	
AVG_PRICE	1.16201224	-97.39615288	-30.46050499	-0.454512407	-30.50083035	-724.8204284	-10.09067561	4.484565552	-48.35179219	84.41955616

PROJECT REPORT – TERRO'S REAL ESTATE AGENCY

Covariance basically signifies the direction of linear relationship between two variables. The direction usually refers to whether the variables vary directly or inversely to each other.

- Average price has positive covariance only with crime rate and average room.
- Average room has only 2 positive covariance with average price and crime rate.
- Lstat has negative covariance only with crime rate and average room.
- Age has negative covariance only with average room and average price.
- Nox and PTRatio has all positive covariance except with average room and average price.
- Indus, Distance and Tax has negative covariance with crime rate, average room and average price.

4) Create a correlation matrix of all the variables (Use Data analysis tool pack). (5 marks) 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	-0.737662726	1

a) Green coloured cells are top 3 positively correlated pairs:

1. TAX and DISTANCE (91%).
2. NOX and INDUS (76%).
3. NOX and AGE (73%).

b) Red coloured cells are top 3 negatively correlated pairs:

1. AVG_PRICE and LSTAT (-74%).
2. LSTAT and AVG_PRICE (-61%).
3. AVG_PRICE and PTRATIO (-51%).

5) Build an initial regression model with AVG_PRICE as 'y' (Dependent variable) and LSTAT variable as Independent Variable. Generate the residual plot. (8 marks)

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?

PROJECT REPORT – TERRO'S REAL ESTATE AGENCY

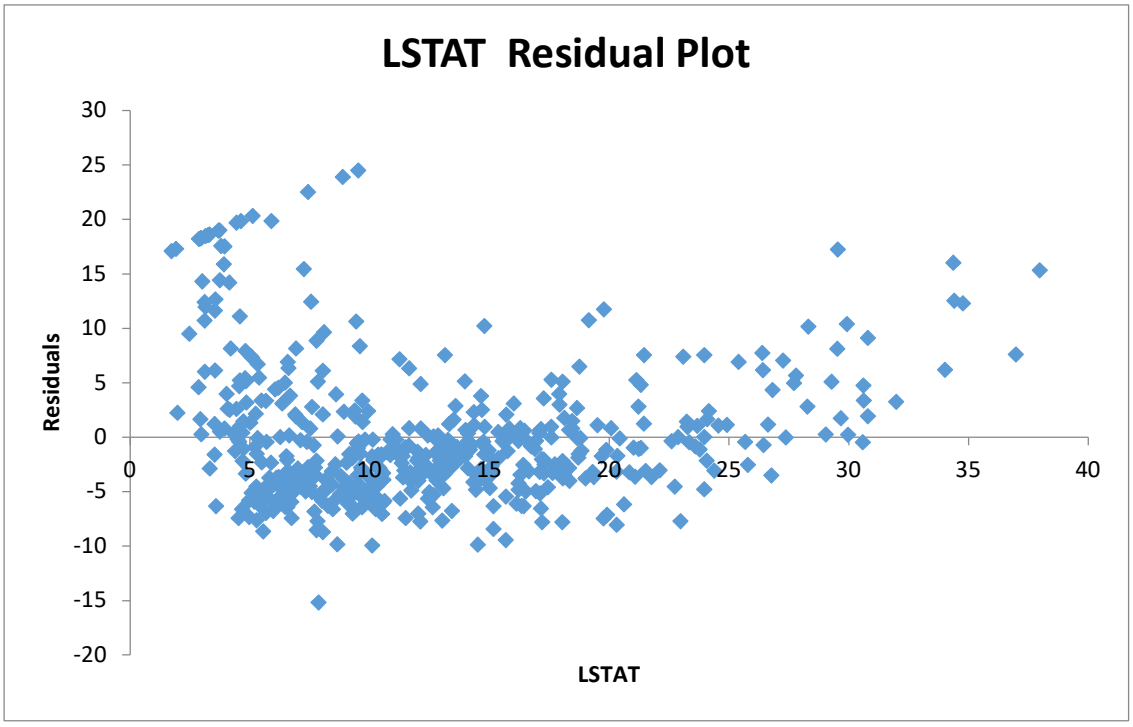
SUMMARY OUTPUT

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	35.6592247
LSTAT	-0.950049354	0.038733416	-24.52789985	5.0811E-88	-1.0261482	-0.873950508	-1.0261482	-0.87395051



a) **i.** From the Regression summary output, it can be inferred that R Square value is 0.54 Which means that 54% of variance in dependent variable (AVG_PRICE) is explained by the independent variable (LSTAT) in this model.

ii. The coefficient value of LSTAT is -0.95 Which means that with increase in value of LSTAT by 1 there is decrease in value of AVG_PRICE by 0.95.

iii. The intercept value is 34.55 Which means that When LSTAT is zero, the value of AVG_PRICE is 34.55K USD.

iv. The residual plot indicates that the residual data points are randomly distributed around the X axis and are not following any particular pattern. So, the linear model is an appropriate model.

b) Yes, the LSTAT variable is significant for analysis in this model. The p value is 5.08E-88 which is lower than 0.05, thus eliminating the null hypothesis for this variable.

PROJECT REPORT – TERRO'S REAL ESTATE AGENCY

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	0.799100498
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	-7.59190028	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	-0.72827717	-0.5564395

The p-value of intercept is 0.6687 which is greater than 0.05. so, it is not significant for this model. So we can eliminate the intercept value from this model.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.973885353
R Square	0.948452681
Adjusted R Square	0.946366278
Standard Error	5.53576654
Observations	506

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	284181.4056	142090.7028	4636.712087	0
Residual	504	15444.93444	30.64471119		
Total	506	299626.34			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
AVG_ROOM	4.906906071	0.070193339	69.90557997	1.6137E-261	4.768998482	5.044813661	4.768998482	5.044813661
LSTAT	-0.655739993	0.030558561	-21.45847115	4.81185E-73	-0.715777847	-0.595702138	-0.715777847	-0.595702138

PROJECT REPORT – TERRO'S REAL ESTATE AGENCY

a) From the Regression summary output, it can be inferred that the regression equation is

Avg_Price = 4.9069 * Avg_Room - 0.6557 * LSTAT

If a new house in this locality has 7 rooms on an average and L-STAT value is 20 means, then the value of AVG_PRICE will be

AVG_ROOM = 7

LSTAT = 20

AVG_PRICE = 4.9069*7 – 0.6557*20 = **21.2343K** or **21234 USD**

If the company is quoting a value of 30000 USD for this locality. Yes, the company is **overcharging**, the AVG_PRICE of this house is 21234 USD as per the model.

b) Yes, the performance of this model is better than the previous model, the adjusted R Square value of this model is 0.63 or 63% which is higher than 54% obtained in the previous model as this model can explain 63% of variance in AVG_PRICE by the independent variables AVG_ROOM and LSTAT. Also, the p value of both the variables are less than 0.05 and are significant variables for this model.

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	0.832978824
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

- From the Regression summary output, it can be inferred that the Adjusted R Square value is 0.6882 or 68.82% which is higher than the previous two regression models. Thus 68% of variance in AVG_PRICE is explained by the independent variables in this model.
- The intercept value is 29.24 which means that the AVG_PRICE is 29.24 USD when the value of all the independent variables is zero.

- It is observed that the coefficients of CRIME_RATE, AGE, INDUS, DISTANCE and AVG_ROOM are positive so whenever there is increase in value of these variables, the AVG_PRICE also increases accordingly. On the other hand, it is observed that the coefficients of NOX, TAX, PTRATIO and LSTAT are negative and hence the AVG_PRICE decreases with increase in the value of these variables.
- It is observed from the p value of variables, that the p value of CRIME_RATE is 0.5346 which is greater than 0.05 with respect to AVG_PRICE, so the null hypothesis cannot be rejected for this particular variable and thus becomes an insignificant in the value of these variables.
- All other independent variables such as AGE, INDUS, DISTANCE, NOX, TAX, PTRATIO, LSTAT and AVG_ROOM have a "p value" less than 0.05 with respect to AVG_PRICE. So, this variables are significant and the null hypothesis is rejected.

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	0.832835773
R Square	0.693615426
Adjusted R Square	0.688683682
Standard Error	5.131591113
Observations	506

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	8	29628.68142	3703.585178	140.6430411	1.911E-122
Residual	497	13087.61399	26.33322735		
Total	505	42716.29542			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
NOX	-10.27270508	3.890849222	-2.640221837	0.008545718	-17.9172457	-2.628164466	-17.9172457	-2.62816447
PTRATIO	-1.071702473	0.133453529	-8.030529271	7.08251E-15	-1.333905109	-0.809499836	-1.333905109	-0.80949984
LSTAT	-0.605159282	0.0529801	-11.42238841	5.41844E-27	-0.70925186	-0.501066704	-0.70925186	-0.5010667
TAX	-0.014452345	0.003901877	-3.703946406	0.000236072	-0.022118553	-0.006786137	-0.022118553	-0.00678614
AGE	0.03293496	0.013087055	2.516605952	0.012162875	0.007222187	0.058647734	0.007222187	0.05864773
INDUS	0.130710007	0.063077823	2.072202264	0.038761669	0.006777942	0.254642071	0.006777942	0.25464207
DISTANCE	0.261506423	0.067901841	3.851242024	0.000132887	0.128096375	0.394916471	0.128096375	0.39491647
AVG_ROOM	4.125468959	0.44248544	9.323400461	3.68969E-19	3.256096304	4.994841615	3.256096304	4.99484161
Intercept	29.42847349	4.804728624	6.124898157	1.84597E-09	19.98838959	38.8685574	19.98838959	38.8685574

a) From the output it is inferred that all the variables are significant as the p value of all the independent variables is less than 0.05, hence rejecting the null hypothesis. The Adjusted R Square value is 68.86% which means 68.86 % variance in AVG_PRICE is explained by the independent variables in this model.

b) From the Regression summary output, it can be inferred that the Adjusted R Square value is 0.6886 or 68.86% which is highest when compared to previous regression models. So, this model performs better than previous models.

c) After sorting, the values of coefficients in ascending order are, as follows:

NOX (-10.27) < PTRATIO (-1.07) < LSTAT (-0.60) < TAX (-0.01) < AGE (0.03) < INDUS (0.13) < DISTANCE (0.26) < AVG_PRICE (4.12)

If the value of NOX is more in locality, the NOX is increase by 1 point then the AVG_PRICE decreases by 10.27 points. So, basically the AVG_PRICE of house will be lesser in a locality having higher concentration of nitric oxides (pollutants). More the pollution, leads to lesser be the average price of the property.

d) The Regression equation for this model is:

AVG_PRICE = 0.0329 * AGE + 0.1307 * INDUS - 10.2727 * NOX + 0.2615 * DISTANCE - 0.0144 * TAX 1.0717 * PTRATIO + 4.1254 * AVG_ROOM - 0.6051 * LSTAT + 29.4284