Assignment – 2

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*	
Mean	4.871976	Mean	68.5749	Mean	11.13678
Standard Error	0.12986	Standard Error	1.25137	Standard Error	0.30498
Median	4.82	Median	77.5	Median	9.69
Mode	3.43	Mode	100	Mode	18.1
Standard Deviation	2.921132	Standard Deviation	28.14886	Standard Deviation	6.860353
Sample Variance	8.533012	Sample Variance	792.3584	Sample Variance	47.06444
Kurtosis	-1.18912	Kurtosis	-0.96772	Kurtosis	-1.23354
Skewness	0.021728	Skewness	-0.59896	Skewness	0.295022
Range	9.95	Range	97.1	Range	27.28
Minimum	0.04	Minimum	2.9	Minimum	0.46
Maximum	9.99	Maximum	100	Maximum	27.74
Sum	2465.22	Sum	34698.9	Sum	5635.21
Count	506	Count	506	Count	506

NOX		*DISTANCE*		*TAX*	
Mean	0.554695	Mean	9.549407	Mean	408.2372
Standard Error	0.005151	Standard Error	0.387085	Standard Error	7.492389
Median	0.538	Median	5	Median	330
Mode	0.538	Mode	24	Mode	666
Standard Deviation	0.115878	Standard Deviation	8.707259	Standard Deviation	168.5371
Sample Variance	0.013428	Sample Variance	75.81637	Sample Variance	28404.76
Kurtosis	-0.06467	Kurtosis	-0.86723	Kurtosis	-1.14241
Skewness	0.729308	Skewness	1.004815	Skewness	0.669956
Range	0.486	Range	23	Range	524
Minimum	0.385	Minimum	1	Minimum	187
Maximum	0.871	Maximum	24	Maximum	711
Sum	280.6757	Sum	4832	Sum	206568
Count	506	Count	506	Count	506

PTRATIO		*AVG_ROOM*		*LSTAT*	
Mean	18.45553	Mean	6.284634	Mean	12.65306
Standard Error	0.096244	Standard Error	0.031235	Standard Error	0.317459
Median	19.05	Median	6.2085	Median	11.36
Mode	20.2	Mode	5.713	Mode	8.05
Standard Deviation	2.164946	Standard Deviation	0.702617	Standard Deviation	7.141062
Sample Variance	4.686989	Sample Variance	0.493671	Sample Variance	50.99476
Kurtosis	-0.28509	Kurtosis	1.8915	Kurtosis	0.49324
Skewness	-0.80232	Skewness	0.403612	Skewness	0.90646
Range	9.4	Range	5.219	Range	36.24
Minimum	12.6	Minimum	3.561	Minimum	1.73
Maximum	22	Maximum	8.78	Maximum	37.97
Sum	9338.5	Sum	3180.025	Sum	6402.45
Count	506	Count	506	Count	506

AVG_PRICE	
Mean	22.53281
Standard Error	0.408861
Median	21.2
Mode	50
Standard Deviation	9.197104
Sample Variance	84.58672
Kurtosis	1.495197
Skewness	1.108098
Range	45
Minimum	5
Maximum	50
Sum	11401.6
Count	506

COV	**VALUE**
CRIME_RATE	0.599578
AGE	0.410483
INDUS	0.616009
NOX	0.208903
DISTANCE	0.911812
TAX	0.412841
PTRATIO	0.117306
AVG_ROOM	0.111799
LSTAT	0.564374
AVG_PRICE	0.408165

- From the coefficient of variation, DISTANCE (0.91181) has the highest spread and variables such as CRIME_RATE (0.59958), INDUS (0,61601), LSTAT (0.56437) have spread greater than optimal range (0.2 to 0.5). Variables like PTRATIO (0.11731), AVG_ROOM (0.1118) have spread lower than 0.2 which means the spread is low. AGE (0.41048), NOX (0.2089), TAX (0.41284) and AVG_PRICE (0.40817) has optimal spread that ranges between 0.2 to 0.5.

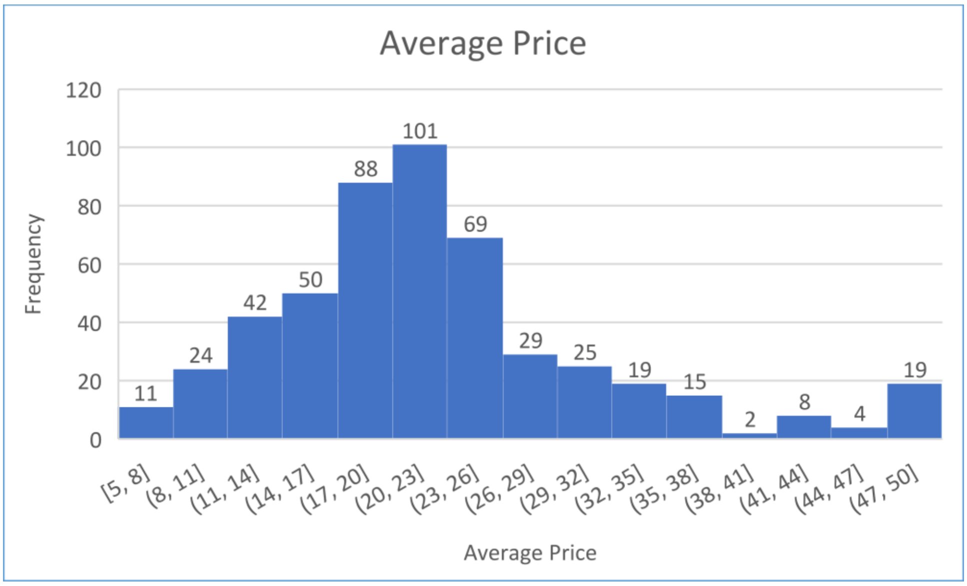
- CRIME_RATE have kurtosis value of -1.18912 which shows that the distribution is platykurtic and is flat. There is no kurtosis value greater than 3 which means that there are no leptokurtic variables in the given data.

- The AGE variable has negative skewness value of -0.59896 and PTRATIO (-0.80232) which shows that the distribution is skewed to the left and more data on the right of the mean value. All other variables are positively skewed and more data are on the left of the mean and skewed to the right.

- The difference between Mean (408.237) and Median (330) of the TAX variable is 78.237 which is very high.

- The median and Mode of NOX is equal which is 0.538.

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

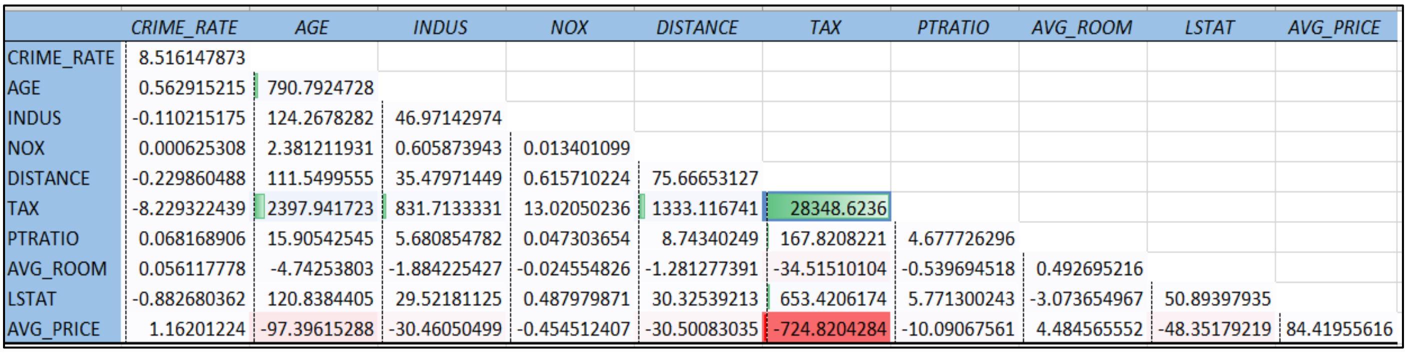


AVG_PRICE	*Values*
Mean	22.532806
Standard Error	0.4088611
Median	21.2
Mode	50
Standard Deviation	9.1971041
Sample Variance	84.586724
Kurtosis	1.4951969
Skewness	1.1080984

Range	45
Minimum	5
Maximum	50
Sum	11401.6
Count	506
Coefficient of Variation	0.4081651

- ❖ The median 21.2 is slightly lesser than the mean value 22.53.
- ❖ The histogram shows that the Average price is positively skewed with skewness value 1.108. This means that there is more data at the left of the mean and lesser at the right of the mean. There is a tail at the right of the distribution that affects the mean and it is skewed to the right.
- ❖ This data has slight Positive Kurtosis of 1.49519 which is lesser than 3 so we can say that the distribution is flat, that is platykurtic.
- ❖ The coefficient of variation is 0.4082 which is between 0.2 to 0.5 which means the spread is normal.

3) Compute the covariance matrix. Share your observations.

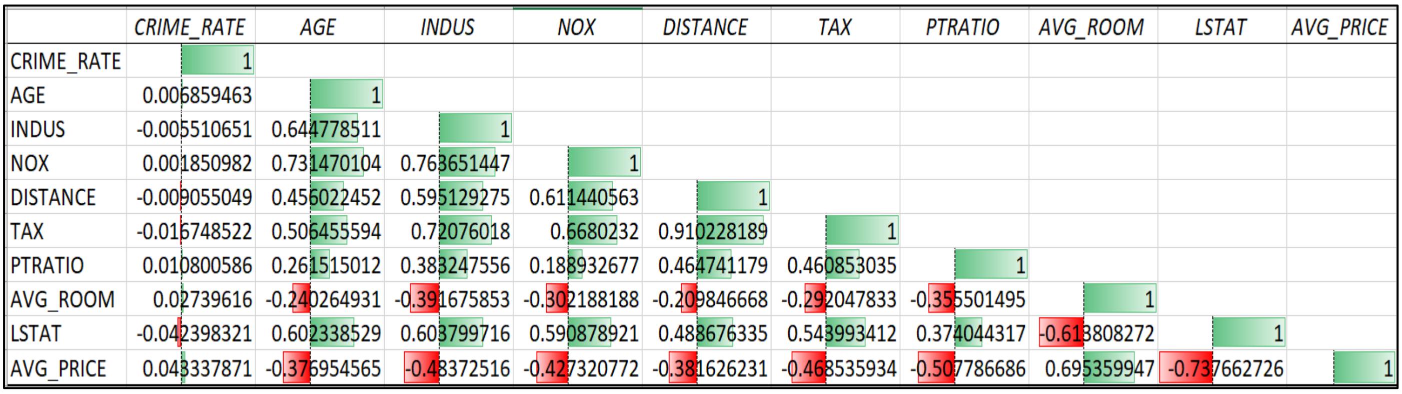


	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RATE	8.516147873									
AGE	0.562915215	790.7924728								
INDUS	-0.110215175	124.2678282	46.97142974							
NOX	0.000625308	2.381211931	0.605873943	0.013401099						
DISTANCE	-0.229860488	111.5499555	35.47971449	0.615710224	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

From the covariance matrix we can infer that AGE_PRICE vs CRIME_RATE (1.16201), AVG_PRICE vs AVG_ROOM(4.48457) has positive covariance. Thus when these independent variables increases or decreases, AVG_PRICE also increases or decreases respectively.

AGE vs AVG_PRICE(-97.39615) , INDUS vs AVG_PRICE(-30.46050), NOX vs AVG_PRICE(-0.454512), DISTANCE vs AVG_PRICE(-30.5008), TAX vs AVG_PRICE(-724.8204), PTRATIO vs AVG_PRICE(-10.090677) and LSTAT vs AVG_PRICE(-48.35179) have negative covariance. Thus the relationship of AVG_PRICE with these variables can't be easily predicted.

4) Create a correlation matrix of all the variables (Use Data analysis tool pack).
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

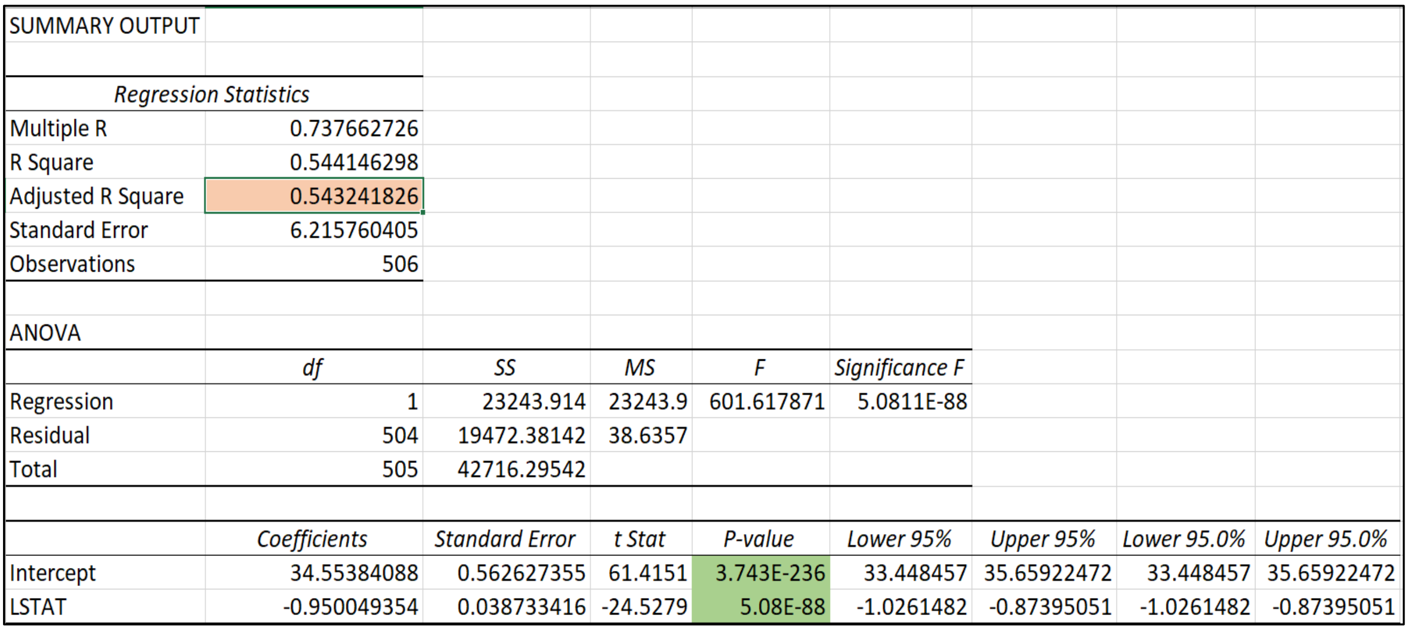
a) The top three positively correlated pairs in descending order are as follows.
1. Tax vs Distance with correlation value 0.910228
2. NOX vs INDUS with correlation value 0.763651
3. NOX vs AGE with correlation value 0.73147

b) The top three negatively correlated pairs in ascending order are as follows.
1. AVG_PRICE vs LSTAT with correlation value -0.7376627
2. LSTAT vs AVG_ROOM with correlation value -0.613808
3. AVG_PRICE vs PTRATIO with correlation value -0.507787

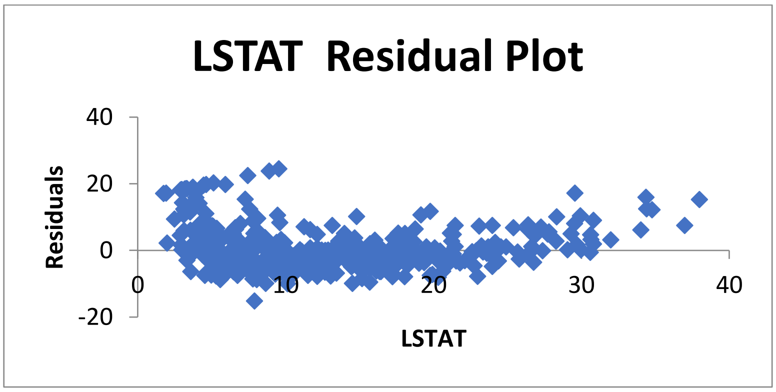
5) Build an initial regression model with AVG_PRICE as 'y' (Dependent variable) and LSTAT variable as Independent Variable. Generate the residual plot.

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	0.737662726							
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	23243.9	601.617871	5.0811E-88			
Residual	504	19472.38142	38.6357					
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	61.4151	3.743E-236	33.448457	35.65922472	33.448457	35.65922472
LSTAT	-0.950049354	0.038733416	-24.5279	5.08E-88	-1.0261482	-0.87395051	-1.0261482	-0.87395051



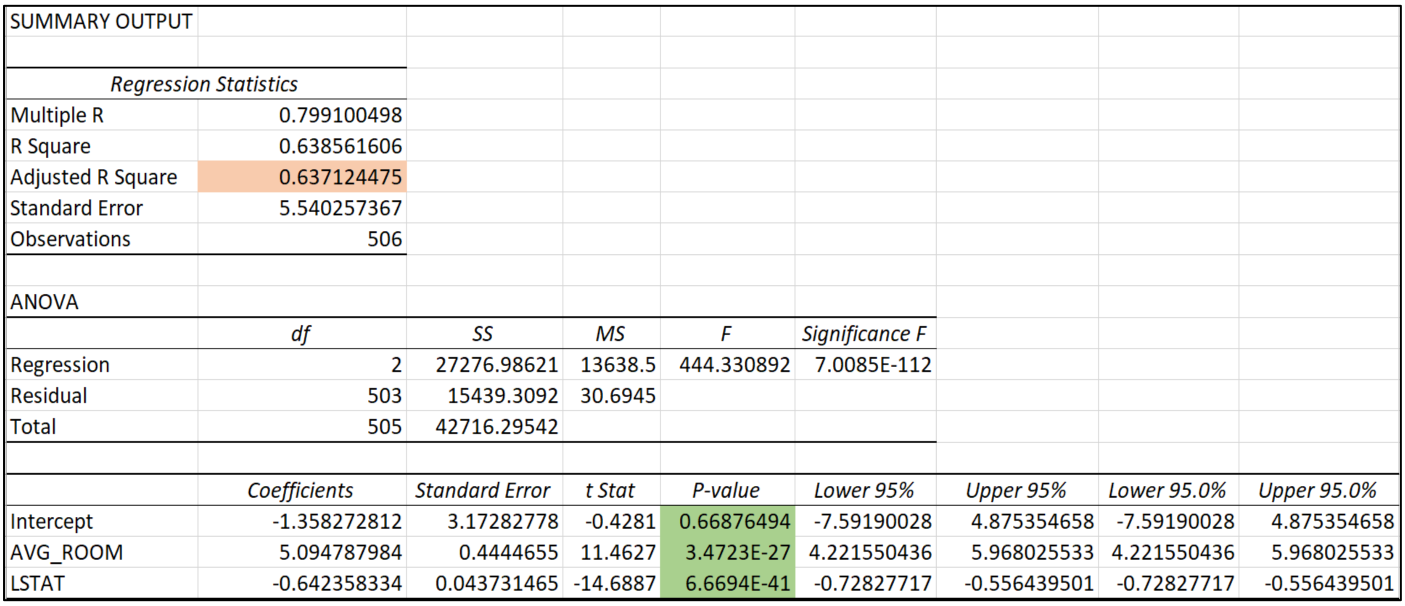
a) From Adjusted R Square value **0.5432418**, we can say that with LSTAT as independent variable, we can predict the dependent variable Average price with **54.32% accuracy**. The **intercept value 34.5538** says that if the independent variable LSTAT is 0, then the **average price is 34.5538 ($34,553.8)**. The coefficient of LSTAT value says that a unit increase in LSTAT value Average price is changed by **-0.9500**, that is if LSTAT is increased be **1%**, the Average price decreases by **$9,500**. The residual plot doesn't show any pattern which means the error is random.

b) From SLR we can infer that the P-value is lesser than 0.05 that is 5.08E-88 which proves that the Alternate hypothesis is true. Adjusted R square is greater than 0.5 and there is no pattern on residual plot, which means the error is random. So, the LSTAT variable is significant for the analysis.

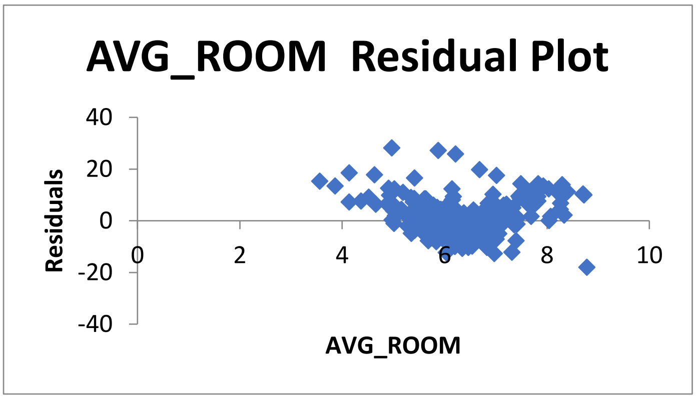
6) Build a new Regression model including LSTAT and AVG_ROOM together as independent variables and AVG_PRICE as dependent variable.

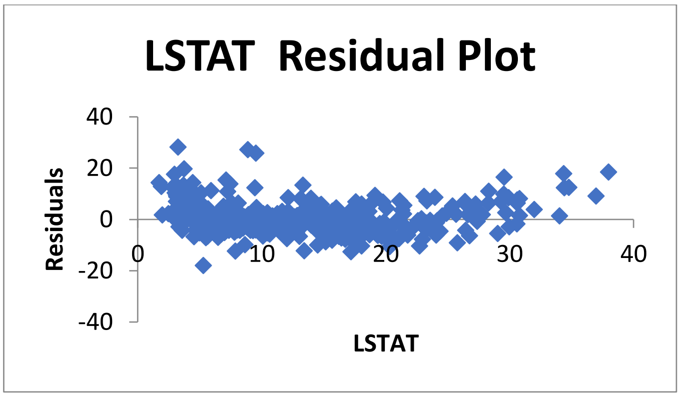
a) Write the Regression equation. If a new house in this locality has 7 rooms (on an average) and has a value of 20 for L-STAT, then what will be the value of AVG_PRICE? How does it compare to the company quoting a value of 30000 USD for this locality? Is the company Overcharging/ Undercharging?

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	13638.5	444.330892	7.0085E-112			
Residual	503	15439.3092	30.6945					
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	-0.4281	0.66876494	-7.59190028	4.875354658	-7.59190028	4.875354658
AVG_ROOM	5.094787984	0.4444655	11.4627	3.4723E-27	4.221550436	5.968025533	4.221550436	5.968025533
LSTAT	-0.642358334	0.043731465	-14.6887	6.6694E-41	-0.72827717	-0.556439501	-0.72827717	-0.556439501





a) The regression equation derived from the model is as follows.

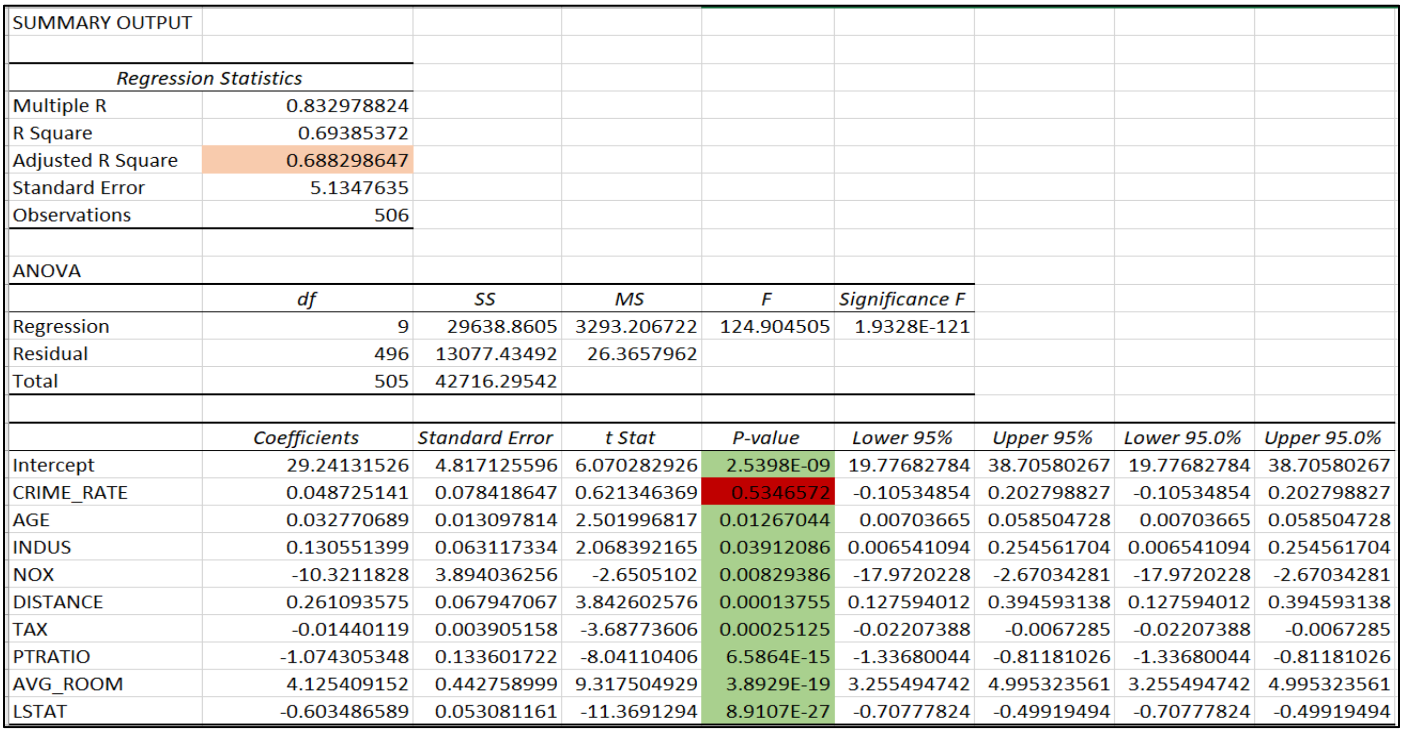
Average Price = intercept + coefficient of AVG_ROOM* AVG_ROOM + coefficient of LSTAT* LSTAT

If a new house in this locality has 7 rooms (on an average) and has a value of 20 for L-STAT, the value of AVG_PRICE will be as follows.

From the regression equation we can say that the average price of a house with the average of 7 rooms and L-STAT value of 20 has average price of $21,458. If a company quotes for the value of $30,000 for this locality, it is overcharging.

b) The performance of this model is better than the previous model as the adjusted R square value of this model is 0.63712 which is greater than the previous model with adjusted R square value 0.54324.

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.



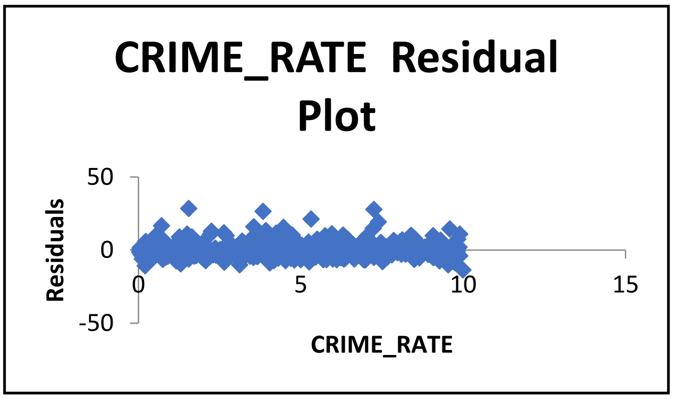
SUMMARY OUTPUT

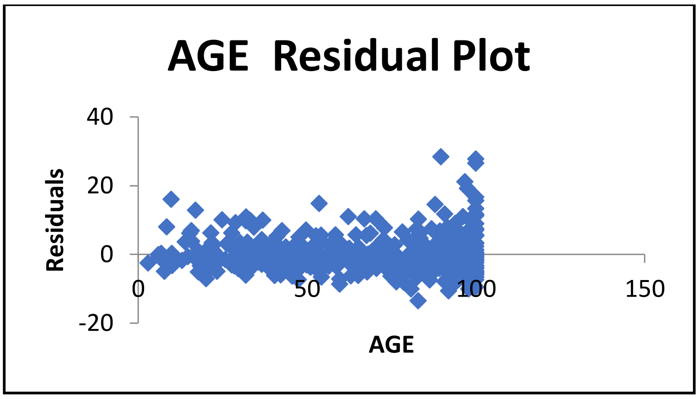
Regression Statistics	
Multiple R	0.832978824
R Square	0.69385372
Adjusted R Square	0.688298647
Standard Error	5.1347635
Observations	506

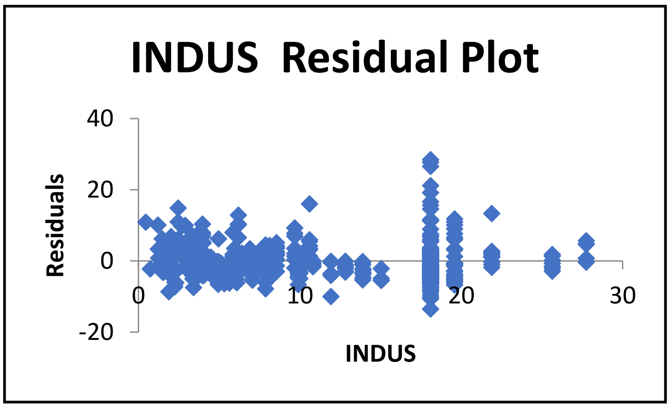
ANOVA

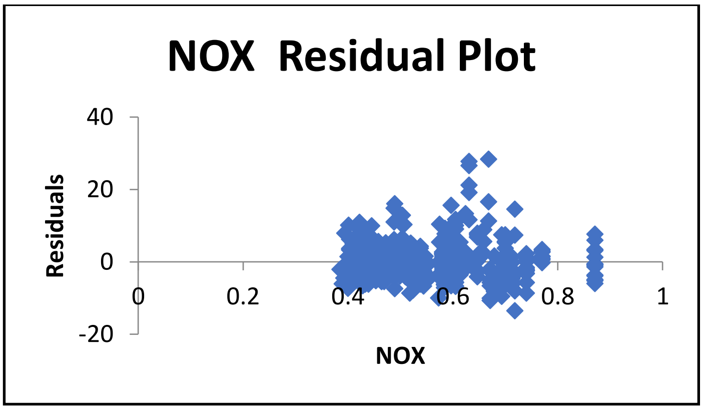
	df	*SS*	*MS*	*F*	*Significance F*
Regression	9	29638.8605	3293.206722	124.904505	1.9328E-121
Residual	496	13077.43492	26.3657962		
Total	505	42716.29542			

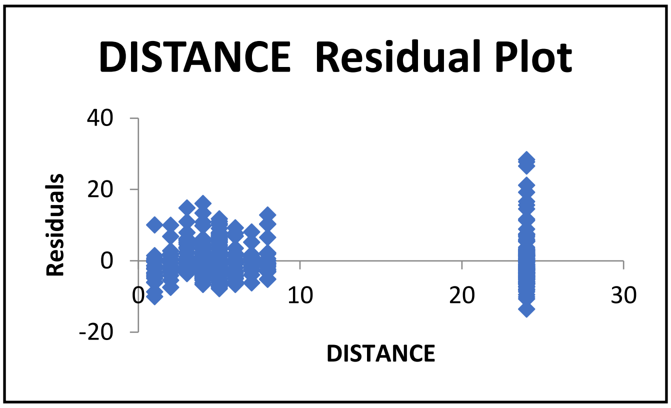
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	29.24131526	4.817125596	6.070282926	2.5398E-09	19.77682784	38.70580267	19.77682784	38.70580267
CRIME_RATE	0.048725141	0.078418647	0.621346369	0.5346572	-0.10534854	0.202798827	-0.10534854	0.202798827
AGE	0.032770689	0.013097814	2.501996817	0.01267044	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	-2.6505102	0.00829386	-17.9720228	-2.67034281	-17.9720228	-2.67034281
DISTANCE	0.261093575	0.067947067	3.842602576	0.00013755	0.127594012	0.394593138	0.127594012	0.394593138
TAX	-0.01440119	0.003905158	-3.68773606	0.00025125	-0.02207388	-0.0067285	-0.02207388	-0.0067285
PTRATIO	-1.074305348	0.133601722	-8.04110406	6.5864E-15	-1.33680044	-0.81181026	-1.33680044	-0.81181026
AVG_ROOM	4.125409152	0.442758999	9.317504929	3.8929E-19	3.255494742	4.995323561	3.255494742	4.995323561
LSTAT	-0.603486589	0.053081161	-11.3691294	8.9107E-27	-0.70777824	-0.49919494	-0.70777824	-0.49919494

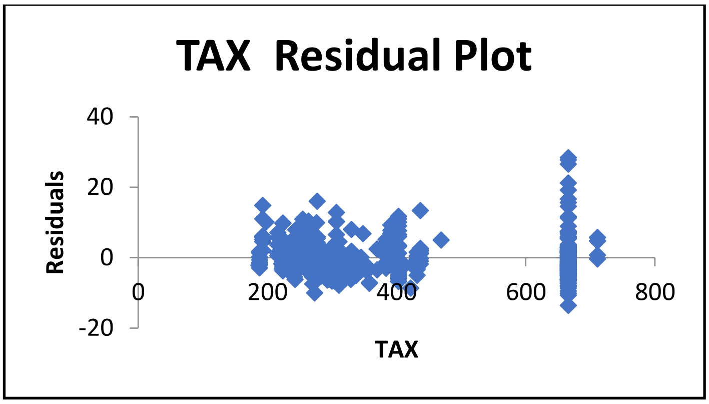


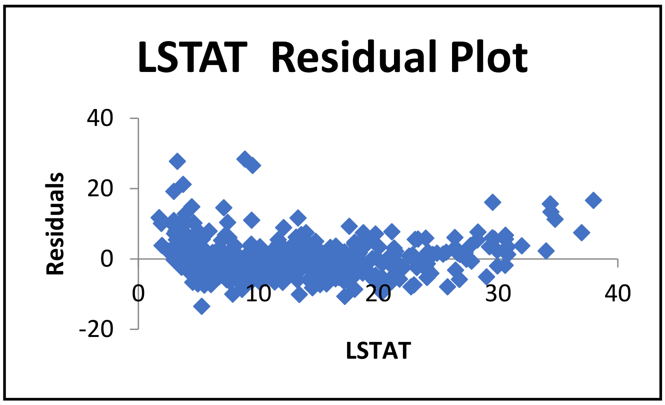


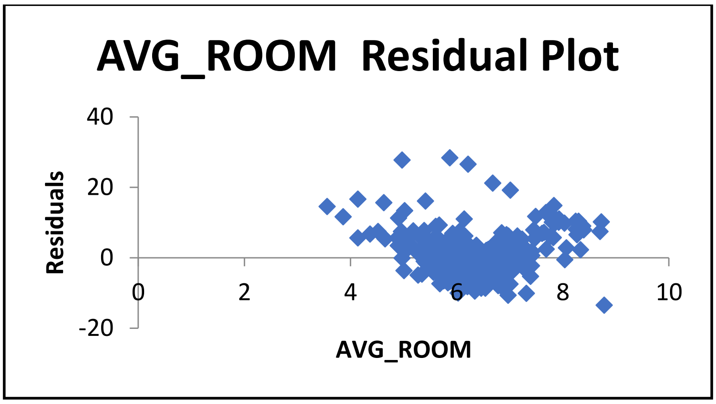


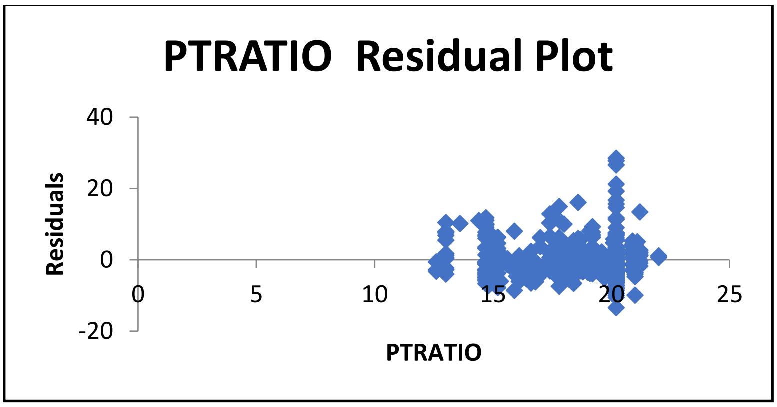












The performance of this model is better than the previous model as **the adjusted R square** value of this model is **0.68829** which is greater than the previous model with adjusted R square value **0.63712**. However, the **P-Value of CRIME_RATE is 0.534657** which is greater than 0.05, which proves that the null hypothesis is true. So, this model cannot be used for prediction of Average Price.

The P-values of the variables Age, INDUS, NOX, Distance, TAX, PTRATIO, AVG_ROOM, LSTAT is lesser than 0.05, so we can create further MLR models by omitting CRIME_RATE variable.

The intercept value is 29.24131 which means if all the independent variables are 0, the Average price would be $29,241.

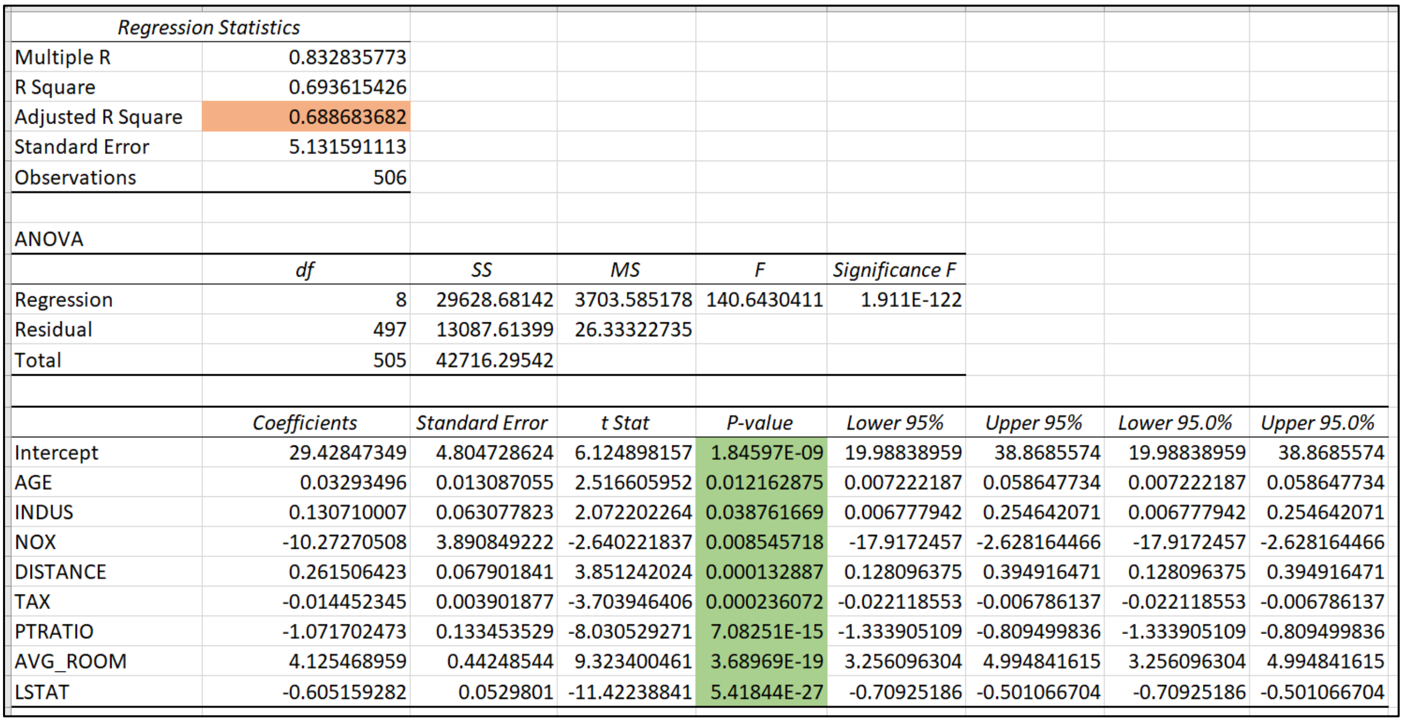
8) Pick out only the significant variables from the previous question. Make another instance of the Regression model using only the significant variables you just picked and answer the questions below:

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.

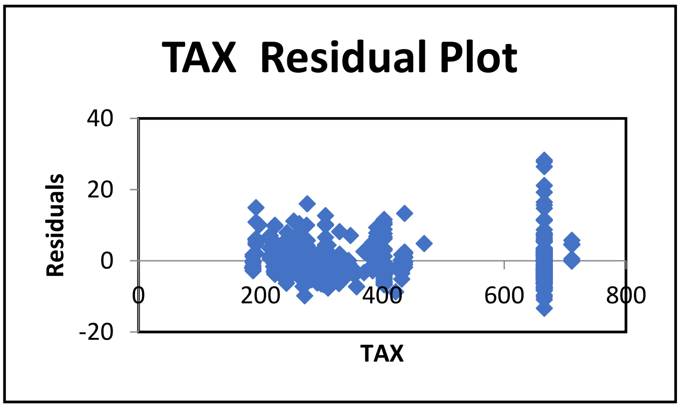


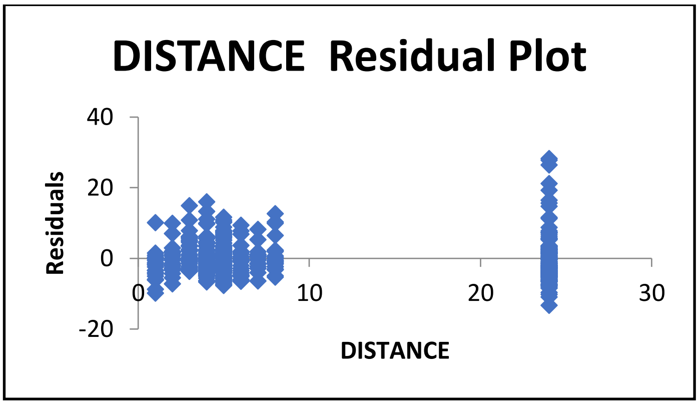
Regression Statistics	
Multiple R	0.832835773
R Square	0.693615426
Adjusted R Square	0.688683682
Standard Error	5.131591113
Observations	506

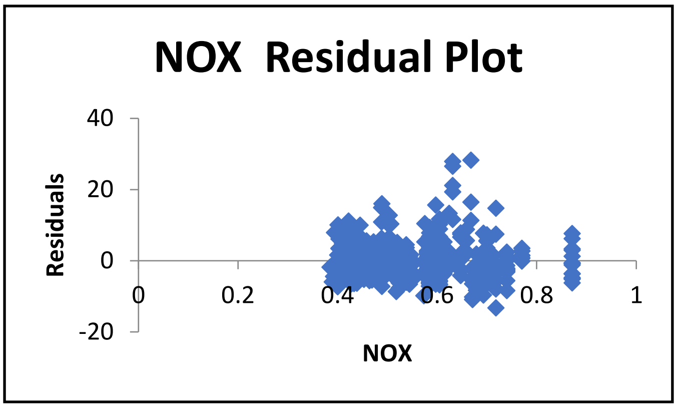
ANOVA

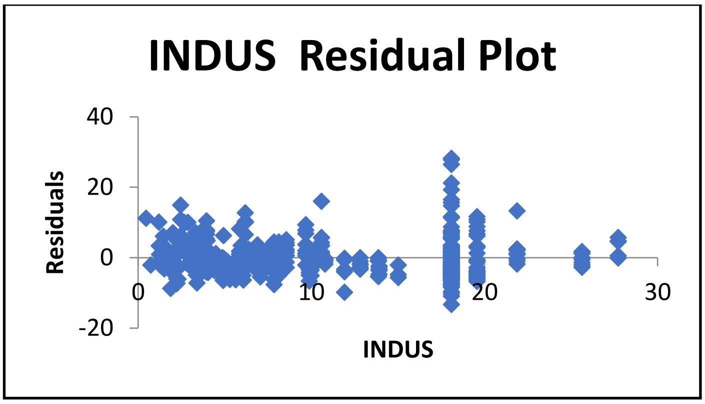
	df	*SS*	*MS*	*F*	*Significance F*
Regression	8	29628.68142	3703.585178	140.6430411	1.911E-122
Residual	497	13087.61399	26.33322735		
Total	505	42716.29542			

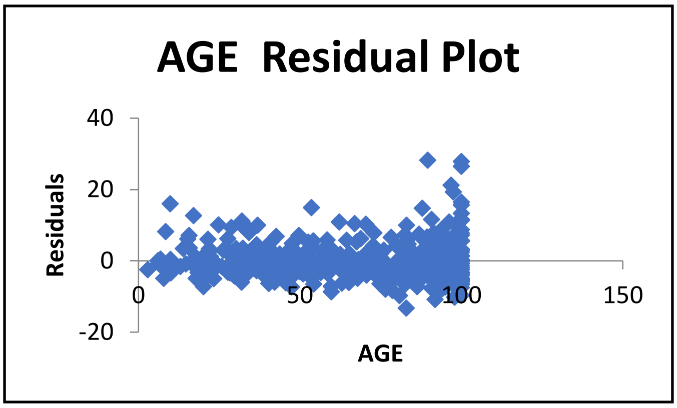
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	29.42847349	4.804728624	6.124898157	1.84597E-09	19.98838959	38.8685574	19.98838959	38.8685574
AGE	0.03293496	0.013087055	2.516605952	0.012162875	0.007222187	0.058647734	0.007222187	0.058647734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

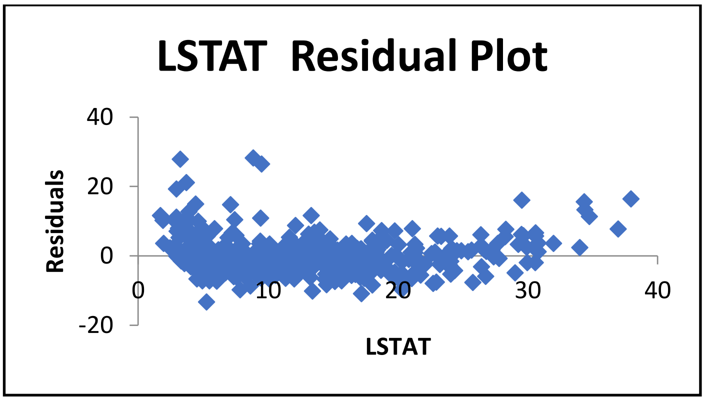


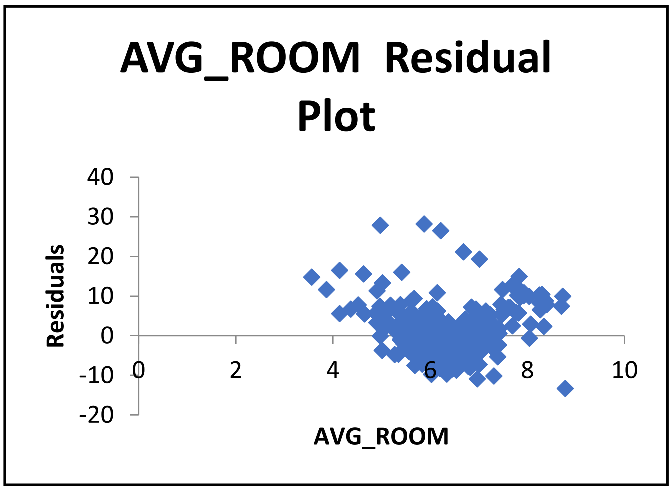


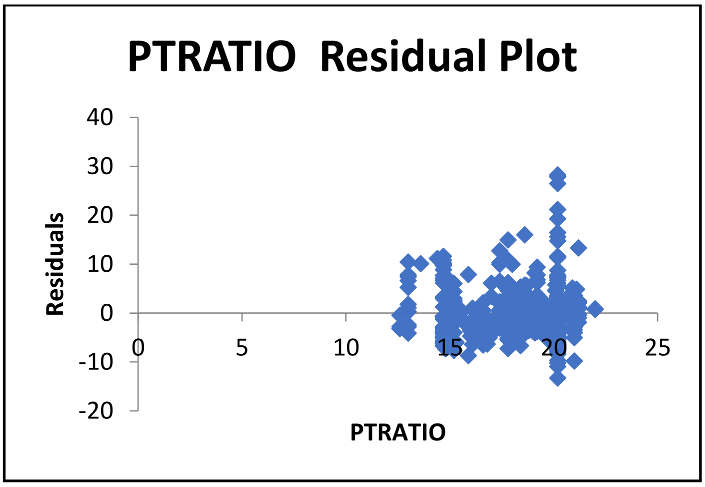












a) The significant variables to build the model are AGE, INDUS, NOX, DISTANCE, TAX, PTRATIO, AVG_ROOM, LSTAT, all with P-value lesser than 0.05. This proves the Alternate hypothesis and the adjusted R square value is 0.6886. The residual plots also don't show any pattern and the residuals are spread out. So, this model can be used to predict the independent variable Average Price.

b) The performance of this model is better than the previous model as the adjusted R square value of this model is **0.68868** which is **slightly greater than the previous model** with adjusted R square value 0.68829. The **accuracy (68.86%)** of this model is better than the previous one.

c) The values of the Coefficients in ascending order as follows

Variables	Coefficients
NOX	-10.27270508
PTRATIO	-1.071702473
LSTAT	-0.605159282
TAX	-0.014452345
AGE	0.03293496
INDUS	0.130710007
DISTANCE	0.261506423
AVG_ROOM	4.125468959

From the coefficient value of NOX that is **-10.2727**, we can infer that a unit increase in NOX value can decrease the Average Price by $10,272.

d) The regression equation can be written as follows

Average price = intercept + coefficient of AGE*AGE + coefficient of INDUS*INDUS + coefficient of DISTANCE*DISTANCE + coefficient of AVG_ROOM*AVG_ROOM +coefficient of NOX*NOX + coefficient of PTRATIO*PTRATIO +coefficient of LSTAT*LSTAT+ coefficient of TAX*TAX
Average Price = 29.42847 + 0.032935 * AGE + 0.13071*INDUS -10.2727*NOX + 0.261506*DISTANCE - 0.014452 * TAX -1.07170*PTRATIO + 4.12547* AVG_ROOM -0.60516 * LSTAT