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:** The mean crime rate is approximately 4.872 which represent the average crime rate in the dataset the standard deviation is approximately 2.291 a moderate amount of variability of dispersion in the dataset and average of crime is 4.82, skewness is nearly 0 which says curve follows normal distribution.

- **Age:** The mean age is approximately is 68.5749 which represent the age in the dataset the standard deviation is approximately 28.14886 a amount of variability of dispersion in the dataset and average of age is 77.5 , skewness is negative value -0 which says flatter distribution for this variable.

- **Indus:** The mean indus is approximately is 11.1367 which represent the indus in the dataset the standard deviation is approximately 6.86 amount of variability of dispersion in the dataset and average of indus is 9.69 , negative kurtosis value -1.2 so flatter distribution is variable.

- **NOX:** The mean nox is approximately is 0.55469 which represent the nox in the dataset the standard deviation is approximately 0.115878 amount of variability of dispersion in the dataset and average of nox is 0.538, skewness is nearly is 0.72 which says curve follow normal distribution. Kurtosis is slightly negative value -0.6 so we can say normal distribution.

- **Distance :** the mean distance is approximately is 9.5494 which represent the distance in the dataset the standard deviation is approximately 8.7072 amount of variability of Dispersion in the dataset and average distance is 5 , kurtosis is negative value is -0.86 and skewness is positive value .

- **Tax:** The mean tax is approximately is 408.23 which represent the tax in the dataset the standard deviation is approximately 2.1649 amount of variability of dispersion in the dataset and average is 330 kurtosis is negative value is -1.14241 and skewness is a positive value.

- **PTRATIO:** The mean Pupil teacher ratio is 18.455 which represent the pupil teacher ratio in the dataset the standard deviation is approximately 2.1649 amount of variability of dispersion in the dataset and average is 19.05 and kurtosis and skewness both are negative value.

- **AVG_Room :** The mean Average room ratio is 6.2846 which represent the Avg room ratio in the dataset the standard deviation is approximately 0.702617 amount of variability of dispersion in the dataset and average is 6.2085 and kurtosis and skewness is both are positive value.

- **LSTAT:** The mean lower state average ratio is 12.65 which represent the Lstate in the dataset the standard deviation is approximately 7.141062 amount of variability of dispersion in the dataset and average is 11.36 and kurtosis and skewness is both are positive value.

- **AVG_Price:** The mean AVG_Price is 22.53 which represent the Average price in the dataset the standard deviation is approximately 9.197 amount of variability of dispersion in the dataset and average is 21.2 and kurtosis and skewness is both are positive value.

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

- A histogram is an approximate representation of the distribution of numerical data . and average price of the is (21,25) has a very few data and it is a rightly skewed with a leptokurtic kurtosis.

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	46.97143							
NOX	0.000625308	2.381211931	0.605874	0.013401						
DISTANCE	-0.229860488	111.5499555	35.47971	0.61571	75.66653					
TAX	-8.229322439	2397.941723	831.7133	13.0205	1333.117	28348.62				
PTRATIO	0.068168906	15.90542545	5.680855	0.047304	8.743402	167.8208	4.677726296			
AVG_ROOM	0.056117778	-4.74253803	-1.88423	-0.02455	-1.28128	-34.5151	-0.53969452	0.492695		
LSTAT	-0.882680362	120.8384405	29.52181	0.48798	30.32539	653.4206	5.771300243	-3.073654967	50.89398	
AVG_PRICE	1.16201224	-97.39615288	-30.4605	-0.45451	-30.5008	-724.82	-10.0906756	4.484565552	-48.3518	84.41955616

- **Observation:**
  - The most positive covariance value is 2397.94 (age)
  - The most negative covariance value is -724.82(tax)

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

	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	-0.3555015	1		
LSTAT	-0.042398321	0.602339	0.6038	0.590879	0.488676	0.543993	0.37404432	-0.613808272	1	
AVG_PRICE	0.043337871	-0.37695	-0.48373	-0.42732	-0.38163	-0.46854	-0.5077867	0.695359947	-0.73766	1

a) Which are the top 3 positively correlated pairs.

- Tax (0.543993)
- Distance(0.910228)
- Indus (0.763651)

b) Which are the top 3 negatively correlated pairs.

- Avg_room (0.613808272)
- PT ratio (-0.5077867)
- lSTAT (-0.73766)

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

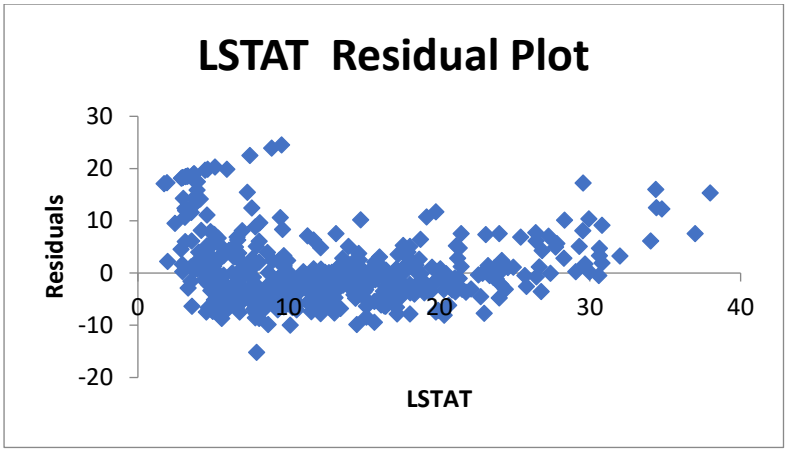
SUMMARY OUTPUT

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	23243.91	23243.91	601.6178711	5.0811E-88
Residual	504	19472.38	38.63568		
Total	505	42716.3			

	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	34.55384088	0.562627	61.41515	3.7431E-236	33.44845704	35.65922472	33.44845704	35.65922472
LSTAT	-0.950049354	0.038733	-24.5279	5.0811E-88	-1.0261482	-0.873950508	-1.0261482	-0.873950508

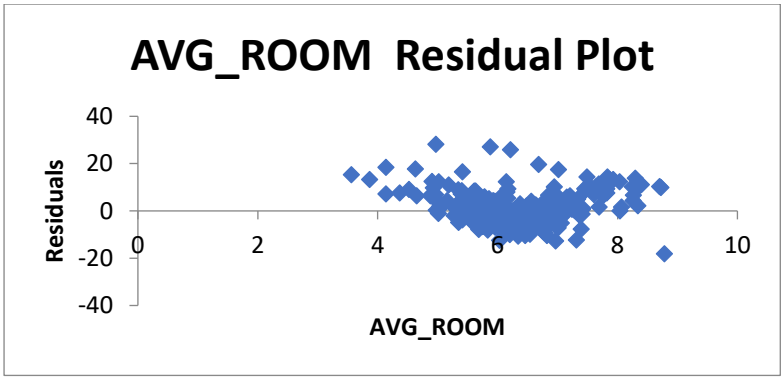


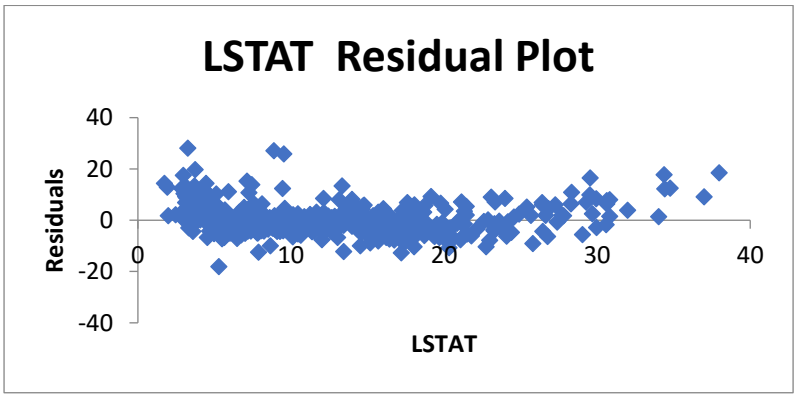
a) **What do you infer from the Regression Summary output in terms of variance explained, coefficient value, Intercept, and the Residual plot?**
   - The value of variance less than 0.5 so the model is highly significant.
   - From the residual plot we can see that is some outlier in a data.

b) **Is LSTAT variable significant for the analysis based on your model?**
   - Yes, LSTAT Variable significant for the analysed model.

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

A	B	C	D	E	F	G	H	I
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	13638.49	444.3309	7.0085E-112			
Residual	503	15439.3092	30.69445					
Total	505	42716.29542						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Upper 95.0%*
Intercept	-1.358272812	3.17282778	-0.4281	0.668765	-7.591900282	4.875355	-7.5919	4.875354658
AVG_ROOM	5.094787984	0.4444655	11.46273	3.47E-27	4.221550436	5.968026	4.22155	5.968025533
LSTAT	-0.642358334	0.043731465	-14.6887	6.67E-41	-0.728277167	-0.55644	-0.72828	-0.556439501





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?

- **Y=B0+b1x1+b2x2+….bnXn**
  Avg price=-1.358272+5.094787*7+-0.64235*20
  = -1.358272+35.66352+-12.8472
  =21.45808 -30000
  Average difference = 29978.54
- By using regression equation the avg value is 21.4581 company quoting a value the company **overcharging .**

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.

- Previous model adjusted R square is =0.543241826
- Model adjusted R square is = 0.637124475

- The previous model adjusted better than new model adjusted more than comparing better value for adjusted R square.

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

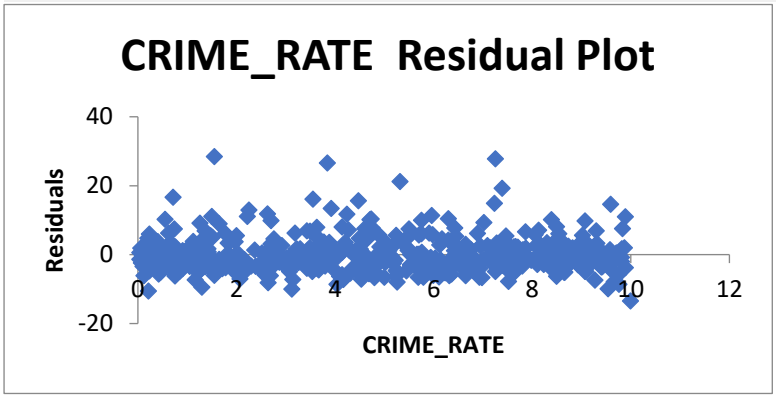
Regression Statistics	
Multiple R	0.832978824
R Square	0.69385372
Adjusted R Square	0.688298647
Standard Error	5.1347635
Observations	506

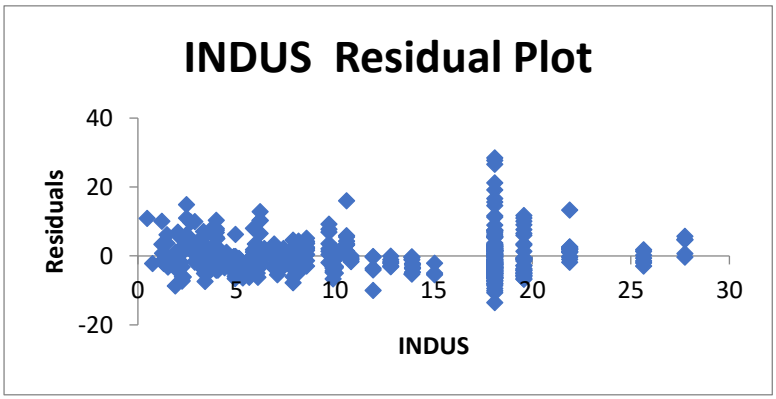
Chart Area

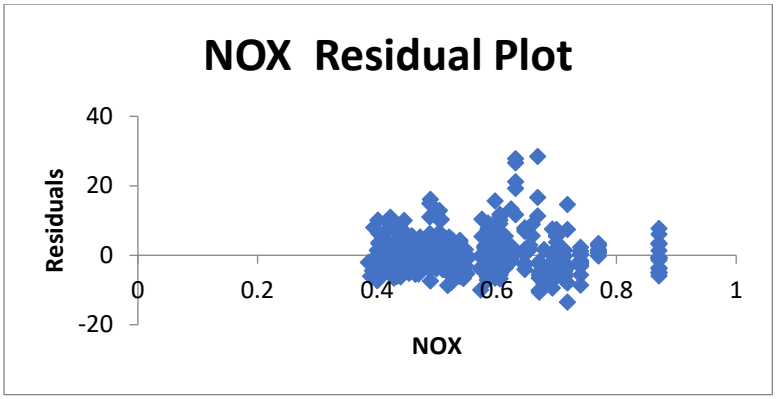
ANOVA

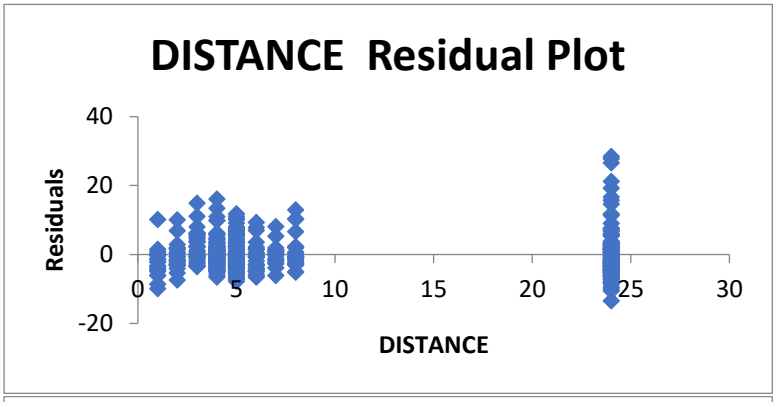
	df	*SS*	*MS*	*F*	*Significance F*
Regression	9	29638.8605	3293.207	124.9045	1.9328E-121
Residual	496	13077.43492	26.3658		
Total	505	42716.29542			

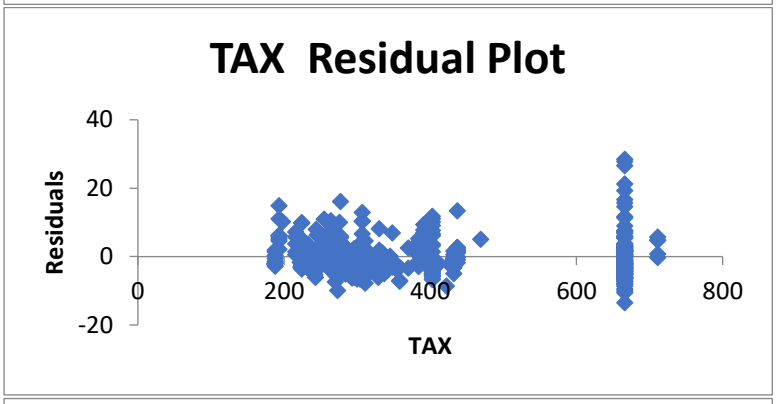
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Upper 95.0%*
Intercept	29.24131526	4.817125596	6.070283	2.54E-09	19.77682784	38.7058	19.77683	38.70580267
CRIME_RATE	0.048725141	0.078418647	0.621346	0.534657	-0.105348544	0.202799	-0.10535	0.202798827
AGE	0.032770689	0.013097814	2.501997	0.01267	0.00703665	0.058505	0.007037	0.058504728
INDUS	0.130551399	0.063117334	2.068392	0.039121	0.006541094	0.254562	0.006541	0.254561704
NOX	-10.3211828	3.894036256	-2.65051	0.008294	-17.97202279	-2.67034	-17.972	-2.670342809
DISTANCE	0.261093575	0.067947067	3.842603	0.000138	0.127594012	0.394593	0.127594	0.394593138
TAX	-0.01440119	0.003905158	-3.68774	0.000251	-0.022073881	-0.00673	-0.02207	-0.0067285
PTRATIO	-1.074305348	0.133601722	-8.0411	6.59E-15	-1.336800438	-0.81181	-1.3368	-0.811810259
AVG_ROOM	4.125409152	0.442758999	9.317505	3.89E-19	3.255494742	4.995324	3.255495	4.995323561
LSTAT	-0.603486589	0.053081161	-11.3691	8.91E-27	-0.70777824	-0.49919	-0.70778	-0.499194938

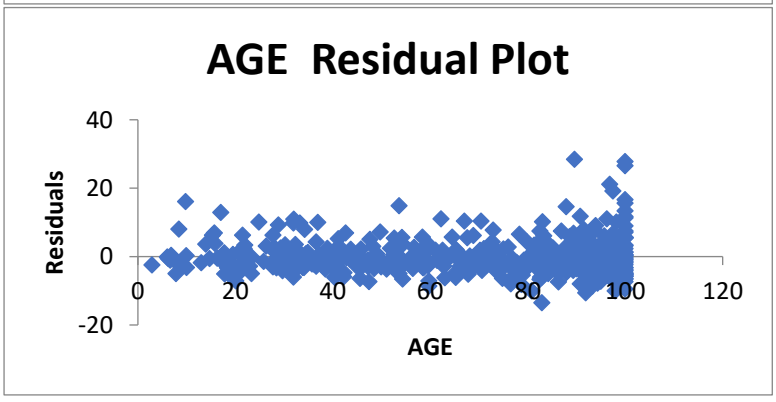


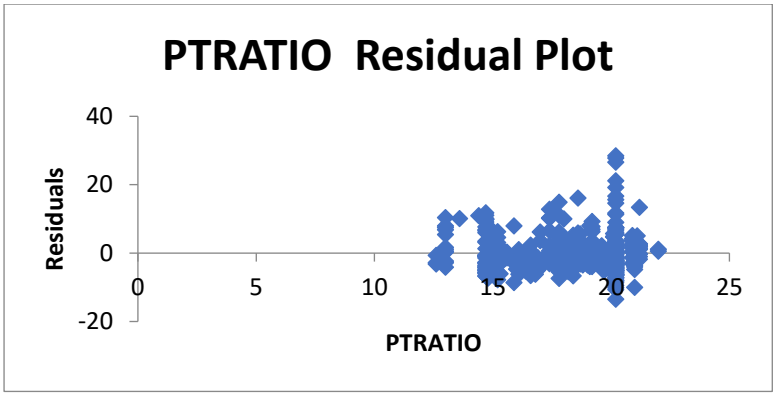


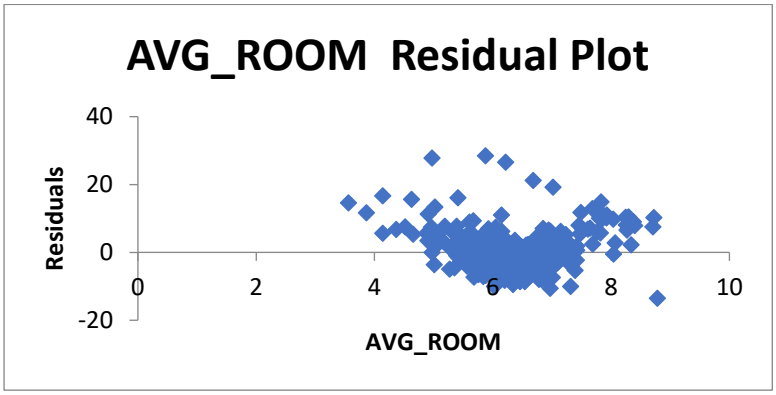

NOX Residual Plot
Residuals
40
20
0
-20
0
0.2
0.4
0.6
0.8
1
NOX

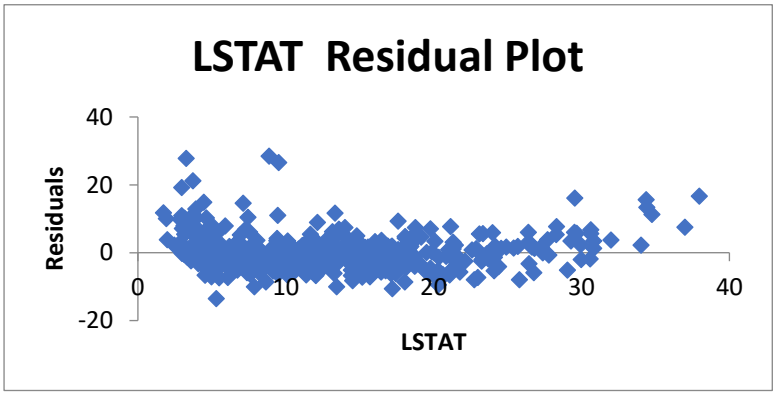

DISTANCE Residual Plot
Residuals
40
20
0
-20
0
5
10
15
20
25
30
DISTANCE


TAX Residual Plot
Residuals
40
20
0
-20
0
200
400
600
800
TAX


AGE Residual Plot
Residuals
40
20
0
-20
0
20
40
60
80
100
120
AGE



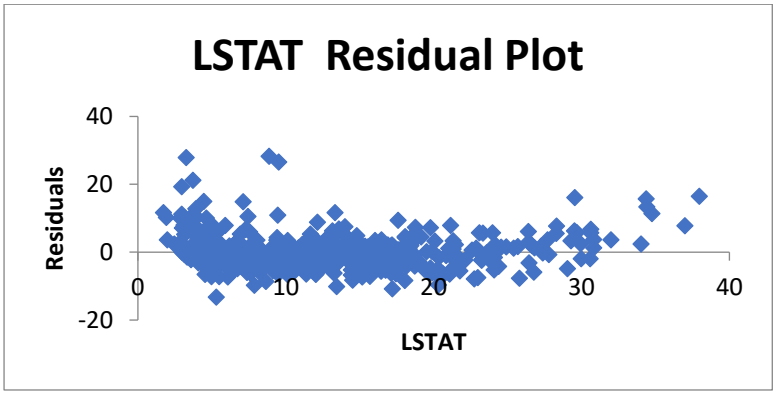


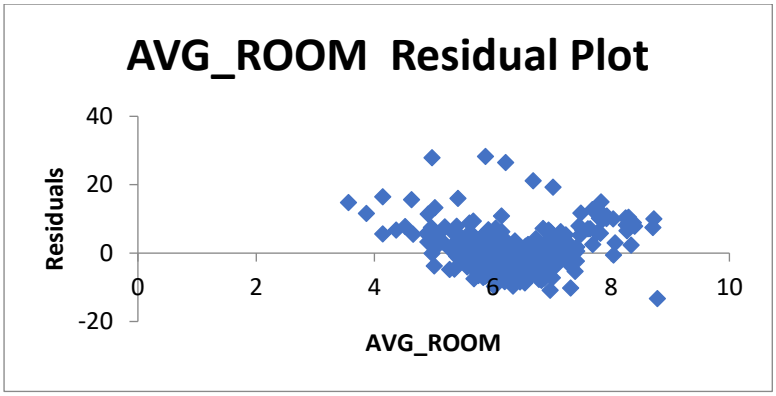


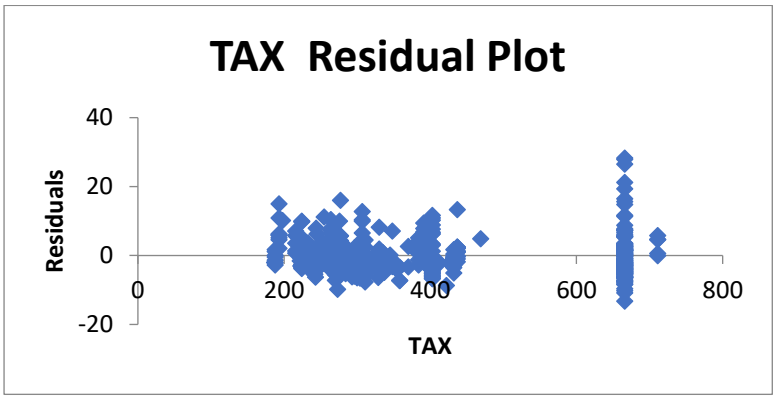
- The average price and average room is a directly proportional . and Nox, and average price is proportional to each other significant variable is age, indus , nox,distance , tax.

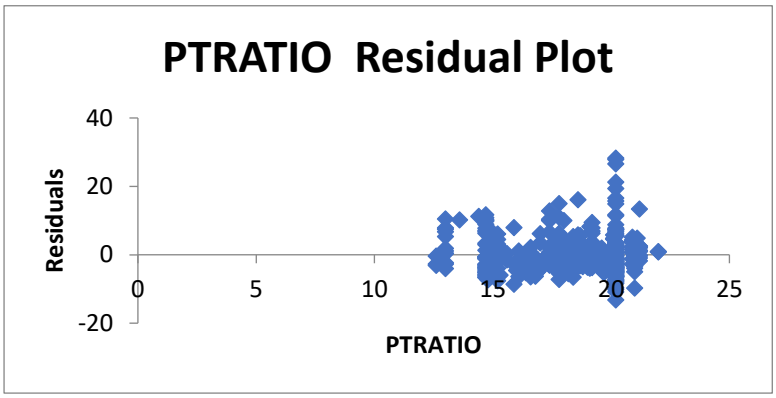
8) Pick out only the significant variables from the previous question. Make another instance of the Regression model using only the significant variables you just picked and answer the questions below:

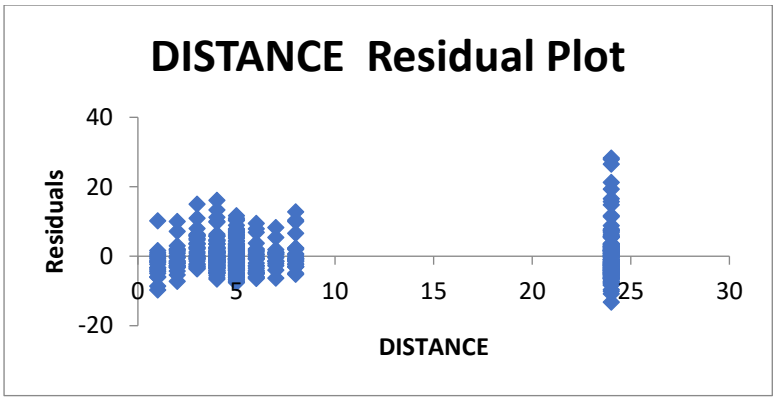
	A	B	C	D	E	F	G	H	I	J
1	SUMMARY OUTPUT									
2										
3	*Regression Statistics*									
4	Multiple R	0.832836								
5	R Square	0.693615								
6	Adjusted R	0.688684							Plot Area	
7	Standard E	5.131591								
8	Observatic	506								
9										
10	ANOVA									
11		*df*	*SS*	*MS*	*F*	*ignificance F*				
12	Regressior	8	29628.68	3703.585	140.643	1.9E-122				
13	Residual	497	13087.61	26.33323						
14	Total	505	42716.3							
15										
16		*Coefficients*	*tandard Erı*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*	
17	Intercept	29.42847	4.804729	6.124898	1.85E-09	19.98839	38.86856	19.98839	38.86856	
18	AGE	0.032935	0.013087	2.516606	0.012163	0.007222	0.058648	0.007222	0.058648	
19	INDUS	0.13071	0.063078	2.072202	0.038762	0.006778	0.254642	0.006778	0.254642	
20	NOX	-10.2727	3.890849	-2.64022	0.008546	-17.9172	-2.62816	-17.9172	-2.62816	
21	DISTANCE	0.261506	0.067902	3.851242	0.000133	0.128096	0.394916	0.128096	0.394916	
22	TAX	-0.01445	0.003902	-3.70395	0.000236	-0.02212	-0.00679	-0.02212	-0.00679	
23	PTRATIO	-1.0717	0.133454	-8.03053	7.08E-15	-1.33391	-0.8095	-1.33391	-0.8095	
24	AVG_ROO	4.125469	0.442485	9.3234	3.69E-19	3.256096	4.994842	3.256096	4.994842	
25	LSTAT	-0.60516	0.05298	-11.4224	5.42E-27	-0.70925	-0.50107	-0.70925	-0.50107	

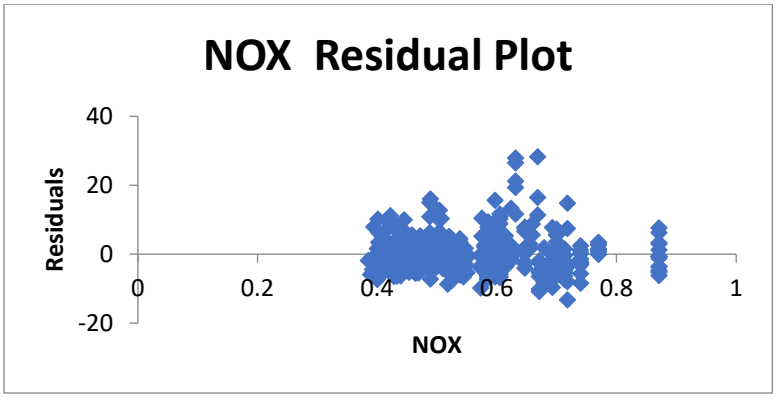


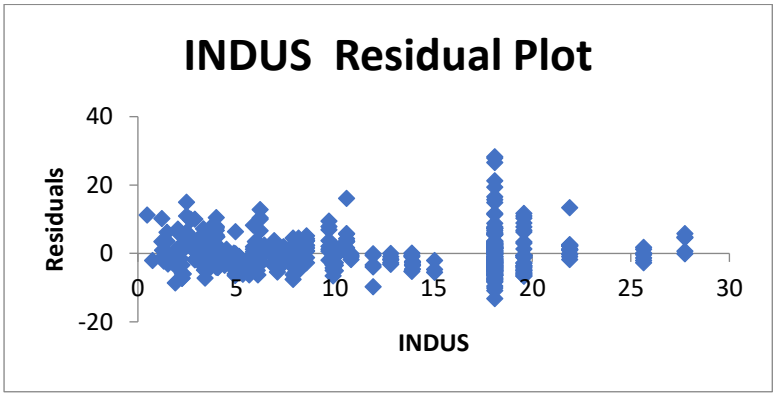

AVG_ROOM Residual Plot
Residuals
40
20
0
-20
0
2
4
6
8
10
AVG_ROOM

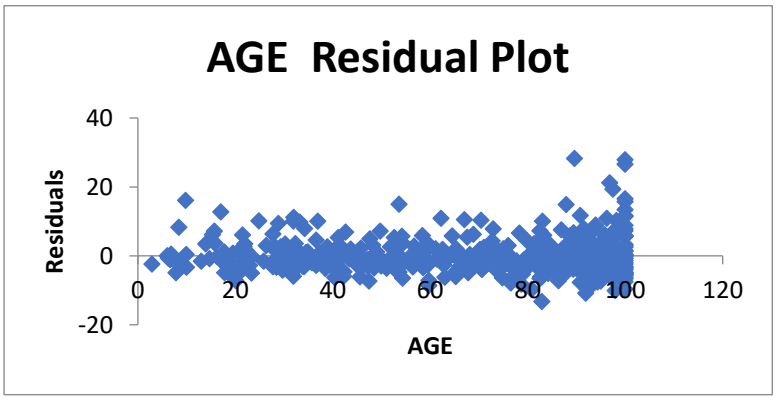

TAX Residual Plot
Residuals
40
20
0
-20
0
200
400
600
800
TAX


PTRATIO Residual Plot
Residuals
40
20
0
-20
0
5
10
15
20
25
PTRATIO


DISTANCE Residual Plot
Residuals
40
20
0
-20
0
5
10
15
20
25
30
DISTANCE







a) **Interpret the output of this model.**
   - The model Is adjusted R square is a interpret the model 68.8684.

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?

   - Adjusted R – square is 0.693615 to 0.68829 in the previous model performs better value of significant value.

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?**

Column1	Column2
AGE	0.032935
INDUS	0.13071
NOX	-10.272705
DISTANCE	0.2615064
TAX	-0.0144523
PTRATIO	-1.0717025
AVG_ROOM	4.125469
LSTAT	-0.6051593

- Therefore value of coefficient value is average price and average room is high and negative value is Nox .

d) Write the regression equation from this model.

- Multiply linear regression:
  Y = 29.42847349 + (0.03293496 * X1) + (0.130710007 * X2) - (10.27270508 * X3) + (0.261506423 * X4) - (0.014452345 * X5) - (1.071702473 * X6) + (4.125468959 * X7) - (0.605159282 * X8).