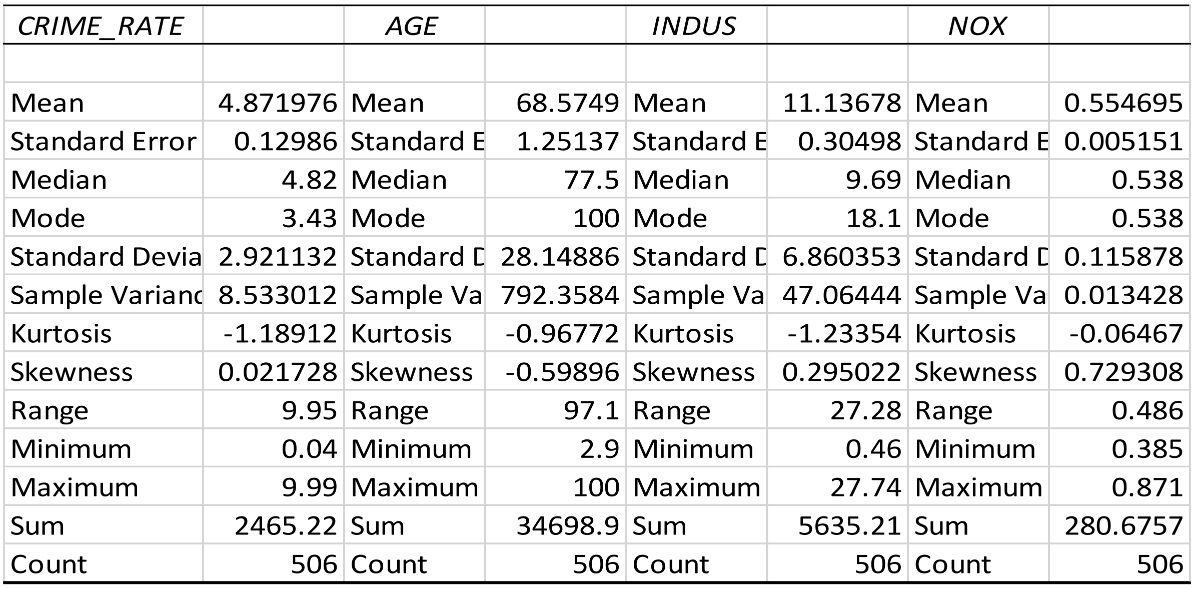
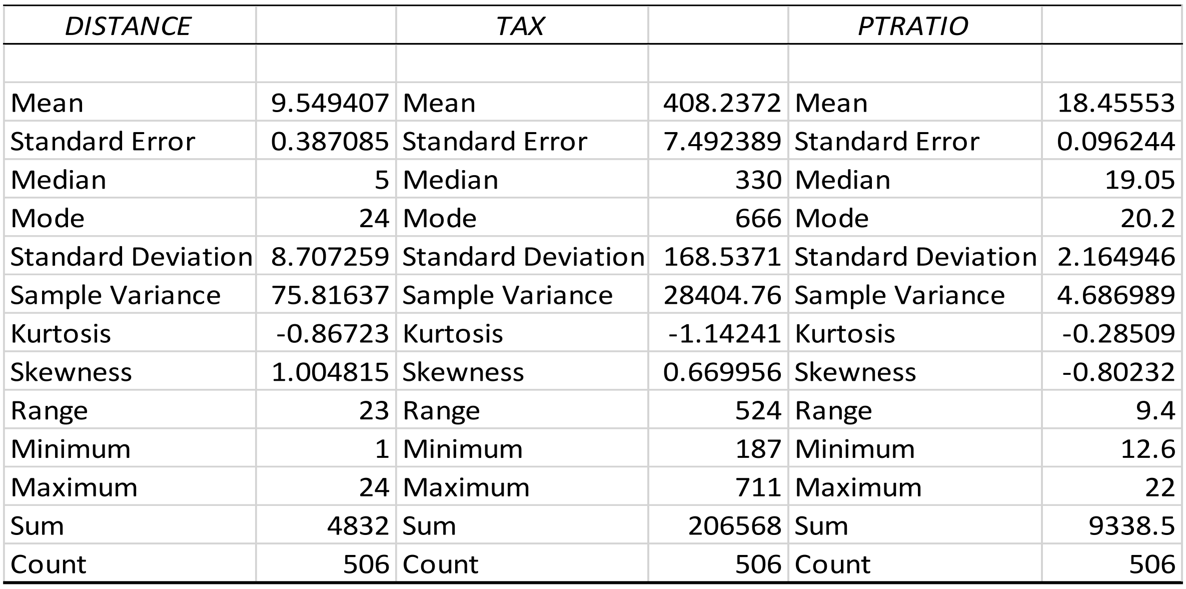
TERRO’S REAL ESTATE DATA ANALYSIS PROJECT

Q1: Generate the summary statistics for each variable in the table. (Use Data analysis tool pack). Write down your observation.

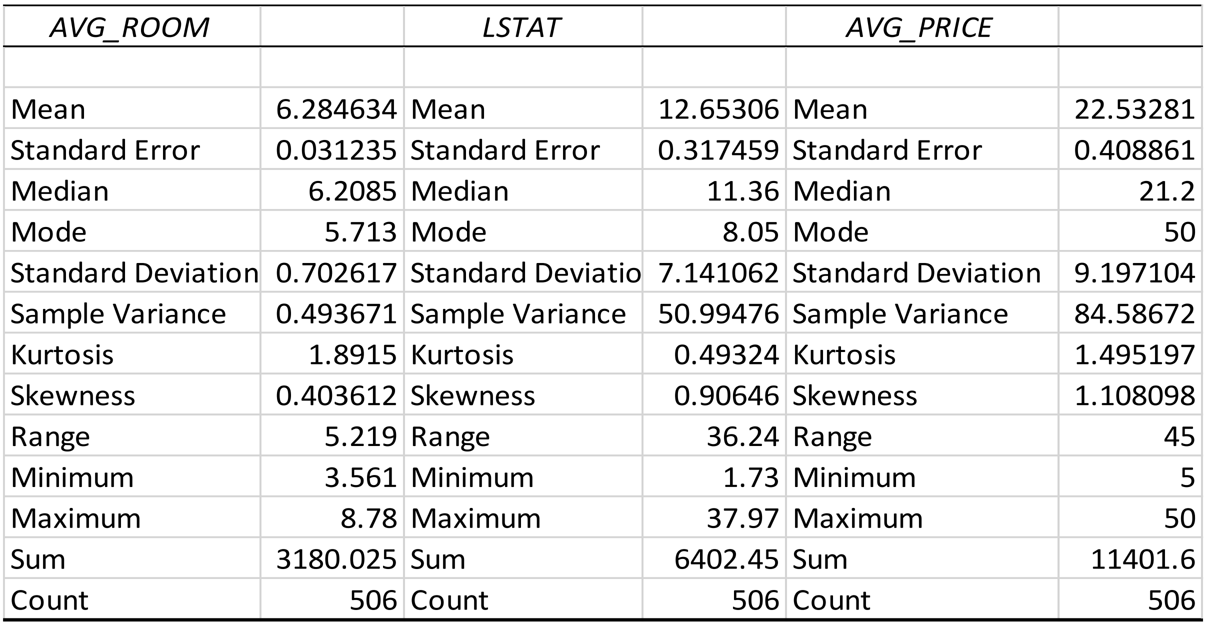


CRIME_RATE		*AGE*		*INDUS*		*NOX*	
Mean	4.871976	Mean	68.5749	Mean	11.13678	Mean	0.554695
Standard Error	0.12986	Standard E	1.25137	Standard E	0.30498	Standard E	0.005151
Median	4.82	Median	77.5	Median	9.69	Median	0.538
Mode	3.43	Mode	100	Mode	18.1	Mode	0.538
Standard Devia	2.921132	Standard D	28.14886	Standard D	6.860353	Standard D	0.115878
Sample Varianc	8.533012	Sample Va	792.3584	Sample Va	47.06444	Sample Va	0.013428
Kurtosis	-1.18912	Kurtosis	-0.96772	Kurtosis	-1.23354	Kurtosis	-0.06467
Skewness	0.021728	Skewness	-0.59896	Skewness	0.295022	Skewness	0.729308
Range	9.95	Range	97.1	Range	27.28	Range	0.486
Minimum	0.04	Minimum	2.9	Minimum	0.46	Minimum	0.385
Maximum	9.99	Maximum	100	Maximum	27.74	Maximum	0.871
Sum	2465.22	Sum	34698.9	Sum	5635.21	Sum	280.6757
Count	506	Count	506	Count	506	Count	506



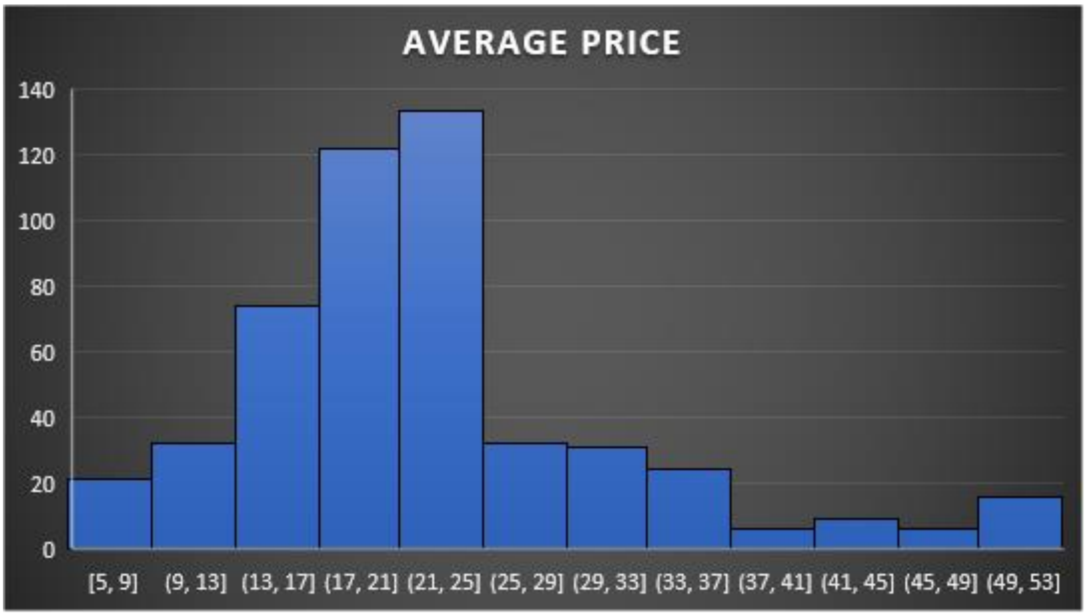
DISTANCE		*TAX*		*PTRATIO*	
Mean	9.549407	Mean	408.2372	Mean	18.45553
Standard Error	0.387085	Standard Error	7.492389	Standard Error	0.096244
Median	5	Median	330	Median	19.05
Mode	24	Mode	666	Mode	20.2
Standard Deviation	8.707259	Standard Deviation	168.5371	Standard Deviation	2.164946
Sample Variance	75.81637	Sample Variance	28404.76	Sample Variance	4.686989
Kurtosis	-0.86723	Kurtosis	-1.14241	Kurtosis	-0.28509
Skewness	1.004815	Skewness	0.669956	Skewness	-0.80232
Range	23	Range	524	Range	9.4
Minimum	1	Minimum	187	Minimum	12.6
Maximum	24	Maximum	711	Maximum	22
Sum	4832	Sum	206568	Sum	9338.5
Count	506	Count	506	Count	506

From the above table we summarise the statistics with the required variables Mean shows the average the column ,Standard error shows the error value ,Median and Mode shows the centre and most frequent values ,Standard Deviation shows the deviation of the value from -ve to +ve and Kurtosis shows the curve and its peakness and skewness shows the most of the value lies in the curve ,Min and Max shows the maximum and minimum value from the column ,range shows the total value lies and Sum and Count shows the total and count of the value present in the column or variable.



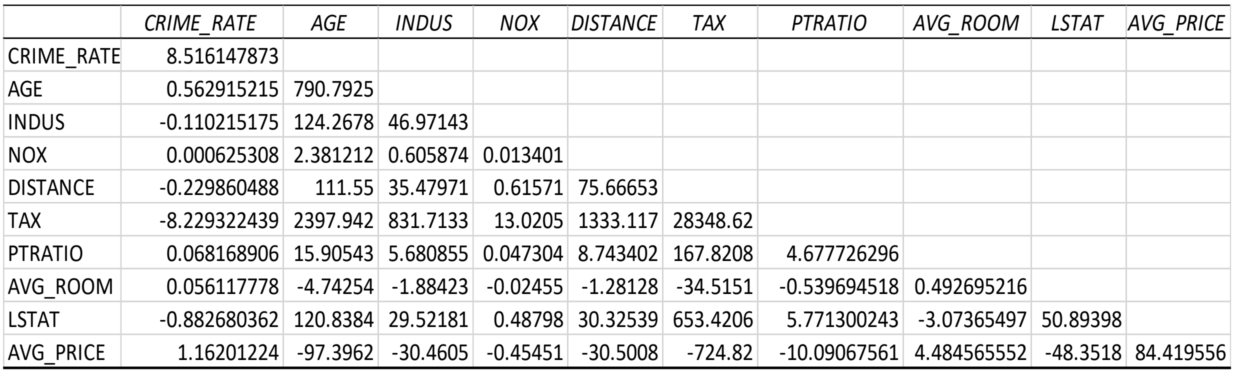
AVG_ROOM		*LSTAT*		*AVG_PRICE*	
Mean	6.284634	Mean	12.65306	Mean	22.53281
Standard Error	0.031235	Standard Error	0.317459	Standard Error	0.408861
Median	6.2085	Median	11.36	Median	21.2
Mode	5.713	Mode	8.05	Mode	50
Standard Deviation	0.702617	Standard Deviatio	7.141062	Standard Deviation	9.197104
Sample Variance	0.493671	Sample Variance	50.99476	Sample Variance	84.58672
Kurtosis	1.8915	Kurtosis	0.49324	Kurtosis	1.495197
Skewness	0.403612	Skewness	0.90646	Skewness	1.108098
Range	5.219	Range	36.24	Range	45
Minimum	3.561	Minimum	1.73	Minimum	5
Maximum	8.78	Maximum	37.97	Maximum	50
Sum	3180.025	Sum	6402.45	Sum	11401.6
Count	506	Count	506	Count	506

Q2: Plot a histogram of the Avg_Price variable. What do you infer?



It as a left skewness ,so on the left of the mean has more value.So it is positive.

Q3: Compute the covariance matrix. Share your observations.

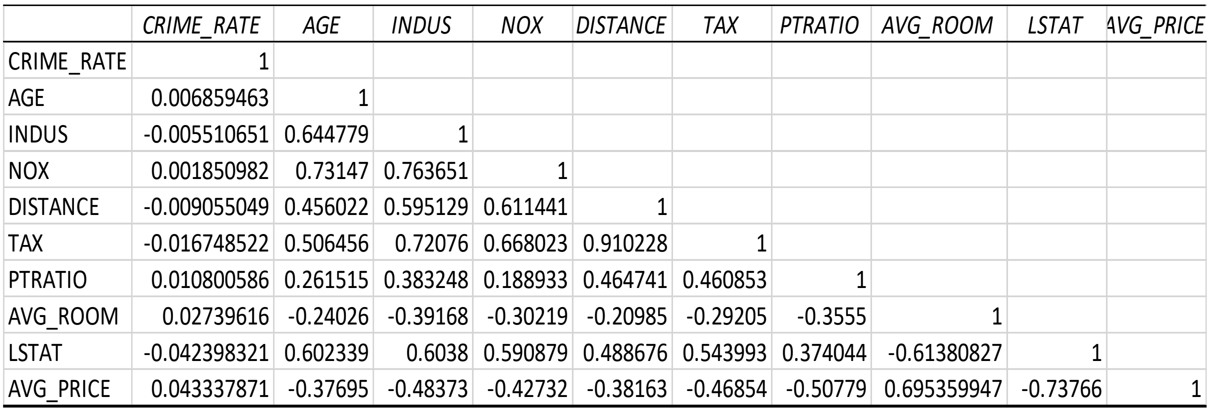


	CRIME_RATE	*AGE*	*INDUS*	*NOX*	*DISTANCE*	*TAX*	*PTRATIO*	*AVG_ROOM*	*LSTAT*	*AVG_PRICE*
CRIME_RATE	8.516147873									
AGE	0.562915215	790.7925								
INDUS	-0.110215175	124.2678	46.97143							
NOX	0.000625308	2.381212	0.605874	0.013401						
DISTANCE	-0.229860488	111.55	35.47971	0.61571	75.66653					
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	-3.07365497	50.89398	
AVG_PRICE	1.16201224	-97.3962	-30.4605	-0.45451	-30.5008	-724.82	-10.09067561	4.484565552	-48.3518	84.419556

Positive value shows that the above variables has a co-variance between x and y.

Negative value shows that the above variables has not have co-variance or less between x and y.

Q4: 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	-0.3555	1		
LSTAT	-0.042398321	0.602339	0.6038	0.590879	0.488676	0.543993	0.374044	-0.61380827	1	
AVG_PRICE	0.043337871	-0.37695	-0.48373	-0.42732	-0.38163	-0.46854	-0.50779	0.695359947	-0.73766	1

a) Which are the top 3 positively correlated pairs

0.910228, 0.763651, 0.73147 are the 3 positively correlated pairs .

Positively correlated values are find according to the combination values less than or equal to one .

Tax and Distance are highly correlated compare to other variables, Nox and Indus are 2nd highly correlated compare to other variables then Nox and Age as 3rd highly correlated to other variables.

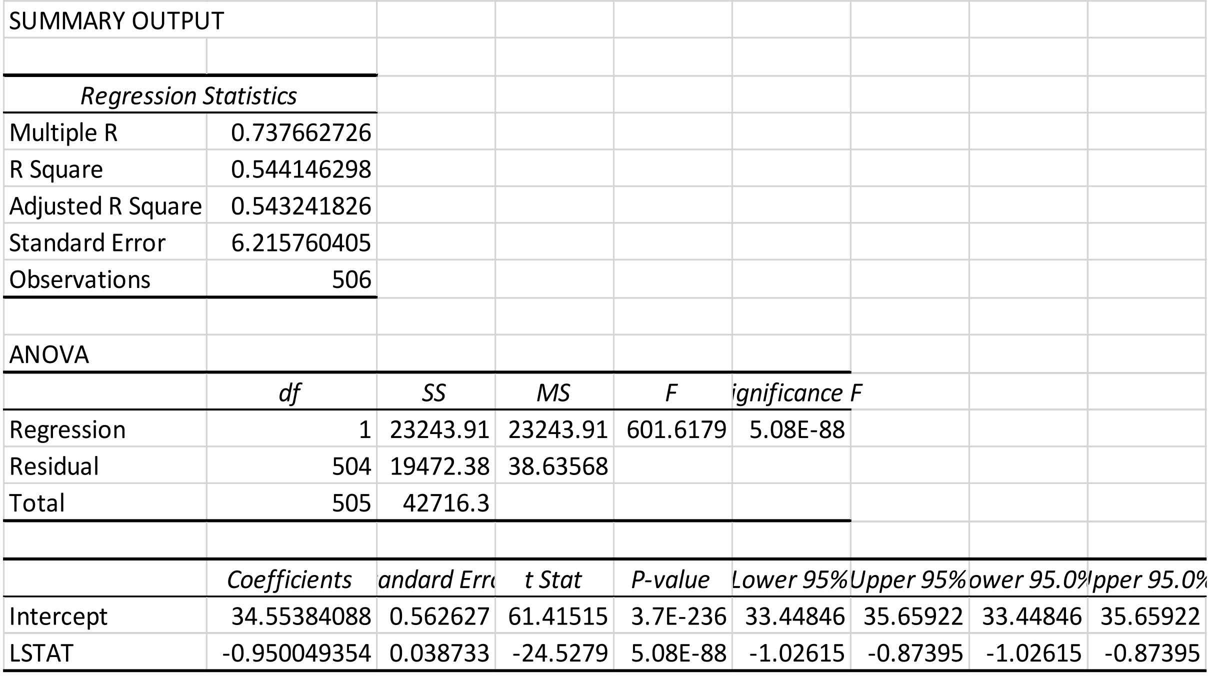
b) Which are the top 3 negatively correlated pairs.

Negatively correlated values are find according to the combination values less than zero.

-0.73766, -0.50779, -0.48373 are 3 negatively correlated pairs.

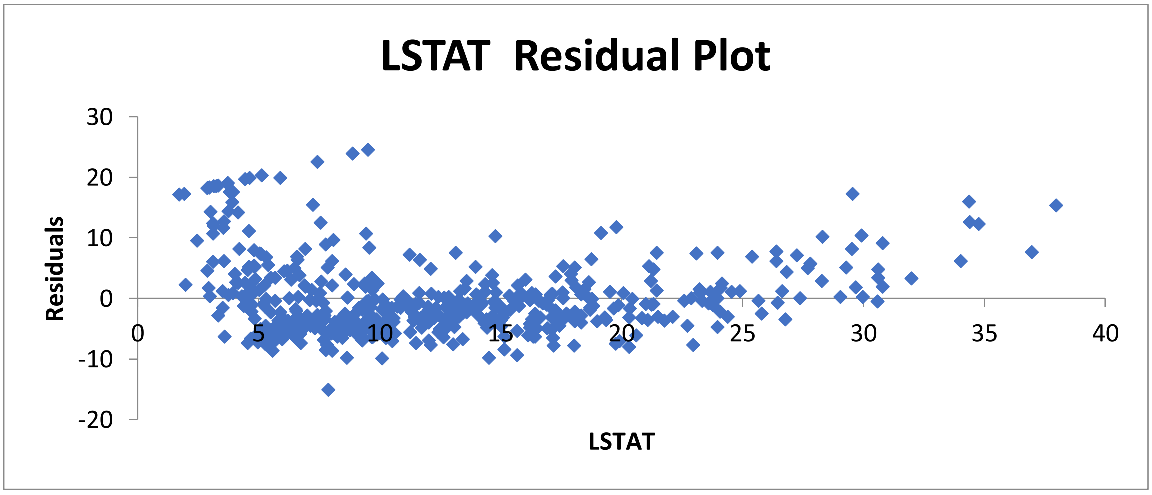
Lstat and Avg_price are high negatively correlated compare to other variables, PTratio and Avg_price are 2nd negatively correlated compare to other variables, then Indus and Avg_price are 3rd negatively correlated to other variables.

Q5: 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*	*ignificance F*			
Regression	1	23243.91	23243.91	601.6179	5.08E-88			
Residual	504	19472.38	38.63568					
Total	505	42716.3						
	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	34.55384088	0.562627	61.41515	3.7E-236	33.44846	35.65922	33.44846	35.65922
LSTAT	-0.950049354	0.038733	-24.5279	5.08E-88	-1.02615	-0.87395	-1.02615	-0.87395

Graph



a) What do you infer from the Regression Summary output in terms of variance explained, coefficient value, Intercept, and Residual plot?

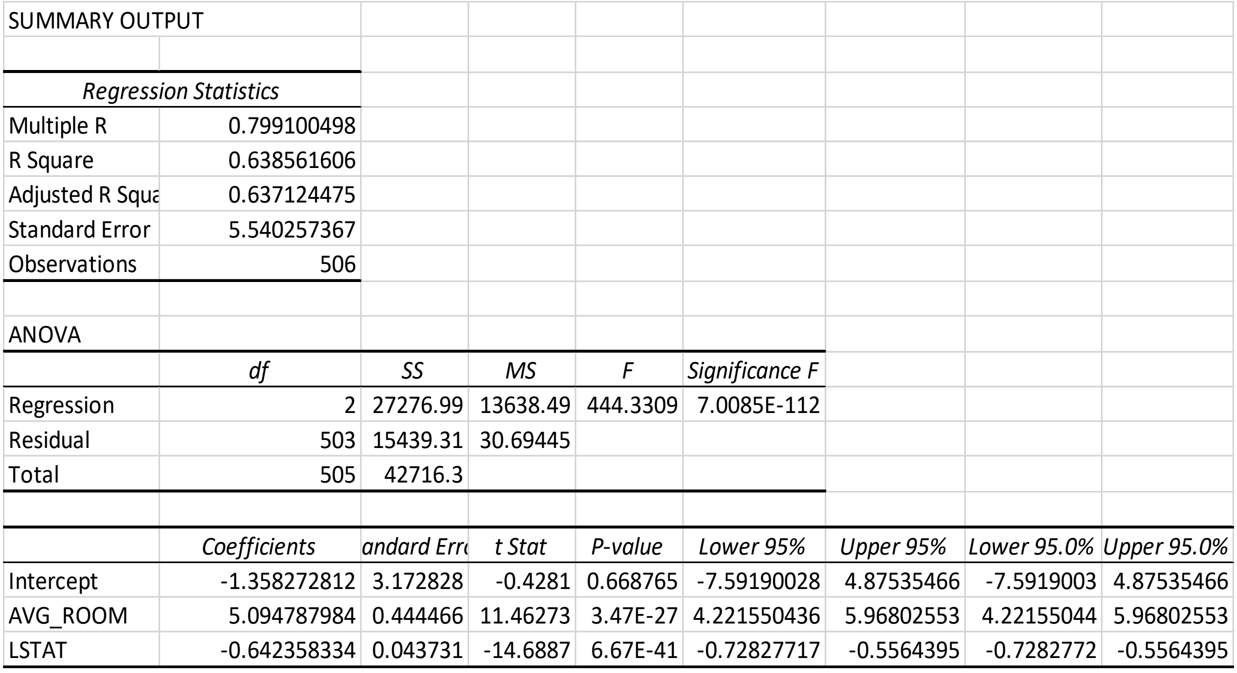
The R-square value of this model is below 0.7 ,So it is not good prediction because the R-square value have to cross atleast 0.7 but the best is 1 or 0.9.

The coefficient of Lstat is -0.95005 .it is inferred that each 1000 increases but there is a decrease in population.

b) Is LSTAT variable significant for the analysis based on your model?

The p-value for Lstat variable is 5.08E-88 which is good value it has p-value less than 0.05.So this variable is significant value for the analysis.

Q6: Build a new Regression model including LSTAT and AVG_ROOM together as Independent variables and AVG_PRICE as dependent variable



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.799100498
R Square	0.638561606
Adjusted R Squa	0.637124475
Standard Error	5.540257367
Observations	506

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	27276.99	13638.49	444.3309	7.0085E-112
Residual	503	15439.31	30.69445		
Total	505	42716.3			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-1.358272812	3.172828	-0.4281	0.668765	-7.59190028	4.87535466	-7.5919003	4.87535466
AVG_ROOM	5.094787984	0.444466	11.46273	3.47E-27	4.221550436	5.96802553	4.22155044	5.96802553
LSTAT	-0.642358334	0.043731	-14.6887	6.67E-41	-0.72827717	-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?

Regression equation =intercept+(coefficient of Avg_room*Avg_room value) +(coefficient of Lstat*Lstat value)

=B17+7*B18+20*B19   (in excel)

=21.4580764 USD

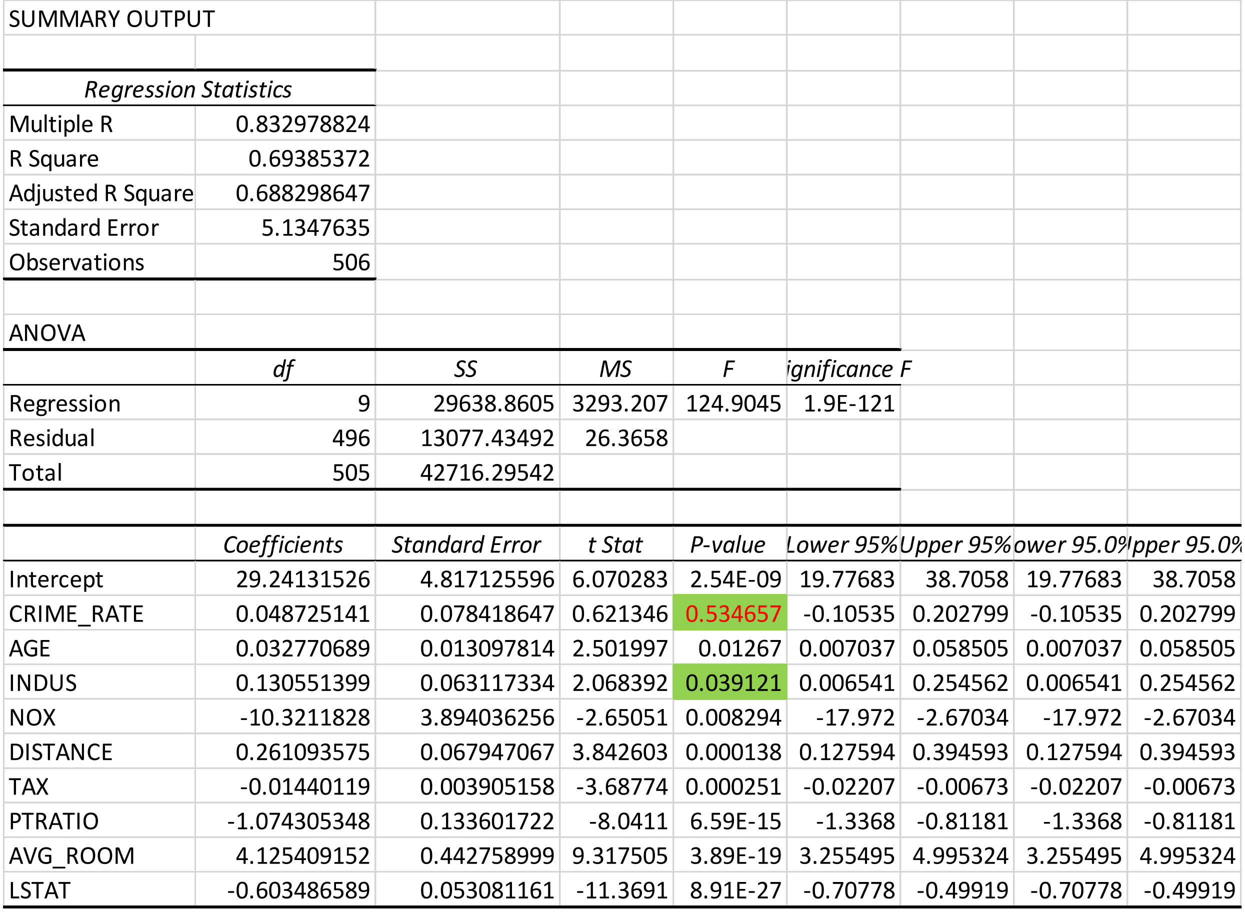
Therefore the value get from the regression model is 21.4580764 USD

Compared to the company quoting a value of 30000 USD and we can say that the company quoted value is overcharged, because 30000 is greater than 210000 USD.

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

yes ,the performance of this model is better than previous model ,because the R-square and Adjusted R-square value are greater in this model compare to the previous model.

Q7: 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*	*ignificance F*
Regression	9	29638.8605	3293.207	124.9045	1.9E-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%*	*pper 95.0%*
Intercept	29.24131526	4.817125596	6.070283	2.54E-09	19.77683	38.7058	19.77683	38.7058
CRIME_RATE	0.048725141	0.078418647	0.621346	0.534657	-0.10535	0.202799	-0.10535	0.202799
AGE	0.032770689	0.013097814	2.501997	0.01267	0.007037	0.058505	0.007037	0.058505
INDUS	0.130551399	0.063117334	2.068392	0.039121	0.006541	0.254562	0.006541	0.254562
NOX	-10.3211828	3.894036256	-2.65051	0.008294	-17.972	-2.67034	-17.972	-2.67034
DISTANCE	0.261093575	0.067947067	3.842603	0.000138	0.127594	0.394593	0.127594	0.394593
TAX	-0.01440119	0.003905158	-3.68774	0.000251	-0.02207	-0.00673	-0.02207	-0.00673
PTRATIO	-1.074305348	0.133601722	-8.0411	6.59E-15	-1.3368	-0.81181	-1.3368	-0.81181
AVG_ROOM	4.125409152	0.442758999	9.317505	3.89E-19	3.255495	4.995324	3.255495	4.995324
LSTAT	-0.603486589	0.053081161	-11.3691	8.91E-27	-0.70778	-0.49919	-0.70778	-0.49919

Crime_rate: The variable crime_rate as very high significant value which is greater than p-value (i.e,) negative that affect the R-square value less than 0.5. This reduce the Adjusted R-square.

Age: The variable age as perfect significant value (i.e,) p-value is less than 0.05 is positive, so it contribute in the growth of Adjusted R-square.

Indus: The variable age as perfect significant value (i.e,) p-value is less than 0.05 is positive, so it contribute in the growth of Adjusted R-square.

Iox: The variable age as perfect significant value (i.e,) p-value is less than 0.05 is positive, so it contribute in the growth of Adjusted R-square.

Distance: The variable age as perfect significant value (i.e,) p-value is less than 0.05 is positive, so it contribute in the growth of Adjusted R-square.

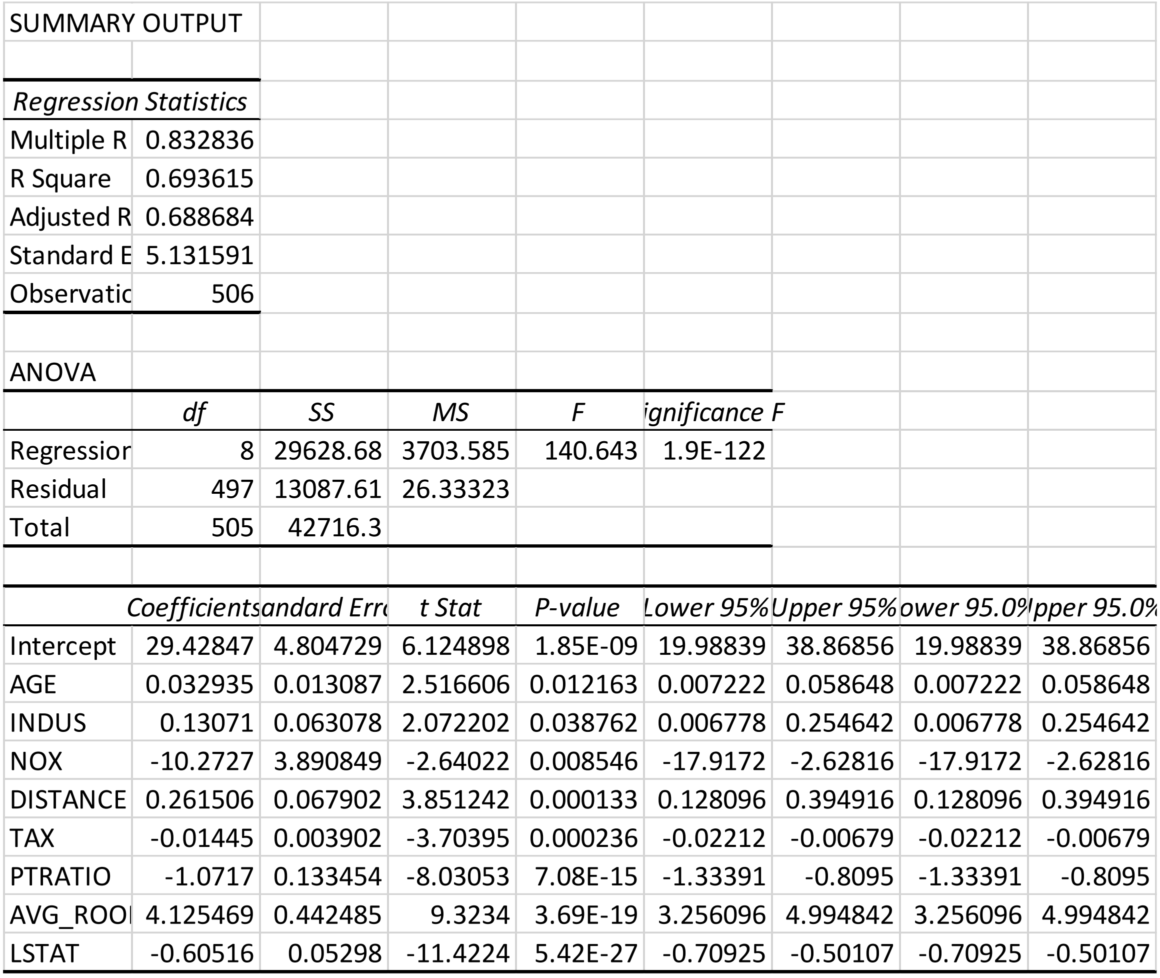
Tax: The variable age as perfect significant value (i.e,) p-value is less than 0.05 is positive, so it contribute in the growth of Adjusted R-square.

PTratio: The variable age as perfect significant value (i.e,) p-value is less than 0.05 is positive, so it contribute in the growth of Adjusted R-square.

Avg_room: The variable age as perfect significant value (i.e,) p-value is less than 0.05 is positive, so it contribute in the growth of Adjusted R-square.

Avg_price: The variable age as perfect significant value (i.e,) p-value is less than 0.05 is positive, so it contribute in the growth of Adjusted R-square.

Q8) 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:



SUMMARY OUTPUT

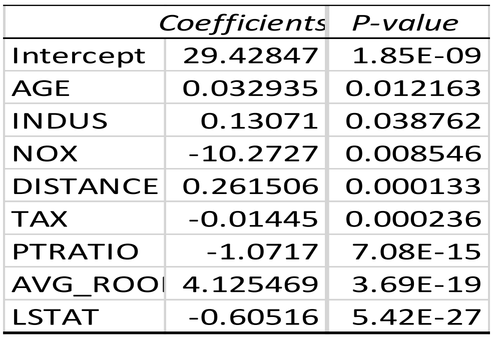
Regression Statistics	
Multiple R	0.832836
R Square	0.693615
Adjusted R	0.688684
Standard E	5.131591
Observatic	506

ANOVA

	df	SS	MS	F	ignificance F
Regressior	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	29.42847	4.804729	6.124898	1.85E-09	19.98839	38.86856	19.98839	38.86856
AGE	0.032935	0.013087	2.516606	0.012163	0.007222	0.058648	0.007222	0.058648
INDUS	0.13071	0.063078	2.072202	0.038762	0.006778	0.254642	0.006778	0.254642
NOX	-10.2727	3.890849	-2.64022	0.008546	-17.9172	-2.62816	-17.9172	-2.62816
DISTANCE	0.261506	0.067902	3.851242	0.000133	0.128096	0.394916	0.128096	0.394916
TAX	-0.01445	0.003902	-3.70395	0.000236	-0.02212	-0.00679	-0.02212	-0.00679
PTRATIO	-1.0717	0.133454	-8.03053	7.08E-15	-1.33391	-0.8095	-1.33391	-0.8095
AVG_ROOI	4.125469	0.442485	9.3234	3.69E-19	3.256096	4.994842	3.256096	4.994842
LSTAT	-0.60516	0.05298	-11.4224	5.42E-27	-0.70925	-0.50107	-0.70925	-0.50107

a) Interpret the output of this model.



	Coefficients	P-value
Intercept	29.42847	1.85E-09
AGE	0.032935	0.012163
INDUS	0.13071	0.038762
NOX	-10.2727	0.008546
DISTANCE	0.261506	0.000133
TAX	-0.01445	0.000236
PTRATIO	-1.0717	7.08E-15
AVG_ROOI	4.125469	3.69E-19
LSTAT	-0.60516	5.42E-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?

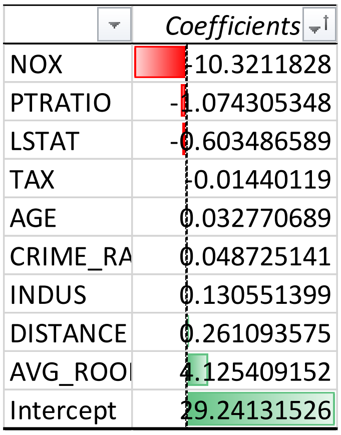
Adjusted R-square = 0.688684

Adjusted R-square = 0.688298

The above are adjusted R-square value this models value is more efficient than the previous model,

Because it has some not significant value in the previous and this model has a perfect Adjusted R-square value , it reaches almost 0.7.

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?



	Coefficients
NOX	-10.3211828
PTRATIO	-1.074305348
LSTAT	-0.603486589
TAX	-0.01440119
AGE	0.032770689
CRIME_RA	0.048725141
INDUS	0.130551399
DISTANCE	0.261093575
AVG_ROOI	4.125409152
Intercept	29.24131526

The red range shows the Nox value is more locality in this town, the value in green range is less locality in this town .

d) Write the regression equation from this model.

Regression Equation

Avg_price=Intercept+(coeff of Age*Age value)+ (coeff of Indus*Indus value) +(coeff of *Nox

value) +(coeff of Diatance*Distance value) +(coeff of Tax*Tax value) +(coeff of

PTratio*PTratio value) +(coeff of Avg_room*Avg_room value) +

(coeff of Lstat*Lstat value)