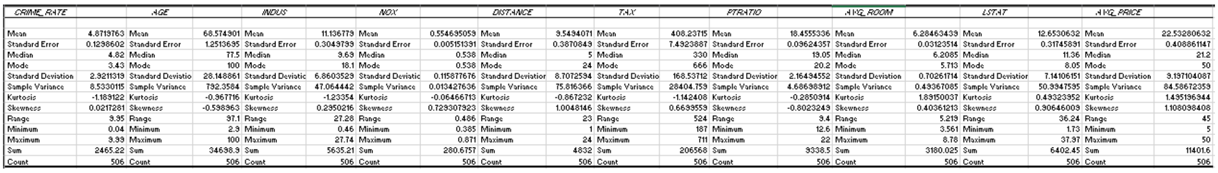
Business report

Project-Terro's Real Estate Agency

Mohammed Toufiq

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?*



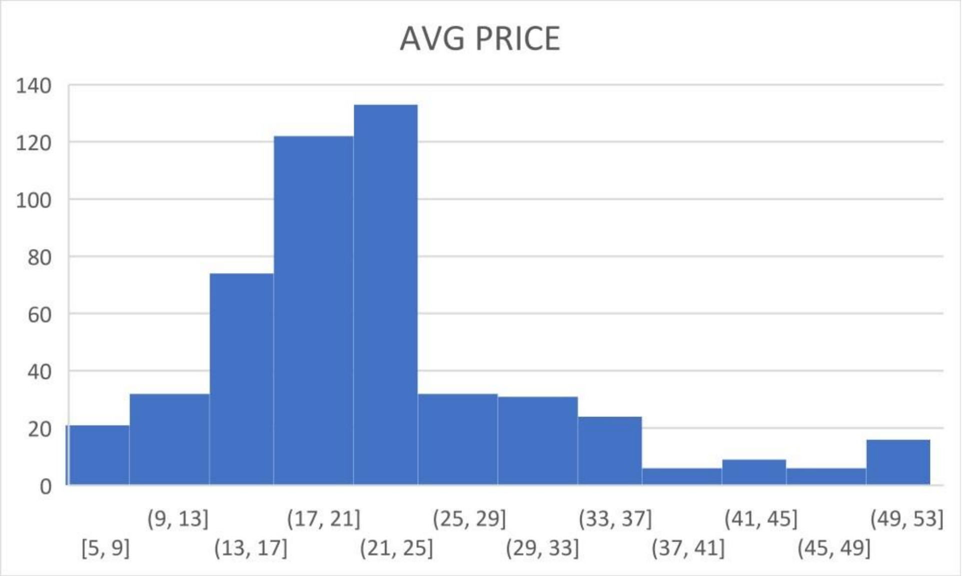
CRIME_RATE		AGE		INDUS		NOX		DISTANCE		TAX		PTRATIO		AVG_ROOM		LSTAT		AVG_PRICE	
Mean	4.8719763	Mean	68.574901	Mean	11.136779	Mean	0.554695059	Mean	9.5494071	Mean	408.23715	Mean	18.4555336	Mean	6.28463439	Mean	12.6530632	Mean	22.53280632
Standard Error	0.1298602	Standard Error	1.2513695	Standard Error	0.3049799	Standard Error	0.005151391	Standard Error	0.3870849	Standard Error	7.4923887	Standard Error	0.09624357	Standard Error	0.03123514	Standard Error	0.31745891	Standard Error	0.408861147
Median	4.82	Median	77.5	Median	9.69	Median	0.538	Median	5	Median	330	Median	19.05	Median	6.2085	Median	11.36	Median	21.2
Mode	3.43	Mode	100	Mode	18.1	Mode	0.538	Mode	24	Mode	666	Mode	20.2	Mode	5.713	Mode	8.05	Mode	50
Standard Deviation	2.9211319	Standard Deviatio	28.148861	Standard Deviatic	6.8603529	Standard Deviatic	0.115877676	Standard Deviatior	8.7072594	Standard Deviatio	168.53712	Standard Deviatior	2.16494552	Standard Deviation	0.70261714	Standard Deviatior	7.14106151	Standard Deviation	9.197104087
Sample Variance	8.5330115	Sample Variance	792.3584	Sample Variance	47.064442	Sample Variance	0.013427636	Sample Variance	75.816366	Sample Variance	28404.759	Sample Variance	4.68698912	Sample Variance	0.49367085	Sample Variance	50.9947595	Sample Variance	84.58672359
Kurtosis	-1.189122	Kurtosis	-0.967716	Kurtosis	-1.23354	Kurtosis	-0.06466713	Kurtosis	-0.867232	Kurtosis	-1.142408	Kurtosis	-0.2850914	Kurtosis	1.89150037	Kurtosis	0.49323952	Kurtosis	1.495196944
Skewness	0.0217281	Skewness	-0.598963	Skewness	0.2950216	Skewness	0.729307923	Skewness	1.0048146	Skewness	0.6699559	Skewness	-0.8023249	Skewness	0.40361213	Skewness	0.90646009	Skewness	1.108098408
Range	9.95	Range	97.1	Range	27.28	Range	0.486	Range	23	Range	524	Range	9.4	Range	5.219	Range	36.24	Range	45
Minimum	0.04	Minimum	2.9	Minimum	0.46	Minimum	0.385	Minimum	1	Minimum	187	Minimum	12.6	Minimum	3.561	Minimum	1.73	Minimum	5
Maximum	9.99	Maximum	100	Maximum	27.74	Maximum	0.871	Maximum	24	Maximum	711	Maximum	22	Maximum	8.78	Maximum	37.97	Maximum	50
Sum	2465.22	Sum	34698.9	Sum	5635.21	Sum	280.6757	Sum	4832	Sum	206568	Sum	9338.5	Sum	3180.025	Sum	6402.45	Sum	11401.6
Count	506	Count	506	Count	506	Count	506	Count	506	Count	506	Count	506	Count	506	Count	506	Count	506

We have plotted the summary statistics for all the variables above

Key observations

- There are total 506 observations
- Average of Average price of houses is 22532.80$.
- On an average 68% of the houses are built before 1940.

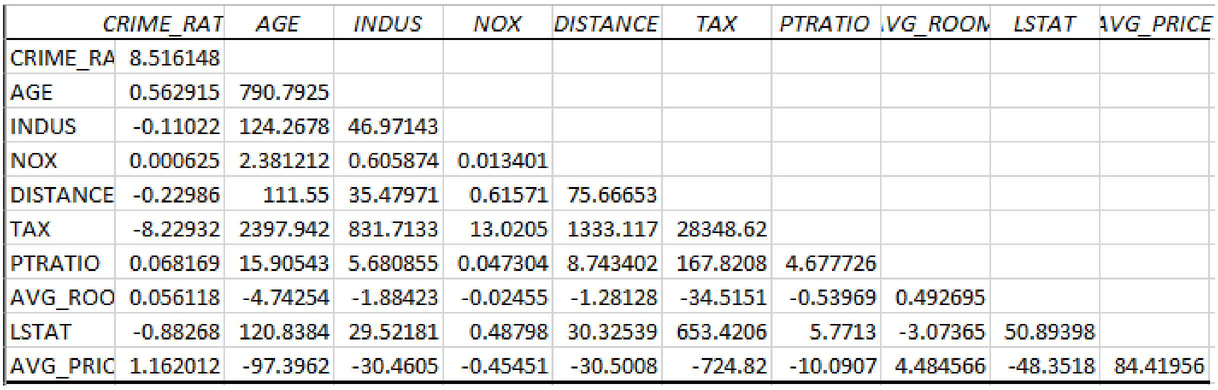
2. *Plot the histogram of the Avg Price Variable. What do you infer?*



From the histogram above we infer

- Most of the houses in the city of Boston has the average price in between 210000-250000 USD

3. Compute the covariance matrix. Share your observations.

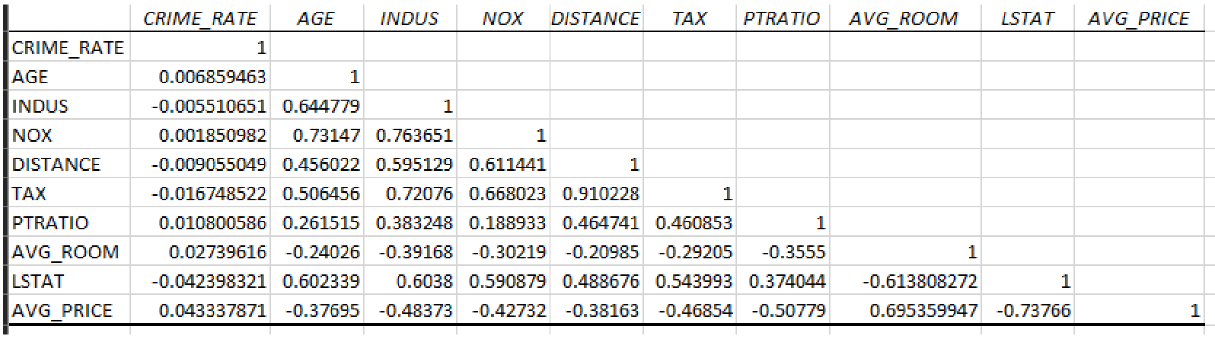


	CRIME_RAT	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	VG_ROOM	LSTAT	AVG_PRICE
CRIME_RA	8.516148									
AGE	0.562915	790.7925								
INDUS	-0.11022	124.2678	46.97143							
NOX	0.000625	2.381212	0.605874	0.013401						
DISTANCE	-0.22986	111.55	35.47971	0.61571	75.66653					
TAX	-8.22932	2397.942	831.7133	13.0205	1333.117	28348.62				
PTRATIO	0.068169	15.90543	5.680855	0.047304	8.743402	167.8208	4.677726			
AVG_ROO	0.056118	-4.74254	-1.88423	-0.02455	-1.28128	-34.5151	-0.53969	0.492695		
LSTAT	-0.88268	120.8384	29.52181	0.48798	30.32539	653.4206	5.7713	-3.07365	50.89398	
AVG_PRIC	1.162012	-97.3962	-30.4605	-0.45451	-30.5008	-724.82	-10.0907	4.484566	-48.3518	84.41956

From the above covariance matrix, we infer that

- There are negative and positive affected numbers in the matrix.
- Positive variables say the positive relation among the variables and negative numbers relate negativity among the variables
- The numbers closer to zero infer no relation among the variables.

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

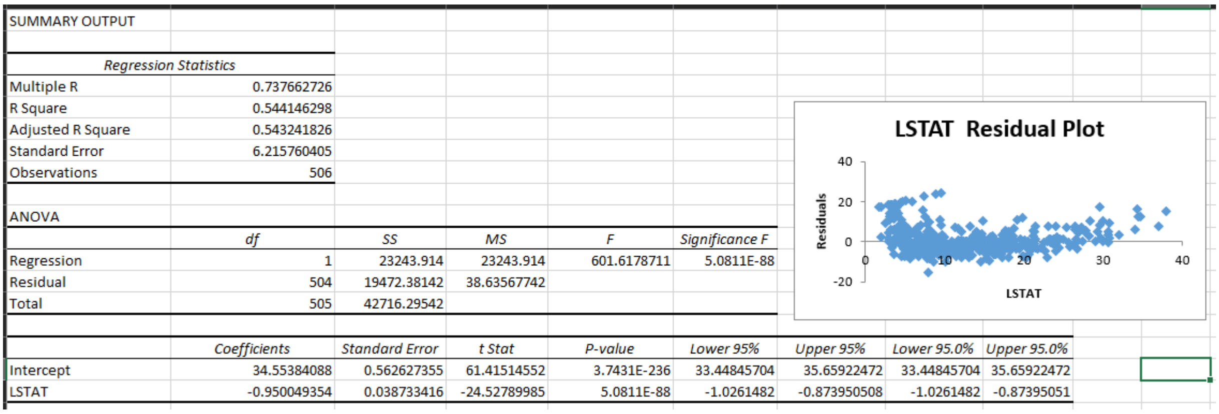
Top 3 positively`` correlated pairs

1. 0.91 - Distance vs TAX
2. 0.76 - Indus vs NOX
3. 0.73 - Age vs NOX

Top 3 negatively correlated pairs

1. -0.74 - LSTAT vs Avg Price:
2. -0.61 – Avg Room vs LSTAT:
3. -0.50 - PRATIO vs Avg Price

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



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	35.65922472
LSTAT	-0.950049354	0.038733416	-24.52789985	5.0811E-88	-1.0261482	-0.873950508	-1.0261482	-0.87395051

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

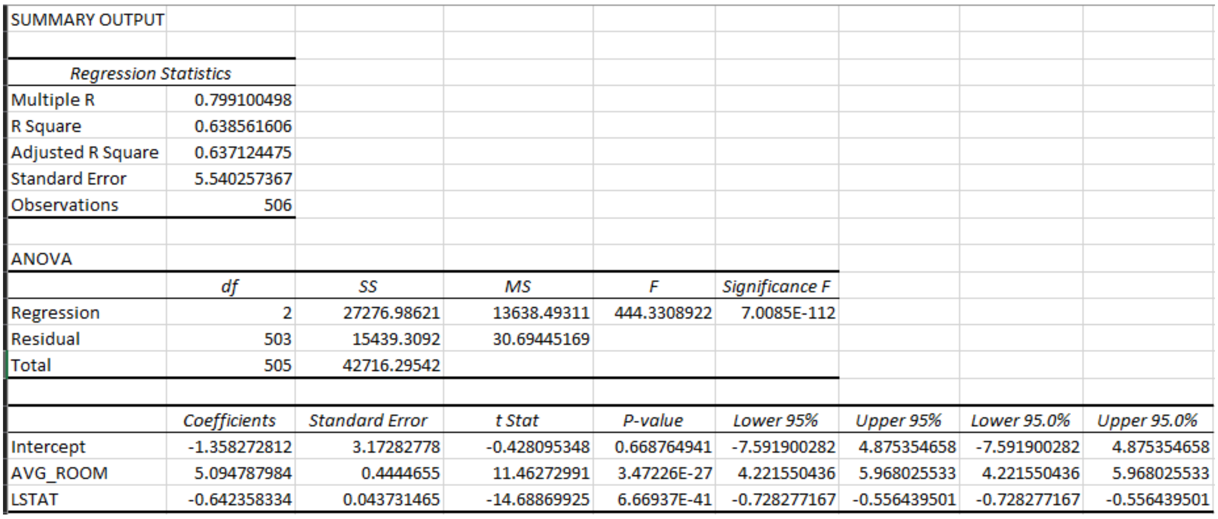
We infer that

- LSTAT is highly related to average price
- The governing equation of the regression line is ○ **AVG_price = -0.95005*(LSTAT)**
- The Y intercept is 34.55384 which means, at LSTAT = 0 our regression line assumes the average price to be 34553.84$.
- **Residual plot**- The residuals lie between positives and negatives of the Y axis

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

Yes, LSTAT value is a significant variable in deciding average price as its p value is way lower than 5%.

6. Build another instance of the Regression model but this time including LSTAT and AVG_ROOM together as independent variables and AVG_PRICE as the dependent variable.



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	-7.591900282	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	-0.728277167	-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?

The regression equation governing this relation would be

Average price = 5.09478*(Average room) – 0.64236*(LSTAT) – 1.35827

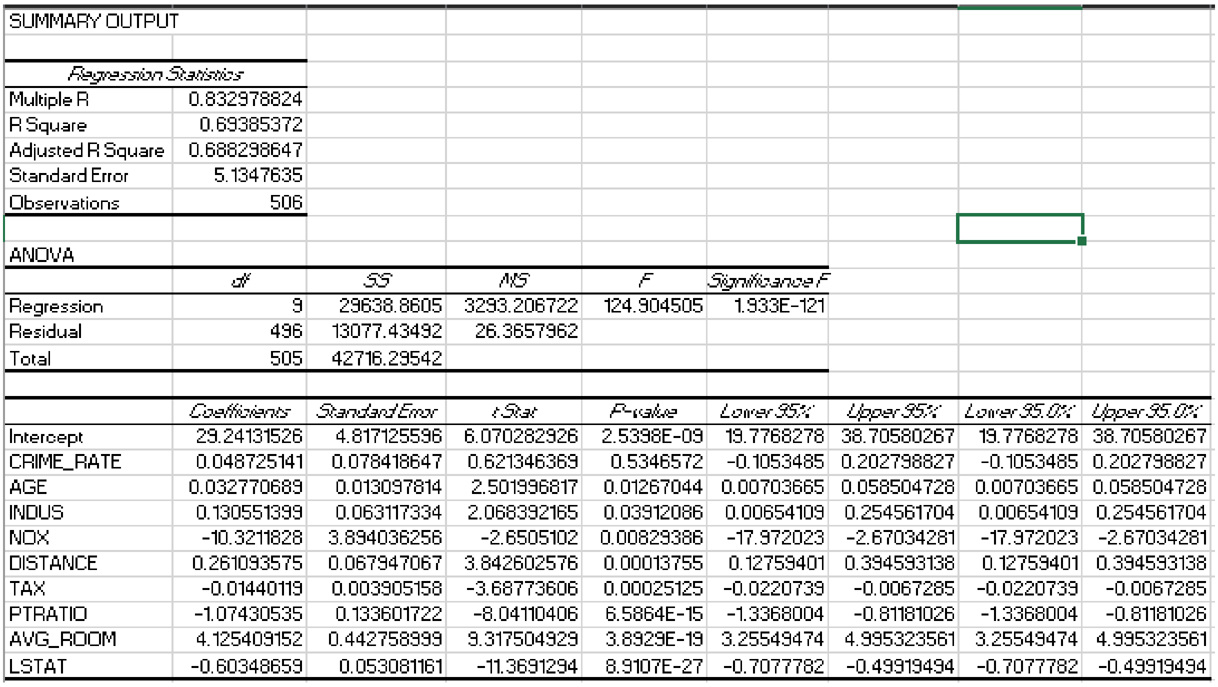
Average price for the above given conditions according to this equation would be 21458 $

The company is Overcharging by 8451 $.

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.

Yes, the performance of this model is clearly better than the previous model. The Adjusted R-square value of this model is 0.637 while for the previous model it was 0.544. The greater the R-square value the better is the fit of the linear regression line.

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	124.904505	1.933E-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	19.7768278	38.70580267	19.7768278	38.70580267
CRIME_RATE	0.048725141	0.078418647	0.621346369	0.5346572	-0.1053485	0.202798827	-0.1053485	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	0.00654109	0.254561704	0.00654109	0.254561704
NOX	-10.3211828	3.894036256	-2.6505102	0.00829386	-17.972023	-2.67034281	-17.972023	-2.67034281
DISTANCE	0.261093575	0.067947067	3.842602576	0.00013755	0.12759401	0.394593138	0.12759401	0.394593138
TAX	-0.01440119	0.003905158	-3.68773606	0.00025125	-0.0220739	-0.0067285	-0.0220739	-0.0067285
PTRATIO	-1.07430535	0.133601722	-8.04110406	6.5864E-15	-1.3368004	-0.81181026	-1.3368004	-0.81181026
AVG_ROOM	4.125409152	0.442758999	9.317504929	3.8929E-19	3.25549474	4.995323561	3.25549474	4.995323561
LSTAT	-0.60348659	0.053081161	-11.3691294	8.9107E-27	-0.7077782	-0.49919494	-0.7077782	-0.49919494

The adjusted R-square value for this model is very high which is 0.688. Hence the linear regression model is a good fit for our data.

The governing equation for this regression model is

Average price = 29.24 + 0.048*(crime rate) + 0.032*(Age) + 0.13*(INDUS) - 10.32*(NOX)+0.26*(Distance)-0.0144*(TAX)-1.074*(PTRATIO) + 4.125*(AVG_ROOM) – 0.6034*(LSTAT)

The variable is insignificant if its p- value is more than 5%, hence here our insignificant variable is

Crime rate

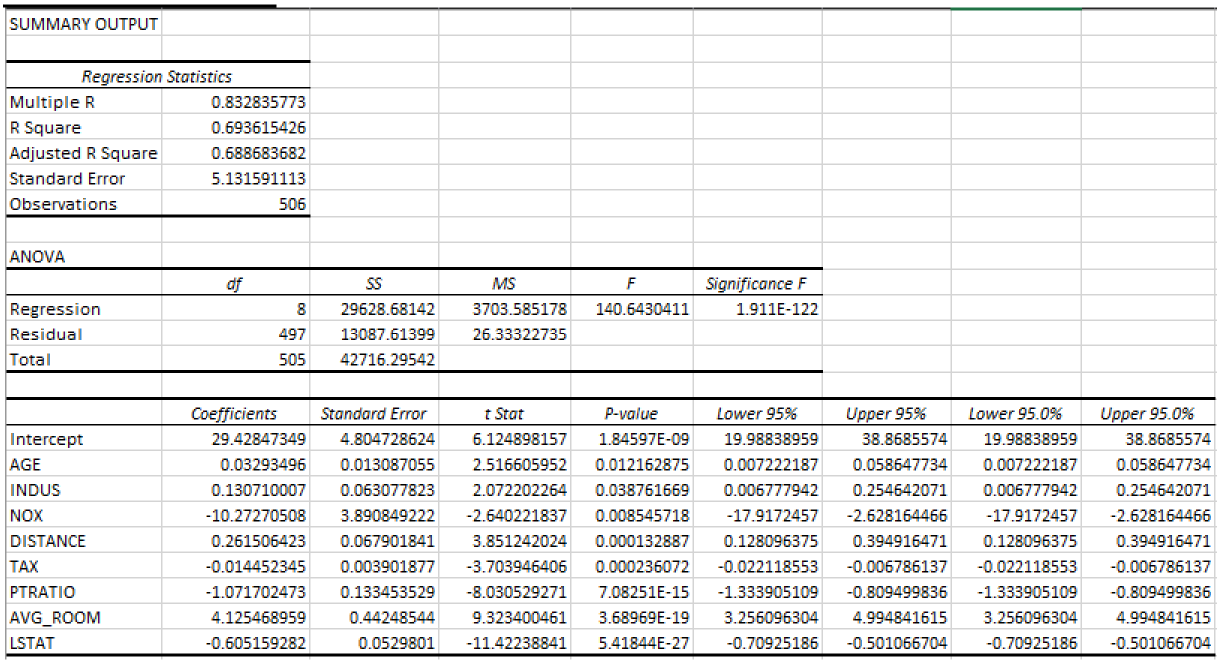
The variable is significant if its p- value is Less than 5%, hence here our significant variables are

- AGE
- INDUS
- NOX
- DISTANCE
- TAX
- PTRATIO
- AVG_ROOM
- LSTAT

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

The variable is significant if its p- value is Less than 5%, hence here our significant variables are

- AGE
- INDUS
- NOX
- DISTANCE
- TAX
- PTRATIO
- AVG_ROOM
- LSTAT



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%*
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. Interpret the output of this model.

The equation governing this model is

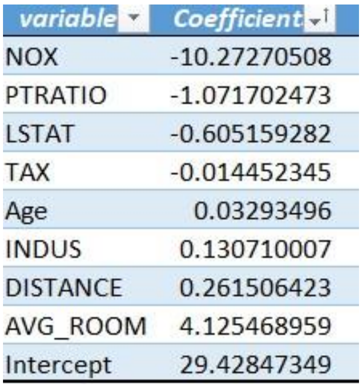
Average price = 29.42 + 0.032*(Age) + 0.13*(INDUS) -10.27*(NOX)+0.26*(Distance)-0.014*(TAX)-1.07*(PTRATIO) + 4.12*(AVG_ROOM) – 0.605*(LSTAT)

This model has all p values less than 5%

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?

The Adjusted R-square value of this model is 0.6888684 while for previous model it was 0.688299. Since the newer model with ignored insignificant values has a higher adjusted RSquared value than the older model. The newer model performs better.

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?***



variable	Coefficient
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
Intercept	29.42847349

Since Coefficient of NOX is lowest (-10.27). **If the Value of NOX is increased, then the average price of house will have a significant amount of decrease.**

d. ***Write the regression equation from this model.***

The equation governing this model is

Average price = 29.42 + 0.032*(Age) + 0.13*(INDUS) -10.27*(NOX)+0.26*(Distance)0.014*(TAX)-1.07*(PTRATIO) + 4.12*(AVG_ROOM) – 0.605*(LSTAT)