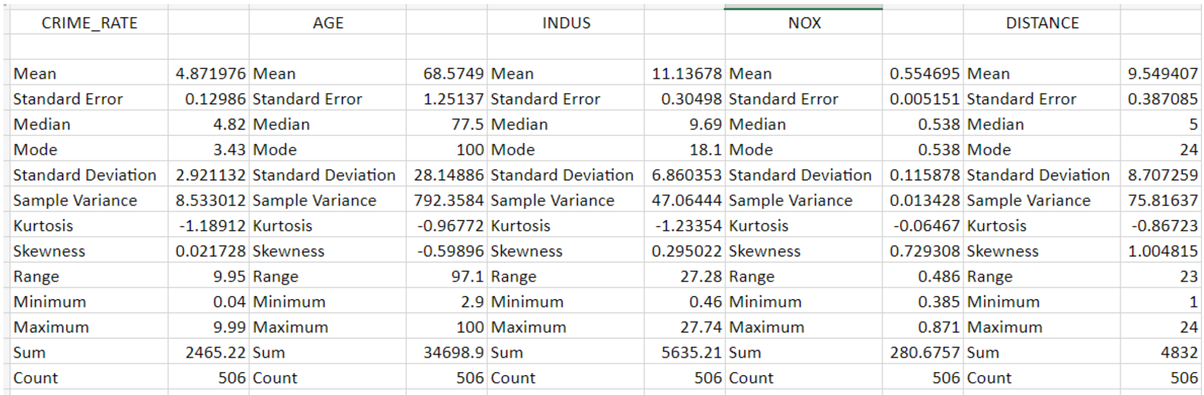
Terro's Real Estate Agency

Problem Statement:

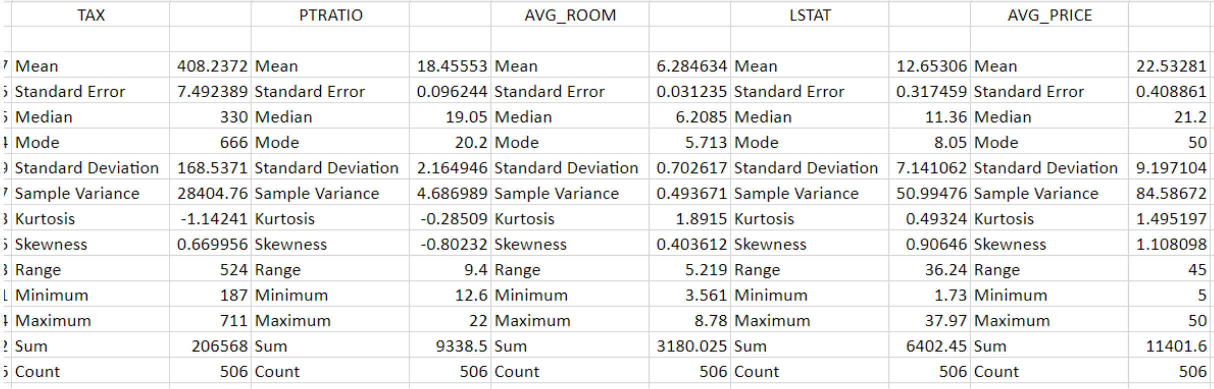
You have been hired at a Terro's Real Estate Agency in the capacity of an Auditor. One of the jobs that the auditor of this agency does is to map all the relevant features of the properties along with the information related to the geography around it. The agency wants to understand the relevance of the parameters that they collect in relation to the value of the house (Avg_Price). You have been given a dataset of 506 houses in Boston.

Question 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	
Mean	4.871976	Mean	68.5749	Mean	11.13678	Mean	0.554695	Mean	9.549407
Standard Error	0.12986	Standard Error	1.25137	Standard Error	0.30498	Standard Error	0.005151	Standard Error	0.387085
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 Deviation	2.921132	Standard Deviation	28.14886	Standard Deviation	6.860353	Standard Deviation	0.115878	Standard Deviation	8.707259
Sample Variance	8.533012	Sample Variance	792.3584	Sample Variance	47.06444	Sample Variance	0.013428	Sample Variance	75.81637
Kurtosis	-1.18912	Kurtosis	-0.96772	Kurtosis	-1.23354	Kurtosis	-0.06467	Kurtosis	-0.86723
Skewness	0.021728	Skewness	-0.59896	Skewness	0.295022	Skewness	0.729308	Skewness	1.004815
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



TAX		PTRATIO		AVG_ROOM		LSTAT		AVG_PRICE	
Mean	408.2372	Mean	18.45553	Mean	6.284634	Mean	12.65306	Mean	22.53281
Standard Error	7.492389	Standard Error	0.096244	Standard Error	0.031235	Standard Error	0.317459	Standard Error	0.408861
Median	330	Median	19.05	Median	6.2085	Median	11.36	Median	21.2
Mode	666	Mode	20.2	Mode	5.713	Mode	8.05	Mode	50
Standard Deviation	168.5371	Standard Deviation	2.164946	Standard Deviation	0.702617	Standard Deviation	7.141062	Standard Deviation	9.197104
Sample Variance	28404.76	Sample Variance	4.686989	Sample Variance	0.493671	Sample Variance	50.99476	Sample Variance	84.58672
Kurtosis	-1.14241	Kurtosis	-0.28509	Kurtosis	1.8915	Kurtosis	0.49324	Kurtosis	1.495197
Skewness	0.669956	Skewness	-0.80232	Skewness	0.403612	Skewness	0.90646	Skewness	1.108098
Range	524	Range	9.4	Range	5.219	Range	36.24	Range	45
Minimum	187	Minimum	12.6	Minimum	3.561	Minimum	1.73	Minimum	5
Maximum	711	Maximum	22	Maximum	8.78	Maximum	37.97	Maximum	50
Sum	206568	Sum	9338.5	Sum	3180.025	Sum	6402.45	Sum	11401.6
Count	506	Count	506	Count	506	Count	506	Count	506

From descriptive statistics of the given dataset, we can get few observations as:

There are total of 506 records in the dataset

If we look at the “Distance” variable, we can see that the maximum distance is 24 and the mode is 24. According to which, most houses are located away from the highway.

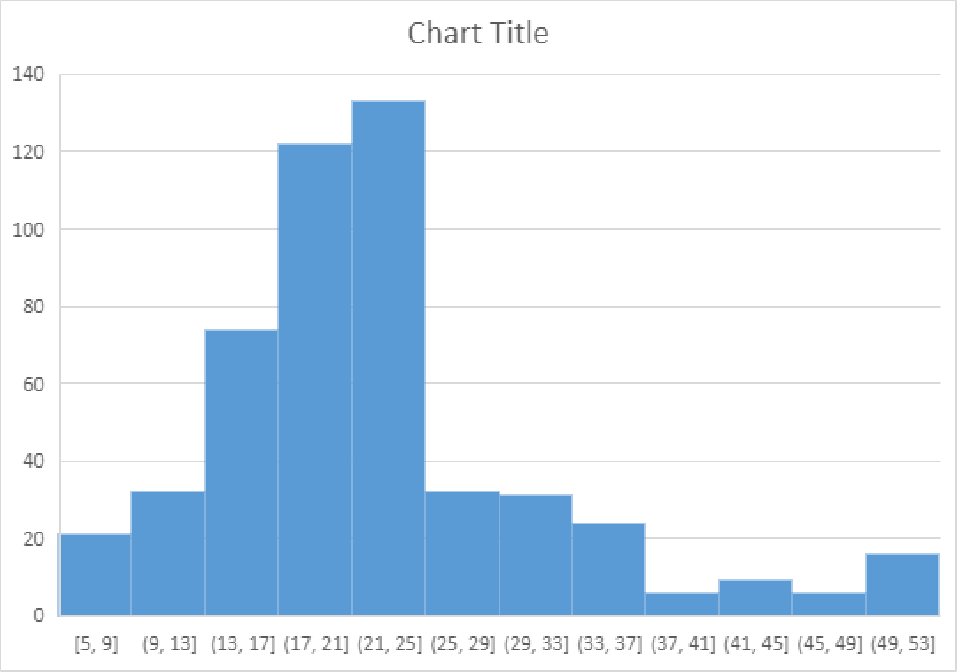
With the tax range of 524, the average tax paid is 408.2

From the skewness of variables, we can say that dataset is highly skewed.

If we take the “Age” into consideration, we observe that maximum age is 100 and mode age is 100 too. This indicates that most of the houses are of age 100 and above

Question 2

Plot the histogram of the Avg_Price Variable. What do you infer?



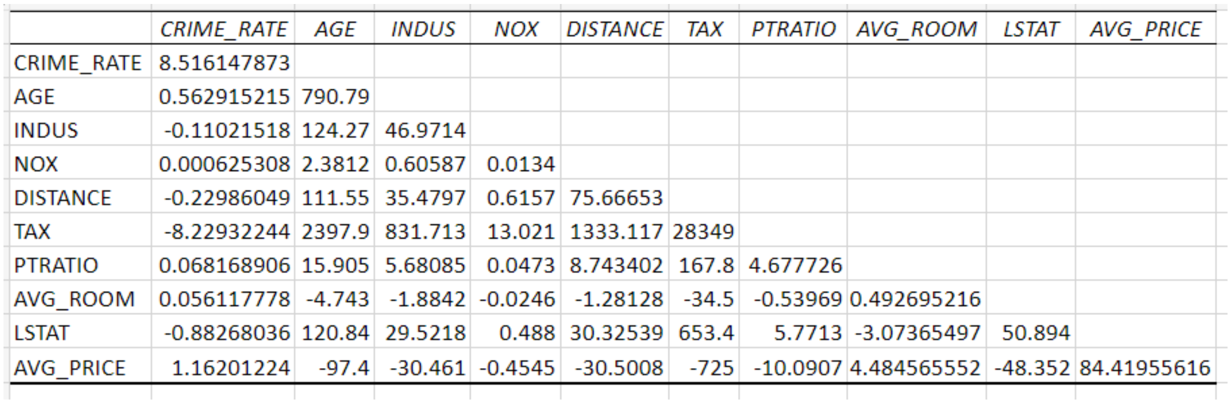
From the histogram:

Majority of the houses lie in the range $21,000 to $25,000

Least number of hiuses range from $37,000 to $41,000 and from $45,000 to $49,000

Question 3:

Compute the covariance matrix. Share your observations.

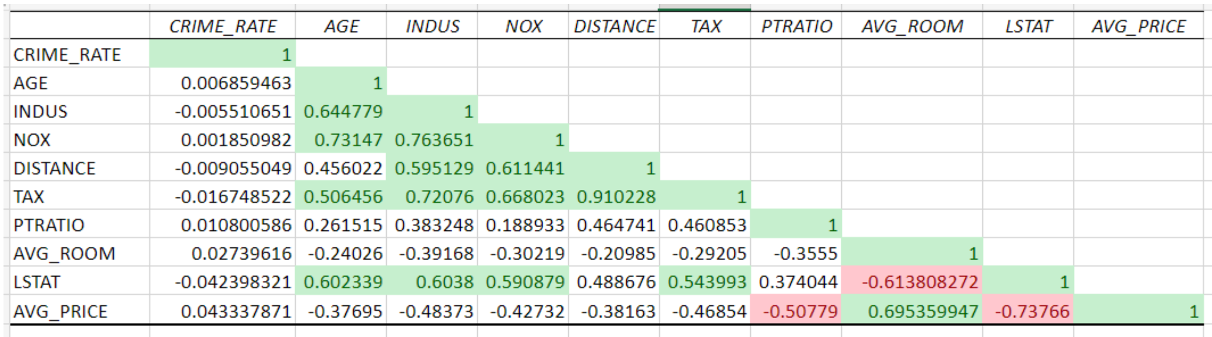


	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RATE	8.516147873									
AGE	0.562915215	790.79								
INDUS	-0.11021518	124.27	46.9714							
NOX	0.000625308	2.3812	0.60587	0.0134						
DISTANCE	-0.22986049	111.55	35.4797	0.6157	75.66653					
TAX	-8.22932244	2397.9	831.713	13.021	1333.117	28349				
PTRATIO	0.068168906	15.905	5.68085	0.0473	8.743402	167.8	4.677726			
AVG_ROOM	0.056117778	-4.743	-1.8842	-0.0246	-1.28128	-34.5	-0.53969	0.492695216		
LSTAT	-0.88268036	120.84	29.5218	0.488	30.32539	653.4	5.7713	-3.07365497	50.894	
AVG_PRICE	1.16201224	-97.4	-30.461	-0.4545	-30.5008	-725	-10.0907	4.484565552	-48.352	84.41955616

- CRIME_RATE has a relatively high correlation with AGE, INDUS, and LSTAT, indicating that these variables frequently move together. This shows that places with greater crime rates may also have older homes, more industrial land, and a higher proportion of low-income occupants.
- Strong correlation between TAX and INDUS shows that places with more industrial land typically have higher real estate taxes.
- DISTANCE has a strong negative covariance with AVG_ROOM, indicating that houses closer to certain amenities tend to have more rooms.
- LSTAT has a strong negative covariance with AVG_PRICE, suggesting that areas with a higher percentage of lower-income residents tend to have lower housing prices.
- Strong correlation between AVG_PRICE and TAX indicates that average home prices are generally higher in locations with higher property taxes.
- NOX has a strong correlation with AGE, INDUS, and DISTANCE, indicating that locations with greater nitrogen oxide emissions are more likely to have older homes, more industrial land, and be farther from certain amenities.

Question 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:

From above correlation matrix we can analyze the top 3 positively correlated pairs as

1.Distance – Tax

2.NOX – Age

3.NOX – Indus

Top 3 negatively correlated pairs

From above correlation matrix we can analyse the top 3 negatively correlated pairs as

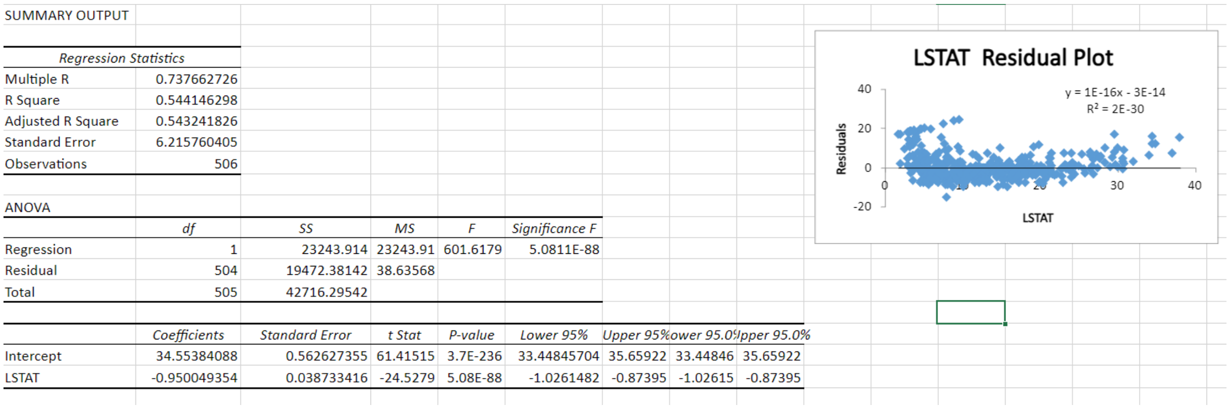
1. LSTAT – Avg_Room

2. Avg_Price – PTRATIO

3. Avg_Price – LSTAT

Question 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	23243.91	601.6179	5.0811E-88
Residual	504	19472.38142	38.63568		
Total	505	42716.29542			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	lpper 95.0%
Intercept	34.55384088	0.562627355	61.41515	3.7E-236	33.44845704	35.65922	33.44846	35.65922
LSTAT	-0.950049354	0.038733416	-24.5279	5.08E-88	-1.0261482	-0.87395	-1.02615	-0.87395

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

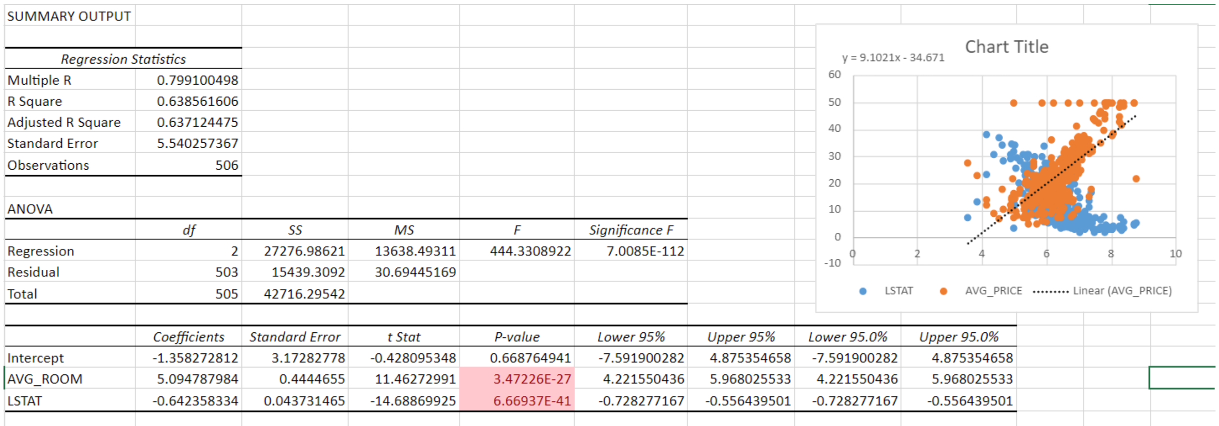
Ans] From this model 54% of the variation in the average price is explained by the LSTAT. The coefficient of LSTAT for the model is -0. 950049354.This says that if LSTAT increases by 0.9 times then average price of house decreases 0.9times. The intercept of LSTAT for the model is 34.55384088.

B] Is LSTAT variable significant for the analysis based on your model?

Ans] Yes, LSTAT is significant variable for the avg_price from this model. As the p-value(5.08E-88) we obtained from this model is away less than 0.05. By this we can say that LSTAT is a significant variable according to this model.

Question 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?

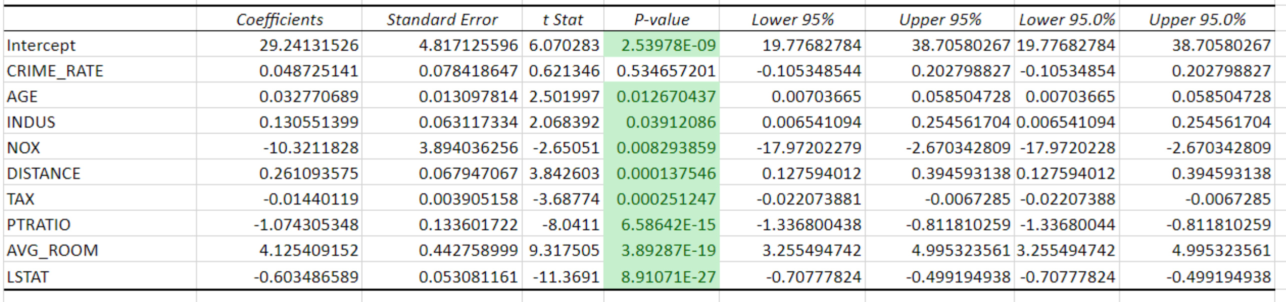
Ans] Regression Equation we obtained for this model is:
y = -1.358 +5.09 X0 -0.642 X1
Where y=Avg_price
X0 = avg_room
X1 = LSTAT
As per the model, avg_price for new house can be calculated as
Y= -1.358 + 5.09(7) -0.642(20) = 21.44
So, the price for the new house is $21440. We can say that company is 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. Explain.

Ans] Yes, the performance of this model performs well compared to previous model.
From this model the linear equation we obtained is
y= -1.35 +5.09a -0.64b (Where a=Avg_room b=LSTAT)
And Value of R square = 0.638561606.
With this we can say that 63% of variability for average price is explained by Avg_room and LSTAT combinedly and we obtained multiple R value as 0.79 which says it is highly correlated.
But in the previous model LSTAT alone describes 54% of variability for average price.

Question 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.



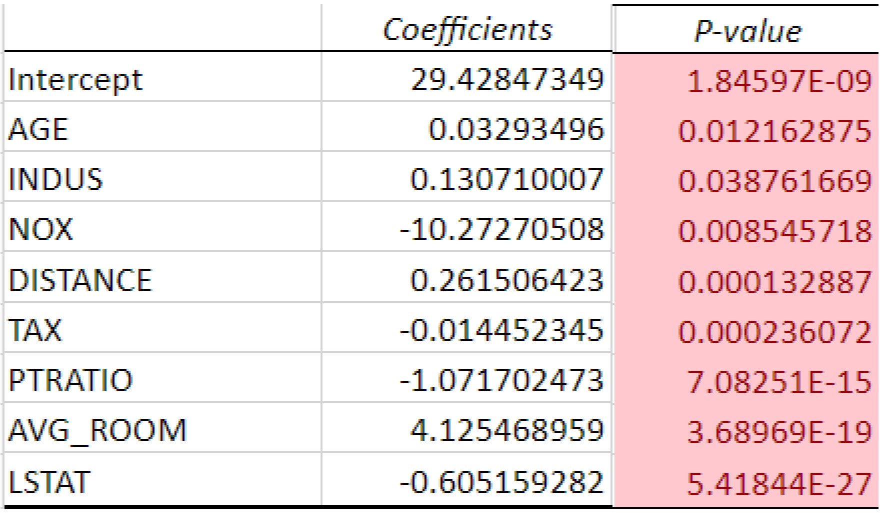
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	29.24131526	4.817125596	6.070283	2.53978E-09	19.77682784	38.70580267	19.77682784	38.70580267
CRIME_RATE	0.048725141	0.078418647	0.621346	0.534657201	-0.105348544	0.202798827	-0.10534854	0.202798827
AGE	0.032770689	0.013097814	2.501997	0.012670437	0.00703665	0.058504728	0.00703665	0.058504728
INDUS	0.130551399	0.063117334	2.068392	0.03912086	0.006541094	0.254561704	0.006541094	0.254561704
NOX	-10.3211828	3.894036256	-2.65051	0.008293859	-17.97202279	-2.670342809	-17.9720228	-2.670342809
DISTANCE	0.261093575	0.067947067	3.842603	0.000137546	0.127594012	0.394593138	0.127594012	0.394593138
TAX	-0.01440119	0.003905158	-3.68774	0.000251247	-0.022073881	-0.0067285	-0.02207388	-0.0067285
PTRATIO	-1.074305348	0.133601722	-8.0411	6.58642E-15	-1.336800438	-0.811810259	-1.33680044	-0.811810259
AVG_ROOM	4.125409152	0.442758999	9.317505	3.89287E-19	3.255494742	4.995323561	3.255494742	4.995323561
LSTAT	-0.603486589	0.053081161	-11.3691	8.91071E-27	-0.70777824	-0.499194938	-0.70777824	-0.499194938

From this we can say that crime rate is not a significant variable for average price of an house as p-value is greater than 0.5.
All the features combinely explains 69% of variability for average price of a house.
NOX, TAX, PTRATIO and LSTAT have negative coefficients which says that increase in these features will result decrease in price of the house and viceversa.

Question 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.

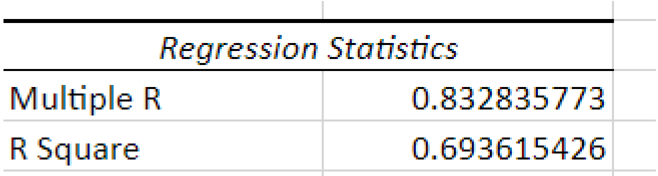
A] Interpret the output of this model.



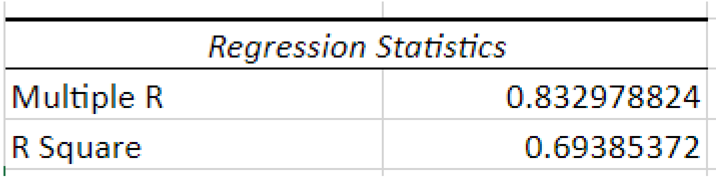
	Coefficients	*P-value*
Intercept	29.42847349	1.84597E-09
AGE	0.03293496	0.012162875
INDUS	0.130710007	0.038761669
NOX	-10.27270508	0.008545718
DISTANCE	0.261506423	0.000132887
TAX	-0.014452345	0.000236072
PTRATIO	-1.071702473	7.08251E-15
AVG_ROOM	4.125468959	3.68969E-19
LSTAT	-0.605159282	5.41844E-27

Ans] From this we can conclude that all the features are significant variables for average price of the house.

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?



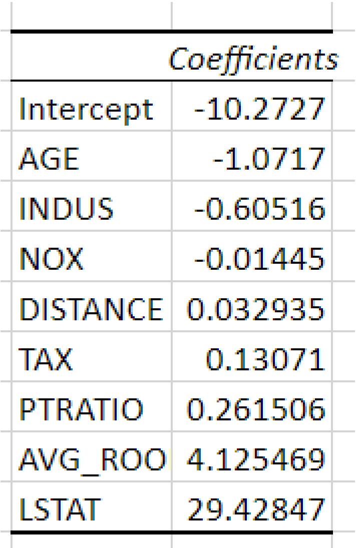
Regression Statistics	
Multiple R	0.832835773
R Square	0.693615426



Regression Statistics	
Multiple R	0.832978824
R Square	0.69385372

Ans] By comparing Multiple R and R square values for both the models we can conclude that both models perform well.

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
Intercept	-10.2727
AGE	-1.0717
INDUS	-0.60516
NOX	-0.01445
DISTANCE	0.032935
TAX	0.13071
PTRATIO	0.261506
AVG_ROO	4.125469
LSTAT	29.42847

Ans] If NOX is more in the locality, according to this model average price of the house will decrease by 10 times.

D] Write the regression equation from this model.

Ans] Y=0.03293496 X0 + 0.130710007 X1 -10.27270508 X3 +0.261506423 X4 - 0.014452345 X5 -1.071702473 X6 + 4.125468959 X7 -0.605159282 X8 +29.42847349

Where Y = average_Price

X0 = Age

X1 = Indus

X2 = NOX

X3 = Distance

X4 = TAX

X5 = PTRATIO

X6 = Avg_room

X7 = LSTAT

Summary: From this Analysis, we can conclude that all the features play a vital role in estimating the average price of the house excluding crime rate. And a few features have

negative coefficients which say that increase rate in those features will decrease the average price of the house like NOX, PTRATIO, TAX and LSTAT.