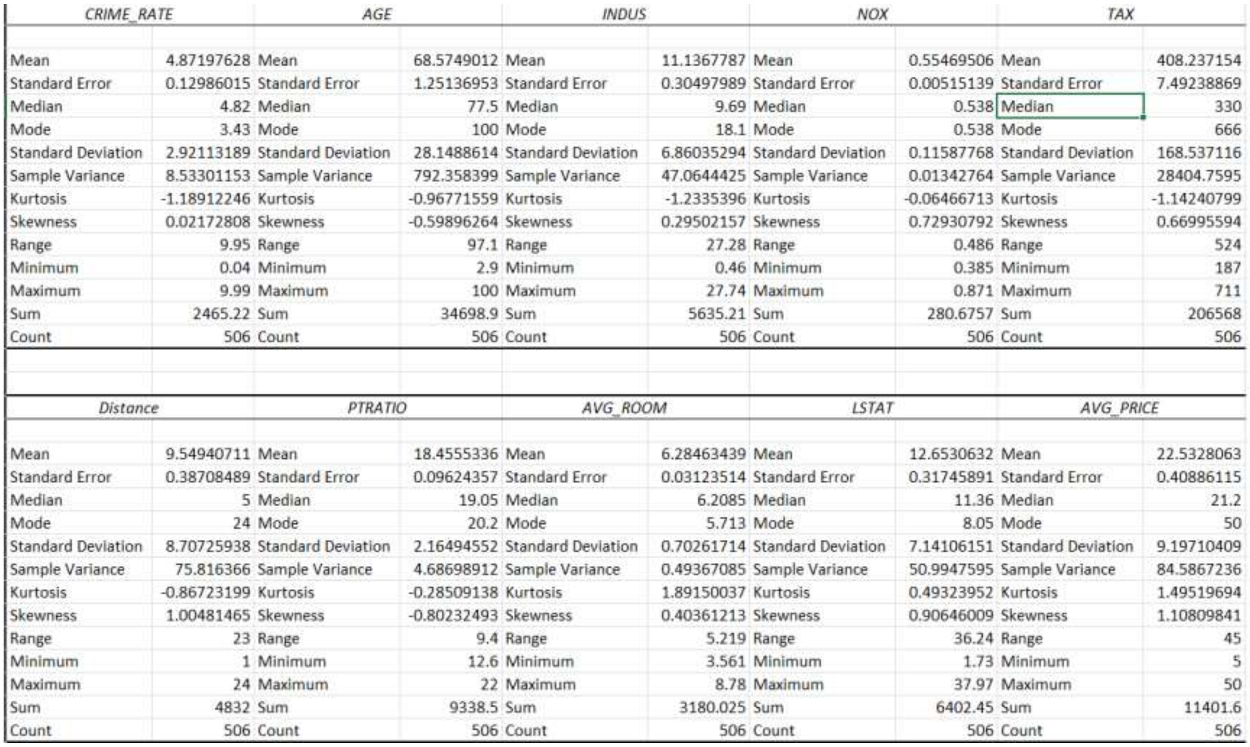
Terro’s Real Estate Agency

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?

Solution:



CRIME_RATE		AGE		INDUS		NOX		TAX	
Mean	4.87197628	Mean	68.5749012	Mean	11.1367787	Mean	0.55469506	Mean	408.237154
Standard Error	0.12986015	Standard Error	1.25136953	Standard Error	0.30497989	Standard Error	0.00515139	Standard Error	7.49238869
Median	4.82	Median	77.5	Median	9.69	Median	0.538	Median	330
Mode	3.43	Mode	100	Mode	18.1	Mode	0.538	Mode	666
Standard Deviation	2.92113189	Standard Deviation	28.1488614	Standard Deviation	6.86035294	Standard Deviation	0.11587768	Standard Deviation	168.537116
Sample Variance	8.53301153	Sample Variance	792.358399	Sample Variance	47.0644425	Sample Variance	0.01342764	Sample Variance	28404.7595
Kurtosis	-1.18912246	Kurtosis	-0.96771559	Kurtosis	-1.2335396	Kurtosis	-0.06466713	Kurtosis	-1.14240799
Skewness	0.02172808	Skewness	-0.59896264	Skewness	0.29502157	Skewness	0.72930792	Skewness	0.66995594
Range	9.95	Range	97.1	Range	27.28	Range	0.486	Range	524
Minimum	0.04	Minimum	2.9	Minimum	0.46	Minimum	0.385	Minimum	187
Maximum	9.99	Maximum	100	Maximum	27.74	Maximum	0.871	Maximum	711
Sum	2465.22	Sum	34698.9	Sum	5635.21	Sum	280.6757	Sum	206568
Count	506	Count	506	Count	506	Count	506	Count	506

Distance		PTRATIO		AVG_ROOM		LSTAT		AVG_PRICE	
Mean	9.54940711	Mean	18.4555336	Mean	6.28463439	Mean	12.6530632	Mean	22.5328063
Standard Error	0.38708489	Standard Error	0.09624357	Standard Error	0.03123514	Standard Error	0.31745891	Standard Error	0.40886115
Median	5	Median	19.05	Median	6.2085	Median	11.36	Median	21.2
Mode	24	Mode	20.2	Mode	5.713	Mode	8.05	Mode	50
Standard Deviation	8.70725938	Standard Deviation	2.16494552	Standard Deviation	0.70261714	Standard Deviation	7.14106151	Standard Deviation	9.19710409
Sample Variance	75.816366	Sample Variance	4.68698912	Sample Variance	0.49367085	Sample Variance	50.9947595	Sample Variance	84.5867236
Kurtosis	-0.86723199	Kurtosis	-0.28509138	Kurtosis	1.89150037	Kurtosis	0.49323952	Kurtosis	1.49519694
Skewness	1.00481465	Skewness	-0.80232493	Skewness	0.40361213	Skewness	0.90646009	Skewness	1.10809841
Range	23	Range	9.4	Range	5.219	Range	36.24	Range	45
Minimum	1	Minimum	12.6	Minimum	3.561	Minimum	1.73	Minimum	5
Maximum	24	Maximum	22	Maximum	8.78	Maximum	37.97	Maximum	50
Sum	4832	Sum	9338.5	Sum	3180.025	Sum	6402.45	Sum	11401.6
Count	506	Count	506	Count	506	Count	506	Count	506

OBSERVATIONS:

Crime Rate:

- The average crime rate in the town is around 4.87.
- The skewness for this it's around 0 so that we can say that this distribution is normal distribution.
- The data deviates 2.9 standard distribution from the mean.

Age:

- On average the houses built in the town it's around 68 years.
- The negative kurtosis value gives us your flatter distribution for this data.
- Most of The houses are around 100 years.

Indus:

- On average 11.13 % belongs to non-retail businesses.
- Negative kurtosis shows that that the values are spread across the mean value.Nox:
- On average nitric oxide concentration is around 0.55ppm.
- It gives us a positive skew , shows us that most number of houses have below 0.55 ppm.

Distance:

- On an average, distance from highway is around 9.5 miles.
- Most houses are 24 miles from the highway.
- It gives us a positive skew which shows that most of the houses is below 9.5 miles.

Tax:

- The average tax on the house is 408 dollars.
- Most of the houses have $666 as the tax .

Ptratio:

- On an average , pupil teacher ratio is 18.45.
- Most houses hav3 20.2 as ptraio.

Avg_room:

- There are 6.3 rooms on average.
- Positive kurtosis shows that most values is concentrated in the median . mean and median have
- close values.
- Positive skew gives most houses have less than 6.3 rooms.

Lstat:

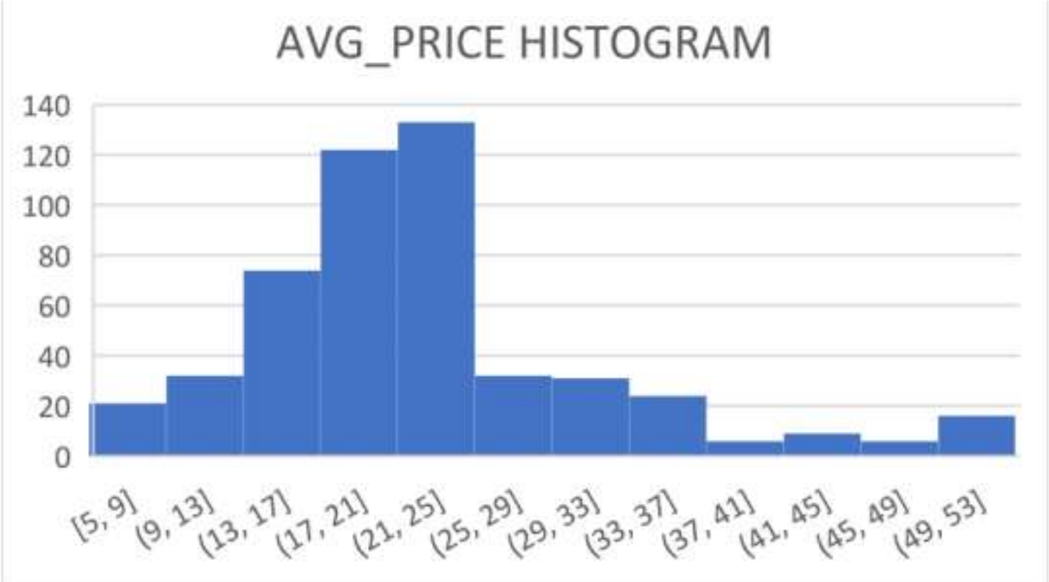
- On an average , 12% of population has lower status.
- Positive skewness tells us that more number of houses have less than 12% lower status
- population.

Avg_price:

- Average value of price of house is around $22500.
- Maximum houses have price aorund $50000.

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

SOLUTION:

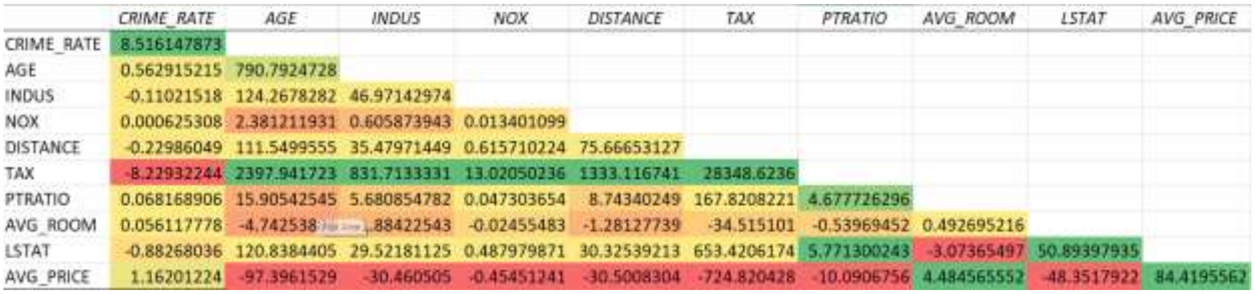


INFERENCE:

- The graph shows us that it is right skew (positive Skew) where the shape of the distribution tails right side of the graph.
- This shows that most values lie in the left side, most of the values lie between (21,25) and (17,21) range which lies below median value 21.2.

3. Compute the covariance matrix. Share your observations.

SOLUTION:



	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RATE	8.516147873									
AGE	0.562915215	790.7924728								
INDUS	-0.11021518	124.2678282	46.97142974							
NOX	0.000625308	2.381211931	0.605873943	0.013401099						
DISTANCE	-0.22986049	111.5499555	35.47971449	0.615710224	75.66653127					
TAX	-8.22932244	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	-4.742538[illegible]	[illegible].88422543	-0.02455483	-1.28127739	-34.515101	-0.53969452	0.492695216		
LSTAT	-0.88268036	120.8384405	29.52181125	0.487979871	30.32539213	653.4206174	5.771300243	-3.07365497	50.89397935	
AVG_PRICE	1.16201224	-97.3961529	-30.460505	-0.45451241	-30.5008304	-724.820428	-10.0906756	4.484565552	-48.3517922	84.4195562

Covariance gives a measure of relationship between two variables .
Green and Red shows positive and negative values respectively

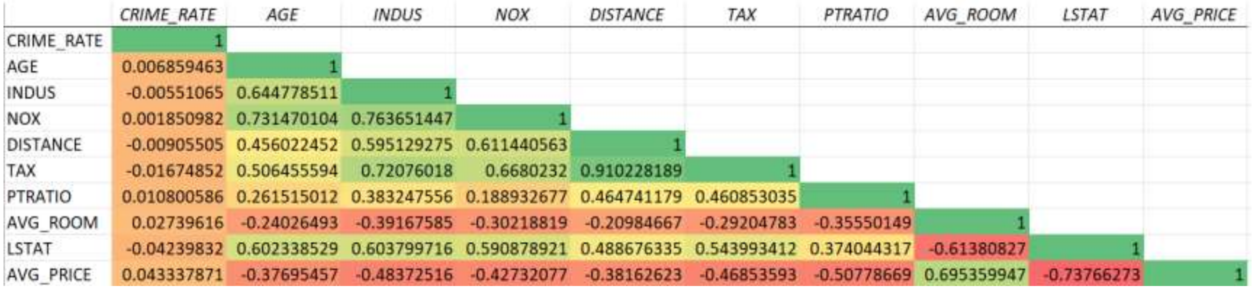
INFERENCE:

- Crime rate shows us negative correlation with the tax, which says that crime rate is less when tax rate is high

- Crime rate shows almost zero correlation with other variables, hence insignificant relationship between them.
- Property Tax is higher for those houses built prior to 1940(positive Relation).
- Industry, nox, distance also has positive relationship with tax
- Age, Industry, nox, distance, tax, ptratio, lstat has negative relationship with average price of the house

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.

SOLUTION:



	CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
CRIME_RATE	1									
AGE	0.006859463	1								
INDUS	-0.00551065	0.644778511	1							
NOX	0.001850982	0.731470104	0.763651447	1						
DISTANCE	-0.00905505	0.456022452	0.595129275	0.611440563	1					
TAX	-0.01674852	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	-0.24026493	-0.39167585	-0.30218819	-0.20984667	-0.29204783	-0.35550149	1		
LSTAT	-0.04239832	0.602338529	0.603799716	0.590878921	0.488676335	0.543993412	0.374044317	-0.61380827	1	
AVG_PRICE	0.043337871	-0.37695457	-0.48372516	-0.42732077	-0.38162623	-0.46853593	-0.50778669	0.695359947	-0.73766273	1

Correlation gives how strongly they are related to each other . They lie between -1 and +1
Top 3 positively correlated pairs

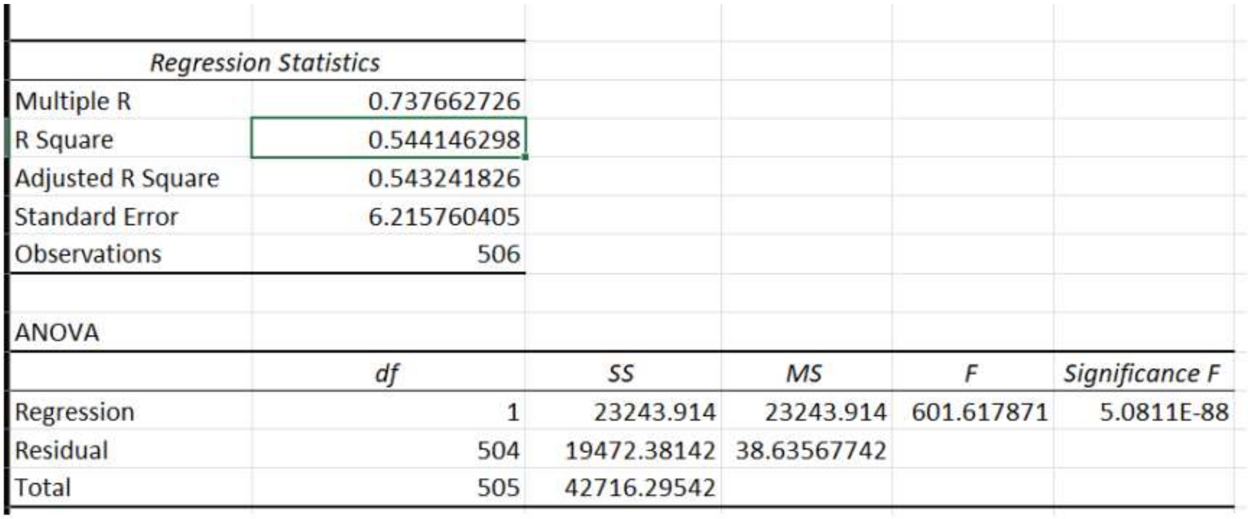
- Tax and Distance
- Nox and Industry
- Indus and Age

Top 3 negatively correlated pairs

- Avg_price and Lstat
- Lstat and Avg_room
- Avg_price and Pt ratio

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.

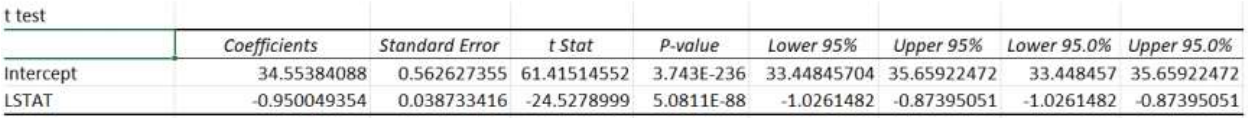
Solution:



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	601.617871	5.0811E-88
Residual	504	19472.38142	38.63567742		
Total	505	42716.29542			



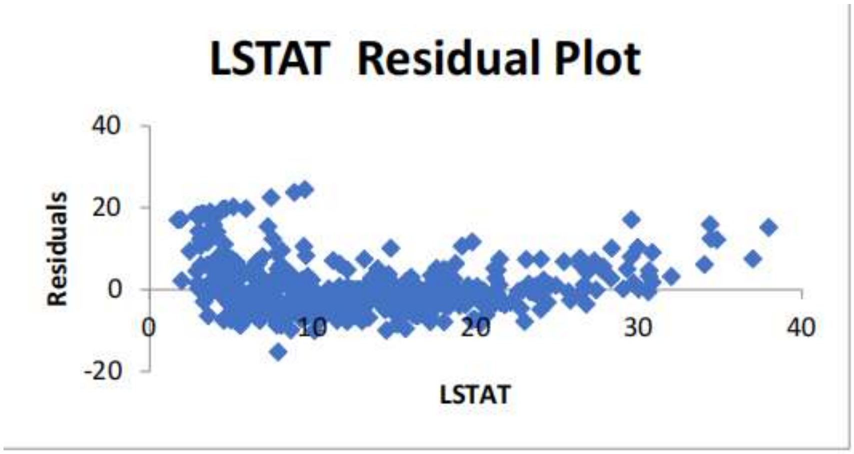
t test

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	34.55384088	0.562627355	61.41514552	3.743E-236	33.44845704	35.65922472	33.448457	35.65922472
LSTAT	-0.950049354	0.038733416	-24.5278999	5.0811E-88	-1.0261482	-0.87395051	-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?

SOULUTION:

- From the regression stats, R square value shows that 54% of variation in Avg_price is explained by Lstat.
- From the p-value which is <0.05 , null hypothesis is rejected and alternate hypothesis is accepted i.e avg_price (dependent variable) is dependent in the Lstat(independent Variable) in calculating linear regression.



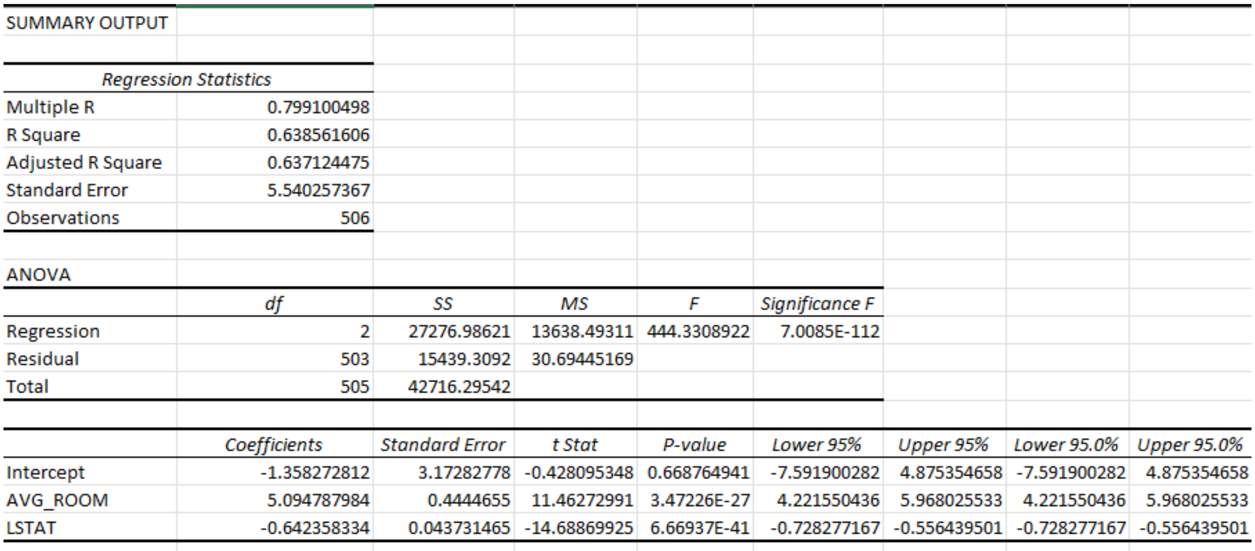
b. Is LSTAT variable significant for the analysis based on your model?

Solution:

From the regression summary ,
p-value is < 0.005 and r square is 54% which shows us that Lstat is significant variable for avg_price

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.

Solution:



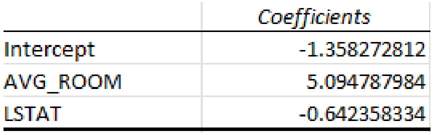
SUMMARY OUTPUT

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



	Coefficients
Intercept	-1.358272812
AVG_ROOM	5.094787984
LSTAT	-0.642358334

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?

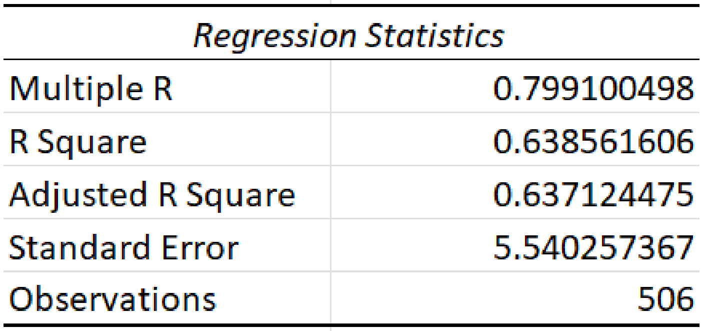
Solution:

Equation :
Avg_price = -1.35 + 5.095(avg_room) – 0.64(lstat)
For,
Avg_room=7
Lstat=20
Avg_price=21.515*1000
=$ 21515 which is < $ 30000

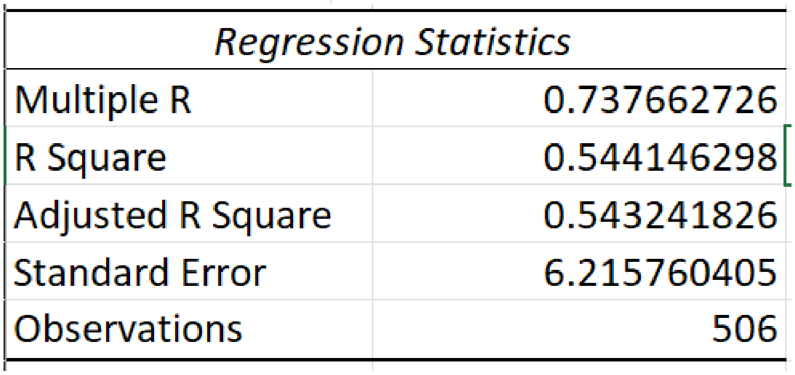
Therefore, the company is overcharging.

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.

Solution:



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



Regression Statistics	
Multiple R	0.737662726
R Square	0.544146298
Adjusted R Square	0.543241826
Standard Error	6.215760405
Observations	506

Yes, the performance of this model better than the previous model I built in Question 5 because

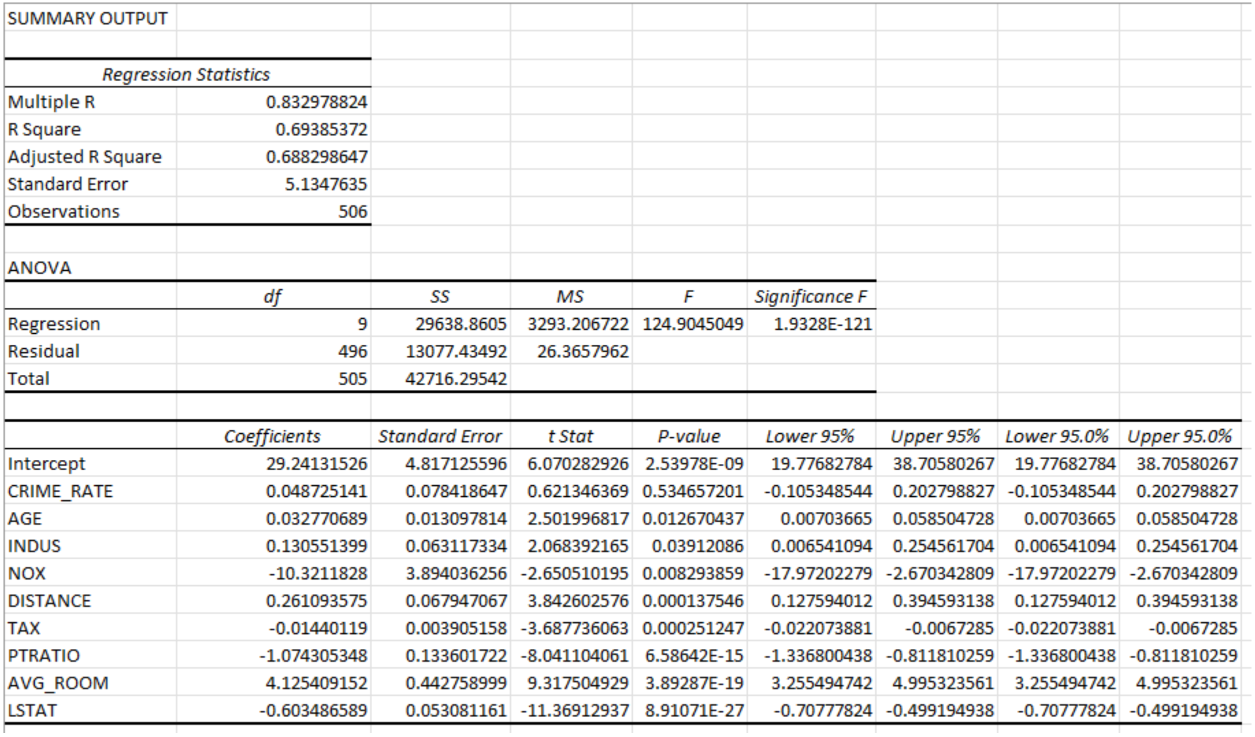
R square value in previous ques is 0.54 and

R square value in this ques is 0.64.

So it says that 64% of variation in avg_price is due to Avg_room and Lstst.

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.

Solution:



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	124.9045049	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	2.53978E-09	19.77682784	38.70580267	19.77682784	38.70580267
CRIME_RATE	0.048725141	0.078418647	0.621346369	0.534657201	-0.105348544	0.202798827	-0.105348544	0.202798827
AGE	0.032770689	0.013097814	2.501996817	0.012670437	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	-2.650510195	0.008293859	-17.97202279	-2.670342809	-17.97202279	-2.670342809
DISTANCE	0.261093575	0.067947067	3.842602576	0.000137546	0.127594012	0.394593138	0.127594012	0.394593138
TAX	-0.01440119	0.003905158	-3.687736063	0.000251247	-0.022073881	-0.0067285	-0.022073881	-0.0067285
PTRATIO	-1.074305348	0.133601722	-8.041104061	6.58642E-15	-1.336800438	-0.811810259	-1.336800438	-0.811810259
AVG_ROOM	4.125409152	0.442758999	9.317504929	3.89287E-19	3.255494742	4.995323561	3.255494742	4.995323561
LSTAT	-0.603486589	0.053081161	-11.36912937	8.91071E-27	-0.70777824	-0.499194938	-0.70777824	-0.499194938

From the table,

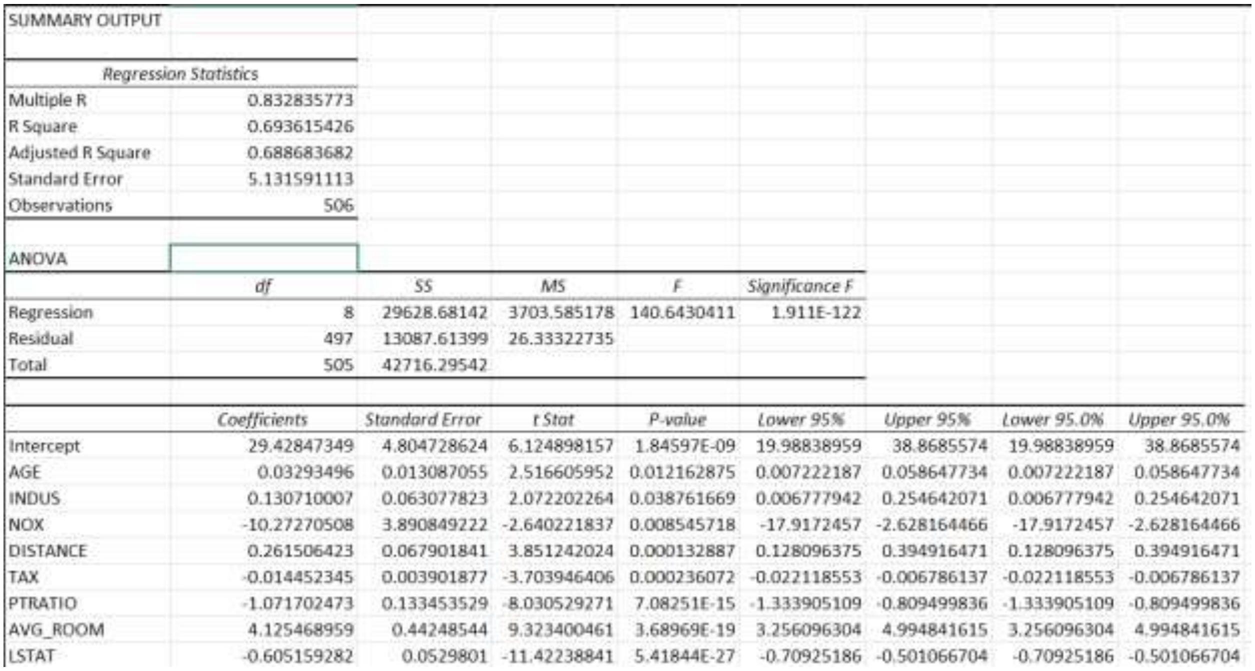
- P-value of Crime rate is > significance level (0.05) which shows that it does not reject null hypothesis .Thus coefficients could be 0 which could be seen in lower95% and upper 95& column.
- Crime rate could be independent variable so that adjusted r square is less than the r square value as shown in table (independent variables will be deleted in adjusted R square)
- The coefffient of avg_room is higher when compared to other variables, so it has high significance on avg_price
- The intercept value is 29.24 if all coefficients are 0
- The p-value of Indus and nox is < 0.05 but little closer to it.so, it has less significance on the avg_price.
- Nox,tax,ptratio,lstat has negative coefficiants.so it has negative relation to avg_price.

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.

<u>Solution:</u>

Here crime rate has p-value greater than 0.05 so , we will neglect this variable



SUMMARY OUTPUT

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

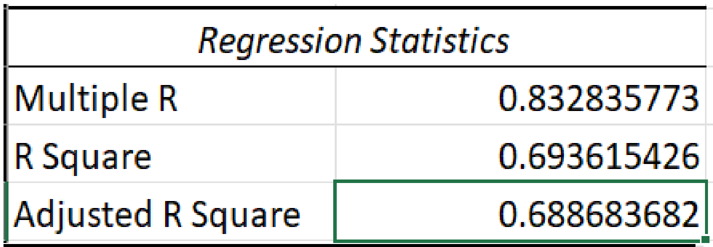
ANOVA

	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			

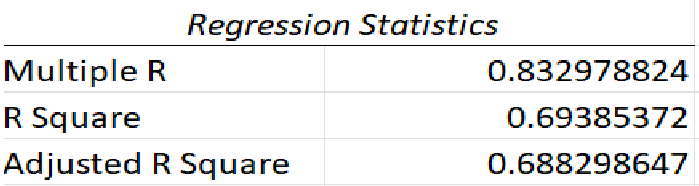
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	29.42847349	4.804728624	6.124898157	1.84597E-09	19.98838959	38.8685574	19.98838959	38.8685574
AGE	0.03293496	0.013087055	2.516605952	0.012162875	0.007222187	0.058647734	0.007222187	0.058647734
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

b.While comparing the two models, one in this question and the other in previous one, we can see there is slight increment in adjusted R-square value for the model where insignificant variable is removed to make a new regression model

Old



Regression Statistics	
Multiple R	0.832835773
R Square	0.693615426
Adjusted R Square	0.688683682

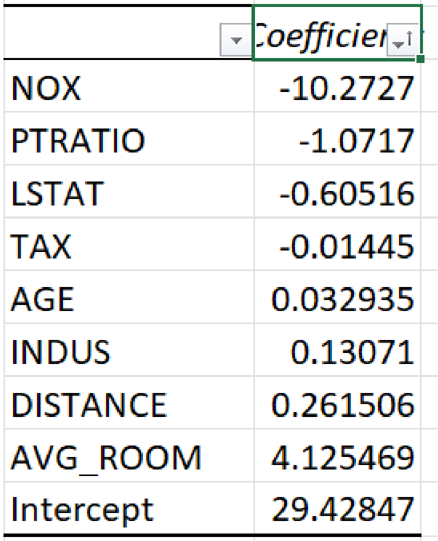


Regression Statistics	
Multiple R	0.832978824
R Square	0.69385372
Adjusted R Square	0.688298647

<u>(new)</u>

This shows there is very little difference in r square values.

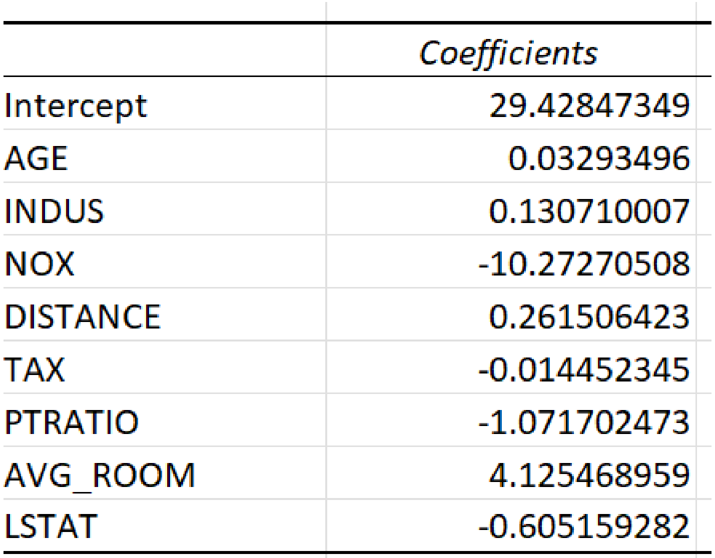
c. Sort the values of the Coefficients in ascending order. What will happen to the average price if the value of NOX is more in a locality in this town?



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

NOX and avg_price are in negative relation to eachother . so when nox increases the price of house
decreases.

d. Write the regression equation from this model



	Coefficients
Intercept	29.42847349
AGE	0.03293496
INDUS	0.130710007
NOX	-10.27270508
DISTANCE	0.261506423
TAX	-0.014452345
PTRATIO	-1.071702473
AVG_ROOM	4.125468959
LSTAT	-0.605159282

Avg_price= 29.42 + 0.033*(AGE) + 0.13*(INDUS) - 10.27*(NOX) + 0.26*(DISTANCE) - 0.014*(TAX) - 1.07*(PTRATIO) + 4.13*(AVG_ROOM) - 0.60*(LSTAT)