Terro's Real Estate Agency

MARCH 31

Great Learning
Authored by: Nutan Parab

Great Learning

Case Study: Terro's Real Estate Agency

(TOPICS COVERED: Descriptive Statistics, Covariance, Correlations, Simple Linear Regression, Multiple Linear Regression)

You have been hired at a Terro's Real Estate Agency in the capacity of an Auditor. One of the jobs that the auditors of this agency do is to map all the relevant features for 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).

Data Dictionary:

CRIME RATE: *per capita crime rate by town*

INDUSTRY: *the proportion of non-retail business acres per town (in percentage terms)*

NOX: *nitric oxides concentration (parts per 10 million)*

AVG_ROOM: *average number of rooms per house*

AGE: *the proportion of houses built prior to 1940 (in percentage terms)*

DISTANCE: *distance from highway (in miles)*

TAX: *full-value property-tax rate per $10,000*

PTRATIO: *pupil-teacher ratio by town*

LSTAT: *% lower status of the population*

AVG PRICE: *Average value of houses in $1000's*

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	
Mean	4.87198	Mean	68.57490119	Mean	11.13677866	Mean	0.554695059
Standard Error	0.12986	Standard Error	1.251369525	Standard Error	0.304979888	Standard Error	0.005151391
Median	4.82	Median	77.5	Median	9.69	Median	0.538
Mode	3.43	Mode	100	Mode	18.1	Mode	0.538
Standard Deviation	2.92113	Standard Deviation	28.14886141	Standard Deviation	6.860352941	Standard Deviation	0.115877676
Sample Variance	8.53301	Sample Variance	792.3583985	Sample Variance	47.06444247	Sample Variance	0.013427636
Kurtosis	-1.18912	Kurtosis	-0.967715594	Kurtosis	-1.233539601	Kurtosis	-0.064667133
Skewness	0.02173	Skewness	-0.59896264	Skewness	0.295021568	Skewness	0.729307923
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

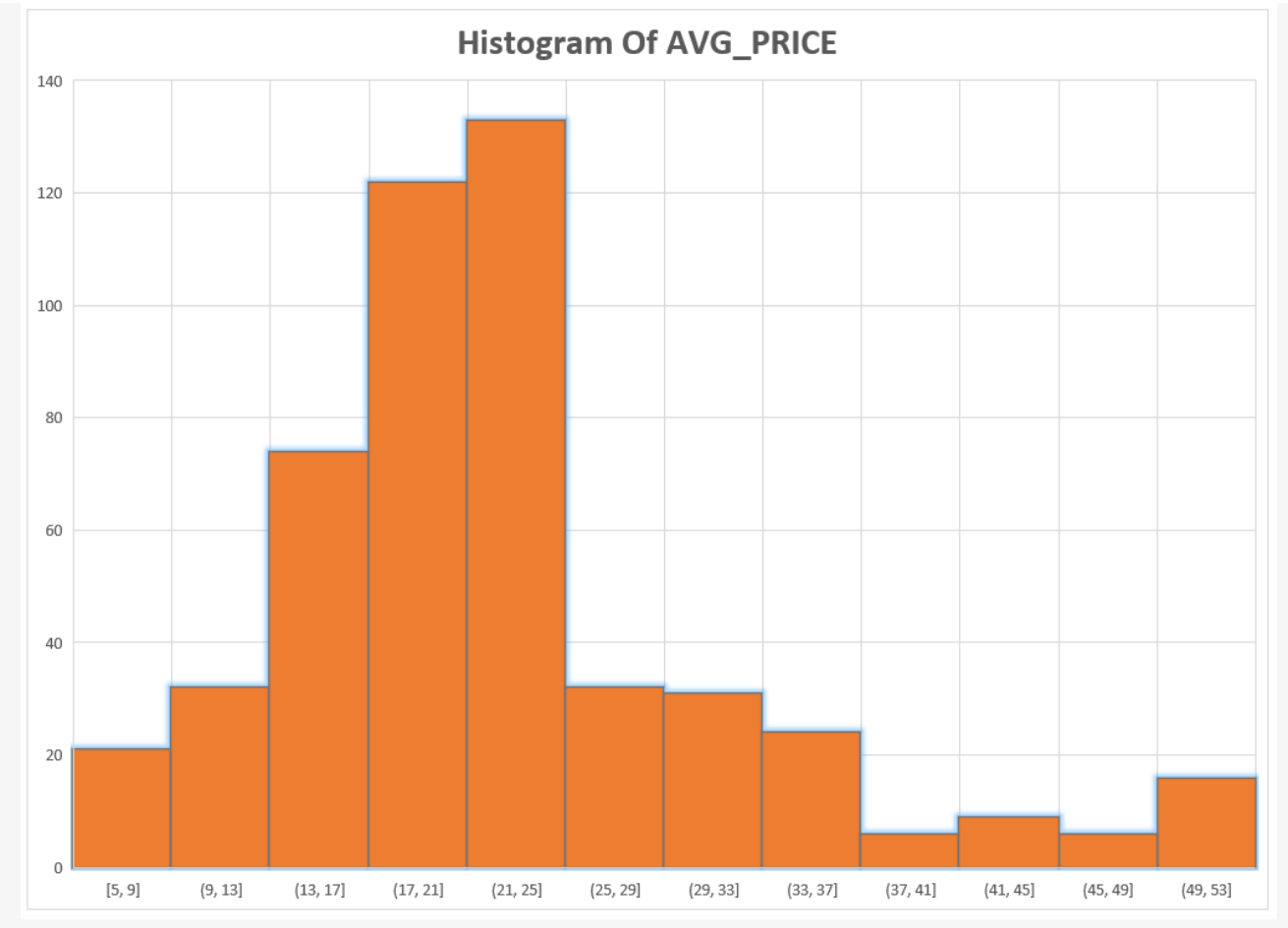
TAX		PTRATIO		AVG_ROOM		LSTAT	
Mean	408.237	Mean	18.4555336	Mean	6.284634387	Mean	12.65306324
Standard Error	7.49239	Standard Error	0.096243568	Standard Error	0.031235142	Standard Error	0.317458906
Median	330	Median	19.05	Median	6.2085	Median	11.36
Mode	666	Mode	20.2	Mode	5.713	Mode	8.05
Standard Deviation	168.537	Standard Deviation	2.164945524	Standard Deviation	0.702617143	Standard Deviation	7.141061511
Sample Variance	28404.8	Sample Variance	4.686989121	Sample Variance	0.49367085	Sample Variance	50.99475951
Kurtosis	-1.14241	Kurtosis	-0.285091383	Kurtosis	1.891500366	Kurtosis	0.493239517
Skewness	0.66996	Skewness	-0.802324927	Skewness	0.403612133	Skewness	0.906460094
Range	524	Range	9.4	Range	5.219	Range	36.24
Minimum	187	Minimum	12.6	Minimum	3.561	Minimum	1.73
Maximum	711	Maximum	22	Maximum	8.78	Maximum	37.97
Sum	206568	Sum	9338.5	Sum	3180.025	Sum	6402.45
Count	506	Count	506	Count	506	Count	506

DISTANCE		AVG_PRICE	
Mean	9.549407115	Mean	22.53280632
Standard Error	0.387084894	Standard Error	0.408861147
Median	5	Median	21.2
Mode	24	Mode	50
Standard Deviation	8.707259384	Standard Deviation	9.197104087
Sample Variance	75.81636598	Sample Variance	84.58672359
Kurtosis	-0.86723199	Kurtosis	1.495196944
Skewness	1.004814648	Skewness	1.108098408
Range	23	Range	45
Minimum	1	Minimum	5
Maximum	24	Maximum	50
Sum	4832	Sum	11401.6
Count	506	Count	506

OBSERVATIONS:

- The total records recorded in dataset are 506.
- Most of the variables are positively skewed.
- For the "CRIME_RATE" variable, the mean crime rate is 4.87, with a standard deviation of 2.92. The range of crime rates is from 0.04 to 9.95.
- The "AGE" variable represents the average age of the population. The mean age is 68.57, with a standard deviation of 28.15. The range of ages is from 2 to 100.
- The "INDUS" variable represents the proportion of non-retail business acres per town. The mean proportion is 11.14, with a standard deviation of 6.86. The range of proportions is from 0.04 to 27.28
- The "NOX" variable represents the nitric oxides concentration. The mean concentration is 0.55, with a standard deviation of 0.12. The range of concentrations is from 0.54 to 1. 5.
- The "DISTANCE" variable represents the distance to five Boston employment centers from highway. The mean distance is 9.55, with a standard deviation of 8.71. The range of distances is from 5 to 24.
- Tax has the highest variance among all as it depends on various other factors.

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



OBSERVATONS:

- The horizontal axis line indicates the price range and vertical axis line indicates number of houses.
- Less number of houses ranges between the price range of $31k to $41k and $45k to $4k.
- Most number of houses ranges between the price range of $21k to $25k.
- This shows positive 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	46.97142974							
NOX	0.000625308	2.381211931	0.605873943	0.013401099						
DISTANCE	-0.229860488	111.5499555	35.47971449	0.615710224	75.66653127					
TAX	-8.229322439	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	-4.74253803	-1.884225427	-0.024554826	-1.281277391	-34.51510104	-0.539694518	0.492695216		
LSTAT	-0.882680362	120.8384405	29.52181125	0.487979871	30.32539213	653.4206174	5.771300243	-3.073654967	50.89397935	
AVG_PRICE	1.16201224	-97.39615288	-30.46050499	-0.454512407	-30.50083035	-724.8204284	-10.09067561	4.484565552	-48.35179219	84.4195562

OBSERVATIONS:

- The covariance between CRIME_RATE and AGE is 0.5629, indicating a positive relationship between crime rate and the age of the area's population.
- The covariance between INDUS and NOX is 0.6059, suggesting a positive relationship between the proportion of non-retail business acres per town and the concentration of nitrogen oxides.
- The covariance between DISTANCE and TAX is -8.2293, indicating a negative relationship between the weighted distances to five Boston employment centers and the full-value property tax rate.
- The covariance between PTRATIO and AVG_ROOM is -0.5397, suggesting a negative relationship between the pupil-teacher ratio and the average number of rooms per dwelling.

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	0.644778511	1							
NOX	0.001850982	0.731470104	0.763651447	1						
DISTANCE	-0.009055049	0.456022452	0.595129275	0.611440563	1					
TAX	-0.016748522	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	-0.240264931	-0.391675853	-0.302188188	-0.209846668	-0.292047833	-0.355501495	1		
LSTAT	-0.042398321	0.602338529	0.603799716	0.590878921	0.488676335	0.543993412	0.374044317	-0.613808272	1	
AVG_PRICE	0.043337871	-0.376954565	-0.48372516	-0.427320772	-0.381626231	-0.468535934	-0.507786686	0.695359947	-0.737662726	1
		Top 3 Positive correlation	1) Tax,Distance	2) NOX,Indus	3) NOX,Age					
		Top 3 Negative correlation	1) Avg_price,LSTAT	2) LSTAT,Avg_room	3) Avg_price,PTRATIO					

- Top 3 positively correlated pairs:
  1. Tax – Distance
  2. NOX – Indus
  3. NOX – Age

- Top 3 negatively correlated pairs:
  1. Avg_price – LSTAT
  2. LSTAT – Avg_room
  3. Avg_price – PTRATIO

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.

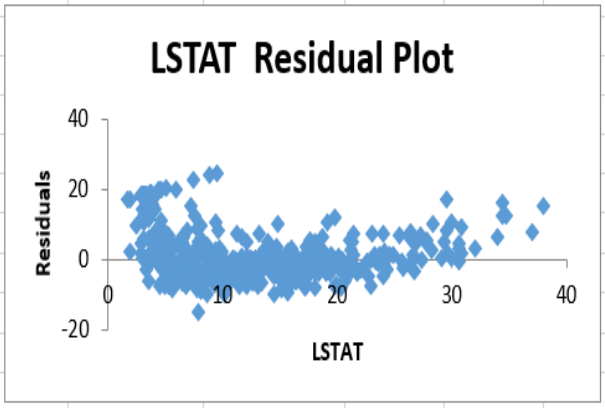
SUMMARY OUTPUT

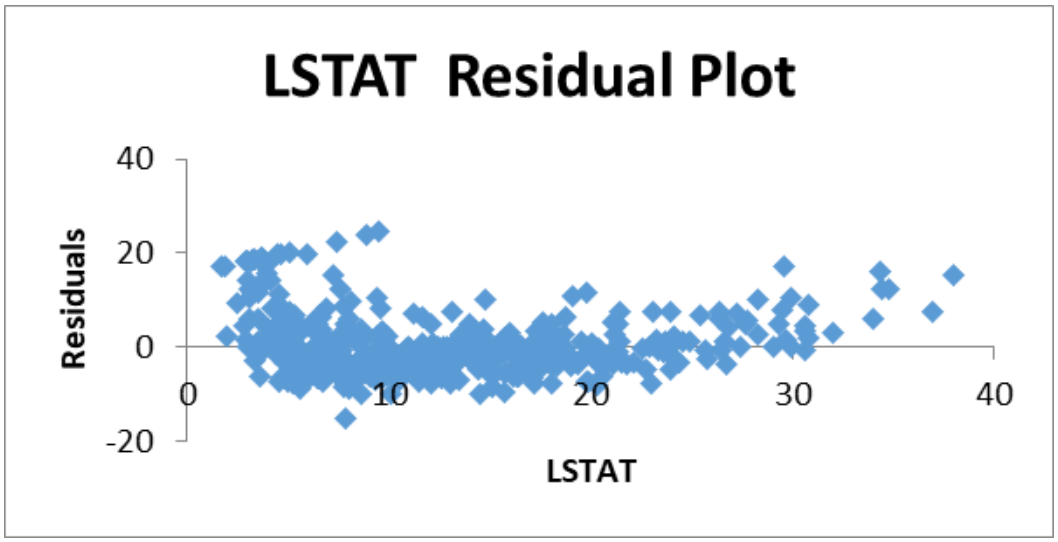
Regression Statistics	
Multiple R	0.737662726
R Square	0.544146298
Adjusted R	0.543241826
Standard E	6.215760405
Observatic	506

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regressior	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	-0.873950508





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

- Variance explained: The R-squared value of 0.5441 indicates that approximately 54.41% of the total variation in the dependent variable is explained by the independent variable(s) in the model.
- Coefficient values: The coefficient for the LSTAT variable is -0.9500. This means that for every one unit increase in the LSTAT variable (percentage of lower status of the population), the predicted average price decreases by approximately $9500, assuming all other variables are held constant.
- Intercept: The intercept value is 34.5538, which represents the estimated average price when all independent variables are zero. In this case, it would mean the estimated average price when the LSTAT variable is zero.
- Residuals: The residual values represent the differences between the observed average prices and the predicted average prices from the regression model. They indicate the degree of error in the model's predictions. For example, the first observation has a predicted average price of 29.8226, but the actual value is 24, resulting in a residual of -5.8226.

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

- Yes, the LSTAT variable is significant for the analysis based on the model. It has a coefficient value of -0.9500, meaning it has a strong impact on the predicted average 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.

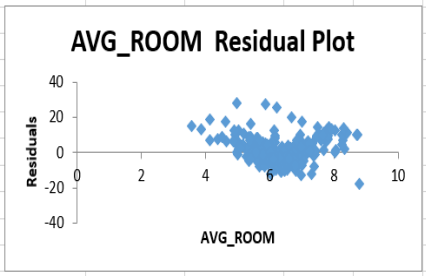
SUMMARY OUTPUT

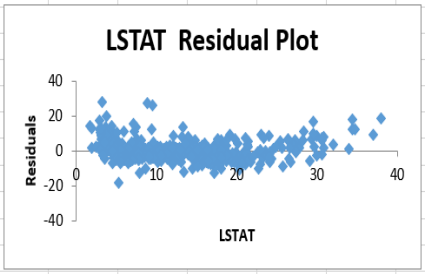
Regression Statistics	
Multiple R	0.799100498
R Square	0.638561606
Adjusted R Square	0.637124475
Standard Error	5.540257367
Observations	506

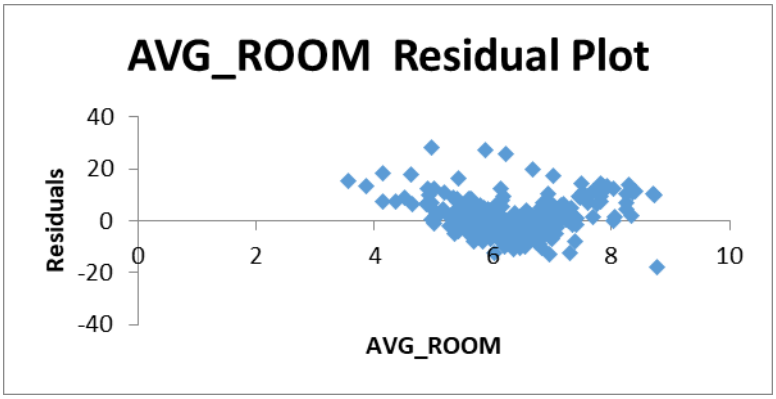
ANOVA

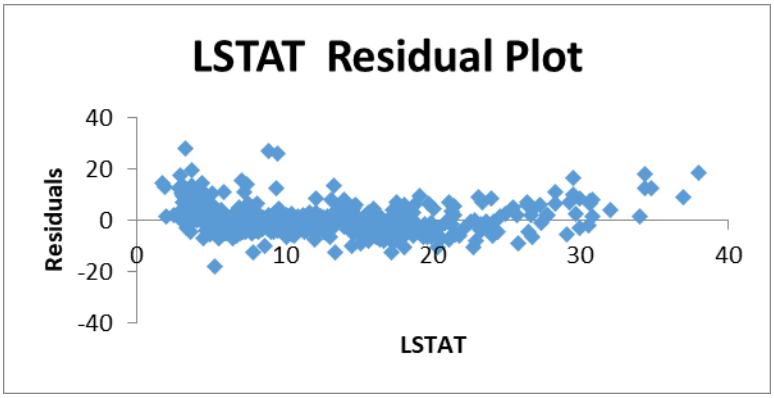
	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	27276.98621	13638.49311	444.3308922	7.0085E-112
Residual	503	15439.3092	30.69445169		
Total	505	42716.29542			

	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 is:
- AVG_PRICE = Intercept + (AVG_ROOM * Number of rooms) + (LSTAT * L-STAT value)
- Substituting the values:
- AVG_PRICE = -1.358272812 + (5.094787984 * 7) + (-0.642358334 * 20)
- Calculating this, the predicted AVG_PRICE for a house with 7 rooms and an L-STAT value of 20 is **21.458**
- If the predicted AVG_PRICE is higher than 30000 USD, the company might be undercharging. If it's lower, they might be overcharging.
- Since the predicted value is significantly lower than the company's quote, it seems like the company might be overcharging for houses in this locality.

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.

- When comparing the performance of two models, the adjusted R-square can help us determine which model fits the data better. The higher the adjusted R-square, the better the model explains the variability in the data.
- The recent model with an adjusted R-square of 0.6371 performs better than the previous model with an adjusted R-square of 0.5432. The higher the adjusted R-square, the more variation in the data is explained by the model. So, the recent model explains a larger portion of the variability in the data compared to the previous model.

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	124.9045049	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	2.54E-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

- The adjusted R-square value of 0.688 suggests that approximately 68.8% of the variability in AVG_PRICE can be explained by the independent variables in the model.
- The intercept value is 29.241, which represents the predicted value of AVG_PRICE when all independent variables are zero.

❖ **For the other variables:**

- CRIME_RATE has a coefficient of 0.0487, but it is not statistically significant (p-value of 0.5347).
- AGE has a coefficient of 0.0328, which is statistically significant (p-value of 0.0127). This suggests that as the AGE variable increases, there is a positive impact on AVG_PRICE.
- INDUS has a coefficient of 0.1306, which is statistically significant (p-value of 0.0391). This indicates that as the INDUS variable increases, there is a positive impact on AVG_PRICE.
- NOX has a coefficient of -10.3212, which is statistically significant (p-value of 0.0083). This suggests that as the NOX variable increases, there is a negative impact on AVG_PRICE.
- Crime_rate is non-significant variable
- NOX, Tax, PTRATIO and LSTAT have negative coefficients.

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.

(HINT: Significant variables are those whose p-values are less than 0.05. If the p-value is greater than 0.05 then it is insignificant)

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

- Crime_rate is non-significant variable.
- Age, Indus, NOX, Distance, Tax, PTRATIO, Avg_room, LSTAT are significant variables.

a. Interpret the output of this model.

- The output of the model provides us with valuable information about the relationship between the independent variables and the predicted outcome, AVG_PRICE. It helps us understand how changes in the independent variables impact the average price.
- The significance of these relationships depends on the p-values associated with each coefficient. A lower p-value indicates a more statistically significant relationship.
- By looking at the coefficients, we can see the direction and significance of these relationships. As the AGE of a property increases by one unit, the predicted average price increases by 0.033 units. Similarly, as the INDUS (industrial proportion of land) increases by one unit, the predicted average price increases by 0.131 units. On the other hand, as the NOX (nitric oxides concentration) increases by one unit, the predicted average price decreases by 10.273 units. Finally, as the DISTANCE to employment centers increases by one unit, the predicted average price increases by 0.262 units.
- These coefficients help us understand the impact of each variable on the predicted average price.

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 current model has a slightly higher adjusted R-square value. This indicates that the current model may perform slightly better in explaining the variability of the dependent variable based on the independent variables. However, the difference between the two values is quite small, so the performance improvement may not be significant.

Q7 Adjusted R Square

Adjusted R Square	0.688298647

Q8 Adjusted R Square

Adjusted R Square	0.688683682

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?

The coefficients in ascending order are: -

NOX: -10.272

PTRATIO: -1.0717

LSTAT: -0.605

TAX: -0.014

AGE: 0.0329

INDUS: 0.1307

DISTANCE: 0.2615

AVG_ROOM: 4.125

	Coefficients
NOX	-10.2727051
PTRATIO	-1.07170247
LSTAT	-0.60515928
TAX	-0.01445235
AGE	0.03293496
INDUS	0.130710007
DISTANCE	0.261506423
AVG_ROOM	4.125468959
Intercept	29.42847349

- Based on the coefficients, we can see that NOX has the largest negative coefficient. This suggests that as the value of NOX (nitric oxides concentration) increases in a locality, the average price (AVG_PRICE) is expected to decrease.
- Higher NOX levels are often associated with increased pollution, which can negatively impact property values.
- So, if the value of NOX is more in a locality in this town, it is likely to have a negative impact on the average price.

d. Write the regression equation from this model.

AVG_PRICE = -10.27(NOX) - 1.07(PTRATIO) - 0.61(LSTAT) - 0.01(TAX) + 0.03(AGE) + 0.13(INDUS) + 0.26(DISTANCE) + 4.13(AVG_ROOM) + 29.43(Intercept)