Terro's Real Estate Agency

Business Report

Lohitha Mada
Great Learning
GLCA DA August 2023

Abstract

Terro's real-estate is an agency that estimates the pricing of houses in a certain locality. They do this by looking at different things about the houses. To do this, they hired an Auditor who studies various geographic features of a property like pollution level (NOX), crime rate, education facilities (pupil to teacher ratio), connectivity (distance from highway), etc. This helps in determining the price of a property.

The dataset includes the following variables and features:

CRIME_RATE: The crime rate in the area.

AGE: The age of the property.

INDUS: The proportion of non-retail business acres per town.

NOX: The nitric oxide concentration in parts per 10 million.

DISTANCE: The distance from the highway (in miles).

TAX: The property tax rate.

PTRATIO: The pupil-teacher ratio in schools.

AVG_ROOM: The average number of rooms per house.

LSTAT: The percentage of the lower status population.

AVG_PRICE: The average price of the property.

Our main goal is to look at all these things one by one and see how much they affect the price of a property in the locality. We want to understand how important each of these things is when it comes to deciding how much a property is worth. This will help us make better decisions when we want to know the price of a property in the real estate market.

Question 1: The first step to any project is understanding the data. For this step, Generate the summary statistics for each of the variables. What do you observe?

Answer:

CRIME_RATE		*AGE*		*INDUS*		*NOX*		*DISTANCE*	
Mean	4.871976285	Mean	68.57490119	Mean	11.13677866	Mean	0.554695059	Mean	9.549407115
Standard Error	0.129860152	Standard Error	1.251369525	Standard Error	0.304979888	Standard Error	0.005151391	Standard Error	0.387084894
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	2.921131892	Standard Deviation	28.14886141	Standard Deviation	6.860352941	Standard Deviation	0.115877676	Standard Deviation	8.707259384
Sample Variance	8.533011532	Sample Variance	792.3583985	Sample Variance	47.06444247	Sample Variance	0.013427636	Sample Variance	75.81636598
Kurtosis	-1.189122464	Kurtosis	-0.967715594	Kurtosis	-1.233539601	Kurtosis	-0.064667133	Kurtosis	-0.867231994
Skewness	0.021728079	Skewness	-0.59896264	Skewness	0.295021568	Skewness	0.729307923	Skewness	1.004814648
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	408.2371542	Mean	18.4555336	Mean	6.284634387	Mean	12.65306324	Mean	22.53280632
Standard Error	7.492388692	Standard Error	0.096243568	Standard Error	0.031235142	Standard Error	0.317458906	Standard Error	0.408861147
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	168.5371161	Standard Deviation	2.164945524	Standard Deviation	0.702617143	Standard Deviation	7.141061511	Standard Deviation	9.197104087
Sample Variance	28404.75949	Sample Variance	4.686989121	Sample Variance	0.49367085	Sample Variance	50.99475951	Sample Variance	84.58672359
Kurtosis	-1.142407992	Kurtosis	-0.285091383	Kurtosis	1.891500366	Kurtosis	0.493239517	Kurtosis	1.495196944
Skewness	0.669955942	Skewness	-0.802324927	Skewness	0.403612133	Skewness	0.906460094	Skewness	1.108098408
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

Inference:

Based on the summary statistics provided by the given dataset, here are my few observations:

There are 506 records for each of the variables.

1. **Crime rate:**
   The mean crime rate is 4.87.
   The median crime rate is 4.82, which is close to the mean.
   The skewness is nearly 0 indicating that the data is normally distributed.
2. **Age:**
   The average age of the properties is around 68.57 years.
   The median age is 77.5 years.
   The mode age is 100 years indicating there are of older properties.

The skewness is approximately -0.60, indicating a slight left-skewed distribution, indicating that there are more older properties in the dataset.

3. **Indus:**
   The average proportion of non-retail business acres per town is approximately 11.13.
   The median is 9.69.
   The sample variance, a measure of data spread is 47.06.
   The skewness is approximately 0.30, indicating a slightly right-skewed distribution, that there are more towns with lower proportions of non-retail businesses.
   The negative kurtosis value (-1.23) indicates that the distribution has lighter tails and is less peaked than a normal distribution.
4. **NOX:**
   The mean of nitric oxide concentration (NOX) is 0.55.
   The median is 0.538, which is close to the mean.
   The skewness is approximately 0.729, indicating a right-skewed distribution, that there are more data points with higher nitric oxide concentrations.
   The range of NOX values a lowest value of 0.385 to a highest value of 0.871, representing a total range of 0.486.
5. **Distance:**
   The mean distance is 9.54
   The median is 5 units which is lower than the mean.
   The skewness is approximately 1.00, indicating a right-skewed distribution. This suggests that there are more data points with longer distances.
   The range of distance values from a minimum of 1 unit to a maximum of 24 units, with a total range of 23 units.
6. **Tax rate:**
   The mean property tax rate (TAX) is approximately 408.24.
   The median is 330, which is lower than the mean.
   The mode is 666, indicating a repeated presence of certain properties with this tax rate.
   The skewness is approximately 0.66, indicating a right-skewed distribution.
7. **PTRATIO:**
   The mean pupil-teacher ratio is approximately 18.45.
   The median is 19.05, which is closer to the mean.
   The skewness is approximately -0.80, indicating a left-skewed distribution.
   The sum of all PTRATIO values in the dataset is 9,338.5.
8. **AVG_ROOM:**
   The mean number of average rooms per property is approximately 6.28.
   The median is 6.2085.
   The skewness is 0.40, indicating a slightly right-skewed distribution, says that more number of houses have less than 6 rooms.
   The kurtosis value is 1.891 indicating that the distribution has heavier tails and is more peaked than a normal distribution
9. **LSTAT:**
   The mean percentage of lower-status population (LSTAT) is approximately 12.65%.
   The median is 11.36.

The positive kurtosis value (0.493) indicates that the distribution has slightly heavier tails and is less peaked than a normal distribution.

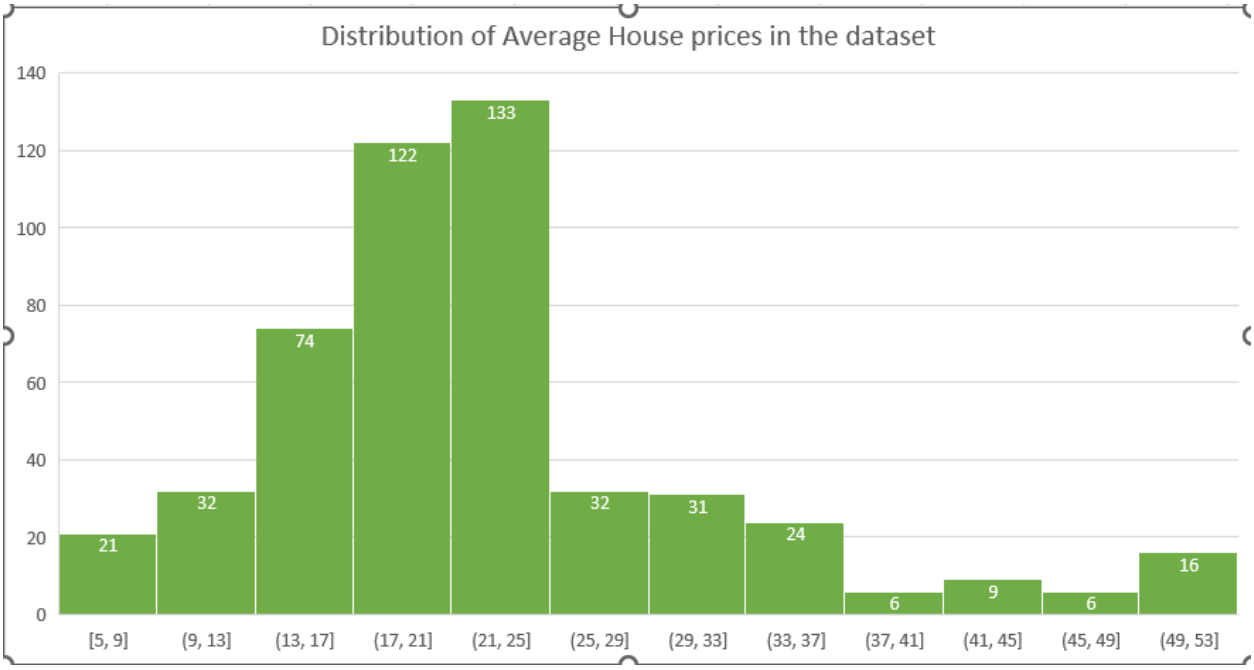
10. **Avg Price:**

The mean of average house prices (AVG_PRICE) is 22.53.
The median is 21.2
The skewness is 1.11, indicating a right-skewed distribution.
The positive kurtosis value 1.495 indicates that the distribution has heavier tails and is more peaked than a normal distribution.

Among the all-variable analysis, "Tax" has higher mean, median, and mode compared to other variables. On the other hand, "Average Price" exhibits a higher degree of skewness compared to the other variables, indicating that its distribution lies more towards higher price values.

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

Answer:



Inference:

- From the above histogram graph, we can see that most of the average house prices are in the range of $21000 to $25000
- We have less number of average house prices in between $37000 to $41000 and $45000 to $49000.
- The graph is right- tailed indicating a positive skewness.

Question 3: Compute the covariance matrix. Share your observations.

Answer:

	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	84.41955616

1. **Crime rate:**
   - Crime rate has high positive covariance with avg price with 1.162, which tells us the positive relation between crime rate and average price. As average price increase, crime rates may increase.
   - Crime rate has negative covariance with tax indicating that when property tax rates are higher, there is a tendency for crime rates to be lower.
2. **Age:**
   - Age has positive covariance with tax of 2397.94, indicating that when property tax rates are higher, older properties are associated with it.
   - Age has negative covariance with average price indicating that older properties tend to have lower average price.
3. **Indus:**
   - Indus has positive covariance with tax indicating that higher property tax rates are associated with a greater proportion of industrial land.
   - Indus has negative covariance with average price indicating that higher proportion of industrial land is associated with lower average property prices.
4. **Nox:**
   - Nox has positive covariance with tax indicating that higher nitrogen oxide concentration is associated with higher property tax rates.
   - Nox has negative covariance with average price indicating that higher proportion of nitrogen oxide concentration is results in lower average property prices.
5. **PTRATIO:**
   - PTRATIO has high positive covariance with LSTAT with 5.771 indicating that a higher pupil-teacher ratio is associated with a higher percentage of lower-income population.
   - PTRATIO has negative covariance with average price indicating that higher pupil-teacher ratio is associated with lower average property prices.

By analyzing the covariance values for each variable, we can observe that Tax has high covariance with almost every other feature except crime rate, which tells us that changes in the

property tax rate results in significant variations in other features, indicating its importance in understanding the dataset.

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.

Answer:

	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

Based on the correlation matrix of all the variables, we can observe that,

The top 3 positively correlated pairs are:

- Tax and Distance with a coefficient of **0.910**.
- Nox and Indus with a coefficient of **0.763**.
- Nox and Age with a coefficient of **0.731**.

The top 3 negatively correlated pairs are:

- Average price with LSTAT with a coefficient of **-0.737**.
- LSTAT with average room with a coefficient of **-0.613**.
- Average price with PTRATIO with a coefficient of **-0.507**.

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.

Answer:

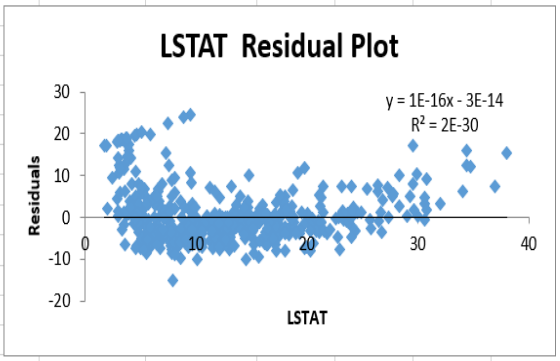
SUMMARY OUTPUT

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	-0.873950508



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

- From this model 54% of the variation in the average price is explained by the LSTAT.
- The coefficient of LSTAT in this model is -0.950049354. This says that if LSTAT for every 1-unit increase, the average price of the house is expected to decrease by 0.950049354 units.
- Intercept of LSTAT in this model is 34.55384088. Which tells us that if LSTAT becomes zero, the price will become 34.55.
- Based on the residual plot, most of the points are having uniform variance. But some errors are there in between 0-10 and 30-40, i.e., plot is not uniform between 0-10 and 30-40.

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

Answer:

Yes, LSTAT is significant variable for the average price from this model. As the p-value (**5.08E-88**) obtained from the model, which is significantly less than the conventional significance threshold of 0.05. We can say that LSTAT is 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.

Answer:

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?

Answer:

Regression Equation we obtained from this model is:

Y = 5.09 * Average room - 0.642 * LSTAT -1.358

Here Y = Average price (Dependent variable)

Y = 5.09 * 7 - 0.642 * 20 - 1.358

Y = 21.432

S0, the price for new house is $21.43.

As company is quoting a value of 30000 USD for this locality, 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.

Answer: Yes, the performance of this model is better than the previous model because the adjusted R-Square in this model (**0.637**) is higher than the adjusted R-square of the previous model (**0.543**). Also, we can see that 63% of variability for average price is explained by Avg_room and LSTAT combinely which says it is highly correlated. But in previous model LSTAT alone describes 54% of variability for average price.

Question 7: 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.

Answer:

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

Inference:

- Based on this model, we can say that CRIME_RATE is not a significant variable for average price of a house as p-value (Crime rate coefficient) is greater than 0.5. Which tells us that higher p-value of CRIME_RATE variable does not have a statistically significant impact on average house prices in this model.
- PTRATIO, AVG_ROOM and LSTAT have negative coefficients which says that increase in these features will result decrease in price of the house and vice versa.
- All the features combinely explains 69% of variability for average price of a house.

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.

Answer:

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

From this model, we can say that all the features are statistically 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?

Answer:

Adjusted R-square of this model is:

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

Adjusted R-square of previous model is:

Regression Statistics	
Multiple R	0.832978824
R Square	0.69385372
Adjusted R Square	0.688298647
Standard Error	5.1347635
Observations	506

Inference:

Comparing the adjusted R-squared values for both models, we can conclude that there is not a significant difference between the two models. Additionally, the variable CRIME_RATE remains statistically insignificant even when it's excluded from the model. This tells us that CRIME_RATE doesn't have an impact on the model's ability to explain variability in average house prices, regardless of whether it's included or not.

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?

Answer:

	Coefficients
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

Inference:

According to this model, if the concentration of nitrogen oxide (NOX) in the locality increases, the average price of the house is predicted to decrease by 10 times. This tells that for every unit increase in NOX concentration, that the average house price is expected to decrease significantly.

d. Write the regression equation from this model.

Y = 0.0329 * AGE + 0.1307 * INDUS - 10.272 * NOX + 0.261 * DISTANCE - 0.014 * TAX - 1.071 * PTRATIO + 4.125 * AVG_ROOM - 0.605 * LSTAT + 29.428

Where Y = Average price (Dependent variable)

Conclusion:

Based on overall analysis of complete dataset, we can conclude that:

- All the features in the model are important for estimating the average price of a house, except CRIME_RATE which has less significance compared to other features.
- Some features have negative coefficients (NOX, PTRATIO, AVG_ROOM and LSTAT) which says that an increase in the values of these features may decrease the average price of the house.