Terro's Real Estate Agency Project Report

Problem Statement (Situation):

“Finding out the most relevant features for pricing of a house” Terro’s real-estate is an agency that estimates the pricing of houses in a certain locality. The pricing is concluded based on different features / factors of a property. This also helps them in identifying the business value of a property. To do this activity the company employs 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**.**

Objective (Task):

Your job, as an auditor, is to analyze the magnitude of each variable to which it can affect the price of a house in a particular locality.

Question1: Generate the summary statistics for each variable in the table. (Use Data analysis tool pack). Write down your observation.

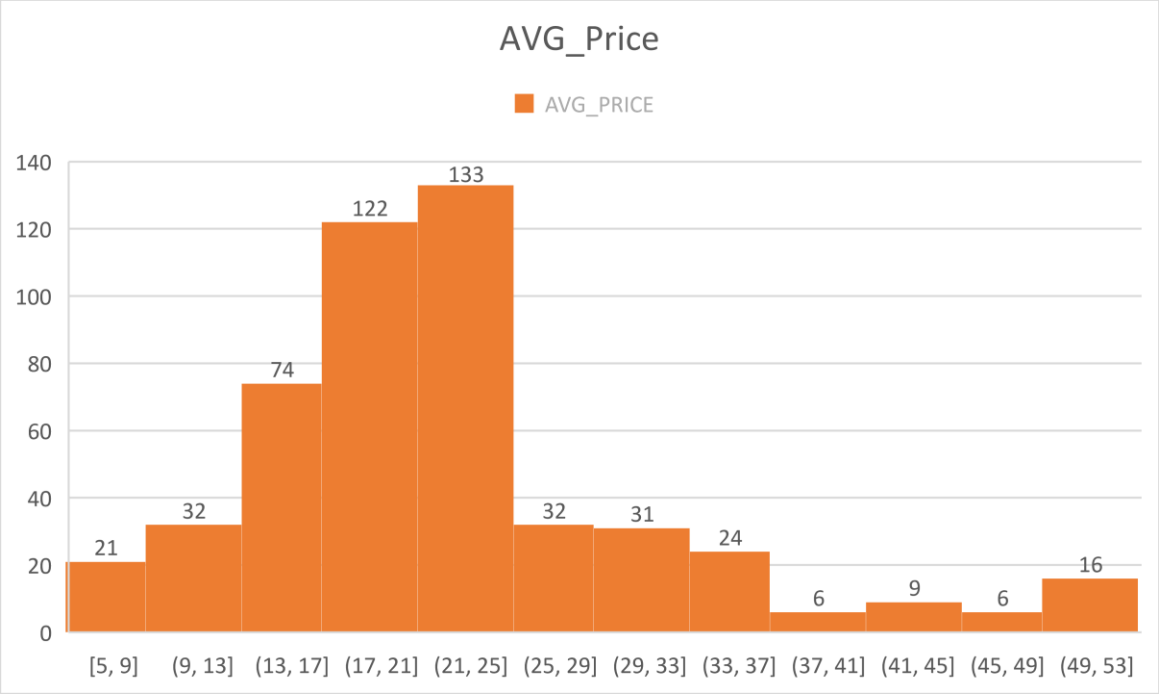
	CRIME_RATE		AGE		INDUS		NOX		DISTANCE		TAX		PTRATIO		AVG_ROOM		LSTAT		AVG_PRICE	
3	Mean	4.87198	Mean	68.5749	Mean	11.1368	Mean	0.5547	Mean	9.54941	Mean	408.237	Mean	18.4555	Mean	6.28463	Mean	12.6531	Mean	22.5328
4	Standard Error	0.12986	Standard Error	1.25137	Standard Error	0.30498	Standard Error	0.00515	Standard Error	0.38708	Standard Error	7.49239	Standard Error	0.09624	Standard Error	0.03124	Standard Error	0.31746	Standard Error	0.40886
5	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
6	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
7	Standard Deviation	2.92113	Standard Deviation	28.1489	Standard Deviation	6.860353	Standard Deviation	0.11586	Standard Deviation	8.70726	Standard Deviation	168.537	Standard Deviation	2.16495	Standard Deviation	0.70262	Standard Deviation	7.14106	Standard Deviation	9.1971
8	Sample Variance	8.53301	Sample Variance	792.358	Sample Variance	47.0644	Sample Variance	0.01343	Sample Variance	75.8164	Sample Variance	28404.8	Sample Variance	4.68699	Sample Variance	0.49367	Sample Variance	50.9948	Sample Variance	84.5867
9	Kurtosis	-1.1891	Kurtosis	-0.9677	Kurtosis	-1.2338	Kurtosis	-0.0647	Kurtosis	-0.8672	Kurtosis	-1.1424	Kurtosis	-0.2851	Kurtosis	1.8915	Kurtosis	0.49324	Kurtosis	1.4952
10	Skewness	0.02173	Skewness	-0.599	Skewness	0.29502	Skewness	0.72931	Skewness	1.00481	Skewness	0.66996	Skewness	-0.8023	Skewness	0.40361	Skewness	0.90646	Skewness	1.1081
11	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
12	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
13	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
14	Sum	2465.22	Sum	34698.9	Sum	5635.21	Sum	280.676	Sum	4832	Sum	206568	Sum	9338.5	Sum	3180.03	Sum	6402.45	Sum	11401.6
15	Count	506	Count	506	Count	506	Count	506	Count	506	Count	506	Count	506	Count	506	Count	506	Count	506

We may get a few observations from the descriptive statistics of the dataset by:

- There are 506 records total in the collection.
- First, if we look at the Distance variable, we can see that the maximum and mean values are 24 and 24, respectively. This states that the majority of residences are located away from highways.
- Tax range is 524, with a 408.2 average tax paid per taxpayer.
- We may infer that the dataset is substantially skewed based on the skewness of the variables.
- And if we take the age variable into account, the maximum age is 100 and the median is also 100, indicating that most residences are older than 100.
- Nitric oxide concentration (NOX) has the lowest standard error of 0.00515 and TAX has the largest standard error of 7.49 among the available data; as a result, sample means for NOX are widely dispersed about the population mean whereas sample means for TAX are narrowly distributed around the population mean.
- The data are more dispersed about the mean for TAX and are concentrated around the mean for NOX because TAX has the highest 168.537 standard deviation and NOX has the lowest 0.116.
- In the same way, when it comes to sample variance, NOX has the lowest sample variance (0.0134) and TAX has the most (28404.759), making NOX's data the most consistent and TAX's the least consistent.
- Kurtosis is a distribution that has a maximum value for AVG ROOM of 1.891 and a minimum value for INDUS of -1.233. This means that the distribution is too peaked for AVG ROOM and too flat (has light tails or lacks outliers) for INDUS.

- The distribution is longer on the right side of the peak for AVG PRICE and longer on the left side of the peak for PTRATIO. AVG PRICE has a high skewness of 1.108 and PTRATIO has a low skewness of -0.802.

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



- By looking at the data in the histogram, we can infer that the data is positively skewed because it is more evenly distributed towards the left side of the histogram, or has a long left tail.
- In summary, the majority of the homes are between $21,000 and $25,000.
- We have the fewest number of homes in the $37,000–$41,000 and $45,000–$49,000 range.

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	790.7925								
INDUS	-0.110215175	124.2678	46.97142974							
NOX	0.000625308	2.381212	0.605873943	0.013401099						
DISTANCE	-0.229860488	111.55	35.47971449	0.615710224	75.66653127					
TAX	-8.229322439	2397.942	831.7133331	13.02050236	1333.116741	28348.6236				
PTRATIO	0.068168906	15.90543	5.680854782	0.047303654	8.74340249	167.8208221	4.677726296			
AVG_ROOM	0.056117778	-4.74254	-1.884225427	-0.024554826	-1.281277391	-34.51510104	-0.539694518	0.492695216		
LSTAT	-0.882680362	120.8384	29.52181125	0.487979871	30.32539213	653.4206174	5.771300243	-3.073654967	50.89398	
AVG_PRICE	1.16201224	-97.3962	-30.46050499	-0.454512407	-30.50083035	-724.8204284	-10.09067561	4.484565552	-48.3518	84.41955616

- The data sets "Age vs Tax," "Indus vs Tax," and "Distance vs Tax" display significant covariance, indicating a direct positive relationship between them—when one variable increases, so does the other.
- The data sets "Tax vs Avg Price," "Age vs Avg Price," and "Lstat vs Avg Price" exhibit minimal covariance, signifying an inverse relationship. In other words, when one variable experiences an increase, the other data tends to decrease.

Question 4: Create a correlation matrix of all the variables (Use Data analysis tool pack).

	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

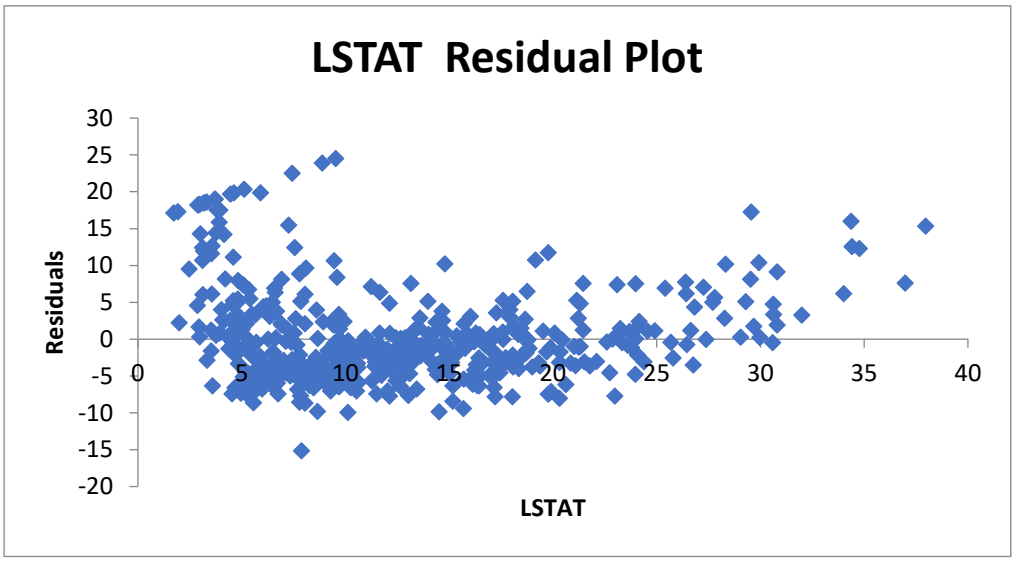
a) Which are the top 3 positively correlated pairs

Distance vs Tax, Nox vs Indus, Nox vs Age are top 3 positively corelated pairs

b) Which are the top 3 negatively correlated pairs.

Avg price vs Lstat, Lstat vsAvg room, Avg price vs Ptratio are top 3 negatively corelated pairs

Question 5: Build an initial regression model with AVG_PRICE as 'y' (Dependent variable) and LSTAT variable as Independent Variable. Generate the residual plot.



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

	Coefficients
Intercept	34.55384088
LSTAT	-0.950049354

Upon inspecting the coefficient value and the intercept value, it becomes apparent that a one-unit increase in the coefficient results in a corresponding decrease of approximately 0.95 in the average price. This indicates a somewhat negative (inverse) relationship between the coefficient and the average price. Additionally, the positive intercept value signifies a consistent increase in the price across all instances.

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

Based on the model, the LSTAT variable exhibits a remarkably low p-value, specifically 5.081 * 10^(-88), which is significantly smaller than the conventional significance threshold of 5%(0.05). This suggests that the LSTAT variable holds substantial statistical significance and can be confidently employed for further analysis.

Furthermore, when we examine the correlation, it becomes evident that LSTAT predominantly maintains a negative correlation with the other variables, indicating an inverse relationship. This implies that changes in the LSTAT variable are likely to have a notable impact on the average price.

Question 6: Build a new Regression model including LSTAT and AVG_ROOM together as Independent variables and AVG_PRICE as dependent variable.

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 you've provided is:

y = -1.358 + 5.094 * X0 - 0.6423 * X1

Where:

y = Avg_price

X0 = avg_room

X1 = LSTAT

Based on this model, when you calculate the average price (y) for a new house with 7 rooms (X0 = 7) and an LSTAT of 20 (X1 = 20), we obtain:

Y = -1.358 + 5.094 * 7 - 0.6423 * 20 = 21.454

The company's $30,000 quote for a property in this locality surpasses our model's prediction of $21,454, indicating a potential overcharge.

b) Is the performance of this model better than the previous model you built in Question 5? Compare in terms of adjusted R-square and explain.

Comparing the two models, the current one boasts an adjusted R-squared value of 0.637, while the previous model(Question 5) had a value of 0.543. This suggests that the current model outperforms the previous one, as its higher adjusted R-squared value signifies a better overall fit to the data.

Question 7: Build another Regression model with all variables where AVG_PRICE alone be the Dependent Variable and all the other variables are independent. Interpret the output in terms of adjusted R square, coefficient and Intercept values. Explain the significance of each independent variable with respect to AVG_PRICE.

- When examining the adjusted R-squared values, it's evident that this model possesses the highest adjusted R-squared value, which stands at 0.688. This indicates superior performance compared to the other models, as it demonstrates a stronger fit to the data.
- Upon interpreting the coefficients and the intercept, we can discern that variables such as Crime rate, Age, Indus, Distance, and Avg room exhibit a direct positive relationship with Average price. Conversely, Nox, Tax, Ptratio, and Lstat demonstrate an inverse (negative) relationship with Average price.
- Upon examination of the p-values, it becomes evident that the only variable lacking statistical significance in this model is Crime rate, as indicated by its p-value of 0.535. All other variables are statistically significant in relation to Average price because their p-values are less than 5% or 0.05.
- Upon reviewing the correlation table, it becomes apparent that Crime rate and Avg room are the variables with positive correlation, while all other variables exhibit negative correlations with Average price. Therefore, we can conclude that Crime rate and Avg room are the significant variables influencing Average price in this context.

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 and answer the questions below:

a) Interpret the output of this model.

- With an accuracy rate of 81.52%, this model surpasses the performance of other models, signifying its superior predictive capability.
- The output of this model includes only the variables that are statistically significant in predicting AVG_PRICE.
- The coefficient values represent the change in AVG_PRICE for a one-unit change in the corresponding independent variable, holding all other variables constant.
- The intercept represents the estimated AVG_PRICE when all independent variables are zero.

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?

- When comparing the R-squared value of this model (0.6886) with the previous one (0.6882), it is evident that this model has a slightly higher R-squared value. This suggests that this model is marginally better for making predictions compared to the previous one, as it explains a slightly greater proportion of the variance in the data.

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
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

Indeed, with a negative coefficient (-10.272) and a negative correlation value (-0.427), it is evident that NOX and AVG price have an inverse (negative) relationship. As the value of NOX increases, the value of AVG price tends to decrease by 10times approx..

d) Write the regression equation from this model.

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