BUSINESS REPORT

TERRO'S REAL ESTATE AGENCY

Vishwanath Anand Sahu 12/10/23 Auditor

PROBLEM STATEMENT

“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.

ABOUT THE DATASET

Data Dictionary:

	Attribute Description
CRIME RATE	per capita crime rate by town
INDUSTRY	proportion of non-retail business acres per town (in percentage terms)
AVG_ROOM	average number of rooms per house
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
PTRATIO	pupil-teacher ratio by town
AGE	proportion of houses built prior to 1940 (in percentage terms)
DISTANCE	distance from highway (in miles)
NOX	nitric oxides concentration (parts per 10 million)
AGE	proportion of houses built prior to 1940 (in percentage terms)

BUSINESS REPORT

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

The observation that we infer from the summary is the following

CRIME RATE

It shows that the average per capita crime rate is around 4.87 and the most occurred rate is around 3.43, the data is positively skewed and is platykurtic which means the observations are trailing towards the right and as it has negative kurtosis it shows that the observations are more concentrated and few outliers are there .

AGE

In this the average age is around 68.57 and higher half of the age is around 77.5 and the most frequent age turns out to be 100. The minimum age is 2.9 whereas the maximum is around 100 which is the most frequent .

INDUS

In case of proportion of non-retail business the average turns out to be 11.13% , , the data is positively skewed and is platykurtic which means the observations are trailing towards the right and as it has negative kurtosis it shows that the observations are more concentrated and few outliers are there.

NOX

The average nitric oxide concentration turns out to be 0.55 part per 10 million and the most frequent turns out to be 18.1 , the minimum concentration turns out to be 0.46 whereas the maximum turns out to be 27.74.

DISTANCE

The average distance of the property from the highway turns out to be 9.55 miles, and the higher 50% population has a distance of 5 miles from the highway , the minimum miles turns out to be 1 and maximum is around 24.

TAX

The average tax turns out to be 408.23 per $10,000 , the minimum tax turns out to be around 187 and the maximum turns out to be 711.

PTRATIO

The average pupil to teacher ratio per town turns out to be 18.45 which means for every 1 teacher there are 18 students the higher 50% of the population has 19 students per 1 teacher , in this case the data is negatively skewed and platykurtic , which means the data is flat and trailing towards the left ,and has less outliers and all the data is concentrated towards the between.

AVG PRICE

The average price of the houses turns out to be 22.53 ($1000) and the higher 50% of the price is around 21.2 ($1000 in this case the data is positively sewed and leptokurtic , which means that the data is trailing towards the right and has many outliers .

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

1) Most of the Houses In Boston Have The Average Price Between $21,000-$25000
2) 50% of Family that lives in Boston have house value under 25k and rest have value above 25k.
3) The Least Count Of the house is between The Price Range of $37000-$41000 &$45000-$49000

From the above histogram data we infer that the starting price range of the house turns out to be $5,000 the average price of the house is around 22.53 ($1000) but the median is 21.2, indicating that the distribution is not symmetric, the data is positively skewed and leptokurtic which means it has many outliers towards the higher price .

3. Compute the covariance matrix. Share your observations.

	CRIME_RATE	*AGE*	*INDUS*	*NOX*	*DISTANCE*	*TAX*	*PTRATIO*
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
LSTAT	0.882680362	120.8384405	29.52181125	0.487979871	30.32539213	653.4206174	5.771300243
AVG_PRICE	-1.16201224	-97.39615288	-30.46050499	-0.454512407	-30.50083035	-724.8204284	-10.09067561

The covariance matrix is a square matrix that measures the degree of linear relationship between two or more variables. A positive covariance indicates that two variables have positive relation with each other , while a negative covariance indicates vice -versa. A zero covariance means that there is no linear relationship between the variables.

The pair of variables with the highest positive covariance is

- age-tax
- distance-tax
- industry-tax

This means that these variables tend to increase and decrease together, suggesting a possible causal relationship or common factor influencing them.

The pair of variables with the highest negative covariance is:

- tax-avg price
- age-avg price
- lstat-avg price

This means that these variables tend to move in opposite directions, implying an inverse relationship or trade-off between them.

4. Create a correlation matrix of all the variables (Use Data analysis tool pack).
a) Which are the top 3 positively correlated pairs and
b) Which are the top 3 negatively correlated pairs.

	CRIME_RATE	*AGE*	*INDUS*	*NOX*	*DISTANCE*	*TAX*	*PTRATIO*	*AVG_ROOM*	*LSTAT*	*AVG_PRICE*
CRIME_RATE	1									
AGE	0.006859	1								
INDUS	-0.00551	0.644779	1							
NOX	0.001851	0.73147	0.763651	1						
DISTANCE	-0.00906	0.456022	0.595129	0.611441	1					
TAX	-0.01675	0.506456	0.72076	0.668023	0.910228	1				
PTRATIO	0.010801	0.261515	0.383248	0.188933	0.464741	0.460853	1			
AVG_ROOM	-0.02739	0.240626	0.391685	0.302019	0.209085	0.292055	0.355	1		
LSTAT	0.602339	0.603	0.590879	0.488676	0.543993	0.374044	-0.61381	1		
AVG_PRICE	-0.04333	0.376	0.483	0.427	0.381	0.468	0.507	0.69536	0.73766	1

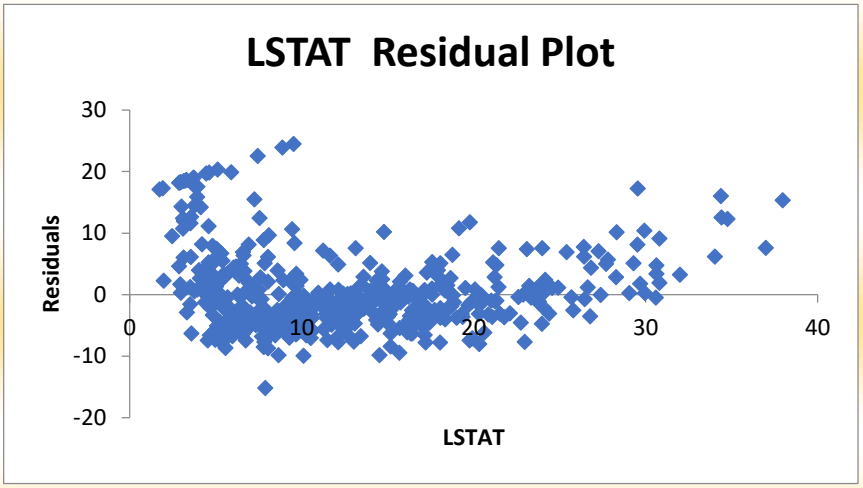
a) Top 3 positively co-related pairs are
DISTANCE & TAX
INDUSTRY & NOX
AGE & NOX

b) Top 3 negatively co-related pairs are
PTRATIO & AVG PRICE
LSTAT & AVG PRICE
AVG ROOM & LSTAT

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?

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



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

a) **R-squared**

This value tells us the proportion of the variance in the dependent variable (AVG_PRICE) that is predictable from the independent variable (LSTAT). If R-squared is greater than 50% it indicate a better fit to the model.

INTERCEPT

It tells us the avg price when LSTAT is zero .It's the point where regression line crosses the y-axis .

RESIDUAL PLOT

This plot shows the difference between the observed and predicted values (residuals) for each observation. If the residuals are randomly scattered around zero, it suggests that the model is a good fit. If the plot shows a linear pattern it suggests that the model is missing some information .

b) Significance on the basis of the LSTAT variable , The p-value for LSTAT in the regression summary tells us whether LSTAT is a significant predictor of AVG_PRICE. If the p-value is less than 0.05, we can conclude that LSTAT is a significant predictor.

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? 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

a) Regression equation for two independent variable is

Y = b + b1 * X1 + b2 * X2

b = -1.35

b1 = 5.09

b2 = -0.64

Average Rooms = 7, L-STAT = 20

AVG_PRICE = -1.35 + 5.09*7 + (-0.64)*20 = 21.48

The predicted value according to the model is 21,480 which is less than 30,000 so in this case we can conclude that company is overcharging .

b) Since the adjusted R square value for **this model is 0.6371** which is higher then **0.5432 in question 5**, in question 5 there was only one independent variable so if one more independent value is added and the adjusted r square is increasing so it shows the model in question 6 is better

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.

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

	Coefficients	*P-value*
Intercept	29.24131526	2.53978E-09
CRIME_RATE	0.048725141	0.534657201
AGE	0.032770689	0.012670437
INDUS	0.130551399	0.03912086
NOX	-10.3211828	0.008293859
DISTANCE	0.261093575	0.000137546
TAX	-0.01440119	0.000251247
PTRATIO	-1.074305348	6.59E-15
AVG_ROOM	4.125409152	3.89287E-19
LSTAT	-0.603486589	8.91071E-27

In this case where we took all the variables to create the regression output the adjusted r square came out to be 0.68 whereas the r-square came out to be 0.69 ,this means that most of the predictor in this model are adding some values to the model
The intercept represents the estimated average price when all other predictors are zero. In this case, the intercept is statistically significant, suggesting that even when other predictors are zero, the average price is significantly different from zero.
Significance of each variable in respect to the AVG price is :

CRIME_RATE
Significance: Not significant (p-value = 0.5347)
Interpretation: The coefficient for CRIME_RATE is not statistically significant at the 0.05 significance level. This suggests that, based on the available data, there is not enough evidence to conclude that CRIME_RATE has a significant linear relationship with AVG_PRICE.

AGE
Significance: Significant (p-value = 0.0127)
Interpretation: The coefficient for AGE is statistically significant. A one-unit increase in AGE is associated with an increase in AVG_PRICE by 0.0328 units.

INDUS
Significance: Significant (p-value = 0.0391)
Interpretation: The coefficient for INDUS is statistically significant. A one-unit increase in INDUS is associated with an increase in AVG_PRICE by 0.1306 units.

NOX
Significance: Significant (p-value = 0.0083)
Interpretation: The coefficient for NOX is statistically significant. A one-unit increase in NOX is associated with a decrease in AVG_PRICE by 10.3212 units.

DISTANCE
Significance: Highly significant (p-value = 0.0001)
Interpretation: The coefficient for DISTANCE is highly statistically significant. A one-unit increase in DISTANCE is associated with an increase in AVG_PRICE by 0.2611 units.

TAX
Significance: Significant (p-value = 0.0003)
Interpretation: The coefficient for TAX is statistically significant. A one-unit increase in TAX is associated with a decrease in AVG_PRICE by 0.0144 units.

PTRATIO
Significance: Highly significant (p-value = 6.58642E−15)
Interpretation: The coefficient for PTRATIO is highly statistically significant. A one-unit increase in PTRATIO is associated with a decrease in AVG_PRICE by 1.0743 units.

AVG_ROOM
Significance: Highly significant (3.89287E-15)
Interpretation: The coefficient for AVG_ROOM is highly statistically significant. A one-unit increase in AVG_ROOM is associated with an increase in AVG_PRICE by 4.1254 units.

LSTAT
Significance: Highly significant (8.9107E-27)
Interpretation: The coefficient for LSTAT is highly statistically significant. A one-unit increase in LSTAT is associated with a decrease in AVG_PRICE by 0.6035 units.

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.
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?
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?
d) Write the regression equation from this model?

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

a) This model is significant enough as the R square value is 0.6936 , and the adjusted r square is also closer to the actual r square which shows that the predictor in this model are significant predictors .

b) The adjusted r square of the previous model was 0.688298647 whereas the current model adjusted r square is 0.688683682 which shows that the current model excluding crime rate is minutely better as compared to the previous model.

VARIABLE	COEFFICIENT
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

c) From the table data we can see if we sort the coefficient value in an ascending order we got to know that NOX is the most negative coefficient , so it will have a negative impact on the price .
if the NOX concentration increases the average price of the house will decrease.

d) The regression equation of this model will be
AVERAGE PRICE = 29.42 + (0.032*AGE) + (0.130 * INDUS) + (-10.27 * NOX) + (0.261 * DISTANCE) +(0.0144 * TAX) + (-1.0717 * PTRATIO) + (4.125 * AVG_ROOM**)** + (-0.605*LSTAT)