Assignment – Terro's real estate agency

Real estate data analysis – Exploratory data analysis, Linear Regression

BUSINESS 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 analyse the magnitude of each variable to which it can affect the price of a house in a particular locality.

The agency has provided a dataset of 506 houses in Boston. Following are the details of 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)*
NOX	*nitric oxides concentration (parts per 10 million)*
AVG_ROOM	*average number of rooms per house*
AGE	*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*

INTRODUCTION

In the dynamic landscape of real estate, making informed decisions is paramount for success. The Terro's Real Estate Data Analysis project delves into a comprehensive exploration of key variables influencing property prices. Leveraging statistical methods and regression model, this analysis aims to uncover patterns, relationships, and valuable insights within the Terro's real estate market.

Through a careful examination of variables such as crime rates, distance from essential amenities, tax rates, and more, this project seeks to empower stakeholders with actionable intelligence. The exploration spans from uncovering correlations between different factors to constructing predictive models that contribute to a understanding of property valuation.

In this project I have to analyse the magnitude of each variable to which it can affect the price of a house in a particular locality.

To do the analysis, we are expected to solve these questions: -

1) 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	
Mean	4.871976	Mean	68.5749	Mean	11.13678	Mean	0.554695	Mean	9.549407
Standard Error	0.12986	Standard E	1.25137	Standard E	0.30498	Standard E	0.005151	Standard E	0.387085
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	2.921132	Standard D	28.14886	Standard D	6.860353	Standard D	0.115878	Standard D	8.707259
Sample Variance	8.533012	Sample Va	792.3584	Sample Va	47.06444	Sample Va	0.013428	Sample Va	75.81637
Kurtosis	-1.18912	Kurtosis	-0.96772	Kurtosis	-1.23354	Kurtosis	-0.06467	Kurtosis	-0.86723
Skewness	0.021728	Skewness	-0.59896	Skewness	0.295022	Skewness	0.729308	Skewness	1.004815
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
Largest(1)	9.99	Largest(1)	100	Largest(1)	27.74	Largest(1)	0.871	Largest(1)	24
Smallest(1)	0.04	Smallest(1	2.9	Smallest(1	0.46	Smallest(1	0.385	Smallest(1	1
Confidence Level(95.0%)	0.255133	Confidenc	2.458531	Confidenc	0.599186	Confidenc	0.010121	Confidenc	0.760495

TAX		PTRATIO		AVG_ROOM		LSTAT		AVG_PRICE	
Mean	408.2372	Mean	18.45553	Mean	6.284634	Mean	12.65306	Mean	22.53281
Standard E	7.492389	Standard E	0.096244	Standard E	0.031235	Standard E	0.317459	Standard E	0.408861
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 D	168.5371	Standard D	2.164946	Standard D	0.702617	Standard D	7.141062	Standard D	9.197104
Sample Va	28404.76	Sample Va	4.686989	Sample Va	0.493671	Sample Va	50.99476	Sample Va	84.58672
Kurtosis	-1.14241	Kurtosis	-0.28509	Kurtosis	1.8915	Kurtosis	0.49324	Kurtosis	1.495197
Skewness	0.669956	Skewness	-0.80232	Skewness	0.403612	Skewness	0.90646	Skewness	1.108098
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
Largest(1)	711	Largest(1)	22	Largest(1)	8.78	Largest(1)	37.97	Largest(1)	50
Smallest(1	187	Smallest(1	12.6	Smallest(1	3.561	Smallest(1	1.73	Smallest(1	5
Confidenc	14.72009	Confidenc	0.189087	Confidenc	0.061367	Confidenc	0.623703	Confidenc	0.803278

From Descriptive statistics of the given datasets we can get few observations: -

- *From the above summary statistics it can be observed that the houses in the dataset have a wide age range from 2.9 to 100 years. The majority of the houses are on the older side, with the median age of 77.5 .*
- *The average tax paid is 408.2 and tax range is 524.*
- *The average age of houses is 68.5749*
- *It can be observed that the average crime rate is 4.87, but some areas are experiencing high crime rate reaching up to 9.99.*
- *The mode of the average price is 50*
- *The average house value is approximately is 22.53 , some houses have the higher value contributing to the positive skewness .*
- *on average , about 11.14% of land is used for non- retail business , some town have a high percentage for non retail business areas, contributing to the positive skewness.*

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

Skewness: - 1.10481

Median: - 1.2

Min: - 5

Max: - 50

A skewness value of 1.104811 indicates a positive skewness in the distribution of "AVG_PRICE" Variable . A positive skewness indicates that the distribution is skewed to the right , meaning that the tail of the distribution is longer on the right side . Suggesting that the majority of houses have prices clustered towards the lower end, with a few higher priced outliers causing the tail of the distribution to extend towards higher value .

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	84.41955616

Observation: -

- *The covariance value of 28348.6236 between "TAX" and " TAX " has the highest covariance value of all i.e TAX with itself . It is important to note that covariance of a variable with itself is equal to its variance .*
- *The second highest covariance value is 2397.941723 which is between "TAX" and "AGE" which indicates a relative strong relationship between the property tax rate ("TAX") and the age of Houses ("AGE").*
- *The lowest covariance value in the provided matrix is -724.8204284, and it corresponds to the covariance between "AVG_PRICE" and "DISTANCE". Which indicates negative covariance i.e as the distance from highway increases , there tends to be decrease in the average house price.*

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

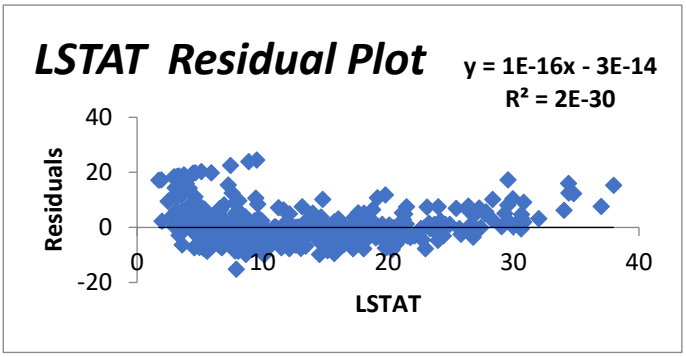
(a) Which are the top 3 positively correlated pairs

- *DISTANCE and TAX (0.9102) - As the distance from highway increase, the property tax tends to increase.*
- *INDUS and NOX (0.7637) –There is a strong positive correlation between the proportion of non-retail business acres per town (INDUS) and nitric oxides concentration (NOX).*
- *AGE and NOX (0.7315) –There is a positive correlation between the proportion of houses built prior to 1940 (AGE) and nitric oxides concentration (NOX).*

(b) Which are the top 3 negatively correlated pairs.

- *LSTAT and AVG_PRICE(-0.737662726) - This implies areas with a higher percentage of lower status have lower house prices.*
- *AVG_ROOM and LSTAT(-0.6954) - This implies may be higher percentage of lower status population tend to have houses with fewer rooms.*
- *AVG_PRICE and PTRATIO(-0.507786686) - May be this implies higher pupil- teacher ratios tends to have lower averages house price .*

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



SUMMARY OUTPUT	
MODEL "A"	
Regression Statistics	
Multiple R	0.737663
R Square	0.544146
Adjusted R Square	0.543242
Standard Error	6.21576
Observations	506

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

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	34.55384088	0.562627355	61.41515	3.7E-236
LSTAT	-0.950049354	0.038733416	-24.5279	5.08E-88

- *The R-squared value of 0.54 indicates that approximately 54% of the variability in AVG_PRICE . This suggest model is moderate level good fit.*
- *The coefficient for LSTAT is -0.95, implying that for each one-unit increase in LSTAT, AVG_PRICE is expected to decrease by 0.95 units. This negative coefficient indicates an inverse relationship between LSTAT and AVG_PRICE.*
- *The intercept is 34.55, representing the estimated AVG_PRICE when LSTAT is zero. This intercept provides the baseline value.*
- *To fully assess the model's performance, examining the residual plot is crucial. If the residuals are randomly scattered around zero, it suggests that the model is capturing the underlying patterns in the data. By further exploring the residual plot we can see that the points are randomly dispersed , we can conclude that a linear model ia an appropriate model.*

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

- *yes, the LSTAT variable appears to be significant for the analysis based on the regression model. The t-statistic for LSTAT is -24.53, and its associated p-value is very close to zero (5.08e-88). So according to this we can say that LSTAT is a significant variable according to the model.*
- *In practical terms, this implies that the percentage of lower-status population (LSTAT) is a meaningful predictor of the average housing price (AVG_PRICE). The negative coefficient further indicates that as the percentage of lower-status population increases, the average housing price tends to decrease.*

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

	Coefficients	andard Err	t Stat	P-value
Intercept	-1.358272812	3.172828	-0.4281	0.668765
AVG_ROO	5.094787984	0.444466	11.46273	3.47E-27
LSTAT	-0.642358334	0.043731	-14.6887	6.67E-41

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?***

*AVG_PRICE = -1.358272812 +(-0.642358334*LSTAT) +(5.094787984 * AVG_ROOM)*
AVG_PRICE = -1.358272812 +(-0.642358334 20) +(5.094787984 * 7)*
AVG_PRICE = 21.44

Comparing this to the company quoting a value of 30000 USD, we observe that the predicted AVG_PRICE of 21.44 is significantly lower. This suggests that the company quoting 30000 USD might be overcharging, as per the model's prediction.

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.***

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

MODEL "B"	
Regression Statistics	
Multiple R	0.7991
R Square	0.638562
Adjusted R Square	0.637124
Standard Error	5.540257
Observations	506

The adjusted R-square in the new model is 0.637124475, while in the previous model (Question 5), it was 0.543241826. A higher adjusted R-square indicates a better fit of the model to the data. Therefore, the new model performs better than the previous one.

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 Rsquare, coefficient and Intercept values. Explain the significance of each independent variable with respect to AVG_PRICE.***

	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	6.58642E-15
AVG_ROOM	4.125409152	3.89287E-19
LSTAT	-0.603486589	8.91071E-27

From this we can say that crime rate is not a significant variable for average price of an house as p-value is greater than 0.5 . All the features combinely explains 69% of variability for average price of a house. NOX, TAX, PTRATIO and LSTAT have negative coefficients which says that increase in these features will result decrease in price of the house and vice- versa.

CRIME_RATE: - *Coefficient (0.0487): For each unit increase in CRIME_RATE, AVG_PRICE is expected to increase by 0.0487 units .P-Value (0.53): The p-value is relatively high (greater than 0.05), i.e CRIME_RATE may not be a statistically significant predictor of AVG_PRICE in this model***.**

AGE: - *Coefficient (0.0328): For each additional year of AGE, AVG_PRICE is expected to increase by 0.0328 units. P-Value (0.01): The low p-value indicates that AGE is statistically significant in predicting AVG_PRICE.*

INDUS: - *Coefficient (0.1306): For each additional unit of INDUS, AVG_PRICE is expected to increase by 0.1306 units. P-Value (0.04): The low p-value suggests that INDUS is statistically significant in predicting AVG_PRICE*

NOX: - *Coefficient (-10.3212): For each unit increase in NOX, AVG_PRICE is expected to decrease by 10.3212 units. P-Value (0.008): The low p-value indicates that NOX is statistically significant in predicting AVG_PRICE.*

DISTANCE: - *Coefficient (0.2611): For each unit increase in DISTANCE, AVG_PRICE is expected to increase by 0.2611 units. P-Value (0.0001): The very low p-value indicates that DISTANCE is highly statistically significant in predicting AVG_PRICE.*

TAX: - *Coefficient (-1.0743): For each additional unit of TAX, AVG_PRICE is expected to decrease by 1.0743 units. P-Value (0.0025): The low p-value indicates that TAX is statistically significant in predicting AVG_PRICE.*

PTRATIO: - *Coefficient (-1.0743): For each additional unit of PTRATIO, AVG_PRICE is expected to decrease by 1.0743 units. P-Value (6.59e-15): The very low p-value indicates that PTRATIO is highly statistically significant in predicting AVG_PRICE.*

AVG_ROOM: - *Coefficient (4.1254): For each additional room in AVG_ROOM, AVG_PRICE is expected to increase by 4.1254 units. P-Value (3.89e-19): The very low p-value indicates that AVG_ROOM is highly statistically significant in predicting AVG_PRICE.*

LSTAT: - *Coefficient (-0.6035): For each unit increase in LSTAT, AVG_PRICE is expected to decrease by 0.6035 units. P-Value (8.91e-27): The very low p-value indicates that LSTAT is highly statistically significant in predicting 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 and answer the questions below:

a) Interpret the output of this model.

	Coefficients	*P-value*
Intercept	29.4284735	1.846E-09
AGE	0.03293496	0.01216288
INDUS	0.13071001	0.03876167
NOX	-10.2727051	0.00854572
DISTANCE	0.26150642	0.00013289
TAX	-0.01445235	0.00023607
PTRATIO	-1.07170247	7.0825E-15
AVG_ROOM	4.12546896	3.6897E-19
LSTAT	-0.60515928	5.4184E-27

In this model, AVG_PRICE is significantly influenced by factors such as NOX, PTRATIO, LSTAT, AVG_ROOM, and others. The estimated AVG_PRICE changes based on variations in these variables, providing insights into their impact on housing prices.

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?

MODEL "D" (QUE 8)	
Regression Statistics	
Multiple R	0.83283577
R Square	0.69361543
Adjusted R Square	0.6887

MODEL "C" (QUE 7)	
Regression Statistics	
Multiple R	0.83297882
R Square	0.69385372
Adjusted R Square	0.6883

The adjusted R-square value for this model (0.6887) is slightly higher than the previous model (0.6883), indicating a marginally better fit. It suggests that the selected significant variables explain a slightly larger proportion of the variability in AVG_PRICE compared to the model with all variables.

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

Sorting the coefficients in ascending order, the variable with the least impact is INDUS, while the variable with the most impact is AVG_ROOM. If the value of NOX increases in a locality, according to this model, it would lead to a decrease in the average price.

d) *Write the regression equation from this model*

$$AVG\_PRICE = 29.43 + 0.03 \times AGE + 0.13 \times INDUS - 10.27 \times NOX + 0.26 \times DISTANCE - 0.01 \times TAX - 1.07 \times PTRATIO + 4.13 \times AVG\_ROOM - 0.61 \times LSTAT$$

This equation represents the estimated AVG_PRICE based on the values of the significant variables.

SUMMARY AND OUTCOMES

In the Terro's Real Estate Project, we took a deep dive into understanding what makes Terro's properties tick. We looked at things like crime rates, how close properties are to important places, and the layout of rooms. It was like solving a puzzle to figure out the secrets behind property prices.

KEY OUTCOMES: -

1. ***Exploratory Data Analysis (EDA):*** *Through statistical analyses and visualizations, the project unveiled patterns and relationships within the dataset, providing a holistic understanding of Terro's real estate.*
2. ***Covariance and Correlation Matrices:*** *The identification of covariance and correlation relationships shed light on the interplay between different variables, guiding towards a nuanced understanding of their impact on property prices.*
3. ***Regression Models:*** *The construction of regression models, both single and multiple variable, enabled the prediction of property prices based on significant factors. The models not only offered predictive capabilities but also showcased the significance of individual variables.*
4. ***Residual Analysis:*** *The project included a detailed examination of residuals, providing valuable insights into the accuracy and reliability of the regression models.*
5. ***Model Comparisons:*** *By comparing models with different variables, the project assessed the efficacy of each in explaining the variance in property prices, allowing for informed decision-making.*

CONCLUSION: -

The Terro's Real Estate Project isn't just about numbers; it's about making sense of the real estate world. We found clues, made predictions, and now have a better guide for anyone interested in Terro's properties. It's like having a treasure map for the real estate adventure!