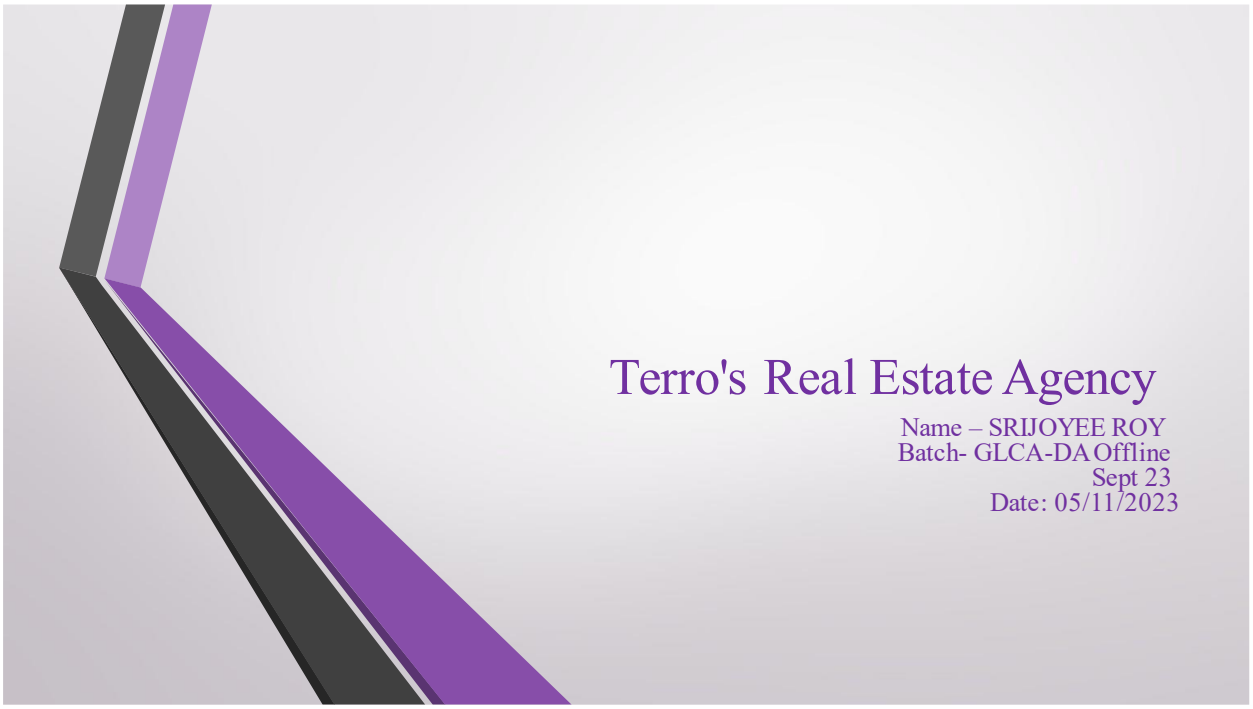
Terro's Real Estate Agency
Name – SRIJOYEE ROY
Batch- GLCA-DA Offline
Sept 23
Date: 05/11/2023

Table of contents –

Content –

List of Figures –

List of Tables –

Executive Summary –

A wide range of properties and land types are handled by the Terro Real Estate Agency. The placements of the characteristics in the dataset are used to categorize them. The property's numerous features and attributes are taken into account while determining its price. In this issue statement, we'll look at the property's attributes and how they impact the asking price.

Introduction –

This entire exercise aims to investigate the dataset. as well as analyze the exploratory data. Examine the dataset by adjusting different variables and analytic tools. There are 10 distinct features out of 506 distinct attributes in the data. Analyzing the property's many qualities can aid in determining the average price of the property.

Data description –

CRIME RATE	per capita crime rate by town
INDUSTRY	proportion of non-retail business acres per town (in percentage terms)
NOX	nitric oxide 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 the 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

Table 1. Data Description

Sample Dataset –

CRIME_RATE	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	AVG_ROOM	LSTAT	AVG_PRICE
6.32	65.2	2.31	0.538	1	296	15.3	6.575	4.98	24
4.31	78.9	7.07	0.469	2	242	17.8	6.421	9.14	21.6
7.87	61.1	7.07	0.469	2	242	17.8	7.185	4.03	34.7
6.47	45.8	2.18	0.458	3	222	18.7	6.998	2.94	33.4
5.24	54.2	2.18	0.458	3	222	18.7	7.147	5.33	36.2
9.75	58.7	2.18	0.458	3	222	18.7	6.43	5.21	28.7
9.42	66.6	7.87	0.524	5	311	15.2	6.012	12.43	22.9
2.76	96.1	7.87	0.524	5	311	15.2	6.172	19.15	27.1
7.66	100	7.87	0.524	5	311	15.2	5.631	29.93	16.5
1.12	85.9	7.87	0.524	5	311	15.2	6.004	17.1	18.9

Table 2. Dataset sample

Ten different parameters and features with the five different properties of the dataset. The property's price is established based on its attributes.

Exploratory Dataset –

Check the types of variables in the dataset –

Crime rate	Float64
Age	Float64
Indus	Float64
NOX	Float64
Distance	Int64
Tax	Int64
PTRatio	Float64
Avg_Room	Float64
LSTAT	Float64
Avg_Price	Float64

Table 3. Exploratory Dataset

Check for the missing values in the dataset –

The dataset has no missing values.

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

CRIME_RATE	
Mean	4.871976
Standard Error	0.12986
Median	4.82
Mode	3.43
Standard Deviation	2.921132
Sample Variance	8.533012
Kurtosis	-1.18912
Skewness	0.021728
Range	9.95
Minimum	0.04
Maximum	9.99
Sum	2465.22
Count	506

AGE	
Mean	68.5749
Standard Error	1.25137
Median	77.5
Mode	100
Standard Deviation	28.14886
Sample Variance	792.3584
Kurtosis	-0.96772
Skewness	-0.59896
Range	97.1
Minimum	2.9
Maximum	100
Sum	34698.9
Count	506

INDUS	
Mean	11.13678
Standard Error	0.30498
Median	9.69
Mode	18.1
Standard Deviation	6.860353
Sample Variance	47.06444
Kurtosis	-1.23354
Skewness	0.295022
Range	27.28
Minimum	0.46
Maximum	27.74
Sum	5635.21
Count	506

NOX	
Mean	0.554695
Standard Error	0.005151
Median	0.538
Mode	0.538
Standard Deviation	0.115878
Sample Variance	0.013428
Kurtosis	-0.06467
Skewness	0.729308
Range	0.486
Minimum	0.385
Maximum	0.871
Sum	280.6757
Count	506

DISTANCE	
Mean	9.549407
Standard Error	0.387085
Median	5
Mode	24
Standard Deviation	8.707259
Sample Variance	75.81637
Kurtosis	-0.86723
Skewness	1.004815
Range	23
Minimum	1
Maximum	24
Sum	4832
Count	506

TAX	
Mean	408.2372
Standard Error	7.492389
Median	330
Mode	666
Standard Deviation	168.5371
Sample Variance	28404.76
Kurtosis	-1.14241
Skewness	0.669956
Range	524
Minimum	187
Maximum	711
Sum	206568
Count	506

PTRATIO	
Mean	18.45553
Standard Error	0.096244
Median	19.05
Mode	20.2
Standard Deviation	2.164946
Sample Variance	4.686989
Kurtosis	-0.28509
Skewness	-0.80232
Range	9.4
Minimum	12.6
Maximum	22
Sum	9338.5
Count	506

AVG_ROOM	
Mean	6.284634
Standard Error	0.031235
Median	6.2085
Mode	5.713
Standard Deviation	0.702617
Sample Variance	0.493671
Kurtosis	1.8915
Skewness	0.403612
Range	5.219
Minimum	3.561
Maximum	8.78
Sum	3180.025
Count	506

LSTAT	
Mean	12.65306
Standard Error	0.317459
Median	11.36
Mode	8.05
Standard Deviation	7.141062
Sample Variance	50.99476
Kurtosis	0.49324
Skewness	0.90646
Range	36.24
Minimum	1.73
Maximum	37.97
Sum	6402.45
Count	506

AVG_PRICE	
Mean	22.53281
Standard Error	0.408861
Median	21.2
Mode	50
Standard Deviation	9.197104
Sample Variance	84.58672
Kurtosis	1.495197
Skewness	1.108098
Range	45
Minimum	5
Maximum	50
Sum	11401.6
Count	506

Table 4. Summary Statistics

Insights –

The descriptive statistics of the dataset provide the following observations:

i. To begin with, the Distance variable shows that, with a maximum distance of 24 and a mode of 24, the majority of dwellings are situated distant from the highway.

ii. The skewness of the variables indicates that the dataset is significantly skewed.

iii. There are 506 items in the entire dataset.

iv. Examining the age variable, we see that the average age of the apartments is 100, with 100 serving as the maximum and mean ages.

v. versus. The average tax paid is 408.2, while the tax range is 524.

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

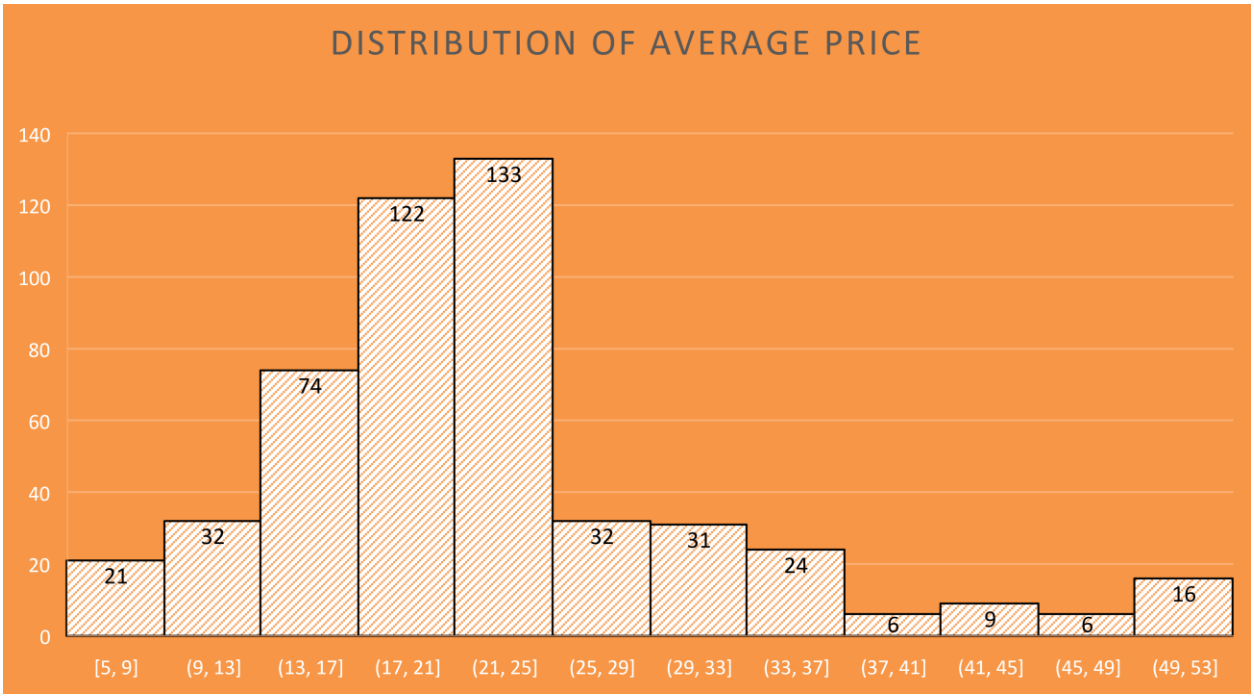

Figure 1. Histogram

Insights -

According to the Histogram above,

i. We have the fewest number of dwellings between the ranges of 45 to 49 and 37 to 41.

ii. Based on our findings, the most of the residences are located in the 17–25 age range.

Q.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	790.792								
INDUS	-0.110215175	124.268	46.97143							
NOX	0.000625308	2.381	0.60587	0.013401						
DISTANCE	-0.229860488	111.550	35.47971	0.61571	75.66653					
TAX	-8.229322439	2397.942	831.71333	13.0205	1333.11674	28348.624				
PTRATIO	0.068168906	15.905	5.68085	0.047304	8.74340	167.821	4.6777263			
AVG_ROOM	0.056117778	-4.743	-1.88423	-0.02455	-1.28128	-34.515	-0.5396945	0.4926952		
LSTAT	-0.882680362	120.838	29.52181	0.48798	30.32539	653.421	5.7713002	-3.0736550	50.893979	
AVG_PRICE	1.16201224	-97.396	-30.46050	-0.45451	-30.50083	-724.820	-10.0906756	4.4845656	-48.351792	84.419556

Table 5. Covariance matrix

Insights –

The following assumptions may be drawn from the above-mentioned matrix:

i. We can see that the tax variable has strong correlation values with all other characteristics, except the crime rate. Therefore, a sizable amount of the variability seen in other variables may be attributed to taxation.

ii. It is evident that several of the characteristics have high covariance values, which point to a significant relationship between their variability and that of the other features.

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

	CRIME_RATE	*AGE*	*INDUS*	*NOX*	*DISTANCE*	*TAX*	*PTRATIO*	*AVG_ROOM*	*LSTAT*	*AVG_PRICE*
CRIME_RATE	1									
AGE	0.006859	1								
INDUS	-0.005511	0.644779	1							
NOX	0.001851	0.731470	0.763651	1						
DISTANCE	-0.009055	0.456022	0.595129	0.611441	1					
TAX	-0.016749	0.506456	0.720760	0.668023	0.910228	1				
PTRATIO	0.010801	0.261515	0.383248	0.188933	0.464741	0.460853	1			
AVG_ROOM	0.027396	-0.240265	-0.391676	-0.302188	-0.209847	-0.292048	-0.355501	1		
LSTAT	-0.042398	0.602339	0.603800	0.590879	0.488676	0.543993	0.374044	-0.613808	1	
AVG_PRICE	0.043338	-0.376955	-0.483725	-0.427321	-0.381626	-0.468536	-0.507787	0.695360	-0.737663	1

Table 6. Correlation matrix

a. Which are the top 3 positively correlated pairs?

From the above correlation matrix, we can analyze the top 3 positively correlated pairs:

i. Distance – Tax
ii. NOX – Indus
iii. NOX – Age

b. Which are the top 3 negatively correlated pairs?

From the above correlation matrix, we can analyze the top 3 negatively correlated pairs:

i. Avg_Price – LSTAT
ii. Avg_Room – LSTAT
iii. Avg_Price – PTRATIO

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

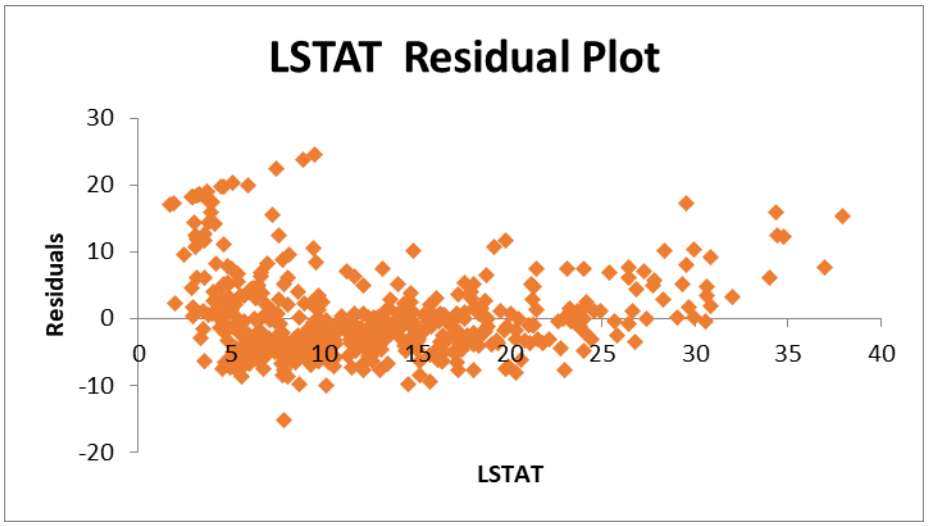

Figure 2. LSTAT Residual Plot

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

i. This model indicates that 54% of the variation in the average price can be explained by the LSTAT.

ii. The LSTAT coefficient for the model is -0.95004935. This means that for every 0.9 rise in LSTAT, the average price of a property decreases by 0.9 times.

iii. 34.55384088 is the LSTAT intercept for the model.

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

A key variable for the avg_price in this model is LSTAT. This model produced a p-value that is considerably less than 0.05, 5.0811E-88.

This indicates that LSTAT is a meaningful variable for this model.

Q.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 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?

For this model, we discovered the regression equation shown below:

y equals 5.09 X0 - 0.642 X1 + -1.358.

where y = Average Price,

X0 = Average Room,

and X1 = LSTAT

Using the model, the average cost of a new home may be found with the calculation Y= -1.358 + 5.09(7) -0.642(20) = 21.44.

The new home will thus cost $21440.

We may argue that the business 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 and explain.

Indeed, this model performs better than the previous one.

With a matching R square value of 0.638561606, the linear equation that we generated from this model is y= -1.35 +5.09a -0.64b,

where a=Avg_room

and b=LSTAT.

This suggests that 63% of the variation in average price can be explained by the interaction of Avg_room and LSTAT, and the multiple R-values of 0.79 suggest a high degree of correlation. On the other hand, under the previous model, LSTAT alone explains 54% of the average price variation.

Q.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.

	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

Table 7. P-value of all variables

We may infer from this that the average price of a property is not significantly influenced by the crime rate because the p-value is greater than 0.5.

69% of the variation in the average house price can be explained when all the components are included.

The negative coefficients for LSTAT, PTRATIO, NOX, and TAX show that an increase in these factors will cause the price of the property to decrease and vice versa.

Q.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	29.42847349	1.84597E-09
AGE	0.03293496	0.012162875
INDUS	0.130710007	0.038761669
NOX	-10.27270508	0.008545718
DISTANCE	0.261506423	0.000132887
TAX	-0.014452345	0.000236072
PTRATIO	-1.071702473	7.08251E-15
AVG_ROOM	4.125468959	3.68969E-19
LSTAT	-0.605159282	5.41844E-27

Table 8. P-value of significant variables

This leads to the conclusion that each feature has a significant impact on the average house 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?

Regression stats from the previous model

Regression Statistics	
Multiple R	0.832978824
R Square	0.69385372

Table 9. Regression stat of the previous model

Regression stats for this model

Regression Statistics	
Multiple R	0.832835773
R Square	0.693615426

Table 10. Regression stat of this model

By comparing the Multiple R and R square values, it is possible to ascertain whether both models operate as intended.

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

Table 11. Coefficient of variables

According to this model, an increase in NOX in the neighborhood will result in a ten-times decrease in the average home price.

d). Write the regression equation from this model.

Y=0.03293496 X0 + 0.130710007 X1 -10.27270508 X2 +0.261506423 X3 -

0.014452345 X4 -1.071702473 X5 + 4.125468959 X6 -0.605159282 X7

+29.42847349

Where Y = AVG_Price

X0 = Age

X1 = Indus

X2 = NOX

X3 = Distance

X4 = TAX

X5 = PTRATIO

X6 = Avg_Room

X7 = LSTAT

Conclusion –

The average property price in the neighborhood will drop 10 times in response to an increase in NOX.

The average price of the residence will decrease if the rates of a few qualities—like NOX, PRATIO, TAX, and LSTAT—are raised. These characteristics have negative coefficients. The business ought to prioritize these elements.

When it comes to predicting average price using the relevant factors, the final regression model is almost precise.

THE END.