
TERRO'S REAL ESTATE AGENCY
NAME- SAYANTAN DEY
GLCA-DA OFFLINE
SEPT 23
DATE: 05/11/2023

Table of contents

Contents

List of Figures –

List of Tables –

Executive Summary –

The Terro Real Estate Agency handles a variety of properties and land kinds. The attributes in the dataset are categorized according to their respective positions. The price of the property is determined by considering its many aspects and traits. We will examine the characteristics of the property and how they affect the asking price in this problem statement.

Introduction –

The purpose of this whole exercise is to explore the dataset. And do the exploratory data analysis. Explore the dataset using different analysis tools and other parameters. The data consists of 506 different properties with 10 unique characteristics. Analysis of the different attributes of the property can help in analyzing 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

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

Table 1. Dataset sample

The dataset has 5 different properties with 10 different variables/ characteristics. Based on the characteristics price of the property is determined.

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 2. Exploratory Dataset

Check for the missing values in the dataset –

There are no missing values in the dataset.

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

CRIME_RATE	*STAT*	*AGE*	*STAT*	*INDUS*	*STAT*	*NOX*	*STAT*	*DISTANCE*	*STAT*
Mean	4.871976	Mean	68.5749	Mean	11.13678	Mean	0.554695	Mean	9.549407
Standard Error	0.12986	Standard Error	1.25137	Standard Error	0.30498	Standard Error	0.005151	Standard Error	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 Deviation	28.14886	Standard Deviation	6.860353	Standard Deviation	0.115878	Standard Deviation	8.707259
Sample Variance	8.533012	Sample Variance	792.3584	Sample Variance	47.06444	Sample Variance	0.013428	Sample Variance	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

TAX	*STAT*	*PTRATIO*	*STAT*	*AVG_ROOM*	*STAT*	*LSTAT*	*STAT*	*AVG_PRICE*	*STAT*
Mean	408.2372	Mean	18.45553	Mean	6.284634	Mean	12.65306	Mean	22.53281
Standard Error	7.492389	Standard Error	0.096244	Standard Error	0.031235	Standard Error	0.317459	Standard Error	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 Deviation	168.5371	Standard Deviation	2.164946	Standard Deviation	0.702617	Standard Deviation	7.141062	Standard Deviation	9.197104
Sample Variance	28404.76	Sample Variance	4.686989	Sample Variance	0.493671	Sample Variance	50.99476	Sample Variance	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

Table 3. Descriptive Statistics

Observation:

A few observations may be obtained from the provided dataset's descriptive statistics:

To start, if we look at the Distance variable, we can see that most homes are located far from the highway, with a maximum distance of 24 and a mode of 24.

ii. The dataset has 506 entries in total.

iii. The tax range is 524, with an average tax paid of 408.2.

iv. The dataset is substantially skewed based on the skewness of the variables.

v. When looking at the age variable, we find that the majority of the residences have an age of 100, with 100 serving as both the maximum and mean.

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

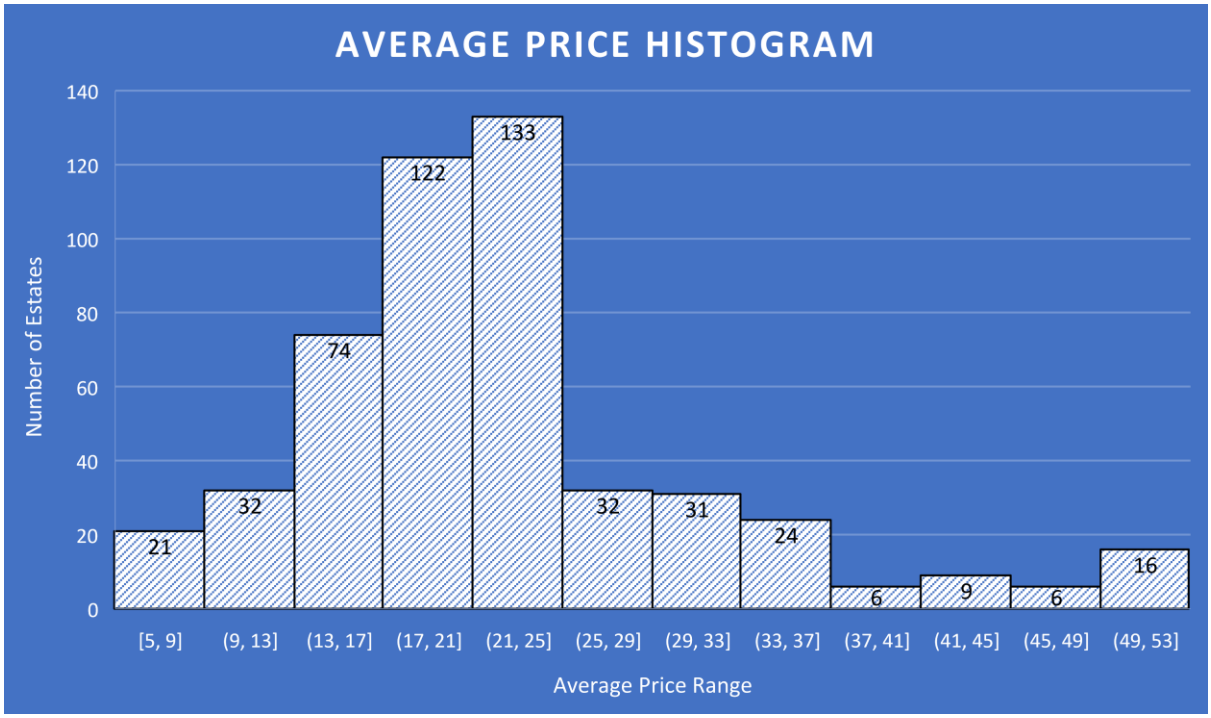

Figure 1. Histogram

Observation:

Based on the Histogram above,

i. we can conclude that the majority of the houses fall between 17 and 25.
ii. Out of the ranges of 37 to 41 and 45 to 49, we have the fewest number of residences.

Q.3 **Compute the covariance matrix. Share your observations.**

	CRIME_RATE	*AGE*	*INDUS*	*NOX*	*DISTANCE*	*TAX*	*PTRATIO*	*AVG_ROOM*	*LSTAT*	*AVG_PRICE*
CRIME_RATE	8.51615									
AGE	0.56292	790.79247								
INDUS	-0.11022	124.26783	46.97143							
NOX	0.00063	2.38121	0.60587	0.01340						
DISTANCE	-0.22986	111.54996	35.47971	0.61571	75.66653					
TAX	-8.22932	2397.94	831.71	13.02	1333.12	28348.62				
PTRATIO	0.06817	15.90543	5.68085	0.04730	8.74340	167.82082	4.67773			
AVG_ROOM	0.05612	-4.74254	-1.88423	-0.02455	-1.28128	-34.51510	-0.53969	0.49270		
LSTAT	-0.88268	120.83844	29.52181	0.48798	30.32539	653.42062	5.77130	-3.07365	50.89398	
AVG_PRICE	1.16201	-97.39615	-30.46050	-0.45451	-30.50083	-724.82043	-10.09068	4.48457	-48.3518	84.41956

Table 4. Covariance matrix

Observation:

The assumptions may be obtained from the above matrix as follows -

i. Except crime rate, we can observe that the tax variable has strong covariance values with all other features. Thus, taxes account for a significant portion of the variability observed in other variables.

ii. Some of the features have large covariance values, as can be seen, indicating a strong correlation between them and the variability of the other features.

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

	CRIME_RAT	AGE	INDUS	NOX	DISTANCE	TAX	PTRATIO	VG_ROOM	LSTAT	AVG_PRICE
CRIME_RATE	1									
AGE	0.00686	1								
INDUS	-0.00551	0.64478	1							
NOX	0.00185	0.73147	0.76365	1						
DISTANCE	-0.00906	0.45602	0.59513	0.61144	1					
TAX	-0.01675	0.50646	0.72076	0.66802	0.91023	1				
PTRATIO	0.01080	0.26152	0.38325	0.18893	0.46474	0.46085	1			
AVG_ROOM	0.02740	-0.24026	-0.39168	-0.30219	-0.20985	-0.29205	-0.35550	1		
LSTAT	-0.04240	0.60234	0.60380	0.59088	0.48868	0.54399	0.37404	-0.61381	1	
AVG_PRICE	0.04334	-0.37695	-0.48373	-0.42732	-0.38163	-0.46854	-0.50779	0.69536	-0.73766	1

Table 5. 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.**

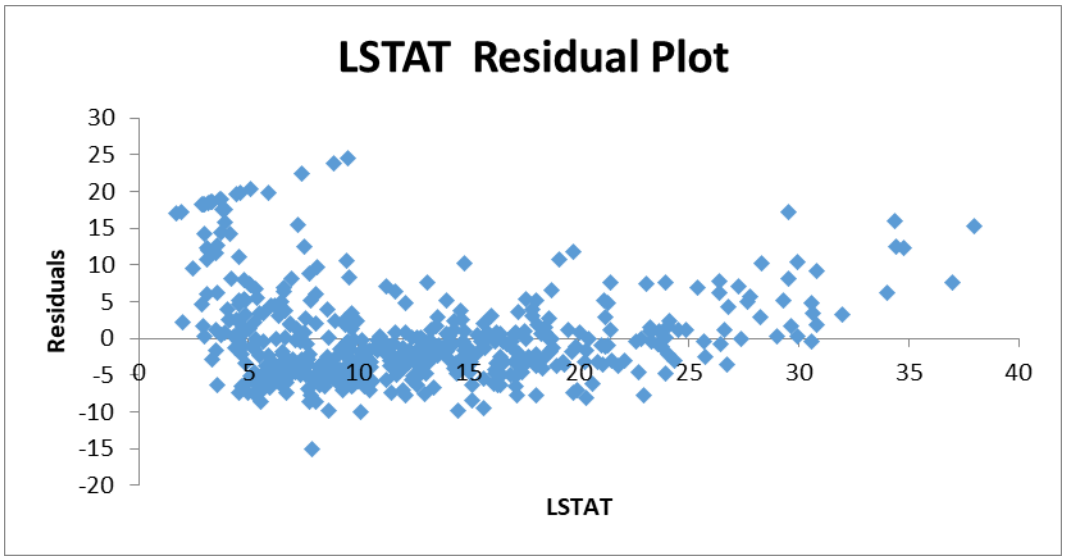

Figure 2. LSTAT Residual Plot

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

i. For the model, the LSTAT coefficient is -0.95004935. According to this, the average price of a property drops by 0.9 times if LSTAT increases by 0.9 times.
ii. The model's LSTAT intercept is 34.55384088.
iii. According to this model, the LSTAT accounts for 54% of the variance in the average price.

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

For the avg_price in this model, LSTAT is an important variable. Since this model yielded a p-value of 5.0811E-88, it is significantly less than 0.05.

This means that, in light of this model, LSTAT is a relevant variable.

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?

We found the following regression equation for this model:

y = -1.358 + 5.09 X0 - 0.642 X1.

where X0 = avg_room,

X1 = LSTAT,

and y=Avg_price

The formula for calculating the average price of a new house using the model is

Y= -1.358 + 5.09(7) -0.642(20)

= 21.44.

Thus, the new house will cost $21440.

We could say that the 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 and explain.

Indeed, this model's performance is superior to that of the preceding model.

The linear equation we derived from this model is,

y= -1.35 +5.09a -0.64b (where a=Avg_room and b=LSTAT), with a corresponding R square value of 0.638561606.

This indicates that the combination of Avg_room and LSTAT accounts for 63% of the fluctuation in average pricing, and the multiple R-value of 0.79 indicates a strong degree of correlation. However, LSTAT by itself accounts for 54% of the average price fluctuation in the prior model.

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 6. P-value of all the variables

Because the p-value is more than 0.5, we may conclude from this that the crime rate is not a significant determinant of the average price of a property.

When all the factors are taken into account, 69% of the variation in the average house price can be explained.

The negative coefficients for NOX, TAX, PTRATIO, and LSTAT indicate that a rise in these characteristics will lead to a fall in the house's price and the reverse.

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 7. P-value of significant variables

This leads to the conclusion that every feature affects the average price of the house in a meaningful way.

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 8. Regression stats of the previous model

Regression stats for this model,

Regression Statistics	
Multiple R	0.832835773
R Square	0.693615426

Table 9. Regression stats of the current model

It can determine that both models function properly by comparing the Multiple R and R square values.

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 10. coefficients of variables

This model predicts that the average property price will drop by ten times if there is more NOX in the area.

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 & recommendation –

i. A few characteristics, such as NOX, PRATIO, TAX, and LSTAT, have negative coefficients, which indicates that raising the rate of those features will lower the average price of the home. The company should focus on these aspects.

ii. All of the research's findings are significant in figuring out the typical cost of a house, except the crime rate. This data leads me to believe that the corporation should focus more on the other factors influencing the average price rather than assuming that a higher or lower crime rate has no impact on the average property price.

THE END!