

# EXPLORATORY DATA ANALYSIS PROJECT

Terro's Real Estate Agency

Deepthi Sree.G

GLCA March 2023

## Table of Contents

|  |    |
|--|----|
| Problem:.....  | 3  |
| Objective: .....   | 3  |
| Question 1 Generate the summary statistics for each variable in the table. (Use Data analysis tool pack). Write down your observation. ....  | 4  |
| Question 2 Plot a histogram of the Avg_Price variable. What do you infer?.....   | 6  |
| Question 3 Compute the covariance matrix. Share your observations. ....  | 7  |
| Question 4 Create a correlation matrix of all the variables (Use Data analysis tool pack). ....  | 8  |
| a) Which are the top 3 positively correlated pairs.....  | 8  |
| b) Which are the top 3 negatively correlated pairs.....  | 8  |
| Question 5 Build an initial regression model with AVG_PRICE as 'y' (Dependent variable) and LSTAT variable as Independent Variable. Generate the residual plot. ....   | 9  |
| a) What do you infer from the Regression Summary output in terms of .....  | 9  |
| variance explained, coefficient value, Intercept, and the Residual plot? .....   | 9  |
| b) Is LSTAT variable significant for the analysis based on your model?.....  | 9  |
| Question 6 Build a new Regression model including LSTAT and AVG_ROOM together as Independent variables and AVG_PRICE as dependent variable. ....   | 10 |
| 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?.....                        | 10 |
| 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.....   | 10 |
| 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 Rsquare, coefficient and Intercept values. Explain the significance of each independent variable with respect to AVG_PRICE. .... | 11 |
| 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: .....  | 12 |
| a) Interpret the output of this model.....   | 12 |
| 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? .....   | 12 |
| 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. ....  | 13 |
| d) Write the regression equation from this model. ....   | 13 |
| Summary:.....  | 14 |
| Thank You!! .....  | 14 |

### Problem:

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:

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

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

| <i>CRIME_RATE</i>  |          | <i>AGE</i>         |          | <i>INDUS</i>       |          |
|--------------------|----------|--------------------|----------|--------------------|----------|
| Mean               | 4.871976 | Mean               | 68.5749  | Mean               | 11.13678 |
| Standard Error     | 0.12986  | Standard Error     | 1.25137  | Standard Error     | 0.30498  |
| Median             | 4.82     | Median             | 77.5     | Median             | 9.69     |
| Mode               | 3.43     | Mode               | 100      | Mode               | 18.1     |
| Standard Deviation | 2.921132 | Standard Deviation | 28.14886 | Standard Deviation | 6.860353 |
| Sample Variance    | 8.533012 | Sample Variance    | 792.3584 | Sample Variance    | 47.06444 |
| Kurtosis           | -1.18912 | Kurtosis           | -0.96772 | Kurtosis           | -1.23354 |
| Skewness           | 0.021728 | Skewness           | -0.59896 | Skewness           | 0.295022 |
| Range              | 9.95     | Range              | 97.1     | Range              | 27.28    |
| Minimum            | 0.04     | Minimum            | 2.9      | Minimum            | 0.46     |
| Maximum            | 9.99     | Maximum            | 100      | Maximum            | 27.74    |
| Sum                | 2465.22  | Sum                | 34698.9  | Sum                | 5635.21  |
| Count              | 506      | Count              | 506      | Count              | 506      |

| <i>NOX</i>         |          | <i>DISTANCE</i>    |          | <i>TAX</i>         |          |
|--------------------|----------|--------------------|----------|--------------------|----------|
| Mean               | 0.554695 | Mean               | 9.549407 | Mean               | 408.2372 |
| Standard Error     | 0.005151 | Standard Error     | 0.387085 | Standard Error     | 7.492389 |
| Median             | 0.538    | Median             | 5        | Median             | 330      |
| Mode               | 0.538    | Mode               | 24       | Mode               | 666      |
| Standard Deviation | 0.115878 | Standard Deviation | 8.707259 | Standard Deviation | 168.5371 |
| Sample Variance    | 0.013428 | Sample Variance    | 75.81637 | Sample Variance    | 28404.76 |
| Kurtosis           | -0.06467 | Kurtosis           | -0.86723 | Kurtosis           | -1.14241 |
| Skewness           | 0.729308 | Skewness           | 1.004815 | Skewness           | 0.669956 |
| Range              | 0.486    | Range              | 23       | Range              | 524      |
| Minimum            | 0.385    | Minimum            | 1        | Minimum            | 187      |
| Maximum            | 0.871    | Maximum            | 24       | Maximum            | 711      |
| Sum                | 280.6757 | Sum                | 4832     | Sum                | 206568   |
| Count              | 506      | Count              | 506      | Count              | 506      |

| <i>PTRATIO</i>     |          | <i>AVG_ROOM</i>    |          | <i>LSTAT</i>       |          | <i>AVG_PRICE</i>   |          |
|--------------------|----------|--------------------|----------|--------------------|----------|--------------------|----------|
| Mean               | 18.45553 | Mean               | 6.284634 | Mean               | 12.65306 | Mean               | 22.53281 |
| Standard Error     | 0.096244 | Standard Error     | 0.031235 | Standard Error     | 0.317459 | Standard Error     | 0.408861 |
| Median             | 19.05    | Median             | 6.2085   | Median             | 11.36    | Median             | 21.2     |
| Mode               | 20.2     | Mode               | 5.713    | Mode               | 8.05     | Mode               | 50       |
| Standard Deviation | 2.164946 | Standard Deviation | 0.702617 | Standard Deviation | 7.141062 | Standard Deviation | 9.197104 |
| Sample Variance    | 4.686989 | Sample Variance    | 0.493671 | Sample Variance    | 50.99476 | Sample Variance    | 84.58672 |
| Kurtosis           | -0.28509 | Kurtosis           | 1.8915   | Kurtosis           | 0.49324  | Kurtosis           | 1.495197 |
| Skewness           | -0.80232 | Skewness           | 0.403612 | Skewness           | 0.90646  | Skewness           | 1.108098 |
| Range              | 9.4      | Range              | 5.219    | Range              | 36.24    | Range              | 45       |
| Minimum            | 12.6     | Minimum            | 3.561    | Minimum            | 1.73     | Minimum            | 5        |
| Maximum            | 22       | Maximum            | 8.78     | Maximum            | 37.97    | Maximum            | 50       |
| Sum                | 9338.5   | Sum                | 3180.025 | Sum                | 6402.45  | Sum                | 11401.6  |
| Count              | 506      | Count              | 506      | Count              | 506      | Count              | 506      |

From descriptive statistics of the given dataset we can get few observations as:

The number of records given in the dataset are 506.

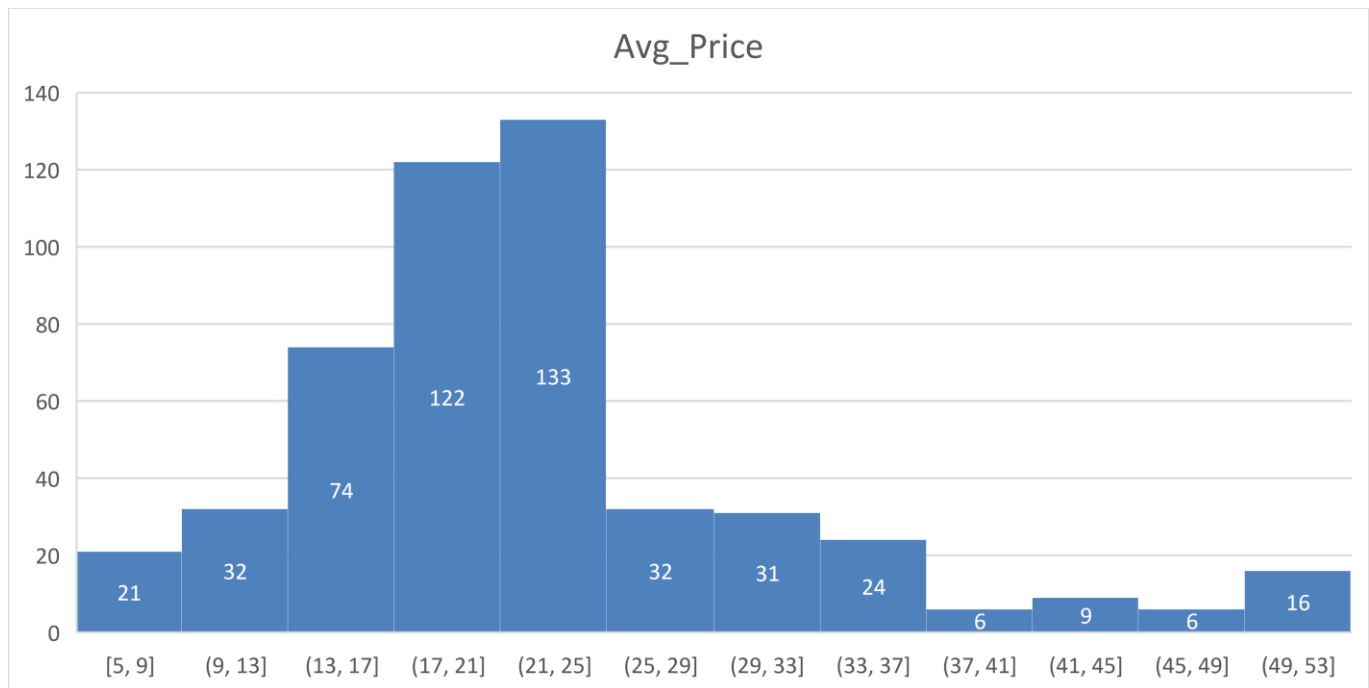
Firstly if we consider Distance variable we can analyse that maximum distance is 24 and has mode as 24. Which says that most of the houses are away from Highway.

The average tax paid is 408.2 and tax range is 524.

From the skewness of variables we can say that dataset is highly skewed.

And if we consider age variable the maximum age is 100 and mode is also 100 which says that most of the houses has age of 100.

Question 2 Plot a histogram of the Avg\_Price variable. What do you infer?



From above Histogram,

We can summarise that most of the houses are from range \$21000 to \$25000.

We have least count of houses from range \$37000 to \$41000 and \$45000 to \$49000.

### Question 3 Compute the covariance matrix. Share your observations.

|            | <i>CRIME_RATE</i> | <i>AGE</i> | <i>INDUS</i> | <i>NOX</i> | <i>DISTANCE</i> | <i>TAX</i> | <i>PTRATIO</i> | <i>AVG_ROOM</i> | <i>LSTAT</i> | <i>AVG_PRICE</i> |
|------------|-------------------|------------|--------------|------------|-----------------|------------|----------------|-----------------|--------------|------------------|
| CRIME_RATE | 8.516147873       |            |              |            |                 |            |                |                 |              |                  |
| AGE        | 0.562915215       | 790.79     |              |            |                 |            |                |                 |              |                  |
| INDUS      | -0.11021518       | 124.27     | 46.9714      |            |                 |            |                |                 |              |                  |
| NOX        | 0.000625308       | 2.3812     | 0.60587      | 0.0134     |                 |            |                |                 |              |                  |
| DISTANCE   | -0.22986049       | 111.55     | 35.4797      | 0.6157     | 75.66653        |            |                |                 |              |                  |
| TAX        | -8.22932244       | 2397.9     | 831.713      | 13.021     | 1333.117        | 28349      |                |                 |              |                  |
| PTRATIO    | 0.068168906       | 15.905     | 5.68085      | 0.0473     | 8.743402        | 167.8      | 4.677726       |                 |              |                  |
|            |                   |            |              | -          |                 |            |                |                 |              |                  |
| AVG_ROOM   | 0.056117778       | -4.743     | -1.8842      | 0.0246     | -1.28128        | -34.5      | -0.53969       | 0.492695216     |              |                  |
| LSTAT      | -0.88268036       | 120.84     | 29.5218      | 0.488      | 30.32539        | 653.4      | 5.7713         | -3.07365497     | 50.894       |                  |
|            |                   |            |              | -          |                 |            |                |                 | -            |                  |
| AVG_PRICE  | 1.16201224        | -97.4      | -30.461      | 0.4545     | -30.5008        | -725       | -10.0907       | 4.484565552     | 48.352       | 84.41955616      |

From above matrix we can get assumptions as :

As we can see there is high covariance value for some of the features which tells that they are highly correlated and explains the variability of the other features.

We can see that tax variable has high covariance values with each other feature except crime rate. That means tax explains a very good variability with other features.

## 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 | 0.6448 | 1      |       |          |       |          |             |       |           |
| NOX        | 0.001850982 | 0.7315 | 0.7637 | 1     |          |       |          |             |       |           |
| DISTANCE   | 0.009055049 | 0.456  | 0.5951 | 0.611 | 1        |       |          |             |       |           |
| TAX        | 0.016748522 | 0.5065 | 0.7208 | 0.668 | 0.910228 | 1     |          |             |       |           |
| PTRATIO    | 0.010800586 | 0.2615 | 0.3832 | 0.189 | 0.464741 | 0.461 | 1        |             |       |           |
| AVG_ROOM   | 0.02739616  | 0.2403 | -0.392 | -0.3  | -0.20985 | 0.292 | -0.3555  | 1           |       |           |
| LSTAT      | 0.042398321 | 0.6023 | 0.6038 | 0.591 | 0.488676 | 0.544 | 0.374044 | -0.61380827 | 1     |           |
| AVG_PRICE  | 0.043337871 | -0.377 | -0.484 | -0.43 | -0.38163 | 0.469 | -0.50779 | 0.695359947 | 0.738 | 1         |

a) Which are the top 3 positively correlated pairs.

From above correlation matrix we can analyse the top 3 positively correlated pairs as

1.Distance – Tax

2.NOX – Age

3.NOX – Indus

b) Which are the top 3 negatively correlated pairs.

From above correlation matrix we can analyse the top 3 negatively correlated pairs as

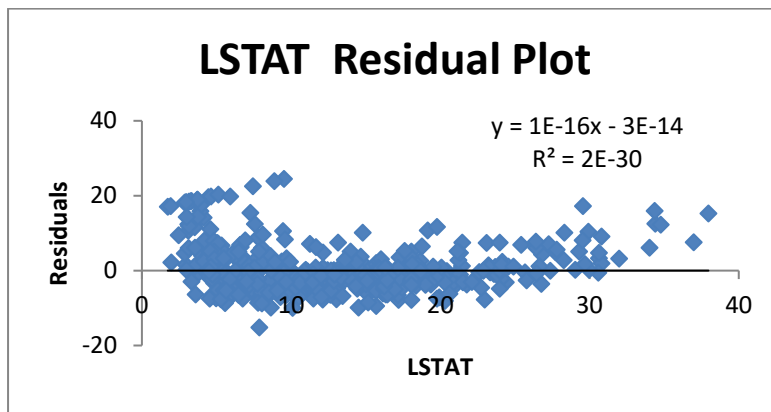
1.LSTAT – Avg\_Room

2.Avg\_Price – PTRATIO

3.Avg\_Price – LSTAT



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?

From this model 54% of the variation in the average price is explained by the LSTAT.

The coefficient of LSTAT for the model is -0.950049354. This says that if LSTAT increases by 0.9 times then average price of house decreases 0.9 times.

Intercept of LSTAT for the model is 34.55384088.

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

Yes, LSTAT is significant variable for the avg\_price from this model.

As the p-value(5.08E-88) we obtained from this model is away less than 0.05.

By this we can say that LSTAT is a significant variable according to this model.

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?

Regression Equation we obtained for this model is :

$$y = -1.358 + 5.09 X_0 - 0.642 X_1$$

Where  $y = \text{Avg\_price}$

$$X_0 = \text{avg\_room}$$

$$X_1 = \text{LSTAT}$$

As per the model, avg\_price for new house can be calculated as

$$Y = -1.358 + 5.09(7) - 0.642(20) = 21.44$$

So ,the price for the new house is \$21440 .

we can say that 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.

Yes, the performance of this model performs well compared to previous model.

From this model the linear equation we obtained is

$$y = -1.35 + 5.09a - 0.64b \quad (\text{Where } a = \text{Avg\_room} \\ b = \text{LSTAT})$$

And Value of R square = 0.638561606 .

With this we can say that 63% of variability for average price is explained by Avg\_room and LSTAT combinely and we obtained multiple R value as 0.79 which says it is highly correlated. But in previous model LSTAT alone describes 54% of variability for average price.

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 Rsquare, coefficient and Intercept values. Explain the significance of each independent variable with respect to AVG\_PRICE.

From the model we can obtain coefficients and p values as below:

|            | <i>Coefficients</i> | <i>P-value</i> |
|------------|---------------------|----------------|
| Intercept  | 29.24131526         | 2.54E-09       |
| CRIME_RATE | 0.048725141         | 0.534657       |
| AGE        | 0.032770689         | 0.01267        |
| INDUS      | 0.130551399         | 0.039121       |
| NOX        | -10.3211828         | 0.008294       |
| DISTANCE   | 0.261093575         | 0.000138       |
| TAX        | -0.01440119         | 0.000251       |
| PTRATIO    | -1.074305348        | 6.59E-15       |
| AVG_ROOM   | 4.125409152         | 3.89E-19       |
| LSTAT      | -0.603486589        | 8.91E-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.

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.

|           | <i>Coefficients</i> | <i>P-value</i> |
|-----------|---------------------|----------------|
| Intercept | 29.42847349         | 1.85E-09       |
| AGE       | 0.03293496          | 0.012163       |
| INDUS     | 0.130710007         | 0.038762       |
| NOX       | -10.27270508        | 0.008546       |
| DISTANCE  | 0.261506423         | 0.000133       |
| TAX       | -0.014452345        | 0.000236       |
| PTRATIO   | -1.071702473        | 7.08E-15       |
| AVG_ROOM  | 4.125468959         | 3.69E-19       |
| LSTAT     | -0.605159282        | 5.42E-27       |

From this we can conclude that all the features are significant variables for average price of the house.

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 previous model

| <i>Regression Statistics</i> |             |
|------------------------------|-------------|
| Multiple R                   | 0.832978824 |
| R Square                     | 0.69385372  |

Regression stats for this model.

| <i>Regression Statistics</i> |             |
|------------------------------|-------------|
| Multiple R                   | 0.832835773 |
| R Square                     | 0.693615426 |

By comparing Multiple R and R square values for both the models we can conclude that both models perform well.

- 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.

|           | <i>Coefficients</i> |
|-----------|---------------------|
| 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         |

If NOX is more in the locality, according to this model average price of the house will decrease by 10 times.

- d) Write the regression equation from this model.

$$Y = 0.03293496 X_0 + 0.130710007 X_1 - 10.27270508 X_2 + 0.261506423 X_3 - 0.014452345 X_4 - 1.071702473 X_5 + 4.125468959 X_6 - 0.605159282 X_7 + 29.42847349$$

Where Y = average\_Price

X0 = Age

X1 = Indus

X2 = NOX

X3 = Distance

X4 = TAX

X5 = PTRATIO

X6 = Avg\_room

X7 = LSTAT

### Summary:

From this Analysis ,we can conclude that all the features play a vital role in estimating the average price of the house excluding crime rate.

And few features have negative coefficients which say that increase rate in those features will decrease the average price of the house like NOX, PTRATIO, TAX and LSTAT.

Thank You!!

GREAT LEARNING!!!!