Assignment – Terro’s real estate agency

Real estate data analysis – Exploratory data analysis, Linear Regression



# 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.

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 |

# Objective (Task):

Your job, as an auditor, is to analyze the magnitude of each variable to which it can affect the price of a house in a particular locality.

To do the analysis, you 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. **(5 marks)**

Solution: The average value of owner-occupied houses (AVG\_Price) in the data set is 22.53 ‘000 USD), and the median value is 21.20.

The range of values is between 5 to 50. The data has a bit of skewness. The average AGE is 68.5 and the median age is 77.5 suggesting negative skewness in this variable, it means that there are extreme values towards the lower end of the spectrum. Mean of AVG\_Rooms is 6.28 and median is 6.2 suggesting that this variable could be normally distributed (more analysis would be required to know the exact picture). Most frequent value of AVG\_ROOMS is 5.7.

|  |  |  |
| --- | --- | --- |
| **Column1** | **Kurtosis** | **Skewness** |
| CRIME\_RATE | -1.189122464 | 0.021728079 |
| AGE | -0.967715594 | -0.59896264 |
| INDUS | -1.233539601 | 0.295021568 |
| NOX | -0.064667133 | 0.729307923 |
| DISTANCE | -0.867231994 | 1.004814648 |
| TAX | -1.142407992 | 0.669955942 |
| PTRATIO | -0.285091383 | -0.802324927 |
| AVG\_ROOM | 1.891500366 | 0.403612133 |
| LSTAT | 0.493239517 | 0.906460094 |
| AVG\_PRICE | 1.495196944 | 1.108098408 |

1. Plot a histogram of the Avg\_Price variable. What do you infer? **(5 marks)**

Solution: The histogram shows that the variable, AVG\_PRICE Is positively skewed as there is a tail towards the right, indicating the presence of some very high values in the data. Most of the houses are priced in the 17.88 to 24.49 bracket (prices are in ‘000 USD).

1. Compute the covariance matrix. Share your observations. **(5 marks)**

Solution: CRIM\_RATE and AVG\_PRICE, AVG\_ROOM and AVG\_PRICE are positively related, rest all variables are negatively related with AVG\_PRICE.

1. Create a correlation matrix of all the variables (Use Data analysis tool pack). **(5 marks)**
   1. Which are the top 3 positively correlated pairs and

Solution a) Top 3 positively correlated pairs – TAX and Distance (0.891), NOX and INDUS (0.76) and NOX and AGE (0.73).

* 1. Which are the top 3 negatively correlated pairs.

Solution b) Top 3 negative correlations – LSTAT and AVG\_Price (-0.74), LSTAT and AVG\_ROOM (-0.61) and PTRATIO and AVG\_PRICE (-0.51).

1. Build an initial regression model with AVG\_PRICE as **‘y’** (Dependent variable) and LSTAT variable as Independent Variable. Generate the residual plot. **(8 marks)**
   1. What do you infer from the Regression Summary output in terms of variance explained, coefficient value, Intercept, and the Residual plot?

Solution: The model has a R-squared value of 0.544 which suggests that this explains 54.4% of the variance in the AVG\_PRICE. The intercept is 34.55 which suggests that even if LSTAT is 0, the value of AVG\_PRICE will be positive, i.e. 34.55. Looking t the residual plot, we see more concentration of points towards the lower values of LSTAT, visually it suggests that there might be a pattern here, so we should explore more models we could get a better model.

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

Solution: LSTAT has a significance value very close to 0, but it cannot be absolute 0. Since it is less than the significance level of 0.05, this variable LSTAT is significant and should be retained in our analysis.

1. Build a new Regression model including LSTAT and AVG\_ROOM together as Independent variables and AVG\_PRICE as dependent variable. **(6 marks)**

Solution: Regression Equation

y = B1X1 + B2X2 + c

y = (5.09 \* X1) – (0.64 \* X2) - 1.36

X1 = 7 (AVG\_ROOM) X2 = 20 (LSTAT)

X2 = 20 (LSTAT)

Y = (5.09 \* 7) – (0.64 \* 20) – 1.36

Y = 35.63 – 12.8 – 1.36 = 21.47

The company is quoting a value of 30 against a prediction of 21.47, which suggests that the company is overcharging.

* 1. 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.

Solution b) The R-squared value here is 0.64 as compared to 0.54 of the previous model, this means by adding AVG\_ROOMS to our existing model, we are able to capture additional 10% of the variance in AVG\_Price, because of which this is a better model than the previous one.

1. 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. **(8 marks)**

Solution: This particular model has an R-squared value of 0.6939 against a R-squared value of 0.64 in the previous model (With LSTAT and AVG\_ROOM), this model captures more variance as compared to the previous model.

Also, here the adjusted R-square value is 0.6883 suggesting that the significant variables are contributing to 68.83% of the variance.

The intercept value is 29.24, suggesting that even if all the independent variables were zero, the AVG\_PRICE would be 29.24.

Looking at the p-values, CRIM\_RATE should be dropped as its p-value is more than 0.05. Rest all variables are significant.

1. 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: **(8 marks)**
   1. Interpret the output of this model.

Solution a) This model explains 69.36% of the variance in AVG\_PRICE. The intercept value is 29.42 suggesting that if all independent variables are 0, then the value of the house would be 29.42. All variables are significant here. This model is acceptable as it has a decent R-square and all variables are significant.

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

* 1. 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?

Solution b) Adjusted R-square for this model is 0.6887 vs 0.6883 in the previous model. Although adjusted R-square value is not up drastically, but we have all significant variables here, so we consider these two factors together, then this model is a better model than the previous one.

* 1. 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?

Solution c) NOX and AVG\_Price are negatively related. If the value of NOX increases, then value of AVG\_PRICE falls, more specifically every 1- unit increase in the value of NOX decreased the value of AVG\_PRICE by 10.27.

Coefficients in ascending order:

|  |  |
| --- | --- |
| **Column1** | **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 |

* 1. Write the regression equation from this model.

Solution d) Regression Equation:

Y = 29.4285 + 0.0329 \* X1 + 0.1307 \* X2 + -10.2727 \* X3 + 0.2615 \* X4 – 0.0145 \* X5 – 1.0717 \* X6 + 4.1255 \* X7 – 0.6052 \* X8

**HINT**: *Significant variables are those whose p-values are less than 0.05. If the p-value is greater than*

* 1. *then it is insignificant*

# Learning Outcome (Result):

* + - Implementation of Exploratory Data Analysis helps you to understand the nature of different data-attributes
    - You will understand how to use various statistical/analytical tools in MS Excel like Summary statistics, Histogram, correlation table, Regression analysis (using Data analysis tool pack)