

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)**
- 2) Plot a histogram of the Avg_Price variable. What do you infer? **(5 marks)**
- 3) Compute the covariance matrix. Share your observations. **(5 marks)**
- 4) Create a correlation matrix of all the variables (Use Data analysis tool pack). **(5 marks)**
 - a) Which are the top 3 positively correlated pairs and
 - b) Which are the top 3 negatively correlated pairs.

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

- 5) Build an initial regression model with AVG_PRICE as 'y' (Dependent variable) and LSTAT variable as Independent Variable. Generate the residual plot. **(8 marks)**
 - a) What do you infer from the Regression Summary output in terms of variance explained, coefficient value, Intercept, and the Residual plot?
 - b) Is LSTAT variable significant for the analysis based on your model?
- 6) Build a new Regression model including LSTAT and AVG_ROOM together as Independent variables and AVG_PRICE as dependent variable. **(6 marks)**
 - 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?
 - 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.
- 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. **(8 marks)**
- 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: **(8 marks)**
 - a) Interpret the output of this model.
 - 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?
 - 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?
 - d) Write the regression equation from this model.

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