Data Analytics Using Excel Project

Graded Project – 70 Marks

Problem Statement:

You have been hired at a Terro's Real Estate Agency in the capacity of an Auditor. One of the jobs that the auditors of this agency do is to map all the relevant features for the properties along with the information related to the geography around it. The agency wants to understand the relevance of the parameters that they collect in relation to the value of the house (Avg_Price).

:

Domain: Rea Estate Sales and Marketing

Data Description: You have been given a dataset of 506 houses in Boston. Please refer to the data dictionary below. The data set contains the following fields:

CRIME_RATE: per capita crime rate by town

AGE: proportion of house built prior to 1940 (in % Terms)

INDUS: Proportion of non-retail business acres per town (in % terms)

NOX: Nitric Oxide concentration (Parts per 10 million)

DISTANCE: distance from Highway (in miles)

TAX: full value property tax rate per $10,000

PTRATIO: Pupil-Teacher ratio by town

AVG_ROOM: average number of rooms per house

LSTAT: % lower status of Population

AVG_PRICE: Average value of houses in $1000's

Tasks/ Questions to be answered: Your key job is to analyse the extent and magnitude of each variable relative to the value of the house. For this, you have the following deliverables to execute.

1. Understanding the data. (10 Marks)
   a. Generate the summary statistics for each of the variables and note what do you observe?
   b. Plot the histogram of the Avg_Price Variable and note what do you infer?
   c. Compute the covariance matrix. Share your observations.
   d. Create a correlation matrix of all the variables
   e. State top 3 positively correlated pairs and top 3 negatively correlated pairs.

2. Simple Linear Regression with one variable (10 Marks)
    a. Build an initial regression model with AVG_PRICE as the y or the Dependent variable and LSTAT variable as the Independent Variable.
    b. Generate the residual plot for the Repressor line
    c. Interpret the Regression Summary Output in terms of adjusted R-square, variance explained, coefficient value, Intercept and the Residual plot
    d. Is LSTAT variable significant for the analysis based on your 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)
3. Multilinear Regression with 2 variables (10 Marks)
    a. Build another instance of the Regression model but this time including LSTAT and AVG_ROOM together as Independent variables and AVG_PRICE as the dependent variable.
    b. Interpret the output in terms of adjusted R-square, coefficient and Intercept values, Significance of variables with respect to AVG_price.
    c. Is the performance of this model better than the previous model you built in Question 2? (Hint: Compare in terms of RMSE, R-square and adjusted R-square and explain the equations of Q2,Q3,Q4 & Q5)
4. Multilinear Regression & Finding Significant Variables (15 Marks)
    a. Build a Regression model with all variables. Please note that the AVG_PRICE shall be the Dependent Variable and others will be independent variables.
    b. Interpret the output in terms of adjusted R-square, coefficient and Intercept values, Significance of variables with respect to AVG_price.
    c. Is the performance of this model better than the previous models you built in Question 2 & 3? (Hint: Compare in terms of RMSE, R-square and adjusted R-square and explain the equations of Q2,Q3,Q4 & Q5)
    d. List out the significant variables from the previous question. (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)
5. Best fit Model (10 Marks)
    a. Make another new instance of the Regression model using only the significant variables selected in Question 4
    b. Interpret the output in terms of adjusted R-square, coefficient and Intercept values, Significance of variables with respect to AVG_price
    c. Is the performance of this model better than the previous models you built in Question 2, 3 & 4? (Hint: Compare in terms of RMSE, R-square and adjusted R-square and explain the equations of Q2,Q3,Q4 & Q5)
6. Answer the questions below based on the bestfit model/ final model(2.5 Marks)
    a. Sort the values of the Coefficients in ascending order.
    b. What will happen to the “average price” if the “value of NOX” is **more** in a locality in this town? (Hint: Define the relation of NOX value with Avg_Price using the coefficient matrix and explain the impact of change in NOX to the price)

7. Based on the above best fit Model, write down the Regression equation and answer the following (2.5 Marks)
    a. 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?
    b. Based on the outputs arrived from the above questions, is the value of $30000 USD Quoted by the company for this locality is Overcharging/Undercharging? (Hint: provide detailed inferences from analysis along with data points to explain your conclusion. It is expected to be part of business report along with the workbook)
8. Business Report & detailed Explanation of Inferences for all the Tasks(10 Marks)
    a. Please note the inferences, Observations and Conclusion should be available both in the workbook and business report.
    b. In this, you need to submit all the answers to all the questions in a sequential manner
    c. The Business report should include a detailed explanation of the approach used, insights, inferences, all outputs like graphs, tables, etc. The Level of detail and explanation in business report should be deeper than just bullet points of observations/inferences.
    d. Your report should not be filled with Excel tables. Only use important metrics or values obtained from analysis