# Week 4 Project

## Case Study: Terro's Real Estate Agency

(TOPICS COVERED: Descriptive Statistics, Covariance, Correlations, Simple Linear Regression, Multiple Linear Regression)

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).
You have been given a dataset of 506 houses in Boston. Please refer to the data dictionary below:

**Data Dictionary:**

**CRIME_RATE**: per capita crime rate by town
**INDUSTRY**: the 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**: the 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

Your key job is to analyze the extent and magnitude of each variable relative to the value of the house. For this, you have the following deliverables to execute.

1. The first step to any project is understanding the data. So for this step, generate the summary statistics for each of the variables. What do you observe? (5 marks)
2. Plot the 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 as shown in the Videos and various case studies. State top 3 positively correlated pairs and top 3 negatively correlated pairs. (5 marks)
5. Build an initial regression model with AVG_PRICE as the y or the Dependent variable and LSTAT variable as the Independent Variable. Generate the residual plot too. (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 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.

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. Explain.

7. Now, build a Regression model with all variables. AVG_PRICE shall be the Dependent Variable. Interpret the output in terms of adjusted R-square, coefficient and Intercept values, Significance of variables with respect to AVG_price. Explain. (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. (8 marks)

(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)
Answer the questions below:

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.

**Important:**

**Reflect on what you learnt while working on this project and fill the Reflection Report (Non-Graded) –Reflection Report**