Simple Linear Regression

1. Machine learning models can be classified into following two categories on the basis of learning algorithm:
    - **Supervised learning method:** Past data with labels is available to build the model
        - **Regression:** The output variable is continuous in nature
        - **Classification:** The output variable is categorical in nature
    - **Unsupervised learning method:** Past data with labels in not available
        - **Clustering:** No pre-defined notion of labels is there
2. Past data set is divided into two parts during supervised learning method:
    - **Training data** is used for the model to learn during modelling
    - **Testing data** is used by the trained model for prediction and model evaluation
3. Linear regression models can be classified into two types depending upon the number of independent variables:
    - **Simple linear regression:** When the number of independent variables is 1
    - **Multiple linear regression:** When the number of independent variables is more than 1
4. The equation of the best fit regression line $Y = \beta_0 + \beta_1 X$ can be found by minimising the cost function (RSS in this case, using the Ordinary Least Squares method) which is done using the following two methods:
    - **Differentiation**
    - **Gradient descent method**
5. The strength of a linear regression model is mainly explained by R^2, where $R^2 = 1 - (RSS / TSS)$
    - **RSS:** Residual Sum of Squares
    - **TSS:** Total Sum of Squares

Multiple Linear Regression

1. **Understand the business objective**

Recall that the first step towards solving any problem is to understand the business objective, followed by understanding the given data set.

2. **Data preparation**

The next important step is to prepare the data set by treating the missing values, outliers, etc. and also creating dummy variables to convert categorical variables into numerical variables. In addition, you also need to split the data into training and testing data.

3. **Model development**

The next step is to build the model, where the most important step is to identify the most significant variables and remove the insignificant variables.

Our SME took the backward elimination approach, where first the model was built containing all the variables. Then, we checked for multicollinearity and VIF (variance inflation factor). We removed the variables which had a higher p-value in order of their insignificance. Finally, we built the regression model predicting the house price and used it for the prediction of prices on the test data. We then checked the accuracy of our model on the test data set.