Problem Statement

In our previous module we had imported the House Sale Price data, explored it a bit using descriptive statistics, created some charts and did some simple calculations with the data. From this module onwards, our journey of real machine learning begins where we are going to learn to build different ML models.

As we have seen, the dataset contains the sale price of houses and their characteristics like No. of Bedrooms, No. of Bathrooms, Flat Area, Basement Area, etc.

We would like to build a model that will be able to predict the Sale Price of a house based on its characteristics which can be useful for house owners/real estate agents. When they wish to sell their house now or in near future, the model should help them in predicting the price they should expect based on the house characteristics. The same model can also be used by buyers to ensure they are paying a fair price for the house they are purchasing.

Here the Sale Price is our Target Variable (or dependent variable) and house characteristics are independent variables on which the sales price of a house may depend.

Since we have a target variable for this model, this is an example of supervised learning as we have learnt earlier. And since our target variable is a numeric continuous variable, we will build a linear regression model which we will cover in greater detail subsequently.

So, our problem statement is to build a linear regression model to predict the sale price of a house based on its characteristics.