Introduction/Business Problem

As a part of Data Science Capstone, I am going to implement the linear regression model for predicting the houses price. The linear regression technique is used for estimating linear relationships between various features and a continuous target value. An example scenario, could be one can estimate the selling price of the house based on different parameters like number of bedrooms, number of bathrooms, sq. ft., number of floors, number of years built, etc. if a linear regression model can be built on housing dataset. The linear regression is an example of supervised machine learning algorithm.

The business problems of this capstone project are:

1. How much more price one can sell his/her house with additional bedroom/bathroom?
2. Do houses near hospitals, schools and shopping malls are expensive compared to others?
3. What is the impact of number of years built for the pricing of houses?

Data Collection/Preprocessing

The housing dataset is downloaded from the Kaggle and it contains a number of attributes like:

1. Price
2. Number of Bedrooms
3. Number of Bathrooms
4. Living Room Area
5. Number of Floors
6. Waterfront
7. View
8. Condition of the House
9. Year built
10. Year renovated, etc

The dataset is to be split into training and test samples. The training samples are used to fit the model of housing price prediction using linear regression. As a rule of thumb, 70% of the data will be used to train the model and remaining 30% will be used for testing the model.