**Experiment No. 4**

**Title:** To Implement Linear Regression

**Tools Required: RStudio/Anaconda Navigator**

**Concept :**

Linear regression answers a simple question: Can you measure an exact relationship between one target variables and a set of predictors?

The simplest of probabilistic models is the straight line model:

https://www.guru99.com/images/r_programming/032918_1024_RSimpleMult7.jpg

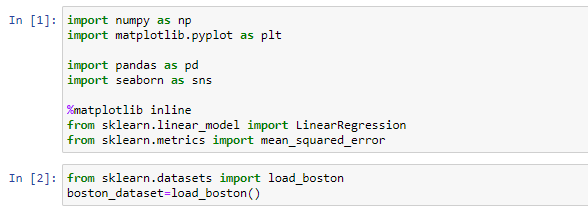
where

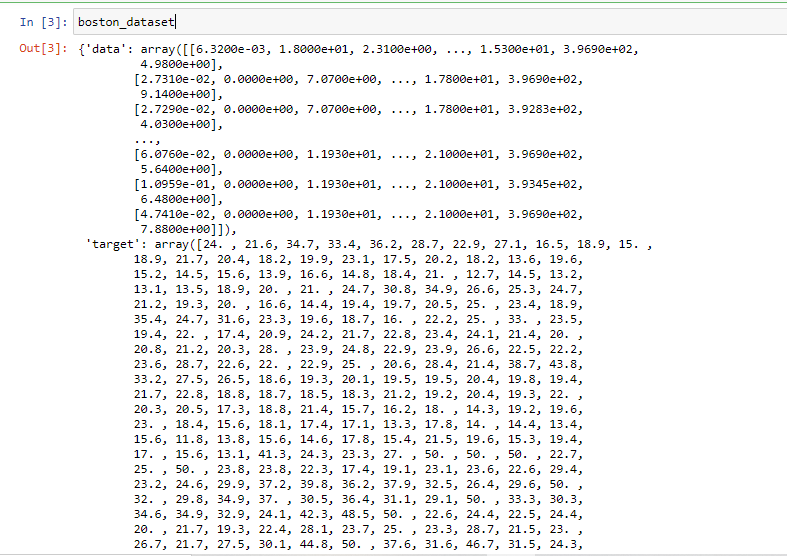
* y = Dependent variable
* x = Independent variable
* [https://www.guru99.com/images/r_programming/032918_1024_RSimpleMult8.jpg](https://www.guru99.com/images/r_programming/032918_1024_RSimpleMult8.jpg)= random error component
* [https://www.guru99.com/images/r_programming/032918_1024_RSimpleMult9.jpg](https://www.guru99.com/images/r_programming/032918_1024_RSimpleMult9.jpg)= intercept
* [https://www.guru99.com/images/r_programming/032918_1024_RSimpleMult10.jpg](https://www.guru99.com/images/r_programming/032918_1024_RSimpleMult10.jpg)= Coefficient of x

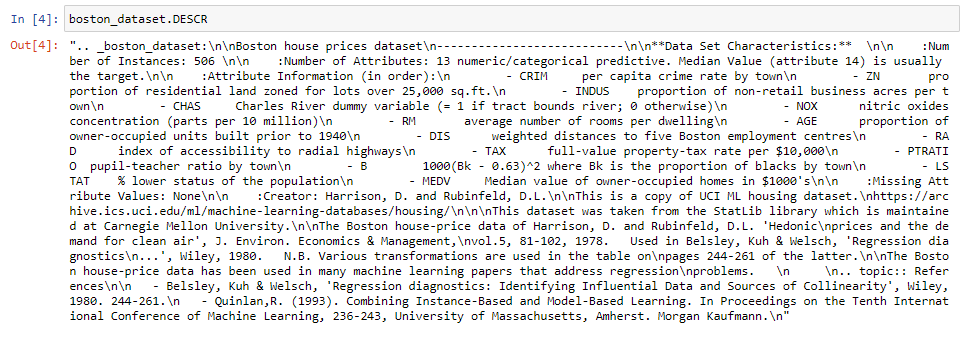
## **Example Problem**

For this analysis, We will take the Housing dataset which contains information about different houses in Boston. We can also access this data from the scikit-learn library. There are 506 samples and 13 feature variables in this dataset. The objective is to predict the value of prices of the house(MEDV) using the given features.

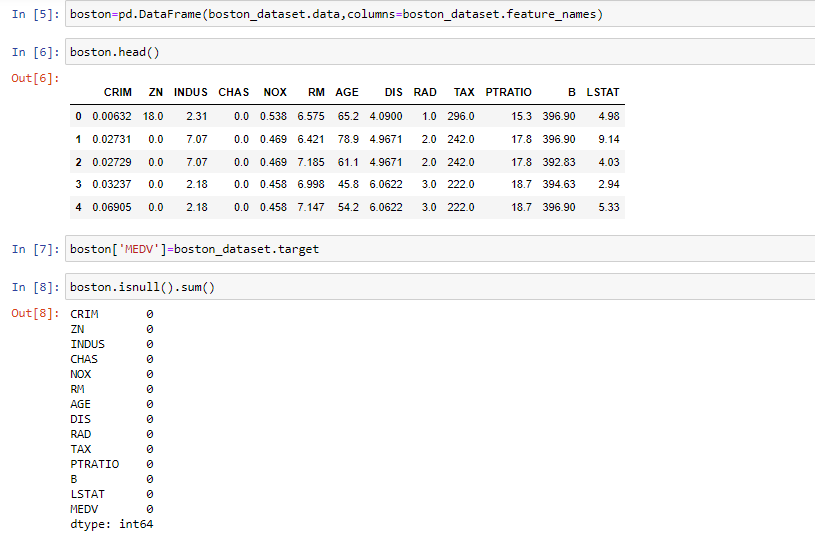
1. **Load the data**

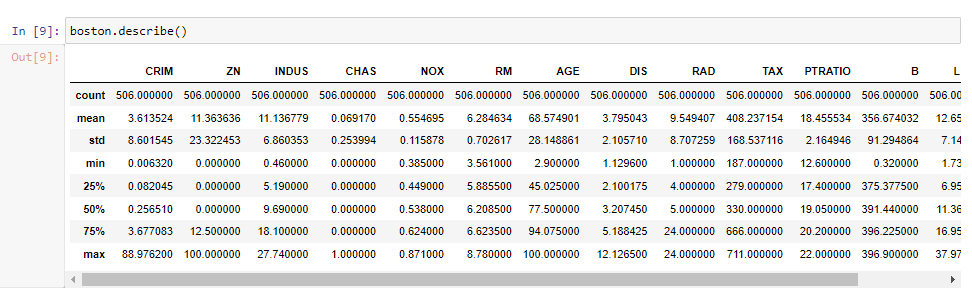






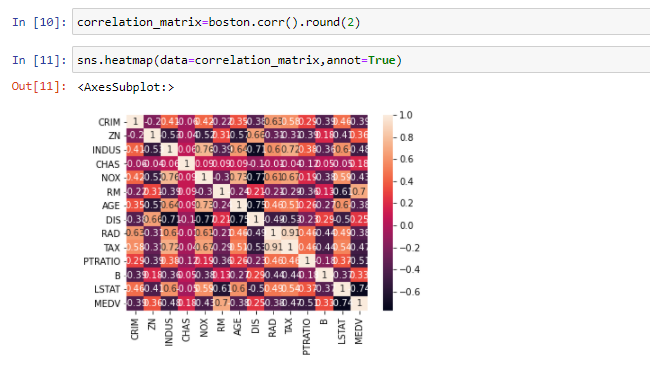
1. **Data preprocessing**:-After loading the data, it’s a good practice to see if there are any missing values in the data

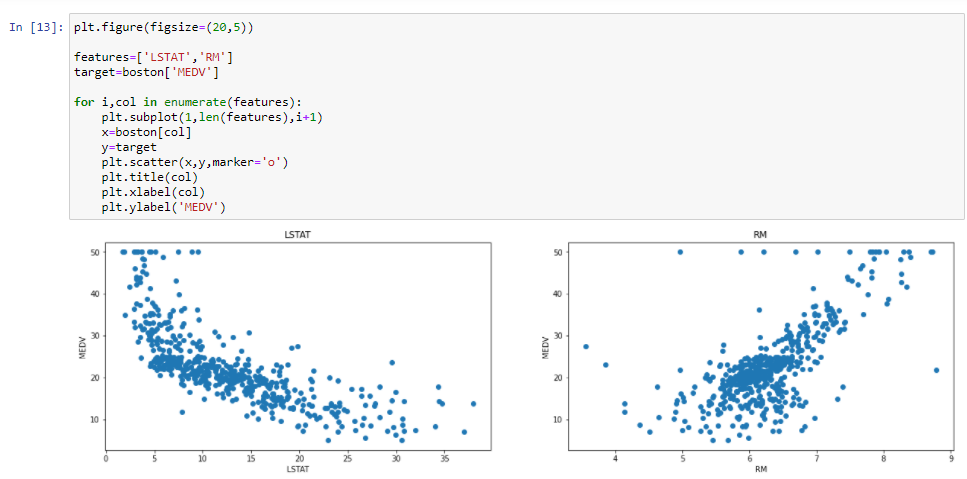




1. **Exploratory Data Analysis:** Exploratory Data Analysis is a very important step before training the model. In this section, we will use some visualizations to understand the relationship of the target variable with other features.
   * **Distribution plot**: Visualize the distribution of target variable
   * **Heatmap**: we create a correlation matrix that measures the linear relationships between the variables. The correlation matrix can be formed by using the corr function from the pandas dataframe library. We will use the heatmap function from the seaborn library to plot the correlation matrix.

Correlation is a statistical measure that suggests the level of linear dependence between two variables, Correlation can take values between -1 to +1. there is a high positive correlation between them if value is between 0 and 1. The opposite is true for an inverse relationship, in which case, the correlation between the variables will be close to -1. A value closer to 0 suggests a weak relationship between the variables.

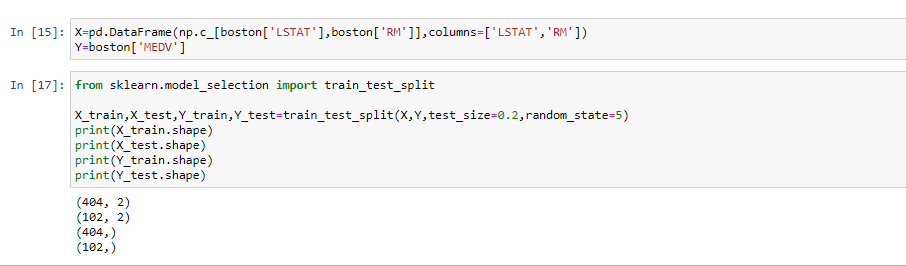


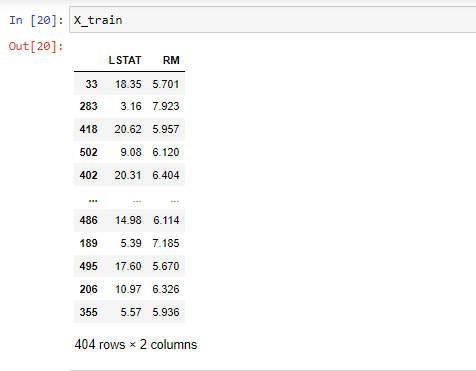


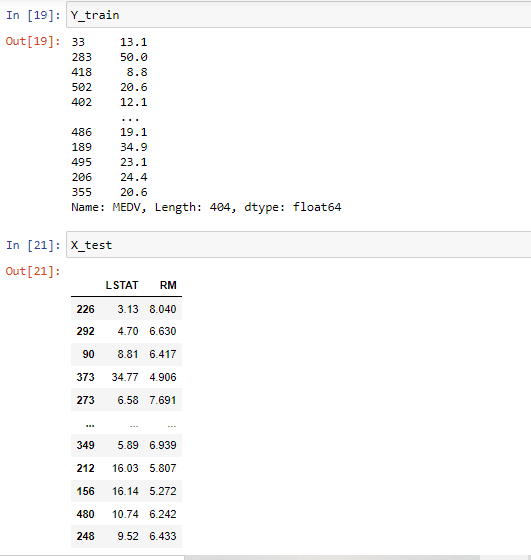
1. **Build and Predicting Linear Models:**

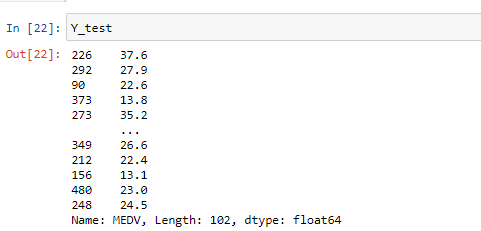
Step 1: Create the training (development) and test (validation) data samples from original data.

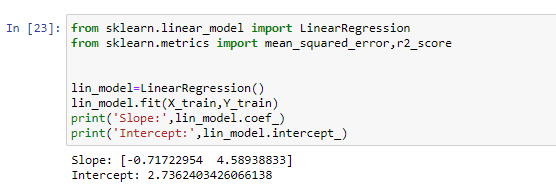
Step 2: Develop the model on the training data and use it to predict the target variable on test data











1. **Model evaluation:-**We will evaluate our model using RMSE and R2-score.

