Data Exploration

1. **Output in C++**

Open file Boston.csv
Reading line 1
Heading: rm,medv
New length 506
Closing file Boston.csv
Number of records: 506

Stats for rm
Sum: 3180.03
Mean: 6.28463
Median: 6.2085
Range: 5.219

Stats for medv
Sum: 11401.6
Mean: 22.5328
Median: 21.2
Range: 45

Covariance = 4.49345

Correlation = 0.696737

Program terminated

2. **Experience using built-in functions in R vs functions in C++**
   - The result between 2 versions is very similar. There is just some slightly different due to the rounding errors. On the stats output, they are the matching up 4 digits after the decimal point. The covariance output also matching up to 4 digits after decimal point. The correlation output is only matching up to 4 digits after decimal point. It is because it involve more on power by 2 then squareroot the intermediate values.

3. **Descriptive statistical of mean, median and range. And their usefulness**
   - Mean: is the average of data set
   - Median is the medium value of the set of number
   - Range: is the difference between max and min (range = max – min)
   - These values are very important in data exploration because they give us the general status of the data set. They can show us how the data was distributed and what is the boundary of the data set. They can also be used for data visualization.

4. **Describe covariance and correlation in statistic. What information they provide? How useful are they in machine learning?**
   - Covariance measure the direction of relationship between two variables. Positive value mean, they move in the same direction. Negative value mean they moving in opposite direction
   - Correlation is actually covariance, scaled to [-1,1]. So it not only show the direction relationship between 2 variables but it also show how closely the value of 2 variables are when they changing
   - By using covariance and correlation, we can confirm if a machine learning model is correct and give us most better prediction on the data. By using the data visualization, plotting the data point and the correlation function, we can see how the ML model fit into the dataset. There are usually 3 cases: underfit, goodfit, and overfit/