

Code Output:

Closing file Boston.csv.
Number of records: 506

Stats for rm
Sum = 3180.03
Mean = 6.28463
Median = 6.2085
Range = 3.561, 8.78

Stats for medv
Sum = 11401.6
Mean = 22.5328
Median = 21.2
Range = 5, 50

Covariance = 4.49345
Correlation = 0.696737
Program terminated.

Describe your experience using built-in functions in R versus coding your own functions in C++

It took a lot longer to code my own functions for data manipulation and statistics in C++ than it would be to use R functions. I appreciate that in R, I can simply use the `cov()` function to compute covariance of a data set instead of making my own function each time.

Describe the descriptive statistical measures mean, median, and range, and how these values might be useful in data exploration prior to machine learning

- Mean is the average of a data set. Calculated as the sum of the data divided by the number of data elements.
- Median is the “middle” of a data set. Calculated by sorting the data from least to greatest then either picking the middle element or by averaging the two “most center” elements.
- Range is the lowest element of the data set and the highest element of the data set as a tuple. This is useful to know what the “range” is for how low a data set can go and how high it can go as well.
- These values can be useful to categorize different numerical data sets and get a broader idea of what the data is – regardless of the size.

Describe the covariance and correlation statistics, and what information they give about two attributes. How might this information be useful in machine learning?

Correlation will tell us whether two data sets are linearly dependent on each other and gives us a numerical value between -1 and 1 to tell us how dependent they are. Covariance will also tell us whether data sets are dependent on each other linearly. It will only tell us in either 0, -1, or 1

C++ Data Exploration

values. These can be useful to know whether there is a relation between data – and what that relationship is. For machine learning, this could be particularly useful in training data.