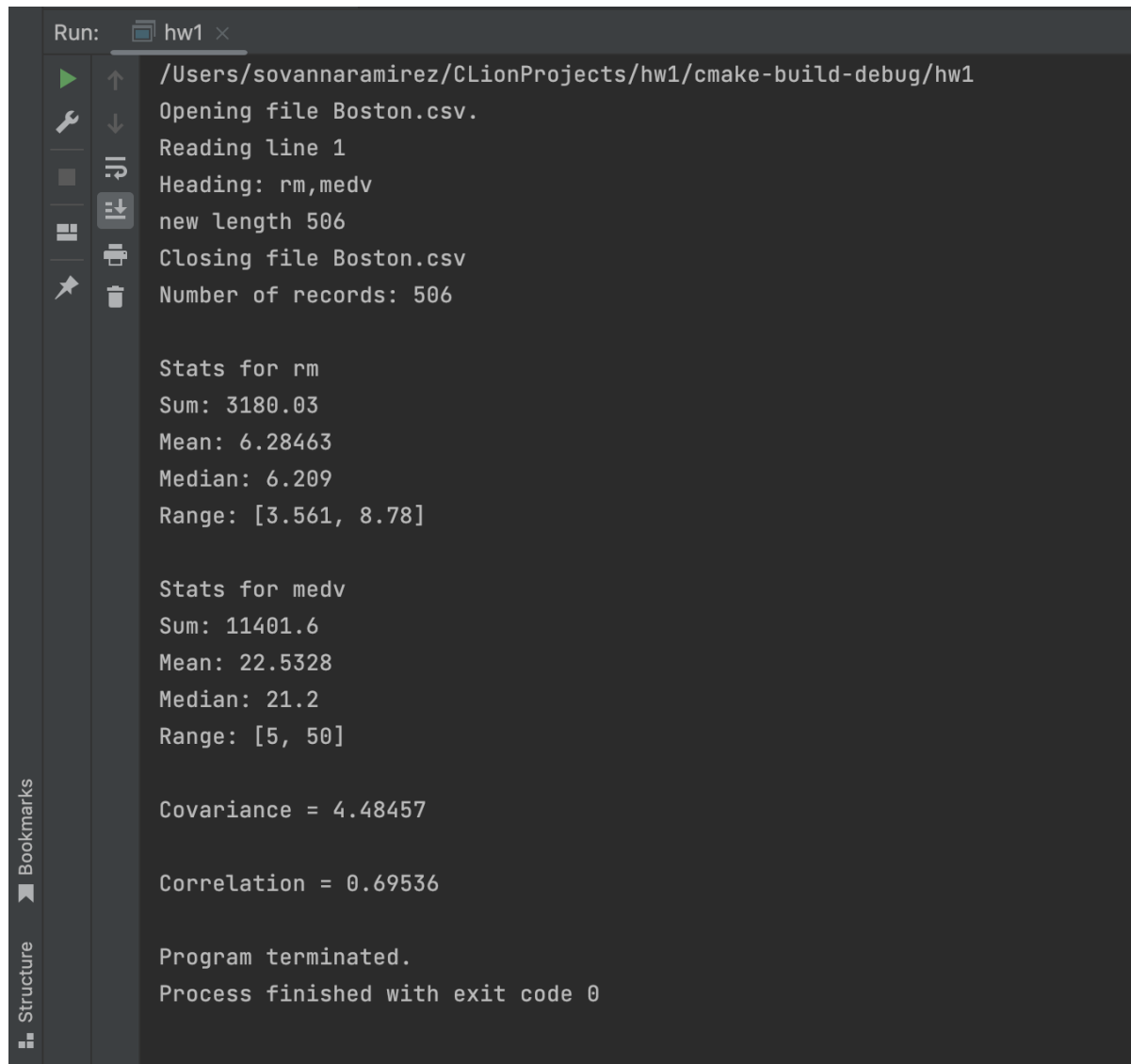


Data Exploration Report

A screenshot of a terminal window showing the output of a C++ program. The window title is 'Run: hw1 x'. The output text is as follows:

```
/Users/sovannaramirez/CLionProjects/hw1/cmake-build-debug/hw1
Opening 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: 6.209
Range: [3.561, 8.78]

Stats for medv
Sum: 11401.6
Mean: 22.5328
Median: 21.2
Range: [5, 50]

Covariance = 4.48457

Correlation = 0.69536

Program terminated.
Process finished with exit code 0
```

The terminal window has a dark background with light gray text. On the left side, there is a vertical toolbar with icons for running, debugging, and other IDE functions. Below the toolbar, the words 'Structure', 'Bookmarks', and 'Run' are visible, indicating the IDE's interface.

b. describing your experience using built-in functions in R versus coding your own functions in C++

In this homework assignment, I wrote functions to calculate the sum, mean, average, and range for two separate vectors labeled `rm` and `medv` in a `Boston.csv` file. Additionally, I wrote functions to find the covariance and correlation statistics between `rm` and `medv`. It has been quite a while since I wrote in C++, so I had to spend some time refreshing my memory on C++. When it came time to write the functions for summation, mean, median, and range the pieces started to come together. I don't really remember doing a lot of work with vectors, so I appreciated that I could go through and understand how to work with them in C++. On the other hand, R's built in functions may be easier and may look simple, but I believe there is still a

lot to learn before I can use the functions on a whim. Moreover, I will most likely just look the functions up to check what parameters are required per function.

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

In statistics, mean, median, and range are used to summarize datasets into concise information which can further be used for analysis purposes.

- mean is also known as the average. We find the mean by taking the summation of all the values in the dataset and dividing them by the total number of values.
- median is found by sorting the data in increasing order and finding the middle value.
- range is the minimum value in a dataset and the maximum value in a dataset.

All of these statistical measures are useful in data exploration because it allows people to understand the data, make better decisions, and/or come to conclusions.

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

Covariance and correlation are both used in statistics to describe a relationship between two attributes. Covariance determines how much two attributes differ from each other. Whereas, correlation determines how much two attributes are related to each other. Both components are important and useful in machine learning because they provide an understanding of the relationship between variables.