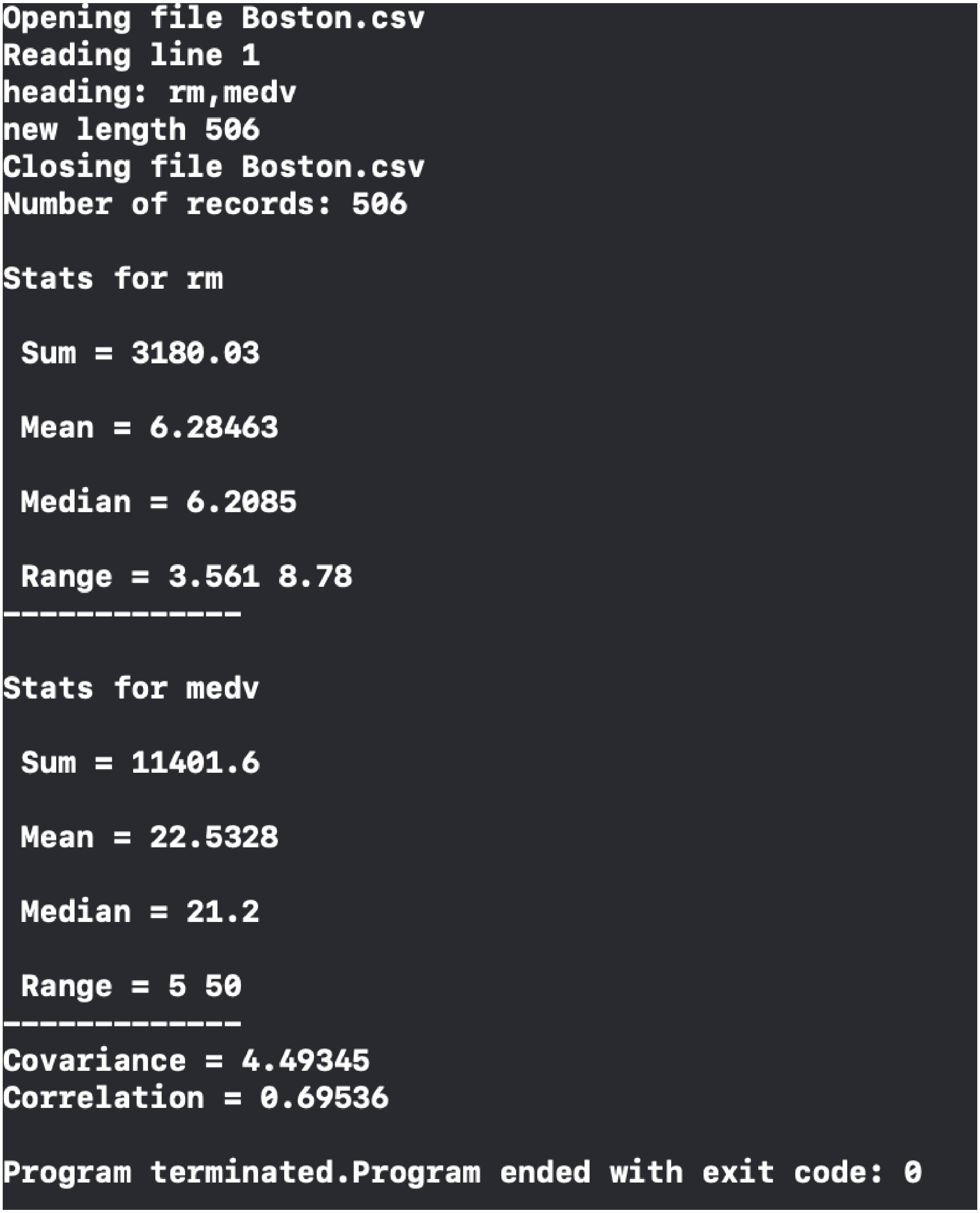
C++ Data Exploration Writeup



```
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 = 6.2085

 Range = 3.561 8.78
-------------

Stats for medv

 Sum = 11401.6

 Mean = 22.5328

 Median = 21.2

 Range = 5 50
-------------
Covariance = 4.49345
Correlation = 0.69536

Program terminated.Program ended with exit code: 0
```

The good thing about using the built-in functions in R is that the implementation is already done for us, and we do not have to worry about the math behind the functions. However, coding our functions in C++ allows us to understand how these functions work behind the scenes. However, in C++, we are adding each value of rm and medv to separate vectors, while in R, it converts this CSV to a data frame which we can manipulate with built-in R functions. The mean represents an average of all the values in a dataset. The median gives the middle value in that dataset as long as the data is sorted. In the case of R, the range of a dataset gives us the minimum and maximum values of a set of data rather than the difference between these values explicitly. These statistical measures are essential in understanding how our data is distributed to identify outliers and trends in the data that might not have been discovered otherwise.

On the other hand, covariance allows us to see how the two attributes are changing concerning each other. If both attributes increase or decrease, this is considered a positive covariance. However, if one attribute increases while the other decreases in value or vice versa, the attributes have a negative covariance. The covariance will be zero if there is no relation to be found. Correlation is related to covariance in that it allows us to understand the strong relationship between these attributes, represented as a value between -1 and 1. If the correlation is precisely -1, the attributes are negatively related. If the correlation is 1, the attributes are positively related. If the correlation is 0, the attributes have no relation to each other. This information might be helpful in machine learning because if we can better understand the relationship between two attributes, it can indicate the best attributes to include in the machine learning model. Rather than including attributes that are unrelated to each other, including attributes that are related to each other might give us a better understanding of what we can expect to predict a target column of the data frame.