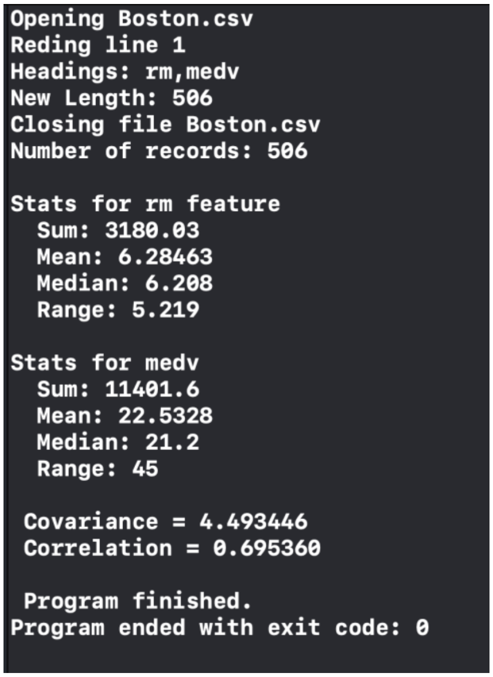
Fernando Colman
Fec190000
9/11/2022

HW 2

Outputs



```
Opening Boston.csv
Reding line 1
Headings: rm,medv
New Length: 506
Closing file Boston.csv
Number of records: 506

Stats for rm feature
  Sum: 3180.03
  Mean: 6.28463
  Median: 6.208
  Range: 5.219

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

 Covariance = 4.493446
 Correlation = 0.695360

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

Experience with R vs. C++

Coding with R, specifically for linear regression, is definitely a lot easier than using C++. In R, you can simply call functions for mean, median, covariance, correlation, and even reading from a csv file is a lot easier. However, for C++ we had to build every function by hand to do what requires one line in R. The one thing I do appreciate about C++ in this scenario is that it gave me a deeper understanding of how R was calculating all of these metrics on our datasets. For example, C++ made it clear how covariance are correlation are so clearly related.

Statistical Measures

The first measure we calculated was the sum of all our observations. This could be useful in unison with the mean or the number of elements to give a rough estimate of how much data has been collected. The second metric, mean, is a lot more useful because it allows us to see the average of all the observations for a specific feature. Median is slightly less useful but it does allow us to see where the middle of the datapoints lie, depending on how close the median is to mean could give a less-than-accurate approximation of how much the data points differ from each other. The range also works in unison with the mean to show just how spread out the observations are. In conclusion, all of these metrics are not very descriptive by themselves but all together can help paint a better picture of a dataset if there is not visualization.

Covariance and Correlation

Covariance and Correlation are metrics to measure how different or similar two variables are from each other. But by difference we don’t mean absolute difference, we mean if the greater values of one variable have a corresponding greater value in the other value. If this happens then covariance and correlation are high, the same is true for two lesser corresponding lesser values. However, if the greater values of one variable have a corresponding lesser value in the other variable (or vice-versa) then the correlation and covariance are low. Covariance is not scaled to anything so it might be a bit hard to interpret without more information on the data, put simply it’s the raw variances of the two variables together. Correlation is a much more useful statistic because it has a scale of -1 to 1. The closer to 1 the correlation is the more of a linear relationship the two variables have, the closer to -1 the opposite is true.