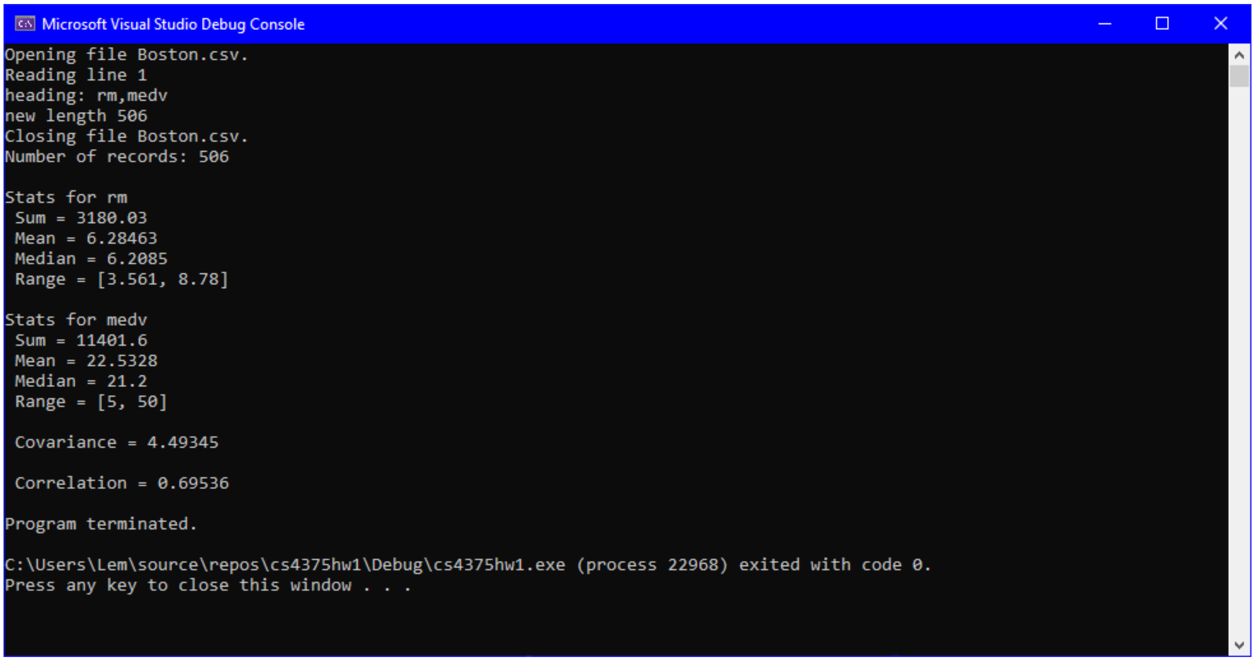
Zuhayr Ali

Dr. Karen Mazidi

CS 4375

12 September 2022

Data Exploration



```
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.

C:\Users\Lem\source\repos\cs4375hw1\Debug\cs4375hw1.exe (process 22968) exited with code 0.
Press any key to close this window . . .
```

Building the functions for this exercise was definitely more tedious than using the analogous functions in R. Each statistic is one function call away in R as a native component of the language; it reduces the process to thoughtlessness and allows one to focus on other tasks. Only one of my equivalent functions in C++ was a single line until I optimized it for faster calculations. Between learning the equations, structuring the equations per C++, managing the data passed to functions, bug fixes, and optimization, the process in C++ involved much more forethought and frustration. However, having designed the C++ functions I know how much memory and speed they take, characteristics that are encapsulated away from my view in R.

Here will be discussed the simpler statistical measures designed in this exercise. The mean of an attribute is equivalent to the average, the sum of all observations divided by the number of observations. The median of an attribute is the middle value of the attribute when sorted by value, which is an average of the two middlemost values when the attribute has an equal number of observations. The range of an attribute is a two-value statistic displaying the lowest and highest values in the attribute, with all other observations existing between the two. The range can give a good understanding of how much an attribute varies in observations, while the mean can indicate the most likely observation to expect and comparing the mean to the median could indicate how uniform the attribute is in its range.

Now will be discussed the more complex statistical measures of this exercise. The covariance of a dataset is a measure of the relationship between two attributes, specifically how much one attribute matches the increase or decrease of the other per observation. The correlation of a dataset is the covariance standardized as a value with range [-1, 1], much quicker at indicating the type and strength of the relationship between two attributes. This information is greatly useful to machine learning as it allows a model to understand how predictable observations for a dataset are as well as accurately predict observations for a dataset by understanding how attributes interact with each other.