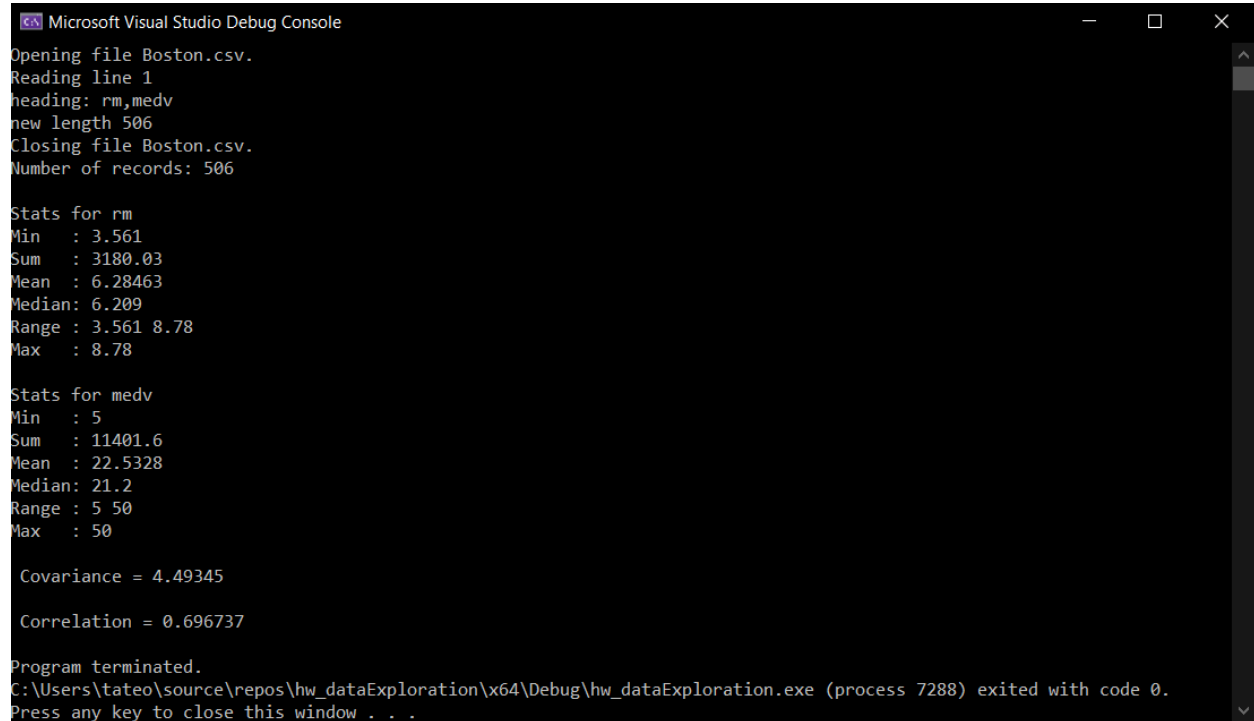


Hw1 Data Exploration

Pst170030 / Parker Tate

CS4375.004

(a) Code Output



```
Microsoft Visual Studio Debug Console

Opening file Boston.csv.
Reading line 1
heading: rm,medv
new length 506
Closing file Boston.csv.
Number of records: 506

Stats for rm
Min : 3.561
Sum : 3180.03
Mean : 6.28463
Median: 6.209
Range : 3.561 8.78
Max : 8.78

Stats for medv
Min : 5
Sum : 11401.6
Mean : 22.5328
Median: 21.2
Range : 5 50
Max : 50

Covariance = 4.49345

Correlation = 0.696737

Program terminated.
C:\Users\tateo\source\repos\hw_dataExploration\x64\Debug\hw_dataExploration.exe (process 7288) exited with code 0.
Press any key to close this window . . .
```

- (b) In my experience recreating the built-in R functions in C++, R has more tools readily available than what a simple function can provide. Most notably, R has the capacity for many additional parameters to account for variables such as NA data points, which would require a lot of custom function overloading. Managing different data sets is also much easier in R as we don't have to rely on file streams, although that may also be my unfamiliarity with the concept in C++.

Although not tested in C++, I imagine generating different graphs and plots corresponding to data sets is much easier in R.

- (c) **Mean** is the mathematical average value of all data values in a given data set. This is useful for finding the general expected value for any data value, as well as calculating other useful metrics such as the standard deviation.

Median is another form of averaging, but by finding the middle data value in a sorted data set. This is useful in cases where there are a couple of extreme outliers that may disproportionately skew the mean.

Range is the minimum and maximum values of a given data set. This shows the general scope of the data.

- (d) Covariance and correlation are measures of the relationship between two variables.

Covariance measures the strength and direction of the relationship, with a positive covariance indicating that when one variable shifts more positive, the other variable will as well. Negative covariance indicates the opposite.

Correlation is the same as covariance but standardized onto a value ranging between -1 and 1. A correlation of 1 implies a perfect positive linear relationship between variables, while a correlation of -1 implies a perfect negative linear relationship. A correlation of 0 implies no linear relationship.

These measurements are useful in machine learning because it allows us to create more accurate predictive models.