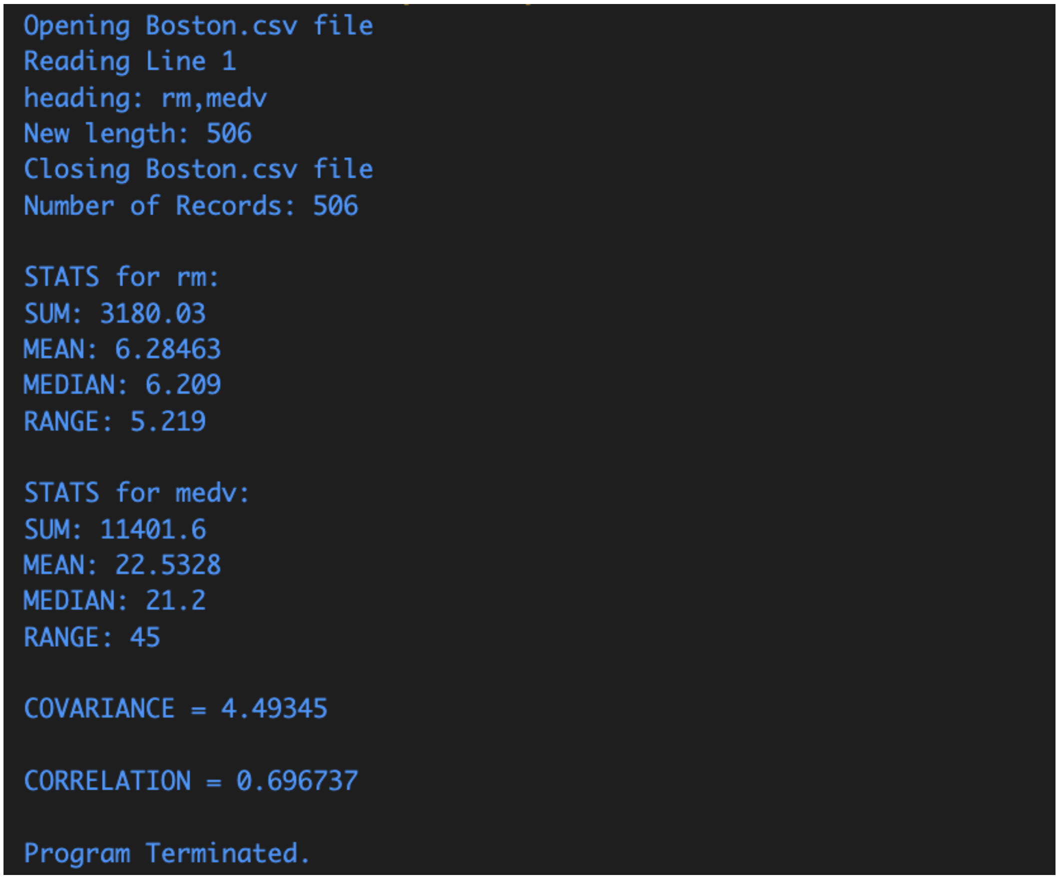
Sanjana Jadhav
CS 4375.004

C++ Data Exploration Report

- **Output of the Code**



```
Opening Boston.csv file
Reading Line 1
heading: rm,medv
New length: 506
Closing Boston.csv file
Number of Records: 506

STATS for rm:
SUM: 3180.03
MEAN: 6.28463
MEDIAN: 6.209
RANGE: 5.219

STATS for medv:
SUM: 11401.6
MEAN: 22.5328
MEDIAN: 21.2
RANGE: 45

COVARIANCE = 4.49345

CORRELATION = 0.696737

Program Terminated.
```

- **Built-in Functions vs. Coding my Own Functions**
  As I was coding, the significance of all the values involved in the calculations became clear. It is definitely more efficient to use the in-built functions in R, but coding my own functions in C++ helped me solidify my understanding of the covariance and correlation formulas. As someone who loves math, I can confidently say that I enjoyed this assignment!

- **Data Exploration & Mean, Median, and Range**
  The **mean** is the average (center) of all the values in a dataset. This is found by summing the values and dividing by the number of values. The **median** is the middle value of the dataset (in ascending order). Sometimes, the median is more useful in determining the average if the dataset is skewed. The **range** provides a measure of the distribution of the data. A smaller value means that the data is clustered, and a bigger value indicates that it is more spread out.

- **Covariance and Correlation**

  The **covariance** measures how one variable changes in accordance with another variable. Positive values indicate that both variables are increasing/decreasing, while negative values indicate that one variable is increasing while the other is decreasing (and vice versa). The **correlation** is a measure of the covariance, but it is scaled in the range of [-1,1]. Therefore, this indicates how strong the relationship between the two variables is. Correlation is high if the value is close to +/- 1 (sign indicates direction), while it is low if the value is closer to 0. If the value is 0, then there is no correlation. As a result of covariance and correlation, we can construct a model of the data (**y = wX + b**). The model can then predict future y values for new data (**X**). This is called **linear regression**.

- **Sources**
  - ML Textbook
  - Median
  - Range.
  - Covariance
  - Correlation