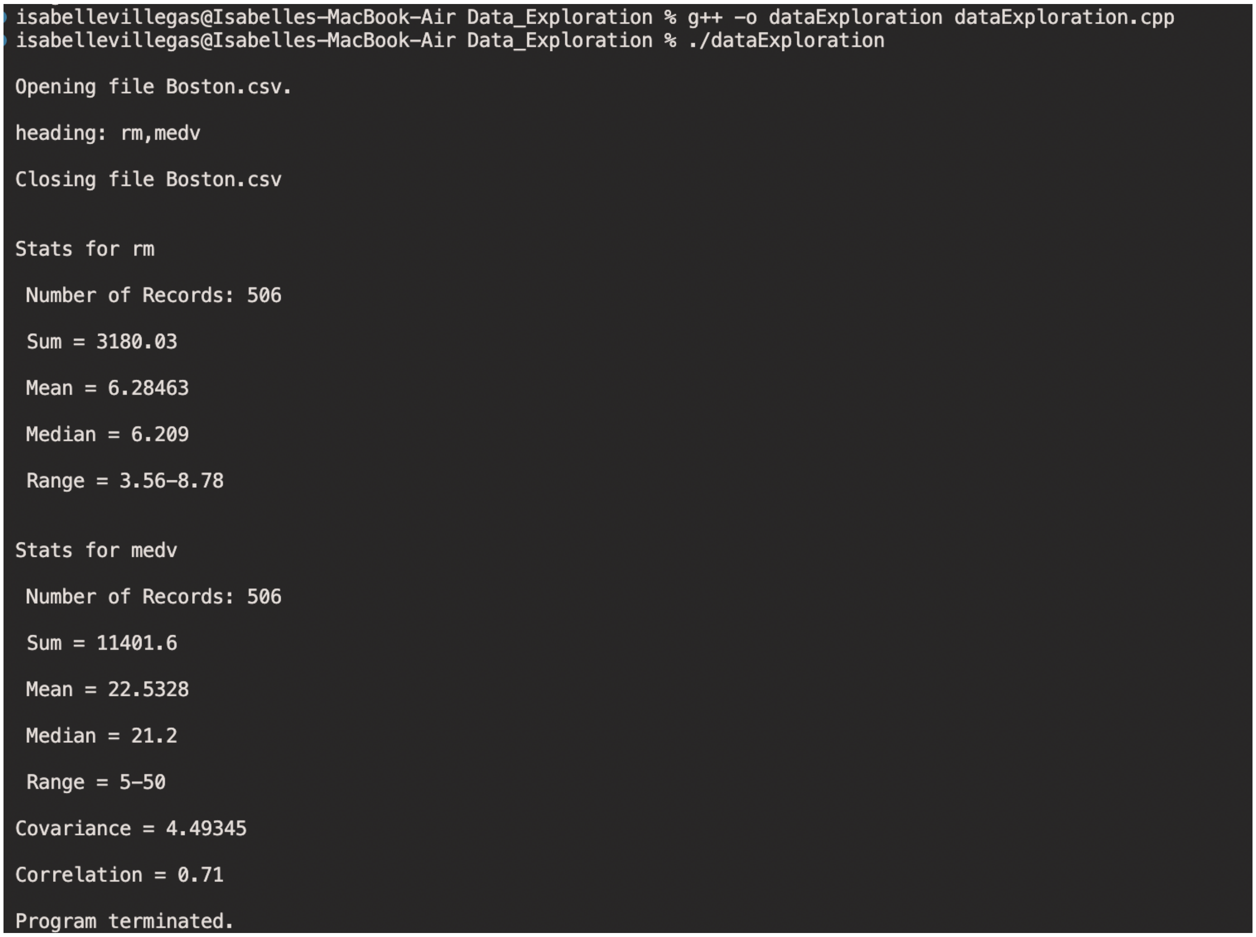
CS 4375.003

Isabelle Villegas

iav180000

Data Exploration

a. Here is my program running on VS Code



```
isabellevillegas@Isabelles-MacBook-Air Data_Exploration % g++ -o dataExploration dataExploration.cpp
isabellevillegas@Isabelles-MacBook-Air Data_Exploration % ./dataExploration

Opening file Boston.csv.

heading: rm,medv

Closing file Boston.csv


Stats for rm

 Number of Records: 506

 Sum = 3180.03

 Mean = 6.28463

 Median = 6.209

 Range = 3.56-8.78


Stats for medv

 Number of Records: 506

 Sum = 11401.6

 Mean = 22.5328

 Median = 21.2

 Range = 5-50

Covariance = 4.49345

Correlation = 0.71

Program terminated.
```

b. Using the built in functions are really nice in RStudio, simply because they are streamlined and easy to use, however, I found that while working in C++, I was able to understand each function better and see how they worked in the code.

c. Mean is the average of a data set so taking each value and adding them together and then dividing them by the total elements in that set. The Median is found by ordering a data set from lowest to highest and then taking the middle of the set. The Range was simply taking the smallest and highest value in that data set. These are all super important in Machine Learning since we have to essentially make all data that we are given quantifiable and able to be measured in different ways. If there is a consistent number being shown in a data set, that has value to it and can show patterns in data.

d. Covariance is good in ML since it shows the direction of the relationship between two data sets while correlation shows how similar they are to each other, if they trend the same way then they are going to have a high correlation. This is useful in ML since it can show the relationships in data and how they relate to one another.