Data Exploration

```
PS C:\Users\rjgag\OneDrive\Documents\School\Fall 2022 Junior\CS 4375\Data Exploration> .\main
Opening file
Reading line one:
new length: 506
closing file
number of records: 506

Summation of rm: 2923
Summation of medv: 11190

Mean of rm: 5.77668
Mean of medv: 22.1146

Median of rm: 6.2085
Median of medv: 21.2

Range of rm: 5.219
Range of medv: 45

Covariance of the two: 4.70628

Correlation of the two: 0.589401
PS C:\Users\rjgag\OneDrive\Documents\School\Fall 2022 Junior\CS 4375\Data Exploration>
```

1. Describing your experience using built-in functions in R versus coding your own functions in C++
    a. Using the built in functions of R makes things much simpler as opposed to creating your own in C++ because everything is readily available for you to use. However, when building your own functions, it ensures that you fully understand the math behind each function and each variable forcing you to learn and understand the data and what you are trying to accomplish completely.
2. Describe the descriptive statistical measures mean, median, and range, and how these values might be useful in data exploration prior to machine learning
    a. Mean is simply the average of all the points of data combined, median is the “true center” that counts from each side until the middle point is reached, and range is the difference between the highest value and lowest value in the data set.
    b. The importance of these variables cannot be understated. These are the basics of analyzing data and spotting trends in it, as well as simply understanding the data better as a whole. Without these attempting to discover anything to do with the data would be near impossible.
3. Describe the covariance and correlation statistics, and what information they give about two attributes. How might this information be useful in machine learning?
    a. Covariance is measuring how much a change in one variable will affect the other. It shows the amount that the variables seem to change or follow patterns with one

another. Correlation is just taking the covariance and putting in a metric that is more understandable and ranges from -1 to 1. These are both very important for machine learning because it shows how much one data set interacts with another, showing us if there will be a bias when comparing them, if it will skew the results, etc. It is crucial to know just how much two data sets are correlated, even if just by chance, to avoid algorithms with too much bias or not enough.