**a. copy/paste runs of your code showing the output**

Using C++:

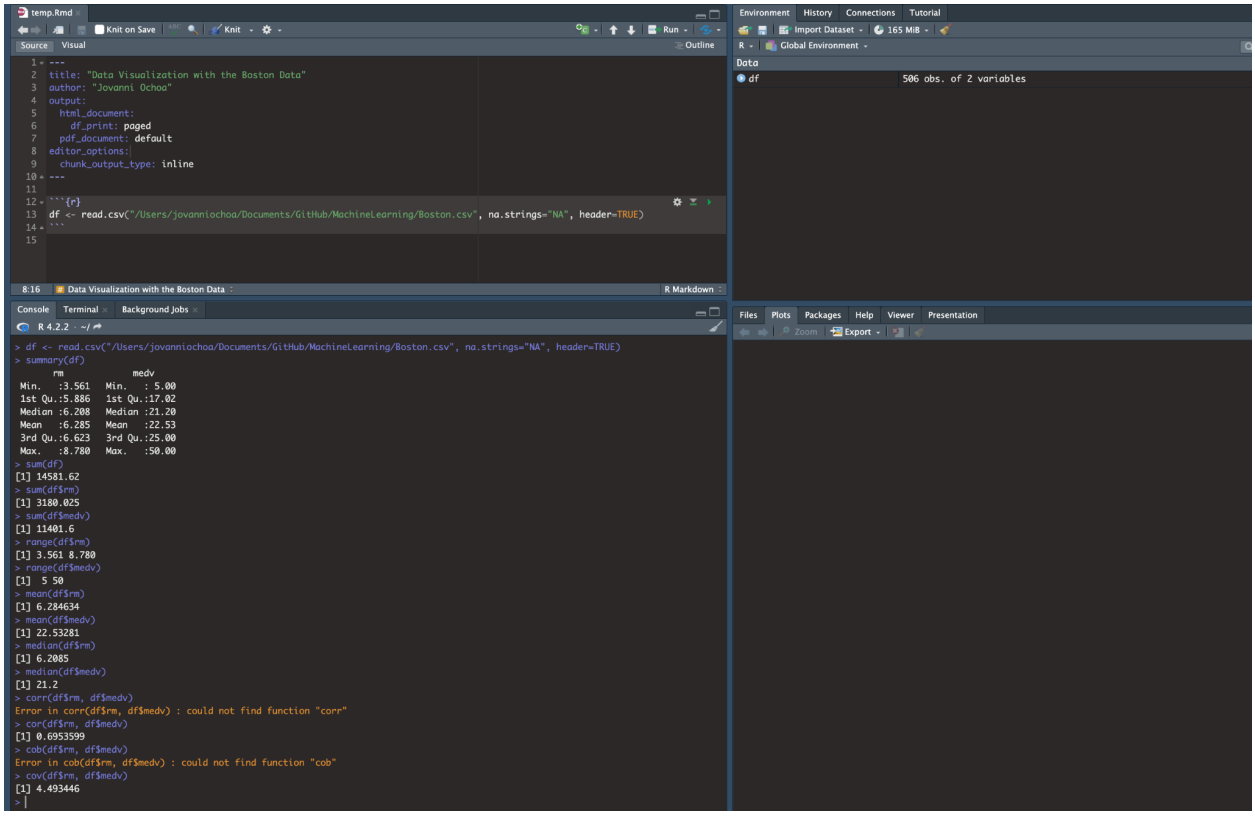
```
Loaded '/usr/lib/system/libsystem_product_info_filter.dylib'. Cannot find or open the symbol file.
Loaded '/usr/lib/system/libsystem_sandbox.dylib'. Symbols loaded.
Loaded '/usr/lib/system/libsystem_secinit.dylib'. Symbols loaded.
Loaded '/usr/lib/system/libsystem_kernel.dylib'. Symbols loaded.
Loaded '/usr/lib/system/libsystem_platform.dylib'. Symbols loaded.
Loaded '/usr/lib/system/libsystem_pthread.dylib'. Symbols loaded.
Loaded '/usr/lib/system/libsystem_symptoms.dylib'. Symbols loaded.
Loaded '/usr/lib/system/libsystem_trace.dylib'. Symbols loaded.
Loaded '/usr/lib/system/libunwind.dylib'. Symbols loaded.
Loaded '/usr/lib/system/libxpc.dylib'. Symbols loaded.
Loaded '/usr/lib/libobjc.A.dylib'. Symbols loaded.
Loaded '/usr/lib/liboah.dylib'. Symbols loaded.
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
2
Stats for medv
Sum = 11401.6
Mean = 22.5328
Median = 21.2
Range =  5 50
2
 Covariance = 4.49258

 Correlation = 0.695226

Program terminated.
The program '/Users/jovanniochoa/Documents/GitHub/MachineLearning/DataExploration' has exited with code 0 (0x00000000).
```

Using R:



**b. describing your experience using built-in functions in R versus coding your own functions in C++**

Using built in functions is a lot easier to do in R as opposed to creating them and using them in C++. Not only is it easier, but there's more you can do with the final output of these in R like creating graphs of the data without the necessities for tons of libraries. I did see a couple of problems such as with rounding, so I'm not sure which is more accurate, but I assume R ended up rounding one of my numbers and therefore I got a somewhat different result at the third decimal place since I didn't find an issue with C++.

**c. describe the descriptive statistical measures mean, median, and range, and how these values might be useful in data exploration prior to machine learning**

Mean is useful when trying to find the average of a group of numbers. The mean gets the sum of all the data and then divides it by the number of entries of data. The problem with this average is that it tends to skew to the direction of outliers. In other words, if the data isn't given a good sample size, or if the data has a large outlier, then the average using mean would not be the greatest. This is where median comes in. Median sorts the data and takes from the top number and bottom number until there is either only one number left in the middle, or two numbers left in the middle to which it calculates the average from. The median is however less accurate to get the average of data when there is less skewness. This means, the median is not optimal for normal data sets. Finally the range is important because it tells you the maximum and the minimum of the data in question. This is useful so that you know the distribution between the absolute lowest and highest.

**d. describe the covariance and correlation statistics, and what information they give about**
**two attributes. How might this information be useful in machine learning?**

The covariance and correlation statistics work to show how two values or groups of values relate to each other. This is to say that if one group is moving in a certain fashion and another is moving similarly, then the two groups have a positive covariance and correlation. How similar they are to each other, the better the numbers. For correlation this is shown as a -1 or +1. Covariance works around the mean of the value, and the closer it is to the determined mean, the stronger the correlation.

This information might be useful in machine learning to determine if the two groups correlate or not. That's important because it can tell you how accurate the machine may be in trying to predict something. If there is a low correlation between the original and the machine learning model, then the model didn't do a good job at performing. This can mean a variety of different things from the math being messed up to the machine learning model not being fed enough data.