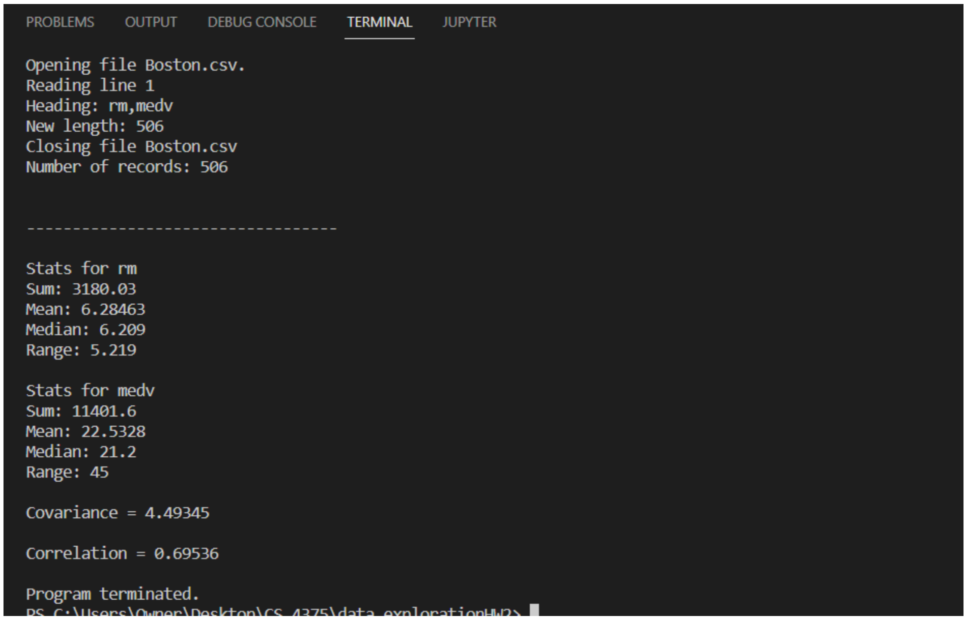
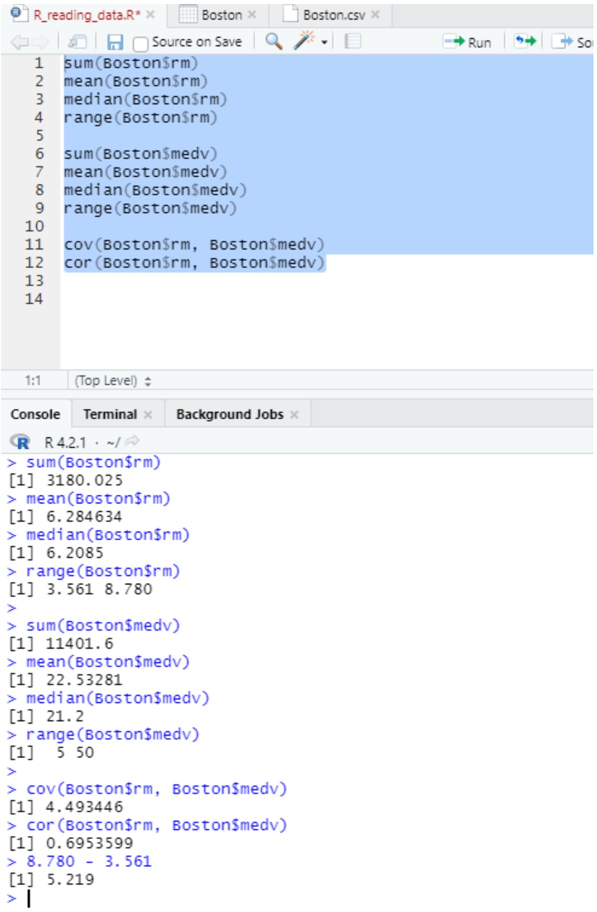
C++ & R: Data Exploration Documentation

A. Results from C++ file & R:



```
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: 6.209
Range: 5.219

Stats for medv
Sum: 11401.6
Mean: 22.5328
Median: 21.2
Range: 45

Covariance = 4.49345

Correlation = 0.69536

Program terminated.
PS C:\Users\Owner\Desktop\CS 4375\data explorationHW2>
```



```
sum(Boston$rm)
mean(Boston$rm)
median(Boston$rm)
range(Boston$rm)

sum(Boston$medv)
mean(Boston$medv)
median(Boston$medv)
range(Boston$medv)

cov(Boston$rm, Boston$medv)
cor(Boston$rm, Boston$medv)


```

```
> sum(Boston$rm)
[1] 3180.025
> mean(Boston$rm)
[1] 6.284634
> median(Boston$rm)
[1] 6.2085
> range(Boston$rm)
[1] 3.561 8.780
>
> sum(Boston$medv)
[1] 11401.6
> mean(Boston$medv)
[1] 22.53281
> median(Boston$medv)
[1] 21.2
> range(Boston$medv)
[1]  5 50
>
> cov(Boston$rm, Boston$medv)
[1] 4.493446
> cor(Boston$rm, Boston$medv)
[1] 0.6953599
> 8.780 - 3.561
[1] 5.219
> |
```

B. From my experience using built-in functions in R is a much easier way to gather and analyze data compared to coding my own functions in C++. Coding the functions in C++ has the advantage of understanding the process of how each statistic is being calculated, though it does take a considerable amount of time to create. If I had to choose which method I prefer more, I would say using R because of the time and simplicity advantages.

C. Mean, the average value found in a set of data, is important in data exploration as it tells you the most frequent data point using every data in the set. Median, the value at the middle of the data set, is important in data exploration especially in conditions where there are extreme outliers within the data set. Range, the lowest and highest values in a data set, is important in data exploration because it tell you the scope of the entire data set.

D. If there is a high covariance between two attributes, then they both appear to have a strong relationship in how their data varies. Whereas if they have a low covariance, the relationship is weaker. Correlation is used more to determine if one attribute affects another attribute’s results and to determine what future results may be. The higher the correlation, the closer the relationship. These statistics are important because machine learning primarily is about predicting future results and making sure that the predictions are as accurate as possible; correlation and covariance are significant statistics to make sure the accuracy remains.