Ved Nigam

Data Exploration in C++

Output of my code:

```
Opening Boston.csv
Reading in line 1
heading rm,medv
New length is 506
Closing file
Number of records is 506

 Stats for rm
Sum is 3180.03
Mean is 6.28463
Median is 6.209
Range is 5.219

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

 Covariance is 4.49345

 Correlation is 0.69536

 Program ended
```

After having knowledge of built-in functions in R for statistics, coding the above functions seemed very tedious and had a greater room for error. The first thing I did at the start of this assignment was compute these statistics in R because it is difficult to calculate incorrectly because the functions are already built in. After coding in R, I moved to C++. It was surprising how many times I had forgotten to add parentheses or forgotten the order of operations used in C++ coding. If the primary task requires data manipulation/computation, R is the safer language to work with compared to C++.

The computed statistics provide us with a lot of insight of quantitative data. The mean gives us the average of the data set, the median gives us the midpoint of the data irrespective of other values, and the range gives us the difference between the smallest and the largest value in the data. Such information can be very important for summarizing a dataset rather than having to analyzing each dataset. Large CSVs and lists of data can be very intimidating to use, but if we focus on such statistics, we can analyze data more efficiently. With this, we can also more efficiently pick which datasets we want our machine learning models to learn from.

Two other statistics that we calculated in this project was covariance and correlation. Covariance measures the variability of the two variables assuming that they are Random Variables. This means that the two variables do not depend on each other at all, but if the variance between datapoints from the two sets has a linear relationship, the covariance value will indicate that. Being a linear indicator, it does matter if the value is positive or negative. The correlation value indicates whether there is or is not a relationship between the two variables. It ranges from -1 to 1: 0 being there is no relationship, positive value meaning there is a positive relationship, and a negative value meaning there is a negative relationship.