Swarn Singh
02/01/23

Assignment 2 - Portfolio Component 1: Data Exploration

a.

```cpp
assignment2.cpp  ✕
assignment2.cpp > ...
1    #include <iostream>
2    #include <vector>
3    #include <fstream>
4    #include <numeric>
5    #include <algorithm>
6    #include <cmath>
7
8    using namespace std;
9
10   // vector<double> rm, medv;
11
12   double sum_vector(vector<double> v)
13   {
14       return accumulate(v.begin(), v.end(), 0.0);
15   }
16
17   double mean_vector(vector<double> v)
18   {
19       return sum_vector(v) / v.size();
```

```
TERMINAL    GITLENS

swarn@MacBook-Pro-4 Assignment2 % ./a.out
--------------------------
Reading line 1
heading: rm,medv
new length: 506
Closing file Boston.csv.
Number of records: 506
--------------------------
rm:
sum: 3180.03
mean: 6.28463
median: 6.2085
range: 5.219
--------------------------
medv:
sum: 11401.6
mean: 22.5328
median: 21.2
range: 45
--------------------------
covariance: 4.48457
correlation: 0.69536
--------------------------
swarn@MacBook-Pro-4 Assignment2 %
```

b. After using R in class and C++ for this project, I can understand why R isa preferred language of choice for machine learning. There is a lot less fundamental work required to get basic data from a file required as opposed to C++. The out of the box functionality that R provides lets me focus on the actual data results rather than trying to spending time to simply generate the results. C++ does provide more control over the code and can result in faster performance. In C++, you have the flexibility to write custom functions that fit your specific needs and can optimize the code for performance.

c. The mean, median, and range are descriptive statistical measures that provide information about the central tendency and spread of a dataset. The mean is the average value of the data, calculated by summing all the values and dividing by the number of observations. The median is the middle value of the data when it is sorted in ascending or descending order. The range is the difference between the maximum and minimum values in the data.
These values can be useful in data exploration prior to machine learning because they provide a basic understanding of the characteristics of the data. The mean and median can be used to understand the central tendency of the data while the range can provide information about the spread of the data.

d. Covariance and correlation are statistical measures that describe the relationship between two attributes. Covariance is a measure of the joint variability of two variables. It is a scalar value that describes the degree to which two variables change together. The covariance can be positive, negative, or zero, indicating the direction of the relationship between the two variables. This information can be useful in machine learning because it can provide insight into the relationships between the variables in the data. For example, high covariance or correlation between two variables might indicate that one of the variables can be dropped from the model without affecting the model performance. On the other hand, low covariance or correlation might indicate that the variables are independent and provide unique information, and therefore both should be included in the model.

Resources:

Textbook

https://www.fireblazeaischool.in/blogs/covariance-and-correlation-in-machine-learning/