Aarya Patil
CS 4375
Professor Mazidi
02/01/23

Assignment Rundown: Portfolio Component 1 (Data Exploration)

Output from running the code:

```
Opening input file: Boston.csv


Reading info from file...
  File heading: rm,medv
  Resized length of vectors rm and medv: 506


Closing input file: Boston.csv


STATISTICS
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 ended with exit code: 0
```

Experience using built-in functions in R vs. coding own functions in C++:
Using the built-in functions in R is definitely preferred when obtaining the data and statistics is your main goal. It’s much faster and allows you to be able to manipulate in more ways to draw more conclusions from the results. However, to really understand what all the functions really do to the data, coding in C++ is better. Being forced to build the functions from scratch really allows you to grasp what the purpose of each function is and what values and data it uses to obtain its results. That being said, writing your own functions is much more tedious and can be frustrating if you are unable to figure out how to fix your function. But once you do, it’s very rewarding.

Describe mean, median, and range and how they might be useful in data exploration prior to machine learning:
Mean is the average of a dataset. It is obtained by calculating the sum and dividing it by the number of values present within that set. Median is the middle of an ordered dataset. You must first sort your data set from least to greatest and the middle value represents your median. The range is the difference between the largest and smallest values within the dataset. Similar to median, you need to sort your dataset first, and then subtract the first value from the last value to get the range.
These values must have been very important to data exploration prior to machine learning as they allowed people to understand the nature of a dataset. You could also calculate these values for a particular data set repeatedly over a certain period of time to study how the data changes. This ability to analyze the data would have been very important as the concept of being able to obtain data and adapt accordingly did not exist back then.

Describe covariance and correlation, what information they give about two attributes, and how this information may be useful in machine learning:
Covariance is the statistic that measures how changes in one variable are associated with changes in a second variable. Correlation is very similar to covariance except its scaled to [-1, 1]. Both of these values show the dependency between two attributes, but correlation also shows the strength of the dependency, the closer the number is to 1 (or -1) the stronger the relationship.
Obtaining these values can prove very useful in machine learning as they can help us understand how changes to one attribute affect another. This can help people understand how changes to a small sample can affect a population.