Portfolio Component 1: Data Exploration

Michael Stinnett
CS 4375 Intro to Machine Learning

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

```
PS D:\UTD\Fall 2022\CS 4375.003 - Introduction to Machine Learning - F22\Intro-ML>
.\Data_Exploration_Boston.exe
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: 5.219

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

 Covariance = 4.48808

 Correlation = 0.695542

 Program terminated.
```

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

Coding my own functions solidified my appreciation for the built-in functions in R. The design of R to easily handle statistical data shined while I was coding my own functions. What I did in 15 lines of code could be done in R in 1 and it would be better. However, a benefit of writing my own functions is that I knew exactly what it was doing under the hood. Whereas the R functions might have their own quirks and edge cases that I would not know about until I ran into them.

Portfolio Component 1: Data Exploration

Michael Stinnett
CS 4375 Intro to Machine Learning

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

When first getting your feet wet with a new data set, the classic descriptive statistical measures are good starting points for finding patterns in what your data is telling you. A mean can help illustrate an average case for the data when the data has few outliers. A median can sometimes be a better example of an average case when the data does have several outliers. A range can help show point out outliers in a data set.

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

Covariance tries to the relationship between the changes in one variable and how that affects the other. The correlation statistic is just the covariance statistic scaled down to the ranges 1 to -1. With 1 showing a perfect positive correlation and -1 showing the perfect negative correlation. Finding data that is correlated is important when trying to predict a target in a model. If the model has too many highly correlated data sets, then there might a bias towards those data sets when predicting the target. If the model doesn't have any correlated data sets, then the predicted target will likely range wildly from the actual measured data point.