Charles Wallis
CS4375 Intro to Machine Learning
September 11, 2022

Portfolio Component 1: Data Exploration

I enjoyed using C++ rather than R for this exploration assignment, mainly because I was able to create reusable functions, and create a flow that fits my needs as I go. R is definitely powerful as it has many built in functions that are easy to use, but with C++ and coding the functions myself, I know better how the numbers are calculated exactly (covariance and correlation data for example)

Statistical measures that I used here like the mean, median, and range are all important and can be useful in any type of data exploration. Mean data can be useful when trying to find an "unobserved" data within the dataframe, as an 'expected number'. Similarly with the median data, this can provide the same information but it will likely exclude the outlier numbers within the data that might be skewing the average. Range shows how far off the minimum and maximum data points are apart, and it can be used to better visualize what numbers are in the dataset along with the mean and median data.

Correlation and Covariance is important in machine learning, as it is often used to measure how correlated the variables are. If the correlation value is close to 0, it means there is little correlation and 1 indicates high correlation. Covariance is similar as it measures how changes in one variable are associated with changes in a second variable, but it is not in a scale from 0 to 1, and this number can range from anywhere, which is why the correlation metric is more preferred in data science. Higher correlation data can be weighed heavier when creating a model, and the closer to 1, it will mean the model is more accurate.

data_exploration.cpp input: Boston.csv
data_exploration.cpp output:

```
Number of Records: 506
|=====RM Stats=====|
|Sum      : 3180.03
|Mean     : 6.28463
|Median : 6.2085
|Range    : 5.219
|====MEDV Stats====|
|Sum      : 11401.6
|Mean     : 22.5328
|Median : 21.2
|Range    : 45
|=====RM and MEDV====|
|Covariance   : 4.49345
|Correlation : 0.69536
```