Meinhard Capucao
9/10/22
CS 4375.003

```
PS C:\Users\meinc\Documents\C++> cd "c:\Users\meinc\Documents\C++\" ; if ($?) { g++ dataExploration.cpp -o dataExploration } ; if ($?)
 { .\dataExploration }
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: 6.209
Range: 5.219

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

 Covariance = 4.49345

 Correlation = 0.69536

Program terminated.
PS C:\Users\meinc\Documents\C++>
```

I found that the built-in functions in R were very easy to learn and implement. Even though the syntax for R was new for me, the syntax was simple and all I had to do was call the functions. For example, the four basic functions for this project are sum(), mean(), median(), and range(). It's extremely simplified and all I need is the parameter, or some nuances with is.na. However, the C++ was very challenging for me. I never was strong to begin with, and it's been years since I used C++ for a full project. Therefore, even if this data exploration project itself is pretty simple, it took me a while to complete since I needed to refine and reacquire my basic knowledge of C++. I had to look up syntax for the math functions, and even writing functions themselves again! However, this served as a good refresher for me and helped me brush up on my fundamentals.

Mean is the average of all numbers from a set of numbers. Median is the middle number from a sequence of sorted numbers by size. Range is the difference between the lowest and highest values from a set of numbers. I feel like these values are integral in data exploration even before diving deep into machine learning. These three factors help describe and analyze overall data. For example, the mean can help find generalizations in data by showing the average. It can show where certain categories or events trend towards. The median is similar but can be used in skewed distributions, which can occur often. The range tells people where to look and gives a boundary to help analyze data in the right area.

Correlation tells us how variables are correlated. It is measured by the covariance of two variables, let's say x and y, divided by the standard deviation of x multiplied by the standard deviation of y. Correlation is covariance, scaled to [-1, 1]. Then, covariance measures how changes in one variable are associated with changes in another. In covariance, the numbers can range wildly. Below I put an image for the formulas of correlation and covariance from the Machine Learning Handbook by Dr. Karen Mazidi. These are useful because these show association between variables. If two variables have strong correlation, then we know they can predict each other. In machine learning, learning the correlations between variables help with predicting future data, and that is a big part of machine learning!

$$cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Formula for correlation

$$\rho_{x,y} = Corr(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

Formula for covariance