David Favela Corella

Dxf200002

CS 4375.003

C++ Data Exploration

1. Output from the code:

[Running] cd "c:\Users\david\OneDrive\Documents\UT Dallas\UT DALLAS\Fall 2022\CS 4375\CS4375\_Portfolio\CS4375\_Portfolio\" && g++ dataexploration.cpp -o dataexploration && "c:\Users\david\OneDrive\Documents\UT Dallas\UT DALLAS\Fall 2022\CS 4375\CS4375\_Portfolio\CS4375\_Portfolio\"dataexploration

Opening Boston.csv

Reading line 1

heading: rm,medv

rm stats

Sum: 3180.03

Mean: 6.28463

Median: 6.2085

Range: 5.219

mdev stats

Sum: 11401.6

Mean: 22.5328

Median: 21.2

Range: 45

Covariance of rm and medv: 4.49345

Correlation of rm and medv: 0.696737

[Done] exited with code=0 in 0.725 seconds

1. My experience using the built=in functions in R is not much but from what we’ve done in coursework, I can say it is much simpler to use them and they are much more reliable. I’ve been learning and using C++ for several years now, so implementing the functions is not much trouble. The trouble comes in how the data is handled. In my case, I noticed that some of the results were off by a few floating points because of how C++ handles operations with these types of data. One does not worry about accuracy in data when using the built-in functions in R.
2. The mean is known to be the average in a set of data. Median is the median value in the set of data, meaning the value in the middle of a sorted data set. Lastly, the range is the difference between the highest value and lowest value in the data set. These values help greatly in data exploration as we based a lot of statistic on them. Whether it is to calculate the average grade point in a school year, calculating variations of weight and height in a population, these formulas let us know some descriptive information on the data sets without using machine learning.
3. Covariance in statistics measures the relationship between two data sets and how different they are. The closer values are to 0, the lesser the covariance. The correlation is a measure of the dependency between two different data sets. It determines how closely a set of data moves along another one. In terms of attributes, it provides a measure of how related and dependent those two attributes are. This helps machine learning create patterns in the data sets and know how might the data move according to something that is fed to it.