Colton Townsend
2/4/2023
CS 4375.004

Data Exploration

a. Expected output:

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 of vector: 3180.03

Mean of vector: 6.28463

Median of vector: 6.2085

Range of vector:

Min: 3.561

Max: 8.78

Stats for medv

Sum of vector: 11401.6

Mean of vector: 22.5328

Median of vector: 21.2

Range of vector:

Min: 5

Max: 50

Covariance = 5

Correlation = 0.437132

Program terminated.

b.

Using built-in functions in R comes with the convenience of not having to build your own functions. For completing simple tasks, R has a significant advantage over C++ due to the time saved by using built-in functions. C++

c.

Both mean and median help us understand the "average" value of a set of values. If mean and median differs significantly then it suggests that outliers exist or that the set of data heavily favors one extreme end over the other. Range helps us understand these ends by defining the minimum and maximum values observed in the set, but it can also reveal outliers that are influencing the value of the mean. In machine learning, it is important that we understand what data we are feeding the machine in order to get the desired prediction behavior.

d.

Covariance measures the relationship between two variables. Positive values suggest that a great value of one variable tends to correspond with a greater value of the other variable. Negative values suggest that a greater value of one variable tends to correspond with a lesser value of the other variable. Correlation measures the relationship between two variables in a different manner; measuring how variables are linearly related. A correlation coefficient is +1 in the case of a perfect linear relationship and -1 in the case of a perfect inverse linear relationship. Covariance and correlation help us understand how two sets of data are related to each other. Understanding how sets of data are related is important for machine learning as we want the machine to recognize patterns between sets of data when they exist.