Umar Ali-Salaam

Professor Mazidi

CS 4375.003

11 September 2022

Data Exploration

a)

Opening file...

Read line 1

heading: rm,medv

new length 506

Closing file..

Number of records: 506

STATS FOR RM

Sum: 3180.03

Mean: 6.28463

Median: 6.2085

Range: 5.219 (Max: 8.78 Min: 3.561)

STATS FOR MEDV

Sum: 11401.6

Mean: 22.5328

Median: 21.2

Range: 45 (Max: 50 Min: 5)

Covariance = 4.49345

Correlation = 0.69536

Program terminated.

C:\Users\ualis\UAHW1\Debug\UAHW1.exe (process 17972) exited with code 0.

To automatically close the console when debugging stops, enable Tools->Options->Debugging->Automatically close the console when debugging stops.

Press any key to close this window . . .

b) The built-in functions on R are far faster to use versus actually having to code and come up with the equation myself in C++. I did a large data science project on NBA Statistics, using R, and ended up having to write a little over 1800 lines of code. I don't even want to think about how many lines of code I would've had to write if I did it in C++.

c) Mean is a very useful tool for any statistical analysis. It provides a central tendency or typical value for a dataset, and allows the statistician to access more equations. Median similar to mean also provides a central tendency for a dataset, but can be influenced by any skewing of the data, this allows statisticians to get a better understanding of how skewed their data could potentially be. Range allows you to see the spread of a dataset, which is important for calculating things like accuracy, precision, variance, and other equations.

d) Covariance helps you better understand the directional relationship of two variables, while correlation measures how meaningful and relevant the linear relationship between different variables are. These two calculations can be used in Machine Learning to help the computer better understand the relationship of multiple variables, and then remember that and use that data to help make decisions in the future.