Dylan Kapustka

CS 4375.003

Dr. Mazidi

September 11, 2022

Data Exploration

a)

```
rm:
The sum of rm is: 3180.03
The mean of rm is: 6.28463
The median of rm is: 6.2085
The range of rm is: 5.219

medv:
The sum of medv is: 11401.6
The mean of medv is: 22.5328
The median of medv is: 21.2
The range of medv is: 45

medv and rm:
The covariance of medv and rm is: 4.49345
The correlation of medv and rm is: 0.69536

Process finished with exit code 0
```

b)

Built-in R functions make life a whole lot easier. With R I am able to get all of this data and more with simple function calls that already exist. However, in C++ I have to take time to write custom functions to retrieve data that R returns seamlessly. Further, I have to ensure all of the custom functions I am writing are returning the proper values. Working with C++ is more tedious than working with R, from built-in functions to reading in a file, R seems to be the best bet

c)

Mean is the most commonly used method for finding the average in a dataset. **Median** is the value(s) exactly in the middle of a dataset. **Range** shows how far apart the “two extreme” (high and low) values are within a dataset. All three of these descriptive statistical measures help us better understand the data that we are given. We want to be able to understand the data that we are using to train a model, but also we need to understand this data to properly interpret the testing results of different models. For this, statistical methods are required.

(d)

Covariance measures how changes in one variable are associated with changes in a second variable, and Correlation is **Covariance** scaled [-1,1]. Covariance indicates the direction of the linear relationship between variables, while Correlation measures both the strength and direction of the linear relationship between two variables. Correlation has standardized values, while Covariance does not.