Ally de Vera

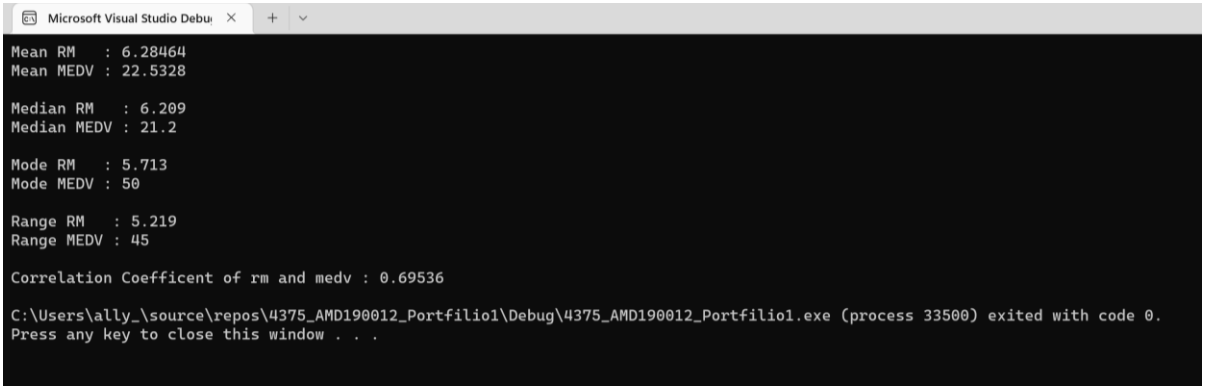
AMD190012

CS 4375

Dr. Karen Mazidi

Portfolio 1 : Data Exploration in C++

a.



b. I definitely took RStudio's built in functions for granted. I think the biggest thing didn't ever take into consideration is just how much the built-ins do, every call cycles through hundreds if not thousands of observations. While programming I did look at it a few times thinking "there has to be a better way to do that" but there was never a work around of looking at everything.

c.

Mean : the average, the mean can be affected by any outliers. In ML it's helpful to know what our most 'basic' observation would be.

Median : the middle, unaffected by outliers, in normal distribution models/ mostly normal distribution models, the median represents the center most point of the data. In ML the mean isn't always reliable due to outliers so if we were to replace N/A's for example, it may be more accurate to use the median rather than the mean, especially if there is a heavy skew in the model.

Mode : the most occurring value of a attribute. This is useful in ML because it is applicable to quantitative and qualitative variables, and allows us to recognize patterns among those most occurring days.

Range : the spread of the data, the max – min statistic. This is useful in ML because it can show just how much your dataset varies.

d.

Covariance : How changing variable a affects variable b, this is a big part of ML, this is where we look at our data, recognize patterns, and this is the foundation of making predictions.

Correlation : How reliable is it that changing a will also change b on a [-1,1] scale, meaning if I have a correlation of .23, I would be more hesitant to make predictions using whatever model that correlation is stemming from, but if I have a correlation of .98, we can say that we are confident in the predictions were making. This is important in ML because making predictions with a low correlation is unreliable.