NET-ID: JPM190001

ML DATA EXPLORATION

Jason P. McElhenney | 9/13/2022

```
-----------------------------------
Trying to open Boston.csv... Opened.
reading column labels
new len: 506
closing file Boston.csv... Closed.
-----------------------------------

# records: 506

Stats for rm
 sum    : 3180.03
 mean   : 6.28463
 median : 6.209
 range  : 5.219

Stats for medv
 sum    : 11401.6
 mean   : 22.5328
 median : 21.2
 range  : 45

 Covariance  = 4.49345

 Correlation = 0.69536

End.
```

While I didn't find implementing the common statistical functions sum, mean, median, range, covariance, and correlation terribly difficult, the R language is tailored toward statistical analysis and includes this functionality within its base structure, making it easier to use

The mean value within a data set represents the expected value of the set, and is a good midpoint to your data only when that data is relatively normal, while the median represents the value which divides the set into two equal parts, and is a good midpoint value when the data you are examining is skewed or includes extreme outliers. Range is simply the difference

between the largest value and the smallest value within the set, and it provides a way to determine the spread of that set easily. This is helpful as it gives you the largest potential change between all possible values.

Covariance between two variables gives us a measure of how different these variables are, which is extremely useful as a cost function within a neural network style machine learning system. However, since this covariance can easily become extremely large and unpredictable in domain size, it is helpful for us to normalize this value using the product of the two variable distribution's standard deviations as a divisor to their covariance. This gives us the correlation with a domain of only [-1,1], which is a perfect scalar range.