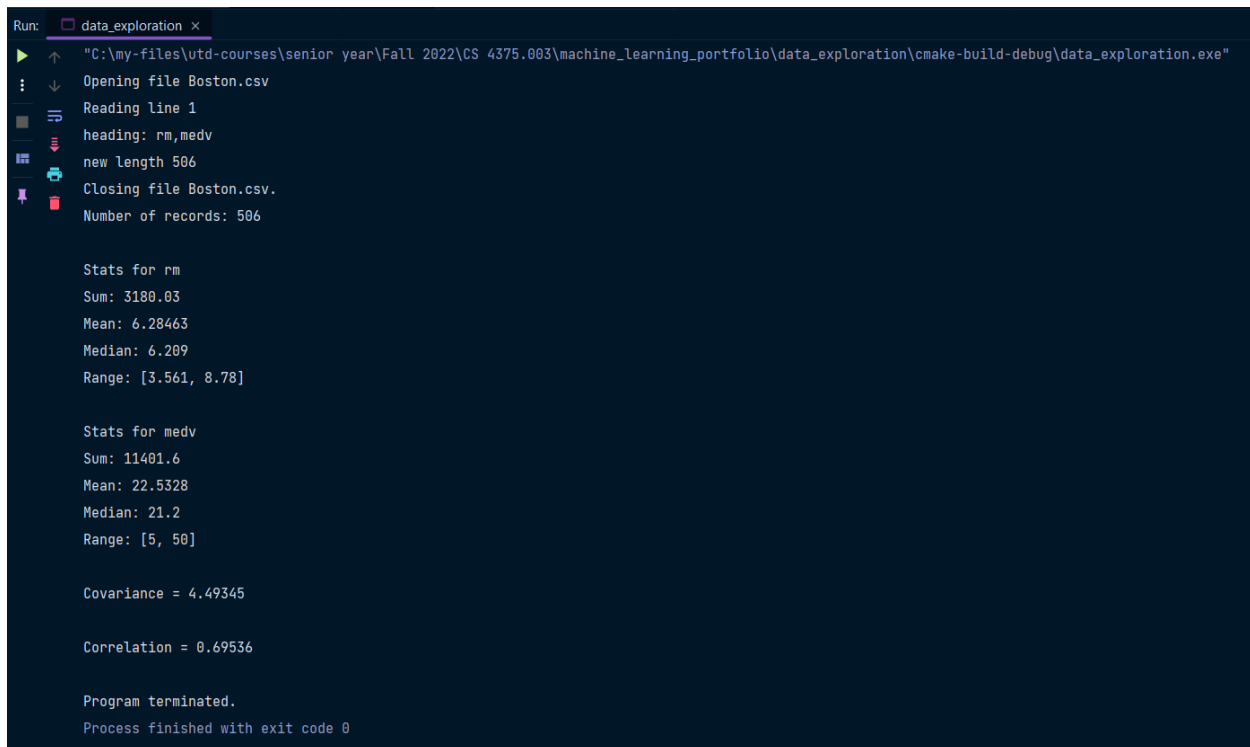# Data Exploration

By: Aditi Chaudhari

In the file titled HW1_data_exploration.cpp, I wrote functions in C++ to calculate the sum, mean, median, and range of each column of data stored in the Boston.csv file. Then, I wrote more functions to calculate the covariance and Pearson correlation coefficient between the two columns of data stored in the Boston.csv file. Here is a sample run of my code:

```
Run:    □ data_exploration ×
▶  ↑   "C:\my-files\utd-courses\senior year\Fall 2022\CS 4375.003\machine_learning_portfolio\data_exploration\cmake-build-debug\data_exploration.exe"
:  ↓   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: 3180.03
       Mean: 6.28463
       Median: 6.209
       Range: [3.561, 8.78]

       Stats for medv
       Sum: 11401.6
       Mean: 22.5328
       Median: 21.2
       Range: [5, 50]

       Covariance = 4.49345

       Correlation = 0.69536

       Program terminated.
       Process finished with exit code 0
```

Coding these six functions in C++ helped me gain a better understanding of how these built-in functions are used in R, especially when it came to the covariance and the correlation statistic. Being able to call these functions in R is incredibly convenient, but it is difficult to understand what actually goes on behind the scene. Prior to coding the sum(), mean(), median(), and range() functions, I was already familiar with summation, mean, median, and range, but coding these functions taught me how to work with and iterate through vectors in C++. Coding the covar() and cor() functions helped me understand the difference between the covariance statistic and the Pearson correlation coefficient, which will definitely help me in my journey in learning R.

Mean, median, and range are incredibly to data exploration prior to machine learning because they help describe the variability of the data (or how spread out the data is). Mean is the sum of all of the data points divided by the total number of data points. Median is the midpoint value of all of the data points sorted in increasing order. Finally, the range identifies the minimum and maximum value in the data. The mean, median, and range can identify potential outliers in the data. It is important to clean up

outliers in the data prior to applying machine learning algorithms in order to make the data more uniform.

Covariance and the Pearson correlation coefficient are also both important to machine learning. Covariance is a numerical measure of how changes in one variable can be associated to changes in another variable. The Pearson correlation coefficient is a statistic between -1 and +1 that gives insight into how strongly two variables are linearly correlated. A Pearson correlation coefficient of -1 would imply a strong negative linear correlation between two variables, whereas a Pearson correlation coefficient of +1 would imply a strong positive linear correlation between two variables. A Pearson correlation coefficient of 0 indicates that there is no correlation between the two variables. The covariance statistic and the Pearson correlation coefficient are both useful to the field of machine learning because they give insight into whether using a linear regression model would be a good fit for a data set or not.