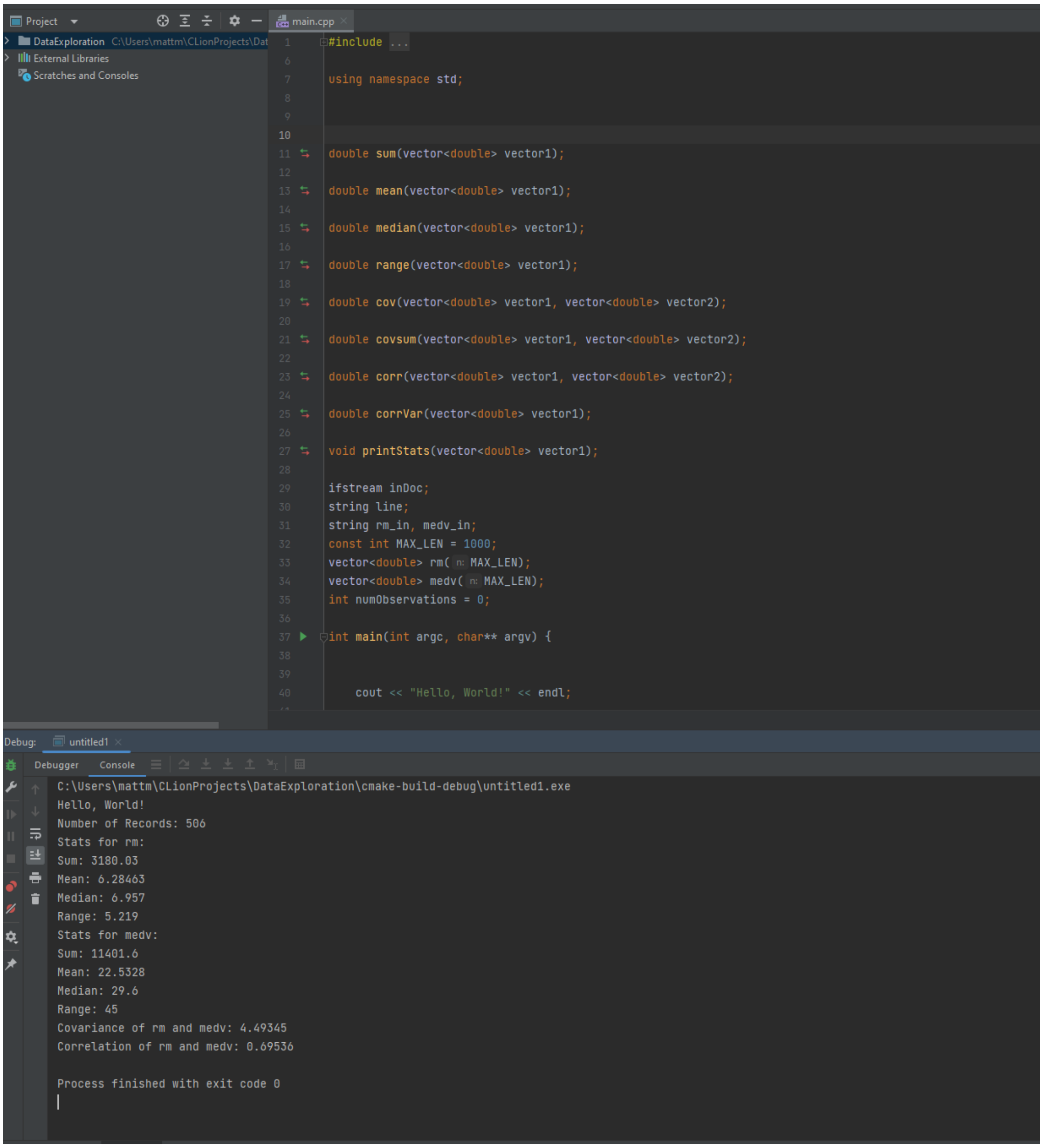
Matthew McCoy

CS 4375.004

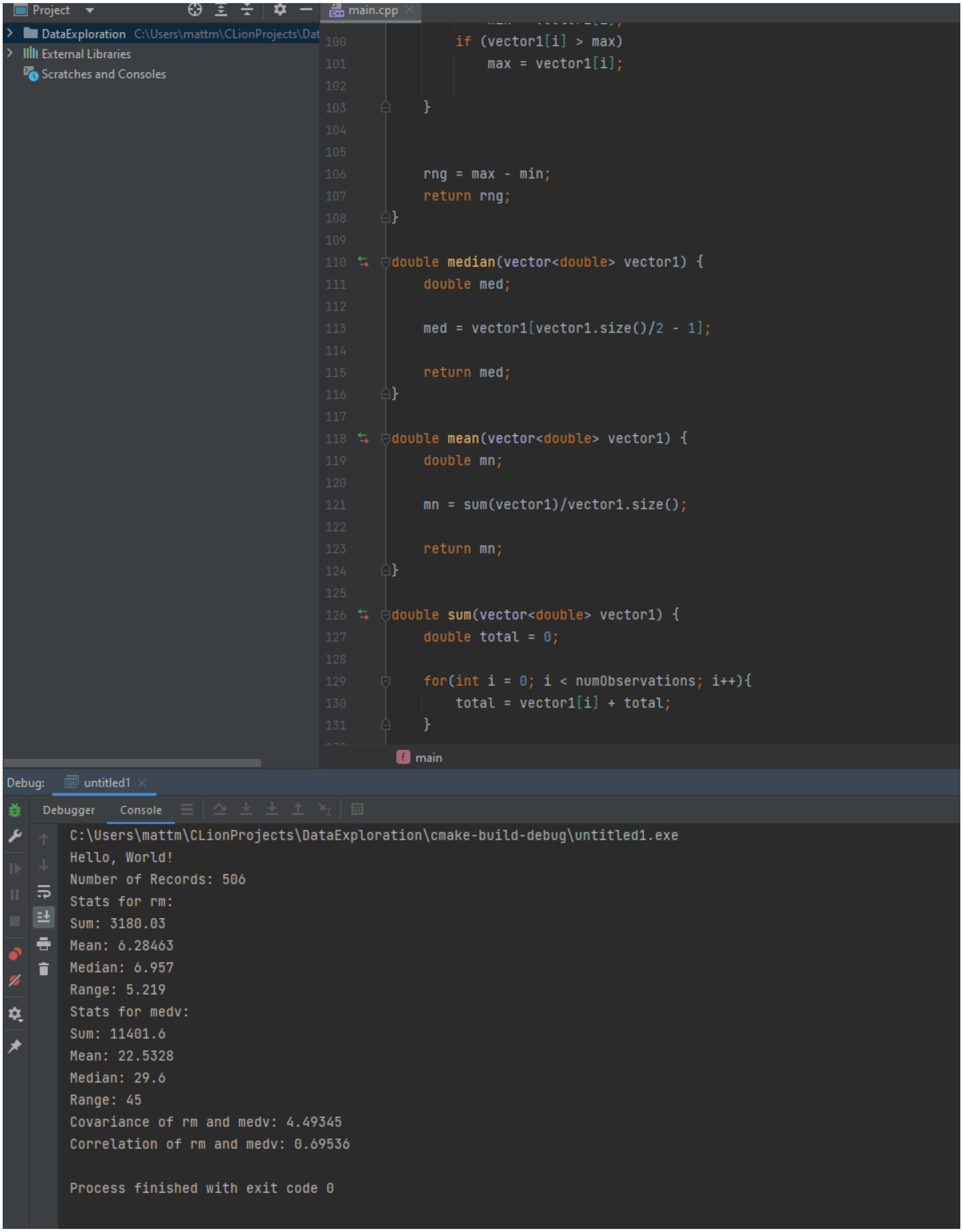
Portfolio Component 1: Data Exploration

a) Runs of the Code:



```cpp
#include ...

using namespace std;



double sum(vector<double> vector1);

double mean(vector<double> vector1);

double median(vector<double> vector1);

double range(vector<double> vector1);

double cov(vector<double> vector1, vector<double> vector2);

double covsum(vector<double> vector1, vector<double> vector2);

double corr(vector<double> vector1, vector<double> vector2);

double corrVar(vector<double> vector1);

void printStats(vector<double> vector1);

ifstream inDoc;
string line;
string rm_in, medv_in;
const int MAX_LEN = 1000;
vector<double> rm( n: MAX_LEN);
vector<double> medv( n: MAX_LEN);
int numObservations = 0;

int main(int argc, char** argv) {


    cout << "Hello, World!" << endl;
```

```
C:\Users\mattm\CLionProjects\DataExploration\cmake-build-debug\untitled1.exe
Hello, World!
Number of Records: 506
Stats for rm:
Sum: 3180.03
Mean: 6.28463
Median: 6.957
Range: 5.219
Stats for medv:
Sum: 11401.6
Mean: 22.5328
Median: 29.6
Range: 45
Covariance of rm and medv: 4.49345
Correlation of rm and medv: 0.69536

Process finished with exit code 0
```



```
        if (vector1[i] > max)
            max = vector1[i];

    }


    rng = max - min;
    return rng;
}

double median(vector<double> vector1) {
    double med;

    med = vector1[vector1.size()/2 - 1];

    return med;
}

double mean(vector<double> vector1) {
    double mn;

    mn = sum(vector1)/vector1.size();

    return mn;
}

double sum(vector<double> vector1) {
    double total = 0;

    for(int i = 0; i < numObservations; i++){
        total = vector1[i] + total;
    }
```

```
C:\Users\mattm\CLionProjects\DataExploration\cmake-build-debug\untitled1.exe
Hello, World!
Number of Records: 506
Stats for rm:
Sum: 3180.03
Mean: 6.28463
Median: 6.957
Range: 5.219
Stats for medv:
Sum: 11401.6
Mean: 22.5328
Median: 29.6
Range: 45
Covariance of rm and medv: 4.49345
Correlation of rm and medv: 0.69536

Process finished with exit code 0
```

b) Coding has always been a bit easier for me since I like to understand how something is working. That has always helped me create a solid foundation for something larger. That being said, RStudio has made using functions a different learning experience and helped to visualize comparisons quickly. The graphs alone would be a choir to code.

c) Mean, median and range are some of the most fundamental statistical measures in mathematics and statistics. They allow us to get a picture of a large data set and its patterns that we would otherwise have trouble recognizing. Helping us see first of all if the information is even viable to be used as a meaningful characteristic of the observed. This is a first step in deciding what to teach in machine learning.

d) Covariance shows us how much one set of data changes in relation to another set. In our assignment we found a covariance of 4.49345. This means there is a positive covariance, when one element increases its counterpart also increases. Correlation tells us how they are linearly related the data sets are and in assignment 1 we have a positive correlation of 0.69536. Machine learning is heavily dependent on making educated guesses based on similar situations and the outcomes they had. Showing a correlation between information will allow the machine to tie two pieces of information together to have more data sets to pull from when making decisions.