

Portfolio Component 1: Data Exploration

Write a short document:

a. copy/paste runs of your code showing the output

```
#include <iostream>
#include <vector>
#include <fstream>
#include <string>
#include <algorithm>

using namespace std;

double sum(vector<double>);
double mean(vector<double>);
double median(vector<double>);
vector<double> range(vector<double>);
double covariance(vector<double>, vector<double>);
double correlation(vector<double>, vector<double>);
double calculateStandardDeviation(vector<double>);
void display(vector<double>);

int main() {
    ifstream inFS;
    string line;
    string rm_in, medv_in;
    const int MAX_LEN = 1000;
    vector<double> rm(MAX_LEN);
    vector<double> medv(MAX_LEN);

    cout << "Opening file Boston.csv" << endl;

    inFS.open("Boston.csv");

    if (!inFS.is_open()) {
        cout << "Could not open file" << endl;
        return 1;
    }

    cout << "Reading line 1" << endl;
    getline(inFS, line);

    cout << "Heading " << line << endl;
```

```
int numberOfObservation = 0;
while (inFS.good()) {
    getline(inFS, rm_in, ',');
    getline(inFS, medv_in, '\n');
    rm.at(numberOfObservation) = stof(rm_in);
    medv.at(numberOfObservation) = stof(medv_in);

    numberOfObservation++;
}
rm.resize(numberOfObservation);
medv.resize(numberOfObservation);

cout << "New Line Size = " << numberOfObservation << endl;

cout << "Closing file Boston.csv" << endl;
inFS.close();

cout << "***** Stats of rm*****" << endl;
display(rm);
cout << "*****" << endl << endl;

cout << "***** Stats of medv*****" << endl;
display(medv);
cout << "*****" << endl << endl;

cout << "Covariance = " << covariance(rm, medv) << endl;

cout << "\nCorrelation = " << correlation(rm, medv) << endl;

cout << "\n Program terminated." << endl;

return 0;
}

double sum(vector<double> temp) {
    double sum = 0.0;
    for (int i = 0; i < temp.size(); i++) {
        sum += temp.at(i);
    }
    return sum;
}

double mean(vector<double> temp) {
    double avg = 0.0;
    int observation = temp.size();
    double sumofVector = sum(temp);
    return (double)sumofVector / (double)observation;
}
```

```
}

double median(vector<double> temp) {
    double median = 0.0;
    int size = temp.size();
    sort(temp.begin(), temp.end());
    if (size % 2 == 0) {
        return (double)(temp.at((size - 1) / 2) + temp.at(size / 2)) / 2.0;
    }else{
        return (double)temp.at(size / 2);
    }
}

vector<double> range(vector<double> temp) {
    vector<double> result;
    double left = (double)*min_element(temp.begin(), temp.end());
    result.push_back(left);
    double right = (double)*max_element(temp.begin(), temp.end());
    result.push_back(right);
    return result;
}

double covariance(vector<double> vect1, vector<double> vect2) {
    double sum = 0;
    double x_mean = mean(vect1);
    double y_mean = mean(vect2);

    for (int i = 0; i < vect1.size(); i++) {
        sum = (double)sum + (vect1.at(i) - x_mean)*(vect2.at(i) - y_mean)/(double)(vect1.size());
    }
    return sum;
}

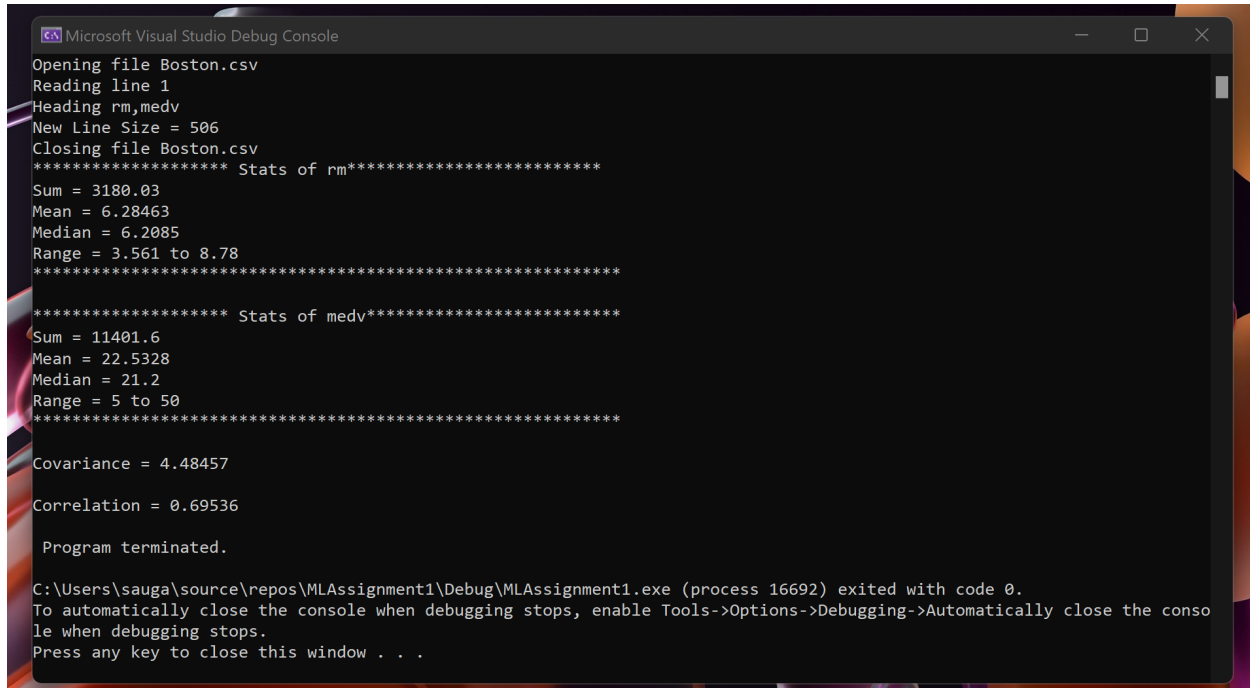
double correlation(vector<double> vect1, vector<double> vect2) {
    double standard_deviation_X = calculateStandardDeviation(vect1);
    double standard_deviation_Y = calculateStandardDeviation(vect2);
    return (double)covariance(vect1, vect2) / (double)(standard_deviation_X * standard_deviation_Y);
}

double calculateStandardDeviation(vector<double> temp) {
    double standardDeviation = 0.0;
    for (int i = 0; i < temp.size(); i++) {
        standardDeviation += pow(temp.at(i) - mean(temp), 2);
    }
    return (double)sqrt(standardDeviation / temp.size());
}

void display(vector<double>temp) {
```

```
cout << "Sum = " << sum(temp) << endl;
cout << "Mean = " << mean(temp) << endl;
cout << "Median = " << median(temp) << endl;
vector<double> rangeofVector = range(temp);
cout << "Range = " << rangeofVector.at(0) << " to " << rangeofVector.at(1) << endl;

}
```



The screenshot shows the Microsoft Visual Studio Debug Console window. The output text is as follows:

```
Opening file Boston.csv
Reading line 1
Heading rm,medv
New Line Size = 506
Closing file Boston.csv
***** Stats of rm*****
Sum = 3180.03
Mean = 6.28463
Median = 6.2085
Range = 3.561 to 8.78
*****
***** Stats of medv*****
Sum = 11401.6
Mean = 22.5328
Median = 21.2
Range = 5 to 50
*****
Covariance = 4.48457
Correlation = 0.69536

Program terminated.

C:\Users\sauga\source\repos\MLAssignment1\Debug\MLAssignment1.exe (process 16692) exited with code 0.
To automatically close the console when debugging stops, enable Tools->Options->Debugging->Automatically close the console when debugging stops.
Press any key to close this window . . .
```

b. describing your experience using built-in functions in R versus coding your functions in C++.

Built-in functions are easier to use because we do not need to worry about the inside mathematics behind them. While coding my own function, I need to care for mathematics and ensure the formula is correct. Since that mathematical function is needed in the maximum problem of machine learning, it is easier for further use.

c. describes the descriptive statistical measures mean, median, and range, and how these values might be useful in data exploration prior to machine learning.

Mean is an average of all the terms which is calculated by adding all the data and dividing it by number of observations.

Median is a middle number in a sorted data set. The median is the same as the mean for normal distribution.

Mode is a number which appears frequently in a data set.

Range is a difference between smallest number and largest number in a data set.

These values are useful in data exploration prior to machine learning so that we can find the central tendency or variability of data. By knowing variability of a data, we can know the dispersion of data. It also gives information about how data is distributed. So, it is very important for analyzing data to increase consistency.

d. describes the covariance and correlation statistics, and what information they give about two attributes. How might this information be useful in machine learning?

Covariance is a statistical term in which it shows the change in one variable reflects on other variables. It shows either a direct relation or inverse relation between variables.

Correlation is a measure of how changes in one variable are associated with changes in a second variable; like covariance but scaled to be in the range $[-1, +1]$. Correlation is a statistical term which describes how the two variables are related to each other. With covariance and correlation, If two variables are correlated to each other we can predict one variable from another variable.