

Francisco Trejo

2/4/2023

CS 4375.004

Karen Mazidi

Portfolio Component 1: Data Exploration

```
1 // Francisco Trejo FDT200000
2 // CS4375.004 with Karen Mazidi
3 // Portfolio Component 1: Data Exploration
4 #include <iostream>
5 #include <fstream>
6 #include <vector>
7 #include <string>
8 #include <algorithm>
9 #include <numeric>
10
11 using namespace std;
12
13 void printStats(const vector<double> &data) {
14     double sum = 0;
15     double mean = 0;
16     double range = 0;
17     double median = 0;
18
19     for (int i = 0; i < data.size(); i++) {
20         sum += data[i];
21     }
22     mean = sum / data.size();
23
24     range = data.back() - data.front();
25
26     sort(data.begin(), data.end());
27     median = data[data.size() / 2];
28 }
```

Microsoft Visual Studio Debug Console

```
Opening file Boston.csv.
Reading line 1
heading: rm,medv
new length 506
Closing file Boston.csv.
Number of records: 506

Stats for rm
The sum is: 3180.03
The mean is: 6.28463
The range is: 3.561 - 8.78
The median is: 6.2085

Stats for medv
The sum is: 11401.6
The mean is: 22.5328
The range is: 5 - 50
The median is: 21.2

Covariance = 4.49345
Correlation = 0.69536

Program terminated.
C:\Users\Paco Trejo\source\repos\ConsoleApplication3\x64\Debug\ConsoleApplication3.exe (process 16448) 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 . . .

'ConsoleApplication3.exe' (Win32): Loaded 'C:\Windows\System32\kernel.appcore.dll'.
'ConsoleApplication3.exe' (Win32): Loaded 'C:\Windows\System32\msvcrt.dll'.
The thread 0x2bec has exited with code 0 (0x0).
The thread 0x46d8 has exited with code 0 (0x0).
```

```
117
118 int main(int argc, char** argv)
119 {
120     ifstream inFS; // Input file stream
121     string line;
122     string rm_in, medv;
123     const int MAX_LEN = 1000;
124     vector<double> rm(MAX_LEN);
125     vector<double> medv(MAX_LEN);
126     cout << "Opening file Boston.csv." << endl;
127     while (getline(inFS, line))
128     {
129         if (!inFS.is_open())
130             return 1;
131         cout << "Reading line 1" << endl;
132         heading: rm, medv
133         new length 506
134         cout << "Closing file Boston.csv." << endl;
135         Number of records: 506
136
137         Stats for rm
138         The sum is: 3180.03
139         The mean is: 6.28463
140         The range is: 3.561 - 8.78
141         The median is: 6.2085
142
143         Stats for medv
144         The sum is: 11401.6
145         The mean is: 22.5328
146         The range is: 5 - 50
147         The median is: 21.2
148
149         Covariance = 4.49345
150         Correlation = 0.69536
151
152         Program terminated.
153         C:\Users\Paco Trejo\source\repos\ConsoleApplication3\x64\Debug\ConsoleApplication3.exe (process 18720) exited with code 0.
154         To automatically close the console when debugging stops, enable Tools->Options->Debugging->Automatically close the console when debugging stops.
155         The thread 0x4404 has exited with code 0 (0x0). Press any key to close this window . . .
156         ConsoleApplication3.exe' (Win32): Loader
157         ConsoleApplication3.exe' (Win32): Loaded 'C:\Windows\System32\msvcrt.dll'.
158         The thread 0x3cd4 has exited with code 0 (0x0).
159         The thread 0x4258 has exited with code 0 (0x0).
160         The program '[18720] ConsoleApplication3.exe' has exited with code 0 (0x0).
```

I feel it was definitely harder to code it on C++ versus coding it in R. In C++ you have to write several functions to accomplish the same thing a few lines of R would do. R really doesn't seem like too much coding and those built in functions are easier to use and understand.

The mean gives you the average value of the elements of the vector. The range gives you the highest and lowest values of the vector and the median value gives you the value that is right in the middle of the sorted vector. These are useful in data exploration because it pretty much gives an overview of the data. The range gives you a finite boundary on where the data limits itself. The mean and median gives you a value that will most likely come close to or come up in that set of data.

The covariance and correlation is how much two attributes are related to each other and gives a numeric representation of their relationship. Covariance gives the direction and extent of the relationship while correlation gives the strength of that relationship. These are important in

machine learning because since both find relationships between two attributes then it can help find trends and reliability of those trends to predict future data.

https://ftrejo2013.github.io/Class_Portfolio/