Final Project: Data Analysis Report

Michael Surdek

Southern New Hampshire University

These data sets contain information regarding temperature (in degrees Celsius) and rainfall (in millimeters) in Great Britain and the United States for every month from January 1991 through December 2015. The United States data set contained a double the amount of observations, so I cross referenced and found out that each row was recorded twice. I simply removed the duplicate data before I continued with on to calculating the mean, minimum, and maximum of the columns in each data set. My first assessment from high level view is that Great Britain not only seems to have consistently higher rainfall, but also seems to have most of the highest amounts of rainfall.

By performing some basic calculations on this data, I was able to determine that Great Britain has a higher average temperature than the United States. However, the United States has both a lower minimum temperature and a higher maximum temperature, which means there is more variation when it comes to temperature in the United States. I was also able to determine that Great Britain has much higher amounts of rainfall. Both the average amount of rainfall per month and the maximum amount of rainfall in one month are significantly higher in Great Britain than in the United States.

The first thing I would do to analyze this data set further would be to use RStudio and group the data by the month column and perform similar calculations on the minimums, maximums, and averages for the two countries in different months. To do this, the data frame structure would be helpful. I would begin with two data frames containing the original data for each country. I would group by country and summarize the data in the data frame by creating additional columns for minimum, maximum, and mean of each month in each country. This would result in a 12 x 4 data frame for each country, and I could manipulate them further to combine the countries into one and compare the information directly. I would be interested to see if Great Britain has higher rainfall than the United States in all months of the year, or if they are similar in some months and have much higher rainfall totals in others.

Having experience with R coming into this course, I struggled more with the Command Line Interface module than any other. I found the syntax somewhat different than I am used to, and since I am not the most familiar with computers in general, I was sometimes confused on what I was supposed to be trying to accomplish. With that being said, I think it will be useful to me going forward. I think understanding how file storage on computers is organized and accessed is essential when working with data. In today’s industry of big data, there is value in the command line interface and those who can operate it efficiently, because companies have very complex file storage systems and they all do it in their own uniquely chaotic way.

Throughout this course, I enjoyed using and analyzing the sample data. Prior to this course, I felt like I had a decent understanding of the data analysis process, but after working through multiple data sets, I realized there is a lot more that goes into it. I believe I had an idea of the general data assessment, validation, discovery, and analysis steps, but I was not aware of how many smaller processes each of those aspects contain. I also learned how important it is to perform them each step in full and in order. Along the same lines, if I could have done one thing differently, I would have always had a pencil and paper by my side to write down notes and ideas. I realized how easy it is to think about something at one point and completely forget it when you move to the next step. I think using notes more diligently would allow me to retain valuable thoughts and ideas and to pursue them further down the line.

My process for source code management is to comment on every code chunk. At the very least, I try to include a comment that contains the basic purpose of that line of code. I think this will be important when line(s) of code become more complex and it can be tougher to understand their purpose. I also never hesitate to add comments with any thoughts that might be important to myself or anyone else reading the code in the future. I think, so far, I have always been able to include comments that are sufficient in communicating the intent of my code. Although I have never worked with others in a professional coding environment, most of my previous experience in coding has been self-learning through the internet, and I have seen firsthand how much easier it is to understand code when it is well commented.

With the data sets that we worked with in this project, I believe a database management system could be appropriate. I do not think it is necessary, because the data is not too complex, but I think a system could be created that would make it easier to turn raw temperature and rainfall data into actionable information. Our data sets contained the monthly average temperature and the monthly total rainfall, but this type of information is discovered daily. For this reason, a database management system could be created to turn the raw daily data into either the data sets that we first worked with, or into some of the data structures that we created and planned to create in the future. One benefit of database management systems is that they can be helpful on both simple and complex data.

Appendix A: Data Assessment

A screenshot of a cell phone

Description automatically generated

Appendix B: File Organization in Linux

A screenshot of a cell phone screen with text

Description automatically generated

Appendix C: Code for R program (File Output)

A screenshot of a cell phone

Description automatically generated

A close up of a piece of paper

Description automatically generated

A screenshot of a social media post

Description automatically generated

A screenshot of a cell phone

Description automatically generated