# **Final Project: Data Analysis Report**

**Matt Schnorr**

**Southern New Hampshire University**

**11/15/2020**

Findings

Two things initially stood out to me. One was regarding temperature. It appeared that Great Britain had a higher average temperature, but the United States had more extreme outliers, meaning there was greater variance in the possible range of temperatures. The second thing I noticed was about the rainfall. Great Britain’s average was almost double that of the USA, and the max rainfall was almost three times higher than the max of the United States.

Informed Conclusion

Based on my findings from the data, I can conclude several things. First, Great Britain has less variance in their temperature, meaning that you would expect the weather to fall within a much smaller range of values than the United States. That makes sense to me, as the United States covers more latitude (or distance from the equator), so temperatures will vary more.

I can also conclude that Great Britain is a much wetter place than the United States as its average rainfall and max rainfall were much higher than the USA over the 25 year period, which is enough of a measurement to confidently reach said conclusion. It would be a better place to launch any product that involved rain gear.

Future Analysis

Although this project seemed relatively simple as it only involved a few separate metrics and countries, the tools used for measurement are very pliable and able to handle much larger, and more complex data sets. Vectors, matrices, and data frames in R can provide in depth statistical analysis of multivariate sets.

Command Line Interface

The command line interface will be vital to me as a data scientist, which is my desired role upon completion of my studies. While I will not use it much in my current role, it is necessary to be interactive with a variety of operating systems when pulling in and analyzing data. It is diverse and allows you to use multiple sources.

Data Assessment and Analysis Process

My experience in analyzing data allows me to look at things from a high level, then narrow down to search for more specific stories and outliers. I am then able to tie my findings into my knowledge of business and industry to be able to use the data to tell a story and plan future actions and make informed business decisions. The data can provide a roadmap for where you want to go if you are able to accurately mine, clean, and analyze the information.

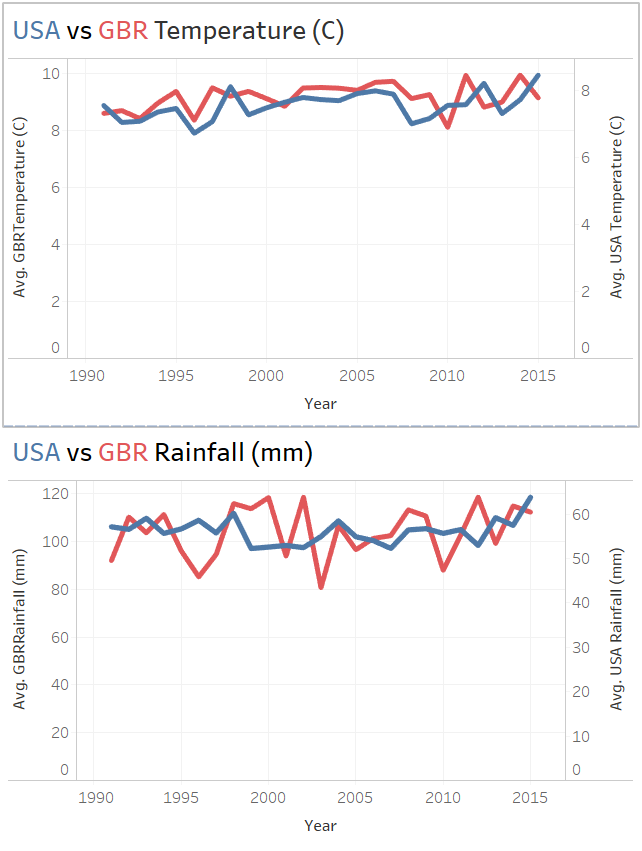
Source Code Management

My intent was to create vectors, matrices, tables, and ultimately data frames in R from the raw data to see if there were stories to be told with simple minimum, maximum, and average values in each of the four columns. I cleaned the data in Excel first into one worksheet and saved it as one CSV so it was easier to import and work with as it was all from one source. The end table matched what I produced in Excel, so I knew that the programming was correct as it was double verified.

Relational Database Management Systems

A database management system can handle this data with no problem. However, I would argue that it would not necessarily be needed in this case with the direction of just finding MIN, AVG, and MAX with such simple data can easily be handled by Excel in a much shorter time. A good data scientist needs to understand all the tools, and when to apply them. The best course is to make things as simple as possible, but not any simpler. Going through R was much more time consuming. However, the foundation is now set for much more complex data, research, and tasks.

DATA VISUALS



Appendix A: Data Assessment

EXCEL

TOP 10 ROWS - Example

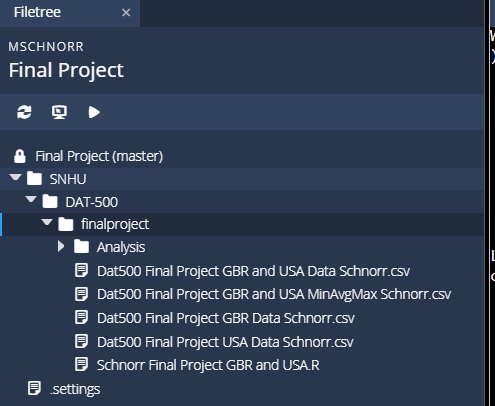
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Year | Month | Country | Temperature (C) | Rainfall (mm) | Country | Temperature (C) | Rainfall (mm) |
| 1991 | 1 | GBR | 2.56628 | 123.511 | USA | -6.0482 | 45.2501 |
| 1991 | 2 | GBR | 1.68146 | 78.1605 | USA | -1.5655 | 34.5711 |
| 1991 | 3 | GBR | 6.69328 | 102.646 | USA | 0.93979 | 68.595 |
| 1991 | 4 | GBR | 7.1204 | 96.5587 | USA | 7.20034 | 63.5456 |
| 1991 | 5 | GBR | 10.0956 | 23.6001 | USA | 13.4703 | 73.3747 |
| 1991 | 6 | GBR | 11.0998 | 105.637 | USA | 17.9005 | 60.212 |
| 1991 | 7 | GBR | 15.9509 | 81.6537 | USA | 20.0899 | 64.8082 |
| 1991 | 8 | GBR | 15.6379 | 46.3049 | USA | 19.2929 | 59.5409 |
| 1991 | 9 | GBR | 13.131 | 90.5544 | USA | 15.0592 | 58.4347 |
| 1991 | 10 | GBR | 9.1101 | 116.375 | USA | 8.28441 | 48.5126 |

MEASUREMENTS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | GBR Temp | GBR Rainfall | USA Temp | USA Rainfall |
| Min | -0.1819 | 19.4244 | -6.8235 | 27.4009 |
| Avg | 9.1550483 | 104.003184 | 7.554503 | 56.1112037 |
| Max | 17.3942 | 233.083 | 21.4416 | 85.1186 |

I ran the MIN, AVERAGE, and MAX functions in Excel to the total range of each of the measurements of both temperature (in Celsius) and rainfall (in mm). I did this for both Great Britain and the United States. It showed the minimum, average, and maximum for each category over a comprehensive 25-year period from 1991 to 2015.

Appendix B: File Organization in Linux



Appendix C: Code for R program (File Output)

R Programming

#Read in Files

library(readr)

#Adjust the file name and file path below for each of the 2 files. Use the same file read code from Milestone Two

GBR <- read\_csv("~/workspace/SNHU/DAT-500/finalproject/Dat500 Final Project GBR Data Schnorr.csv")

View(GBR)

USA <- read\_csv("~/workspace/SNHU/DAT-500/finalproject/Dat500 Final Project USA Data Schnorr.csv")

View(USA)

GBR\_USA <- read\_csv("~/workspace/SNHU/DAT-500/finalproject/Dat500 Final Project GBR and USA Data Schnorr.csv")

View(GBR\_USA)

GBR\_USA\_CALC <- read\_csv("~/workspace/SNHU/DAT-500/finalproject/Dat500 Final Project GBR and USA MinAvgMax Schnorr.csv")

View(GBR\_USA\_CALC)

#Create assignments for Min, Max, and Avg Temperatures and Rainfalls for both GBR and USA

GBRTempMin<-min(GBR$`Temperature (C)`)

GBRTempMax<-max(GBR$`Temperature (C)`)

GBRTempAvg<-mean(GBR$`Temperature (C)`)

GBRRainfallMin<-min(GBR$`Rainfall (mm)`)

GBRRainfallMax<-max(GBR$`Rainfall (mm)`)

GBRRainfallAvg<-mean(GBR$`Rainfall (mm)`)

USATempMin<-min(USA$`Temperature (C)`)

USATempMax<-max(USA$`Temperature (C)`)

USATempAvg<-mean(USA$`Temperature (C)`)

USARainfallMin<-min(USA$`Rainfall (mm)`)

USARainfallMax<-max(USA$`Rainfall (mm)`)

USARainfallAvg<-mean(USA$`Rainfall (mm)`)

#Create Vectors for GBR and USA

GBRV<-c((GBRTempMin),(GBRTempMax),(GBRTempAvg),(GBRRainfallMin),(GBRRainfallMax),(GBRRainfallAvg))

GBRV

USAV<-c((USATempMin),(USATempMax),(USATempAvg),(USARainfallMin),(USARainfallMax),(USARainfallAvg))

USAV

#Create Matrices for GBR and USA

GBRM<-matrix(c((GBRTempMin),(GBRTempMax),(GBRTempAvg),(GBRRainfallMin),(GBRRainfallMax),(GBRRainfallAvg)), nrow=3, ncol=2, byrow = FALSE)

GBRM

namedMat<-GBRM #copy matrix for the purpose of adding a label

dimnames(GBRM)=list(c('GBR Min','GBR Max','GBR Avg'),c('Temperature (C)', 'Rainfall (mm)')) #name the cols and rows

GBRM #print matrix with labels

USAM<-matrix(c((USATempMin),(USATempMax),(USATempAvg),(USARainfallMin),(USARainfallMax),(USARainfallAvg)), nrow=3, ncol=2, byrow = FALSE)

USAM

namedMat<-USAM

dimnames(USAM)=list(c('USA Min','USA Max','USA Avg'),c('Temperature (C)', 'Rainfall (mm)'))

USAM

GBRandUSA<-rbind((GBRM),(USAM))

GBRandUSA

#Create data frames for GBR, USA, and Both

dfGBR<-data.frame(GBRM)

dfGBR

colnames(dfGBR)<-c("Temp", "Rainfall")

dfGBR

dfUSA<-data.frame(USAM)

dfUSA

colnames(dfUSA)<-c("Temp", "Rainfall")

dfUSA

dfGBRUSA<-data.frame(GBRM-USAM)

dfGBRUSA

colnames(dfGBRUSA)<-c("Temp", "Rainfall")

rownames(dfGBRUSA)<-c("GBR\_USA Difference Min", "GBR\_USA Difference Max", "GBR\_USA Difference Avg")

dfGBRUSA

GBRandUSAall<-array(c((GBRM),(USAM),(GBRM-USAM)), c(nrows = 9,ncol = 2))

GBRandUSAall

colnames(GBRandUSAall)<-c("Temp","Rainfall")

rownames(GBRandUSAall)<-c("GBR Min","GBR Max","GBR Avg","USA Min","USA Max","USA Avg","GBR\_USA Difference Min", "GBR\_USA Difference Max", "GBR\_USA Difference Avg")

GBRandUSAall

R TABLE

|  | **X1** | **GBR Temp** | **GBR Rainfall** | **USA Temp** | **USA Rainfall** |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| **1** | Min | -0.181900 | 19.4244 | -6.823500 | 27.4009 |
| **2** | Avg | 9.155048 | 104.0032 | 7.554503 | 56.1112 |
| **3** | Max | 17.394200 | 233.0830 | 21.441600 | 85.1186 |

Showing 1 to 3 of 3 entries