Anthony Dougher

Dat-500

2/9/2024

Final Project

**Data Analysis Report:**

Below, we’ve ran the measures of central tendency and organized them into multiple tables. My first assumption was that the higher the temperature, the more rainfall. Although this would appear true for the United States, it seems as though Great Britain shares a similar temperature average but significantly more rainfall. My informed conclusion after analyzing both data sets is that rainfall is independent of temperature. I came to this conclusion due to having two sets of data to analyze. If I had just analyzed USA’s data I would stick with this conclusion, but the average temperature similar to USA’s and significantly higher rainfall disproves this. If I had additional data, I would leverage the data frame more than anything. Let’s say we get another country to analyze, I would create another matrix and add it as another column of the data frame. That way we could see all the averages, minimums, and maximums right next to each other.

**Reflection:**

The command line interface was extremely intimidating when I began this course. Slowly but surely, I became more comfortable with it. The codio assignments got us familiar with the work but the milestones really solidified and built my confidence with the CLI. The CLI has value in the industry due to how quick it is to analyze data sets and compare them. Although I believe speed is the primary benefit, ease of use is a big plus. Of course, it takes time to learn but it is more organized than using something like excel. I don’t think I would of done anything differently when analyzing the data. I would like some examples of previous data projects and some of the insights they used to drive business decisions. I feel green when it comes to drawing conclusions from data sets but I’m confident that will change over time. I believe my comments effectively communicate the intentions of my code but for the most part it’s self-explanatory. I understand comments are paramount when performing more complex functions but for building basic matrices and data frames I find it hard to add anything additional. It’s good to get into the habit of adding comments and that is what we are doing here. Every time I stepped away for a few minutes I would lose connection with Rstudio and I would have to rerun all the code. That would be the extent of the benefit of the database management system for this project as the scale of the project is not big enough in my opinion. I believe its full potential would be on a larger project with multiple analysts involved.

Appendices

**Section I:**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Section II:**

I had made a mistake in Linux while transferring GBR data to the Analysis folder and accidentally named it the same as the USA data. During the process the GBR file was lost and I was unable to recover it until I loaded the excel file into Rstudio. I lost the commands in Linux as well trying to get the data back so a quick rundown of how to move the file is as follows:

Use mkdir Analysis to create analysis folder

cd to change directories until the file in the SNHU dat500 is reached

Use mv command to transfer file into Analysis directory, then add the new name to the end of the code to rename

A screenshot of a computer

Description automatically generated

**Summary Functions**

A screenshot of a computer

Description automatically generated

The benefit of the summary function in Rstudio is that it is much quicker and gives the measures of central tendency all at once as opposed to having to run functions one by one in excel. It all provides us with a concise vector for the data. If we were presenting data to a third party, the summary function would be easy for them to visualize. After running the summary function we can see that it confirms my calculations in excel. It also enforces my hypothesis that higher temperatures are associated with more rainfall.

**Variables**

**A screenshot of a computer

Description automatically generated**

**Section III:**

**A screenshot of a computer

Description automatically generated**

**R script:**

> summary(USA\_data)

\tYear Month Country Temperature (C) Rainfall (mm)

Min. :1991 Min. : 1.00 Length:600 Min. :-6.8235 Min. :27.40

1st Qu.:1997 1st Qu.: 3.75 Class :character 1st Qu.:-0.7086 1st Qu.:47.64

Median :2003 Median : 6.50 Mode :character Median : 7.3800 Median :56.11

Mean :2003 Mean : 6.50 Mean : 7.5545 Mean :56.11

3rd Qu.:2009 3rd Qu.: 9.25 3rd Qu.:16.6396 3rd Qu.:63.76

Max. :2015 Max. :12.00 Max. :21.4416 Max. :85.12

> summary(GBR\_data)

Year Month Country Temperature (C) Rainfall (mm)

Min. :1991 Min. : 1.00 Length:300 Min. :-0.1819 Min. : 19.42

1st Qu.:1997 1st Qu.: 3.75 Class :character 1st Qu.: 5.2719 1st Qu.: 73.67

Median :2003 Median : 6.50 Mode :character Median : 8.7781 Median : 97.09

Mean :2003 Mean : 6.50 Mean : 9.1550 Mean :104.00

3rd Qu.:2009 3rd Qu.: 9.25 3rd Qu.:13.0889 3rd Qu.:126.41

Max. :2015 Max. :12.00 Max. :17.3942 Max. :233.08

> USAtempMin<-min(USA\_data$`Temperature (C)`)

> USAtempMax<-max(USA\_data$`Temperature (C)`)

> USAtempAvg<-mean(USA\_data$`Temperature (C)`)

>

> USArainMin<-min(USA\_data$`Rainfall (mm)`)

> USArainMax<-max(USA\_data$`Rainfall (mm)`)

> USArainAvg<-mean(USA\_data$`Rainfall (mm)`)

>

> GBRtempMin<-min(GBR\_data$`Temperature (C)`)

> GBRtempMax<-max(GBR\_data$`Temperature (C)`)

> GBRtempAvg<-mean(GBR\_data$`Temperature (C)`)

>

> GBRrainMin<-min(GBR\_data$`Rainfall (mm)`)

> GBRrainMax<-max(GBR\_data$`Rainfall (mm)`)

> GBRrainAvg<-mean(GBR\_data$`Rainfall (mm)`)

> USAv <- c(USArainAvg, USArainMax, USArainMin, USAtempAvg, USAtempMax, USAtempMin)

> # we use c() to ccombine data points into a vector

> # a vector is a list of values, and the first step in building a data structure

> GBRv <- c(GBRrainAvg, GBRrainMax, GBRrainMin, GBRtempAvg, GBRtempMax, GBRtempMin)

>

> #we use matrix function to turn a vector into a list with columns and rows

> USAm = matrix(c(USArainAvg, USArainMax, USArainMin, USAtempAvg, USAtempMax, USAtempMin), nrow = 6, ncol = 1, byrow = TRUE)

>

>

> #we use rownames to change the row names and label the data

> rownames(USAm) <- c("rainAvg", "rainMax", "rainMin", "tempAvg", "tempMax", "tempMin")

>

> colnames(USAm) <- c("USA")

>

>

> GBRm = matrix(c(GBRrainAvg, GBRrainMax, GBRrainMin, GBRtempAvg, GBRtempMax, GBRtempMin), nrow = 6, ncol = 1, byrow = TRUE)

>

>

> colnames(GBRm) <- c("GBR")

>

>

> #we use cbind to combine both matrices into one

>

> USAmGBRm <- cbind(USAm,GBRm)

>

>

> #we use data.frame to build a data fram and compare the differences in the measures of central tendency between USA and GBR

>

> DF <- data.frame(USAmGBRm, USAm - GBRm)

>

> print(DF)

USA GBR USA.1

rainAvg 56.111204 104.003184 -47.891981

rainMax 85.118600 233.083000 -147.964400

rainMin 27.400900 19.424400 7.976500

tempAvg 7.554503 9.155048 -1.600546

tempMax 21.441600 17.394200 4.047400

tempMin -6.823500 -0.181900 -6.641600