Applied Statistical Programming - Projects

2/23/2022

Write the R code to answer the following questions. Write the code, and then show what the computer returns when that code is run. Thoroughly comment your solutions.

You have until the beginning of class 2/28 at 10:00am to complete the assignment below. You may use R, but not any online R documentation. Submit the Rmarkdown and the knitted PDF to Canvas. Have one group member submit the activity with all group members listed at the top.

Project Management

In this exercise, you will plot a Twitter user's activity for a single day distinguishing between their novel content and their retweets. You will load a data set, modify it, and generate figures in a project environment. Download the `Tweets.csv` file from Canvas, and complete the following tasks in your project environment.

1. Subset the data to the user `a_silberberg` for tweets occurring on November 4, 2015.
2. Write the subset data as a CSV file into a `Data` sub-folder of your project.
3. Generate a `plot()` where the Y-axis is a count of Twitter activity and the X-axis is the time of day the activity took place. Use the `IsRetweet` variable to distinguish whether the activity was a retweet or new content generated by the user. Your plot must have a title, labeled axes, and a legend for the two types of Twitter activity ("Tweet" versus "Retweet").
4. Write the plot as a PDF to a `Figures` sub-folder of your project.

You will need to generate count variables to make this plot. You will also need to use the `lubridate` package to parse time from the full date-time stamp. Assuming you have already subsetted the data to only focus on the user `a_silberberg`, you can create a new variable in the data for the time with the following code. You will also need to wrap `newtime` in a `as.POSIXct()` statement so R knows how to handle the time data.

```
# Loading data
tweets <- read.csv("Tweets.csv")
only_a_Silberberg <- tweets[tweets$ScreenName == "a_silberberg",]

## This came with the rmd folder; it's not my own
# Remove eval=FALSE to have this code block run.
library(lubridate)
# Assume the "only_a_Silberberg" subsetted data already exists.
dates <- as.POSIXct(only_a_Silberberg$CreatedTime, format = "%Y-%m-%d %H:%M:%S")
# Extra the day from the full time stamp
days <- format(dates, format = "%Y-%m-%d")
# Subset the data again so only tweets on November 4 are included.
newData <- only_a_Silberberg[which(days == "2015-11-04"),]
# Make a new variable in the data that is only the time of the tweet, not the day
```

```
newData$newtime <- format(as.POSIXct(newData$CreatedTime, format = "%Y-%m-%d %H:%M:%S"),
                          format = "%H%M%S")

# outputting the new data
write.csv(newData, "./Data/newData.csv")

# making the plot
pdf(file="./Figures/tweetplot.pdf")

plot(x = as.numeric(newData$newtime),
     y = newData$RetweetCount + newData$FavoritesCount,
     col = factor(newData$IsRetweet),
     main = "This is my plot",
     sub = "retweets in RED; everything else is BLACK",
     xlab = "time (lol not really)",
     ylab = "activity (sorta)")

dev.off()
```