Applied Statistical Programming - Projects

Berta Diaz, Zion Little, Alma Velazquez

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.

Read in data, explore variables

```
tweets <- read.csv("Tweets.csv")
head(tweets$ScreenName)
```

```
## [1] "a_silberberg" "a_silberberg" "a_silberberg" "a_silberberg" "a_silberberg"
## [6] "a_silberberg"
```

```
head(tweets$CreatedTime)
```

```
## [1] "2018-04-28 22:26:48" "2018-04-28 22:13:38" "2018-04-28 19:42:14"
## [4] "2018-04-28 14:42:20" "2018-04-28 13:58:39" "2018-04-28 13:53:21"
```

```
length(unique(tweets$ScreenName))
```

```
## [1] 572
```

```r
only_a_Silberberg <- tweets[tweets$ScreenName == "a_silberberg", ]
```

Prepare dates

```r
# Remove eval=FALSE to have this code block run.
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```r
# Assume the "only_a_Silberberg" subsetted data already exists.
only_a_Silberberg$dates <- as.POSIXct(only_a_Silberberg$CreatedTime, format = "%Y-%m-%d %H:%M:%S")
# Extra the day from the full time stamp
only_a_Silberberg$days <- format(only_a_Silberberg$dates, format = "%Y-%m-%d")
# Subset the data again so only tweets on November 4 are included.
newData <- only_a_Silberberg[which(only_a_Silberberg$days == "2015-11-04"),]
# Make a new variable in the data that is only the time of the tweet, not the day
newData$newtime <- format(as.POSIXct(newData$CreatedTime, format = "%Y-%m-%d %H:%M:%S"),
                          format = "%H:%M:%S")
```

Write to CSV

```r
write.csv(newData, "Data/a_silberg_110415.csv")
```

Creating count variables (unused)

```r
# str(newData$IsRetweet)
#
# count_var <- function(i){
#   return(nrow(newData[newData$hour_tweeted == i,]))
# }
#
# hours <- c(0:24)
# counts <- as.vector(unlist(lapply(hours, count_var)))
#
# newData <- merge(newData, cbind(hours, counts), by.x = "hour_tweeted", by.y = "hours")
```

Plots and saving

```r
newData$hour_tweeted <- hour(newData$dates)
labs <- c(paste0(c(12, 1:11), "AM"),paste0(c(12, 1:11), "PM"))

pdf("plots/a_silberberg110415.pdf")
hist(x=newData[newData$IsRetweet==1,]$hour_tweeted, col=rgb(1,0,0,.3), border=NA,xaxt="none", main = pas
hist(x=newData[newData$IsRetweet==0,]$hour_tweeted, col=rgb(0,0,1,.3), border=NA, add=TRUE,xaxt="none",
axis(side=1, at=0:23, labels=labs, cex.axis=0.75)
legend("topright",box.lwd=0,title="Tweet Type",c("Retweet", "Original Tweet"), fill = c(rgb(1,0,0,.3), r
dev.off()
```

```
## pdf
```

```
##   2
```

Testing hist() outputs

```
nrow(newData[which(newData$hour_tweeted<=5 & newData$IsRetweet == 0), ])
```

```
## [1] 10
```