HW4 - Exploring Social Networks

Logan Bartels
October 25, 2020

Note about R scripts

Most of my R scripts were tested on a copy of the Week-03-InfoVis-R Google Collab notebook. The png versions of the graphs shown in this report were saved from the Collab notebook. Some lines of code found in the notebook were omitted in my final R scripts for brevity. The code can be found under the "HW4" section here:
https://colab.research.google.com/drive/1BVYc-LunOuLe4bI-1S1xea3nP_1Z_wdH?usp=sharing

Q1

Answer

```
[1] "Number of Friends: 98"
[1] "Mean: 542.673469387755"
[1] "Standard Deviation: 539.433738523966"
[1] "Median: 396"
```

Listing 1: Mean standard deviation and median of Facebook friends found in HW4-friend-count.csv.

```
file <- read.csv("HW4-friend-count.csv")
print(paste("Number of Friends:", nrow(file)))
print(paste("Mean:", mean(file[,2])))
print(paste("Standard Deviation:", sd(file[,2])))
print(paste("Median:", median(file[,2])))
```

Listing 2: R script used to calculate output in listing 1.

```
library(ggplot2)
theme_set(theme_bw())
library(tidyverse)
options(scipen=999)

file2 <- read.csv("user-HW4-friend-count.csv")

ordered <- file2[order(-file2$FRIENDCOUNT),]
ordered$USER <- factor(ordered$USER, levels = ordered$USER)

theme_update(axis.text.x=element_blank())
```

```
ordered %>%
  mutate(highlight_flag = ifelse(USER=="U", T, F)) %>%
  ggplot(aes(x=USER, y=FRIENDCOUNT)) +
  geom_point(aes(color = highlight_flag)) +
  scale_color_manual(values = c('black', 'red')) +
  labs(y="Number of Friends", x="Users", caption="Facebook Data")
```

Listing 3: Rscript used to graph Facebook friends based on their number of friends.

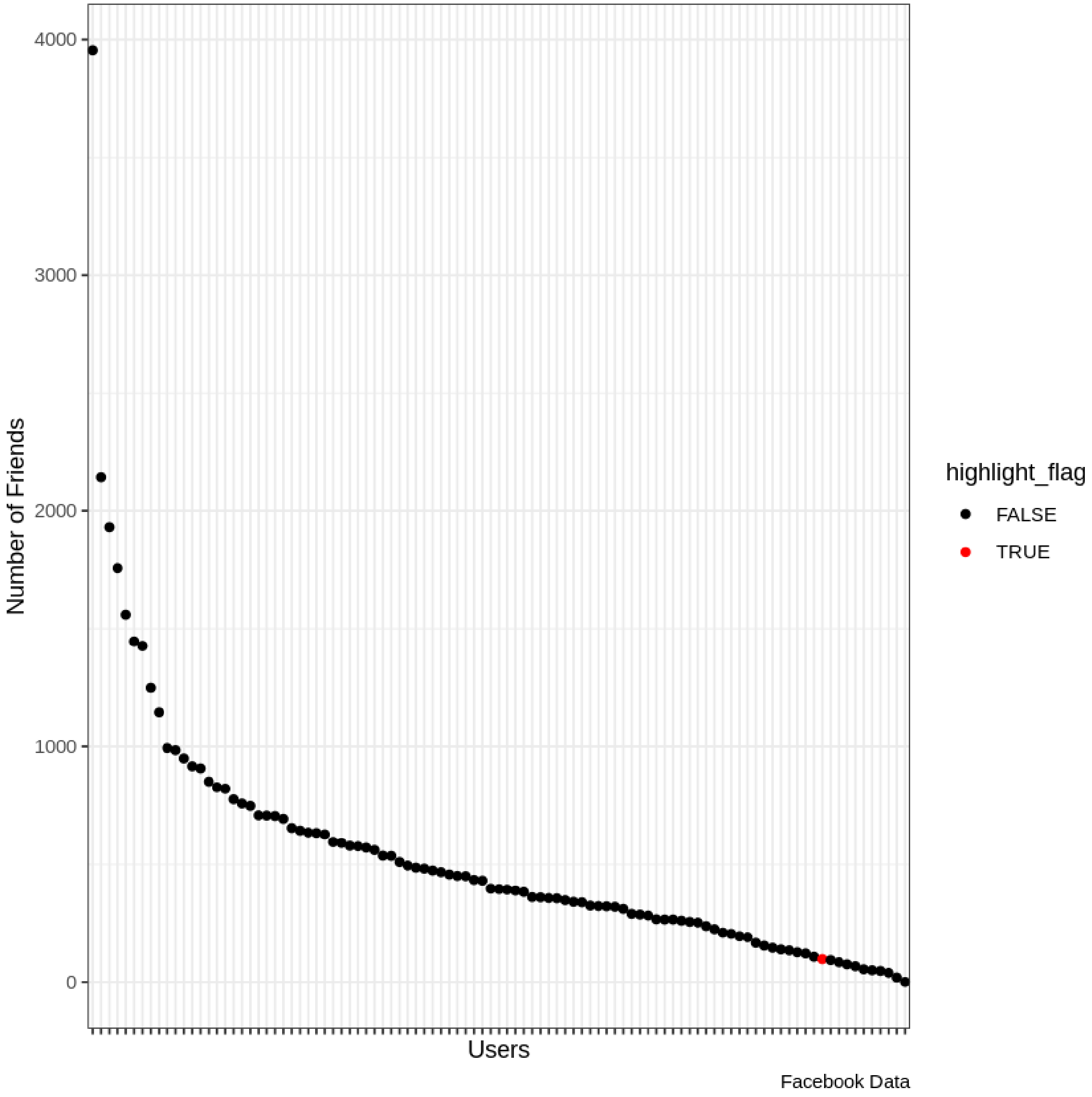

Figure 1: Facebook friends graph.

Discussion

Let’s start with the data in listing 1. This was calculated by the R script in listing 2, and the output was directed to “facebookData.txt” at the command line. While the question does not ask for it, I felt in necessary to include the number of friends the user has by outputting the number of rows in the csv file using the “nrow” function.

Next is the Rscript in listing 3. The first thing of note is the inclusion of the “tidyverse” library. This library allowed me to pipe my dataset into the “ggplot” function. The next thing of note is the “theme_update” function that makes the x-axis blank. You may be asking yourself “Well how did he label the user with a blank x-axis?” Don’t worry, I’m about to answer that. Before the data reaches the “ggplot” function, it is first piped through the “mutate” function. Normally, “mutate” is used to add data, but I used it to highlight data. In this instance, I used it to highlight the user. To do this, I had to add the user to the data set. I manually (using vim) added the user as “U,” as well as their friend count into a separate copy of the “HW4-friend-count.csv” file called “user-HW4-friend-count.csv.” In this particular implementation, the mutate function checks to see if the user is “U.” If it is, then the user is highligted on the graph. The default color scheme of the highlight flag is blue-TRUE and pink-FALSE. I changed it on line 16 to red and black, respectively. This means that the user is the red dot on the graph, and their friends are the black dots.

To conclude this question, it seems that the friendship paradox does hold true for this user. Most of the user’s friends have more friends than they do.

Q2

Answer

```
import tweepy
import csv

auth = tweepy.OAuthHandler("", "")
auth.set_access_token("", "")

api = tweepy.API(auth, wait_on_rate_limit=True,
    wait_on_rate_limit_notify=True)

with open(’Followers.csv’, ’a’) as file:
    writer = csv.writer(file, delimiter = ’\t’)
    writer.writerow([’USER’, ’FOLLOWERS’])
    for follower in tweepy.Cursor(api.followers, screen_name=’weiglemc’
    ).items():
        writer.writerow([follower.screen_name, follower.followers_count
   ])
```

Listing 4: Python script used to get follower data for weiglemc and output to “Followers.csv.”

```
[1] "Number of Followers: 433"
[1] "Mean: 3081.54734411085"
[1] "Standard Deviation: 28070.8917303204"
[1] "Median: 195"
```

Listing 5: Mean standard deviation and median of Twitter followers found in Followers.csv.

```
file <- read.csv("Followers.csv", sep="\t")
print(paste("Number of Followers:", nrow(file)))
print(paste("Mean:", mean(file[,2])))
print(paste("Standard Deviation:", sd(file[,2])))
print(paste("Median:", median(file[,2])))
```

Listing 6: R script used to calculate output in listing 5.

```
import csv

with open(’user-Followers.csv’, ’a’) as file:
    writer = csv.writer(file, delimiter = ’\t’)
    writer.writerow([’U’, ’433’])
```

Listing 7: Python script to append the user and their follower count to “user-Followers.csv.”

```
library(ggplot2)
theme_set(theme_bw())
library(tidyverse)
options(scipen=999)

file3 <- read.csv("user-Followers.csv", sep="\t")

ordered2 <- file3[order(-file3$FOLLOWERS),]
ordered2$USER <- factor(ordered2$USER, levels = ordered2$USER)

theme_update(axis.text.x=element_blank())
ordered2 %>%
  mutate(highlight_flag = ifelse(USER=="U", T, F)) %>%
  ggplot(aes(x=USER, y=FOLLOWERS)) +
  geom_point(aes(color = highlight_flag)) +
  ylim(0, 1000) +
  scale_color_manual(values = c(’black’, ’red’)) +
  labs(y="Number of Followers", x="Users", caption="Twitter Data")
```

Listing 8: R script used to graph Twitter followers based on their follower count.

The graph for this question is figure 2. It is found of page 7.

Discussion

I had to use the account “weiglemc” to get follower data as my account has less than fifty followers. Most of the methods used in Question 1 are used here, with some differences:

- I used the Python script in listing 4 to get follower data for this question.
- I wrote the “Followers.csv file with a tab delimiter.
- Output from listing 6 was directed to “twitterData.txt” at the command line.
- In order to append the user to the “user-Followers.csv” I wrote the Python script in listng 7.
- Because some of weiglemc’s followers had so many followers, I limited the y-axis to 1,000 on line 16 in listing 8. This means that only users 1,000 and fewer followers appear on the graph.
- According to Rstudio and Google Collab, 88 rows of data do not appear in figure 2. This means that 88 of weiglemc’s followers had more than 1,000 followers.

Given that “mweigle” is in the top fifty percent of the ordered users in figure 2, it would seem that the friendship paradox does not hold here.

References

- Tweepy Introduction, `http://docs.tweepy.org/en/latest/getting_started.html#introduction`
- Tweepy API Reference, `http://docs.tweepy.org/en/latest/api.html#API.followers`
- Tweepy Cursor Tutorial, `http://docs.tweepy.org/en/latest/cursor_tutorial.html`
- tweepy count limited to 200?, `https://stackoverflow.com/questions/23460560/tweepy-count-limited-to-200`
- Tweet.py search method does not support pagination, `https://github.com/tweepy/tweepy/issues/1040`
- Getting this error when using Tweepy, `https://stackoverflow.com/questions/38775997/getting-this-error-when-using-tweepy`

- Python CSV, https://www.programiz.com/python-programming/csv
- How to Write R Script Explained with an Awesome Example, https://dzone.com/articles/how-to-write-r-script-explained-with-an-awesome-ex#:~:text=How%20to%20Create%20R%20Script%201%20You%20can,file%20extension%20to%20the%20file.%20More%20items...%20
- Printing and displaying strings, https://riptutorial.com/r/example/1221/printing-and-displaying-strings
- R - Mean, Median and Mode, https://www.tutorialspoint.com/r/r_mean_median_mode.htm
- R Documentaion-sd, https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/sd
- The Number of Rows/Columns of an Array, https://stat.ethz.ch/R-manual/R-devel/library/base/html/nrow.html
- ggplot2 axis ticks : A guide to customize tick marks and labels, http://www.sthda.com/english/wiki/ggplot2-axis-ticks-a-guide-to-customize-tick-marks-and-labels
- How to Set Axis Limits in ggplot2, https://www.statology.org/set-axis-limits-ggplot2/#:~:text=Often%20you%20may%20want%20to%20set%20the%20axis,the%20lower%20and%20upper%20limit%20of%20the%20y-axis.
- How to highlight data in ggplot2, https://www.sharpsightlabs.com/blog/highlight-data-in-ggplot2/#:~:text=There%20are%20a%20few%20ways%20to%20change%20the,the%20variable%20we%20mapped%20to%20the%20fill%20aesthetic.
- Week-03-InfoVis-R Collab Notebook, https://colab.research.google.com/github/cs432-websci-fall20/assignments/blob/master/432_Week_03_InfoVis_R.ipynb

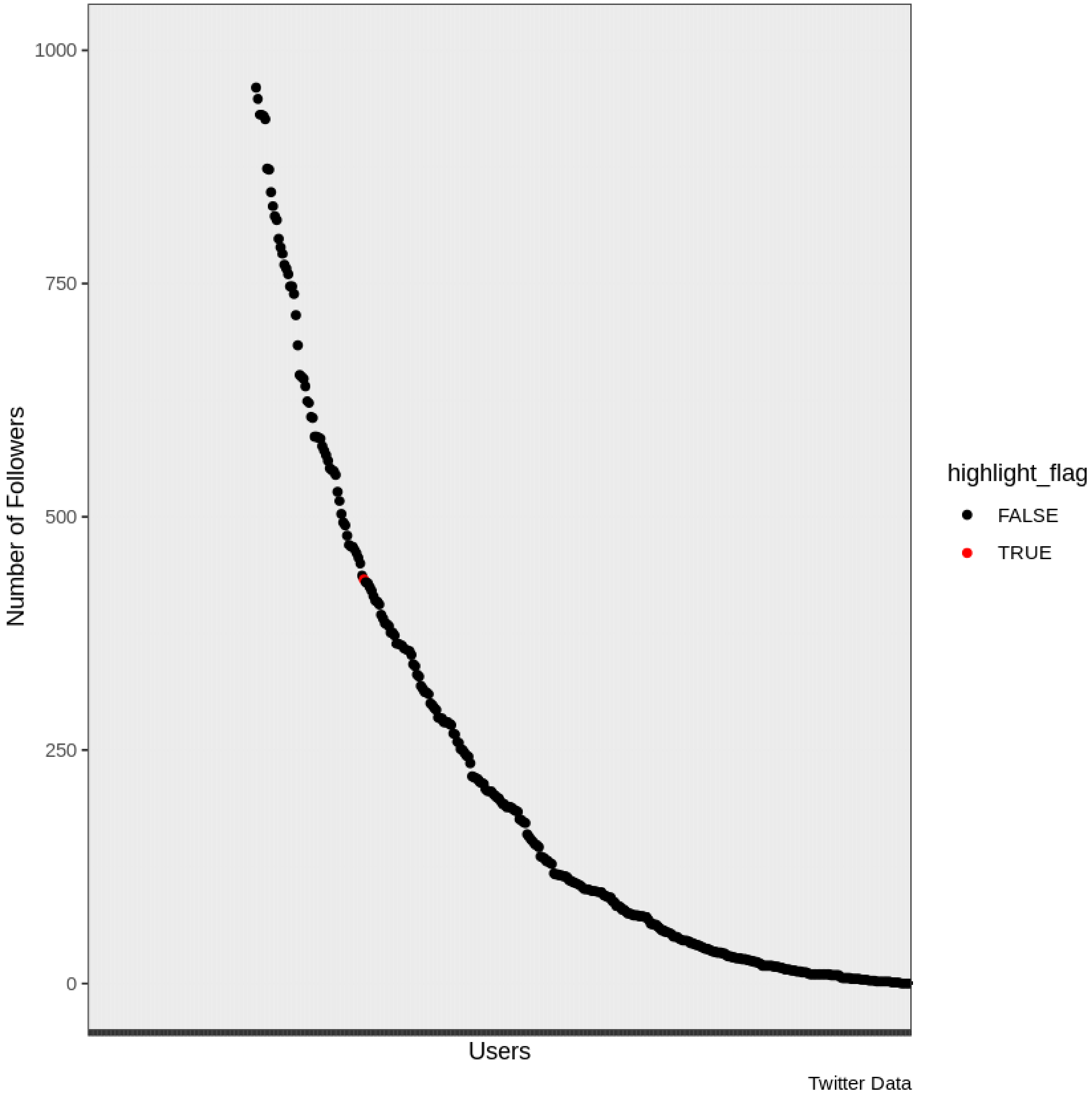

Figure 2: Twitter followers graph.