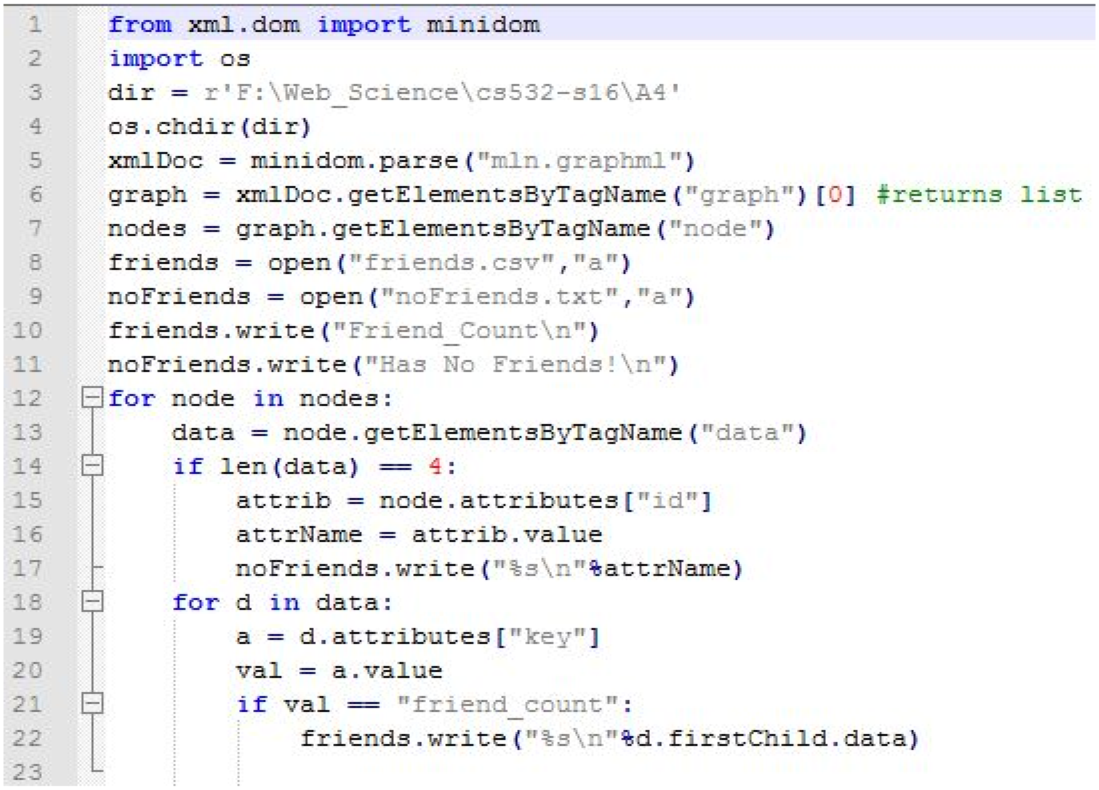
CS 432

Assignment 4

Daniel Moore

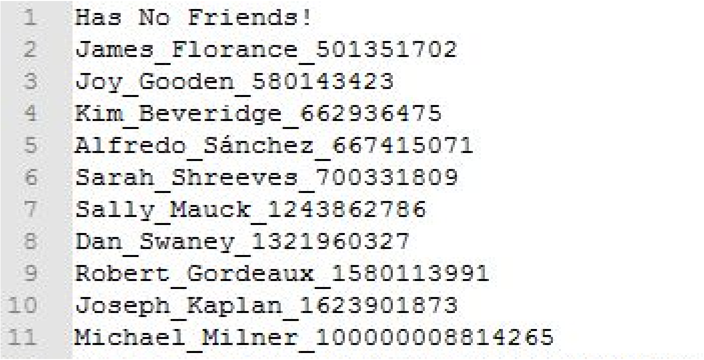
Question 1:

Part one of this assignment involved determining whether the friendship paradox held true for Dr. Nelson's facebook account. In order to do so I used the following script to parse the xml data in a graph which Dr. Nelson provided of his facebook account and print the friend count for each of his friends to a .csv file:



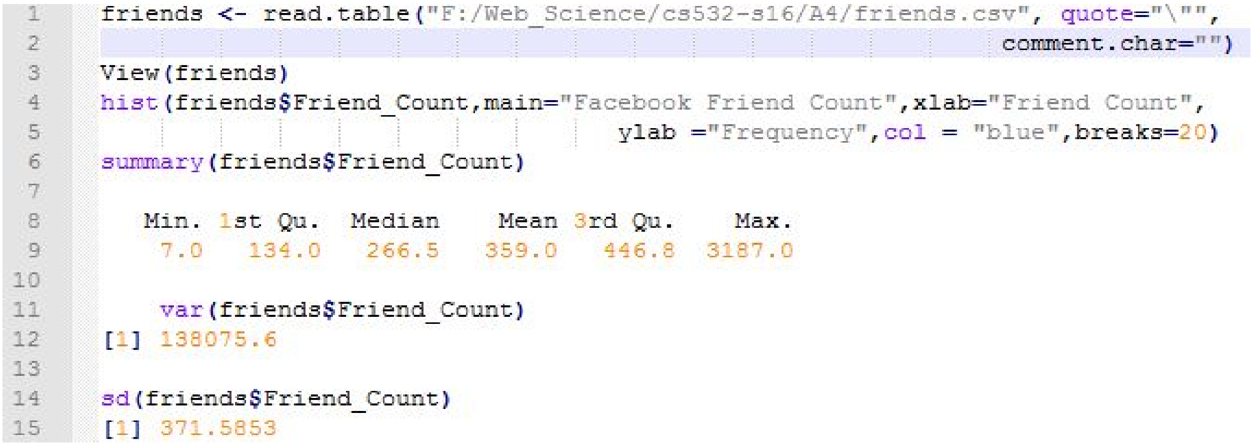
```
from xml.dom import minidom
import os
dir = r'F:\Web_Science\cs532-s16\A4'
os.chdir(dir)
xmlDoc = minidom.parse("mln.graphml")
graph = xmlDoc.getElementsByTagName("graph")[0] #returns list
nodes = graph.getElementsByTagName("node")
friends = open("friends.csv","a")
noFriends = open("noFriends.txt","a")
friends.write("Friend_Count\n")
noFriends.write("Has No Friends!\n")
for node in nodes:
    data = node.getElementsByTagName("data")
    if len(data) == 4:
        attrib = node.attributes["id"]
        attrName = attrib.value
        noFriends.write("%s\n"%attrName)
    for d in data:
        a = d.attributes["key"]
        val = a.value
        if val == "friend_count":
            friends.write("%s\n"%d.firstChild.data)

```

This data as not available for all of Dr. Nelson's friends. I used the above script to post the names of these users into the text file displayed below:

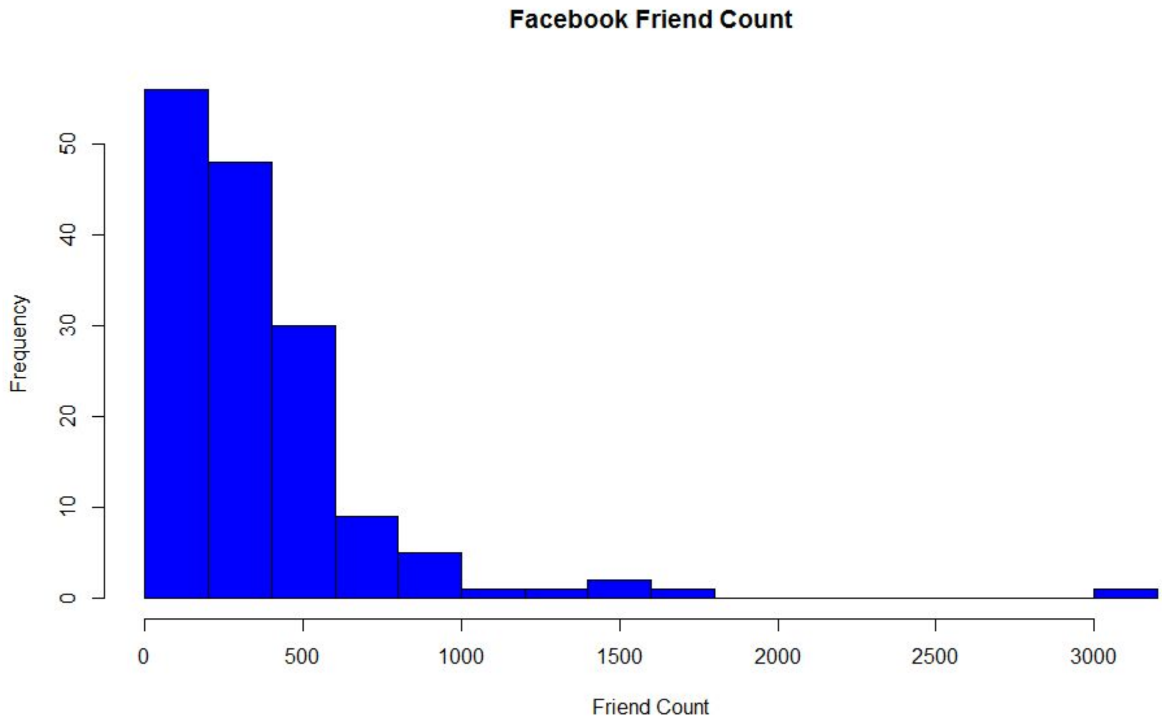


```
Has No Friends!
James_Florance_501351702
Joy_Gooden_580143423
Kim_Beveridge_662936475
Alfredo_Sánchez_667415071
Sarah_Shreeves_700331809
Sally_Mauck_1243862786
Dan_Swaney_1321960327
Robert_Gordeaux_1580113991
Joseph_Kaplan_1623901873
Michael_Milner_100000008814265
```

I then used the following r script to plot the extracted data on a histogram:



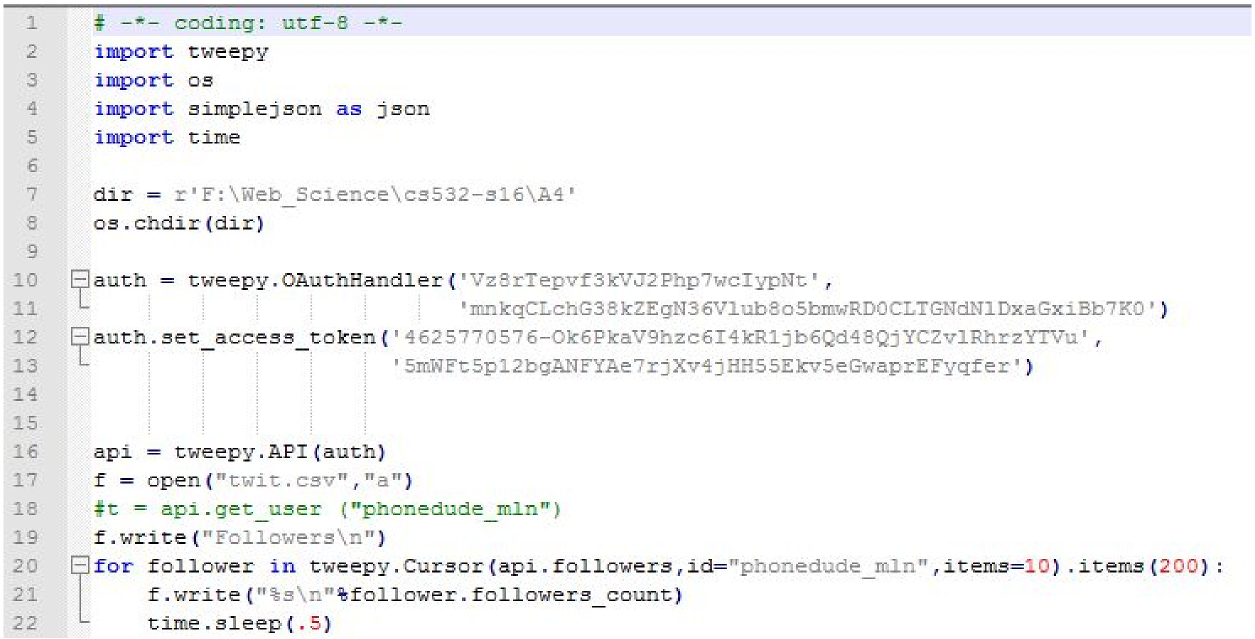
```
friends <- read.table("F:/Web_Science/cs532-s16/A4/friends.csv", quote="\"",
                                                            comment.char="")
View(friends)
hist(friends$Friend_Count,main="Facebook Friend Count",xlab="Friend Count",
                                   ylab ="Frequency",col = "blue",breaks=20)
summary(friends$Friend_Count)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    7.0   134.0   266.5   359.0   446.8  3187.0

    var(friends$Friend_Count)
[1] 138075.6

sd(friends$Friend_Count)
[1] 371.5853
```



The median friend count was 266.5 friends with a mean of 359.0 friends and a standard deviation of 371.6 friends. The minimum value was 7 friends and the maximum value was 3187 friends. Both the mean and median values are greater than Dr. Nelson’s friend count of 165.

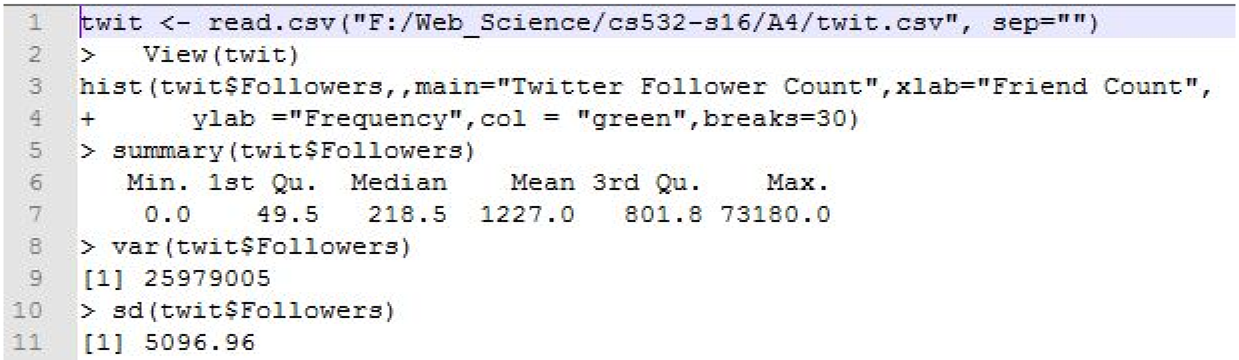
Question 2:

For the second part of this assignment, I was instructed to perform the same tasks for a twitter account. My twitter account only has 2 followers so, once again, I examined Dr. Nelson’s account. I used the following script to extract the followers counts of Dr. Nelson’s followers:

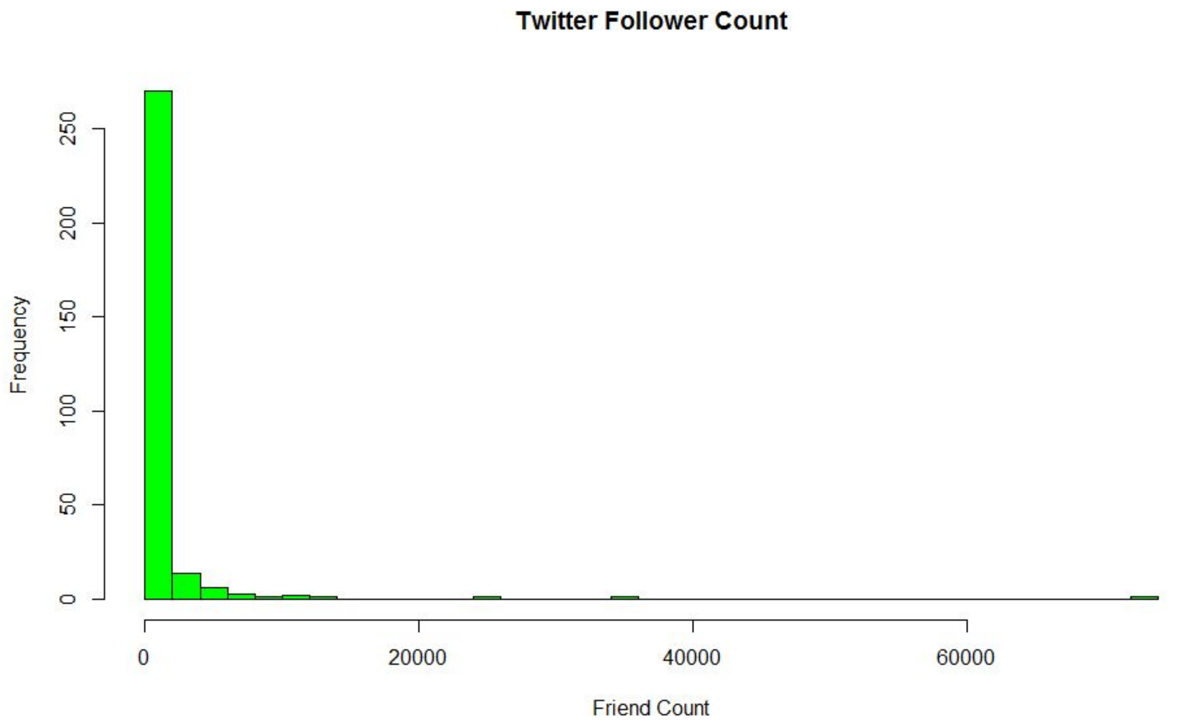


```
# -*- coding: utf-8 -*-
import tweepy
import os
import simplejson as json
import time

dir = r'F:\Web_Science\cs532-s16\A4'
os.chdir(dir)

auth = tweepy.OAuthHandler('Vz8rTepvf3kVJ2Php7wcIypNt',
                           'mnkqCLchG38kZEgN36Vlub8o5bmwRD0CLTGNdNlDxaGxiBb7K0')
auth.set_access_token('4625770576-Ok6PkaV9hzc6I4kR1jb6Qd48QjYCZvlRhrzYTVu',
                      '5mWFt5p12bgANFYAe7rjXv4jHH55Ekv5eGwaprEFyqfer')


api = tweepy.API(auth)
f = open("twit.csv","a")
#t = api.get_user ("phonedude_mln")
f.write("Followers\n")
for follower in tweepy.Cursor(api.followers,id="phonedude_mln",items=10).items(200):
    f.write("%s\n"%follower.followers_count)
    time.sleep(.5)
```

Unfortunately, I repeatedly encountered “Rate_Limit_Errors” and was only able to extract the followers counts for 300 of Dr. Nelson’s 491 followers. Below is the script which I used to plot the followers counts in a histogram:



```
twit <- read.csv("F:/Web_Science/cs532-s16/A4/twit.csv", sep="")
>   View(twit)
hist(twit$Followers,,main="Twitter Follower Count",xlab="Friend Count",
+      ylab ="Frequency",col = "green",breaks=30)
> summary(twit$Followers)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0    49.5   218.5  1227.0   801.8 73180.0
> var(twit$Followers)
[1] 25979005
> sd(twit$Followers)
[1] 5096.96
```



The minimum followers count was 0 and the maximum was 73180 followers. The median value was 218.5 and the mean was 1227 with a standard deviation of 5097.