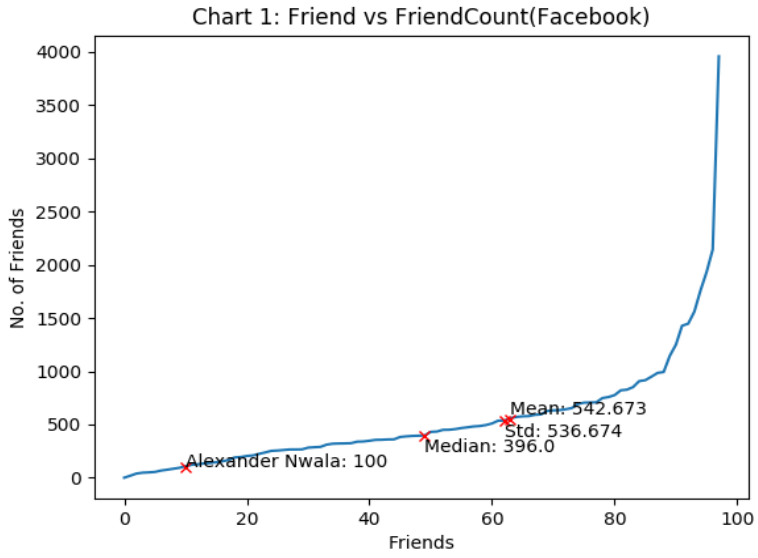
CS 432 Assignment #4
Larry Montgomery 01093132

Question 1:

The friendship paradox is the concept that a person's friends have more friends than they do. When it comes to professor Nwala I can say that this is a true statement when considering his friends. We can see from the figure below that of professor Nwala's 100 friends, the average value is 542 and median value is 396. This shows that the average is far beyond the 100 friends and follows the friendship paradox.



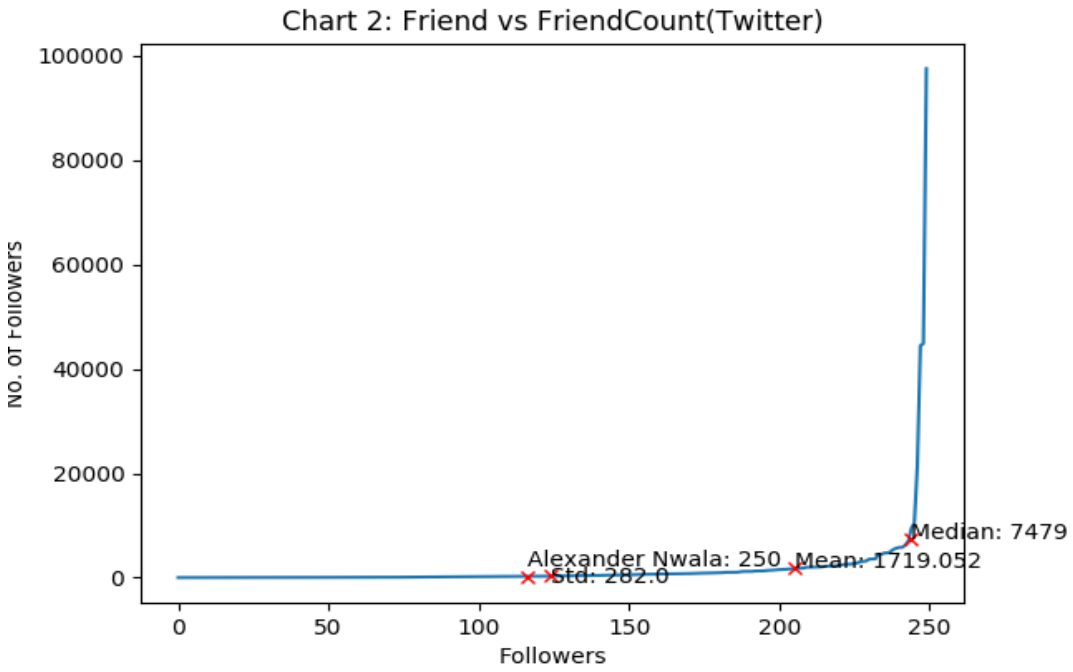
To calculate the values and generate the graph I made use for Matplotlib , CSV module, and Numpy libraries. The csv module was used to retrieve the information from the csv file that we were given. I parsed each row into a tuple of name and friend count. I then sorted each tuple pair based on number of friends. I then put the friend totals into a list and used the numpy library to calculate the mean, median and standard deviation for the data set. The code below represents the parsing and calculating statistical values.

```
def parse_csv(csv_file):
    data_list = []
    with open(csv_file) as f:
        doc_parsed = csv.reader(f, delimiter=',')
        for row in doc_parsed:
            try:
                data_list.append((row[0],int(row[1])))
            except ValueError:
                pass
        return sorted(data_list, key=itemgetter(1))
```

```
def calc_stats(data_list):
    data_mean = round(numpy.mean(data_list), 3)
    data_std = round(numpy.std(data_list),3)
    data_median = round(numpy.median(data_list),3)
    print('curr_mean: ', data_mean)
    print('curr_std: ', data_std)
    print('curr_median: ', data_median)
    # returns mean, standard deviation and median in that order
    return data_mean, data_std, data_median
```

Question 2:

The data collected from professor Nwala's Twitter account and the follower counts of his followers proves that the friendship paradox is true for followers on Twitter. From the 250 followers of Professor Nwala the average follower count was 1719 and the median value is 7479. This large difference means there are a few large outliers but both are extensively beyond the 250 followers of professor Nwala. The graph below represents the distribution of followers per follower of professor Nwala. There was an issue when adjusting the annotation of the standard deviation label since the scaling was so small but the standard deviation is 282.



To collect the data to create this graph I used the Tweepy API to first collect the list of Id's for the people professor Nwala follows. Then I made a request for each user id and collected the number of followers for the response object of the API request. After collecting the 250 following counts, I sorted them and performed the same operations as I did for the Facebook data set in question 1. The code below represents the code used to collect the data used in the graph.

```
user_id = "acnwala"

auth = tweepy.OAuthHandler(ckey, csecret)
auth.set_access_token(atoken, asecret)

api = tweepy.API(auth)

user_list = api.followers_ids(user_id)
follower_counts = []

for each in user_list:

    temp_follower_object = api.get_user(each)
    temp_follower_count = temp_follower_object.followers_count
    follower_counts.append(temp_follower_count)
```