Week 10 Assignment – Twitter and Python

Brian Mallari

Part 1 – Twitter Topics and Tweet Collection

Pick three different topics and/or politicians (possibly controversial!) and collect 180 tweets for each of them. Give a summary (what you collected, how you did it). Don't put actual data. [4 points]

For this assignment, I chose the following politicians:

1. Phil Murphy, Democratic governor of New Jersey
2. Bill De Blasio, Democratic mayor of New York City
3. Andrew Cuomo, Democratic governor of New York State

I chose these three politicians because I figured that the impact of their work would be the most relevant to me as a resident of northern New Jersey.

I utilized code featured in the video podcast, 10.2 - Sentiment analysis with Twitter data, by Dr. Chirag Shah to collect Twitter tweets relating to each of the politicians. For each politician, 180 tweets were collected. For each tweet, I collected the following details:

1. User name
2. Author ID
3. Date the tweet was created
4. Content of the tweet
5. Count of retweets
6. Any hashtags associated with the tweet
7. Number of followers of the user
8. Number of friends of the user
9. Tweet polarity (how positive, neutral, or negative a tweet is)
10. Tweet subjectivity (how subjective or objective a tweet is)

The query code was repeated three times (once for each politician). Each politician's name was written in quotes to be sure that the full name was used for each query. Moreover, the names of the csv files were tailored in such a way that each query for each politician was separate from each other:

1. 'results_phil_murphy.csv' for Phil Murphy
2. 'results_bill_de_blasio.csv' for Bill De Blasio
3. 'results_andrew_cuomo.csv' for Andrew Cuomo.

The original simply had 'results.csv', so each iteration of the query code would have written over the previous file, and only the results of the last query would have been saved.

In order to resolve a 'charmap' codec error in Spyder where certain characters could not be encoded, I had to set the *encoding* parameter in the open() function to "utf-8". However, it appears that there are still some characters that do not show up properly when viewing the csv files with Microsoft Excel. Also, in order to gather 180 tweets for each politician, I had to change the *result_type* parameter in the tweepy.Collect() function from "popular" to "recent".

The Python script for collecting the Twitter tweets was deliberately kept separate from the Python script for running sentimental analysis. In doing so, the first script could be run only once to acquire the tweets while the second script could be run multiple times as the sentiment analysis progressed. Had the two scripts been combined into one, then a new set of tweets would've be collected each time the analysis script was run, thus potentially compromising any attempt at a proper analysis.

Python Script

```
# -*- coding: utf-8 -*-
"""
Created on Wed Mar 28 16:15:05 2018

@author: Brian
"""

from textblob import TextBlob
import csv
import tweepy
import unidecode

# AUTHENTICATION (OAuth)
f = open('auth.k','r')
ak = f.readlines()
f.close()
auth1 = tweepy.auth.OAuthHandler(ak[0].replace("\n",""), ak[1].replace("\n",""))
auth1.set_access_token(ak[2].replace("\n",""), ak[3].replace("\n",""))
api = tweepy.API(auth1)

# Twitter search #1
target_num = 180
query = "Phil Murphy"

csvFile = open('results_phil_murphy.csv','w', encoding = "utf-8")
csvWriter = csv.writer(csvFile)
csvWriter.writerow(["username","author id","created", "text", "retwc", "hashtag", "followers",
"friends","polarity","subjectivity"])
counter = 0

for tweet in tweepy.Cursor(api.search, q = query, lang = "en", result_type = "recent", count =
target_num).items():
    created = tweet.created_at
    text = tweet.text
    text = unidecode.unidecode(text)
    retwc = tweet.retweet_count
    try:
        hashtag = tweet.entities[u'hashtags'][0][u'text'] #hashtags used
```

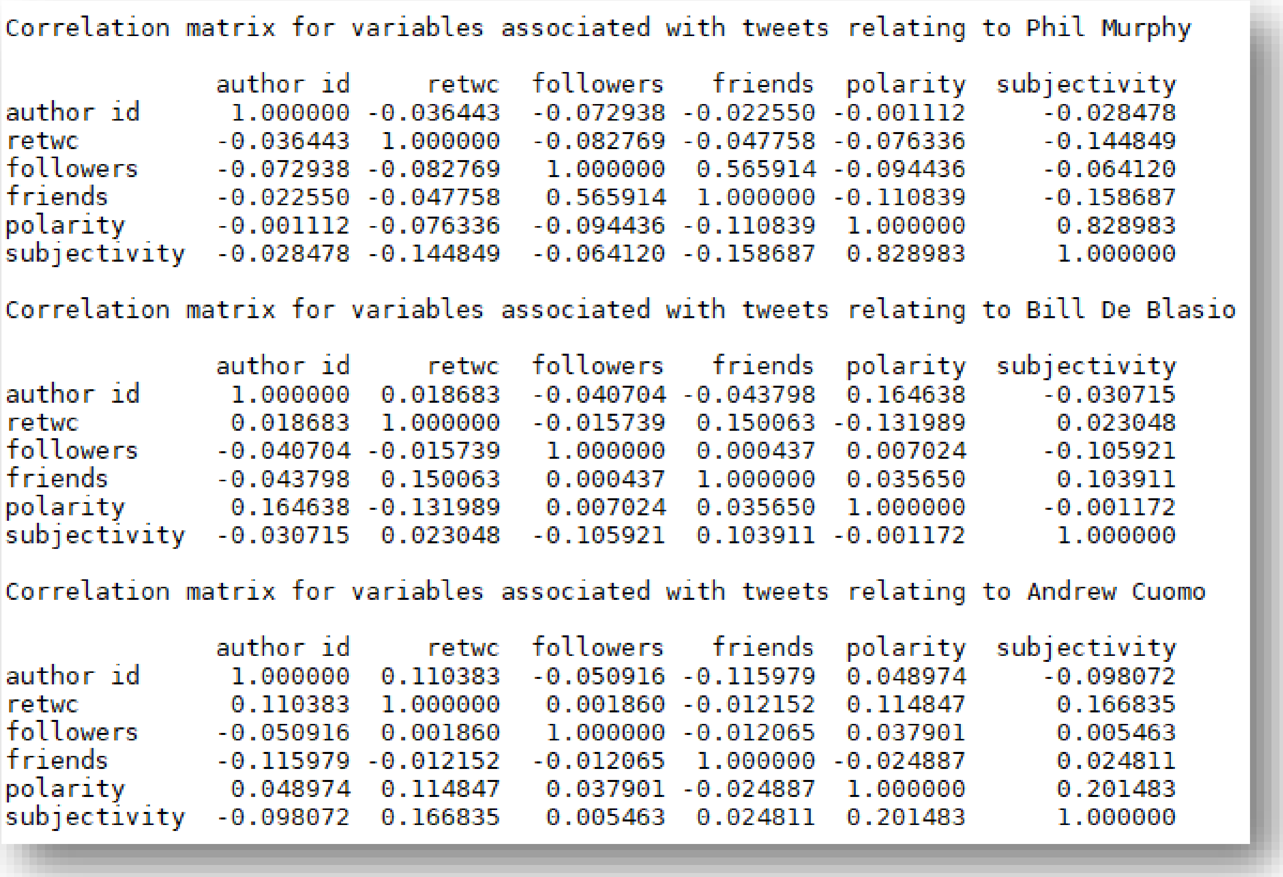
```
   except:
      hashtag = "None"
   username  = tweet.author.name           #author/user name
   authorid  = tweet.author.id             #author/user ID#
   followers = tweet.author.followers_count #number of author/user followers (inlink)
   friends = tweet.author.friends_count     #number of author/user friends (outlink)

   text_blob = TextBlob(text)
   polarity = text_blob.polarity
   subjectivity = text_blob.subjectivity
   csvWriter.writerow([username, authorid, created, text, retwc, hashtag, followers, friends, polarity,
subjectivity])

   counter = counter + 1
   if (counter == target_num):
      break

csvFile.close()

# Twitter search #2
target_num = 180
query = "Bill de Blasio"

csvFile = open('results_bill_de_blasio.csv','w', encoding = "utf-8")
csvWriter = csv.writer(csvFile)
csvWriter.writerow(["username","author id","created", "text", "retwc", "hashtag", "followers",
"friends","polarity","subjectivity"])
counter = 0

for tweet in tweepy.Cursor(api.search, q = query, lang = "en", result_type = "recent", count =
target_num).items():
   created = tweet.created_at
   text = tweet.text
   text = unidecode.unidecode(text)
   retwc = tweet.retweet_count
   try:
      hashtag = tweet.entities[u'hashtags'][0][u'text'] #hashtags used
   except:
      hashtag = "None"
   username  = tweet.author.name           #author/user name
   authorid  = tweet.author.id             #author/user ID#
   followers = tweet.author.followers_count #number of author/user followers (inlink)
   friends = tweet.author.friends_count     #number of author/user friends (outlink)

   text_blob = TextBlob(text)
   polarity = text_blob.polarity
   subjectivity = text_blob.subjectivity
```

```
    csvWriter.writerow([username, authorid, created, text, retwc, hashtag, followers, friends, polarity,
subjectivity])

    counter = counter + 1
    if (counter == target_num):
        break

csvFile.close()

# Twitter search #3
target_num = 180
query = "Andrew Cuomo"

csvFile = open('results_andrew_cuomo.csv','w', encoding = "utf-8")
csvWriter = csv.writer(csvFile)
csvWriter.writerow(["username","author id","created", "text", "retwc", "hashtag", "followers",
"friends","polarity","subjectivity"])
counter = 0

for tweet in tweepy.Cursor(api.search, q = query, lang = "en", result_type = "recent", count =
target_num).items():
    created = tweet.created_at
    text = tweet.text
    text = unidecode.unidecode(text)
    retwc = tweet.retweet_count
    try:
        hashtag = tweet.entities[u'hashtags'][0][u'text'] #hashtags used
    except:
        hashtag = "None"
    username  = tweet.author.name           #author/user name
    authorid  = tweet.author.id             #author/user ID#
    followers = tweet.author.followers_count #number of author/user followers (inlink)
    friends = tweet.author.friends_count     #number of author/user friends (outlink)

    text_blob = TextBlob(text)
    polarity = text_blob.polarity
    subjectivity = text_blob.subjectivity
    csvWriter.writerow([username, authorid, created, text, retwc, hashtag, followers, friends, polarity,
subjectivity])

    counter = counter + 1
    if (counter == target_num):
        break

csvFile.close()
```

Part 2 – Sentiment Analysis

Do exploratory analyses using Python to detect any trends and form hypotheses. Present (1) visual relationships among the variables explored [3 points]; and (2) your hypotheses based on these relationships. [3 points]

Here are the correlation matrices for each of the three politicians selected for this assignment:

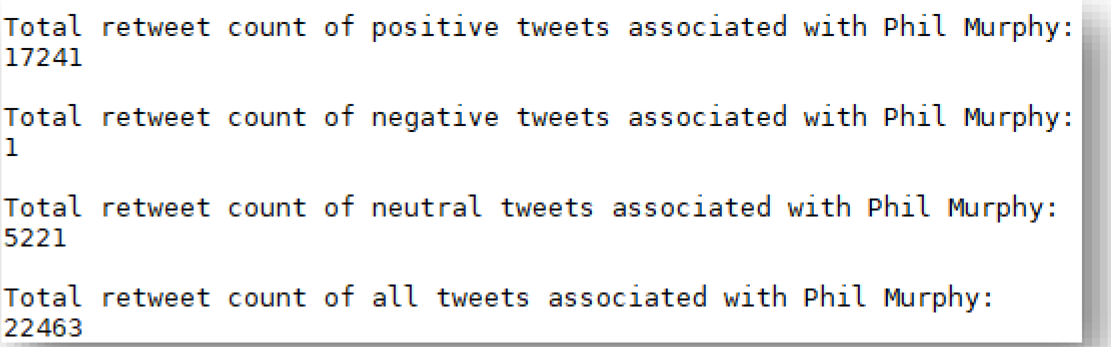


```
Correlation matrix for variables associated with tweets relating to Phil Murphy

              author id     retwc  followers   friends  polarity  subjectivity
author id      1.000000 -0.036443  -0.072938 -0.022550 -0.001112     -0.028478
retwc         -0.036443  1.000000  -0.082769 -0.047758 -0.076336     -0.144849
followers     -0.072938 -0.082769   1.000000  0.565914 -0.094436     -0.064120
friends       -0.022550 -0.047758   0.565914  1.000000 -0.110839     -0.158687
polarity      -0.001112 -0.076336  -0.094436 -0.110839  1.000000      0.828983
subjectivity  -0.028478 -0.144849  -0.064120 -0.158687  0.828983      1.000000

Correlation matrix for variables associated with tweets relating to Bill De Blasio

              author id     retwc  followers   friends  polarity  subjectivity
author id      1.000000  0.018683  -0.040704 -0.043798  0.164638     -0.030715
retwc          0.018683  1.000000  -0.015739  0.150063 -0.131989      0.023048
followers     -0.040704 -0.015739   1.000000  0.000437  0.007024     -0.105921
friends       -0.043798  0.150063   0.000437  1.000000  0.035650      0.103911
polarity       0.164638 -0.131989   0.007024  0.035650  1.000000     -0.001172
subjectivity  -0.030715  0.023048  -0.105921  0.103911 -0.001172      1.000000

Correlation matrix for variables associated with tweets relating to Andrew Cuomo

              author id     retwc  followers   friends  polarity  subjectivity
author id      1.000000  0.110383  -0.050916 -0.115979  0.048974     -0.098072
retwc          0.110383  1.000000   0.001860 -0.012152  0.114847      0.166835
followers     -0.050916  0.001860   1.000000 -0.012065  0.037901      0.005463
friends       -0.115979 -0.012152  -0.012065  1.000000 -0.024887      0.024811
polarity       0.048974  0.114847   0.037901 -0.024887  1.000000      0.201483
subjectivity  -0.098072  0.166835   0.005463  0.024811  0.201483      1.000000
```

With the exception of the diagonal values which are 1 because the variable in question is correlating with itself, there aren't many strong or even moderate linear relationships among the variables. Moreover, the positivity or negativity of a correlation isn't necessarily consistent across the three correlation matrices. That is, a correlation that is negative for one politician, like 'retwc'-'followers' for Phil Murphy, can be positive for another politician, like Andrew Cuomo. It's worth noting that 'author id' isn't really a counted, measured, or calculated value. Instead, it's just a name used by an entity on Twitter. It's also worth noting that these values will change each time the Python script for Part 1 is run because the tweets collected will be different.

Now with all of that said, I see no clear trend among the variables based off of these correlation coefficients. Nevertheless, I'll explore the following hypothesis as a matter of curiosity – for a given politician, positive tweets are retweeted more than negative tweets. For this case, positive tweets are tweets with a polarity greater than 0, and negative tweets are tweets with a polarity less than 0.

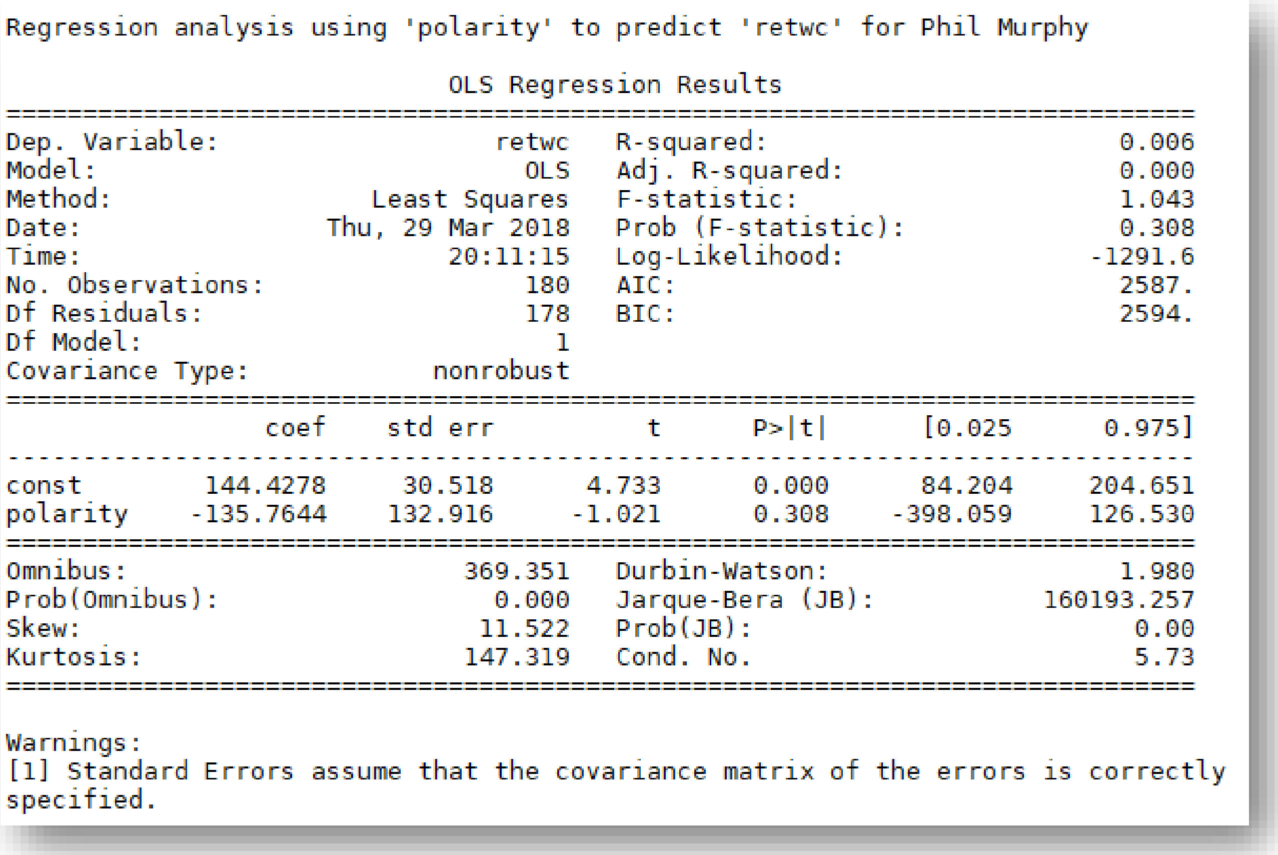
First, here are the retweet counts for Phil Murphy:



```
Total retweet count of positive tweets associated with Phil Murphy:
17241

Total retweet count of negative tweets associated with Phil Murphy:
1

Total retweet count of neutral tweets associated with Phil Murphy:
5221

Total retweet count of all tweets associated with Phil Murphy:
22463
```

In the case of Phil Murphy, the number of retweets for positive polarity tweets is far higher than the number of retweets for negative polarity tweets. Worth noting, though, is the significant number of retweets for neutral tweets.

Now let's look at the regression model to see how well 'polarity' predicts 'retwc' for Phil Murphy:

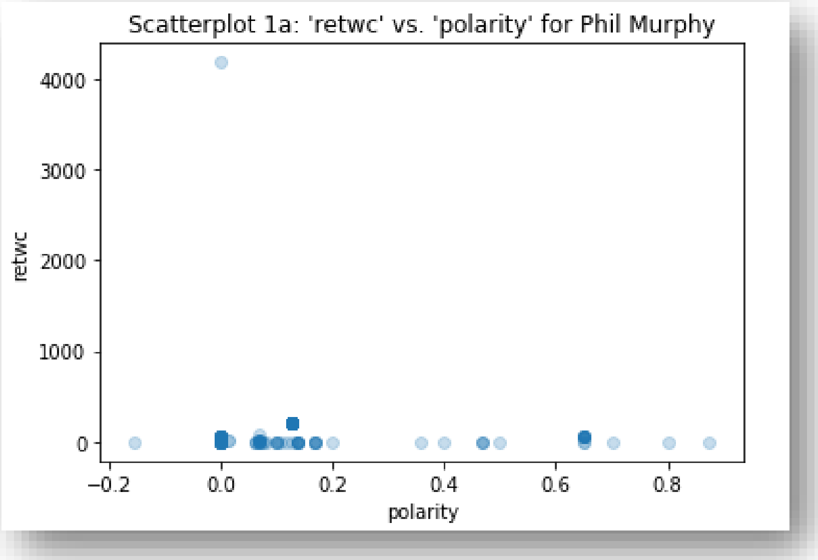


```
Regression analysis using 'polarity' to predict 'retwc' for Phil Murphy

                            OLS Regression Results
==============================================================================
Dep. Variable:                  retwc   R-squared:                       0.006
Model:                            OLS   Adj. R-squared:                  0.000
Method:                 Least Squares   F-statistic:                     1.043
Date:                Thu, 29 Mar 2018   Prob (F-statistic):              0.308
Time:                        20:11:15   Log-Likelihood:                -1291.6
No. Observations:                 180   AIC:                             2587.
Df Residuals:                     178   BIC:                             2594.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        144.4278     30.518      4.733      0.000      84.204     204.651
polarity    -135.7644    132.916     -1.021      0.308    -398.059     126.530
==============================================================================
Omnibus:                      369.351   Durbin-Watson:                   1.980
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           160193.257
Skew:                          11.522   Prob(JB):                         0.00
Kurtosis:                     147.319   Cond. No.                         5.73
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

Regression equation: predicted retwc = 144.4278 + -135.7644*polarity

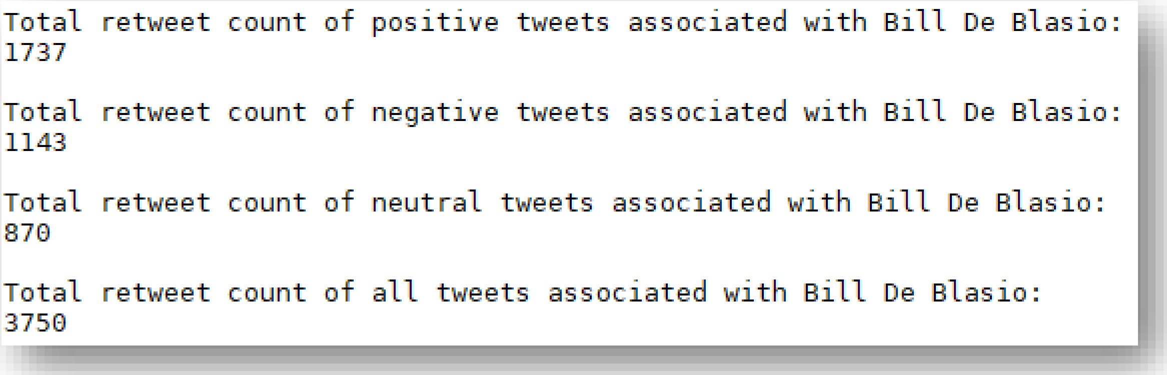
Judging by the R-squared value, it doesn't seem that 'polarity' does a good job at predicting 'retwc'.

Now let's look at the scatterplot of 'retwc' and 'polarity' for any possible insights:



We can see here that majority of the points are distributed towards the bottom of the scatterplot in a horizontal fashion, with the exception being that one point towards the top. The overall combination of points results in the low R-squared value.

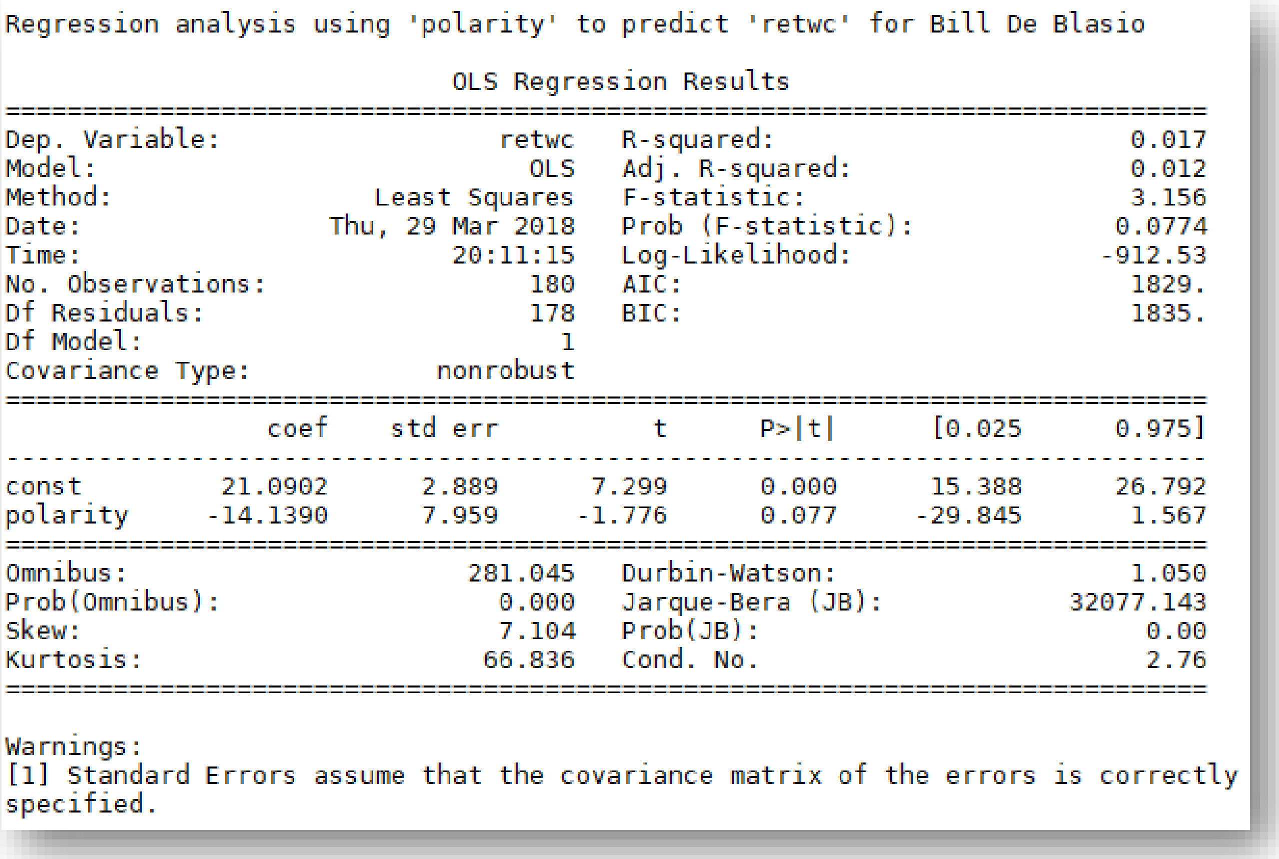
Next, here are the retweet counts for Bill De Blasio:



```
Total retweet count of positive tweets associated with Bill De Blasio:
1737

Total retweet count of negative tweets associated with Bill De Blasio:
1143

Total retweet count of neutral tweets associated with Bill De Blasio:
870

Total retweet count of all tweets associated with Bill De Blasio:
3750
```

In the case of Bill De Blasio, the number of retweets for positive polarity tweets is also higher than the number of retweets for negative polarity tweets, though not as strikingly so as in the case of Phil Murphy. There is also a significant number of retweets for neutral tweets.

Now, let’s look at the regression model to see how well ‘polarity’ predicts ‘retwc’ for Bill De Blasio:

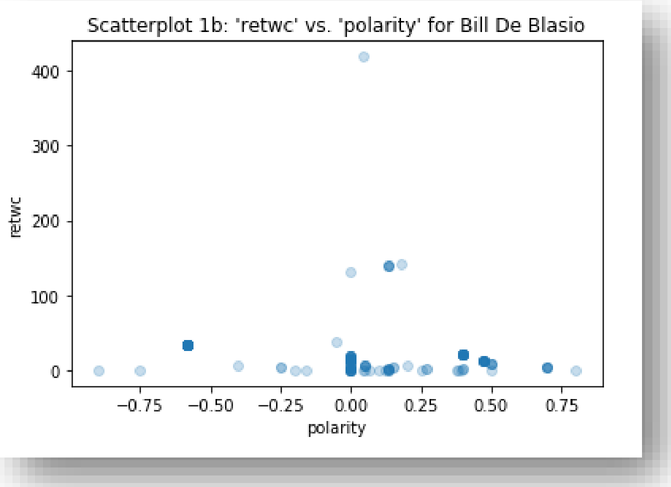


```
Regression analysis using 'polarity' to predict 'retwc' for Bill De Blasio

                            OLS Regression Results
==============================================================================
Dep. Variable:                  retwc   R-squared:                       0.017
Model:                            OLS   Adj. R-squared:                  0.012
Method:                 Least Squares   F-statistic:                     3.156
Date:                Thu, 29 Mar 2018   Prob (F-statistic):             0.0774
Time:                        20:11:15   Log-Likelihood:                -912.53
No. Observations:                 180   AIC:                             1829.
Df Residuals:                     178   BIC:                             1835.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         21.0902      2.889      7.299      0.000      15.388      26.792
polarity     -14.1390      7.959     -1.776      0.077     -29.845       1.567
==============================================================================
Omnibus:                      281.045   Durbin-Watson:                   1.050
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            32077.143
Skew:                           7.104   Prob(JB):                         0.00
Kurtosis:                      66.836   Cond. No.                         2.76
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

Regression equation: predicted retwc = 21.0902 + -14.1390*polarity

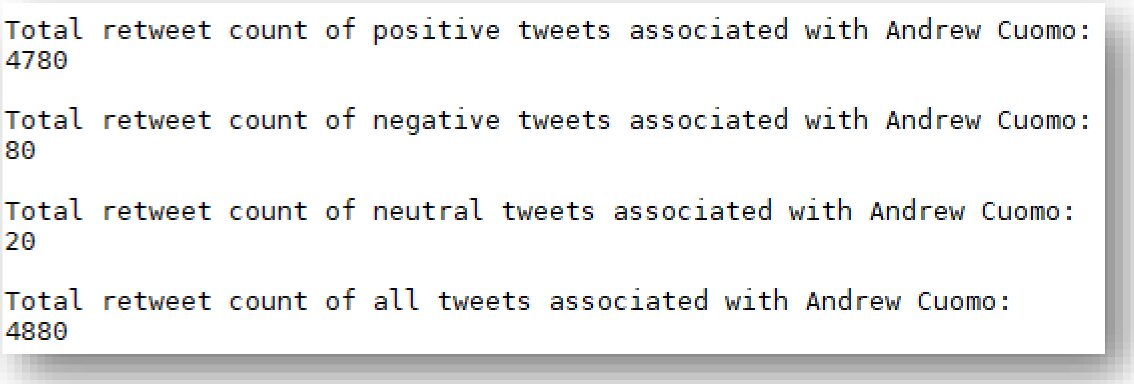
The R-squared value here is higher than in the case of Phil Murphy. However, the R-squared value here is still very low, thus implying that ‘polarity’ doesn’t do a good job at predicting ‘retwc’.

Now let’s look at the scatterplot of ‘retwc’ and ‘polarity’ for any possible insights:



We can see here that several of the points are distributed towards the bottom of the scatterplot in a horizontal fashion. There is also an outlier once again towards the top of the plot. Unlike the case of Phil Murphy, there are some points that are not extreme outliers but do have higher ‘retwc’ values than the low, horizontal-trending points. However, the overall combination of points still results in the low R-squared value.

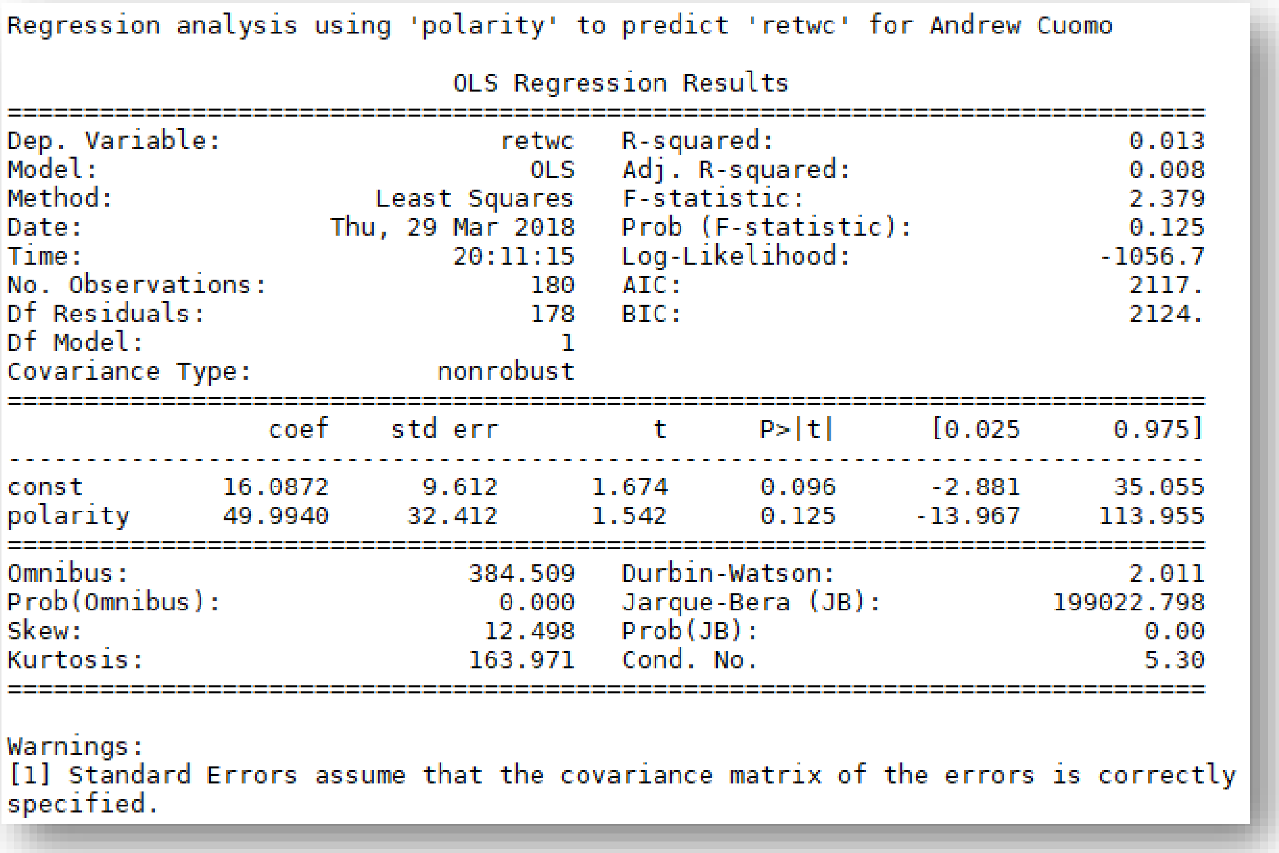
Finally, here are the retweet counts for Andrew Cuomo:



```
Total retweet count of positive tweets associated with Andrew Cuomo:
4780

Total retweet count of negative tweets associated with Andrew Cuomo:
80

Total retweet count of neutral tweets associated with Andrew Cuomo:
20

Total retweet count of all tweets associated with Andrew Cuomo:
4880
```

In the case of Andrew Cuomo, the number of retweets for positive polarity tweets is higher than the number of retweets for negative polarity tweets. However, there is also a significantly low number of retweets for neutral tweets.

Now, let's look at the regression model to see how well 'polarity' predicts 'retwc' for Andrew Cuomo:

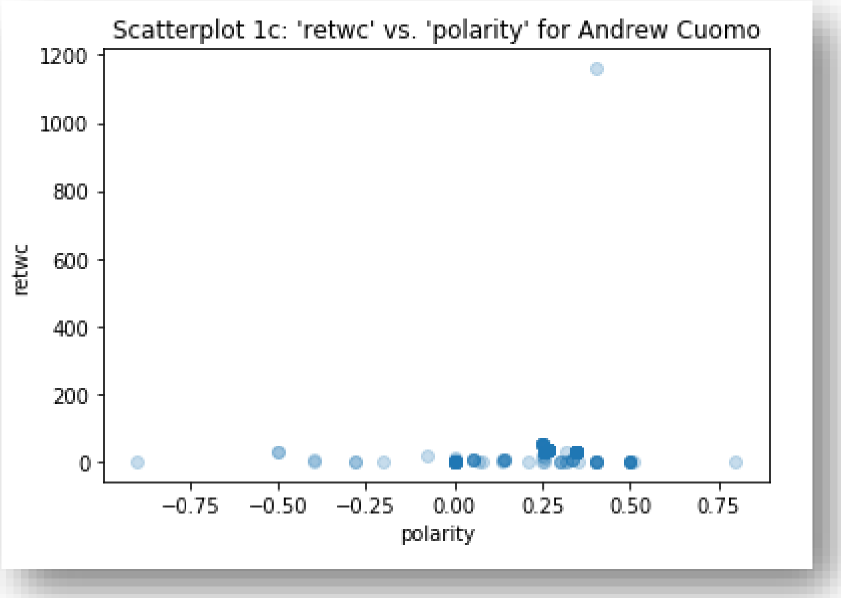


```
Regression analysis using 'polarity' to predict 'retwc' for Andrew Cuomo

                            OLS Regression Results
==============================================================================
Dep. Variable:                  retwc   R-squared:                       0.013
Model:                            OLS   Adj. R-squared:                  0.008
Method:                 Least Squares   F-statistic:                     2.379
Date:                Thu, 29 Mar 2018   Prob (F-statistic):              0.125
Time:                        20:11:15   Log-Likelihood:                -1056.7
No. Observations:                 180   AIC:                             2117.
Df Residuals:                     178   BIC:                             2124.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         16.0872      9.612      1.674      0.096      -2.881      35.055
polarity      49.9940     32.412      1.542      0.125     -13.967     113.955
==============================================================================
Omnibus:                      384.509   Durbin-Watson:                   2.011
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           199022.798
Skew:                          12.498   Prob(JB):                         0.00
Kurtosis:                     163.971   Cond. No.                         5.30
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

Regression equation: predicted retwc = 16.0872 + 49.9940*polarity

The R-squared value here is slightly higher than in the case of Phil Murphy, but lower than in the case of Bill De Blasio. Nevertheless, the R-squared value here is still very low, thus implying that 'polarity' doesn't do a good job at predicting 'retwc'.

Now let's look at the scatterplot of 'retwc' and 'polarity' for any possible insights:



Like in the case of Phil Murphy, we can see here that majority of the points for Andrew Cuomo are distributed towards the bottom of the scatterplot in a horizontal fashion, with an outlier towards the top of the plot. The overall combination of points once again results in the low R-squared value.

Conclusion

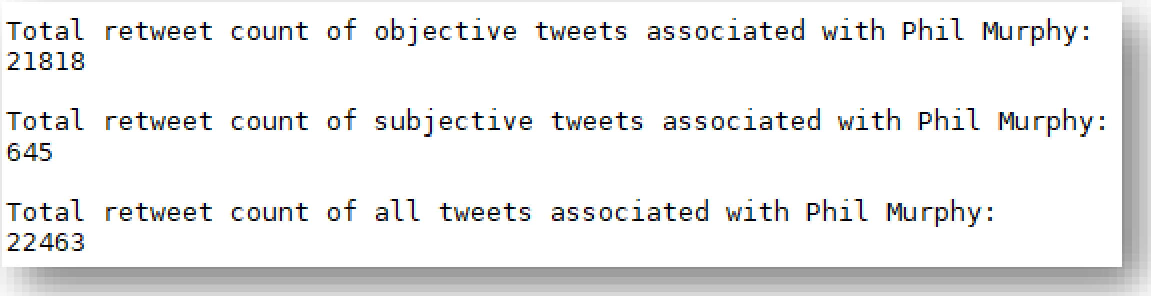
Overall, it would appear that the hypothesis holds true – for a given politician, positive tweets are retweeted more than negative tweets. However, the polarity of a tweet doesn't necessarily predict the number of times the tweet gets retweeted, as evident in the weak linear correlation between 'retwc' and 'polarity', as well as the very low R-squared score in the regression models for each of the politicians.

It is worth noting that the number of tweets used here is only 180 for each politician. However, the p-values associated with the 'polarity' variable in each of the regression models is rather high (above 0.05, which is a common level of significance when performing hypothesis tests), which would make 'polarity' not statistically significant in the regression model. Therefore, more data points won't necessarily improve these regression models.

It is also worth noting that only ‘polarity’ was considered for predicting ‘retwc’. Perhaps different attributes or a different combination of attributes can help provide a better regression model for predicting ‘retwc’.

Another relationship I'd like to explore is the relationship between subjectivity and retweet count. Specifically, I'd like to test the following hypothesis – subjective tweets are retweeted more than objective posts. In this case, objective tweets are tweets with a subjectivity score less than 0.5, while subjective tweets are tweets with subjectivity score greater than or equal to 0.5

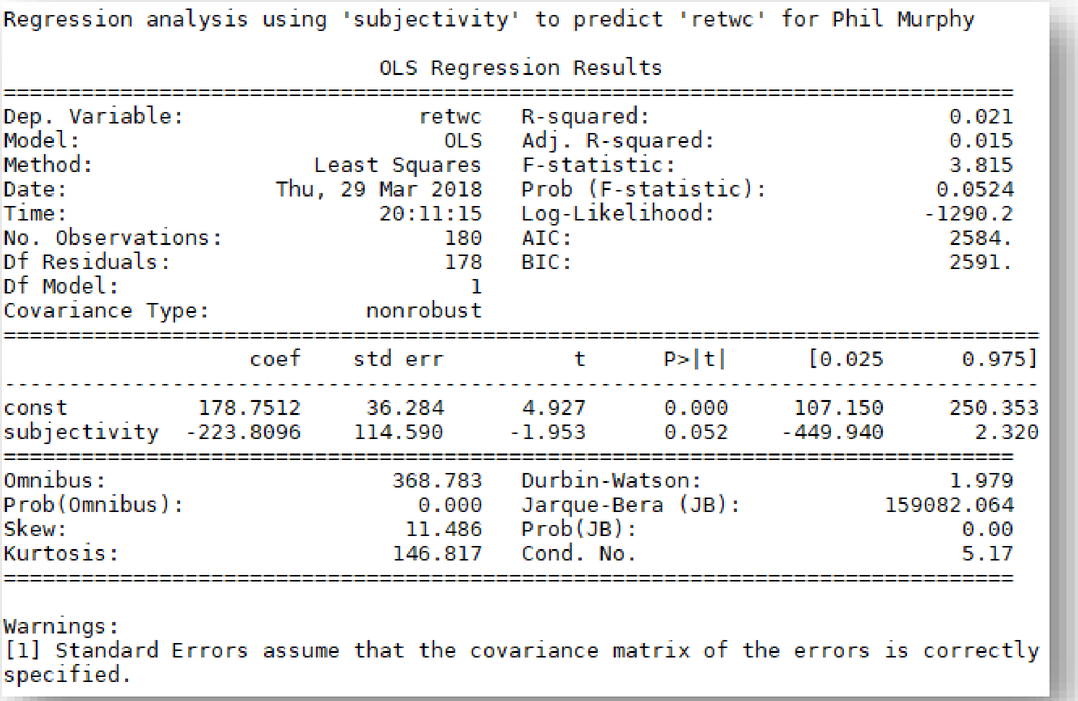
First, here are the retweet counts for Phil Murphy:



```
Total retweet count of objective tweets associated with Phil Murphy:
21818

Total retweet count of subjective tweets associated with Phil Murphy:
645

Total retweet count of all tweets associated with Phil Murphy:
22463
```

Here we can see that in the case of Phil Murphy, objective tweets are retweeted far more than subjective tweets.

Now let's look at the regression model to see how well 'subjectivity' predicts 'retwc' for Phil Murphy:

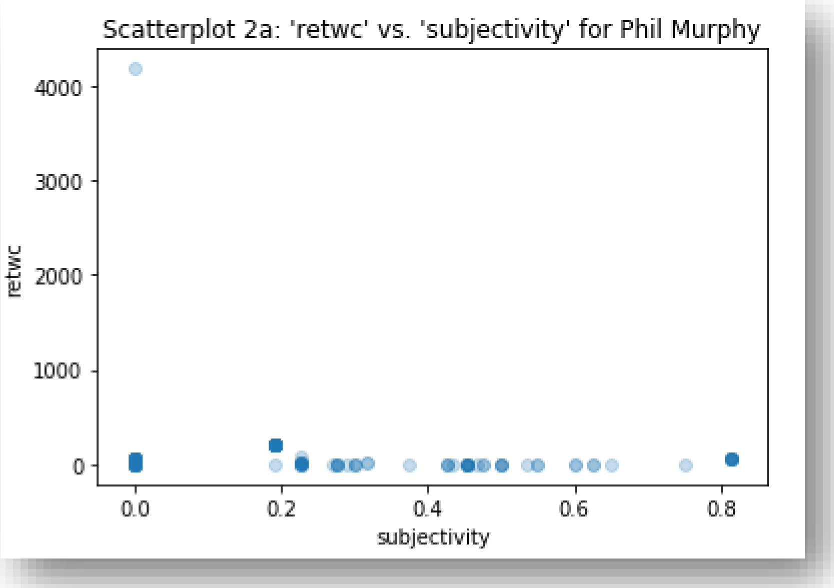


```
Regression analysis using 'subjectivity' to predict 'retwc' for Phil Murphy

                            OLS Regression Results
==============================================================================
Dep. Variable:                  retwc   R-squared:                       0.021
Model:                            OLS   Adj. R-squared:                  0.015
Method:                 Least Squares   F-statistic:                     3.815
Date:                Thu, 29 Mar 2018   Prob (F-statistic):             0.0524
Time:                        20:11:15   Log-Likelihood:                -1290.2
No. Observations:                 180   AIC:                             2584.
Df Residuals:                     178   BIC:                             2591.
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const          178.7512     36.284      4.927      0.000     107.150     250.353
subjectivity  -223.8096    114.590     -1.953      0.052    -449.940       2.320
==============================================================================
Omnibus:                      368.783   Durbin-Watson:                   1.979
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           159082.064
Skew:                          11.486   Prob(JB):                         0.00
Kurtosis:                     146.817   Cond. No.                         5.17
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

Regression equation: predicted retwc = 178.7512 + -223.8096*subjectivity

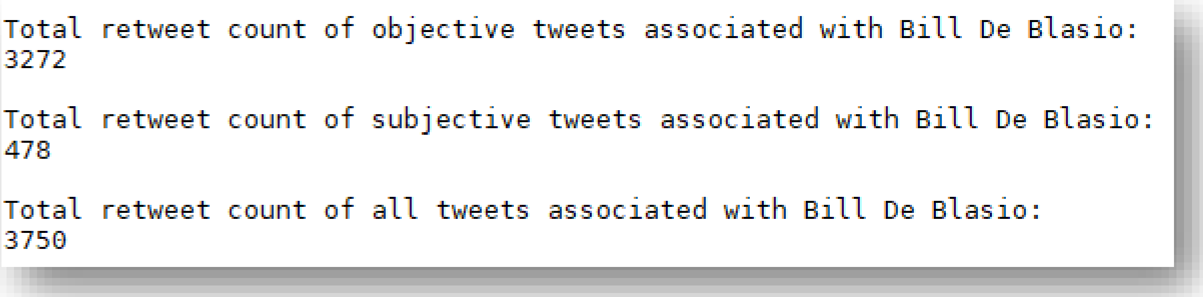
Judging by the R-squared value, it doesn't seem that 'subjectivy does a good job at predicting 'retwc'.

Now let’s look at the scatterplot of ‘retwc’ and ‘subjectivity’ for any possible insights:



Like in the case of ‘retwc’ vs. ‘polarity’ for Phil Murphy, we can see here that majority of the points are distributed towards the bottom of the scatterplot in a horizontal fashion, with the exception being that one point towards the top. The overall combination of points results in the low R-squared value.

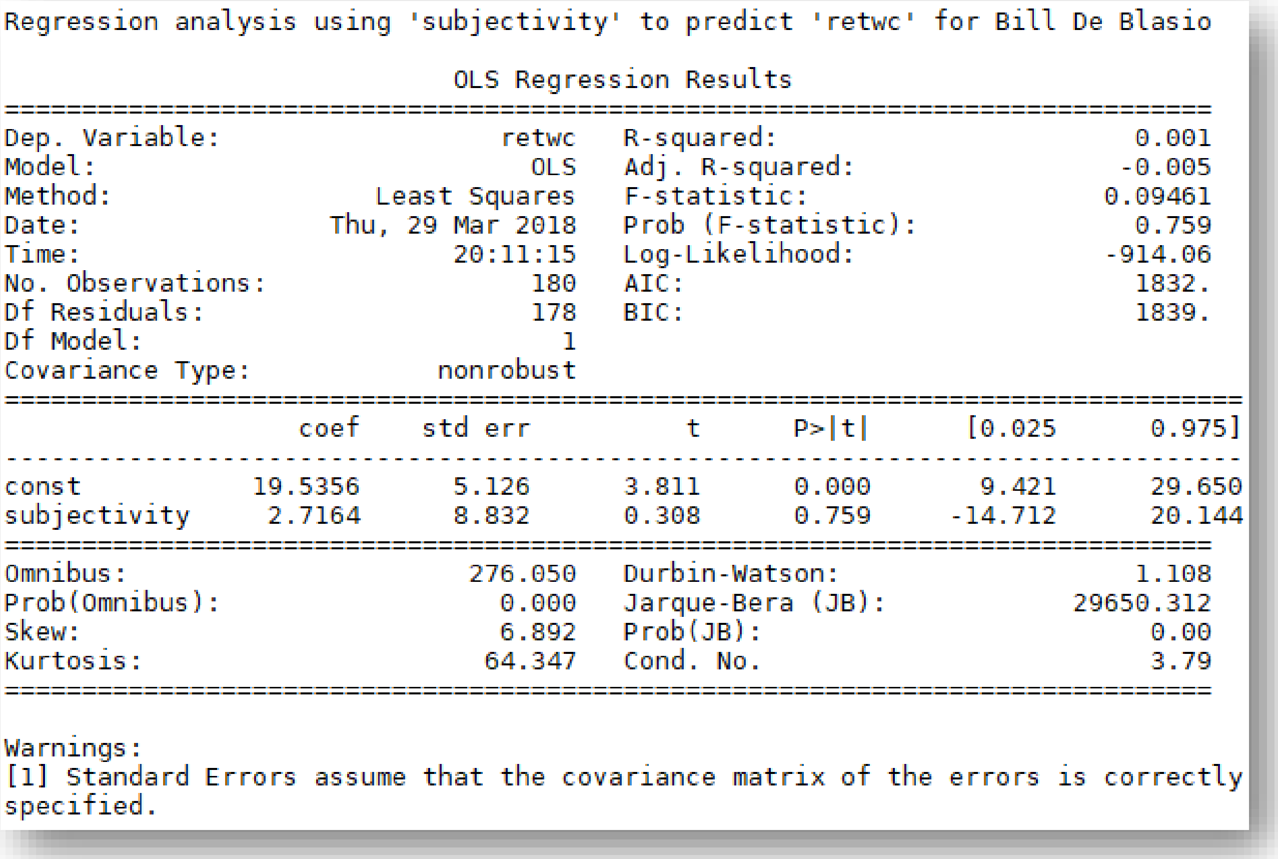
Next, here are the retweet counts for Bill De Blasio:



```
Total retweet count of objective tweets associated with Bill De Blasio:
3272

Total retweet count of subjective tweets associated with Bill De Blasio:
478

Total retweet count of all tweets associated with Bill De Blasio:
3750
```

Like in the case for Phil Murphy, we can see here that in the case of Bill De Blasio, objective tweets are retweeted far more often than subjective tweets.

Now, let's look at the regression model to see how well 'subjectivity' predicts 'retwc' for Bill De Blasio:

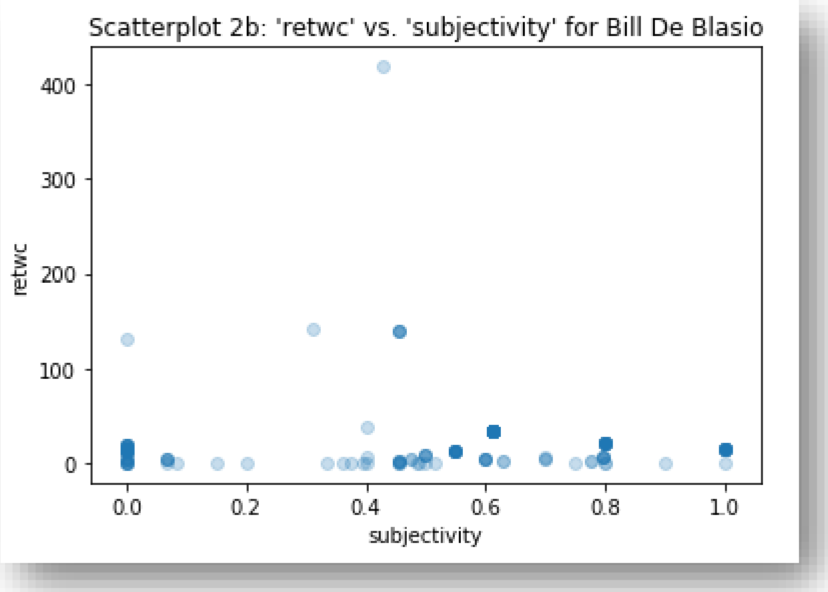


```
Regression analysis using 'subjectivity' to predict 'retwc' for Bill De Blasio

                            OLS Regression Results
==============================================================================
Dep. Variable:                  retwc   R-squared:                       0.001
Model:                            OLS   Adj. R-squared:                 -0.005
Method:                 Least Squares   F-statistic:                   0.09461
Date:                Thu, 29 Mar 2018   Prob (F-statistic):              0.759
Time:                        20:11:15   Log-Likelihood:                -914.06
No. Observations:                 180   AIC:                             1832.
Df Residuals:                     178   BIC:                             1839.
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const           19.5356      5.126      3.811      0.000       9.421      29.650
subjectivity     2.7164      8.832      0.308      0.759     -14.712      20.144
==============================================================================
Omnibus:                      276.050   Durbin-Watson:                   1.108
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            29650.312
Skew:                           6.892   Prob(JB):                         0.00
Kurtosis:                      64.347   Cond. No.                         3.79
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

Regression equation: predicted retwc = 19.5356 + 2.7164*subjectivity

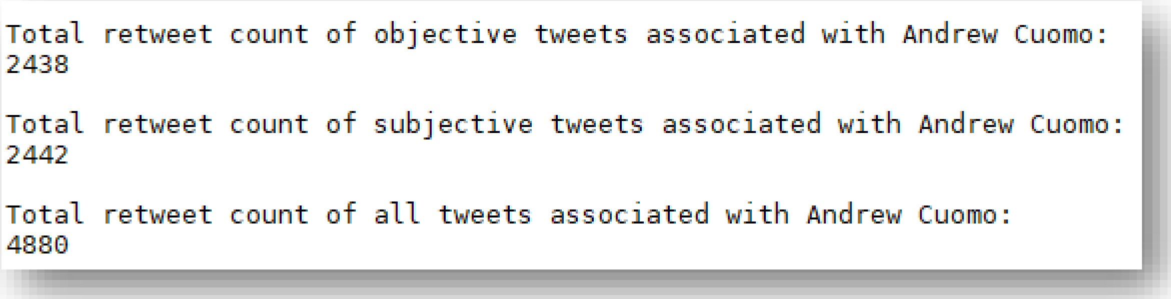
Judging by the R-squared value, it doesn't seem that 'subjectivity does a good job at predicting 'retwc' in the case of Bill De Blasio as well.

Now let’s look at the scatterplot of ‘retwc’ and ‘subjectivity’ for any possible insights:



Like in the case of ‘retwc’ vs. ‘polarity’ for Bill De Blasio, we can see here that several of the points are distributed towards the bottom of the scatterplot in a horizontal fashion. There is an outlier once again towards the top of the plot. There are also some points that are not extreme outliers but do have higher ‘retwc’ values than the low, horizontal-trending points. However, the overall combination of points still results in the low R-squared value.

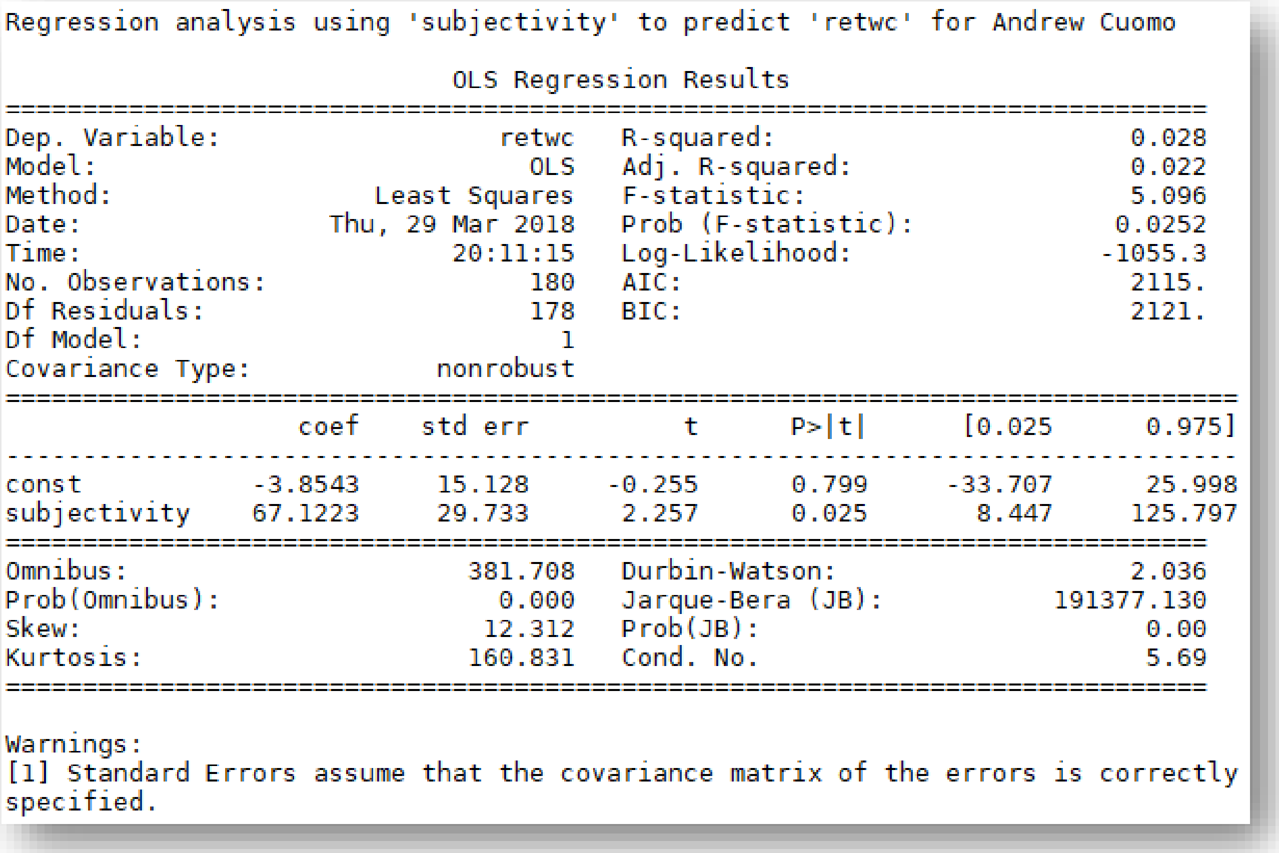
Finally, here are the retweet counts for Andrew Cuomo:



```
Total retweet count of objective tweets associated with Andrew Cuomo:
2438

Total retweet count of subjective tweets associated with Andrew Cuomo:
2442

Total retweet count of all tweets associated with Andrew Cuomo:
4880
```

Unlike in the cases of Phil Murphy or Bill De Blasio, we can see here that for Andrew Cuomo, subjective tweets are retweeted more than objective tweets. However, the difference between the retweet counts is relatively small.

Now, let’s look at the regression model to see how well ‘subjectivity’ predicts ‘retwc’ for Andrew Cuomo:

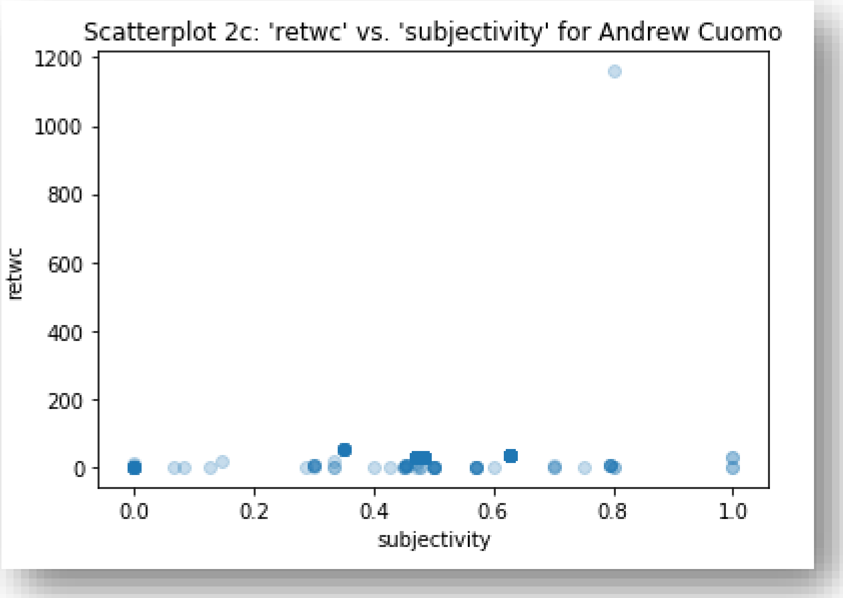


```
Regression analysis using 'subjectivity' to predict 'retwc' for Andrew Cuomo

                            OLS Regression Results
==============================================================================
Dep. Variable:                  retwc   R-squared:                       0.028
Model:                            OLS   Adj. R-squared:                  0.022
Method:                 Least Squares   F-statistic:                     5.096
Date:                Thu, 29 Mar 2018   Prob (F-statistic):             0.0252
Time:                        20:11:15   Log-Likelihood:                -1055.3
No. Observations:                 180   AIC:                             2115.
Df Residuals:                     178   BIC:                             2121.
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const           -3.8543     15.128     -0.255      0.799     -33.707      25.998
subjectivity    67.1223     29.733      2.257      0.025       8.447     125.797
==============================================================================
Omnibus:                      381.708   Durbin-Watson:                   2.036
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           191377.130
Skew:                          12.312   Prob(JB):                         0.00
Kurtosis:                     160.831   Cond. No.                         5.69
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

Regression equation: predicted retwc = -3.8543 + 67.1223*subjectivity

Judging by the R-squared value, it doesn't seem that 'subjectivity does a good job at predicting 'retwc' in the case of Andrew Cuomo.

Now let's look at the scatterplot of 'retwc' and 'subjectivity' for any possible insights:



Like in the case of 'retwc' vs. 'polarity' for Andrew Cuomo, we can see here that majority of the points are distributed towards the bottom of the scatterplot in a horizontal fashion, with an outlier towards the top of the plot. The overall combination of points once again results in the low R-squared value.

Conclusion

Overall, it would appear that the hypothesis doesn't hold true – for a given politician, it is possible for negative tweets to be retweeted more than positive tweets. Moreover, like the case of polarity, the subjectivity of a tweet doesn't necessarily predict the number of times a tweet gets retweeted, as evident in the weak linear correlation between 'retwc' and 'subjectivity, as well as the very low R-squared score in the regression models for each of the politicians.

It is worth noting that the number of tweets used here is still only 180 for each politician. However, unlike the case of polarity where associated p-values were greater than 0.05, the p-values associated with the 'subjectivity' variable in each of the regression models varied from below 0.5 to near 0.5 to even above 0.5. Nevertheless, this doesn't make 'subjectivity' a good predictor variable since it could still be considered statistically insignificant, so more data points won't necessarily improve these regression models.

It is also worth noting that only 'subjectivity' was considered for predicting 'retwc' in this case. Perhaps different attributes (aside from 'polarity') or a different combination of attributes can help provide a better regression model for predicting 'retwc'.

Python Script

```
# -*- coding: utf-8 -*-
"""
Created on Wed Mar 28 17:46:46 2018

@author: Brian
"""

import matplotlib.pyplot as plt
import pandas as pd

# load the rest of the "standard" set of packages
import numpy as np
import statsmodels.api as sm
import matplotlib.pyplot as plt

# correlation for the first Twitter search - Phil Murphy
twitter_data_phil_murphy = pd.read_csv('results_phil_murphy.csv')
print("\nCorrelation matrix for variables associated with tweets relating to Phil Murphy\n")
print(twitter_data_phil_murphy.corr())

# positive polarity tweets for Phil Murphy
twitter_data_phil_murphy_positive = twitter_data_phil_murphy[
    0 < twitter_data_phil_murphy['polarity']
]
print("\nTotal retweet count of positive tweets associated with Phil Murphy:")
print(twitter_data_phil_murphy_positive.retwc.sum())

# negative polarity tweets for Phil Murphy
twitter_data_phil_murphy_negative = twitter_data_phil_murphy[
    twitter_data_phil_murphy['polarity'] < 0
]
print("\nTotal retweet count of negative tweets associated with Phil Murphy:")
print(twitter_data_phil_murphy_negative.retwc.sum())

# neutral polarity tweets for Phil Murphy
twitter_data_phil_murphy_neutral = twitter_data_phil_murphy[
    0 == twitter_data_phil_murphy['polarity']
]
print("\nTotal retweet count of neutral tweets associated with Phil Murphy:")
print(twitter_data_phil_murphy_neutral.retwc.sum())

# all tweets for Phil Murphy
print("\nTotal retweet count of all tweets associated with Phil Murphy:")
print(twitter_data_phil_murphy.retwc.sum())

# regression model using 'polarity' to predict 'retwc'
```

```
phil_murphy_response_1 = twitter_data_phil_murphy.retwc
phil_murphy_predictor_1 = twitter_data_phil_murphy['polarity']
phil_murphy_predictor_1 = sm.add_constant(phil_murphy_predictor_1)
phil_murphy_regression_model_1 = sm.OLS(phil_murphy_response_1,
                                        phil_murphy_predictor_1).fit()
print("\nRegression analysis using 'polarity' to predict 'retwc' for Phil Murphy")
print("\n", phil_murphy_regression_model_1.summary())

# scatterplot of 'retwc' vs. 'polarity'
plt.figure()
plt.scatter(twitter_data_phil_murphy.polarity,
            twitter_data_phil_murphy.retwc, alpha = 0.25)
plt.title("Scatterplot 1a: 'retwc' vs. 'polarity' for Phil Murphy")
plt.xlabel("polarity")
plt.ylabel("retwc")

# ----

print('\n# ----') # an arbitrary divider to make reading the output a little easier

# correlation for the second Twitter search - Bill De Blasio
twitter_data_bill_de_blasio = pd.read_csv('results_bill_de_blasio.csv')
print("\nCorrelation matrix for variables associated with tweets relating to Bill De Blasio\n")
print(twitter_data_bill_de_blasio.corr())

# positive polarity tweets for Bill De Blasio
twitter_data_bill_de_blasio_positive = twitter_data_bill_de_blasio[
    0 < twitter_data_bill_de_blasio['polarity']
]
print("\nTotal retweet count of positive tweets associated with Bill De Blasio:")
print(twitter_data_bill_de_blasio_positive.retwc.sum())

# negative polarity tweets for Bill De Blasio
twitter_data_bill_de_blasio_negative = twitter_data_bill_de_blasio[
    twitter_data_bill_de_blasio['polarity'] < 0
]
print("\nTotal retweet count of negative tweets associated with Bill De Blasio:")
print(twitter_data_bill_de_blasio_negative.retwc.sum())

# neutral polarity tweets for Bill De Blasio
twitter_data_bill_de_blasio_neutral = twitter_data_bill_de_blasio[
    0 == twitter_data_bill_de_blasio['polarity']
]
print("\nTotal retweet count of neutral tweets associated with Bill De Blasio:")
print(twitter_data_bill_de_blasio_neutral.retwc.sum())

# all tweets for Bill De Blasio
print("\nTotal retweet count of all tweets associated with Bill De Blasio:")
print(twitter_data_bill_de_blasio.retwc.sum())

# regression model using 'polarity' to predict 'retwc'
bill_de_blasio_response_1 = twitter_data_bill_de_blasio.retwc
bill_de_blasio_predictor_1 = twitter_data_bill_de_blasio['polarity']
bill_de_blasio_predictor_1 = sm.add_constant(bill_de_blasio_predictor_1)
bill_de_blasio_regression_model_1 = sm.OLS(bill_de_blasio_response_1,
                    bill_de_blasio_predictor_1).fit()
print("\nRegression analysis using 'polarity' to predict 'retwc' for Bill De Blasio")
print("\n", bill_de_blasio_regression_model_1.summary())

# scatterplot of 'retwc' vs. 'polarity'
plt.figure()
plt.scatter(twitter_data_bill_de_blasio.polarity,
        twitter_data_bill_de_blasio.retwc, alpha = 0.25)
plt.title("Scatterplot 1b: 'retwc' vs. 'polarity' for Bill De Blasio")
plt.xlabel("polarity")
plt.ylabel("retwc")

# ----

print('\n# ----') # an arbitrary divider to make reading the output a little easier

# correlation for the third Twitter search - Andrew Cuomo
twitter_data_andrew_cuomo = pd.read_csv('results_andrew_cuomo.csv')
print("\nCorrelation matrix for variables associated with tweets relating to Andrew Cuomo\n")
print(twitter_data_andrew_cuomo.corr())

# positive polarity tweets for Andrew Cuomo
twitter_data_andrew_cuomo_positive = twitter_data_andrew_cuomo[
    0 < twitter_data_andrew_cuomo['polarity']
]
print("\nTotal retweet count of positive tweets associated with Andrew Cuomo:")
print(twitter_data_andrew_cuomo_positive.retwc.sum())

# negative polarity tweets for Andrew Cuomo
twitter_data_andrew_cuomo_negative = twitter_data_andrew_cuomo[
    twitter_data_andrew_cuomo['polarity'] < 0
]
print("\nTotal retweet count of negative tweets associated with Andrew Cuomo:")
print(twitter_data_andrew_cuomo_negative.retwc.sum())

# neutral polarity tweets for Andrew Cuomo
twitter_data_andrew_cuomo_neutral = twitter_data_andrew_cuomo[
    0 == twitter_data_andrew_cuomo['polarity']
]
print("\nTotal retweet count of neutral tweets associated with Andrew Cuomo:")
print(twitter_data_andrew_cuomo_neutral.retwc.sum())
```

```
# all tweets for Andrew Cuomo
print("\nTotal retweet count of all tweets associated with Andrew Cuomo:")
print(twitter_data_andrew_cuomo.retwc.sum())

# regression model using 'polarity' to predict 'retwc'
andrew_cuomo_response_1 = twitter_data_andrew_cuomo.retwc
andrew_cuomo_predictor_1 = twitter_data_andrew_cuomo['polarity']
andrew_cuomo_predictor_1 = sm.add_constant(andrew_cuomo_predictor_1)
andrew_cuomo_regression_model_1 = sm.OLS(andrew_cuomo_response_1,
                        andrew_cuomo_predictor_1).fit()
print("\nRegression analysis using 'polarity' to predict 'retwc' for Andrew Cuomo")
print("\n", andrew_cuomo_regression_model_1.summary())

# scatterplot of 'retwc' vs. 'polarity'
plt.figure()
plt.scatter(twitter_data_andrew_cuomo.polarity,
        twitter_data_andrew_cuomo.retwc, alpha = 0.25)
plt.title("Scatterplot 1c: 'retwc' vs. 'polarity' for Andrew Cuomo")
plt.xlabel("polarity")
plt.ylabel("retwc")

# ----

print('\n# ----') # an arbitrary divider to make reading the output a little easier

# objective tweets for Phil Murphy
twitter_data_phil_murphy_objective = twitter_data_phil_murphy[
        twitter_data_phil_murphy['subjectivity'] < 0.5
]
print("\nTotal retweet count of objective tweets associated with Phil Murphy:")
print(twitter_data_phil_murphy_objective.retwc.sum())

# subjective tweets for Phil Murphy
twitter_data_phil_murphy_subjective = twitter_data_phil_murphy[
        0.5 <= twitter_data_phil_murphy['subjectivity']
]
print("\nTotal retweet count of subjective tweets associated with Phil Murphy:")
print(twitter_data_phil_murphy_subjective.retwc.sum())

# all tweets for Phil Murphy
print("\nTotal retweet count of all tweets associated with Phil Murphy:")
print(twitter_data_phil_murphy.retwc.sum())

# regression model using 'subjectivity' to predict 'retwc'
phil_murphy_response_2 = twitter_data_phil_murphy.retwc
phil_murphy_predictor_2 = twitter_data_phil_murphy['subjectivity']
phil_murphy_predictor_2 = sm.add_constant(phil_murphy_predictor_2)
```

```
phil_murphy_regression_model_2 = sm.OLS(phil_murphy_response_2,
                        phil_murphy_predictor_2).fit()
print("\nRegression analysis using 'subjectivity' to predict 'retwc' for Phil Murphy")
print("\n", phil_murphy_regression_model_2.summary())

# scatterplot of 'retwt' vs. 'subjectivity' for Phil Murphy
plt.figure()
plt.scatter(twitter_data_phil_murphy.subjectivity,
        twitter_data_phil_murphy.retwc, alpha = 0.25)
plt.title("Scatterplot 2a: 'retwc' vs. 'subjectivity' for Phil Murphy")
plt.xlabel("subjectivity")
plt.ylabel("retwc")

# ----

print('\n# ----') # an arbitrary divider to make reading the output a little easier

# objective tweets for Bill De Blasio
twitter_data_bill_de_blasio_objective = twitter_data_bill_de_blasio[
        twitter_data_phil_murphy['subjectivity'] < 0.5
]
print("\nTotal retweet count of objective tweets associated with Bill De Blasio:")
print(twitter_data_bill_de_blasio_objective.retwc.sum())

# subjective tweets for Bill De Blasio
twitter_data_bill_de_blasio_subjective = twitter_data_bill_de_blasio[
        0.5 <= twitter_data_phil_murphy['subjectivity']
]
print("\nTotal retweet count of subjective tweets associated with Bill De Blasio:")
print(twitter_data_bill_de_blasio_subjective.retwc.sum())

# all tweets for Bill De Blasio
print("\nTotal retweet count of all tweets associated with Bill De Blasio:")
print(twitter_data_bill_de_blasio.retwc.sum())

# regression model using 'subjectivity' to predict 'retwc'
bill_de_blasio_response_2 = twitter_data_bill_de_blasio.retwc
bill_de_blasio_predictor_2 = twitter_data_bill_de_blasio['subjectivity']
bill_de_blasio_predictor_2 = sm.add_constant(bill_de_blasio_predictor_2)
bill_de_blasio_regression_model_2 = sm.OLS(bill_de_blasio_response_2,
                        bill_de_blasio_predictor_2).fit()
print("\nRegression analysis using 'subjectivity' to predict 'retwc' for Bill De Blasio")
print("\n", bill_de_blasio_regression_model_2.summary())

# scatterplot of 'retwt' vs. 'subjectivity' for Bill De Blasio
plt.figure()
plt.scatter(twitter_data_bill_de_blasio.subjectivity,
        twitter_data_bill_de_blasio.retwc, alpha = 0.25)
```

```
plt.title("Scatterplot 2b: 'retwc' vs. 'subjectivity' for Bill De Blasio")
plt.xlabel("subjectivity")
plt.ylabel("retwc")

# ----

print('\n# ----') # an arbitrary divider to make reading the output a little easier

# objective tweets for Andrew Cuomo
twitter_data_andrew_cuomo_objective = twitter_data_andrew_cuomo[
        twitter_data_andrew_cuomo['subjectivity'] < 0.5
]
print("\nTotal retweet count of objective tweets associated with Andrew Cuomo:")
print(twitter_data_andrew_cuomo_objective.retwc.sum())

# subjective tweets for Andrew Cuomo
twitter_data_andrew_cuomo_subjective = twitter_data_andrew_cuomo[
        0.5 <= twitter_data_andrew_cuomo['subjectivity']
]
print("\nTotal retweet count of subjective tweets associated with Andrew Cuomo:")
print(twitter_data_andrew_cuomo_subjective.retwc.sum())

# all tweets for Andrew Cuomo
print("\nTotal retweet count of all tweets associated with Andrew Cuomo:")
print(twitter_data_andrew_cuomo.retwc.sum())

# regression model using 'subjectivity' to predict 'retwc'
andrew_cuomo_response_2 = twitter_data_andrew_cuomo.retwc
andrew_cuomo_predictor_2 = twitter_data_andrew_cuomo['subjectivity']
andrew_cuomo_predictor_2 = sm.add_constant(andrew_cuomo_predictor_2)
andrew_cuomo_regression_model_2 = sm.OLS(andrew_cuomo_response_2,
                      andrew_cuomo_predictor_2).fit()
print("\nRegression analysis using 'subjectivity' to predict 'retwc' for Andrew Cuomo")
print("\n", andrew_cuomo_regression_model_2.summary())

# scatterplot of 'retwt' vs. 'subjectivity' for Andrew Cuomo
plt.figure()
plt.scatter(twitter_data_andrew_cuomo.subjectivity,
        twitter_data_andrew_cuomo.retwc, alpha = 0.25)
plt.title("Scatterplot 2c: 'retwc' vs. 'subjectivity' for Andrew Cuomo")
plt.xlabel("subjectivity")
plt.ylabel("retwc")
```

References

1. Information on installing tweepy for Anaconda via Anaconda Prompt was taken from the following webpage: https://www.youtube.com/watch?v=GqdgxA_Bgz8
2. Information on installing tweepy for Python via command prompt was taken from the video podcast, 10.1 - Using Python to collect Twitter data, by Dr. Chirag Shah
3. Information on resolving a 'charmap' codec error in the sentiment script was taken from the following webpage: https://stackoverflow.com/questions/27092833/unicodeencodeerror-charmap-codec-cant-encode-characters
4. Code for this assignment was based off of notes taken for the video podcasts, 10.1 - Using Python to collect Twitter data, and 10.2 - Sentiment analysis with Twitter data, both of which by Dr. Chirag Shah
5. Information on running a Python script through the command prompt was taken from the video podcast, 10.1 - Using Python to collect Twitter data, by Dr. Chirag Shah
6. Information on collecting more Twitter results with tweepy was taken from the following webpage: https://stackoverflow.com/questions/17371652/tweepy-twitter-api-not-returning-all-search-results
7. Python code for correlation, scatterplots, and regression was based off of notes taken for the video podcast, 4.2 Statistical Analysis with Python, by Dr. Chirag Shah
8. Examples of Python code on correlation, scatterplots, and regression were taken from my Week 4 homework assignment on regression models using Python, as well as my Week 8 midterm assignment
9. Information on textblob and sentiment was taken from the following webpage: https://planspace.org/20150607-textblob_sentiment/
10. Information on interpreting correlation coefficients was taken from the following webpage: http://www.dummies.com/education/math/statistics/how-to-interpret-a-correlation-coefficient-r/
11. Information on finding the sum of the values in a column of a pandas data frame was taken from the following webpage: https://stackoverflow.com/questions/41286569/get-total-of-pandas-column
12. Information on logical operators in Python was taken from the following webpage: https://thomas-cokelaer.info/tutorials/python/boolean.html
13. Information on the alpha attribute when plotting with matplotlib.pylot in Python was taken from the following webpage: https://matplotlib.org/api/pyplot_api.html
14. Information on how to nest one Python script inside another Python script was taken from the file, Handout – PS with Twitter, by Dr. Chirag Shah.
15. Information on how to interpret the values for both polarity and subjectivity was taken from the file, Slides – Data problem with Twitter, by Dr. Chirag Shah.
16. Information on the calculation of R-squared was taken from the following webpages:
    a. https://en.wikipedia.org/wiki/Coefficient_of_determination
    b. https://onlinecourses.science.psu.edu/stat501/node/255
17. Information on the calculation of the "line of best fit" using the least squares method was taken from the following webpages:
    a. https://onlinecourses.science.psu.edu/stat501/node/252
    b. https://www.varsitytutors.com/hotmath/hotmath_help/topics/line-of-best-fit

18. Information on statistical significance was taken from the following webpage: https://en.wikipedia.org/wiki/Statistical_significance
19. Information on Phil Murphy was taken from the following webpage: https://en.wikipedia.org/wiki/Phil_Murphy
20. Information on Bill De Blasio was taken from the following webpage: https://en.wikipedia.org/wiki/Bill_de_Blasio
21. Information on Andrew Cuomo was taken from the following webpage: https://en.wikipedia.org/wiki/Andrew_Cuomo