Kevin Wang (604036846)

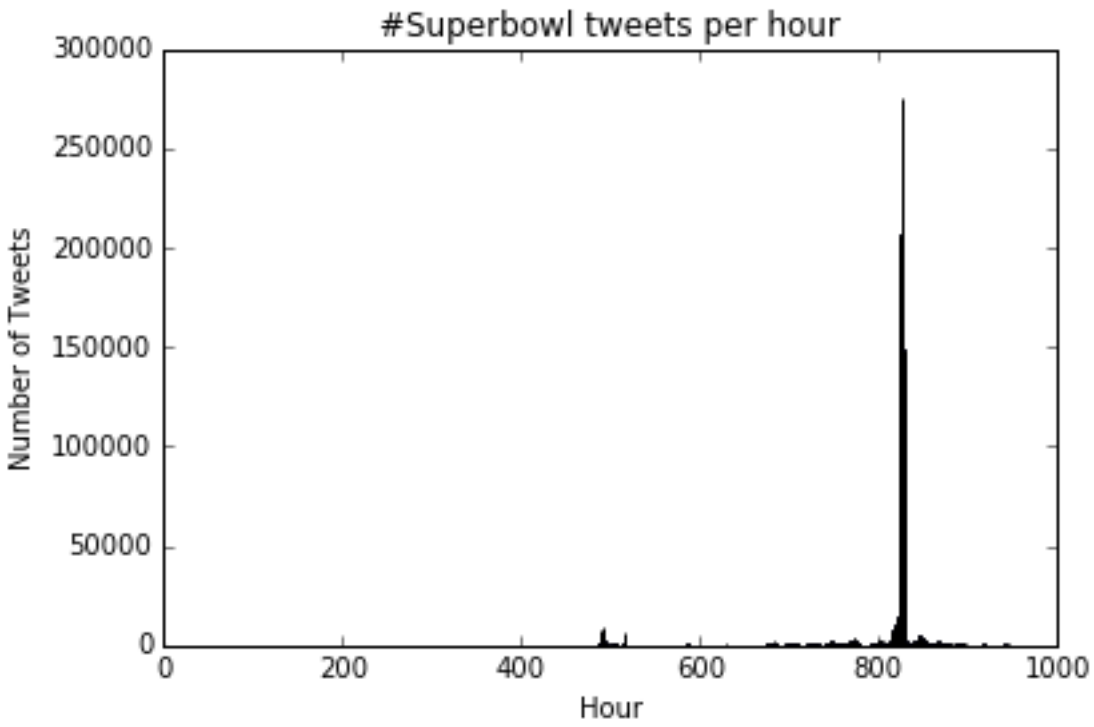
EE 239AS Project 4

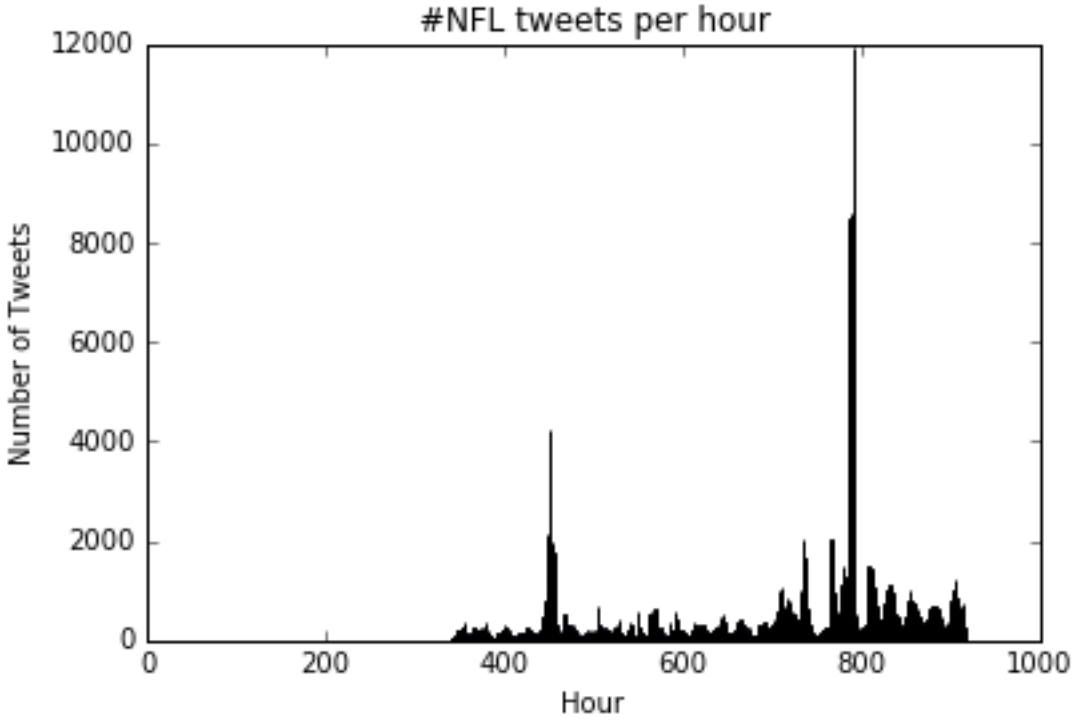
Part 1

Each hashtag data file is iterated through to determine the average number of followers, retweets, and tweets per hour. Since “firstpost_date” is in an int and each number is a second, bin widths of 3600 are used for the histograms and calculating tweets per hour. The complete data for each hashtag could not be stored for all of the hashtags since the files were too large.

Hashtag	Average tweets/hour	Average followers	Average retweets	Average Citations
#Superbowl	1401.245593656886	8858.974662784603	0.13668558023735752	2.3882723999030224
#sb49	1419.8879074871902	10267.31684948685	0.1780129657017163	2.5111487863247035
#Patriots	499.42105160280914	3309.978828415827	0.09146173370933587	1.7828156491659402
#nfl	279.55138019452266	4653.252285502502	0.05093736487738588	1.5385331089011056
#gopatriots	38.38470386915044	1401.8955093016164	0.02683745044220799	1.4000838670326319
#gohawks	193.54482518973285	2203.931767444827	0.20916252072968491	2.014617085512608

The total amount of hours each hashtag data spans is found by finding the difference in firstpost_date between the first and last tweet of each hashtag. This number is then used to calculate average tweets per hour and to specify the amount of bins for the #superbowl and #nfl histograms. The average number of followers was found using the tweet[‘author’][‘followers’] parameter of each tweet. The average retweets was found using the tweet[‘tweet’][‘retweet_count’] parameter of each tweet. The average citations/retweet count is found using the specified tweet['metrics']['citations']['total'] parameter.


#Superbowl tweets per hour
Number of Tweets
Hour
0
50000
100000
150000
200000
250000
300000
0
200
400
600
800
1000


#NFL tweets per hour
Number of Tweets
Hour
0
2000
4000
6000
8000
10000
12000
0
200
400
600
800
1000

Part 2

The required data is first compiled in python before being saved in a csv file. The compiled data is then loaded into R such that linear regression analysis could be performed easier. This is done for each hashtag separately. Note that since we are predicting the next hour's data, the last hour in the samples must be dropped for each part here and in the future since there is no "next hour" data to use.

#superbowl

```
Residuals:
   Min     1Q Median     3Q    Max
-84986   -600    256    498 148926

                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -2.656e+02  5.864e+02  -0.453  0.65069
superbowl_f1$tweets_cur  1.337e+00  2.777e-01   4.814 1.82e-06 ***
superbowl_f1$tot_ret     4.016e-01  1.378e-01   2.914  0.00369 **
superbowl_f1$tot_fol    -2.451e-04  1.420e-05 -17.262  < 2e-16 ***
superbowl_f1$max_fol     1.128e-03  1.245e-04   9.065  < 2e-16 ***
superbowl_f1$time_hour  -2.107e+01  4.226e+01  -0.498  0.61829
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7693 on 692 degrees of freedom
Multiple R-squared:  0.7833,   Adjusted R-squared:  0.7818
F-statistic: 500.4 on 5 and 692 DF,  p-value: < 2.2e-16
```

#sb49

```
Residuals:
   Min     1Q Median     3Q    Max
-29062   -224    -89     62  90610

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.451e+02  3.668e+02   0.668   0.5043
sb49_f1$tweets_cur  1.090e+00  9.927e-02  10.983   <2e-16 ***
sb49_f1$tot_ret    -1.171e-01  9.112e-02  -1.285   0.1994
sb49_f1$tot_fol     3.650e-06  1.448e-05   0.252   0.8011
sb49_f1$max_fol     1.006e-04  4.833e-05   2.081   0.0379 *
sb49_f1$time_hour  -1.920e+01  2.725e+01  -0.705   0.4814
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4511 on 575 degrees of freedom
Multiple R-squared:  0.8038,   Adjusted R-squared:  0.8021
F-statistic: 471.1 on 5 and 575 DF,  p-value: < 2.2e-16
```

#patriots

```
Residuals:
   Min     1Q Median     3Q    Max
-17350   -184   -155   -109  42989

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              1.908e+02  1.316e+02   1.450    0.147
patriots_f1$tweets_cur   1.785e+00  8.570e-02  20.824  < 2e-16 ***
patriots_f1$tot_ret     -8.648e-01  7.189e-02 -12.030  < 2e-16 ***
patriots_f1$tot_fol      1.621e-04  2.343e-05   6.917 9.03e-12 ***
patriots_f1$max_fol     -9.218e-05  8.118e-05  -1.135    0.256
patriots_f1$time_hour   -2.379e+00  9.730e+00  -0.244    0.807
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1971 on 853 degrees of freedom
Multiple R-squared:  0.7129,   Adjusted R-squared:  0.7113
F-statistic: 423.7 on 5 and 853 DF,  p-value: < 2.2e-16
```

#nfl

```
Residuals:
    Min      1Q  Median      3Q     Max
-7060.1   -82.4   -59.6    15.9  8211.3

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        8.643e+01  3.494e+01   2.474 0.013583 *
nfl_f1$tweets_cur  8.765e-01  1.422e-01   6.164 1.12e-09 ***
nfl_f1$tot_ret    -2.338e-01  6.995e-02  -3.343 0.000867 ***
nfl_f1$tot_fol     6.585e-05  2.223e-05   2.962 0.003149 **
nfl_f1$max_fol    -5.240e-05  3.057e-05  -1.714 0.086903 .
nfl_f1$time_hour  -5.059e-01  2.560e+00  -0.198 0.843401
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 501.3 on 799 degrees of freedom
Multiple R-squared:  0.5865,   Adjusted R-squared:  0.5839
F-statistic: 226.7 on 5 and 799 DF,  p-value: < 2.2e-16
```

#gopatriots

```
Residuals:
     Min       1Q   Median       3Q      Max
-2177.23    -3.91    -0.81     1.57  2088.72

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                2.6026147 13.4262928   0.194  0.84636
gopatriots_f1$tweets_cur -0.9082116  0.2419134  -3.754  0.00019 ***
gopatriots_f1$tot_ret     1.8150893  0.2428115   7.475  2.6e-13 ***
gopatriots_f1$tot_fol    -0.0005611  0.0002100  -2.672  0.00772 **
gopatriots_f1$max_fol     0.0002991  0.0002008   1.489  0.13693
gopatriots_f1$time_hour  -0.1987171  0.9965832  -0.199  0.84202
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 172.5 on 628 degrees of freedom
Multiple R-squared:  0.6532,   Adjusted R-squared:  0.6505
F-statistic: 236.6 on 5 and 628 DF,  p-value: < 2.2e-16
```

#gohawks

```
Residuals:
    Min      1Q  Median      3Q     Max
-9808.7   -77.5   -71.2   -42.8 16837.1

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             6.887e+01  4.722e+01   1.458 0.145035
gohawks_f1$tweets_cur   1.296e+00  1.327e-01   9.769  < 2e-16 ***
gohawks_f1$tot_ret     -1.665e-01  4.367e-02  -3.813 0.000146 ***
gohawks_f1$tot_fol     -1.739e-04  6.479e-05  -2.685 0.007378 **
gohawks_f1$max_fol      6.556e-05  1.205e-04   0.544 0.586462
gohawks_f1$time_hour    5.322e-01  3.524e+00   0.151 0.880007
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 755.6 on 965 degrees of freedom
Multiple R-squared:  0.4816,   Adjusted R-squared:  0.4789
F-statistic: 179.3 on 5 and 965 DF,  p-value: < 2.2e-16
```

Hashtag	Multiple R-squared	Significant features (*,**,***)
#Superbowl	0.7833	4
#sb49	0.8038	2
#Patriots	0.7129	3
#nfl	0.5868	3
#gopatriots	0.6532	3
#gohawks	0.4816	3

The significance of the features was determined by looking at the t-value and p-test results from the regression summaries (large t values and low probabilities were deemed significant).

Judging from the R^2 values, for some of the hashtags a linear model may not be the best model for the data. The current amount of tweets was always significant, and the total amount of retweets and followers were almost always significant; the hashtag #sb49 was the most different in that it was the only hashtag in which the two features were not significant. The maximum number of followers in the time interval was only mildly significant in 2 cases, while the time of day was generally not significant. Thus, perhaps not all of the features are linearly related to the amount of tweets in the next hour.

Something that may have affected the results is the fact that during some hours, such as near the beginning of the datasets, there were no tweets and therefore all of the feature values were 0.

Part 3

For this part, some additional features are considered: total retweets (not citations), total times the tweets in the hour were favorited, and total number of replies the tweets in the hour received. These features were extracted tweet['tweet']['favorite_count'], tweet['tweet']['retweet_count'], and tweet['metrics']['citations']['replies']. After some testing, the insignificant features in the models are dropped to complete a final linear model for each hashtag.

The eight features that were tested were (each value for the current hour):

1. Number of tweets
2. Total number of Citations
3. Total number of Followers
4. Maximum number of followers
5. Hour of day
6. Total amount of favorites
7. Total amount of retweets
8. Total amount of replies

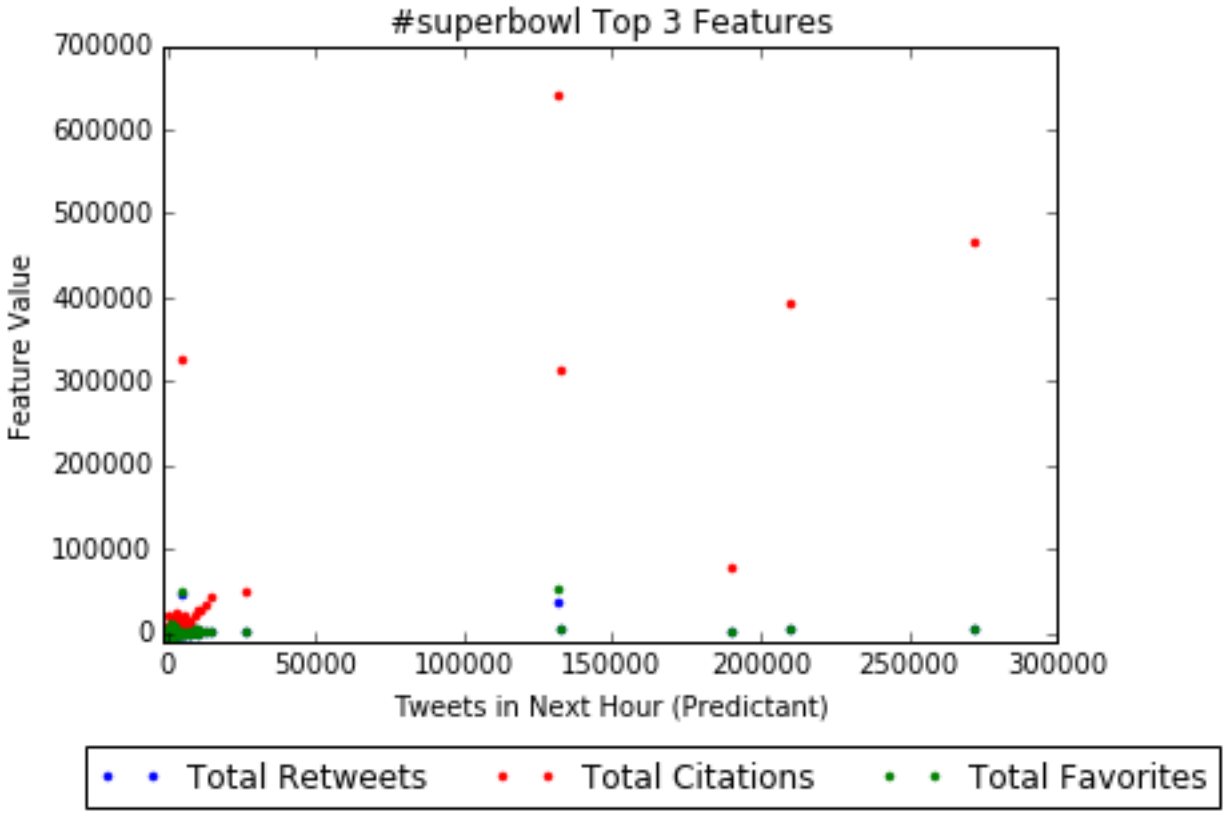
After testing, the following features were kept for the models:

1. Number of tweets
2. Total number of Citations
3. Total number of Followers
4. Total amount of favorites
5. Total amount of retweets

The feature “total amount of replies” was added for this part from part 2, but this value tended to be small often so this is likely why it was not a good feature to use.

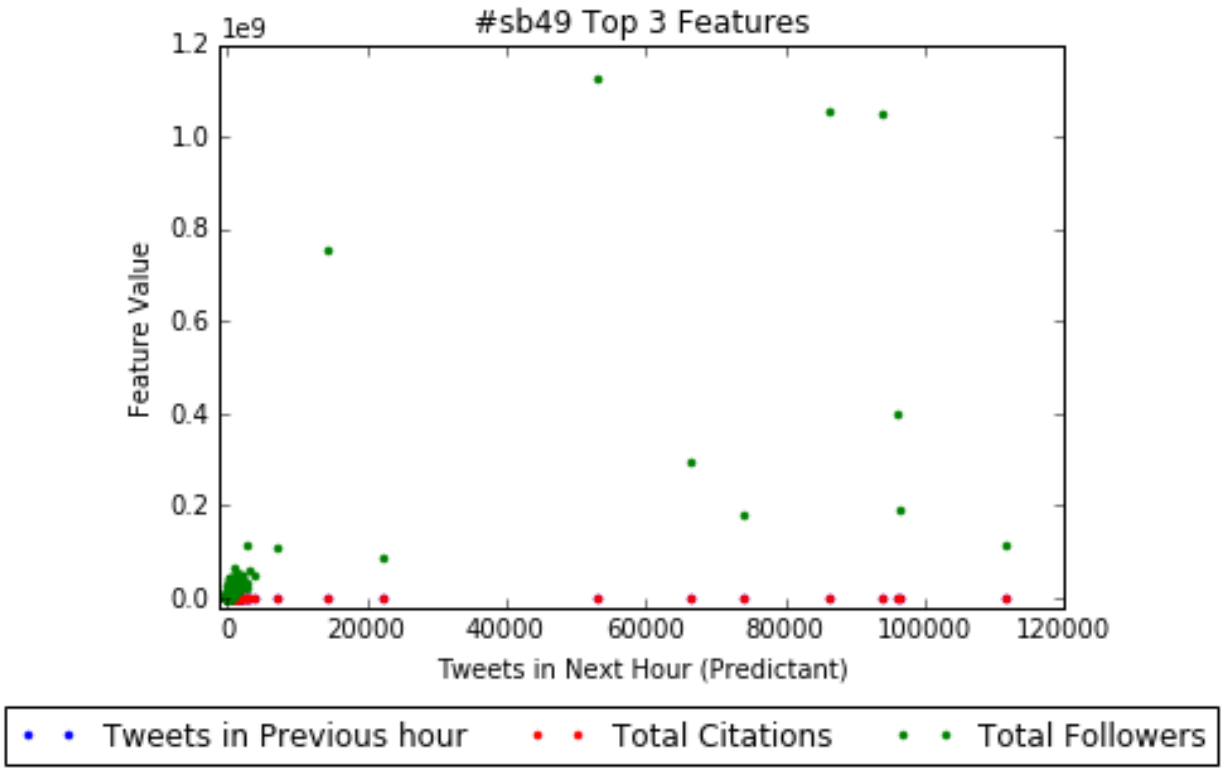
#superbowl

```
Residuals:
     Min       1Q   Median       3Q      Max
-1482.28   -10.85    -6.20    -5.20  1797.15

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              6.203e+00  5.681e+00   1.092  0.27535
superbowl_f2$tweets_cur  5.697e-01  2.149e-01   2.652  0.00821 **
superbowl_f2$tot_ret     9.185e-01  1.717e-01   5.350 1.23e-07 ***
superbowl_f2$tot_fol    -1.963e-04  4.603e-05  -4.264 2.32e-05 ***
superbowl_f2$tot_fav     2.829e+01  4.486e+00   6.307 5.37e-10 ***
superbowl_f2$tot_ret_c  -4.794e+01  4.691e+00 -10.221  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 139.8 on 627 degrees of freedom
Multiple R-squared:  0.7723,   Adjusted R-squared:  0.7705
F-statistic: 425.4 on 5 and 627 DF,  p-value: < 2.2e-16
```

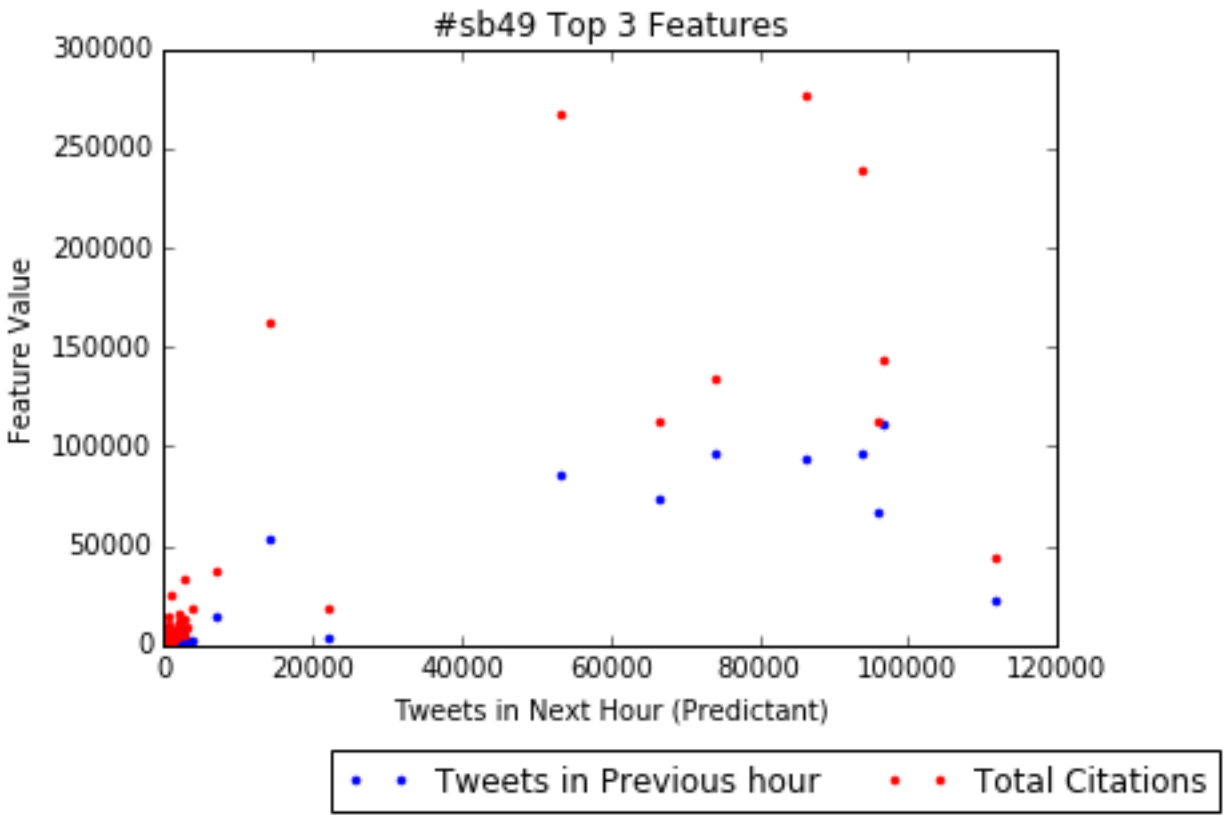


#sb49

```
Residuals:
   Min     1Q Median     3Q    Max
-29655   -226   -222   -193  91416

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.243e+02  1.915e+02   1.171    0.242
sb49_f2$tweets_cur  1.095e+00  1.024e-01  10.698   <2e-16 ***
sb49_f2$tot_ret    -1.133e-01  9.453e-02  -1.199    0.231
sb49_f2$tot_fol     5.990e-06  1.453e-05   0.412    0.680
sb49_f2$tot_fav    -1.001e-01  4.591e-01  -0.218    0.827
sb49_f2$tot_ret_c  -4.202e-02  6.286e-01  -0.067    0.947
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4521 on 575 degrees of freedom
Multiple R-squared:  0.8028,   Adjusted R-squared:  0.8011
F-statistic: 468.3 on 5 and 575 DF,  p-value: < 2.2e-16
```

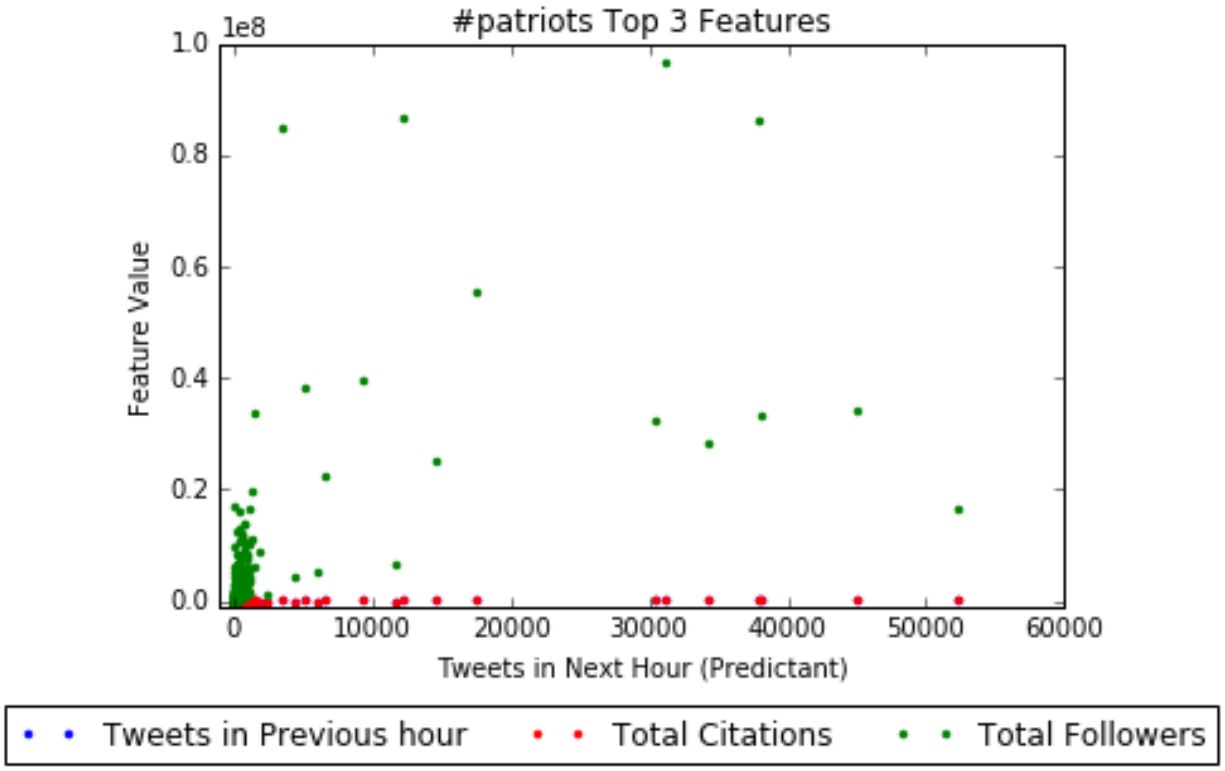


Once again, for #sb49, only number of tweets in previous hour was the only significant feature, with the other features having very small predicted coefficients. The blue points overlapped with the red points in this plot a lot.

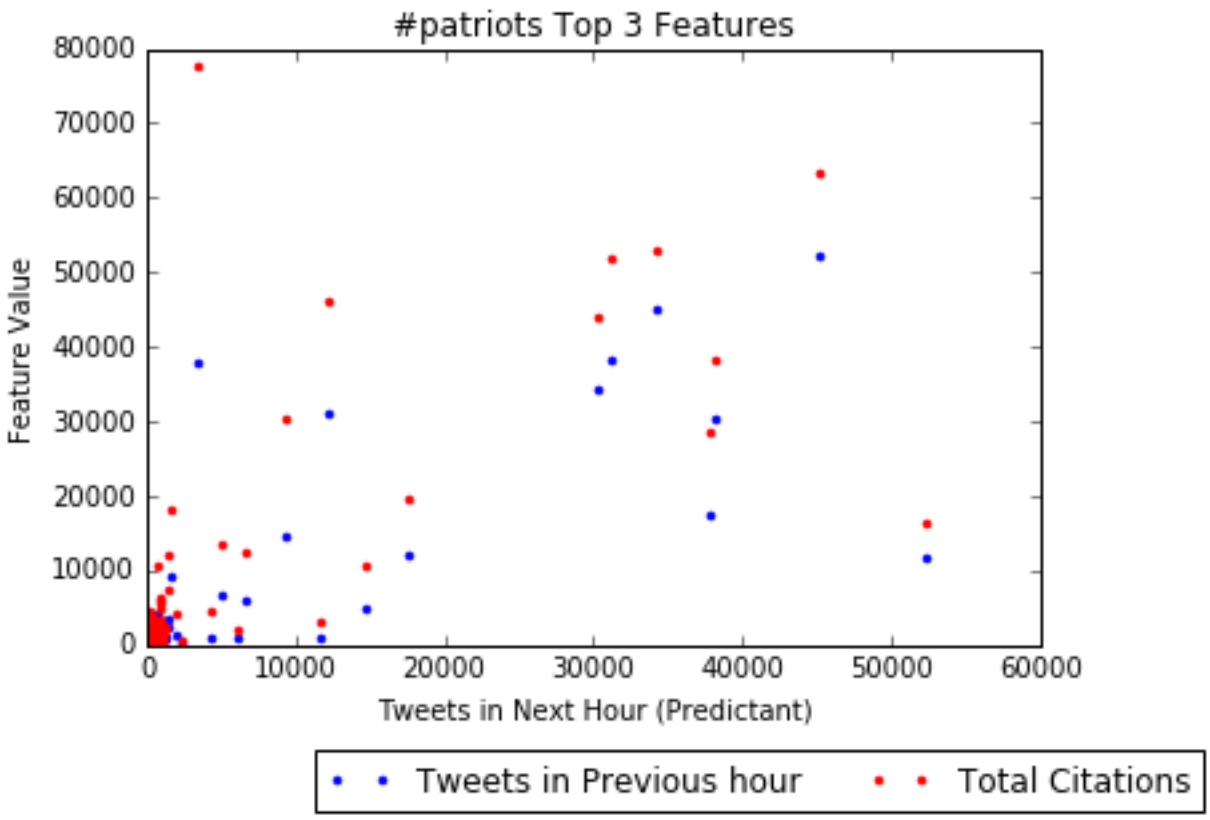

#sb49 Top 3 Features
Feature Value
Tweets in Next Hour (Predictant)
300000
250000
200000
150000
100000
50000
0
0
20000
40000
60000
80000
100000
120000
Tweets in Previous hour
Total Citations

#patriots

```
Residuals:
   Min     1Q Median     3Q    Max
-16792   -151   -140    -98  42884

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              1.398e+02  6.969e+01   2.006   0.0452 *
patriots_f2$tweets_cur   1.685e+00  1.032e-01  16.323  < 2e-16 ***
patriots_f2$tot_ret     -7.686e-01  8.543e-02  -8.997  < 2e-16 ***
patriots_f2$tot_fol      1.384e-04  1.908e-05   7.255 9.05e-13 ***
patriots_f2$tot_fav     -9.454e-01  8.565e-01  -1.104   0.2700
patriots_f2$tot_ret_c    7.975e-01  9.749e-01   0.818   0.4136
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1969 on 853 degrees of freedom
Multiple R-squared:  0.7133,   Adjusted R-squared:  0.7116
F-statistic: 424.4 on 5 and 853 DF,  p-value: < 2.2e-16
```

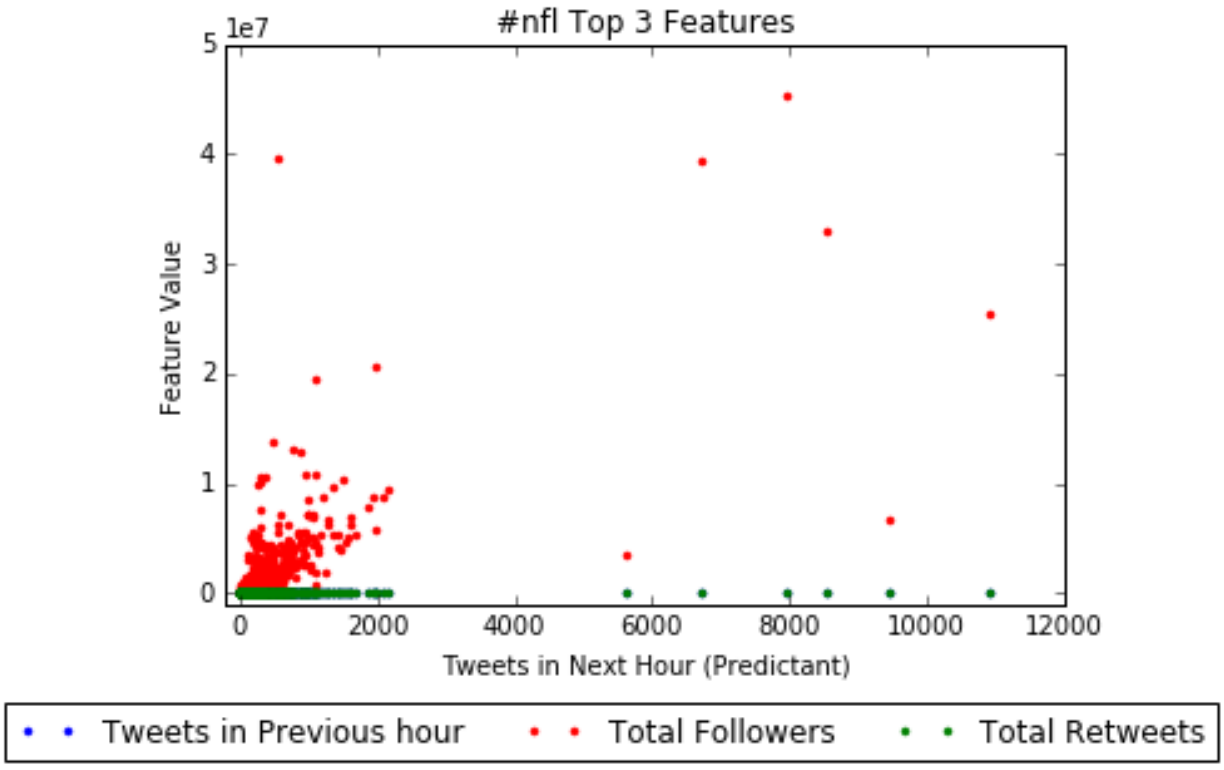


The blue points overlapped with the red points in this plot a lot.

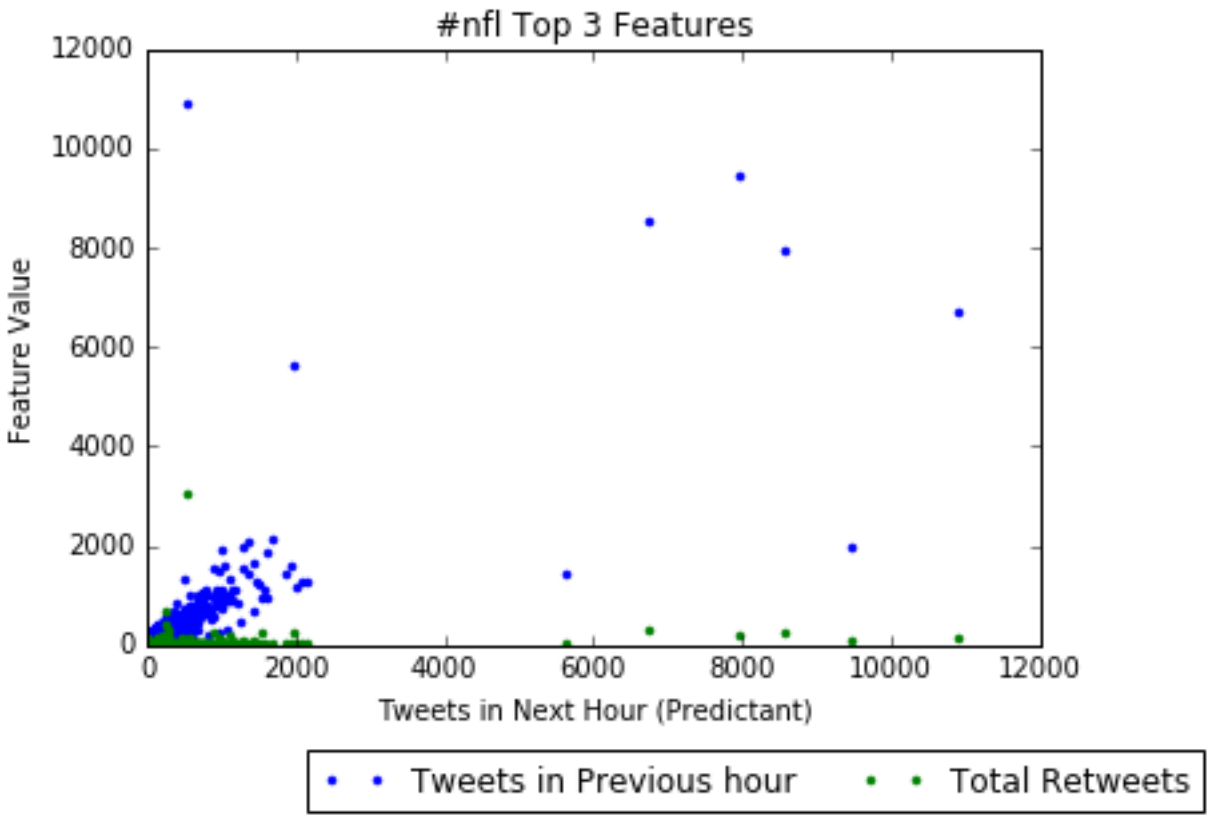

#patriots Top 3 Features
Feature Value
Tweets in Next Hour (Predictant)
0
10000
20000
30000
40000
50000
60000
70000
80000
Tweets in Previous hour
Total Citations

#nfl

```
Residuals:
    Min      1Q  Median      3Q     Max
-3018.0   -62.6   -43.5     3.2  7516.2

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         4.348e+01  1.619e+01   2.686  0.00739 **
nfl_f2$tweets_cur   8.357e-01  1.019e-01   8.197 9.75e-16 ***
nfl_f2$tot_ret      7.429e-02  6.097e-02   1.219  0.22337
nfl_f2$tot_fol      1.329e-05  1.047e-05   1.269  0.20487
nfl_f2$tot_fav     -7.136e-01  7.057e-01  -1.011  0.31227
nfl_f2$tot_ret_c   -2.175e+00  9.036e-01  -2.407  0.01631 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 419.9 on 799 degrees of freedom
Multiple R-squared:  0.7099,   Adjusted R-squared:  0.708
F-statistic:   391 on 5 and 799 DF,  p-value: < 2.2e-16
```

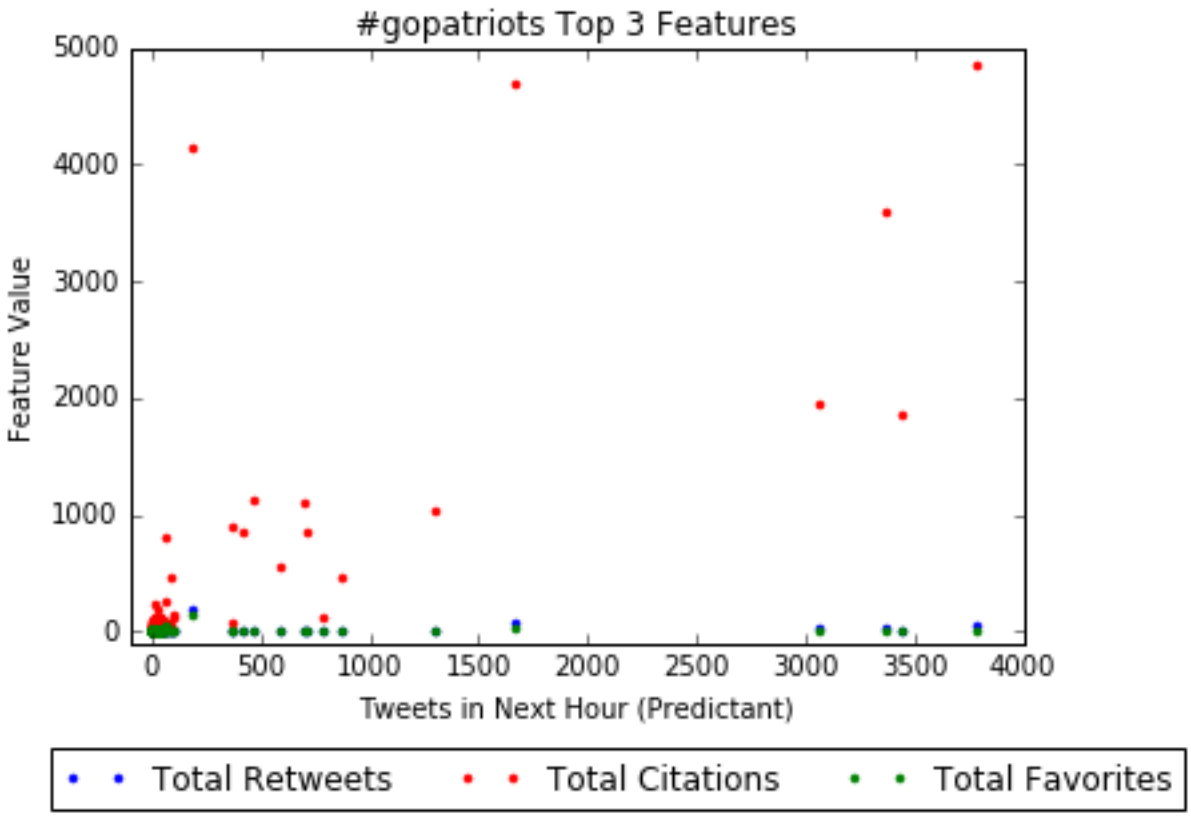


#NFL might have had different behavior since the hashtag would have been relevant during the entire time period due to other NFL games and the NFL playoffs.


#nfl Top 3 Features
Feature Value
Tweets in Next Hour (Predictant)
0
2000
4000
6000
8000
10000
12000
Tweets in Previous hour
Total Retweets

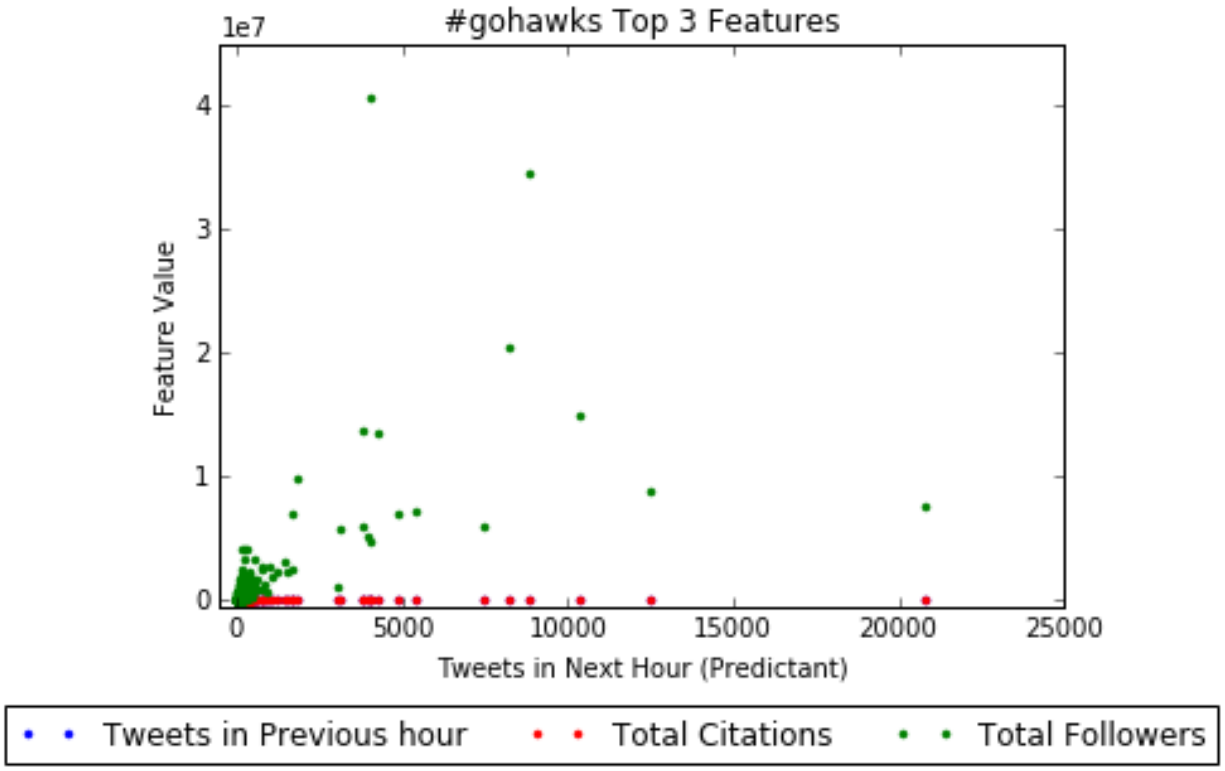
#gopatriots

```
Residuals:
     Min       1Q   Median       3Q      Max
-1482.28   -10.83    -6.19    -5.19  1797.14

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                6.192e+00  5.672e+00   1.092  0.27536
gopatriots_f2$tweets_cur   5.697e-01  2.147e-01   2.653  0.00817 **
gopatriots_f2$tot_ret      9.185e-01  1.715e-01   5.355 1.20e-07 ***
gopatriots_f2$tot_fol     -1.963e-04  4.600e-05  -4.267 2.29e-05 ***
gopatriots_f2$tot_fav      2.829e+01  4.482e+00   6.312 5.20e-10 ***
gopatriots_f2$tot_ret_c   -4.794e+01  4.687e+00 -10.229  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 139.7 on 628 degrees of freedom
Multiple R-squared:  0.7724,   Adjusted R-squared:  0.7705
F-statistic: 426.1 on 5 and 628 DF,  p-value: < 2.2e-16
```

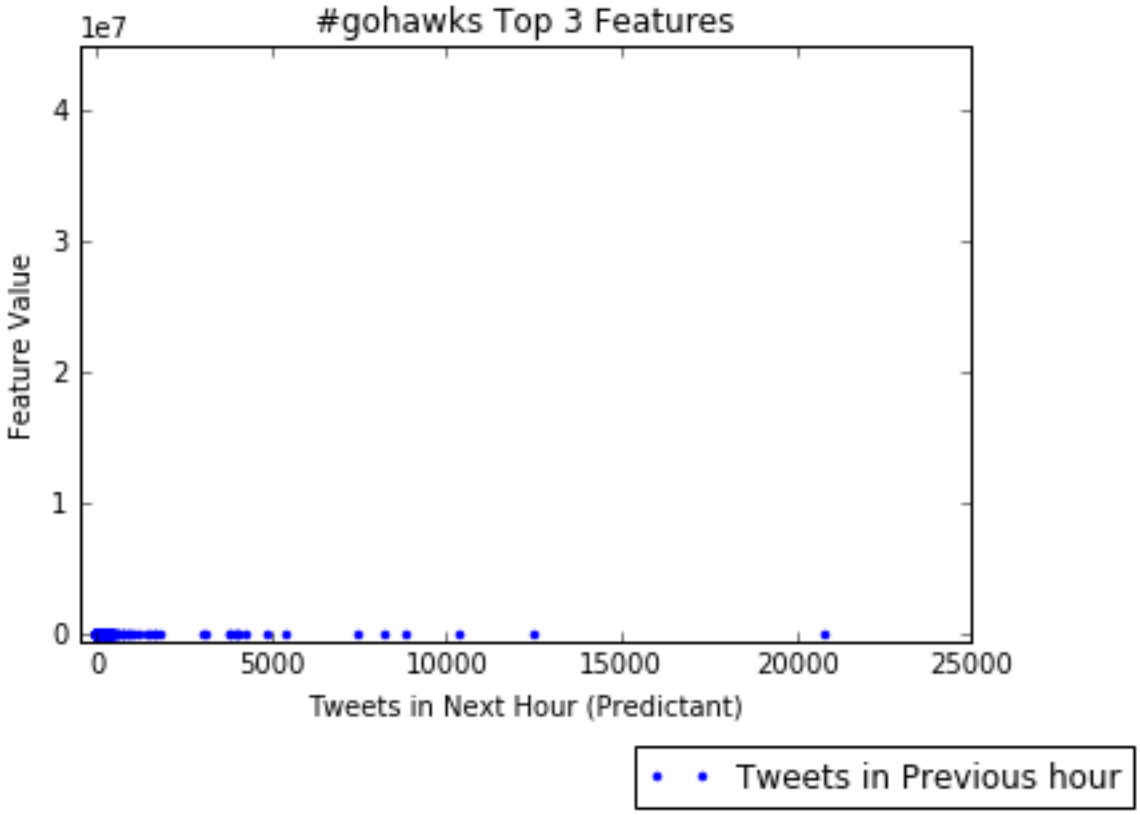


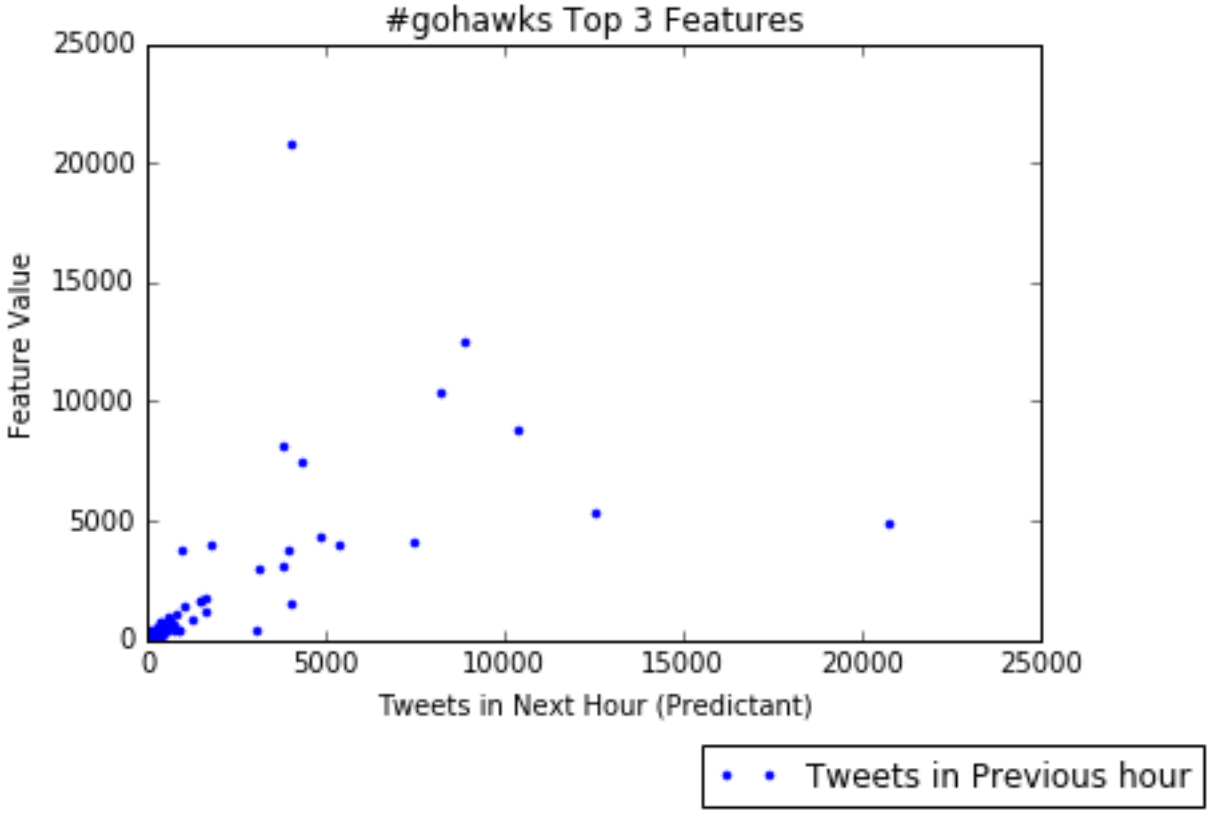
#gohawks

```
Residuals:
    Min      1Q  Median      3Q     Max
-9149.7   -81.7   -80.7   -44.4 16656.8

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)            8.168e+01  2.459e+01   3.322 0.000927 ***
gohawks_f2$tweets_cur  1.564e+00  1.264e-01  12.379  < 2e-16 ***
gohawks_f2$tot_ret    -3.774e-01  6.364e-02  -5.930 4.21e-09 ***
gohawks_f2$tot_fol    -1.262e-04  4.536e-05  -2.783 0.005494 **
gohawks_f2$tot_fav    -8.131e-03  9.076e-02  -0.090 0.928637
gohawks_f2$tot_ret_c   2.649e-01  2.025e-01   1.308 0.191077
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 747.4 on 965 degrees of freedom
Multiple R-squared:  0.4927,   Adjusted R-squared:  0.4901
F-statistic: 187.4 on 5 and 965 DF,  p-value: < 2.2e-16
```



Once again, blue overlapped with red here; sample plots shown below:


#gohawks Top 3 Features
1e7
4
3
2
1
0
Feature Value
0
5000
10000
15000
20000
25000
Tweets in Next Hour (Predictant)
Tweets in Previous hour


#gohawks Top 3 Features
25000
20000
15000
10000
5000
0
Feature Value
0
5000
10000
15000
20000
25000
Tweets in Next Hour (Predictant)
Tweets in Previous hour

Part 4

First, the 10-fold cross validation results are computed across the entire period of time for each hashtag dataset. Then, the datasets are broken up into their respective 3 periods, and the process is completed again. Average absolute errors across each hashtag's 10 tests per period are reported below:

Hashtag	Entire time period	Period 1 (Before)	Period 2 (During)	Period 3 (After)
#Superbowl	2379.06613487	215.014553522	138124.96105	223.553099729
#sb49	995.959091364	37.7935181316	71920.681055	272.220788094
#Patriots	469.681310621	140.628200728	17158.8799081	111.650562321
#nfl	152.939818091	91.088559206	5159.41772408	114.961321746
#gopatriots	39.6164258337	12.7724316486	1664.55465422	3.16106622056
#gohawks	183.679046872	185.969771525	5182.63065425	91.6476900466

For periods 1 and 3 there is noticeable improvement in the absolute error, but for period 2 the error was often very large. This is likely due to the fact that it is only a 12-hour window, so only 11 data points were available to cross-validate, leading to test data sets of 1 hour only. Also, there was a large quantity of tweets posted in the period, so the errors might not be as large as they seem when compared to the volume of tweets during that time period.

Part 5

For this part, first a linear model is computed for all 3 periods for each hashtag. Then, the sample data is processed. To decide which model to use, the text of each tweet is checked, and the amount of tweets containing each hashtag per sample is computed. The text is obtained using tweet['highlight'] since sometimes the other methods of getting the tweet text returned empty strings. The tweet['tweet']['entities']['hashtags'] could also have been checked but checking the text of the tweet was simpler to implement. The hashtag that appeared in most/all of the sample tweets was deemed the hashtag where the sample originated, and the corresponding linear regression model is applied.

Sample	Period	Hashtag	Next Hour Tweets Predicted
1	1	#superbowl	171.49288236916837
2	2	#superbowl	54178.835369524386
3	3	#superbowl	609.71370662760694
4	1	#superbowl	211.92879312142259
5	1	#superbowl	176.91609594649765
6	2	#sb49	63955.510951897464
7	3	#nfl	141.24913418311638
8	1	#nfl	55.984157836553393
9	2	#superbowl	8852.8724399372804
10	3	#nfl	95.771850019671859

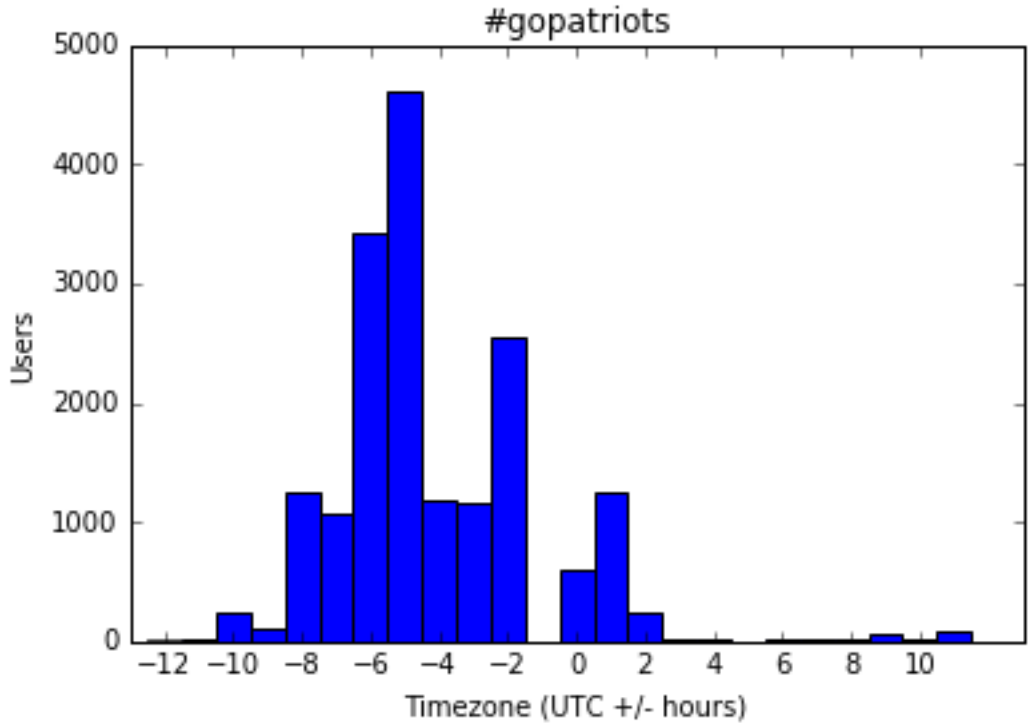
For the first 3 samples, #superbowl appeared in all tweets so that hashtag's models were selected. For samples 4 and 5, there was a mixture of hashtags and #superbowl appeared the most frequently so it was selected again. Sample 6 had #sb49 most frequent, samples 7, 8, and 10 had #nfl most frequent, and sample 9 had #superbowl most frequent once again.

Part 6

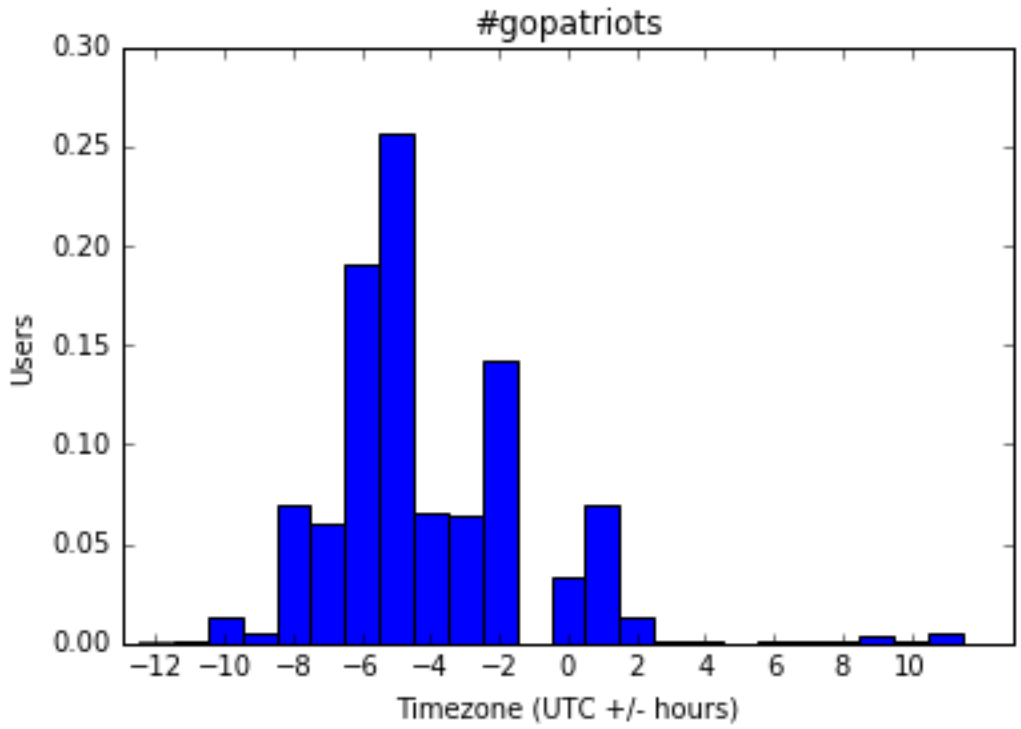
In this free choice part, I have decided to focus on examining relationships between various Twitter user variables located in tweet['tweet']['user'].

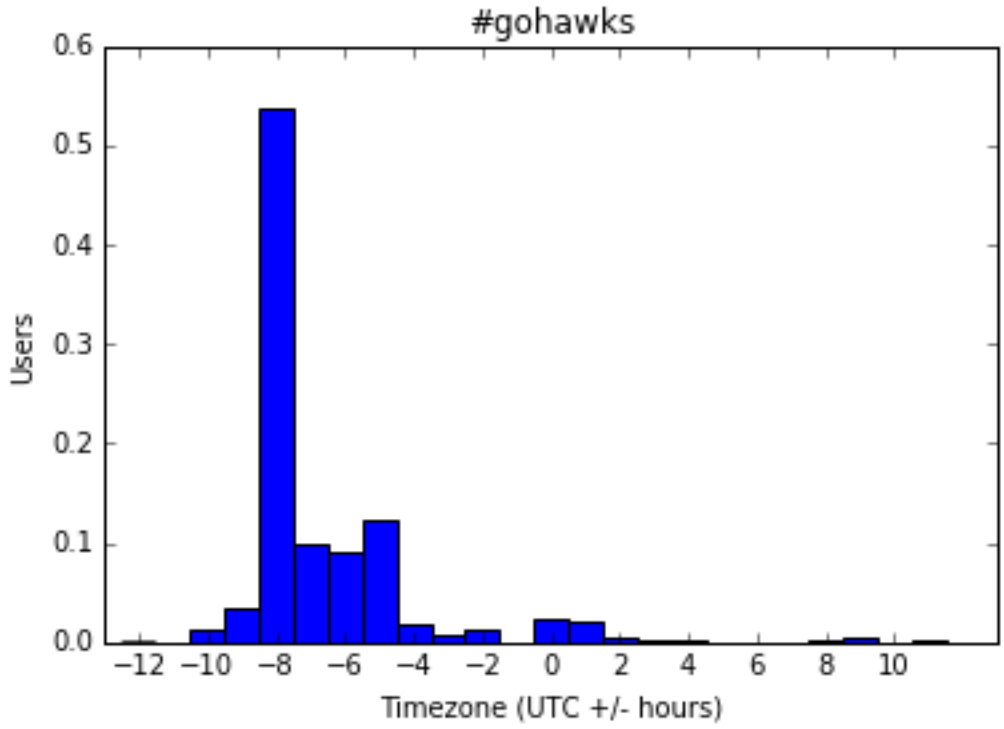
First, I decided to estimate/approximate the distribution of Patriots/Seahawks fans by extracting the timezones that each user who tweeted #gopatriots or #gohawks did. This is accomplished by accessing the tweet['tweet']['user']['utc_offset'] parameters; some are "None" and are ignored. The data is also in seconds, so to convert to hours each value was divided by 3600.

The following is a histogram of the #gopatriots users' timezone distribution in terms of UTC +/- hours:



The plot would be more useful normalized, so we do that:





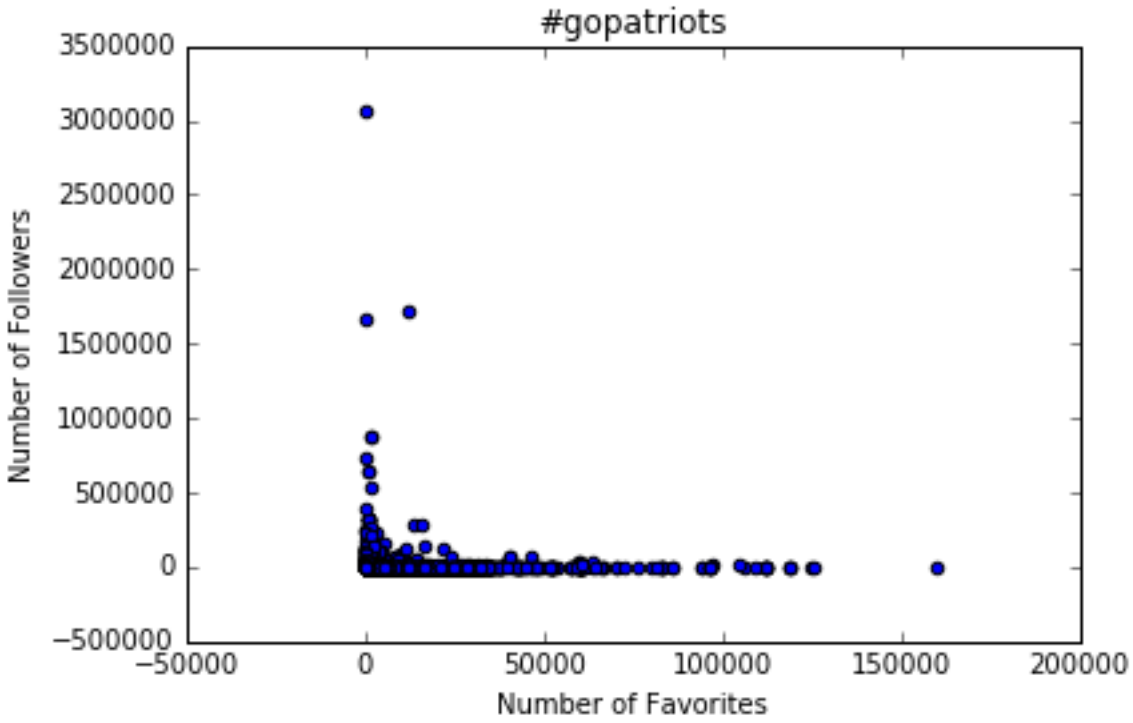
As expected, many Seahawks tweeters were on the west coast, while most Patriots tweeters were on the east coast. Oddly, the #gopatriots data had many users in UTC -2 hours, which appears mostly covers the ocean; perhaps a single or group of users in this timezone contributed most of the tweets. Other than North American timezones, there were also small clusters that represented Europe and Australia/New Zealand.

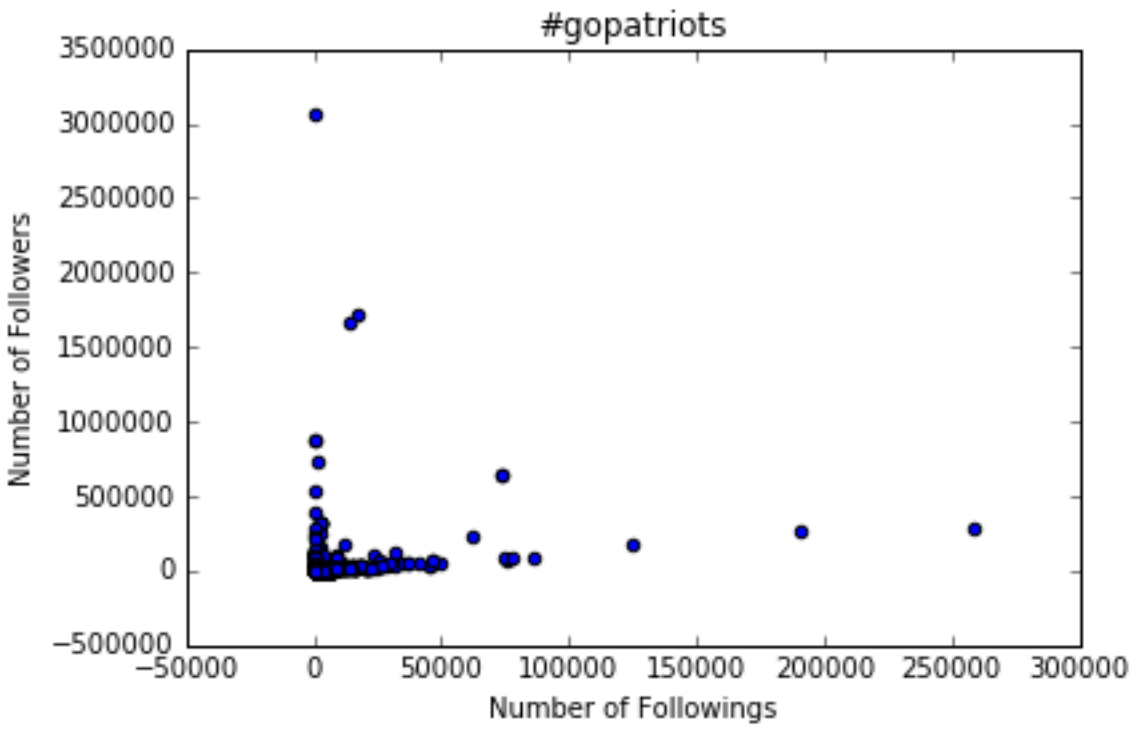
Now, we will try to determine if there are correlations between different variables. I have decided to try and find out if how active a Twitter user is correlates with how many followers that account has. To represent how active a user is on Twitter, the following variables are extracted:

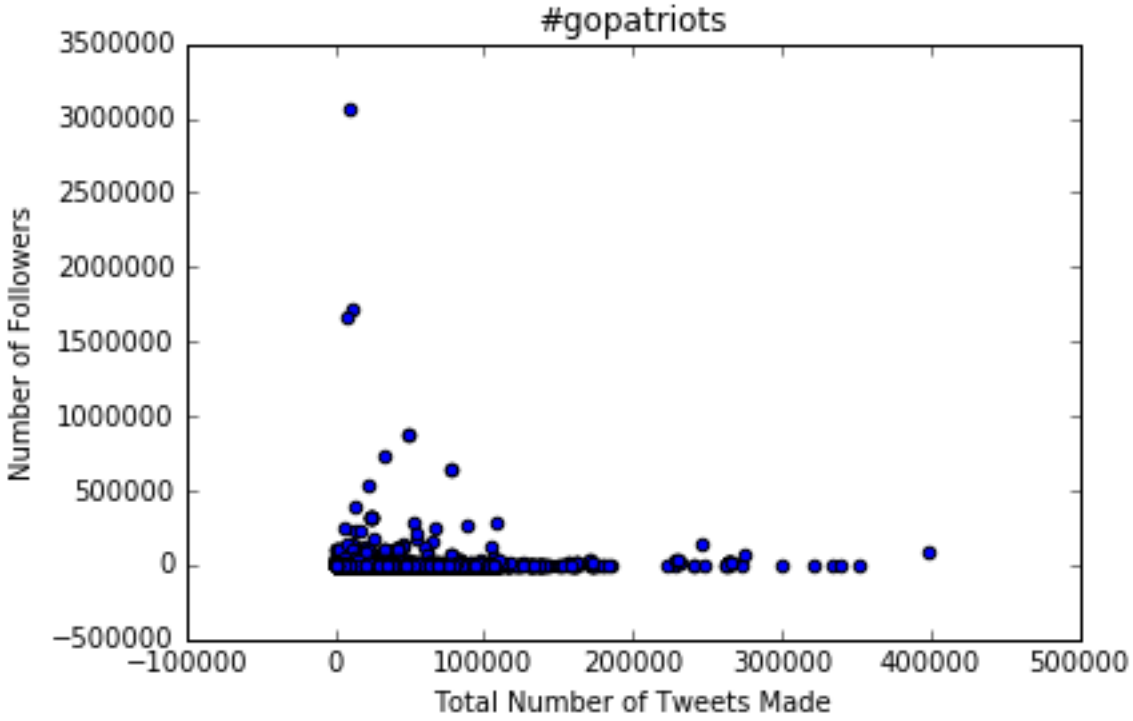
Number of tweets user has saved as favorites	tweet['tweet']['user']['favourites_count']
Number of others the user follows	tweet['tweet']['user']['friends_count']
Total number of Tweets user has made	tweet['tweet']['user']['statuses_count']
Number of Followers	tweet['tweet']['user']['followers_count']

Since the #gopatriots data is the smallest, we will initially work with that data to get a basic idea of the correlations.

First, each variable was plotted against number of followers:







From examining the graphs, the number of followings seems to have a slightly positive relationship with number of followers.

A multiple linear regression is performed in R; the results are as follows:

```
Residuals:
    Min      1Q  Median      3Q     Max
-226743   -1174    -548    -236 3066428

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                49.904113 194.861011   0.256    0.798
gopatriots6$fav_count      -0.041236   0.029031  -1.420    0.156
gopatriots6$n_statuses      0.040481   0.009109   4.444 8.85e-06 ***
gopatriots6$n_followings    1.930518   0.061296  31.495  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27470 on 26228 degrees of freedom
Multiple R-squared:  0.03984,  Adjusted R-squared:  0.03973
F-statistic: 362.8 on 3 and 26228 DF,  p-value: < 2.2e-16
```

It appears that number of followings has a clear, significant positive relationship with number of followers, while number of statuses has a slightly positive significant relationship. Number of favorites made was not significant. Although there is evidence of positive correlations, the R^2 value is low, so a linear model is likely inappropriate for predicting the actual number of followers.

A new question is, would performing the regressions on 1 pair of variables at a time affect the significance of the relationships? The results are shown below:

```
Residuals:
    Min      1Q  Median      3Q     Max
 -11535   -1431   -1331   -1099 3066815

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.463e+03  1.821e+02   8.034 9.83e-16 ***
gopatriots6$fav_count 7.061e-02  2.869e-02   2.461   0.0139 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28030 on 26230 degrees of freedom
Multiple R-squared:  0.0002309,       Adjusted R-squared:  0.0001928
F-statistic: 6.057 on 1 and 26230 DF,  p-value: 0.01385

Residuals:
    Min      1Q  Median      3Q     Max
-233705   -1099    -668    -458 3066501

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              343.44896  173.97403   1.974   0.0484 *
gopatriots6$n_followings   1.97101    0.06032  32.674   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27480 on 26230 degrees of freedom
Multiple R-squared:  0.03911,  Adjusted R-squared:  0.03907
```

```
F-statistic:  1068 on 1 and 26230 DF,  p-value: < 2.2e-16

Residuals:
    Min      1Q  Median      3Q     Max
 -29050   -1122    -775    -695 3066675

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           7.545e+02  1.939e+02   3.892 9.98e-05 ***
gopatriots6$n_statuses 8.561e-02  8.887e-03   9.634  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27980 on 26230 degrees of freedom
Multiple R-squared:  0.003526, Adjusted R-squared:  0.003488
F-statistic: 92.81 on 1 and 26230 DF,  p-value: < 2.2e-16
```

When completed separately, all of the variables became somewhat significant. However, the R^2 values were still low, meaning a linear model would not be appropriate in actually predicting number of followers from the 3 variables. Interestingly, number of favorites now has a positive coefficient when regression is performed on it alone. Even then, the slope of it and number of statuses remained small numbers. Only number of followings had a clear positive relationship to number of followers, so it is likely that as the number of followings an account has made goes up, the more followers it will likely have. Perhaps this is affected by “follow for follow” type interactions between users.

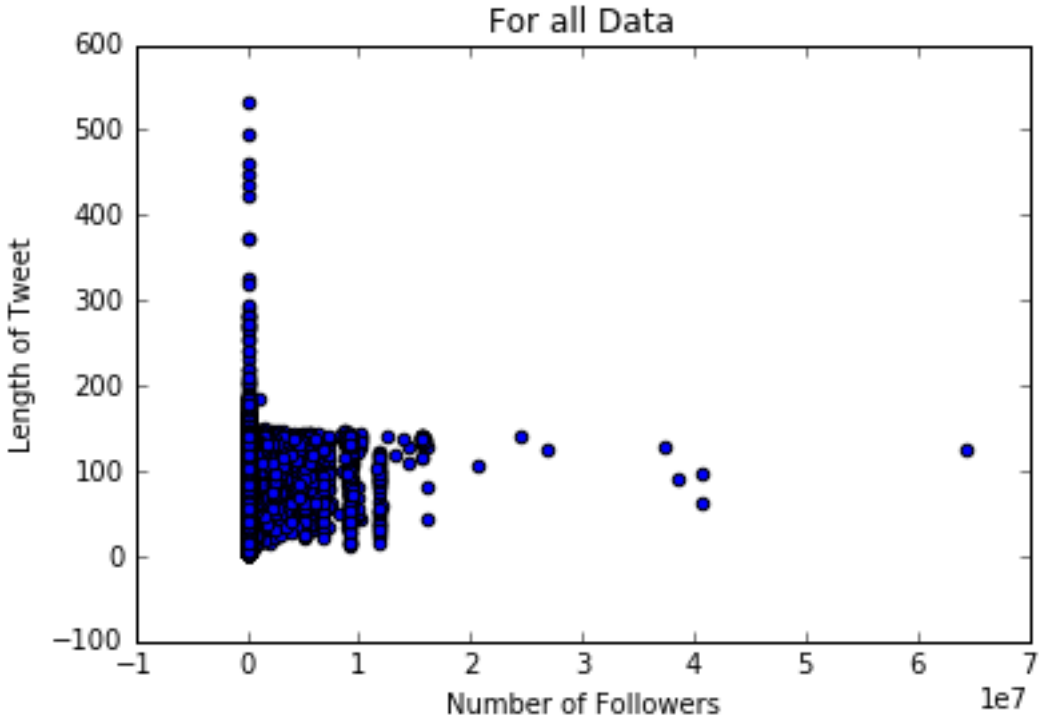
The dataset that is analyzed is then expanded to include all of the users from across all of the hashtag datasets. The results are shown below:

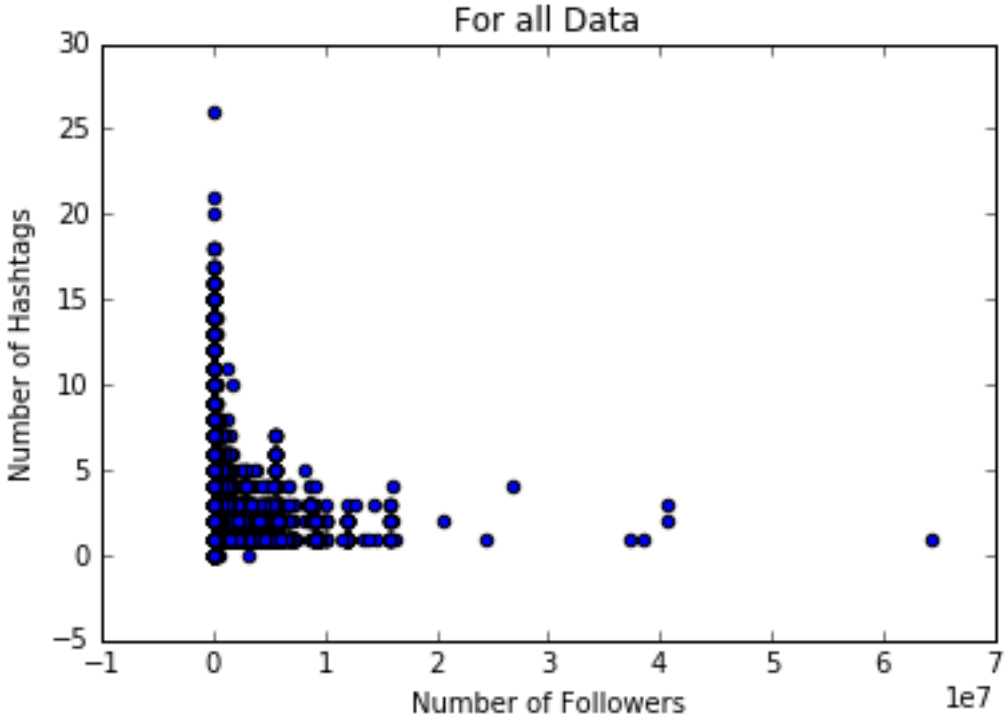
```
Residuals:
     Min       1Q   Median       3Q      Max
-1243685    -5983    -3436    -2454 64315778

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        2003.45146  101.48672  19.741  < 2e-16 ***
data6$fav_count      -0.10201    0.01456  -7.006 2.46e-12 ***
data6$n_statuses      0.26630    0.00239 111.448  < 2e-16 ***
data6$n_followings    2.48509    0.01475 168.528  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 162900 on 3138819 degrees of freedom
Multiple R-squared:  0.01472,  Adjusted R-squared:  0.01472
F-statistic: 1.563e+04 on 3 and 3138819 DF,  p-value: < 2.2e-16
```

```
Residuals:
     Min       1Q   Median       3Q      Max
 -140333    -7646    -7543    -7144 64311764

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     7.615e+03  9.705e+01   78.46   <2e-16 ***
data6$fav_count 2.399e-01  1.455e-02   16.49   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 164100 on 3138821 degrees of freedom
Multiple R-squared:  8.659e-05,       Adjusted R-squared:  8.627e-05
F-statistic: 271.8 on 1 and 3138821 DF,  p-value: < 2.2e-16


Residuals:
     Min       1Q   Median       3Q      Max
-1397046    -6667    -5913    -5596 64313865

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        5.420e+03  9.325e+01   58.12   <2e-16 ***
data6$n_followings 2.707e+00  1.462e-02  185.20   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 163200 on 3138821 degrees of freedom
Multiple R-squared:  0.01081,  Adjusted R-squared:  0.01081
F-statistic: 3.43e+04 on 1 and 3138821 DF,  p-value: < 2.2e-16


Residuals:
     Min       1Q   Median       3Q      Max
 -629777    -6010    -3981    -3496 64314217

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.483e+03  9.844e+01   35.39   <2e-16 ***
data6$n_statuses 3.197e-01  2.362e-03  135.35   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 163600 on 3138821 degrees of freedom
Multiple R-squared:  0.005802, Adjusted R-squared:  0.005802
F-statistic: 1.832e+04 on 1 and 3138821 DF,  p-value: < 2.2e-16
```

With the expanded dataset, it is apparent that there are positive correlations between each variable and number of followers. However, the consistently low R^2 values cement the idea that linear models would not be appropriate in predicting number of followers, and should only be used to check for positive/negative relationships between the features.

The last thing that I decided to test was the relationship between number of followers and the length of the tweet and number of hashtags used. Would being more popular lead to a larger or smaller number of hashtags used? The following are scatterplots of number of followers vs. length of tweet and number of hashtags in the tweet for all hashtags:





It is clear that there are more samples from users with few followers, so we perform the same regression to check for +/- coefficients:

```
Residuals:
    Min      1Q  Median      3Q     Max
-268.07  -20.68   -8.32   27.68  445.68

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       8.632e+01  1.834e-02 4705.62   <2e-16 ***
data6$n_followers 4.785e-06  1.116e-07   42.86   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32.46 on 3138821 degrees of freedom
Multiple R-squared:  0.0005849,       Adjusted R-squared:  0.0005846
F-statistic:  1837 on 1 and 3138821 DF,  p-value: < 2.2e-16


Residuals:
    Min      1Q  Median      3Q     Max
-2.2689 -1.2687 -0.2689  0.7311 23.7311

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.269e+00  8.193e-04 2769.37   <2e-16 ***
data6$n_followers -1.256e-07  4.986e-09  -25.19   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.45 on 3138821 degrees of freedom
Multiple R-squared:  0.0002021,       Adjusted R-squared:  0.0002017
F-statistic: 634.4 on 1 and 3138821 DF,  p-value: < 2.2e-16
```

Once again, the R^2 values are bad, but it is apparent that the coefficients are significant. Even though they are significant, the coefficients are incredibly small, so it is likely that there is no correlation between how popular a user is and how they structure their tweets.