Project 5 Report

Mahmoud Essalat, Sherry He, Weitang Sun, Siqi Huang
ID: 005034839, 805040110, 904946260, 504490530

Part 1.

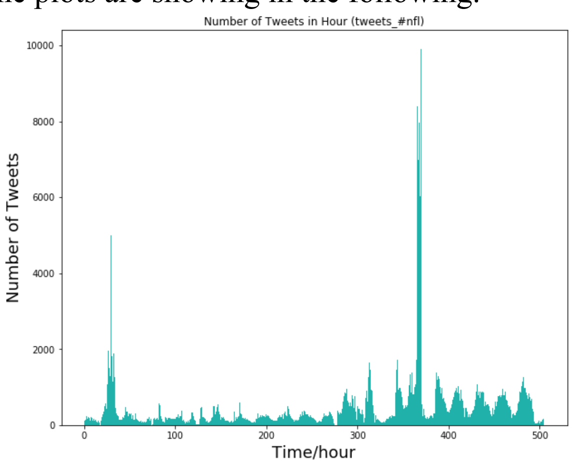
Question 1.

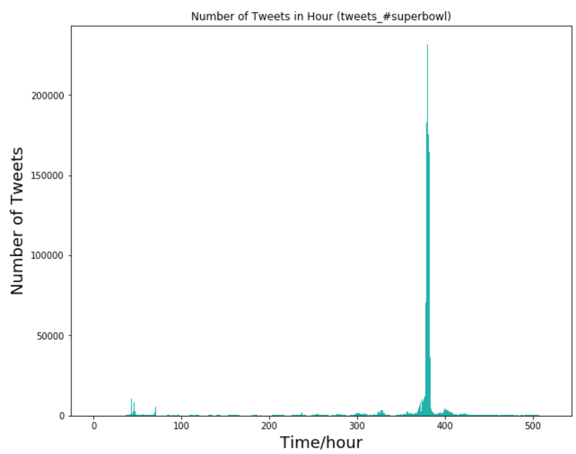
The statistics are showing in the following table.

	Average number of tweets per hour	Average number of followers	Average number of retweets
#gohawks	292.49	2217.92	2.01
#gopatriots	40.95	1427.25	1.41
#nfl	397.02	4662.38	1.53
#patriots	750.89	3280.46	1.79
#sb49	1276.86	10374.16	2.53
#superbowl	2072.12	8814.97	2.39

Question 2.

The plots are showing in the following.





Question 3.

There are five features used in this regression exercise: x1 - number of tweets, x2 – total number of retweets, x3 – sum of the number of followers of the users posting the hashtag, x4 – maximum number of followers of the users posting the hashtag, and x5 – time of the day. We also include the constant in the regression.

#GoHawks

RMSE of #gohawks for the linear regression model is: 870.95 and R-squared is 0.476. For *p*-value, we can see that the parameters of constant, x2, x3 and x4 are significant. *t*-test suggests that the features explains the dependent variable.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.476
Model:                            OLS   Adj. R-squared:                  0.472
Method:                 Least Squares   F-statistic:                     104.1
Date:                Sun, 17 Mar 2019   Prob (F-statistic):           5.01e-78
Time:                        17:46:17   Log-Likelihood:                -4733.0
No. Observations:                 578   AIC:                             9478.
Df Residuals:                     572   BIC:                             9504.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         95.0899     70.543      1.348      0.178     -43.465     233.645
x1             1.6195      5.325      0.304      0.761      -8.839      12.078
x2             1.2827      0.164      7.831      0.000       0.961       1.604
x3            -0.1364      0.043     -3.138      0.002      -0.222      -0.051
x4            -0.0002      8e-05     -2.429      0.015      -0.000   -3.72e-05
x5          6.154e-05      0.000      0.413      0.680      -0.000       0.000
==============================================================================
Omnibus:                      916.585   Durbin-Watson:                   2.216
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           783084.769
Skew:                           8.690   Prob(JB):                         0.00
Kurtosis:                     182.481   Cond. No.                     5.13e+06
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.13e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

#GoPatriots

RMSE of gopatriots for the linear regression model is: 166.09 and R-square is 0.629. For *p*-value, we can see that only the parameter of x3 is significant. *t*-test suggests that the features explains the dependent variable.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.629
Model:                            OLS   Adj. R-squared:                  0.626
Method:                 Least Squares   F-statistic:                     192.9
Date:                Sun, 17 Mar 2019   Prob (F-statistic):          6.97e-120
Time:                        17:46:18   Log-Likelihood:                -3749.1
No. Observations:                 574   AIC:                             7510.
Df Residuals:                     568   BIC:                             7536.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          9.2632     13.562      0.683      0.495     -17.375      35.901
x1            -0.1991      1.017     -0.196      0.845      -2.197       1.799
x2             0.3054      0.285      1.073      0.284      -0.254       0.865
x3             0.4947      0.191      2.590      0.010       0.120       0.870
x4            -0.0001      0.000     -0.525      0.599      -0.001       0.000
x5         -1.937e-05      0.000     -0.089      0.929      -0.000       0.000
==============================================================================
Omnibus:                      486.048   Durbin-Watson:                   1.909
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           291015.311
Skew:                           2.526   Prob(JB):                         0.00
Kurtosis:                     113.192   Cond. No.                     7.48e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.48e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

#nfl

RMSE of #nfl for the linear regression model is: 519.58 and R-square is 0.571. For *p*-value, we can see that only the parameter of x2 is not significant. The parameters of other features are significant. *t*-test suggests that the features explains the dependent variable.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.571
Model:                            OLS   Adj. R-squared:                  0.567
Method:                 Least Squares   F-statistic:                     154.3
Date:                Sun, 17 Mar 2019   Prob (F-statistic):          4.56e-104
Time:                        17:46:21   Log-Likelihood:                -4495.8
No. Observations:                 586   AIC:                             9004.
Df Residuals:                     580   BIC:                             9030.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        123.9278     42.889      2.889      0.004      39.691     208.165
x1             0.4058      3.155      0.129      0.898      -5.791       6.603
x2             0.5671      0.135      4.203      0.000       0.302       0.832
x3            -0.1653      0.064     -2.592      0.010      -0.291      -0.040
x4             0.0001     2.5e-05      4.578      0.000    6.53e-05       0.000
x5            -0.0001    3.31e-05     -3.524      0.000      -0.000   -5.16e-05
==============================================================================
Omnibus:                      670.042   Durbin-Watson:                   2.373
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           350954.169
Skew:                           4.595   Prob(JB):                         0.00
Kurtosis:                     122.537   Cond. No.                     8.60e+06
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.6e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

#Patriots

RMSE of #patriots for the linear regression model is: 2276.16 and R-square is 0.668. For *p*-value, we can see that the parameters of constant, x2, x3 and x5 are significant. *t*-test suggests that the features explains the dependent variable.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.668
Model:                            OLS   Adj. R-squared:                  0.666
Method:                 Least Squares   F-statistic:                     233.8
Date:                Sun, 17 Mar 2019   Prob (F-statistic):          1.91e-136
Time:                        17:46:26   Log-Likelihood:                -5361.4
No. Observations:                 586   AIC:                         1.073e+04
Df Residuals:                     580   BIC:                         1.076e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        180.1751    183.925      0.980      0.328    -181.066     541.416
x1            -5.8597     13.765     -0.426      0.670     -32.896      21.176
x2             0.9145      0.071     12.937      0.000       0.776       1.053
x3            -0.0681      0.058     -1.178      0.239      -0.181       0.045
x4         -1.098e-05   2.63e-05     -0.417      0.677   -6.27e-05    4.07e-05
x5             0.0001   9.17e-05      1.340      0.181   -5.72e-05       0.000
==============================================================================
Omnibus:                      887.682   Durbin-Watson:                   1.998
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           690539.222
Skew:                           7.937   Prob(JB):                         0.00
Kurtosis:                     170.420   Cond. No.                     1.60e+07
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.6e+07. This might indicate that there are
strong multicollinearity or other numerical problems.
```

#sb49

RMSE of #sb49 for the linear regression model is: 4023.48 and R-square is 0.805. For *p*-value, we can see that the parameters of x2, x3, x4 and x5 are significant. *t*-test suggests that the features explains the dependent variable.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.805
Model:                            OLS   Adj. R-squared:                  0.803
Method:                 Least Squares   F-statistic:                     474.3
Date:                Sun, 17 Mar 2019   Prob (F-statistic):         1.66e-201
Time:                        17:46:37   Log-Likelihood:                -5656.4
No. Observations:                 582   AIC:                         1.132e+04
Df Residuals:                     576   BIC:                         1.135e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        206.5402    328.558      0.629      0.530    -438.777     851.857
x1           -15.9597     24.434     -0.653      0.514     -63.950      32.031
x2             1.1363      0.087     13.020      0.000       0.965       1.308
x3            -0.1605      0.079     -2.039      0.042      -0.315      -0.006
x4          9.719e-06   1.25e-05      0.777      0.438   -1.49e-05    3.43e-05
x5          9.449e-05   4.36e-05      2.166      0.031    8.79e-06       0.000
==============================================================================
Omnibus:                     1179.269   Durbin-Watson:                   1.674
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          2205207.514
Skew:                          14.582   Prob(JB):                         0.00
Kurtosis:                     303.143   Cond. No.                     1.57e+08
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.57e+08. This might indicate that there are
strong multicollinearity or other numerical problems.
```

#Superbowl

RMSE of #superbowl for the linear regression model is: 7244.55 and R-square is 0.800. For *p*-value, we can see that the parameters of x2, x3, x4 and x5 are significant. *t*-test suggests that the features explains the dependent variable.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.800
Model:                            OLS   Adj. R-squared:                  0.798
Method:                 Least Squares   F-statistic:                     463.5
Date:                Sun, 17 Mar 2019   Prob (F-statistic):         6.72e-200
Time:                        17:46:51   Log-Likelihood:                -6039.9
No. Observations:                 586   AIC:                         1.209e+04
Df Residuals:                     580   BIC:                         1.212e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -149.5572    605.382     -0.247      0.805   -1338.565    1039.451
x1           -20.4965     43.624     -0.470      0.639    -106.177      65.184
x2             2.2766      0.080     28.537      0.000       2.120       2.433
x3            -0.2543      0.046     -5.544      0.000      -0.344      -0.164
x4            -0.0001    2.2e-05     -6.265      0.000      -0.000   -9.47e-05
x5             0.0007      0.000      4.889      0.000       0.000       0.001
==============================================================================
Omnibus:                      973.862   Durbin-Watson:                   2.283
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1787388.254
Skew:                           9.272   Prob(JB):                         0.00
Kurtosis:                     272.925   Cond. No.                     2.21e+08
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.21e+08. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Question 4-5.

We used the following features: x1 - number of tweets, x2 – time of the day, x3 – sum of the number of followers of the users posting the hashtag, x4 – total number of retweets, x5 - number of URLs in tweets, x6 - number of mentioned users of tweets, x7 - number of favorites in tweets, x8 - the ranking score of tweets and x9 - number of hashtags in tweets. We also include the constant in the regression.

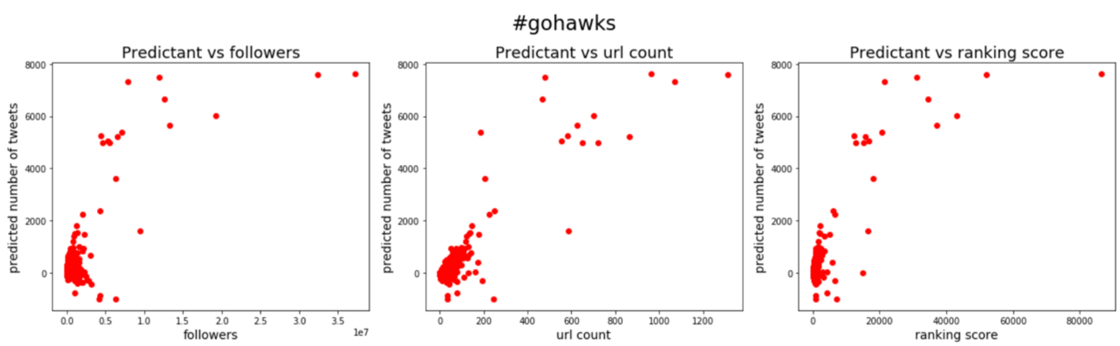
#GoHawks

RMSE of #gohawks for the linear regression model is: 727.71.

```
RMSE of gohawks for the linear regression model is: 727.711953678
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.635
Model:                            OLS   Adj. R-squared:                  0.629
Method:                 Least Squares   F-statistic:                     109.6
Date:                Mon, 18 Mar 2019   Prob (F-statistic):          5.02e-118
Time:                        18:43:48   Log-Likelihood:                -4629.1
No. Observations:                 578   AIC:                             9278.
Df Residuals:                     568   BIC:                             9322.
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -57.4453     60.321     -0.952      0.341    -175.926      61.035
x1             1.5745      4.506      0.349      0.727      -7.276      10.425
x2           -40.6142      4.176     -9.726      0.000     -48.816     -32.412
x3            -0.0553      0.054     -1.017      0.310      -0.162       0.052
x4            -0.0003   4.65e-05     -6.924      0.000      -0.000      -0.000
x5             8.0273      1.499      5.354      0.000       5.082      10.972
x6             1.6909      0.492      3.435      0.001       0.724       2.658
x7             0.0519      0.023      2.229      0.026       0.006       0.098
x8             8.4778      0.870      9.747      0.000       6.769      10.186
x9             0.7790      0.338      2.304      0.022       0.115       1.443
==============================================================================
Omnibus:                      991.584   Durbin-Watson:                   2.213
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           905698.058
Skew:                          10.284   Prob(JB):                         0.00
Kurtosis:                     195.831   Cond. No.                     5.18e+06
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.18e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The scatter plots of top 3 features are shown in the following.



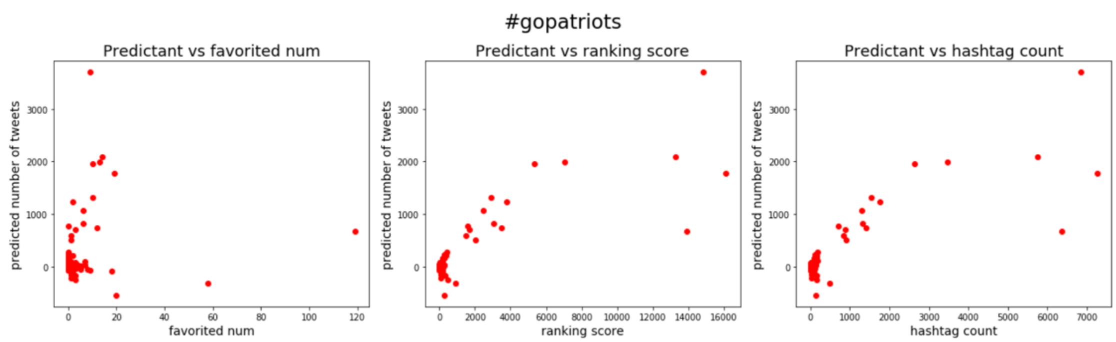
#Gopatriots

RMSE of gopatriots for the linear regression model is: 100.59.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.864
Model:                            OLS   Adj. R-squared:                  0.862
Method:                 Least Squares   F-statistic:                     398.3
Date:                Mon, 18 Mar 2019   Prob (F-statistic):          7.90e-238
Time:                        20:11:25   Log-Likelihood:                -3461.2
No. Observations:                 574   AIC:                             6942.
Df Residuals:                     564   BIC:                             6986.
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -6.9374      8.265     -0.839      0.402     -23.171       9.297
x1             0.3442      0.623      0.552      0.581      -0.880       1.568
x2           -21.1042      1.762    -11.980      0.000     -24.564     -17.644
x3            -1.3881      0.150     -9.278      0.000      -1.682      -1.094
x4          1.168e-05   3.25e-05      0.360      0.719   -5.21e-05    7.55e-05
x5             4.0166      0.640      6.274      0.000       2.759       5.274
x6             5.0759      0.404     12.554      0.000       4.282       5.870
x7            -8.1721      1.067     -7.659      0.000     -10.268      -6.076
x8             3.9893      0.318     12.540      0.000       3.364       4.614
x9             1.4664      0.350      4.185      0.000       0.778       2.155
==============================================================================
Omnibus:                      519.769   Durbin-Watson:                   1.869
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            71270.463
Skew:                           3.352   Prob(JB):                         0.00
Kurtosis:                      57.176   Cond. No.                     7.16e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.16e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The scatter plots of top 3 features are shown in the following.



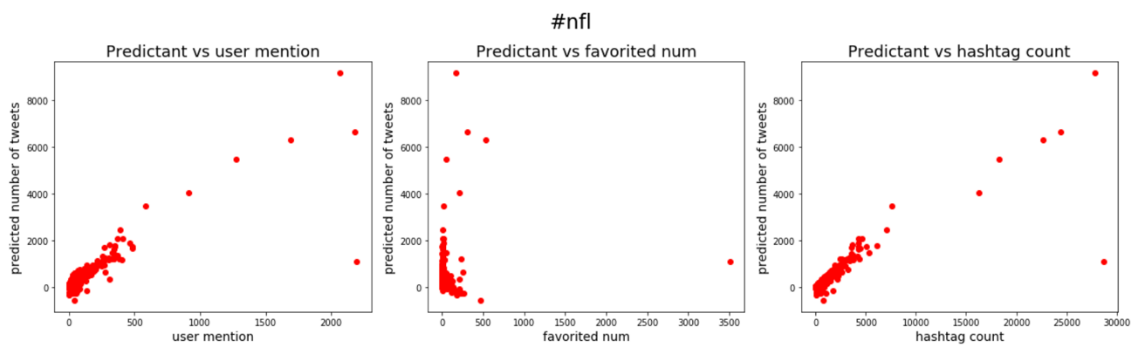
#Nfl

RMSE of #nfl for the linear regression model is: 402.63.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.742
Model:                            OLS   Adj. R-squared:                  0.738
Method:                 Least Squares   F-statistic:                     184.4
Date:                Mon, 18 Mar 2019   Prob (F-statistic):          3.16e-163
Time:                        20:11:30   Log-Likelihood:                -4346.3
No. Observations:                 586   AIC:                             8713.
Df Residuals:                     576   BIC:                             8756.
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         21.3249     36.195      0.589      0.556     -49.766      92.416
x1            -2.2800      2.477     -0.921      0.358      -7.144       2.584
x2            -2.5960      1.447     -1.794      0.073      -5.438       0.246
x3            -0.1690      0.054     -3.103      0.002      -0.276      -0.062
x4         -1.132e-05   1.15e-05     -0.987      0.324   -3.39e-05    1.12e-05
x5             0.4485      0.130      3.447      0.001       0.193       0.704
x6             2.1597      0.532      4.063      0.000       1.116       3.204
x7            -1.9673      0.168    -11.737      0.000      -2.296      -1.638
x8             0.3016      0.299      1.007      0.314      -0.286       0.890
x9             0.6082      0.093      6.542      0.000       0.426       0.791
==============================================================================
Omnibus:                      752.560   Durbin-Watson:                   2.617
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           146029.018
Skew:                           6.225   Prob(JB):                         0.00
Kurtosis:                      79.326   Cond. No.                     9.10e+06
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 9.1e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The scatter plots of top 3 features are shown in the following.



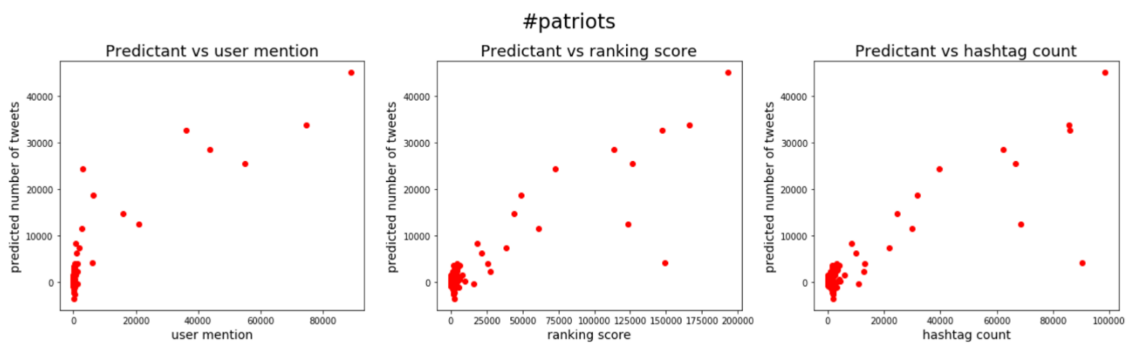
#Patriots

RMSE of #patriots for the linear regression model is: 1712.21.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.812
Model:                            OLS   Adj. R-squared:                  0.809
Method:                 Least Squares   F-statistic:                     277.1
Date:                Mon, 18 Mar 2019   Prob (F-statistic):          8.86e-203
Time:                        20:11:37   Log-Likelihood:                -5194.6
No. Observations:                 586   AIC:                         1.041e+04
Df Residuals:                     576   BIC:                         1.045e+04
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -265.8493    145.074     -1.833      0.067    -550.787      19.089
x1            -0.7834     10.432     -0.075      0.940     -21.274      19.707
x2           -66.5270      4.346    -15.306      0.000     -75.064     -57.990
x3            -0.3918      0.079     -4.938      0.000      -0.548      -0.236
x4             1.2e-05  3.36e-05      0.357      0.721    -5.4e-05     7.8e-05
x5            -3.9497      1.577     -2.504      0.013      -7.047      -0.852
x6             6.6101      0.831      7.953      0.000       4.978       8.242
x7             0.6897      0.293      2.350      0.019       0.113       1.266
x8            12.4921      0.869     14.368      0.000      10.784      14.200
x9             3.7941      0.360     10.549      0.000       3.088       4.500
==============================================================================
Omnibus:                     1066.765   Durbin-Watson:                   1.799
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1172882.536
Skew:                          11.663   Prob(JB):                         0.00
Kurtosis:                     220.927   Cond. No.                     1.66e+07
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.66e+07. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The scatter plots of top 3 features are shown in the following.



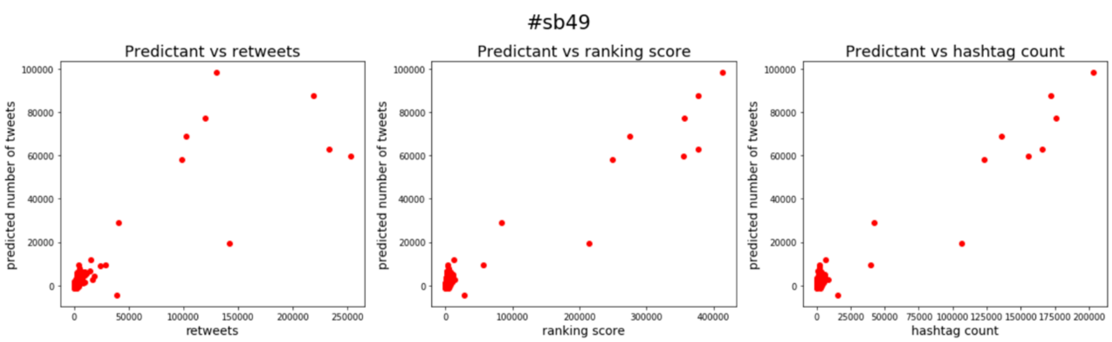
#Sb49

RMSE of #sb49 for the linear regression model is: 3598.65.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.844
Model:                            OLS   Adj. R-squared:                  0.841
Method:                 Least Squares   F-statistic:                     343.0
Date:                Mon, 18 Mar 2019   Prob (F-statistic):          5.97e-224
Time:                        20:11:53   Log-Likelihood:                -5591.4
No. Observations:                 582   AIC:                         1.120e+04
Df Residuals:                     572   BIC:                         1.125e+04
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -156.6960    297.448     -0.527      0.599    -740.920     427.528
x1           -16.6172     21.985     -0.756      0.450     -59.799      26.564
x2           -50.6916      8.128     -6.237      0.000     -66.656     -34.727
x3             0.4060      0.109      3.732      0.000       0.192       0.620
x4          6.705e-05   1.44e-05      4.664      0.000    3.88e-05    9.53e-05
x5            -1.3470      1.365     -0.987      0.324      -4.028       1.334
x6             4.8624      0.638      7.627      0.000       3.610       6.115
x7            -0.2548      0.096     -2.657      0.008      -0.443      -0.066
x8             9.4073      1.688      5.574      0.000       6.092      12.722
x9             2.0936      0.322      6.510      0.000       1.462       2.725
==============================================================================
Omnibus:                     1153.188   Durbin-Watson:                   2.007
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1766217.304
Skew:                          13.951   Prob(JB):                         0.00
Kurtosis:                     271.431   Cond. No.                     1.58e+08
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.58e+08. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The scatter plots of top 3 features are shown in the following.



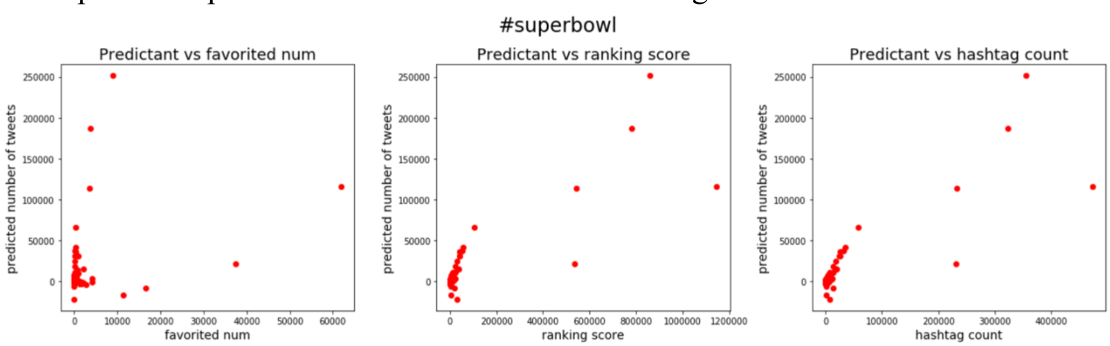
#Superbowl

RMSE of #superbowl for the linear regression model is: 5386.49.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.844
Model:                            OLS   Adj. R-squared:                  0.841
Method:                 Least Squares   F-statistic:                     343.0
Date:                Mon, 18 Mar 2019   Prob (F-statistic):          5.97e-224
Time:                        20:11:53   Log-Likelihood:                -5591.4
No. Observations:                 582   AIC:                         1.120e+04
Df Residuals:                     572   BIC:                         1.125e+04
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -156.6960    297.448     -0.527      0.599    -740.920     427.528
x1           -16.6172     21.985     -0.756      0.450     -59.799      26.564
x2           -50.6916      8.128     -6.237      0.000     -66.656     -34.727
x3             0.4060      0.109      3.732      0.000       0.192       0.620
x4          6.705e-05   1.44e-05      4.664      0.000    3.88e-05    9.53e-05
x5            -1.3470      1.365     -0.987      0.324      -4.028       1.334
x6             4.8624      0.638      7.627      0.000       3.610       6.115
x7            -0.2548      0.096     -2.657      0.008      -0.443      -0.066
x8             9.4073      1.688      5.574      0.000       6.092      12.722
x9             2.0936      0.322      6.510      0.000       1.462       2.725
==============================================================================
Omnibus:                     1153.188   Durbin-Watson:                   2.007
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1766217.304
Skew:                          13.951   Prob(JB):                         0.00
Kurtosis:                     271.431   Cond. No.                     1.58e+08
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.58e+08. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The scatter plots of top 3 features are shown in the following.



The regression coefficients do not always agree with the trend in the plots. The correlation between features might be positive and the corresponding coefficients could be positive or negative.

Question 6-7.

	Window 1 MSE	Window 1 R^2 score	Window 2 MSE	Window 2 R^2 score	Window 3 MSE	Window 3 R^2 score
Gohawks	2911	-0.82	16.43	0.90	1282	0.77
Gopatriots	183.9	-0.16	3.05	0.85	819.2	0.50
Nfl	256.3	0.49	9.38	0.88	1176	0.46
Patriots	1731	-0.18	381.2	0.88	3504	0.80
Sb49	59.08	0.85	848.3	0.87	3546	0.93
Superbowl	1409	-0.66	34.56	0.94	28771	0.78
Aggregate	1191	-0.29	87.40	0.84	31466	0.75

If we use MSE as the measurement, the aggregate model performs moderately compared with the individual models, for both Window 1 and Window 2, and the aggregate model perform the worst. If we use R square score as the measurement, the aggregate model performs moderate for Window 1 and bad for Window 2 and Windom 3. We think it is because the tweets from different hashtags have systematic difference and thus it is easier to model some hashtags than others. In addition, the time windows correspond to the sports event and people share different tweets with different predictable levels under different hashtags.

Question 8.
The optimal parameters for Random Forest Regressors are reported in the following:

	Max Depth	Max Features	Min samples leaf	Min samples split	N estimators	Negated MSE
Random Forest	80	Sqrt	1	5	200	-1.8e+8
Gradient Boosting Regressor	None	Sqrt	2	5	1600	-3.3e+8
OLS	None	Sqrt	1	10	200	-5.0e+5

The errors from random forest and gradient boosting look really bad. There might be a few reasons contributing to that. The data is time varying, and as a result, in each time period, the distribution changes according to the sports events. The errors may indicate that the data varies in a huge range.

Question 9.

	Max Depth	Max Features	Min samples leaf	Min samples split	N estimators	Negated MSE
OLS	None	Sqrt	1	10	200	-5.0e+5

When nonlinear estimators can be flexible and fit the training data well, they risk overfitting. Therefore, random forest and gradient boosting work worse than OLS.

Question 10.

	Max Depth	Max Features	Min samples leaf	Min samples split	N estimators	Negated MSE
Window1	40	Sqrt	4	2	400	-8.3e+5
Window2	20	Sqrt	4	2	200	-7.2e+6
Window3	20	Auto	2	2	200	-2.8e+5

They mostly do not agree with the best parameters for aggregated data. The reason might be that the sport events have fluid time dynamics and thus each time period needs different set of parameters to fit well.

Question 11.

MSE for different hidden layer number and different hidden layer size are reported as follows:

	1 hidden layer	2 hidden layers	3 hidden layers	4 hidden layers
Hidden layer size 5	128835.719	128883.142	128882.808	126990.467
Hidden layer size 50	128883.493	128363.165	127049.786	128776
Hidden layer size 100	128874	128846	127334	127361
Hidden layer size 200	119621	128883	127807	128083

Question 12.

We use one hidden layer and 200 hidden units. The MSE is 0.43. The performance improves dramatically.

Question 13.

The parameters we used are in the following:

```
Neural_grid={
    'hidden_layer_sizes':[(100,),(100,100),(100,100,100),(200,),(200,200),
(200,200,200)],'learning_rate':['adaptive'],'max_iter':[200],'learning_rat
e_init':[0.01],'alpha':[0.0001],'verbose':[10]}
```

The results are reported in the following:

```
window1 MSE for Neural network -0.58
window1 best parameter for Neural Network {'alpha': 0.0001, 'hidden_layer_
sizes': (100,), 'learning_rate': 'adaptive', 'learning_rate_init': 0.01, '
max_iter': 200, 'verbose': 10}

window2 error -0.14
window2 best parameter {'alpha': 0.0001, 'hidden_layer_sizes': (200, 200,
200), 'learning_rate': 'adaptive', 'learning_rate_init': 0.01, 'max_iter':
 200, 'verbose': 10}

window3 error -0.33
window3 best parameter {'alpha': 0.0001, 'hidden_layer_sizes': (100, 100),
 'learning_rate': 'adaptive', 'learning_rate_init': 0.01, 'max_iter': 200,
 'verbose': 10}
```

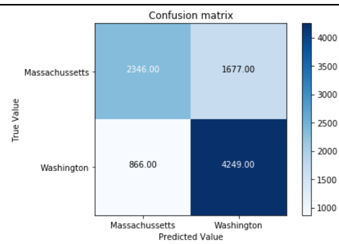
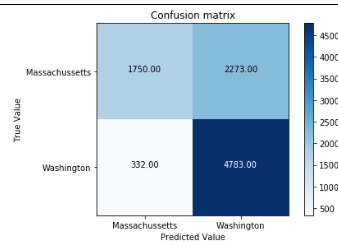
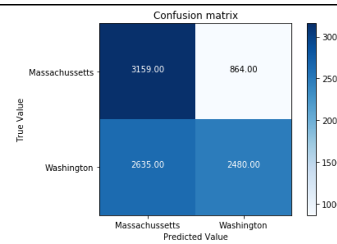
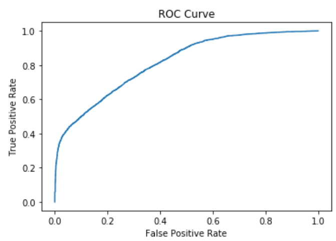
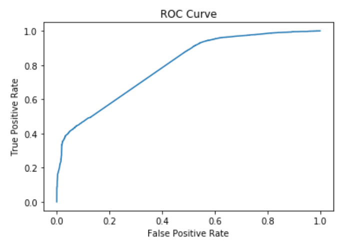
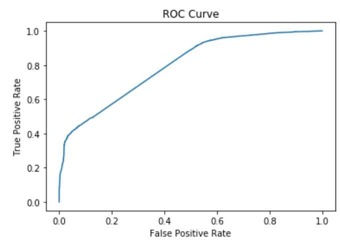
Question 14.

In this exercise, we used three models: linear regression, MLP Regression and random forest regression, as used before. We report the cross validation errors for each file to compare the model performance in the following. We conclude that the random forest model performs the best. Therefore, we report the predicted value from random forest model.

	Linear	MLP	Random Forest	Predicted Value from Random Forest Model
Preactive0	-3964	-11098357515	-307	117
Preactive1	-789	-728848	-40	846
Preactive2	-27	-6394958	-12	98
Active0	-77	-294826	-16	1145
Active1	-335	-7270613	-14	924
Active2	-119	-435	-3.26	212
Postactive0	-55260	-4962846	-15	77
Postactive1	-34	-72584825	-1.26	36
Postactive2	-1.39	-13624441	0.29	33

Part 2
Question 15.

1. Explain the method you use to determine whether the location is in Washington, Massachusetts or neither: I set up the flags for these two locations. For Massachusetts, the flags are “Massachusetts”, “Boston”, “MA”, and for Washington, the flags are “Washington”, “Seattle”, “WA”. We process each tweets, if the location fits the flags exactly, we mark it as what we need for the next part.
2. The three methods we used are Logistic Regression Classifier, Random Forest Classifier and Gaussian Naïve Bayes Classifier. The ROC curve, confusion matrix, accuracy, recall and precision are presented in the following.

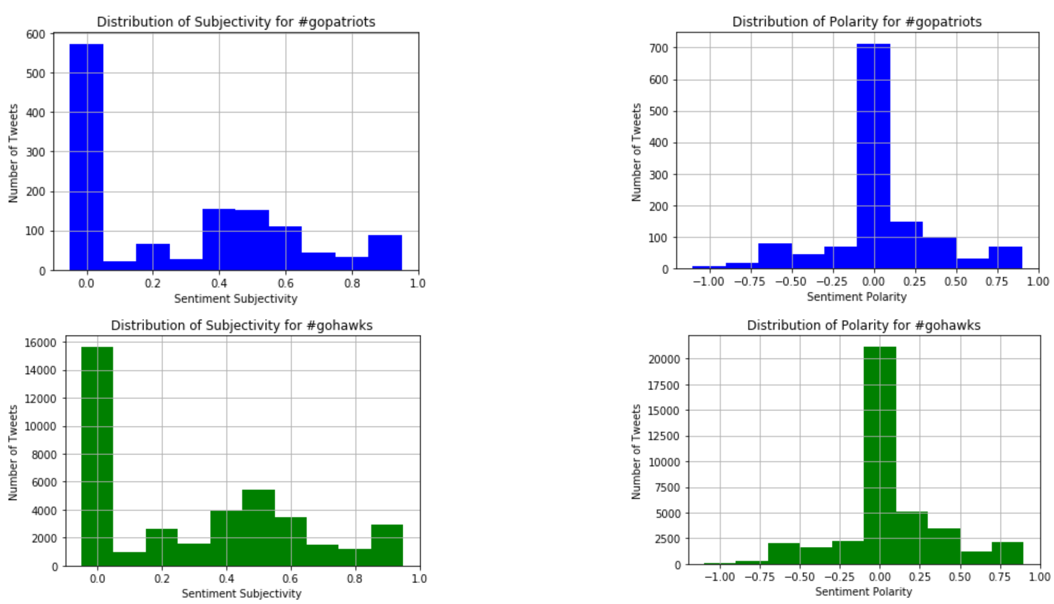
<table>
<tr><th></th><th>Logistic Regression Classifier</th><th>Random Forest Classifier</th><th>Gaussian Naïve Bayes Classifier</th></tr>
<tr><td></td><td>
</td><td>
</td><td>
</td></tr>
<tr><td></td><td>
</td><td>
</td><td>
</td></tr>
<tr><td>Accuracy</td><td>0.7217</td><td>0.7196</td><td>0.6171</td></tr>
<tr><td>Precision</td><td>0.7170</td><td>0.6815</td><td>0.7416</td></tr>
<tr><td>Recall</td><td>0.8307</td><td>0.9370</td><td>0.4848</td></tr>
</table>

Among all the classifiers, we can see that logistic regression and random forest classifiers perform similarly and better than Gaussian Naïve Bayes Classifier.

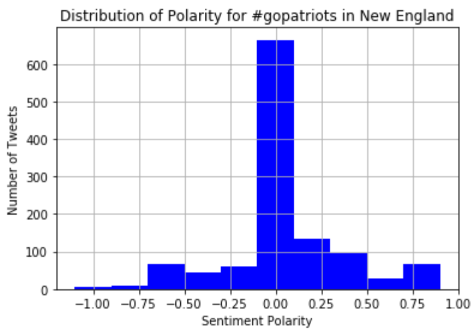
Part 3
Question 16. Define Your Own Project

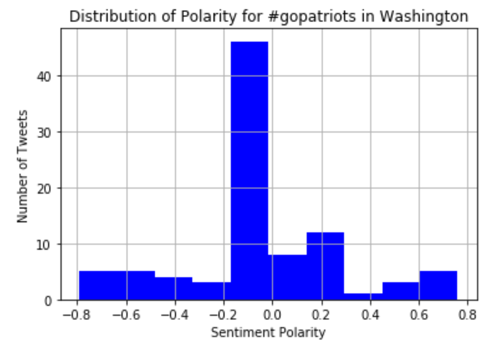
In Super Bowl, fans of the final two teams are region-based: Patriots is a team from New England. Naturally, most of Patriot fans are likely to live in New England area. Similarly, Hawks is a team based in Seattle, Washington. Most of Hawks fan are likely to live in Seattle, Washington. However, there might be some Patriot fans live in Washington and Hawks fan live in New England. We think it would be interesting to see that, as a minority, whether these people will post more negative and subjective tweets online. Another pattern we predict is that, different fans in the same region will display the opposite sentiment in their tweets – imagine there is a good news for the Hawks, when Hawks fans in Seattle post positively and celebrate on Twitter, Patriots fans in Seattle are probably bitter about it and post negatively. We aim to discover some pattern about this sentiment analysis and provide insight in the future prediction exercise.

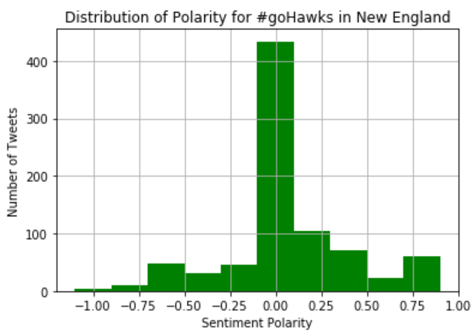
We study two sentiment dimensions – polarity and subjectivity. We predict that, fans of the opposite team will display more negativity and subjectivity in their tweets. The locations to identify for Hawks are Washington, Seattle and WA. As Patriots is based on New England, the locations we selected include six states, Massachusetts, Rhode Island, Connecticut, New Hampshire, Maine, Vermont, MA, VT, NH, RI, CT, ME, and major cities in these states, Worcester, Providence, Springfield, New Haven, Hartford, Stamford and Boston. After we clean up the tweets, the number of #goPatriots tweets is 1270 and the number of #goHawks tweets is 39308. The number of #goPatriots tweets with location Washington is 92 (7.2%) and the numbers of #goHawks tweets with location New England is 820 (2.1%). The distributions of polarity and subjectivity from two hashtags are shown in the following. We can see that, the distributions for two hashtags are similar, even the numbers of tweets are different. It means the samples are comparable. From the above numbers, we can also infer that there are indeed fans of the opposite teams in both New England and Washington.

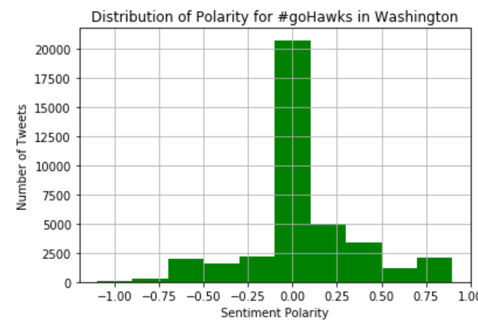


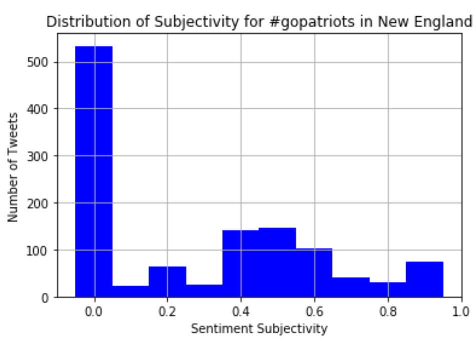
We first plot a few histograms to obtain model free evidence. It is clear that fans of patriots in different locations behave differently in terms of both polarity and subjectivity. However, Hawks fans in different locations behave similarly, in terms of both polarity and subjectivity. In addition, Patriots fans in Washington posted more negative tweets. Subjectivity might not be a good measure of sports tweets in general.

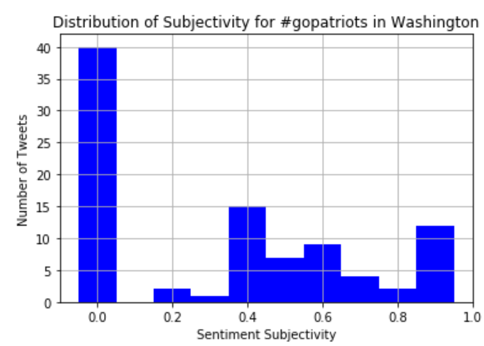


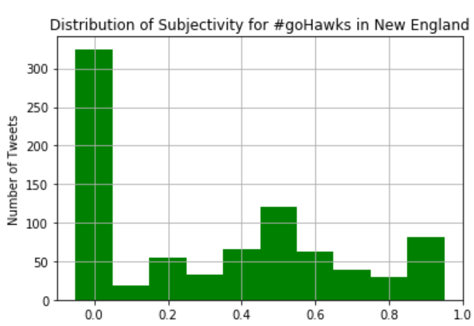


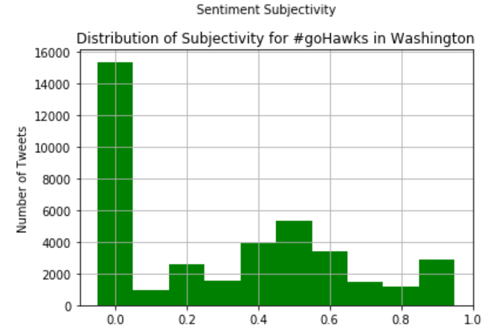












Then the most interesting question would be: can tweets sentiment of the fans in one location predict the fans for the same team in the other location? How about the tweets sentiment for the fans of the opposite team?

We average the tweet sentiments by team, location and dates. For the missing data points, we use zeros and then we plot the four lines for each sentiment measure. In general, the polarities from opposite teams moves oppositely. The subjectivity measurement tends to co-move together. Again, it shows that subjectivity measurement might not be a good measurement for sports tweets.

Not surprisingly, in New England, fans of opposite teams have opposite polarity measure. There might not be enough tweets posted by patriots fans in Washington and thus it is hard to spot the pattern. However, we can see that Hawks fans tend to have similar polarity in both locations. Subjectivity-wise, all of the lines seem to co-move together. Therefore, polarity of tweets from one team one location could probably be predicted by polarity of tweets from the opposite team or another location. The future research could combine multiple sources of data and predict the sentiment of tweets during the sports event.