ECE 219 Large Scale Data Mining

Project 5: Application - Twitter data

Siyuan Liu (404254996)

Zixuan Rong (004288960)

Zhuyun Xiao (504592699)

Zhichun Li (204332387)

Introduction

With 330 million monthly active users, Twitter, the popular social media, is a desirable platform to predict the future popularity of certain subjects or events. Predictions can be made according to current and previous data on twitters, shining light on the trends for the tweet activities with certain hashtags.

For this project, we obtained a set of data collected by querying the hashtags of popular topics related to the 2015 super bowl, time ranging from two weeks before the game to one week after the game. The test data consists of tweets containing hashtags in the specified time window. Various regression algorithms are applied to the data including linear regressions and non-linear regressions. We demonstrated the predictions results using our regression models on different topics.

Part 1: Popularity Prediction

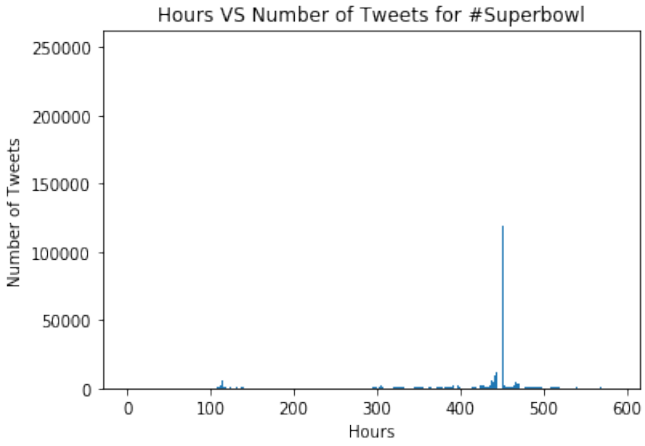
1. A first look at the data

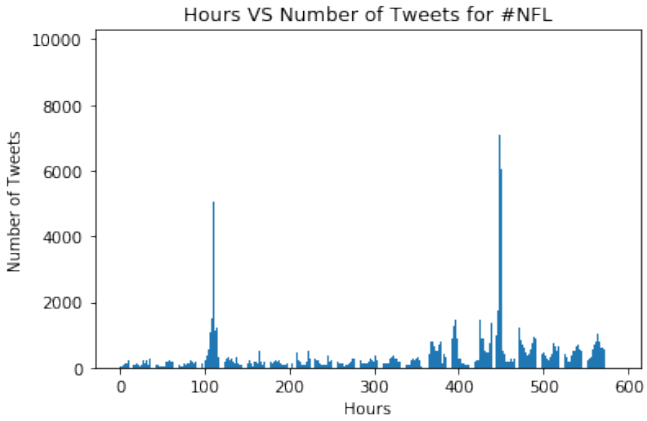
In this section, we obtained our training tweet data. The training data consist of 6 sets of tweet data with different hashtags. We presented the statistics for each hashtag.

QUESTION 1: Report the following statistics for each hashtag.

Hashtag	Average number of tweets per hour	Average number of followers of users posting the tweets	Average number of retweets per tweet
#NFL	396.97103918228277	15652.217242223855	5.15138810851306
#SuperBowl	2067.824531516184	16707.274326499828	4.532093631865189
#GoHawks	288.11243611584325	5208.0543437513015	4.727350985101982
#GoPatriots	40.052810902896084	2035.3087887426457	2.008127615697216
#Patriots	750.6320272572402	4852.411315256763	2.6407693004924786
#SB49	1266.8637137989779	14343.25328288697	3.4940012828497857

QUESTION 2: Plot “number of tweets in hour” over time for #SuperBowl and #NFL





From the plots, we can see that the number of tweets is not evenly distributed. It has several peaks within certain hours.

2. Linear regression

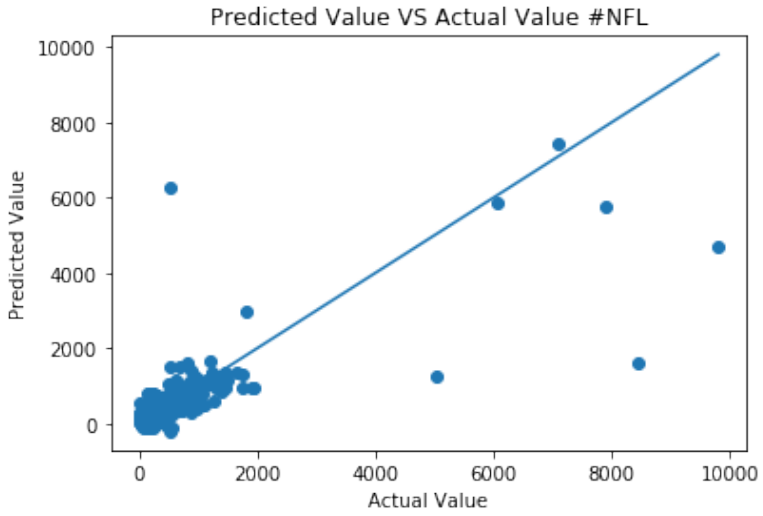
In this part, time windows with a duration of an hour were set up to extract features. We determined the hour slot of the tweets using their timestamp and presented the features for the time slots. We trained a linear regression model using five features that are extracted from tweet data in the previous hour to predict the number of tweets in the next hour. The features include number of tweets, the total number of retweets, the sum of the number of followers of the users posting the hashtag, maximum number of followers of the users posting the hashtag and time of the day. We trained a separate model for each hashtag.

QUESTION 3: For each of your models, report your model's Mean Squared Error (MSE) and R-squared measure. Also, analyze the significance of each feature using the t-test and p-value.

The features we used are Number of tweets (tweet), Total number of retweets (retweet), Sum of the number of followers of the users posting the hashtag (follower), Maximum number of followers of the users posting the hashtag (Maxfollowers) and Time of the day. For time of the day, we first used one hot encoding to convert the time into 24 values in PST time zone.

#NFL:

RMSE is: 503.96521944684315 MSE: 253980.94241210477309



OLS Regression Results

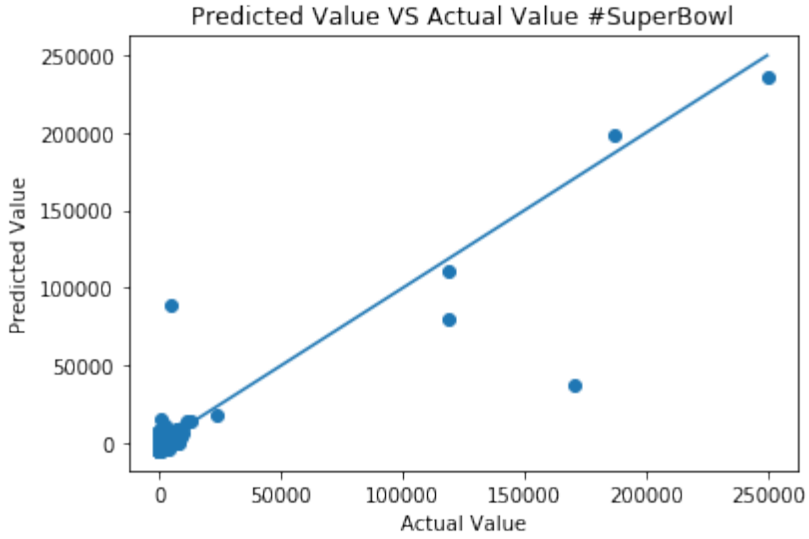
Dep. Variable:	tweet	R-squared:	0.596
Model:	OLS	Adj. R-squared:	0.577
Method:	Least Squares	F-statistic:	30.52
Date:	Thu, 19 Mar 2020	Prob (F-statistic):	5.88e-92
Time:	11:24:58	Log-Likelihood:	-4477.9
No. Observations:	586	AIC:	9012.
Df Residuals:	558	BIC:	9134.
Df Model:	27		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
tweet	0.5597	0.136	4.110	0.000	0.292	0.827
follower	0.0001	2.5e-05	4.596	0.000	6.58e-05	0.000
retweet	-0.1634	0.064	-2.544	0.011	-0.290	-0.037
Maxfollowers	-0.0001	3.31e-05	-3.863	0.000	-0.000	-6.29e-05
hour_0	46.9386	103.751	0.452	0.651	-156.852	250.729
hour_1	62.2417	103.850	0.599	0.549	-141.743	266.227
hour_2	73.1236	103.690	0.705	0.481	-130.547	276.795
hour_3	80.7387	103.675	0.779	0.436	-122.902	284.379
hour_4	81.4481	103.761	0.785	0.433	-122.362	285.258
hour_5	144.2520	103.836	1.389	0.165	-59.705	348.210
hour_6	181.6911	104.220	1.743	0.082	-23.020	386.403
hour_7	141.2195	104.214	1.355	0.176	-63.481	345.920
hour_8	144.7466	104.732	1.382	0.168	-60.971	350.464
hour_9	111.5197	104.276	1.069	0.285	-93.302	316.342
hour_10	218.3326	106.633	2.048	0.041	8.881	427.784
hour_11	229.7065	107.163	2.144	0.033	19.215	440.198
hour_12	90.6092	107.027	0.847	0.398	-119.616	300.835
hour_13	210.7626	108.314	1.946	0.052	-1.990	423.515
hour_14	568.7573	107.525	5.290	0.000	357.555	779.960
hour_15	22.1314	109.649	0.202	0.840	-193.244	237.507
hour_16	242.8060	108.653	2.235	0.026	29.388	456.224
hour_17	212.2537	110.672	1.918	0.056	-5.130	429.638
hour_18	272.6349	109.075	2.500	0.013	58.388	486.882
hour_19	-162.4449	108.502	-1.497	0.135	-375.568	50.678
hour_20	119.0350	107.200	1.110	0.267	-91.531	329.601
hour_21	87.3104	106.488	0.820	0.413	-121.855	296.476
hour_22	34.1922	106.130	0.322	0.747	-174.271	242.655
hour_23	56.5382	105.974	0.534	0.594	-151.619	264.695

Omnibus:	641.756	Durbin-Watson:	2.378
Prob(Omnibus):	0.000	Jarque-Bera (JB):	283134.102
Skew:	4.283	Prob(JB):	0.00
Kurtosis:	110.343	Cond. No.	2.73e+07

#SuperBowl:

RMSE is: 7073.15695692096 MSE: $5.0029549337239375 \times 10^7$



OLS Regression Results

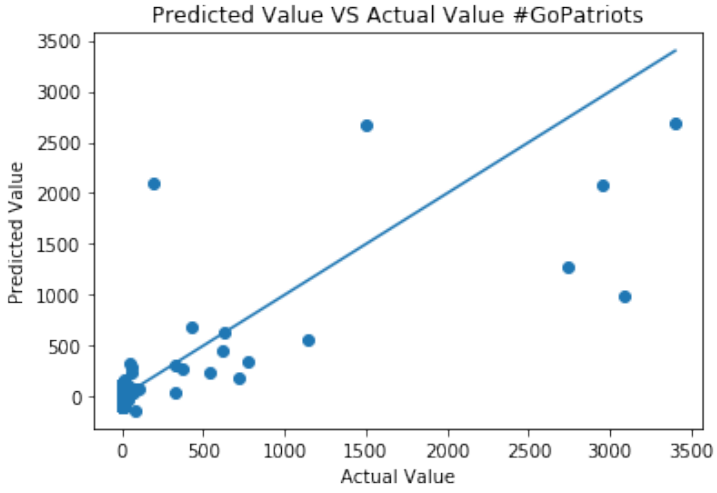
Dep. Variable:	tweet	R-squared:	0.809
Model:	OLS	Adj. R-squared:	0.800
Method:	Least Squares	F-statistic:	87.65
Date:	Thu, 19 Mar 2020	Prob (F-statistic):	4.01e-181
Time:	11:24:58	Log-Likelihood:	-6025.8
No. Observations:	586	AIC:	1.211e+04
Df Residuals:	558	BIC:	1.223e+04
Df Model:	27		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
tweet	2.2820	0.081	28.230	0.000	2.123	2.441
follower	-0.0001	2.24e-05	-6.170	0.000	-0.000	-9.4e-05
retweet	-0.2548	0.047	-5.449	0.000	-0.347	-0.163
Maxfollowers	0.0007	0.000	4.716	0.000	0.000	0.001
hour_0	-627.1772	1461.869	-0.429	0.668	-3498.617	2244.262
hour_1	-544.6397	1456.198	-0.374	0.709	-3404.939	2315.659
hour_2	-358.2494	1454.886	-0.246	0.806	-3215.973	2499.474
hour_3	-303.8166	1461.437	-0.208	0.835	-3174.406	2566.773
hour_4	-151.9279	1453.657	-0.105	0.917	-3007.237	2703.381
hour_5	-328.1518	1461.985	-0.224	0.822	-3199.818	2543.514
hour_6	-290.2115	1463.970	-0.198	0.843	-3165.778	2585.355
hour_7	-574.3447	1465.464	-0.392	0.695	-3452.844	2304.155
hour_8	-90.1660	1469.705	-0.061	0.951	-2976.996	2796.664
hour_9	-504.5772	1474.978	-0.342	0.732	-3401.765	2392.610
hour_10	-773.9733	1508.890	-0.513	0.608	-3737.772	2189.825
hour_11	-324.4393	1499.400	-0.216	0.829	-3269.597	2620.718
hour_12	-392.6190	1503.247	-0.261	0.794	-3345.334	2560.096
hour_13	11.8776	1499.421	0.008	0.994	-2933.322	2957.077
hour_14	5780.5569	1504.402	3.842	0.000	2825.574	8735.540
hour_15	-1423.4358	1517.042	-0.938	0.348	-4403.246	1556.374
hour_16	-75.9232	1528.351	-0.050	0.960	-3077.948	2926.102
hour_17	-766.8710	1560.130	-0.492	0.623	-3831.316	2297.574
hour_18	441.1922	1518.169	0.291	0.771	-2540.833	3423.218
hour_19	-4813.4677	1501.122	-3.207	0.001	-7762.008	-1864.928
hour_20	-464.9572	1491.343	-0.312	0.755	-3394.290	2464.376
hour_21	-718.2255	1496.142	-0.480	0.631	-3656.985	2220.534
hour_22	-508.3395	1485.156	-0.342	0.732	-3425.519	2408.840
hour_23	-773.3304	1491.725	-0.518	0.604	-3703.413	2156.752

Omnibus:	942.241	Durbin-Watson:	2.294
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1498272.212
Skew:	8.682	Prob(JB):	0.00
Kurtosis:	250.106	Cond. No.	6.82e+08

#GoPatriots

RMSE is: 159.86345136919155 MSE: 25556.32308366987166



OLS Regression Results

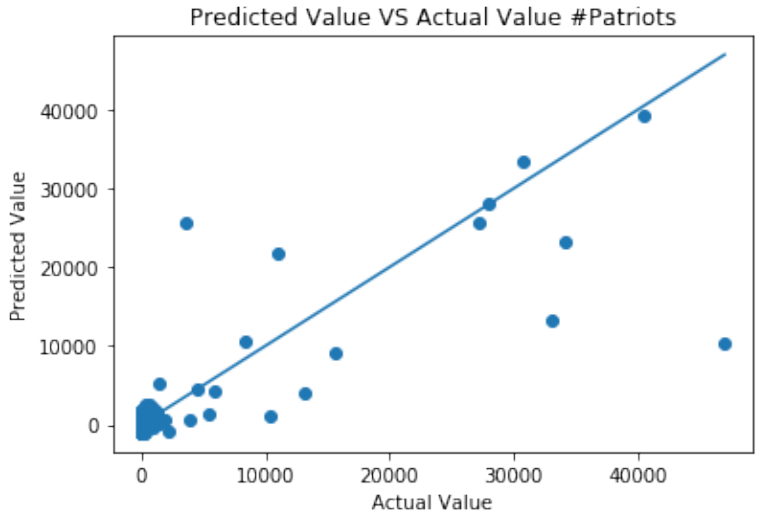
Dep. Variable:	tweet	R-squared:	0.650
Model:	OLS	Adj. R-squared:	0.633
Method:	Least Squares	F-statistic:	38.32
Date:	Thu, 19 Mar 2020	Prob (F-statistic):	1.15e-108
Time:	11:24:59	Log-Likelihood:	-3805.0
No. Observations:	586	AIC:	7666.
Df Residuals:	558	BIC:	7789.
Df Model:	27		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
tweet	0.2981	0.285	1.046	0.296	-0.262	0.858
follower	-6.495e-05	0.000	-0.300	0.764	-0.000	0.000
retweet	0.4645	0.190	2.439	0.015	0.090	0.839
Maxfollowers	-6.906e-05	0.000	-0.316	0.752	-0.000	0.000
hour_0	-0.2054	32.769	-0.006	0.995	-64.570	64.159
hour_1	0.7074	32.767	0.022	0.983	-63.655	65.070
hour_2	-0.2578	32.771	-0.008	0.994	-64.627	64.111
hour_3	1.2990	32.769	0.040	0.968	-63.068	65.666
hour_4	1.2627	32.771	0.039	0.969	-63.107	65.632
hour_5	2.6259	32.777	0.080	0.936	-61.756	67.008
hour_6	-1.2054	32.810	-0.037	0.971	-65.652	63.242
hour_7	3.5069	32.778	0.107	0.915	-60.877	67.891
hour_8	2.4775	32.789	0.076	0.940	-61.928	66.883
hour_9	9.0683	32.816	0.276	0.782	-55.389	73.525
hour_10	6.7564	33.476	0.202	0.840	-58.998	72.511
hour_11	19.0336	33.564	0.567	0.571	-46.895	84.962
hour_12	6.7989	33.555	0.203	0.840	-59.110	72.708
hour_13	36.8736	34.136	1.080	0.281	-30.176	103.924
hour_14	119.8044	33.482	3.578	0.000	54.038	185.571
hour_15	30.4940	34.205	0.892	0.373	-36.693	97.681
hour_16	-51.2266	33.896	-1.511	0.131	-117.806	15.352
hour_17	81.2232	34.096	2.382	0.018	14.251	148.195
hour_18	26.5801	33.773	0.787	0.432	-39.758	92.918
hour_19	-99.1278	33.676	-2.944	0.003	-165.275	-32.981
hour_20	-11.9714	33.584	-0.356	0.722	-77.937	53.995
hour_21	-7.9939	33.678	-0.237	0.812	-74.146	58.158
hour_22	-0.4710	33.453	-0.014	0.989	-66.180	65.238
hour_23	-3.6580	33.467	-0.109	0.913	-69.395	62.079

Omnibus:	462.561	Durbin-Watson:	1.920
Prob(Omnibus):	0.000	Jarque-Bera (JB):	255424.584
Skew:	2.230	Prob(JB):	0.00
Kurtosis:	105.182	Cond. No.	1.99e+06

#Patriots:

RMSE is: 2225.567247490388 MSE: $4.953149573101941 \times 10^6$



OLS Regression Results

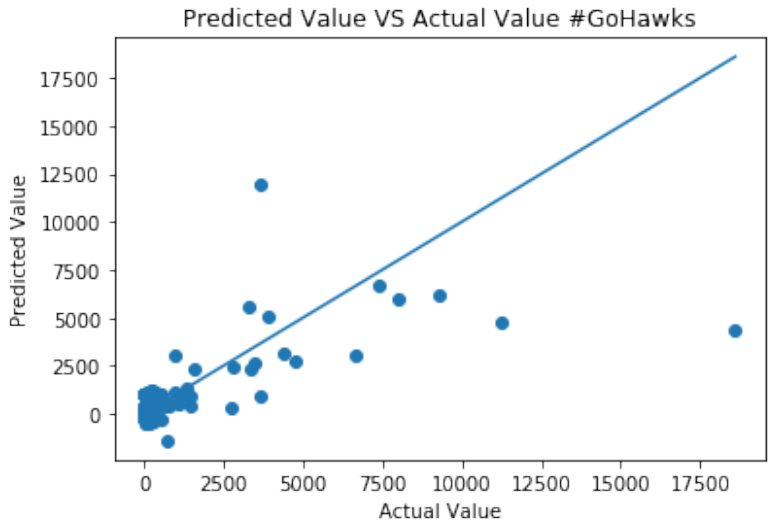
Dep. Variable:	tweet	R-squared:	0.683
Model:	OLS	Adj. R-squared:	0.668
Method:	Least Squares	F-statistic:	44.53
Date:	Thu, 19 Mar 2020	Prob (F-statistic):	1.59e-120
Time:	11:24:59	Log-Likelihood:	-5348.2
No. Observations:	586	AIC:	1.075e+04
Df Residuals:	558	BIC:	1.087e+04
Df Model:	27		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
tweet	0.9198	0.072	12.834	0.000	0.779	1.061
follower	-1.007e-05	2.68e-05	-0.376	0.707	-6.27e-05	4.25e-05
retweet	-0.0709	0.059	-1.210	0.227	-0.186	0.044
Maxfollowers	0.0001	9.4e-05	1.181	0.238	-7.36e-05	0.000
hour_0	3.4148	456.394	0.007	0.994	-893.046	899.875
hour_1	20.8313	456.180	0.046	0.964	-875.208	916.870
hour_2	31.4924	456.235	0.069	0.945	-864.656	927.641
hour_3	56.1884	456.619	0.123	0.902	-840.713	953.090
hour_4	44.8006	456.465	0.098	0.922	-851.799	941.400
hour_5	70.5169	458.229	0.154	0.878	-829.548	970.581
hour_6	77.3183	457.849	0.169	0.866	-822.000	976.637
hour_7	28.0104	459.817	0.061	0.951	-875.174	931.195
hour_8	72.9153	458.208	0.159	0.874	-827.109	972.940
hour_9	359.9939	458.478	0.785	0.433	-540.560	1260.548
hour_10	1660.5986	468.890	3.542	0.000	739.593	2581.605
hour_11	185.4529	471.327	0.393	0.694	-740.339	1111.245
hour_12	-271.9817	474.094	-0.574	0.566	-1203.209	659.246
hour_13	58.5785	473.060	0.124	0.901	-870.617	987.774
hour_14	645.7018	471.115	1.371	0.171	-279.673	1571.077
hour_15	13.5955	480.509	0.028	0.977	-930.231	957.422
hour_16	-500.7354	479.043	-1.045	0.296	-1441.683	440.212
hour_17	694.2742	470.634	1.475	0.141	-230.157	1618.706
hour_18	751.5527	472.321	1.591	0.112	-176.193	1679.298
hour_19	-1094.7975	470.377	-2.327	0.020	-2018.723	-170.872
hour_20	-17.3801	477.693	-0.036	0.971	-955.677	920.916
hour_21	-9.1948	466.857	-0.020	0.984	-926.207	907.817
hour_22	-2.9988	466.383	-0.006	0.995	-919.079	913.082
hour_23	-38.4039	465.920	-0.082	0.934	-953.576	876.768

Omnibus:	862.769	Durbin-Watson:	2.004
Prob(Omnibus):	0.000	Jarque-Bera (JB):	590091.435
Skew:	7.531	Prob(JB):	0.00
Kurtosis:	157.728	Cond. No.	4.69e+07

#GoHawks:

RMSE is: 837.4844185446983 MSE: 701380.15130515140181



OLS Regression Results

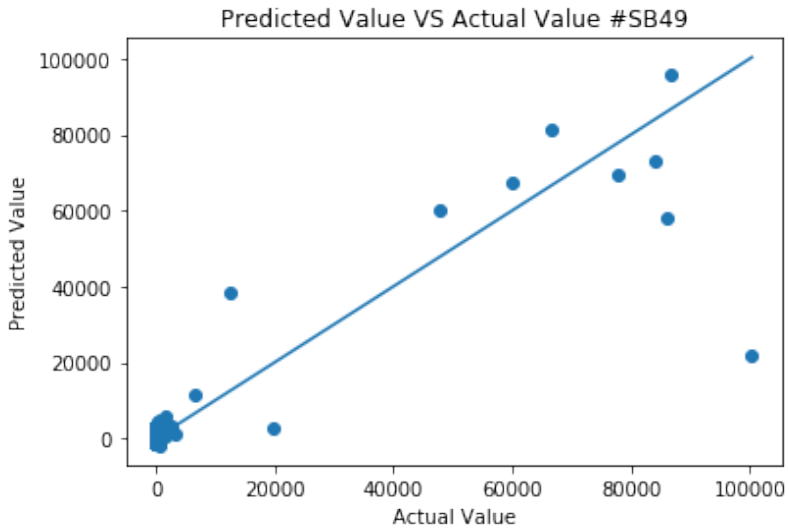
Dep. Variable:	tweet	R-squared:	0.510
Model:	OLS	Adj. R-squared:	0.486
Method:	Least Squares	F-statistic:	21.48
Date:	Thu, 19 Mar 2020	Prob (F-statistic):	3.04e-69
Time:	11:24:59	Log-Likelihood:	-4775.5
No. Observations:	586	AIC:	9607.
Df Residuals:	558	BIC:	9729.
Df Model:	27		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
tweet	1.2478	0.165	7.584	0.000	0.925	1.571
follower	-0.0002	8.01e-05	-2.156	0.032	-0.000	-1.54e-05
retweet	-0.1354	0.043	-3.117	0.002	-0.221	-0.050
Maxfollowers	1.685e-05	0.000	0.112	0.911	-0.000	0.000
hour_0	-2.8752	171.793	-0.017	0.987	-340.316	334.566
hour_1	-0.0936	171.660	-0.001	1.000	-337.272	337.085
hour_2	3.0977	171.672	0.018	0.986	-334.104	340.300
hour_3	8.7027	171.668	0.051	0.960	-328.492	345.898
hour_4	17.9492	171.662	0.105	0.917	-319.233	355.132
hour_5	40.8034	171.683	0.238	0.812	-296.421	378.028
hour_6	105.5482	172.130	0.613	0.540	-232.553	443.649
hour_7	108.8208	171.873	0.633	0.527	-228.775	446.417
hour_8	188.9001	173.744	1.087	0.277	-152.371	530.172
hour_9	164.1382	171.987	0.954	0.340	-173.682	501.959
hour_10	336.7087	177.891	1.893	0.059	-12.710	686.127
hour_11	288.4005	175.810	1.640	0.101	-56.931	633.732
hour_12	22.0898	176.300	0.125	0.900	-324.203	368.383
hour_13	233.9198	176.519	1.325	0.186	-112.804	580.643
hour_14	990.7793	176.682	5.608	0.000	643.737	1337.822
hour_15	-243.6159	181.713	-1.341	0.181	-600.541	113.309
hour_16	182.4294	177.305	1.029	0.304	-165.838	530.696
hour_17	183.3559	177.365	1.034	0.302	-165.029	531.741
hour_18	-49.0485	176.436	-0.278	0.781	-395.609	297.512
hour_19	-14.0769	180.921	-0.078	0.938	-369.446	341.292
hour_20	68.6616	176.422	0.389	0.697	-277.871	415.194
hour_21	133.4715	178.755	0.747	0.456	-217.643	484.586
hour_22	29.0413	176.298	0.165	0.869	-317.247	375.330
hour_23	-16.3524	175.364	-0.093	0.926	-360.807	328.102

Omnibus:	881.339	Durbin-Watson:	2.225
Prob(Omnibus):	0.000	Jarque-Bera (JB):	653123.111
Skew:	7.840	Prob(JB):	0.00
Kurtosis:	165.798	Cond. No.	1.45e+07

#SB49:

RMSE is: 3932.3293340677287 MSE: $1.54632139915695 \times 10^7$



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  tweet   R-squared:                       0.812
Model:                            OLS   Adj. R-squared:                  0.803
Method:                 Least Squares   F-statistic:                     89.31
Date:                Thu, 19 Mar 2020   Prob (F-statistic):          5.96e-183
Time:                        11:24:59   Log-Likelihood:                -5681.8
No. Observations:                 586   AIC:                         1.142e+04
Df Residuals:                     558   BIC:                         1.154e+04
Df Model:                          27
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
tweet          1.1201      0.089     12.560      0.000       0.945       1.295
follower    6.515e-06   1.28e-05      0.510      0.610   -1.86e-05    3.16e-05
retweet       -0.1406      0.080     -1.748      0.081      -0.299       0.017
Maxfollowers 9.141e-05   4.44e-05      2.060      0.040    4.24e-06       0.000
hour_0        -80.3388    807.242     -0.100      0.921   -1665.944    1505.266
hour_1        -44.0681    806.366     -0.055      0.956   -1627.952    1539.816
hour_2         -2.5386    806.255     -0.003      0.997   -1586.205    1581.128
hour_3         16.7774    806.119      0.021      0.983   -1566.621    1600.176
hour_4        -19.6455    806.862     -0.024      0.981   -1604.504    1565.213
hour_5         12.0103    807.986      0.015      0.988   -1575.056    1599.077
hour_6         13.6082    807.595      0.017      0.987   -1572.689    1599.905
hour_7       -103.0860    810.397     -0.127      0.899   -1694.888    1488.716
hour_8        -47.0966    811.095     -0.058      0.954   -1640.269    1546.076
hour_9        607.2890    811.030      0.749      0.454    -985.755    2200.333
hour_10      3248.3419    829.263      3.917      0.000    1619.482    4877.201
hour_11      -571.4947    834.063     -0.685      0.494   -2209.781    1066.792
hour_12      -778.0557    832.835     -0.934      0.351   -2413.930     857.818
hour_13      -521.6144    850.743     -0.613      0.540   -2192.664    1149.435
hour_14      1029.5449    835.368      1.232      0.218    -611.305    2670.395
hour_15       259.8007    837.855      0.310      0.757   -1385.934    1905.535
hour_16       245.6312    830.297      0.296      0.767   -1385.258    1876.520
hour_17      -630.6096    845.654     -0.746      0.456   -2291.665    1030.446
hour_18     -1212.5199    829.305     -1.462      0.144   -2841.461     416.421
hour_19      -306.5597    825.154     -0.372      0.710   -1927.347    1314.227
hour_20      -191.3052    828.341     -0.231      0.817   -1818.353    1435.743
hour_21        19.3785    826.534      0.023      0.981   -1604.120    1642.877
hour_22      -173.6016    827.105     -0.210      0.834   -1798.222    1451.018
hour_23       -77.4107    823.826     -0.094      0.925   -1695.589    1540.768
==============================================================================
Omnibus:                     1153.420   Durbin-Watson:                   1.669
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1889789.449
Skew:                          13.711   Prob(JB):                         0.00
Kurtosis:                     279.849   Cond. No.                     4.44e+08
==============================================================================
```

From the results, we demonstrated that for different hashtag files, the significances of the features are not the same. MSE measures the average of the squares of the errors. In our models, MSE increases as training data gets larger, which means the accuracy decreases. According to p-value and t-test, the predictor with lower p-value is more likely to be a meaningful addition or model. If |t| had larger magnitudes (either negative or positive), the possibility of the parameter is insignificant is decreased. From the results shown above, we concluded that the most significant parameters for each hashtag by choosing the features with lowest p-values. If several features all have very low p-values, the ones with larger absolute values of the t-values are chosen.

Hashtag	Most significant features
#NFL	Number of tweets, Sum of the number of followers, Maximum number of followers
#SuperBowl	Number of tweets, Sum of the number of followers, Total number of retweets
#GoPatriots	Total number of retweets, Total number of retweets
#Patriots	Number of tweets, Total number of retweets
#GoHawks	Number of tweets, Sum of the number of followers, Total number of retweets
#SB49	Number of tweets, Maximum number of followers, Total number of retweets

From the table, we showed that the feature that contributes the most to the linear regression model is Number of tweets.

3. Feature analysis

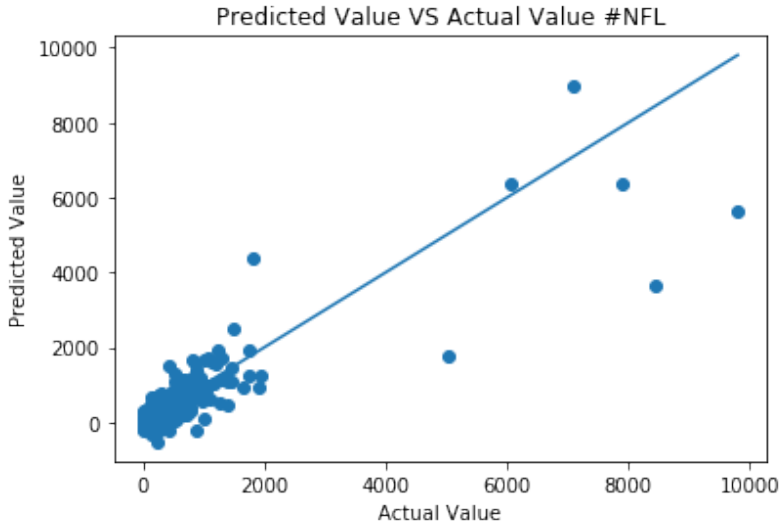
In this session, we applied different features to design a regression model and trained the model with the tweet dataset. For the top 3 features, we presented the scatter plot of the predicted number of tweets for next hour versus value of that feature. The new features include:

1. Hashtag: the total number of hashtags of the tweet
2. reply: the number of the tweets with replies
3. favourite_count: the number of favorites of tweets
4. verified: the number of the tweets with verified users
5. citation: the number of citations
7. ranking: the ranking scores of the users

QUESTION 4: Fit your model on the data of each hashtag and report fitting MSE and significance of features.

#NFL:

RMSE is: 394.5126755879352 MSE: 155640.251199551402



OLS Regression Results

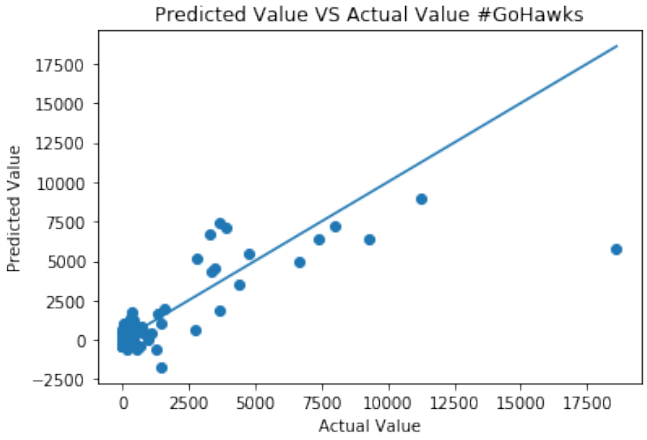
Dep. Variable:	tweet	R-squared:	0.753
Model:	OLS	Adj. R-squared:	0.739
Method:	Least Squares	F-statistic:	54.36
Date:	Thu, 19 Mar 2020	Prob (F-statistic):	1.80e-146
Time:	11:26:03	Log-Likelihood:	-4334.4
No. Observations:	586	AIC:	8733.
Df Residuals:	554	BIC:	8873.
Df Model:	31		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
tweet	-0.7247	0.510	-1.422	0.156	-1.726	0.277
follower	6.258e-06	2.08e-05	0.301	0.763	-3.45e-05	4.71e-05
retweet	-0.1053	0.053	-1.995	0.047	-0.209	-0.002
Maxfollowers	-1.245e-05	2.73e-05	-0.456	0.649	-6.61e-05	4.12e-05
Hashtag	0.6989	0.092	7.609	0.000	0.519	0.879
reply	-0.7247	0.510	-1.422	0.156	-1.726	0.277
favourite_count	-1.9342	0.176	-10.990	0.000	-2.280	-1.588
verified	-0.7247	0.510	-1.422	0.156	-1.726	0.277
citation	0.4386	0.134	3.266	0.001	0.175	0.702
ranking	0.2205	0.314	0.702	0.483	-0.396	0.837
hour_0	-82.5866	83.267	-0.992	0.322	-246.145	80.972
hour_1	-71.3333	82.918	-0.860	0.390	-234.205	91.538
hour_2	-51.4717	82.404	-0.625	0.532	-213.335	110.391
hour_3	-16.9339	82.330	-0.206	0.837	-178.652	144.784
hour_4	12.4898	82.670	0.151	0.880	-149.895	174.874
hour_5	75.9774	82.463	0.921	0.357	-86.000	237.955
hour_6	82.2105	83.651	0.983	0.326	-82.102	246.523
hour_7	27.5142	84.778	0.325	0.746	-139.011	194.039
hour_8	10.8792	85.744	0.127	0.899	-157.544	179.303
hour_9	-16.8073	86.423	-0.194	0.846	-186.564	152.949
hour_10	60.9782	87.702	0.695	0.487	-111.291	233.247
hour_11	39.3336	89.539	0.439	0.661	-136.543	215.210
hour_12	-99.6895	89.532	-1.113	0.266	-275.552	76.173
hour_13	-2.0177	88.929	-0.023	0.982	-176.697	172.662
hour_14	372.8839	88.775	4.200	0.000	198.508	547.260
hour_15	-234.6220	88.834	-2.641	0.008	-409.114	-60.130
hour_16	37.5257	88.820	0.422	0.673	-136.939	211.990
hour_17	29.8052	90.260	0.330	0.741	-147.488	207.098
hour_18	0.1221	89.706	0.001	0.999	-176.084	176.328
hour_19	-84.6402	88.061	-0.961	0.337	-257.614	88.334
hour_20	-69.6344	88.161	-0.790	0.430	-242.805	103.536
hour_21	-83.6191	86.782	-0.964	0.336	-254.082	86.843
hour_22	-147.0253	85.487	-1.720	0.086	-314.944	20.893
hour_23	-100.9054	85.263	-1.183	0.237	-268.383	66.573

Omnibus:	701.310	Durbin-Watson:	2.600
Prob(Omnibus):	0.000	Jarque-Bera (JB):	119289.783
Skew:	5.491	Prob(JB):	0.00
Kurtosis:	72.029	Cond. No.	2.87e+19

#GoHawks

RMSE is: 700.8128346896752 MSE: 491138.629265778019



OLS Regression Results

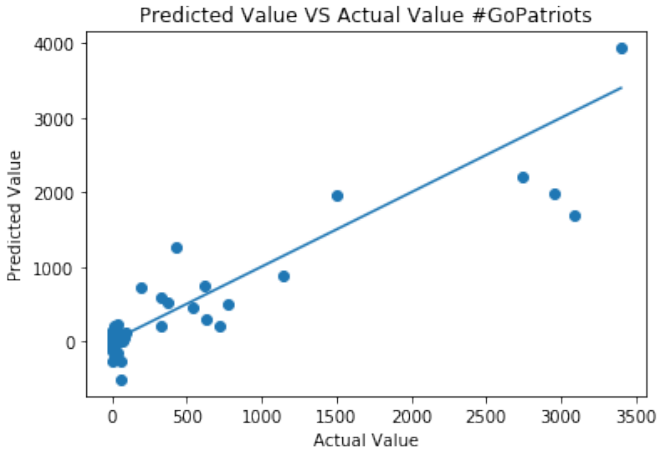
Dep. Variable:	tweet	R-squared:	0.657
Model:	OLS	Adj. R-squared:	0.637
Method:	Least Squares	F-statistic:	34.17
Date:	Thu, 19 Mar 2020	Prob (F-statistic):	6.88e-108
Time:	11:26:28	Log-Likelihood:	-4671.1
No. Observations:	586	AIC:	9406.
Df Residuals:	554	BIC:	9546.
Df Model:	31		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
tweet	-12.8481	1.273	-10.096	0.000	-15.348	-10.348
follower	-0.0005	7.25e-05	-7.149	0.000	-0.001	-0.000
retweet	-0.0169	0.055	-0.309	0.757	-0.124	0.091
Maxfollowers	0.0005	0.000	3.451	0.001	0.000	0.001
Hashtag	0.7487	0.337	2.220	0.027	0.086	1.411
reply	-12.8481	1.273	-10.096	0.000	-15.348	-10.348
favourite_count	0.0399	0.023	1.701	0.089	-0.006	0.086
verified	-12.8481	1.273	-10.096	0.000	-15.348	-10.348
citation	10.6028	1.424	7.445	0.000	7.806	13.400
ranking	8.1778	0.816	10.025	0.000	6.575	9.780
hour_0	8.2288	144.453	0.057	0.955	-275.514	291.972
hour_1	-37.8757	144.214	-0.263	0.793	-321.148	245.397
hour_2	-30.2534	144.230	-0.210	0.834	-313.557	253.050
hour_3	-6.9722	144.212	-0.048	0.961	-290.241	276.297
hour_4	0.0029	144.183	2.04e-05	1.000	-283.210	283.216
hour_5	5.4784	144.319	0.038	0.970	-278.000	288.957
hour_6	5.5405	144.740	0.038	0.969	-278.765	289.846
hour_7	-33.2660	144.851	-0.230	0.818	-317.790	251.258
hour_8	-42.4918	146.927	-0.289	0.773	-331.094	246.110
hour_9	-108.5187	145.920	-0.744	0.457	-395.143	178.105
hour_10	-94.9789	154.007	-0.617	0.538	-397.488	207.530
hour_11	-216.6297	152.654	-1.419	0.156	-516.480	83.221
hour_12	-434.1047	151.321	-2.869	0.004	-731.338	-136.871
hour_13	-98.3096	150.726	-0.652	0.515	-394.374	197.755
hour_14	545.9132	152.090	3.589	0.000	247.170	844.657
hour_15	-298.3471	153.673	-1.941	0.053	-600.201	3.507
hour_16	-13.5677	150.166	-0.090	0.928	-308.533	281.397
hour_17	156.9199	150.532	1.042	0.298	-138.763	452.603
hour_18	-116.0955	148.735	-0.781	0.435	-408.250	176.059
hour_19	116.5992	153.563	0.759	0.448	-185.037	418.236
hour_20	58.9260	149.882	0.393	0.694	-235.480	353.332
hour_21	-36.8429	152.331	-0.242	0.809	-336.060	262.374
hour_22	-37.4291	148.978	-0.251	0.802	-330.061	255.202
hour_23	-36.3372	147.610	-0.246	0.806	-326.281	253.606

Omnibus:	1005.960	Durbin-Watson:	2.196
Prob(Omnibus):	0.000	Jarque-Bera (JB):	944728.186
Skew:	10.306	Prob(JB):	0.00
Kurtosis:	198.620	Cond. No.	1.08e+20

#GoPatriots

RMSE is: 105.96317320325763 MSE: 11228.19407530357586



OLS Regression Results

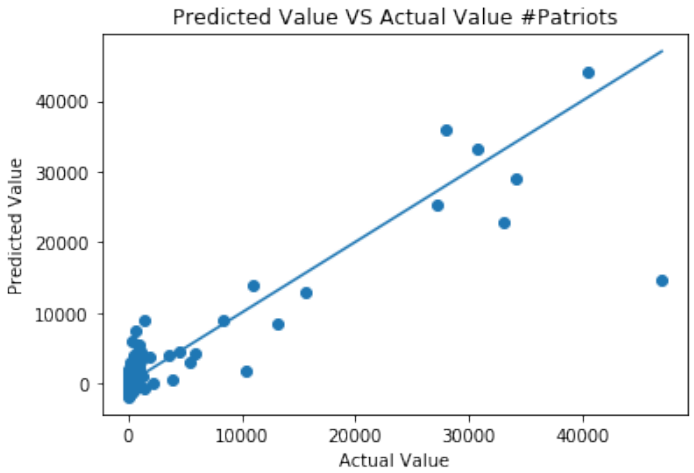
Dep. Variable:	tweet	R-squared:	0.846
Model:	OLS	Adj. R-squared:	0.837
Method:	Least Squares	F-statistic:	98.22
Date:	Thu, 19 Mar 2020	Prob (F-statistic):	8.13e-203
Time:	11:26:38	Log-Likelihood:	-3564.1
No. Observations:	586	AIC:	7192.
Df Residuals:	554	BIC:	7332.
Df Model:	31		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
tweet	-8.8131	0.681	-12.937	0.000	-10.151	-7.475
follower	-0.0012	0.000	-5.968	0.000	-0.002	-0.001
retweet	-1.2831	0.167	-7.685	0.000	-1.611	-0.955
Maxfollowers	0.0013	0.000	6.500	0.000	0.001	0.002
Hashtag	3.0322	0.345	8.792	0.000	2.355	3.710
reply	-8.8131	0.681	-12.937	0.000	-10.151	-7.475
favourite_count	-6.7058	1.271	-5.277	0.000	-9.202	-4.210
verified	-8.8131	0.681	-12.937	0.000	-10.151	-7.475
citation	7.4012	0.919	8.055	0.000	5.596	9.206
ranking	4.7752	0.417	11.454	0.000	3.956	5.594
hour_0	-6.0500	21.802	-0.277	0.782	-48.876	36.776
hour_1	-2.2157	21.808	-0.102	0.919	-45.053	40.622
hour_2	-7.4779	21.806	-0.343	0.732	-50.310	35.355
hour_3	-0.5418	21.808	-0.025	0.980	-43.378	42.294
hour_4	-9.0248	21.811	-0.414	0.679	-51.868	33.818
hour_5	-4.5223	21.811	-0.207	0.836	-47.364	38.320
hour_6	0.3826	21.867	0.017	0.986	-42.570	43.335
hour_7	-15.3860	21.834	-0.705	0.481	-58.273	27.501
hour_8	-16.0104	21.864	-0.732	0.464	-58.956	26.936
hour_9	-4.9237	21.904	-0.225	0.822	-47.948	38.101
hour_10	-7.1451	22.339	-0.320	0.749	-51.025	36.734
hour_11	-10.2376	22.391	-0.457	0.648	-54.219	33.744
hour_12	-6.7911	22.358	-0.304	0.761	-50.708	37.125
hour_13	8.2589	22.776	0.363	0.717	-36.479	52.997
hour_14	82.5229	22.488	3.670	0.000	38.351	126.694
hour_15	15.7035	23.069	0.681	0.496	-29.610	61.017
hour_16	-23.1339	22.701	-1.019	0.309	-67.724	21.457
hour_17	55.4779	22.715	2.442	0.015	10.860	100.096
hour_18	12.3649	22.887	0.540	0.589	-32.591	57.321
hour_19	-17.9576	22.703	-0.791	0.429	-62.553	26.637
hour_20	8.8055	22.567	0.390	0.697	-35.523	53.134
hour_21	20.4790	22.440	0.913	0.362	-23.599	64.557
hour_22	-2.0978	22.275	-0.094	0.925	-45.851	41.656
hour_23	3.4976	22.273	0.157	0.875	-40.253	47.248

Omnibus:	637.926	Durbin-Watson:	2.176
Prob(Omnibus):	0.000	Jarque-Bera (JB):	128635.849
Skew:	4.527	Prob(JB):	0.00
Kurtosis:	75.017	Cond. No.	1.96e+20

#Patriots

RMSE is: 1752.4304812464936 MSE: $3.07101259160181 \times 10^6$



OLS Regression Results

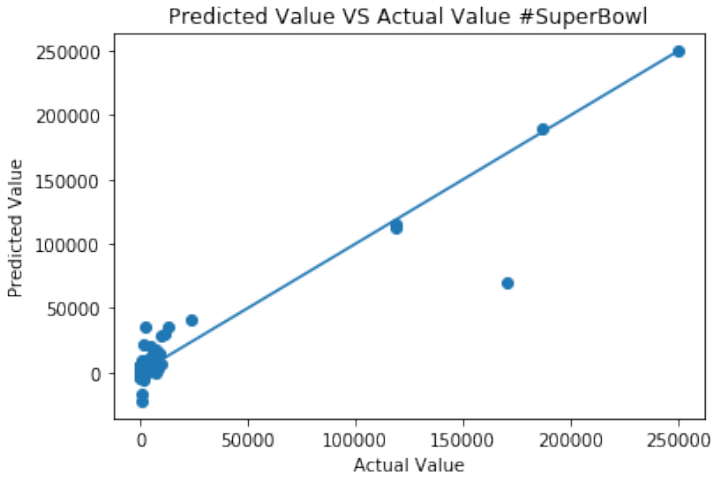
Dep. Variable:	tweet	R-squared:	0.803
Model:	OLS	Adj. R-squared:	0.792
Method:	Least Squares	F-statistic:	73.06
Date:	Thu, 19 Mar 2020	Prob (F-statistic):	9.55e-174
Time:	11:27:40	Log-Likelihood:	-5208.2
No. Observations:	586	AIC:	1.048e+04
Df Residuals:	554	BIC:	1.062e+04
Df Model:	31		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
tweet	-18.1945	1.629	-11.170	0.000	-21.394	-14.995
follower	0.0002	5.23e-05	2.932	0.004	5.06e-05	0.000
retweet	-0.4280	0.085	-5.041	0.000	-0.595	-0.261
Maxfollowers	-0.0003	9.49e-05	-3.358	0.001	-0.001	-0.000
Hashtag	1.1913	0.208	5.732	0.000	0.783	1.599
reply	-18.1945	1.629	-11.170	0.000	-21.394	-14.995
favourite_count	0.7209	0.311	2.315	0.021	0.109	1.333
verified	-18.1945	1.629	-11.170	0.000	-21.394	-14.995
citation	7.8153	0.568	13.748	0.000	6.699	8.932
ranking	11.2195	1.006	11.149	0.000	9.243	13.196
hour_0	-57.6006	361.445	-0.159	0.873	-767.572	652.370
hour_1	-134.6992	361.125	-0.373	0.709	-844.040	574.642
hour_2	-341.8080	362.346	-0.943	0.346	-1053.549	369.933
hour_3	-166.4987	361.366	-0.461	0.645	-876.313	543.316
hour_4	-295.9927	362.477	-0.817	0.415	-1007.990	416.004
hour_5	-257.4393	363.643	-0.708	0.479	-971.727	456.849
hour_6	-712.2856	367.550	-1.938	0.053	-1434.248	9.677
hour_7	-834.3261	370.547	-2.252	0.025	-1562.174	-106.478
hour_8	-849.4562	371.228	-2.288	0.022	-1578.642	-120.270
hour_9	-734.3840	373.901	-1.964	0.050	-1468.822	0.054
hour_10	319.0136	388.827	0.820	0.412	-444.742	1082.770
hour_11	-796.0660	378.843	-2.101	0.036	-1540.210	-51.922
hour_12	-1208.3636	380.849	-3.173	0.002	-1956.449	-460.279
hour_13	-645.7804	377.163	-1.712	0.087	-1386.624	95.063
hour_14	-581.4051	384.882	-1.511	0.131	-1337.412	174.602
hour_15	-843.2169	387.776	-2.174	0.030	-1604.907	-81.527
hour_16	-464.2590	378.928	-1.225	0.221	-1208.570	280.052
hour_17	-137.8268	376.069	-0.366	0.714	-876.522	600.868
hour_18	-152.0236	385.539	-0.394	0.694	-909.321	605.274
hour_19	-560.4756	379.394	-1.477	0.140	-1305.703	184.752
hour_20	-225.5024	379.702	-0.594	0.553	-971.335	520.330
hour_21	-581.3749	374.380	-1.553	0.121	-1316.753	154.003
hour_22	-518.1073	373.310	-1.388	0.166	-1251.383	215.168
hour_23	-247.6477	369.585	-0.670	0.503	-973.608	478.313

Omnibus:	1033.072	Durbin-Watson:	1.688
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1016323.910
Skew:	10.909	Prob(JB):	0.00
Kurtosis:	205.850	Cond. No.	7.59e+19

#SuperBowl

RMSE is: 5285.694572972141 MSE: 81621.389025733998



OLS Regression Results

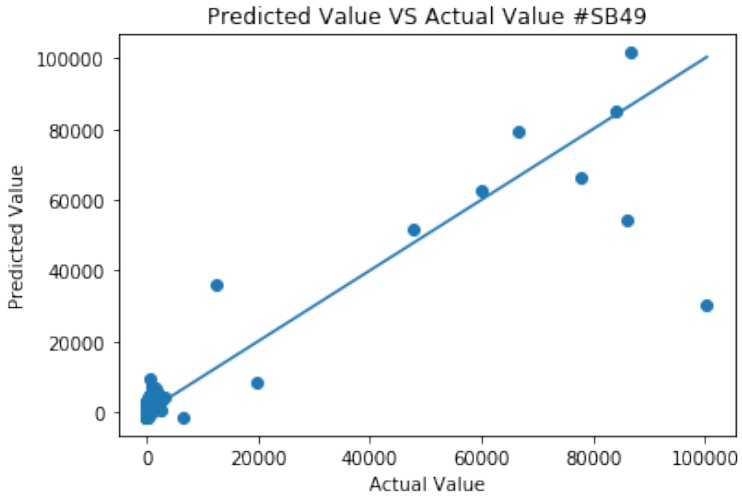
Dep. Variable:	tweet	R-squared:	0.894
Model:	OLS	Adj. R-squared:	0.888
Method:	Least Squares	F-statistic:	151.1
Date:	Thu, 19 Mar 2020	Prob (F-statistic):	1.21e-247
Time:	11:30:28	Log-Likelihood:	-5852.9
No. Observations:	586	AIC:	1.177e+04
Df Residuals:	554	BIC:	1.191e+04
Df Model:	31		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
tweet	-13.1188	1.340	-9.790	0.000	-15.751	-10.487
follower	-9.442e-05	2.16e-05	-4.370	0.000	-0.000	-5.2e-05
retweet	-0.2145	0.070	-3.052	0.002	-0.353	-0.076
Maxfollowers	0.0001	0.000	0.976	0.330	-0.000	0.000
Hashtag	3.7619	0.285	13.183	0.000	3.201	4.322
reply	-13.1194	1.340	-9.790	0.000	-15.752	-10.487
favourite_count	-1.4217	0.224	-6.333	0.000	-1.863	-0.981
verified	-13.1194	1.340	-9.790	0.000	-15.752	-10.487
citation	-1.8196	1.111	-1.638	0.102	-4.002	0.363
ranking	7.5396	0.839	8.985	0.000	5.891	9.188
hour_0	-135.9952	1094.433	-0.124	0.901	-2285.742	2013.751
hour_1	-202.2275	1090.182	-0.185	0.853	-2343.623	1939.168
hour_2	-215.9848	1088.376	-0.198	0.843	-2353.832	1921.863
hour_3	-236.9805	1095.219	-0.216	0.829	-2388.270	1914.309
hour_4	-442.2678	1087.426	-0.407	0.684	-2578.250	1693.714
hour_5	-302.1135	1092.998	-0.276	0.782	-2449.040	1844.813
hour_6	-525.2418	1096.758	-0.479	0.632	-2679.554	1629.070
hour_7	-983.4333	1105.867	-0.889	0.374	-3155.638	1188.771
hour_8	-572.6535	1106.762	-0.517	0.605	-2746.617	1601.310
hour_9	-1272.3270	1115.318	-1.141	0.254	-3463.096	918.442
hour_10	-1643.3521	1136.288	-1.446	0.149	-3875.311	588.607
hour_11	-1387.8126	1136.372	-1.221	0.223	-3619.937	844.312
hour_12	-1606.2934	1133.125	-1.418	0.157	-3832.040	619.453
hour_13	-526.4602	1133.871	-0.464	0.643	-2753.672	1700.752
hour_14	4419.8305	1136.013	3.891	0.000	2188.410	6651.251
hour_15	-1756.8925	1138.924	-1.543	0.124	-3994.030	480.245
hour_16	-671.1969	1143.819	-0.587	0.558	-2917.949	1575.555
hour_17	-21.2607	1171.822	-0.018	0.986	-2323.018	2280.496
hour_18	168.7871	1144.454	0.147	0.883	-2079.213	2416.787
hour_19	-1777.8091	1145.583	-1.552	0.121	-4028.027	472.409
hour_20	-284.3918	1123.991	-0.253	0.800	-2492.198	1923.414
hour_21	-293.7008	1119.930	-0.262	0.793	-2493.529	1906.127
hour_22	-767.8924	1114.304	-0.689	0.491	-2956.670	1420.885
hour_23	-472.3629	1116.184	-0.423	0.672	-2664.833	1720.108

Omnibus:	1054.270	Durbin-Watson:	1.999
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1341397.513
Skew:	11.276	Prob(JB):	0.00
Kurtosis:	236.301	Cond. No.	3.48e+20

#SB49

RMSE is: 3689.5291092083216 MSE: $1.36126250476955 \times 10^7$



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  tweet   R-squared:                       0.835
Model:                            OLS   Adj. R-squared:                  0.825
Method:                 Least Squares   F-statistic:                     90.16
Date:                Thu, 19 Mar 2020   Prob (F-statistic):          3.04e-194
Time:                        11:32:23   Log-Likelihood:                -5644.5
No. Observations:                 586   AIC:                         1.135e+04
Df Residuals:                     554   BIC:                         1.149e+04
Df Model:                          31
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
tweet        -14.3152      2.978     -4.808      0.000     -20.164      -8.466
follower      1.22e-05   1.37e-05      0.893      0.372   -1.46e-05    3.91e-05
retweet         0.4244      0.118      3.606      0.000       0.193       0.655
Maxfollowers -2.137e-05   4.83e-05     -0.443      0.658      -0.000    7.34e-05
Hashtag         0.4477      0.262      1.708      0.088      -0.067       0.963
reply        -14.3157      2.978     -4.808      0.000     -20.164      -8.467
favourite_count  -0.2036      0.102     -1.990      0.047      -0.405      -0.003
verified     -14.3157      2.978     -4.808      0.000     -20.164      -8.467
citation        6.4016      1.041      6.149      0.000       4.357       8.447
ranking         8.8419      1.865      4.740      0.000       5.178      12.506
hour_0        -18.7124    760.385     -0.025      0.980   -1512.302    1474.878
hour_1          7.1097    759.358      0.009      0.993   -1484.464    1498.684
hour_2        -48.7910    759.250     -0.064      0.949   -1540.153    1442.571
hour_3        -52.4291    759.214     -0.069      0.945   -1543.718    1438.860
hour_4        -61.4057    759.868     -0.081      0.936   -1553.981    1431.169
hour_5       -235.6969    761.521     -0.310      0.757   -1731.518    1260.125
hour_6       -224.0208    761.808     -0.294      0.769   -1720.406    1272.364
hour_7       -722.4531    769.062     -0.939      0.348   -2233.087     788.180
hour_8       -567.7173    771.057     -0.736      0.462   -2082.271     946.836
hour_9       -158.6331    773.836     -0.205      0.838   -1678.644    1361.378
hour_10       2480.8486    793.180      3.128      0.002     922.841    4038.856
hour_11      -1280.9149    802.830     -1.595      0.111   -2857.878     296.048
hour_12      -1260.2815    787.573     -1.600      0.110   -2807.276     286.713
hour_13       -788.9334    802.680     -0.983      0.326   -2365.602     787.735
hour_14        751.5397    787.878      0.954      0.341    -796.055    2299.134
hour_15       -507.3473    795.107     -0.638      0.524   -2069.140    1054.445
hour_16         62.5554    788.588      0.079      0.937   -1486.433    1611.544
hour_17       -838.4264    797.845     -1.051      0.294   -2405.597     728.744
hour_18      -1387.9009    786.499     -1.765      0.078   -2932.785     156.983
hour_19        286.5099    789.707      0.363      0.717   -1264.677    1837.696
hour_20         26.2403    792.314      0.033      0.974   -1530.067    1582.548
hour_21       -307.4648    780.197     -0.394      0.694   -1839.971    1225.041
hour_22        -25.1373    781.031     -0.032      0.974   -1559.282    1509.007
hour_23        -50.2639    776.217     -0.065      0.948   -1574.953    1474.425
==============================================================================
Omnibus:                     1092.349   Durbin-Watson:                   2.007
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1342348.644
Skew:                          12.245   Prob(JB):                         0.00
Kurtosis:                     236.189   Cond. No.                     3.83e+20
==============================================================================
```

Hashtag	Average number of tweets per hour	RMSE (before adding new features)	RMSE (after adding new features)	R^2 (before adding new features)	R^2 (after adding new features)
#NFL	396.97103918228277	503.96521944684315	394.5126755879352	0.596	0.753
#SuperBowl	2067.824531516184	7073.15695692096	5285.694572972141	0.809	0.894
#GoHawks	288.11243611584325	837.4844185446983	700.8128346896752	0.510	0.657
#GoPatriots	40.052810902896084	159.86345136919155	105.96317320325763	0.650	0.846
#Patriots	750.6320272572402	2225.567247490388	1752.4304812464936	0.683	0.803
#SB49	1266.8637137989779	3932.3293340677287	3689.5291092083216	0.812	0.835

R-squared provides the relative measure of the percentage of the dependent variable variance that the model explains. Higher R-squared values indicate that the data points are closer to the fitted values. When we add a new feature to our model, the value of its estimated coefficient can either be zero, in which case R-square stays unchanged, or take a nonzero value because it improves the quality of the fit.

From the results, we showed that the R-square value of the linear regression model increases with the new features compared to the model with only five features. In this way, the prediction

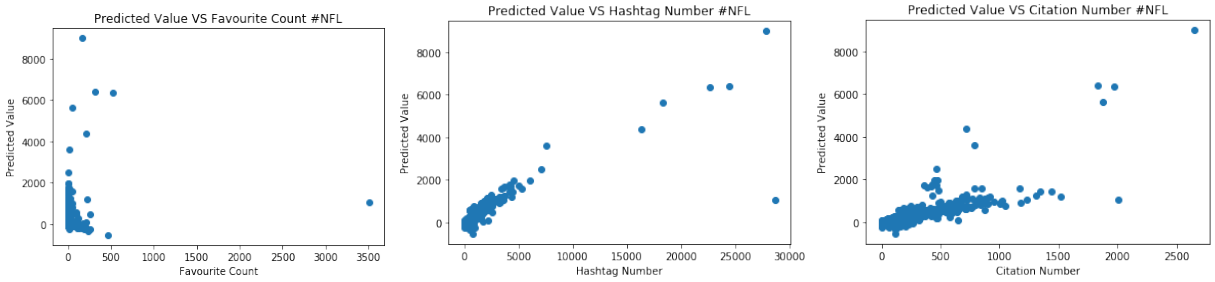
performance improves with the addition of the features. Therefore, the new features are a meaningful addition.

QUESTION 5: For each of the top 3 features (i.e. with the smallest p-values) in your measurements, draw a scatter plot of predictant (number of tweets for next hour) versus value of that feature, using all the samples you have extracted, and analyze it. Do the regression coefficients agree with the trends in the plots? If not, why?

#NFL

The top three features are Favorite count, Hashtag Number and Citation Number.

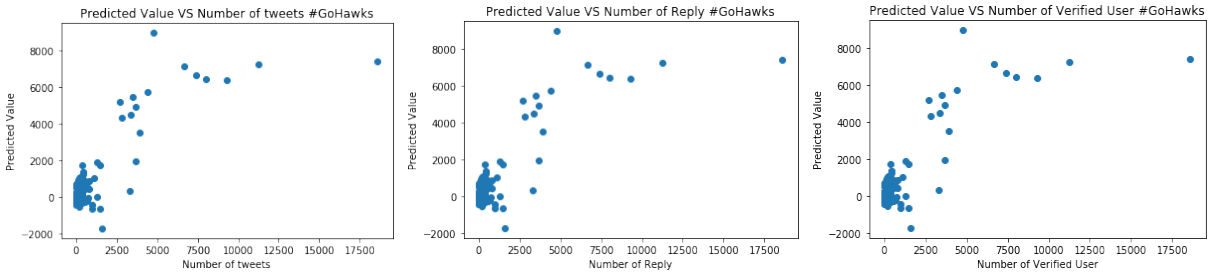
The scatter plots of prediction:



#GoHawks

The top three features are Number of tweets, Number of Reply and Number of Verified User.

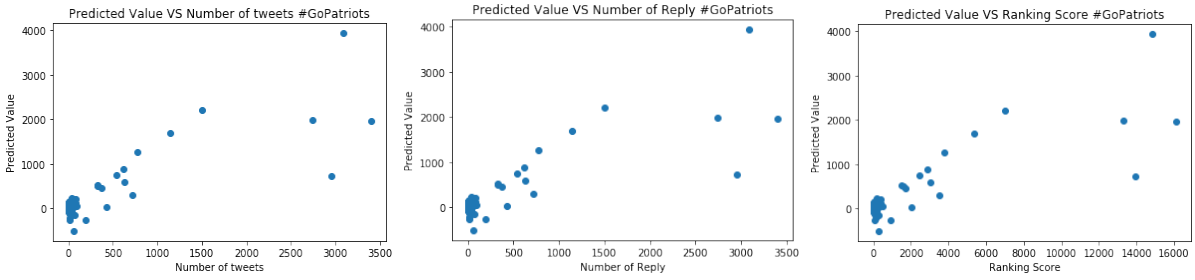
The scatter plots of prediction:



#GoPatriots

The top three features we choose are Number of tweets, Number of Reply and Ranking Score.

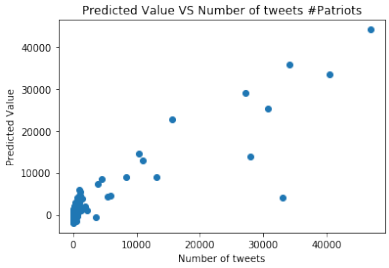
The scatter plots of prediction:

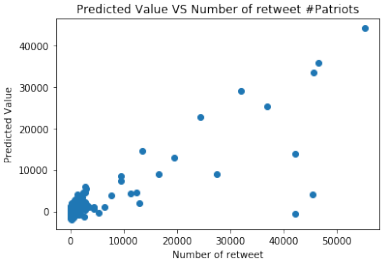


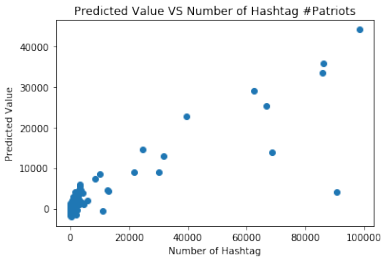
#Patriots

The top three features we choose are Number of tweets, Number of retweet and Number of Hashtag.

The scatter plots of prediction:



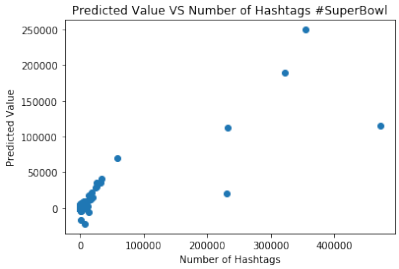


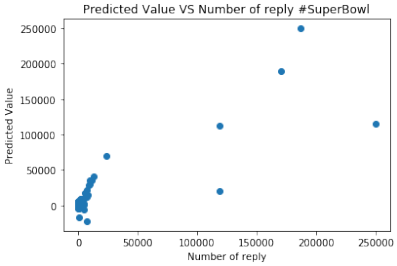


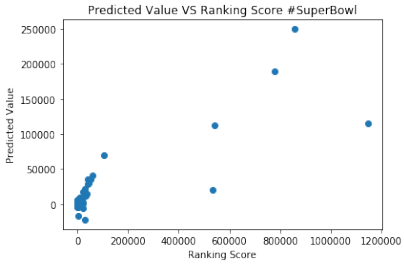
#SuperBowl:

The top three features we choose are Number of Hashtag, Number of Reply and Ranking Score.

The scatter plots of prediction:



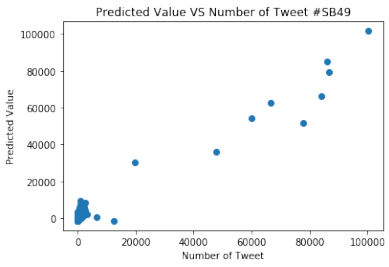


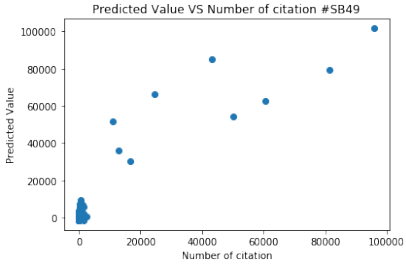


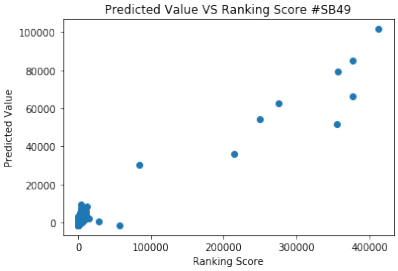
#SB49:

The top three features we choose are Number of Tweet, Number of Citation and Ranking Score.

The scatter plots of prediction:







From the prediction plots for the different hashtags, we showed that when the predicted number of tweets is high, the points become sparse. For different hashtags, different features have different significance. For #GoHawks, #GoPatiots, #NFL and #Patriots, there is a clear linear relationship between the number of tweets for next hour and the value of top features. However, for #SB49 and #SuperBowl, the top 3 features don't present a linear relationship. This is because the number of tweets is very large, and the sparsity of the data influences the linear relationship between top features and target values.

The regression coefficients agree with the trends in the plots. For example, we have the same regression coefficients for the top 3 features in #GoHawks, and we also got the same regression plots for the features. Overall, we can observe a relatively linear relationship between our top features and our target value.

4. Piecewise linear regression

We created different regression models for different periods of time as the Super Bowl's date and time is known. We divided the time duration into three pieces. First time period is when the hashtags haven't become very active, second time period is their active period, and third time period is after they pass their high-activity time. The time periods are defined as below.

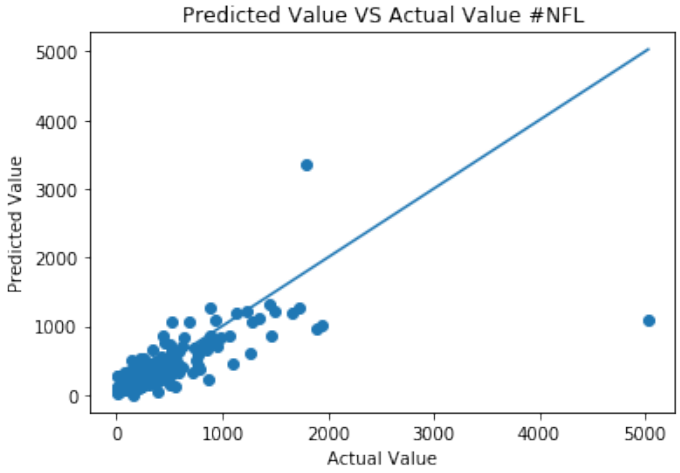
1. Before Feb. 1, 8:00 a.m.: 1-hour window
2. Between Feb. 1, 8:00 a.m. and 8:00 p.m.: 5-minute window
3. After Feb. 1, 8:00 p.m.: 1-hour window

QUESTION 6: For each hashtag, train 3 regression models, one for each of these time periods (the times are all in PST). Report the MSE and R-squared score for each case.

#NFL:

Time period 1:

RMSE is: 248.8577827345286 MSE: 61930.19602754584096266



OLS Regression Results

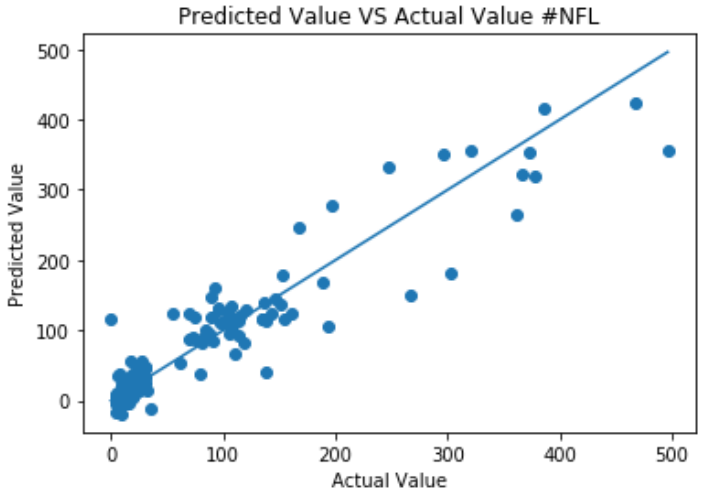
Dep. Variable:	tweet	R-squared:	0.541
Model:	OLS	Adj. R-squared:	0.511
Method:	Least Squares	F-statistic:	17.96
Date:	Thu, 19 Mar 2020	Prob (F-statistic):	9.91e-54
Time:	11:33:50	Log-Likelihood:	-3044.8
No. Observations:	439	AIC:	6146.
Df Residuals:	411	BIC:	6260.
Df Model:	27		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
tweet	0.4769	0.102	4.665	0.000	0.276	0.678
follower	6.055e-05	1.86e-05	3.261	0.001	2.41e-05	9.7e-05
retweet	0.0132	0.047	0.278	0.781	-0.080	0.106
Maxfollowers	-5.847e-05	2.34e-05	-2.494	0.013	-0.000	-1.24e-05
hour_0	41.0611	59.319	0.692	0.489	-75.545	157.667
hour_1	44.5297	59.442	0.749	0.454	-72.319	161.379
hour_2	62.5388	59.282	1.055	0.292	-53.994	179.072
hour_3	63.6052	59.286	1.073	0.284	-52.936	180.146
hour_4	48.4058	59.312	0.816	0.415	-68.187	164.998
hour_5	128.3469	59.345	2.163	0.031	11.689	245.005
hour_6	164.6373	59.750	2.755	0.006	47.184	282.091
hour_7	127.3805	61.538	2.070	0.039	6.412	248.349
hour_8	85.3546	62.006	1.377	0.169	-36.533	207.242
hour_9	110.9681	61.859	1.794	0.074	-10.631	232.568
hour_10	163.6164	61.745	2.650	0.008	42.241	284.992
hour_11	174.6689	62.321	2.803	0.005	52.161	297.177
hour_12	33.5890	62.630	0.536	0.592	-89.527	156.705
hour_13	106.9199	62.476	1.711	0.088	-15.892	229.732
hour_14	273.4295	62.193	4.396	0.000	151.174	395.685
hour_15	-13.8902	63.839	-0.218	0.828	-139.381	111.601
hour_16	118.2011	62.705	1.885	0.060	-5.061	241.463
hour_17	166.0287	63.858	2.600	0.010	40.500	291.557
hour_18	18.3751	62.773	0.293	0.770	-105.021	141.771
hour_19	43.1150	61.596	0.700	0.484	-77.967	164.197
hour_20	44.3381	61.993	0.715	0.475	-77.526	166.202
hour_21	50.4713	61.437	0.822	0.412	-70.299	171.242
hour_22	25.8066	61.061	0.423	0.673	-94.225	145.838
hour_23	55.1748	61.122	0.903	0.367	-64.976	175.326

Omnibus:	714.856	Durbin-Watson:	2.276
Prob(Omnibus):	0.000	Jarque-Bera (JB):	402633.532
Skew:	8.973	Prob(JB):	0.00
Kurtosis:	150.275	Cond. No.	1.92e+07

Time period 2:

RMSE is: 34.34881502837251 MSE: 1179.84109385334919



OLS Regression Results

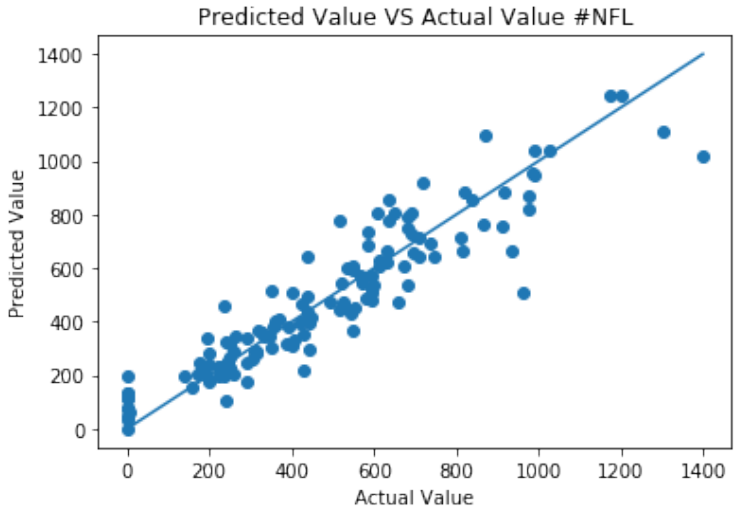
Dep. Variable:	tweet	R-squared:	0.880
Model:	OLS	Adj. R-squared:	0.853
Method:	Least Squares	F-statistic:	33.02
Date:	Sun, 22 Mar 2020	Prob (F-statistic):	3.41e-42
Time:	14:22:50	Log-Likelihood:	-713.59
No. Observations:	144	AIC:	1481.
Df Residuals:	117	BIC:	1561.
Df Model:	26		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
tweet	0.5818	0.129	4.516	0.000	0.327	0.837
follower	7.584e-05	2.91e-05	2.605	0.010	1.82e-05	0.000
retweet	-0.0913	0.031	-2.988	0.003	-0.152	-0.031
Maxfollowers	-0.0001	3.24e-05	-3.251	0.002	-0.000	-4.12e-05
hour_8	-11.0537	10.754	-1.028	0.306	-32.351	10.243
hour_9	-8.9864	11.167	-0.805	0.423	-31.103	13.130
hour_10	-13.3712	10.906	-1.226	0.223	-34.971	8.228
hour_11	-13.6445	11.018	-1.238	0.218	-35.466	8.177
hour_12	-7.5130	10.883	-0.690	0.491	-29.066	14.040
hour_13	-7.0777	10.844	-0.653	0.515	-28.553	14.398
hour_14	0.5209	10.767	0.048	0.961	-20.803	21.845
hour_15	32.5068	13.916	2.336	0.021	4.946	60.067
hour_16	49.3950	12.523	3.944	0.000	24.594	74.196
hour_17	69.4268	17.869	3.885	0.000	34.038	104.816
hour_18	31.6474	11.661	2.714	0.008	8.554	54.741
hour_19	130.5830	32.130	4.064	0.000	66.952	194.214
hour_20	-5.742e-09	3.04e-09	-1.888	0.061	-1.18e-08	2.81e-10
minute_0	51.6680	11.995	4.307	0.000	27.913	75.423
minute_1	-5.0319	12.142	-0.414	0.679	-29.079	19.015
minute_2	15.5996	11.001	1.418	0.159	-6.187	37.386
minute_3	29.4359	11.066	2.660	0.009	7.521	51.351
minute_4	37.4628	11.583	3.234	0.002	14.523	60.403
minute_5	19.4940	12.116	1.609	0.110	-4.502	43.490
minute_6	14.8294	11.507	1.289	0.200	-7.960	37.619
minute_7	15.2921	11.582	1.320	0.189	-7.646	38.230
minute_8	13.2690	11.363	1.168	0.245	-9.236	35.773
minute_9	14.4929	11.110	1.305	0.195	-7.510	36.495
minute_10	13.2628	11.203	1.184	0.239	-8.924	35.449
minute_11	32.6588	11.580	2.820	0.006	9.726	55.592

Omnibus:	33.374	Durbin-Watson:	1.566
Prob(Omnibus):	0.000	Jarque-Bera (JB):	119.234
Skew:	0.778	Prob(JB):	1.28e-26
Kurtosis:	7.178	Cond. No.	1.15e+18

Time period 3:

RMSE is: 109.19394002321302 MSE: 11923.31653779304222



OLS Regression Results

Dep. Variable:	tweet	R-squared:	0.861
Model:	OLS	Adj. R-squared:	0.826
Method:	Least Squares	F-statistic:	24.31
Date:	Thu, 19 Mar 2020	Prob (F-statistic):	5.67e-34
Time:	11:33:50	Log-Likelihood:	-819.02
No. Observations:	134	AIC:	1694.
Df Residuals:	106	BIC:	1775.
Df Model:	27		
Covariance Type:	nonrobust		

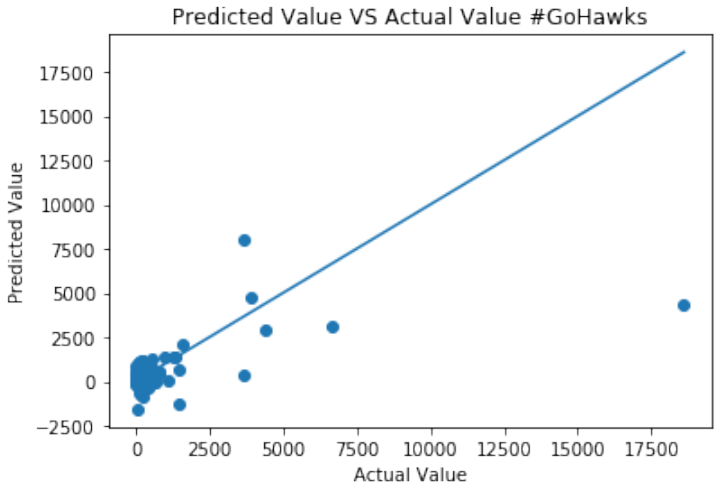
	coef	std err	t	P>|t|	[0.025	0.975]
tweet	0.8436	0.116	7.284	0.000	0.614	1.073
follower	1.112e-05	2.82e-05	0.394	0.695	-4.49e-05	6.71e-05
retweet	-0.0094	0.045	-0.210	0.834	-0.099	0.080
Maxfollowers	-1.396e-05	3.92e-05	-0.356	0.722	-9.16e-05	6.37e-05
hour_0	0.7318	53.116	0.014	0.989	-104.576	106.040
hour_1	46.1674	51.539	0.896	0.372	-56.013	148.348
hour_2	36.4320	52.418	0.695	0.489	-67.491	140.355
hour_3	75.1684	52.142	1.442	0.152	-28.209	178.546
hour_4	79.0007	53.448	1.478	0.142	-26.964	184.966
hour_5	135.3036	53.396	2.534	0.013	29.442	241.165
hour_6	110.2483	55.342	1.992	0.049	0.528	219.968
hour_7	134.1170	56.566	2.371	0.020	21.969	246.265
hour_8	116.5772	57.913	2.013	0.047	1.760	231.395
hour_9	64.5937	60.058	1.076	0.285	-54.478	183.665
hour_10	258.5772	67.439	3.834	0.000	124.873	392.281
hour_11	185.3839	71.992	2.575	0.011	42.653	328.115
hour_12	32.3221	76.532	0.422	0.674	-119.409	184.053
hour_13	99.5261	73.412	1.356	0.178	-46.021	245.073
hour_14	107.7603	70.896	1.520	0.131	-32.798	248.319
hour_15	-22.6289	71.550	-0.316	0.752	-164.483	119.225
hour_16	22.8896	69.618	0.329	0.743	-115.134	160.913
hour_17	98.4369	65.184	1.510	0.134	-30.797	227.671
hour_18	-88.6892	67.772	-1.309	0.193	-223.054	45.675
hour_19	125.3381	60.687	2.065	0.041	5.019	245.657
hour_20	20.6729	60.923	0.339	0.735	-100.114	141.459
hour_21	20.1866	58.586	0.345	0.731	-95.966	136.340
hour_22	-73.5016	57.464	-1.279	0.204	-187.429	40.426
hour_23	-13.1666	52.473	-0.251	0.802	-117.200	90.867

Omnibus:	20.936	Durbin-Watson:	2.082
Prob(Omnibus):	0.000	Jarque-Bera (JB):	42.153
Skew:	0.662	Prob(JB):	7.02e-10
Kurtosis:	5.407	Cond. No.	4.52e+07

#GoHawks:

Time period 1:

RMSE is: 816.6104179136739 MSE: 666852.57464514513879



OLS Regression Results

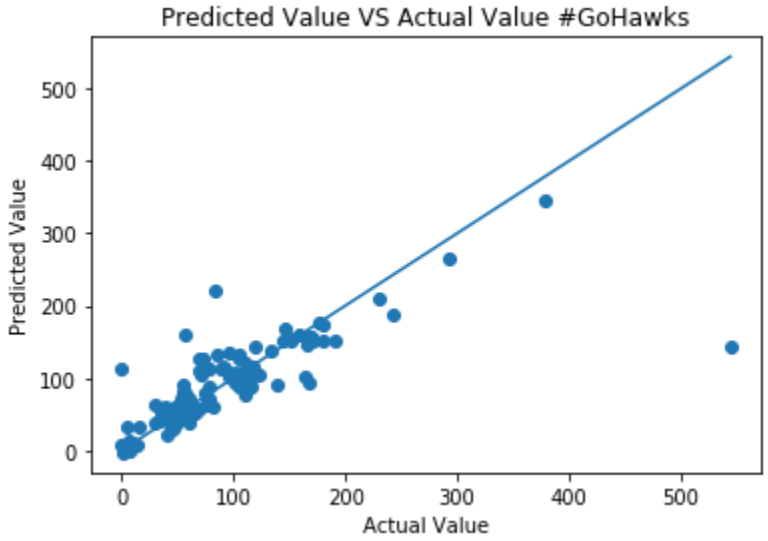
Dep. Variable:	tweet	R-squared:	0.353
Model:	OLS	Adj. R-squared:	0.311
Method:	Least Squares	F-statistic:	8.311
Date:	Thu, 19 Mar 2020	Prob (F-statistic):	2.36e-25
Time:	11:34:20	Log-Likelihood:	-3566.5
No. Observations:	439	AIC:	7189.
Df Residuals:	411	BIC:	7303.
Df Model:	27		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
tweet	2.2386	0.295	7.592	0.000	1.659	2.818
follower	-0.0006	0.000	-3.478	0.001	-0.001	-0.000
retweet	-0.4229	0.115	-3.675	0.000	-0.649	-0.197
Maxfollowers	0.0005	0.000	2.210	0.028	5.09e-05	0.001
hour_0	-16.9656	193.939	-0.087	0.930	-398.201	364.270
hour_1	-3.7885	193.645	-0.020	0.984	-384.447	376.870
hour_2	-5.2445	193.709	-0.027	0.978	-386.029	375.540
hour_3	0.3497	193.669	0.002	0.999	-380.356	381.055
hour_4	16.7627	193.643	0.087	0.931	-363.891	397.416
hour_5	44.0667	193.692	0.228	0.820	-336.683	424.817
hour_6	137.2561	194.416	0.706	0.481	-244.917	519.429
hour_7	174.2155	199.294	0.874	0.383	-217.547	565.978
hour_8	353.4424	203.809	1.734	0.084	-47.195	754.080
hour_9	139.6370	199.544	0.700	0.484	-252.617	531.891
hour_10	526.7610	203.387	2.590	0.010	126.953	926.569
hour_11	403.4644	200.255	2.015	0.045	9.813	797.116
hour_12	38.4597	201.363	0.191	0.849	-357.370	434.290
hour_13	241.8179	200.598	1.205	0.229	-152.507	636.143
hour_14	923.3628	200.455	4.606	0.000	529.318	1317.408
hour_15	-210.0369	205.402	-1.023	0.307	-613.806	193.732
hour_16	73.8700	201.803	0.366	0.715	-322.825	470.565
hour_17	405.4867	205.455	1.974	0.049	1.613	809.361
hour_18	55.6958	200.974	0.277	0.782	-339.369	450.760
hour_19	76.9222	206.698	0.372	0.710	-329.395	483.239
hour_20	65.3365	199.701	0.327	0.744	-327.226	457.900
hour_21	27.0384	202.632	0.133	0.894	-371.286	425.363
hour_22	45.3432	201.485	0.225	0.822	-350.726	441.412
hour_23	-39.0133	199.457	-0.196	0.845	-431.097	353.070

Omnibus:	836.304	Durbin-Watson:	2.228
Prob(Omnibus):	0.000	Jarque-Bera (JB):	856591.609
Skew:	12.310	Prob(JB):	0.00
Kurtosis:	217.996	Cond. No.	1.21e+07

Time period 2:

RMSE is: 42.31315800563716 MSE: 1790.40334041001608



OLS Regression Results

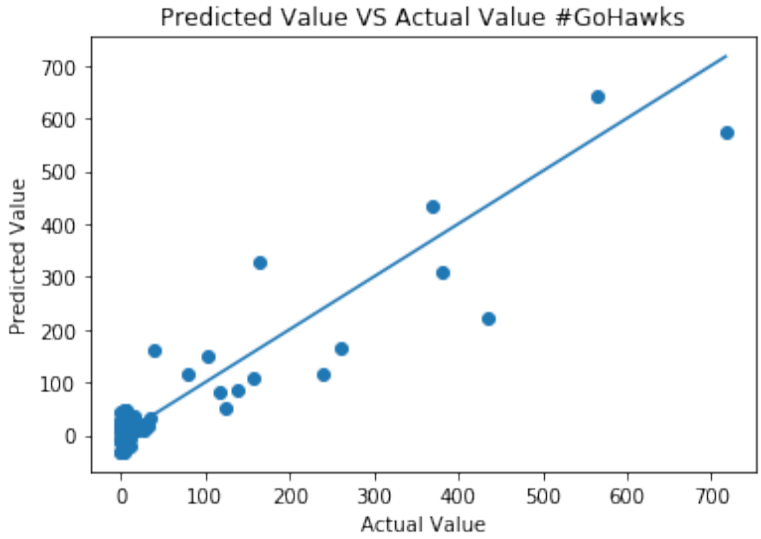
Dep. Variable:	tweet	R-squared:	0.648
Model:	OLS	Adj. R-squared:	0.570
Method:	Least Squares	F-statistic:	8.286
Date:	Sun, 22 Mar 2020	Prob (F-statistic):	2.11e-16
Time:	14:23:10	Log-Likelihood:	-743.62
No. Observations:	144	AIC:	1541.
Df Residuals:	117	BIC:	1621.
Df Model:	26		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
tweet	0.4491	0.186	2.410	0.018	0.080	0.818
follower	-3.458e-05	0.000	-0.310	0.757	-0.000	0.000
retweet	0.0891	0.085	1.044	0.298	-0.080	0.258
Maxfollowers	5.573e-05	0.000	0.380	0.705	-0.000	0.000
hour_8	-13.9744	13.737	-1.017	0.311	-41.179	13.230
hour_9	-13.6783	13.814	-0.990	0.324	-41.036	13.680
hour_10	1.1528	13.187	0.087	0.930	-24.963	27.269
hour_11	1.4231	13.491	0.105	0.916	-25.295	28.141
hour_12	3.1209	13.943	0.224	0.823	-24.493	30.735
hour_13	6.6293	14.432	0.459	0.647	-21.953	35.212
hour_14	19.4473	13.405	1.451	0.150	-7.100	45.995
hour_15	53.0603	20.175	2.630	0.010	13.105	93.016
hour_16	57.3405	14.052	4.080	0.000	29.511	85.170
hour_17	42.9132	23.804	1.803	0.074	-4.230	90.056
hour_18	27.7619	15.519	1.789	0.076	-2.972	58.496
hour_19	13.4032	14.735	0.910	0.365	-15.780	42.586
hour_20	-1.653e-09	8.17e-10	-2.024	0.045	-3.27e-09	-3.53e-11
minute_0	12.8380	15.536	0.826	0.410	-17.930	43.606
minute_1	14.0530	14.261	0.985	0.326	-14.189	42.295
minute_2	10.1839	14.388	0.708	0.480	-18.312	38.680
minute_3	11.8175	13.901	0.850	0.397	-15.712	39.347
minute_4	15.6179	13.691	1.141	0.256	-11.497	42.733
minute_5	13.6508	13.956	0.978	0.330	-13.988	41.289
minute_6	17.1942	13.906	1.237	0.219	-10.345	44.733
minute_7	15.1622	14.023	1.081	0.282	-12.610	42.934
minute_8	8.6316	14.106	0.612	0.542	-19.305	36.569
minute_9	18.1186	14.237	1.273	0.206	-10.078	46.315
minute_10	18.4378	13.950	1.322	0.189	-9.189	46.065
minute_11	42.8941	13.760	3.117	0.002	15.643	70.146

Omnibus:	202.880	Durbin-Watson:	1.982
Prob(Omnibus):	0.000	Jarque-Bera (JB):	18115.472
Skew:	5.391	Prob(JB):	0.00
Kurtosis:	56.880	Cond. No.	8.66e+18

Time period 3:

RMSE is: 38.54608451302766 MSE: 1485.80063128547081



OLS Regression Results

Dep. Variable:	tweet	R-squared:	0.858
Model:	OLS	Adj. R-squared:	0.822
Method:	Least Squares	F-statistic:	23.75
Date:	Thu, 19 Mar 2020	Prob (F-statistic):	1.56e-33
Time:	11:34:20	Log-Likelihood:	-679.49
No. Observations:	134	AIC:	1415.
Df Residuals:	106	BIC:	1496.
Df Model:	27		
Covariance Type:	nonrobust		

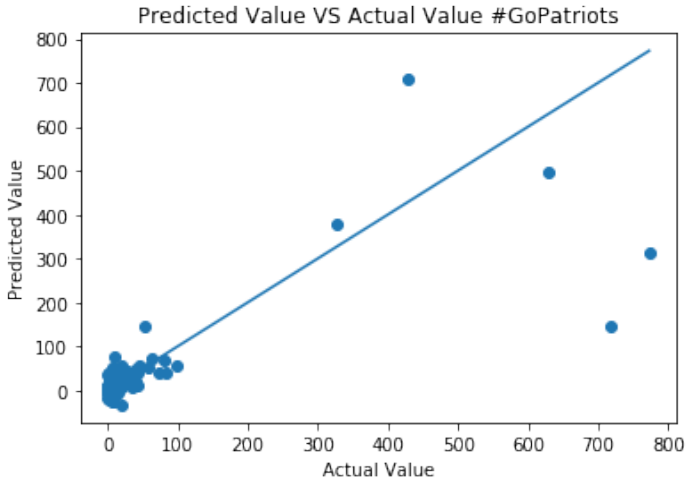
	coef	std err	t	P>|t|	[0.025	0.975]
tweet	0.1750	0.119	1.468	0.145	-0.061	0.411
follower	0.0009	0.000	4.921	0.000	0.001	0.001
retweet	-0.0306	0.010	-3.075	0.003	-0.050	-0.011
Maxfollowers	-0.0009	0.000	-4.603	0.000	-0.001	-0.001
hour_0	8.0104	17.800	0.450	0.654	-27.280	43.301
hour_1	-5.7718	17.783	-0.325	0.746	-41.029	29.486
hour_2	-9.3761	18.064	-0.519	0.605	-45.190	26.438
hour_3	15.0039	17.771	0.844	0.400	-20.229	50.236
hour_4	12.2460	17.740	0.690	0.492	-22.926	47.418
hour_5	10.6516	17.737	0.601	0.549	-24.514	45.817
hour_6	28.6657	17.762	1.614	0.110	-6.550	63.881
hour_7	20.8484	17.878	1.166	0.246	-14.597	56.293
hour_8	45.0436	17.980	2.505	0.014	9.396	80.691
hour_9	17.7479	18.225	0.974	0.332	-18.385	53.881
hour_10	-26.3615	21.368	-1.234	0.220	-68.725	16.002
hour_11	5.7045	19.402	0.294	0.769	-32.761	44.170
hour_12	4.1158	19.432	0.212	0.833	-34.409	42.641
hour_13	4.4738	19.403	0.231	0.818	-33.994	42.942
hour_14	8.9301	19.401	0.460	0.646	-29.534	47.394
hour_15	4.9732	19.414	0.256	0.798	-33.517	43.463
hour_16	5.1405	19.408	0.265	0.792	-33.337	43.618
hour_17	5.2330	19.397	0.270	0.788	-33.224	43.690
hour_18	3.4351	19.463	0.176	0.860	-35.153	42.023
hour_19	4.4469	19.388	0.229	0.819	-33.992	42.885
hour_20	-13.0475	18.689	-0.698	0.487	-50.101	24.006
hour_21	32.5049	18.991	1.712	0.090	-5.147	70.157
hour_22	-12.8890	18.301	-0.704	0.483	-49.173	23.395
hour_23	-31.7929	18.591	-1.710	0.090	-68.650	5.065

Omnibus:	59.570	Durbin-Watson:	1.577
Prob(Omnibus):	0.000	Jarque-Bera (JB):	628.524
Skew:	1.169	Prob(JB):	3.29e-137
Kurtosis:	13.349	Cond. No.	1.73e+06

#GoPatriots

Time period 1:

RMSE is: 40.45812675770317 MSE: 1636.86002074237721



OLS Regression Results

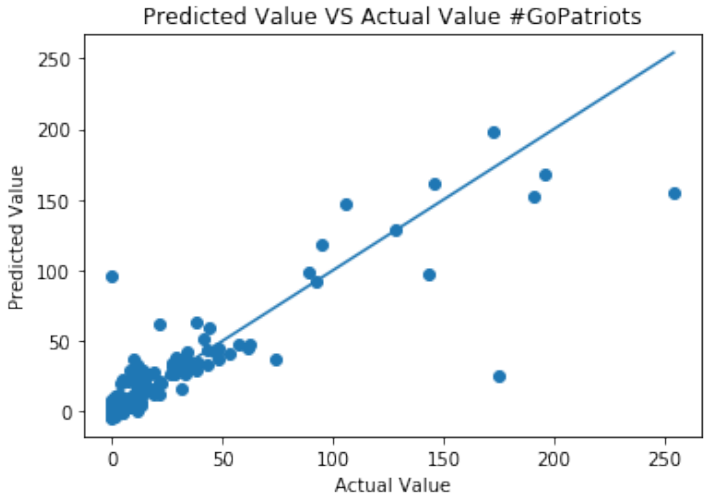
Dep. Variable:	tweet	R-squared:	0.600
Model:	OLS	Adj. R-squared:	0.574
Method:	Least Squares	F-statistic:	22.83
Date:	Thu, 19 Mar 2020	Prob (F-statistic):	2.15e-65
Time:	11:34:29	Log-Likelihood:	-2247.3
No. Observations:	439	AIC:	4551.
Df Residuals:	411	BIC:	4665.
Df Model:	27		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
tweet	1.8682	0.187	9.979	0.000	1.500	2.236
follower	-0.0010	0.000	-6.043	0.000	-0.001	-0.001
retweet	-0.1265	0.118	-1.068	0.286	-0.359	0.106
Maxfollowers	0.0010	0.000	6.138	0.000	0.001	0.001
hour_0	-1.2611	9.595	-0.131	0.895	-20.123	17.601
hour_1	0.0417	9.594	0.004	0.997	-18.818	18.901
hour_2	-0.4147	9.602	-0.043	0.966	-19.290	18.460
hour_3	2.3890	9.600	0.249	0.804	-16.481	21.259
hour_4	-1.1190	9.604	-0.117	0.907	-19.998	17.760
hour_5	3.1962	9.614	0.332	0.740	-15.702	22.094
hour_6	-1.7685	9.637	-0.184	0.854	-20.713	17.176
hour_7	0.5393	9.883	0.055	0.957	-18.888	19.967
hour_8	-0.3656	9.902	-0.037	0.971	-19.831	19.100
hour_9	-5.4397	9.911	-0.549	0.583	-24.922	14.043
hour_10	1.9893	9.882	0.201	0.841	-17.437	21.416
hour_11	3.9459	9.879	0.399	0.690	-15.474	23.366
hour_12	-2.9568	9.931	-0.298	0.766	-22.478	16.564
hour_13	-0.5265	10.173	-0.052	0.959	-20.524	19.471
hour_14	32.9563	9.902	3.328	0.001	13.491	52.422
hour_15	11.0202	10.088	1.092	0.275	-8.811	30.851
hour_16	-1.8384	9.982	-0.184	0.854	-21.461	17.784
hour_17	29.1220	10.143	2.871	0.004	9.183	49.061
hour_18	-20.3921	10.346	-1.971	0.049	-40.729	-0.055
hour_19	-7.7770	10.007	-0.777	0.438	-27.449	11.895
hour_20	-5.8332	9.881	-0.590	0.555	-25.256	13.590
hour_21	-2.5363	9.880	-0.257	0.798	-21.958	16.885
hour_22	-3.1877	9.867	-0.323	0.747	-22.585	16.209
hour_23	-2.2334	9.862	-0.226	0.821	-21.620	17.153

Omnibus:	715.458	Durbin-Watson:	2.142
Prob(Omnibus):	0.000	Jarque-Bera (JB):	329761.986
Skew:	9.098	Prob(JB):	0.00
Kurtosis:	136.030	Cond. No.	1.37e+06

Time period 2:

RMSE is: 20.32109976341389 MSE: 412.94709559462005



OLS Regression Results

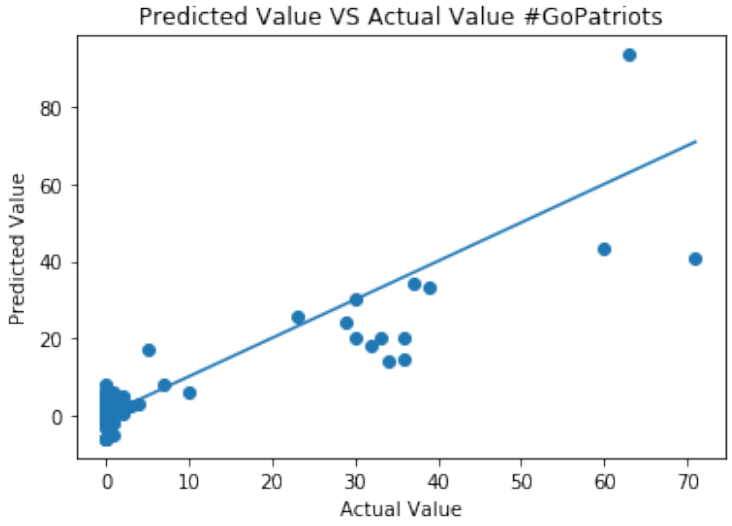
Dep. Variable:	tweet	R-squared:	0.769
Model:	OLS	Adj. R-squared:	0.718
Method:	Least Squares	F-statistic:	15.00
Date:	Thu, 19 Mar 2020	Prob (F-statistic):	2.97e-26
Time:	11:34:29	Log-Likelihood:	-638.01
No. Observations:	144	AIC:	1330.
Df Residuals:	117	BIC:	1410.
Df Model:	26		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
tweet	0.5648	0.175	3.220	0.002	0.217	0.912
follower	0.0002	0.000	1.061	0.291	-0.000	0.001
retweet	-0.0059	0.066	-0.088	0.930	-0.137	0.126
Maxfollowers	-0.0002	0.000	-1.019	0.311	-0.001	0.000
hour_8	-4.2175	6.500	-0.649	0.518	-17.090	8.655
hour_9	-4.0114	6.541	-0.613	0.541	-16.965	8.942
hour_10	-2.5454	6.388	-0.398	0.691	-15.196	10.106
hour_11	-2.2100	6.642	-0.333	0.740	-15.364	10.944
hour_12	-2.1312	6.346	-0.336	0.738	-14.699	10.437
hour_13	-0.4604	6.357	-0.072	0.942	-13.049	12.129
hour_14	2.5025	6.309	0.397	0.692	-9.993	14.998
hour_15	10.3954	8.093	1.284	0.202	-5.632	26.423
hour_16	8.5414	6.883	1.241	0.217	-5.090	22.173
hour_17	5.8409	6.910	0.845	0.400	-7.844	19.526
hour_18	13.4263	6.314	2.126	0.036	0.921	25.932
hour_19	28.9179	18.619	1.553	0.123	-7.957	65.793
hour_20	9.623e-11	1.56e-10	0.616	0.539	-2.13e-10	4.06e-10
minute_0	12.2586	6.567	1.867	0.064	-0.746	25.263
minute_1	0.2279	6.892	0.033	0.974	-13.422	13.878
minute_2	0.3812	6.565	0.058	0.954	-12.621	13.383
minute_3	2.5466	6.477	0.393	0.695	-10.280	15.373
minute_4	3.4751	6.434	0.540	0.590	-9.267	16.217
minute_5	5.8810	6.626	0.888	0.377	-7.241	19.003
minute_6	8.1658	6.563	1.244	0.216	-4.832	21.164
minute_7	5.3405	6.729	0.794	0.429	-7.986	18.667
minute_8	-0.6951	7.114	-0.098	0.922	-14.784	13.394
minute_9	6.0268	6.523	0.924	0.357	-6.892	18.946
minute_10	6.6603	6.538	1.019	0.310	-6.288	19.608
minute_11	3.7798	6.562	0.576	0.566	-9.216	16.776

Omnibus:	130.596	Durbin-Watson:	1.317
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4031.541
Skew:	2.829	Prob(JB):	0.00
Kurtosis:	28.296	Cond. No.	4.73e+17

Time period 3:

RMSE is: 5.9775625555186345 MSE: 35.73125410513846



OLS Regression Results

Dep. Variable:	tweet	R-squared:	0.782
Model:	OLS	Adj. R-squared:	0.727
Method:	Least Squares	F-statistic:	14.09
Date:	Thu, 19 Mar 2020	Prob (F-statistic):	3.82e-24
Time:	11:34:29	Log-Likelihood:	-429.73
No. Observations:	134	AIC:	915.5
Df Residuals:	106	BIC:	996.6
Df Model:	27		
Covariance Type:	nonrobust		

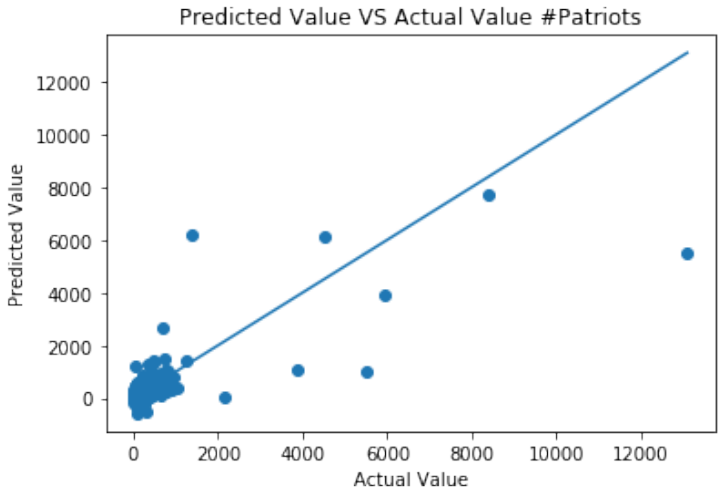
	coef	std err	t	P>|t|	[0.025	0.975]
tweet	0.8763	0.166	5.288	0.000	0.548	1.205
follower	-0.0010	0.000	-2.956	0.004	-0.002	-0.000
retweet	0.0290	0.020	1.421	0.158	-0.011	0.069
Maxfollowers	0.0014	0.000	3.401	0.001	0.001	0.002
hour_0	-0.0725	2.806	-0.026	0.979	-5.636	5.491
hour_1	1.8621	2.768	0.673	0.503	-3.625	7.349
hour_2	2.4663	2.759	0.894	0.373	-3.003	7.936
hour_3	2.7604	2.754	1.002	0.318	-2.699	8.220
hour_4	3.1246	2.773	1.127	0.262	-2.373	8.622
hour_5	0.4994	2.879	0.173	0.863	-5.208	6.207
hour_6	0.9354	2.762	0.339	0.736	-4.541	6.412
hour_7	4.0143	2.771	1.449	0.150	-1.479	9.508
hour_8	4.3498	2.780	1.565	0.121	-1.161	9.861
hour_9	-0.2776	2.793	-0.099	0.921	-5.816	5.261
hour_10	-3.2219	3.165	-1.018	0.311	-9.496	3.052
hour_11	1.1815	3.012	0.392	0.696	-4.790	7.153
hour_12	-0.3209	3.025	-0.106	0.916	-6.318	5.676
hour_13	-1.0241	3.010	-0.340	0.734	-6.993	4.944
hour_14	0.0291	3.006	0.010	0.992	-5.931	5.990
hour_15	0.4374	3.006	0.146	0.885	-5.523	6.398
hour_16	0.2815	3.008	0.094	0.926	-5.682	6.245
hour_17	-0.4839	3.009	-0.161	0.873	-6.450	5.483
hour_18	-0.1596	3.006	-0.053	0.958	-6.120	5.801
hour_19	0.0187	3.006	0.006	0.995	-5.941	5.978
hour_20	-6.1134	2.932	-2.085	0.039	-11.926	-0.301
hour_21	3.5123	2.970	1.183	0.240	-2.375	9.400
hour_22	5.7333	2.805	2.044	0.043	0.173	11.294
hour_23	1.1905	2.828	0.421	0.675	-4.417	6.798

Omnibus:	55.052	Durbin-Watson:	1.219
Prob(Omnibus):	0.000	Jarque-Bera (JB):	656.319
Skew:	0.988	Prob(JB):	3.03e-143
Kurtosis:	13.661	Cond. No.	9.85e+04

#Patriots

Time period 1:

RMSE is: 570.2522769682934 MSE: 325187.65938752320730



OLS Regression Results

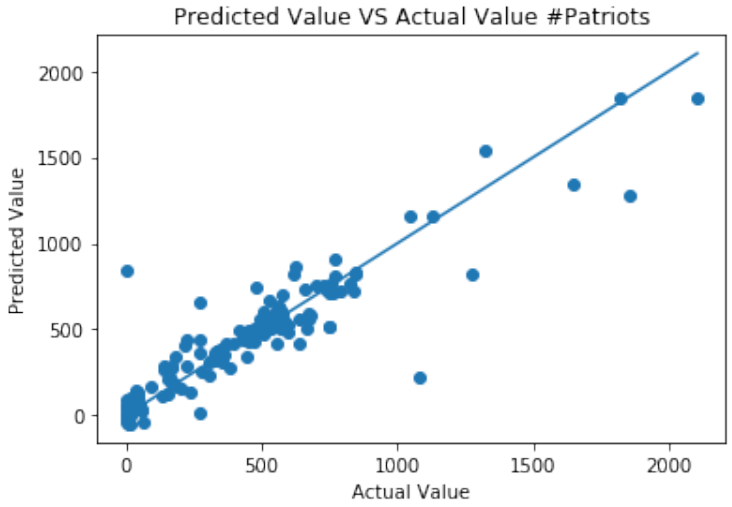
Dep. Variable:	tweet	R-squared:	0.584
Model:	OLS	Adj. R-squared:	0.556
Method:	Least Squares	F-statistic:	21.35
Date:	Thu, 19 Mar 2020	Prob (F-statistic):	5.25e-62
Time:	11:36:07	Log-Likelihood:	-3408.8
No. Observations:	439	AIC:	6874.
Df Residuals:	411	BIC:	6988.
Df Model:	27		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
tweet	-0.2326	0.193	-1.204	0.229	-0.612	0.147
follower	0.0002	3.17e-05	6.867	0.000	0.000	0.000
retweet	0.1194	0.079	1.517	0.130	-0.035	0.274
Maxfollowers	-0.0003	4.97e-05	-6.028	0.000	-0.000	-0.000
hour_0	43.1830	135.345	0.319	0.750	-222.873	309.239
hour_1	46.5200	135.257	0.344	0.731	-219.362	312.402
hour_2	60.8959	135.260	0.450	0.653	-204.992	326.783
hour_3	74.3672	135.640	0.548	0.584	-192.267	341.002
hour_4	64.9851	135.481	0.480	0.632	-201.337	331.307
hour_5	149.4478	135.401	1.104	0.270	-116.717	415.613
hour_6	129.6446	135.938	0.954	0.341	-137.577	396.866
hour_7	67.1302	141.423	0.475	0.635	-210.872	345.133
hour_8	38.0337	140.601	0.271	0.787	-238.353	314.420
hour_9	14.4559	140.343	0.103	0.918	-261.423	290.335
hour_10	124.5324	142.135	0.876	0.381	-154.869	403.934
hour_11	144.4742	140.798	1.026	0.305	-132.299	421.247
hour_12	-161.6128	143.162	-1.129	0.260	-443.035	119.809
hour_13	61.0398	142.242	0.429	0.668	-218.573	340.653
hour_14	190.6455	142.791	1.335	0.183	-90.046	471.337
hour_15	204.4352	144.910	1.411	0.159	-80.423	489.293
hour_16	61.7597	142.417	0.434	0.665	-218.197	341.716
hour_17	473.5905	142.245	3.329	0.001	193.971	753.210
hour_18	94.7277	143.928	0.658	0.511	-188.199	377.654
hour_19	-130.8437	140.820	-0.929	0.353	-407.661	145.974
hour_20	34.5866	140.565	0.246	0.806	-241.729	310.902
hour_21	9.0599	139.254	0.065	0.948	-264.680	282.800
hour_22	41.0195	139.102	0.295	0.768	-232.421	314.460
hour_23	33.1391	139.174	0.238	0.812	-240.442	306.720

Omnibus:	547.578	Durbin-Watson:	2.048
Prob(Omnibus):	0.000	Jarque-Bera (JB):	159513.348
Skew:	5.431	Prob(JB):	0.00
Kurtosis:	95.750	Cond. No.	2.52e+07

Time period 2:

RMSE is: 152.91539600359496 MSE: 23383.11833493626546



OLS Regression Results

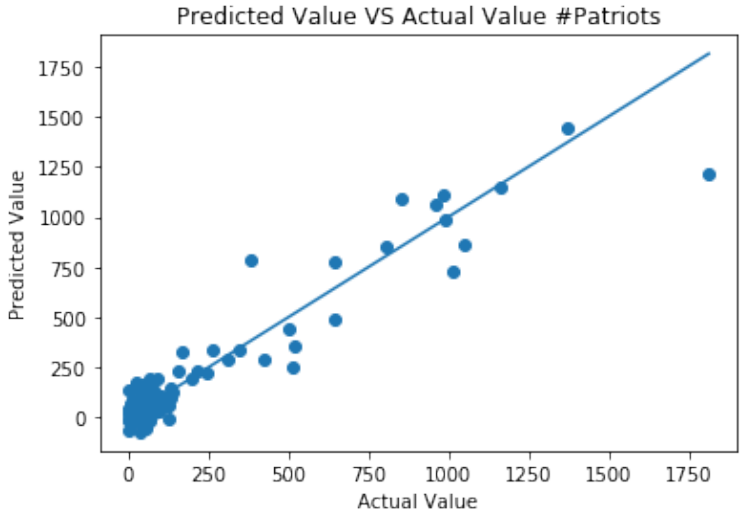
Dep. Variable:	tweet	R-squared:	0.841
Model:	OLS	Adj. R-squared:	0.806
Method:	Least Squares	F-statistic:	23.80
Date:	Sun, 22 Mar 2020	Prob (F-statistic):	2.93e-35
Time:	14:24:13	Log-Likelihood:	-928.63
No. Observations:	144	AIC:	1911.
Df Residuals:	117	BIC:	1991.
Df Model:	26		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
tweet	0.0715	0.192	0.372	0.710	-0.309	0.452
follower	0.0002	5.11e-05	4.638	0.000	0.000	0.000
retweet	0.0076	0.073	0.105	0.917	-0.136	0.151
Maxfollowers	-0.0003	6.13e-05	-4.219	0.000	-0.000	-0.000
hour_8	-118.7008	52.887	-2.244	0.027	-223.440	-13.962
hour_9	-121.3501	53.525	-2.267	0.025	-227.355	-15.346
hour_10	-83.4421	51.352	-1.625	0.107	-185.143	18.258
hour_11	270.5146	63.418	4.266	0.000	144.919	396.111
hour_12	479.7967	101.798	4.713	0.000	278.192	681.402
hour_13	311.8468	75.195	4.147	0.000	162.927	460.766
hour_14	229.2142	62.479	3.669	0.000	105.478	352.950
hour_15	216.3420	70.091	3.087	0.003	77.531	355.153
hour_16	82.3800	50.526	1.630	0.106	-17.685	182.445
hour_17	-55.1538	51.400	-1.073	0.285	-156.950	46.642
hour_18	3.3916	49.301	0.069	0.945	-94.246	101.030
hour_19	317.6876	118.105	2.690	0.008	83.787	551.588
hour_20	-3.961e-08	1.98e-08	-2.004	0.047	-7.87e-08	-4.62e-10
minute_0	62.9782	52.047	1.210	0.229	-40.098	166.054
minute_1	70.2098	55.918	1.256	0.212	-40.533	180.952
minute_2	102.7999	52.409	1.961	0.052	-0.994	206.594
minute_3	115.9717	51.933	2.233	0.027	13.121	218.822
minute_4	111.6351	51.497	2.168	0.032	9.648	213.623
minute_5	169.1130	53.184	3.180	0.002	63.785	274.441
minute_6	214.6869	57.240	3.751	0.000	101.326	328.048
minute_7	178.1135	57.385	3.104	0.002	64.466	291.761
minute_8	136.7307	58.472	2.338	0.021	20.930	252.532
minute_9	125.7382	56.627	2.220	0.028	13.591	237.886
minute_10	162.4094	54.442	2.983	0.003	54.591	270.228
minute_11	82.1402	56.235	1.461	0.147	-29.230	193.510

Omnibus:	51.556	Durbin-Watson:	1.405
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1083.093
Skew:	0.566	Prob(JB):	6.45e-236
Kurtosis:	16.388	Cond. No.	1.33e+19

Time period 3:

RMSE is: 93.86997813654729 MSE: 8811.57279535586623



OLS Regression Results

Dep. Variable:	tweet	R-squared:	0.902
Model:	OLS	Adj. R-squared:	0.877
Method:	Least Squares	F-statistic:	36.14
Date:	Thu, 19 Mar 2020	Prob (F-statistic):	8.75e-42
Time:	11:36:07	Log-Likelihood:	-798.75
No. Observations:	134	AIC:	1654.
Df Residuals:	106	BIC:	1735.
Df Model:	27		
Covariance Type:	nonrobust		

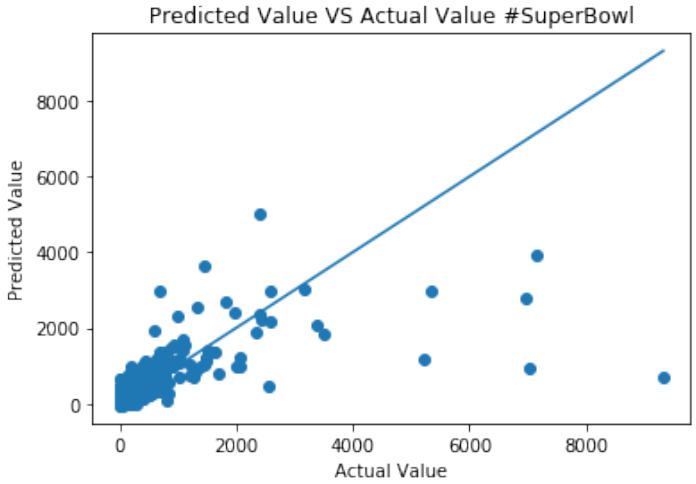
	coef	std err	t	P>|t|	[0.025	0.975]
tweet	0.7776	0.096	8.114	0.000	0.588	0.968
follower	5.521e-05	1.33e-05	4.135	0.000	2.87e-05	8.17e-05
retweet	-0.0520	0.006	-8.898	0.000	-0.064	-0.040
Maxfollowers	-5.618e-05	1.64e-05	-3.430	0.001	-8.86e-05	-2.37e-05
hour_0	37.0977	43.406	0.855	0.395	-48.958	123.153
hour_1	20.2971	43.289	0.469	0.640	-65.529	106.123
hour_2	39.4974	43.402	0.910	0.365	-46.551	125.546
hour_3	-3.7528	44.223	-0.085	0.933	-91.430	83.924
hour_4	72.4330	43.672	1.659	0.100	-14.150	159.016
hour_5	23.9970	44.897	0.534	0.594	-65.015	113.009
hour_6	22.4335	44.071	0.509	0.612	-64.941	109.808
hour_7	-22.6444	44.323	-0.511	0.610	-110.519	65.230
hour_8	127.7876	44.116	2.897	0.005	40.323	215.252
hour_9	-67.7060	44.922	-1.507	0.135	-156.768	21.356
hour_10	-104.7575	48.864	-2.144	0.034	-201.634	-7.881
hour_11	6.3676	48.108	0.132	0.895	-89.011	101.746
hour_12	11.8406	47.775	0.248	0.805	-82.878	106.559
hour_13	42.3102	47.800	0.885	0.378	-52.458	137.078
hour_14	2.0959	48.102	0.044	0.965	-93.270	97.462
hour_15	3.8143	47.924	0.080	0.937	-91.200	98.828
hour_16	6.8983	50.126	0.138	0.891	-92.481	106.278
hour_17	34.9191	47.591	0.734	0.465	-59.435	129.273
hour_18	-17.9062	47.900	-0.374	0.709	-112.873	77.060
hour_19	19.7966	47.434	0.417	0.677	-74.245	113.838
hour_20	1.3473	46.857	0.029	0.977	-91.551	94.246
hour_21	111.3320	44.042	2.528	0.013	24.015	198.649
hour_22	-28.6840	44.809	-0.640	0.523	-117.523	60.155
hour_23	-29.8776	43.880	-0.681	0.497	-116.874	57.119

Omnibus:	75.339	Durbin-Watson:	2.410
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1082.383
Skew:	1.517	Prob(JB):	9.20e-236
Kurtosis:	16.589	Cond. No.	2.10e+07

#SuperBowl

Time period 1:

RMSE is: 701.7720660442512 MSE: 492484.03268001686809



OLS Regression Results

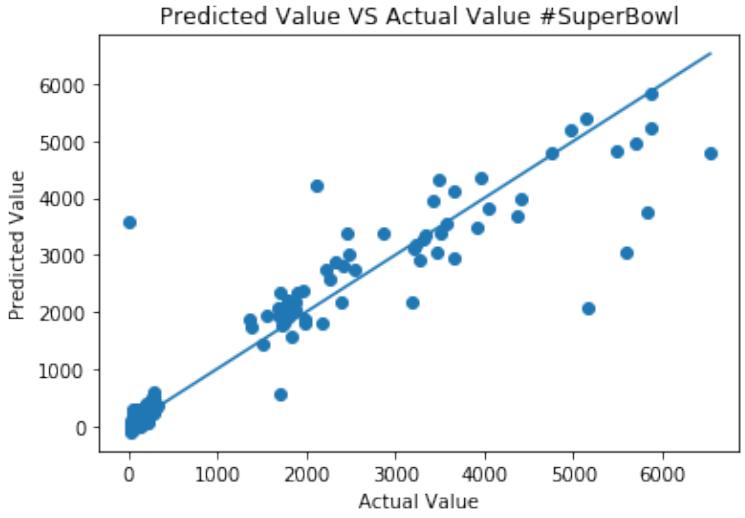
Dep. Variable:	tweet	R-squared:	0.429
Model:	OLS	Adj. R-squared:	0.392
Method:	Least Squares	F-statistic:	11.45
Date:	Thu, 19 Mar 2020	Prob (F-statistic):	1.74e-35
Time:	11:40:25	Log-Likelihood:	-3499.9
No. Observations:	439	AIC:	7056.
Df Residuals:	411	BIC:	7170.
Df Model:	27		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
tweet	0.3555	0.126	2.822	0.005	0.108	0.603
follower	2.629e-05	6.14e-06	4.283	0.000	1.42e-05	3.84e-05
retweet	0.0566	0.061	0.927	0.354	-0.063	0.176
Maxfollowers	-4.431e-05	2.56e-05	-1.731	0.084	-9.46e-05	6e-06
hour_0	61.1748	167.942	0.364	0.716	-268.958	391.308
hour_1	72.5465	167.003	0.434	0.664	-255.740	400.833
hour_2	78.4526	167.008	0.470	0.639	-249.843	406.749
hour_3	150.2787	167.422	0.898	0.370	-178.832	479.389
hour_4	110.3789	167.275	0.660	0.510	-218.443	439.201
hour_5	230.6988	167.794	1.375	0.170	-99.143	560.540
hour_6	373.3387	168.984	2.209	0.028	41.157	705.520
hour_7	182.8593	172.482	1.060	0.290	-156.198	521.917
hour_8	53.5742	173.655	0.309	0.758	-287.789	394.937
hour_9	105.6277	174.486	0.605	0.545	-237.368	448.624
hour_10	171.1519	174.087	0.983	0.326	-171.060	513.363
hour_11	102.1782	173.233	0.590	0.556	-238.355	442.711
hour_12	11.0441	174.537	0.063	0.950	-332.053	354.141
hour_13	52.5513	173.933	0.302	0.763	-289.359	394.461
hour_14	446.8862	174.422	2.562	0.011	104.015	789.758
hour_15	-36.0701	176.348	-0.205	0.838	-382.727	310.587
hour_16	202.2181	173.755	1.164	0.245	-139.341	543.777
hour_17	640.0574	176.523	3.626	0.000	293.057	987.058
hour_18	101.4341	178.339	0.569	0.570	-249.136	452.005
hour_19	-57.8082	176.014	-0.328	0.743	-403.807	288.191
hour_20	22.9888	172.554	0.133	0.894	-316.210	362.187
hour_21	59.6216	171.649	0.347	0.729	-277.798	397.042
hour_22	24.3482	171.749	0.142	0.887	-313.269	361.965
hour_23	52.7155	172.371	0.306	0.760	-286.123	391.554

Omnibus:	595.316	Durbin-Watson:	2.041
Prob(Omnibus):	0.000	Jarque-Bera (JB):	92295.594
Skew:	6.726	Prob(JB):	0.00
Kurtosis:	72.748	Cond. No.	8.86e+07

Time period 2:

RMSE is: 606.1875966790903 MSE: 367463.40236757144930



OLS Regression Results

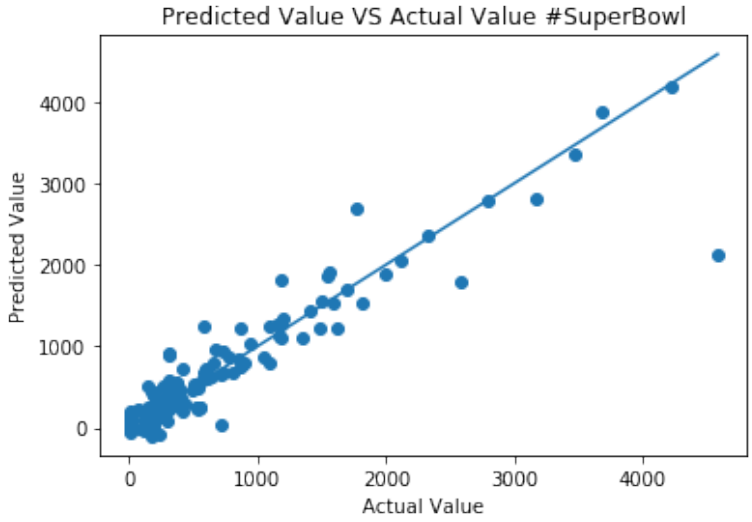
Dep. Variable:	tweet	R-squared:	0.875
Model:	OLS	Adj. R-squared:	0.847
Method:	Least Squares	F-statistic:	31.43
Date:	Sun, 22 Mar 2020	Prob (F-statistic):	3.97e-41
Time:	14:26:47	Log-Likelihood:	-1127.0
No. Observations:	144	AIC:	2308.
Df Residuals:	117	BIC:	2388.
Df Model:	26		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
tweet	0.5588	0.183	3.054	0.003	0.196	0.921
follower	2.051e-05	1.15e-05	1.781	0.077	-2.3e-06	4.33e-05
retweet	-0.0012	0.008	-0.149	0.882	-0.016	0.014
Maxfollowers	-3.181e-05	2.05e-05	-1.553	0.123	-7.24e-05	8.75e-06
hour_8	-195.4944	206.444	-0.947	0.346	-604.347	213.358
hour_9	-239.8574	216.115	-1.110	0.269	-667.862	188.147
hour_10	-191.2782	207.231	-0.923	0.358	-601.688	219.132
hour_11	-176.4272	208.541	-0.846	0.399	-589.432	236.578
hour_12	-168.2396	204.259	-0.824	0.412	-572.764	236.284
hour_13	-145.9300	202.130	-0.722	0.472	-546.239	254.379
hour_14	-18.2414	198.803	-0.092	0.927	-411.960	375.477
hour_15	509.1997	284.913	1.787	0.076	-55.055	1073.454
hour_16	1066.8099	393.079	2.714	0.008	288.338	1845.281
hour_17	839.2106	539.917	1.554	0.123	-230.068	1908.489
hour_18	674.1855	263.078	2.563	0.012	153.174	1195.197
hour_19	1031.2346	741.887	1.390	0.167	-438.033	2500.503
hour_20	-6.099e-07	9.9e-07	-0.616	0.539	-2.57e-06	1.35e-06
minute_0	308.5484	237.968	1.297	0.197	-162.735	779.831
minute_1	112.7101	246.460	0.457	0.648	-375.391	600.811
minute_2	100.8527	217.434	0.464	0.644	-329.764	531.469
minute_3	424.7088	205.682	2.065	0.041	17.367	832.051
minute_4	353.0784	228.694	1.544	0.125	-99.839	805.996
minute_5	216.8121	232.700	0.932	0.353	-244.038	677.662
minute_6	220.8636	226.093	0.977	0.331	-226.901	668.628
minute_7	180.4453	219.008	0.824	0.412	-253.288	614.179
minute_8	179.8738	218.437	0.823	0.412	-252.729	612.477
minute_9	199.7337	212.173	0.941	0.348	-220.463	619.931
minute_10	287.0845	208.885	1.374	0.172	-126.601	700.770
minute_11	400.4608	217.826	1.838	0.069	-30.932	831.854

Omnibus:	47.425	Durbin-Watson:	1.508
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1375.474
Skew:	0.149	Prob(JB):	2.09e-299
Kurtosis:	18.138	Cond. No.	8.35e+17

Time period 3:

RMSE is: 305.15646496688504 MSE: 93120.46811108573673



OLS Regression Results

Dep. Variable:	tweet	R-squared:	0.869
Model:	OLS	Adj. R-squared:	0.836
Method:	Least Squares	F-statistic:	26.07
Date:	Thu, 19 Mar 2020	Prob (F-statistic):	2.58e-35
Time:	11:40:26	Log-Likelihood:	-956.73
No. Observations:	134	AIC:	1969.
Df Residuals:	106	BIC:	2051.
Df Model:	27		
Covariance Type:	nonrobust		

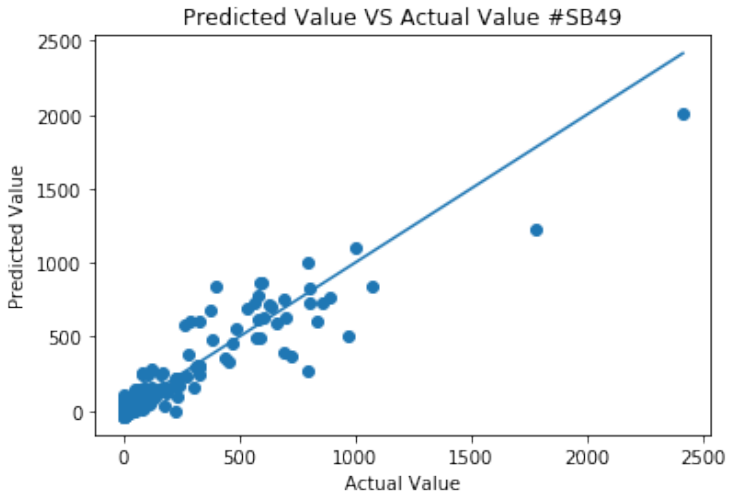
	coef	std err	t	P>|t|	[0.025	0.975]
tweet	0.9219	0.112	8.224	0.000	0.700	1.144
follower	5.087e-06	5.55e-06	0.917	0.361	-5.92e-06	1.61e-05
retweet	-0.0517	0.014	-3.621	0.000	-0.080	-0.023
Maxfollowers	2.806e-05	1.56e-05	1.800	0.075	-2.85e-06	5.9e-05
hour_0	-33.1147	143.443	-0.231	0.818	-317.504	251.275
hour_1	25.3362	144.499	0.175	0.861	-261.147	311.820
hour_2	27.6965	142.525	0.194	0.846	-254.874	310.267
hour_3	194.8438	147.435	1.322	0.189	-97.460	487.147
hour_4	105.9545	143.526	0.738	0.462	-178.601	390.510
hour_5	106.8725	144.861	0.738	0.462	-180.328	394.073
hour_6	20.8595	143.908	0.145	0.885	-264.452	306.171
hour_7	-12.5729	148.581	-0.085	0.933	-307.149	282.003
hour_8	142.0389	148.450	0.957	0.341	-152.278	436.355
hour_9	-70.1721	150.774	-0.465	0.643	-369.097	228.752
hour_10	565.3925	167.237	3.381	0.001	233.829	896.956
hour_11	-107.8623	171.551	-0.629	0.531	-447.980	232.255
hour_12	-145.3834	164.261	-0.885	0.378	-471.047	180.280
hour_13	-43.7385	162.588	-0.269	0.788	-366.085	278.608
hour_14	-3.7524	159.567	-0.024	0.981	-320.110	312.605
hour_15	-75.5941	163.677	-0.462	0.645	-400.100	248.912
hour_16	-36.3324	161.422	-0.225	0.822	-356.367	283.702
hour_17	-220.9468	170.545	-1.296	0.198	-559.069	117.175
hour_18	-40.3413	155.704	-0.259	0.796	-349.040	268.357
hour_19	0.3444	155.644	0.002	0.998	-308.234	308.923
hour_20	-286.7216	151.870	-1.888	0.062	-587.818	14.375
hour_21	35.1822	152.450	0.231	0.818	-267.065	337.430
hour_22	-117.7989	147.747	-0.797	0.427	-410.722	175.124
hour_23	-66.9136	146.449	-0.457	0.649	-357.262	223.435

Omnibus:	147.785	Durbin-Watson:	1.991
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5558.814
Skew:	3.703	Prob(JB):	0.00
Kurtosis:	33.672	Cond. No.	1.83e+08

#SB49

Time period 1:

RMSE is: 79.824298866149 MSE: 6371.91868947227652



OLS Regression Results

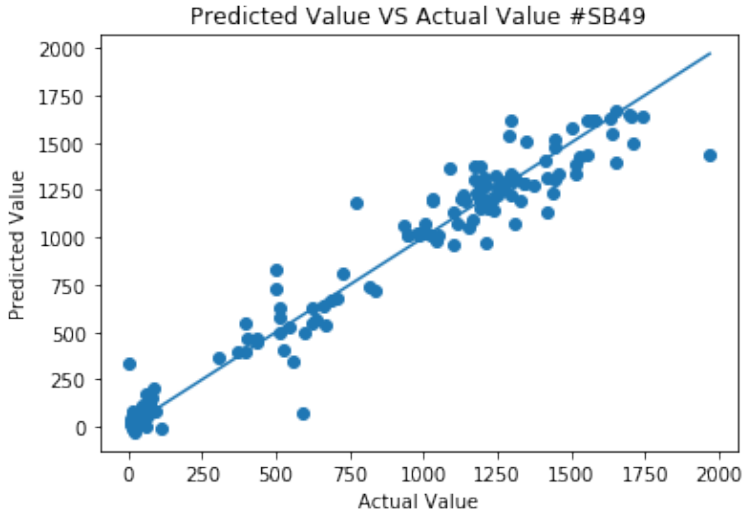
Dep. Variable:	tweet	R-squared:	0.877
Model:	OLS	Adj. R-squared:	0.869
Method:	Least Squares	F-statistic:	108.4
Date:	Thu, 19 Mar 2020	Prob (F-statistic):	1.55e-168
Time:	11:44:32	Log-Likelihood:	-2545.7
No. Observations:	439	AIC:	5147.
Df Residuals:	411	BIC:	5262.
Df Model:	27		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
tweet	1.0446	0.040	25.862	0.000	0.965	1.124
follower	-4.398e-06	1.4e-06	-3.152	0.002	-7.14e-06	-1.65e-06
retweet	0.0240	0.005	4.881	0.000	0.014	0.034
Maxfollowers	3.946e-06	3.43e-06	1.149	0.251	-2.81e-06	1.07e-05
hour_0	-17.4353	18.994	-0.918	0.359	-54.773	19.903
hour_1	-7.5314	18.960	-0.397	0.691	-44.803	29.740
hour_2	6.4359	18.962	0.339	0.734	-30.840	43.711
hour_3	3.6563	18.956	0.193	0.847	-33.606	40.919
hour_4	27.1609	18.972	1.432	0.153	-10.134	64.456
hour_5	41.5225	19.128	2.171	0.031	3.921	79.124
hour_6	78.4137	19.186	4.087	0.000	40.698	116.129
hour_7	23.0639	19.626	1.175	0.241	-15.515	61.643
hour_8	-5.2467	19.602	-0.268	0.789	-43.779	33.286
hour_9	11.0900	19.593	0.566	0.572	-27.425	49.605
hour_10	9.0542	19.681	0.460	0.646	-29.635	47.743
hour_11	-17.2092	19.805	-0.869	0.385	-56.141	21.723
hour_12	-23.1496	19.792	-1.170	0.243	-62.055	15.756
hour_13	17.4860	19.814	0.883	0.378	-21.464	56.436
hour_14	-15.0232	19.644	-0.765	0.445	-53.639	23.593
hour_15	-7.9999	19.930	-0.401	0.688	-47.178	31.178
hour_16	-13.8408	19.620	-0.705	0.481	-52.408	24.727
hour_17	-4.3699	19.784	-0.221	0.825	-43.261	34.521
hour_18	-20.6129	19.604	-1.051	0.294	-59.149	17.924
hour_19	-16.7857	19.613	-0.856	0.393	-55.339	21.768
hour_20	-18.4315	19.648	-0.938	0.349	-57.055	20.192
hour_21	-13.3566	19.548	-0.683	0.495	-51.783	25.069
hour_22	-39.3063	19.569	-2.009	0.045	-77.775	-0.838
hour_23	-3.1885	19.489	-0.164	0.870	-41.499	35.122

Omnibus:	203.658	Durbin-Watson:	1.480
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5978.463
Skew:	1.365	Prob(JB):	0.00
Kurtosis:	20.872	Cond. No.	5.33e+07

Time period 2:

RMSE is: 128.46263732553362 MSE: 16502.64918863158362



OLS Regression Results

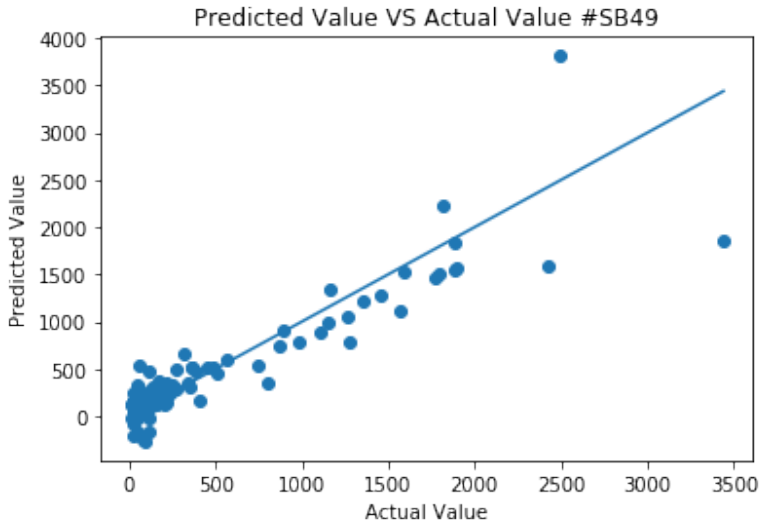
Dep. Variable:	tweet	R-squared:	0.948
Model:	OLS	Adj. R-squared:	0.937
Method:	Least Squares	F-statistic:	82.14
Date:	Sun, 22 Mar 2020	Prob (F-statistic):	4.48e-63
Time:	14:28:35	Log-Likelihood:	-903.54
No. Observations:	144	AIC:	1861.
Df Residuals:	117	BIC:	1941.
Df Model:	26		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
tweet	0.6364	0.094	6.747	0.000	0.450	0.823
follower	1.24e-05	5.19e-06	2.390	0.018	2.12e-06	2.27e-05
retweet	-0.0259	0.019	-1.370	0.173	-0.063	0.012
Maxfollowers	-9.334e-06	7.71e-06	-1.211	0.228	-2.46e-05	5.93e-06
hour_8	-132.3307	51.593	-2.565	0.012	-234.507	-30.154
hour_9	-137.8714	53.606	-2.572	0.011	-244.035	-31.707
hour_10	-91.2214	49.809	-1.831	0.070	-189.866	7.423
hour_11	312.7391	66.798	4.682	0.000	180.450	445.028
hour_12	445.9607	113.098	3.943	0.000	221.977	669.945
hour_13	311.9884	84.426	3.695	0.000	144.788	479.189
hour_14	230.9294	66.100	3.494	0.001	100.022	361.837
hour_15	270.1447	84.804	3.186	0.002	102.194	438.095
hour_16	339.5731	92.948	3.653	0.000	155.495	523.651
hour_17	194.8868	92.921	2.097	0.038	10.862	378.912
hour_18	39.5184	47.354	0.835	0.406	-54.265	133.301
hour_19	-7.6430	40.889	-0.187	0.852	-88.621	73.335
hour_20	-1.037e-09	1.16e-07	-0.009	0.993	-2.32e-07	2.3e-07
minute_0	77.5516	63.212	1.227	0.222	-47.636	202.740
minute_1	99.8384	54.922	1.818	0.072	-8.932	208.609
minute_2	123.2466	51.814	2.379	0.019	20.632	225.861
minute_3	207.5325	50.860	4.080	0.000	106.806	308.259
minute_4	181.9183	55.649	3.269	0.001	71.709	292.127
minute_5	144.9688	56.881	2.549	0.012	32.319	257.618
minute_6	176.5228	55.006	3.209	0.002	67.586	285.460
minute_7	161.4427	56.750	2.845	0.005	49.053	273.832
minute_8	139.1341	55.608	2.502	0.014	29.006	249.262
minute_9	132.5621	53.829	2.463	0.015	25.957	239.168
minute_10	207.9134	53.265	3.903	0.000	102.425	313.402
minute_11	124.0429	56.114	2.211	0.029	12.912	235.174

Omnibus:	23.752	Durbin-Watson:	2.137
Prob(Omnibus):	0.000	Jarque-Bera (JB):	90.534
Skew:	0.453	Prob(JB):	2.19e-20
Kurtosis:	6.777	Cond. No.	1.98e+18

Time period 3:

RMSE is: 249.32054775577703 MSE: 62160.73553324069463



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  tweet   R-squared:                       0.830
Model:                            OLS   Adj. R-squared:                  0.787
Method:                 Least Squares   F-statistic:                     19.23
Date:                Thu, 19 Mar 2020   Prob (F-statistic):           1.33e-29
Time:                        11:44:32   Log-Likelihood:                -929.65
No. Observations:                 134   AIC:                             1915.
Df Residuals:                     106   BIC:                             1996.
Df Model:                          27
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
tweet          0.7893      0.102      7.774      0.000       0.588       0.991
follower   -4.817e-07   5.73e-06     -0.084      0.933   -1.18e-05    1.09e-05
retweet        -0.0302      0.013     -2.397      0.018      -0.055      -0.005
Maxfollowers 2.531e-05   1.23e-05      2.057      0.042    9.13e-07    4.97e-05
hour_0        -55.6483    117.250     -0.475      0.636    -288.108     176.811
hour_1         14.8112    115.805      0.128      0.898    -214.783     244.405
hour_2        118.1703    114.721      1.030      0.305    -109.275     345.616
hour_3        123.7800    115.363      1.073      0.286    -104.939     352.499
hour_4         93.9042    116.105      0.809      0.420    -136.285     324.094
hour_5        127.6977    115.534      1.105      0.272    -101.359     356.755
hour_6         90.0455    117.010      0.770      0.443    -141.937     322.028
hour_7         98.2846    116.819      0.841      0.402    -133.320     329.889
hour_8        181.3135    117.527      1.543      0.126     -51.696     414.323
hour_9         -7.6188    117.933     -0.065      0.949    -241.432     226.194
hour_10       -16.0521    129.527     -0.124      0.902    -272.852     240.748
hour_11        16.6445    128.033      0.130      0.897    -237.193     270.482
hour_12        -2.9686    130.896     -0.023      0.982    -262.483     256.546
hour_13       -26.4329    132.265     -0.200      0.842    -288.661     235.795
hour_14       -69.9276    134.432     -0.520      0.604    -336.453     196.598
hour_15        37.6380    126.943      0.296      0.767    -214.040     289.316
hour_16       -12.4719    130.469     -0.096      0.924    -271.140     246.196
hour_17        73.3702    139.077      0.528      0.599    -202.363     349.103
hour_18       -32.0873    128.548     -0.250      0.803    -286.946     222.772
hour_19        20.6461    126.261      0.164      0.870    -229.678     270.971
hour_20      -330.3532    121.498     -2.719      0.008    -571.236     -89.471
hour_21       232.4904    120.677      1.927      0.057      -6.763     471.743
hour_22      -214.1112    127.423     -1.680      0.096    -466.740      38.518
hour_23       -66.4914    116.681     -0.570      0.570    -297.824     164.841
==============================================================================
Omnibus:                       68.756   Durbin-Watson:                   2.156
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1546.338
Skew:                           1.144   Prob(JB):                         0.00
Kurtosis:                      19.484   Cond. No.                     1.38e+08
==============================================================================
```

Hashtag	RMSE (time period 1)	RMSE (time period 2)	RMSE (time period 3)	R^2 (time period 1)	R^2 (time period 2)	R^2 (time period 3)
#NFL	248.8577827345286	34.34881502837251	109.19394002321302	0.541	0.880	0.861
#SuperBowl	701.7720660442512	606.1875966790903	305.15646496688504	0.429	0.875	0.869
#GoHawks	816.6104179136739	42.31315800563716	38.54608451302766	0.353	0.648	0.858
#GoPatriots	40.45812675770317	20.32109976341389	5.9775625555186345	0.600	0.769	0.782
#Patriots	570.2522769682934	152.91539600359496	93.86997813654729	0.584	0.841	0.902
#SB49	79.824298866149	128.46263732553362	249.32054775577703	0.877	0.948	0.830

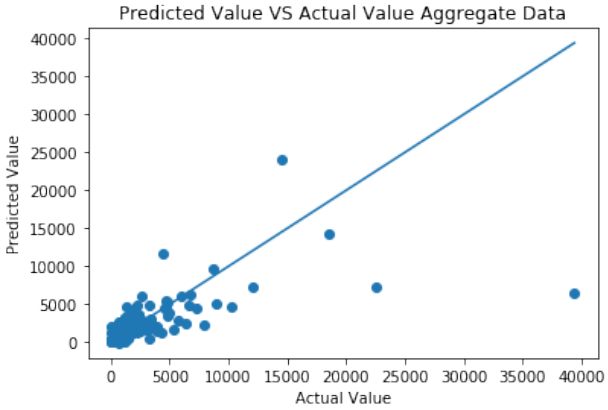
From the MSE results, we found that the RMSE value for piecewise regression is much smaller than the RMSE value from regression models on the entire dataset. It indicates that the error from the piecewise linear regression model is lower and has a better performance.

From the R-squared values, we showed that after the data are divided into different time periods, the performance of the prediction is greatly improved compared with that of training the linear regression model with a whole dataset. The R-squared values increase from 0.6+/- to 0.8+/-, with a highest R-squared value equal to 0.949.

QUESTION 7: Perform the same evaluations on your combined model and compare with models you trained for individual hashtags.

We aggregated the data of all hashtags, and train 3 models for each time period to predict the number of tweets in the next time window on the aggregated data.

Time period 1:
RMSE is: 2043.0891277765588 MSE: $4.17421318403877 \times 10^6$



OLS Regression Results

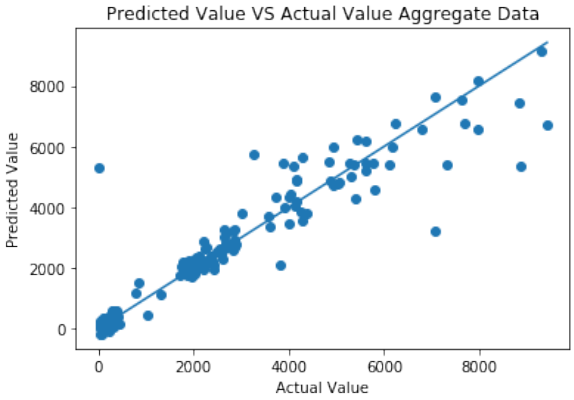
Dep. Variable:	tweet	R-squared:	0.443
Model:	OLS	Adj. R-squared:	0.407
Method:	Least Squares	F-statistic:	12.12
Date:	Thu, 19 Mar 2020	Prob (F-statistic):	1.55e-37
Time:	11:44:33	Log-Likelihood:	-3969.1
No. Observations:	439	AIC:	7994.
Df Residuals:	411	BIC:	8109.
Df Model:	27		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
tweet	0.5751	0.145	3.972	0.000	0.290	0.860
follower	2.192e-05	1.18e-05	1.850	0.065	-1.38e-06	4.52e-05
retweet	-0.0003	0.074	-0.003	0.997	-0.146	0.145
Maxfollowers	-4.439e-05	4.18e-05	-1.062	0.289	-0.000	3.77e-05
hour_0	138.9883	487.711	0.285	0.776	-819.730	1097.707
hour_1	174.2995	486.641	0.358	0.720	-782.317	1130.916
hour_2	223.1661	486.126	0.459	0.646	-732.437	1178.769
hour_3	346.3884	486.267	0.712	0.477	-609.492	1302.269
hour_4	310.1360	486.051	0.638	0.524	-645.320	1265.592
hour_5	689.4738	488.916	1.410	0.159	-271.615	1650.562
hour_6	956.3383	491.375	1.946	0.052	-9.584	1922.261
hour_7	630.2244	504.821	1.248	0.213	-362.128	1622.577
hour_8	426.3455	507.068	0.841	0.401	-570.424	1423.115
hour_9	441.9967	508.204	0.870	0.385	-557.006	1440.999
hour_10	856.3450	508.716	1.683	0.093	-143.666	1856.356
hour_11	879.3836	507.856	1.732	0.084	-118.937	1877.704
hour_12	73.2353	514.093	0.142	0.887	-937.345	1083.815
hour_13	523.7467	514.474	1.018	0.309	-487.581	1535.075
hour_14	2088.9196	512.322	4.077	0.000	1081.822	3096.017
hour_15	-154.7098	533.009	-0.290	0.772	-1202.473	893.053
hour_16	598.2981	514.598	1.163	0.246	-413.275	1609.871
hour_17	1627.6875	527.697	3.085	0.002	590.366	2665.009
hour_18	218.6840	515.651	0.424	0.672	-794.958	1232.326
hour_19	-7.6199	512.702	-0.015	0.988	-1015.465	1000.225
hour_20	160.6065	506.807	0.317	0.751	-835.650	1156.863
hour_21	160.5318	501.023	0.320	0.749	-824.356	1145.420
hour_22	87.9966	501.142	0.176	0.861	-897.124	1073.117
hour_23	138.4222	500.930	0.276	0.782	-846.282	1123.126

Omnibus:	762.915	Durbin-Watson:	2.155
Prob(Omnibus):	0.000	Jarque-Bera (JB):	470171.389
Skew:	10.285	Prob(JB):	0.00
Kurtosis:	162.000	Cond. No.	1.84e+08

Time period 2:

RMSE is: 842.5695104435518 MSE: 709923.37992908654576



OLS Regression Results

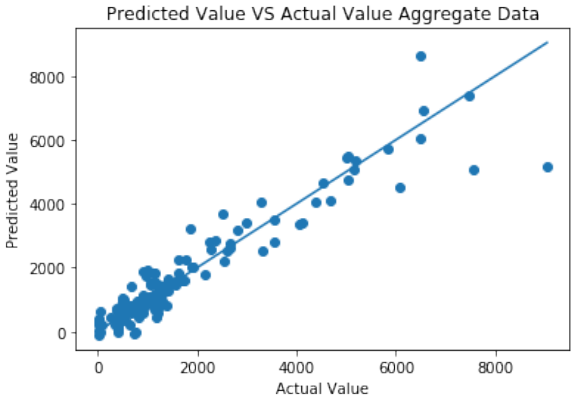
Dep. Variable:	tweet	R-squared:	0.868
Model:	OLS	Adj. R-squared:	0.838
Method:	Least Squares	F-statistic:	29.55
Date:	Sun, 22 Mar 2020	Prob (F-statistic):	8.37e-40
Time:	14:28:36	Log-Likelihood:	-1174.4
No. Observations:	144	AIC:	2403.
Df Residuals:	117	BIC:	2483.
Df Model:	26		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
tweet	0.5014	0.178	2.820	0.006	0.149	0.853
follower	3.637e-05	1.28e-05	2.846	0.005	1.11e-05	6.17e-05
retweet	-0.0012	0.011	-0.107	0.915	-0.023	0.021
Maxfollowers	-5.346e-05	2.32e-05	-2.302	0.023	-9.95e-05	-7.46e-06
hour_8	-459.7044	314.982	-1.459	0.147	-1083.510	164.101
hour_9	-524.4943	334.983	-1.566	0.120	-1187.911	138.923
hour_10	-336.8617	307.141	-1.097	0.275	-945.139	271.416
hour_11	399.3027	274.236	1.456	0.148	-143.807	942.412
hour_12	692.9627	353.137	1.962	0.052	-6.406	1392.331
hour_13	478.7671	305.350	1.568	0.120	-125.963	1083.497
hour_14	475.4824	280.353	1.696	0.093	-79.741	1030.706
hour_15	967.3369	505.991	1.912	0.058	-34.751	1969.425
hour_16	1523.3421	570.292	2.671	0.009	393.908	2652.776
hour_17	932.6404	702.632	1.327	0.187	-458.886	2324.167
hour_18	863.9868	355.817	2.428	0.017	159.309	1568.664
hour_19	1466.2784	956.263	1.533	0.128	-427.551	3360.107
hour_20	5.985e-07	2.36e-06	0.254	0.800	-4.08e-06	5.27e-06
minute_0	356.8672	374.628	0.953	0.343	-385.064	1098.799
minute_1	259.3401	379.382	0.684	0.496	-492.007	1010.687
minute_2	323.6884	329.186	0.983	0.327	-328.247	975.624
minute_3	756.3089	311.356	2.429	0.017	139.685	1372.933
minute_4	731.3616	353.698	2.068	0.041	30.882	1431.841
minute_5	567.6392	359.426	1.579	0.117	-144.186	1279.464
minute_6	640.4700	353.212	1.813	0.072	-59.047	1339.987
minute_7	528.9969	348.535	1.518	0.132	-161.257	1219.251
minute_8	487.2106	353.679	1.378	0.171	-213.233	1187.654
minute_9	499.0456	337.835	1.477	0.142	-170.018	1168.109
minute_10	686.7898	326.967	2.100	0.038	39.248	1334.332
minute_11	641.3210	348.182	1.842	0.068	-48.235	1330.877

Omnibus:	50.342	Durbin-Watson:	1.424
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1324.139
Skew:	-0.400	Prob(JB):	2.93e-288
Kurtosis:	17.834	Cond. No.	9.76e+17

Time period 3:

RMSE is: 603.3893715380818 MSE: 364078.73368512131898



OLS Regression Results

Dep. Variable:	tweet	R-squared:	0.887
Model:	OLS	Adj. R-squared:	0.858
Method:	Least Squares	F-statistic:	30.76
Date:	Thu, 19 Mar 2020	Prob (F-statistic):	1.48e-38
Time:	11:44:33	Log-Likelihood:	-1048.1
No. Observations:	134	AIC:	2152.
Df Residuals:	106	BIC:	2233.
Df Model:	27		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
tweet	0.7614	0.121	6.290	0.000	0.521	1.001
follower	1.395e-05	7.18e-06	1.943	0.055	-2.82e-07	2.82e-05
retweet	-0.0489	0.012	-4.248	0.000	-0.072	-0.026
Maxfollowers	3.323e-05	1.69e-05	1.964	0.052	-3.07e-07	6.68e-05
hour_0	-129.7014	288.457	-0.450	0.654	-701.596	442.193
hour_1	104.7024	287.217	0.365	0.716	-464.734	674.139
hour_2	196.0320	281.282	0.697	0.487	-361.637	753.701
hour_3	374.3960	281.365	1.331	0.186	-183.438	932.230
hour_4	296.6188	286.610	1.035	0.303	-271.614	864.851
hour_5	267.7839	290.651	0.921	0.359	-308.461	844.029
hour_6	174.1913	288.923	0.603	0.548	-398.627	747.010
hour_7	155.7684	294.557	0.529	0.598	-428.219	739.756
hour_8	571.4167	305.155	1.873	0.064	-33.583	1176.416
hour_9	-42.3947	310.038	-0.137	0.891	-657.076	572.286
hour_10	609.9806	335.646	1.817	0.072	-55.469	1275.430
hour_11	-98.3093	332.948	-0.295	0.768	-758.411	561.792
hour_12	-242.0936	331.873	-0.729	0.467	-900.064	415.877
hour_13	-24.8662	333.395	-0.075	0.941	-685.853	636.121
hour_14	-191.5763	326.299	-0.587	0.558	-838.495	455.342
hour_15	-235.7770	322.570	-0.731	0.466	-875.303	403.749
hour_16	-252.2953	328.603	-0.768	0.444	-903.783	399.193
hour_17	-63.8953	330.507	-0.193	0.847	-719.157	591.367
hour_18	-236.7676	317.931	-0.745	0.458	-867.096	393.561
hour_19	173.6681	312.105	0.556	0.579	-445.109	792.446
hour_20	-588.8819	305.403	-1.928	0.057	-1194.373	16.609
hour_21	386.1382	299.744	1.288	0.200	-208.133	980.409
hour_22	-382.9096	314.761	-1.217	0.226	-1006.954	241.135
hour_23	-302.0833	290.071	-1.041	0.300	-877.176	273.010

Omnibus:	93.221	Durbin-Watson:	2.266
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1220.649
Skew:	2.127	Prob(JB):	8.70e-266
Kurtosis:	17.161	Cond. No.	4.08e+08

Hashtag	RMSE (time period 1)	RMSE (time period 2)	RMSE (time period 3)	R^2 (time period 1)	R^2 (time period 2)	R^2 (time period 3)
#NFL	248.8577827345286	34.34881502837251	109.19394002321302	0.541	0.880	0.861
#SuperBowl	701.7720660442512	606.1875966790903	305.15646496688504	0.429	0.875	0.869
#GoHawks	816.6104179136739	42.31315800563716	38.54608451302766	0.353	0.648	0.858
#GoPatriots	40.45812675770317	20.32109976341389	5.9775625555186345	0.600	0.769	0.782
#Patriots	570.2522769682934	152.91539600359496	93.86997813654729	0.584	0.841	0.902
#SB49	79.824298866149	128.46263732553362	249.32054775577703	0.877	0.948	0.830
All hashtags (aggregate data)	**2043.0891277765588**	**842.5695104435518**	**603.3893715380818**	**0.443**	**0.868**	**0.887**

Comparing the results for combined model and models with individual hashtags, we showed that MSE for the aggregate data model increased significantly for all three time periods. It indicates that models based on the dataset of individual hashtags performed much better than the aggregated dataset. This is because different hashtags can have different characteristics and features in the number of tweets, so different models should be used to better track their trends.

5. Nonlinear regressions

Ensemble methods

Ensemble methods is a machine learning technique that combines several base models, producing one optimal predictive model. In this part, we use random forest regression and gradient boosting for regression as ensemble regressors to find the best parameter set. In random forests, each tree in the ensemble is built from a sample drawn with replacement from the training set. Gradient boosting builds an additive model in a forward stage-wise fashion which allows for the optimization of arbitrary differentiable loss functions.

QUESTION 8: Use the following param grid. Analyze the result of the grid search. Do the test errors from cross-validation look good? If not, please explain the reason.

The best Random Forest Regressor has the parameters and cross-validation RMSE as below.

```
The best Random Forest Regressor for Aggregate Data is
RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=70,
                      max_features='auto', max_leaf_nodes=None,
                      min_impurity_decrease=0.0, min_impurity_split=None,
                      min_samples_leaf=1, min_samples_split=2,
                      min_weight_fraction_leaf=0.0, n_estimators=200,
                      n_jobs=None, oob_score=False, random_state=None,
                      verbose=0, warm_start=False)
The RMSE for this Estimator is 15751.87589168475
```

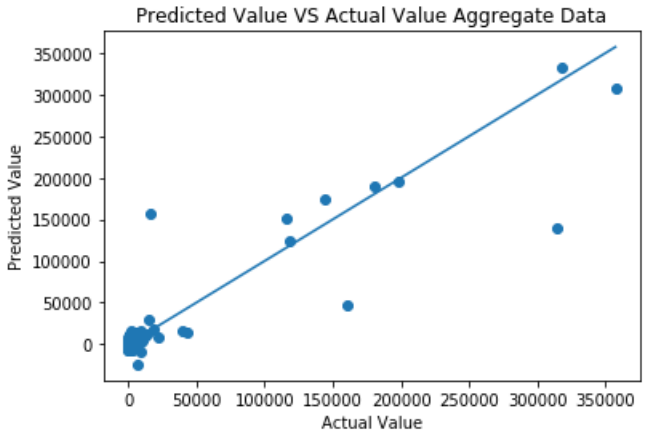
The best Gradient Boosting Regressor has the parameters and cross-validation RMSE as below.

```
The best Gradient Boosting Regressor for Aggregate Data is
GradientBoostingRegressor(alpha=0.9, criterion='friedman_mse', init=None,
                          learning_rate=0.1, loss='ls', max_depth=70,
                          max_features='sqrt', max_leaf_nodes=None,
                          min_impurity_decrease=0.0, min_impurity_split=None,
                          min_samples_leaf=1, min_samples_split=2,
                          min_weight_fraction_leaf=0.0, n_estimators=1200,
                          n_iter_no_change=None, presort='auto',
                          random_state=None, subsample=1.0, tol=0.0001,
                          validation_fraction=0.1, verbose=0, warm_start=False)
The RMSE for this Estimator is 16719.528283058004
```

This cross-validation RMSE has larger value than previous data. The first reason is because the data is not segregated into different time sections to improve the accuracy. The second reason lies within the cross validation method itself. K-fold cross validation tends to have a pessimistic bias and may overestimate generalization error. It gives an unbiased estimate of the generalization error for the surrogate model. As the error of the surrogate model decreases with increasing training sample, the model on average has higher true generalization error than the model trained on the whole data set, which is the model whose error is approximated by the cross validation.

QUESTION 9: Compare the best estimator you found in the grid search with OLS on the entire dataset.

```
RMSE is: 11398.27021655293
```



The RMSE for OLS is 11398.27021655293, which is smaller than the RMSE for Random Forest Regressor and Gradient Boosting Regressor. This does not necessarily mean the OLS outperforms the ensemble methods since cross-validation is not applied on the OLS regression. In general, the RMSE results for OLS and ensemble methods are similar.

QUESTION 10: Does the cross-validation test error change? Are the best parameter set you find in each period agree with those you found above?

```
The best Gradient Boosting Regressor for Aggregate Data Piece 1 is
GradientBoostingRegressor(alpha=0.9, criterion='friedman_mse', init=None,
                          learning_rate=0.1, loss='ls', max_depth=200,
                          max_features='sqrt', max_leaf_nodes=None,
                          min_impurity_decrease=0.0, min_impurity_split=None,
                          min_samples_leaf=1, min_samples_split=10,
                          min_weight_fraction_leaf=0.0, n_estimators=2000,
                          n_iter_no_change=None, presort='auto',
                          random_state=None, subsample=1.0, tol=0.0001,
                          validation_fraction=0.1, verbose=0, warm_start=False)
The RMSE for this Estimator is 2141.2479544016196

The best Gradient Boosting Regressor for Aggregate Data Piece 2 is
GradientBoostingRegressor(alpha=0.9, criterion='friedman_mse', init=None,
                          learning_rate=0.1, loss='ls', max_depth=100,
                          max_features='sqrt', max_leaf_nodes=None,
                          min_impurity_decrease=0.0, min_impurity_split=None,
                          min_samples_leaf=2, min_samples_split=2,
                          min_weight_fraction_leaf=0.0, n_estimators=1000,
                          n_iter_no_change=None, presort='auto',
                          random_state=None, subsample=1.0, tol=0.0001,
                          validation_fraction=0.1, verbose=0, warm_start=False)
The RMSE for this Estimator is 1014.152712497833

The best Gradient Boosting Regressor for Aggregate Data Piece 3 is
GradientBoostingRegressor(alpha=0.9, criterion='friedman_mse', init=None,
                          learning_rate=0.1, loss='ls', max_depth=70,
                          max_features='sqrt', max_leaf_nodes=None,
                          min_impurity_decrease=0.0, min_impurity_split=None,
                          min_samples_leaf=1, min_samples_split=5,
                          min_weight_fraction_leaf=0.0, n_estimators=1000,
                          n_iter_no_change=None, presort='auto',
                          random_state=None, subsample=1.0, tol=0.0001,
                          validation_fraction=0.1, verbose=0, warm_start=False)
The RMSE for this Estimator is 644.4819044650641
```

We performed grid search for each period, we showed that the cross-validation test error for separate time period data is smaller than performing grid search for aggregated data. It is because the trend within each period is more similar, and to predict within each period is more effective. The best parameter sets we find in each period are shown above. The parameters for the whole-piece data are different from the parameters for the piece-wise data. Especially for the whole-piece data, the number of estimators is much larger than the piece-wise data in order to reach its best performance. The whole piece dataset is larger and also the trend during the superbowl is difficult to characterize from the given parameters.

Question 11:

We regressed the aggregated data with MLPRegressor by trying out different architectures (i.e. the structure of the network) with various hidden layer sizes. We tried more than 6 architectures with various numbers of layers and layer sizes. Report the architectures you tried, as well as its MSE of fitting the entire aggregated data.

We pickle.load the previously pickle-dumped data of the aggregate twitter records. We tried six different architectures with different hidden layer lengths and sizes. The maximum iteration of the MLPRegressor is set to 10000 and 5-fold cross validation is performed for each architecture.

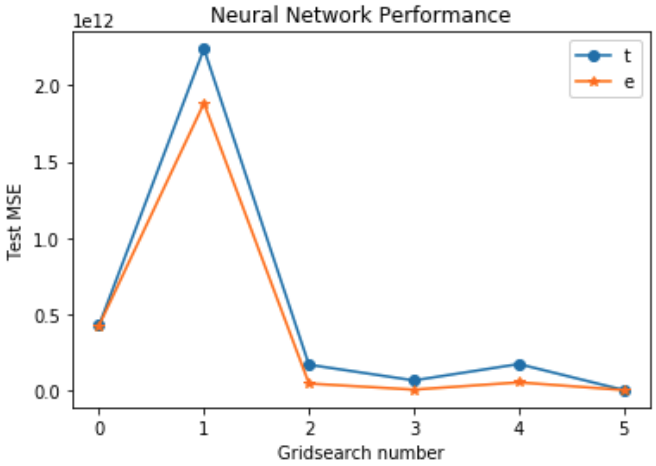
On Jupyter Notebook, we tried the following architectures in the grid search:

grid_NN = {'hidden_layer_sizes':[(200,200,), (300,300,300,), (400,400,400,400,),
(500,500,500,500,500,),(800,800,800,800,800,800,800,800,),
(100, 200, 300, 400, 500, 600, 700, 800, 700, 600, 500, 400, 300, 200, 100,)]
}

```
Best estimator for Neural Network :
 MLPRegressor(activation='relu', alpha=0.0001, batch_size='auto', beta_1=0.9,
             beta_2=0.999, early_stopping=False, epsilon=1e-08,
             hidden_layer_sizes=(100, 200, 300, 400, 500, 600, 700, 800, 700,
                                 600, 500, 400, 300, 200, 100),
             learning_rate='constant', learning_rate_init=0.001, max_iter=10000,
             momentum=0.9, n_iter_no_change=10, nesterovs_momentum=True,
             power_t=0.5, random_state=None, shuffle=True, solver='adam',
             tol=0.0001, validation_fraction=0.1, verbose=False,
             warm_start=False)
Best score for Neural Network :  522467217.406664
Best parameters for Neural Network :  {'hidden_layer_sizes': (100, 200, 300, 400, 500, 600, 7
00, 800, 700, 600, 500, 400, 300, 200, 100)}
```

The MSE of the best architecture (6th one) we have tried is 5.2e8 for fitting the entire aggregated data.

The following plot shows the test MSE for the six different architectures used here.



On Google Colab with GPU, we tried more combinations of the architectures:

```
grid_NN = {'hidden_layer_sizes':[(500,500,500,500,500,),
                                 (100,200,300,400,500,600,700,800,700,600,500,400,300,200,100,),
                                 (1100,1100,1100,1100,1100,1100,1100,1100,1100,1100,1100,)],
           'activation':['logistic', 'relu'],
           'solver':["sgd", "adam"],
           'alpha': [0.00005,0.05, 1]}
NN = MLPRegressor(max_iter = 6000)
```

```
Fitting 5 folds for each of 36 candidates, totalling 180 fits
[CV] activation=logistic, alpha=5e-05, hidden_layer_sizes=(500, 500, 500, 500, 500), solver=sgd
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[CV]  activation=logistic, alpha=5e-05, hidden_layer_sizes=(500, 500, 500, 500, 500), solver=sgd, total=   2.8s
[CV] activation=logistic, alpha=5e-05, hidden_layer_sizes=(500, 500, 500, 500, 500), solver=sgd
[Parallel(n_jobs=1)]: Done   1 out of   1 | elapsed:    2.8s remaining:    0.0s
[CV]  activation=logistic, alpha=5e-05, hidden_layer_sizes=(500, 500, 500, 500, 500), solver=sgd, total=   5.4s
[CV] activation=logistic, alpha=5e-05, hidden_layer_sizes=(500, 500, 500, 500, 500), solver=sgd
[CV]  activation=logistic, alpha=5e-05, hidden_layer_sizes=(500, 500, 500, 500, 500), solver=sgd, total=   3.4s
[CV] activation=logistic, alpha=5e-05, hidden_layer_sizes=(500, 500, 500, 500, 500), solver=sgd
[CV]  activation=logistic, alpha=5e-05, hidden_layer_sizes=(500, 500, 500, 500, 500), solver=sgd, total=   3.3s
[CV] activation=logistic, alpha=5e-05, hidden_layer_sizes=(500, 500, 500, 500, 500), solver=sgd
[CV]  activation=logistic, alpha=5e-05, hidden_layer_sizes=(500, 500, 500, 500, 500), solver=sgd, total=   5.5s
[CV] activation=logistic, alpha=5e-05, hidden_layer_sizes=(500, 500, 500, 500, 500), solver=adam
/usr/local/lib/python3.6/dist-packages/sklearn/neural_network/_multilayer_perceptron.py:571: ConvergenceWarning: Stochastic Optimizer: Maximum iterations (6000) reached and the optimization hasn't converged yet.
  % self.max_iter, ConvergenceWarning)
[CV]  activation=logistic, alpha=5e-05, hidden_layer_sizes=(500, 500, 500, 500, 500), solver=adam, total=19.7min
[CV] activation=logistic, alpha=5e-05, hidden_layer_sizes=(500, 500, 500, 500, 500), solver=adam
/usr/local/lib/python3.6/dist-packages/sklearn/neural_network/_multilayer_perceptron.py:571: ConvergenceWarning: Stochastic Optimizer: Maximum iterations (6000) reached and the optimization hasn't converged yet.
  % self.max_iter, ConvergenceWarning)
[CV]  activation=logistic, alpha=5e-05, hidden_layer_sizes=(500, 500, 500, 500, 500), solver=adam, total=20.0min
[CV] activation=logistic, alpha=5e-05, hidden_layer_sizes=(500, 500, 500, 500, 500), solver=adam
/usr/local/lib/python3.6/dist-packages/sklearn/neural_network/_multilayer_perceptron.py:571: ConvergenceWarning: Stochastic Optimizer: Maximum iterations (6000) reached and the optimization hasn't converged yet.
  % self.max_iter, ConvergenceWarning)
[CV]  activation=logistic, alpha=5e-05, hidden_layer_sizes=(500, 500, 500, 500, 500), solver=adam, total=19.2min
[CV] activation=logistic, alpha=5e-05, hidden_layer_sizes=(500, 500, 500, 500, 500), solver=adam
/usr/local/lib/python3.6/dist-packages/sklearn/neural_network/_multilayer_perceptron.py:571: ConvergenceWarning: Stochastic Optimizer: Maximum iterations (6000) reached and the optimization hasn't converged yet.
  % self.max_iter, ConvergenceWarning)
[CV]  activation=logistic, alpha=5e-05, hidden_layer_sizes=(500, 500, 500, 500, 500), solver=adam, total=18.9min
[CV] activation=logistic, alpha=5e-05, hidden_layer_sizes=(500, 500, 500, 500, 500), solver=adam
/usr/local/lib/python3.6/dist-packages/sklearn/neural_network/_multilayer_perceptron.py:571: ConvergenceWarning: Stochastic Optimizer: Maximum iterations (6000) reached and the optimization hasn't converged yet.
  % self.max_iter, ConvergenceWarning)
[CV]  activation=logistic, alpha=5e-05, hidden_layer_sizes=(500, 500, 500, 500, 500), solver=adam, total=18.8min
[CV] activation=logistic, alpha=5e-05, hidden_layer_sizes=(100, 200, 300, 400, 500, 600, 700, 800, 700, 600, 500, 400, 300, 200, 100), solver=sgd
[CV]  activation=logistic, alpha=5e-05, hidden_layer_sizes=(100, 200, 300, 400, 500, 600, 700, 800, 700, 600, 500, 400, 300, 200, 100), solver=sgd, total=  15.9s
[CV] activation=logistic, alpha=5e-05, hidden_layer_sizes=(100, 200, 300, 400, 500, 600, 700, 800, 700, 600, 500, 400, 300, 200, 100), solver=sgd
[CV]  activation=logistic, alpha=5e-05, hidden_layer_sizes=(100, 200, 300, 400, 500, 600, 700, 800, 700, 600, 500, 400, 300, 200, 100), solver=sgd, total=   7.6s
[CV] activation=logistic, alpha=5e-05, hidden_layer_sizes=(100, 200, 300, 400, 500, 600, 700, 800, 700, 600, 500, 400, 300, 200, 100), solver=sgd
[CV]  activation=logistic, alpha=5e-05, hidden_layer_sizes=(100, 200, 300, 400, 500, 600, 700, 800, 700, 600, 500, 400, 300, 200, 100), solver=sgd, total=  12.0s
[CV] activation=logistic, alpha=5e-05, hidden_layer_sizes=(100, 200, 300, 400, 500, 600, 700, 800, 700, 600, 500, 400, 300, 200, 100), solver=sgd
[CV]  activation=logistic, alpha=5e-05, hidden_layer_sizes=(100, 200, 300, 400, 500, 600, 700, 800, 700, 600, 500, 400, 300, 200, 100), solver=sgd, total=  14.1s
[CV] activation=logistic, alpha=5e-05, hidden_layer_sizes=(100, 200, 300, 400, 500, 600, 700, 800, 700, 600, 500, 400, 300, 200, 100), solver=sgd
[CV]  activation=logistic, alpha=5e-05, hidden_layer_sizes=(100, 200, 300, 400, 500, 600, 700, 800, 700, 600, 500, 400, 300, 200, 100), solver=sgd, total=   9.4s
```

```
[CV]  activation=logistic, alpha=5e-05, hidden_layer_sizes=(100, 200, 300, 400, 500, 600, 700, 800, 700, 600, 500, 400, 300, 200, 100), solver=adam, total=61.9min
[CV] activation=logistic, alpha=5e-05, hidden_layer_sizes=(100, 200, 300, 400, 500, 600, 700, 800, 700, 600, 500, 400, 300, 200, 100), solver=adam
/usr/local/lib/python3.6/dist-packages/sklearn/neural_network/_multilayer_perceptron.py:571: ConvergenceWarning: Stochastic Optimizer: Maximum iterations (6000) reached and the optimization hasn't converged yet.
  % self.max_iter, ConvergenceWarning)
[CV]  activation=logistic, alpha=5e-05, hidden_layer_sizes=(100, 200, 300, 400, 500, 600, 700, 800, 700, 600, 500, 400, 300, 200, 100), solver=adam, total=63.7min
[CV] activation=logistic, alpha=5e-05, hidden_layer_sizes=(100, 200, 300, 400, 500, 600, 700, 800, 700, 600, 500, 400, 300, 200, 100), solver=adam
/usr/local/lib/python3.6/dist-packages/sklearn/neural_network/_multilayer_perceptron.py:571: ConvergenceWarning: Stochastic Optimizer: Maximum iterations (6000) reached and the optimization hasn't converged yet.
  % self.max_iter, ConvergenceWarning)
[CV]  activation=logistic, alpha=5e-05, hidden_layer_sizes=(100, 200, 300, 400, 500, 600, 700, 800, 700, 600, 500, 400, 300, 200, 100), solver=adam, total=63.8min
[CV] activation=logistic, alpha=5e-05, hidden_layer_sizes=(100, 200, 300, 400, 500, 600, 700, 800, 700, 600, 500, 400, 300, 200, 100), solver=adam
/usr/local/lib/python3.6/dist-packages/sklearn/neural_network/_multilayer_perceptron.py:571: ConvergenceWarning: Stochastic Optimizer: Maximum iterations (6000) reached and the optimization hasn't converged yet.
  % self.max_iter, ConvergenceWarning)
[CV]  activation=logistic, alpha=5e-05, hidden_layer_sizes=(100, 200, 300, 400, 500, 600, 700, 800, 700, 600, 500, 400, 300, 200, 100), solver=adam, total=70.4min
[CV] activation=logistic, alpha=5e-05, hidden_layer_sizes=(100, 200, 300, 400, 500, 600, 700, 800, 700, 600, 500, 400, 300, 200, 100), solver=adam
/usr/local/lib/python3.6/dist-packages/sklearn/neural_network/_multilayer_perceptron.py:571: ConvergenceWarning: Stochastic Optimizer: Maximum iterations (6000) reached and the optimization hasn't converged yet.
  % self.max_iter, ConvergenceWarning)
[CV]  activation=logistic, alpha=5e-05, hidden_layer_sizes=(100, 200, 300, 400, 500, 600, 700, 800, 700, 600, 500, 400, 300, 200, 100), solver=adam, total=64.9min
[CV] activation=logistic, alpha=5e-05, hidden_layer_sizes=(1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100), solver=sgd
[CV]  activation=logistic, alpha=5e-05, hidden_layer_sizes=(1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100), solver=sgd, total=  25.7s
[CV] activation=logistic, alpha=5e-05, hidden_layer_sizes=(1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100), solver=sgd
[CV]  activation=logistic, alpha=5e-05, hidden_layer_sizes=(1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100), solver=sgd, total= 1.2min
[CV] activation=logistic, alpha=5e-05, hidden_layer_sizes=(1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100), solver=sgd
[CV]  activation=logistic, alpha=5e-05, hidden_layer_sizes=(1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100), solver=sgd, total= 1.1min
[CV] activation=logistic, alpha=5e-05, hidden_layer_sizes=(1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100), solver=sgd
[CV]  activation=logistic, alpha=5e-05, hidden_layer_sizes=(1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100), solver=sgd, total=  24.7s
[CV] activation=logistic, alpha=5e-05, hidden_layer_sizes=(1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100), solver=sgd
[CV]  activation=logistic, alpha=5e-05, hidden_layer_sizes=(1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100), solver=sgd, total=  47.9s
[CV] activation=logistic, alpha=5e-05, hidden_layer_sizes=(1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100, 1100), solver=adam
```

Due to time constraint, we did not finish running all the 36 architecture combinations. Since the question asks for >=5, we will continue using the result ran from local in Jupyter notebook.

Question 12:

We use the StandardScaler to scale the features before feeding it to MLPRegressor (with the best architecture we have from 11). The MSE drops and the performance increases.

Training MSE using StandardScaler with the best architecture in **11**.: `2298577.66609028 (2.3e6)`

Question 13:

Using grid search, and the best architecture (for scaled data) for each period (with corresponding window length) described in **6**.

1) **Before Feb.1, 8:00 am.: 1-hour window:**

```
GridSearchCV(cv=KFold(n_splits=5, random_state=None, shuffle=True),
             error_score='raise-deprecating',
             estimator=MLPRegressor(activation='relu', alpha=0.0001,
                                    batch_size='auto', beta_1=0.9, beta_2=0.999,
                                    early_stopping=False, epsilon=1e-08,
                                    hidden_layer_sizes=(100,),
                                    learning_rate='constant',
                                    learning_rate_init=0.001, max_iter=10000,
                                    momentum=0.9, n_iter_no_change=1...
                                    validation_fraction=0.1, verbose=False,
                                    warm_start=False),
             iid='warn', n_jobs=1,
             param_grid={'hidden_layer_sizes': [(200, 200), (300, 300, 300),
                                                (400, 400, 400, 400),
                                                (500, 500, 500, 500, 500),
                                                (800, 800, 800, 800, 800, 800,
                                                 800, 800),
                                                (100, 200, 300, 400, 500, 600,
                                                 700, 800, 700, 600, 500, 400,
                                                 300, 200, 100)]},
             pre_dispatch='2*n_jobs', refit=True, return_train_score=False,
             scoring='neg_mean_squared_error', verbose=2)

The best MLPRegressor for Aggregate Data Piece 1 is
MLPRegressor(activation='relu', alpha=0.0001, batch_size='auto', beta_1=0.9,
             beta_2=0.999, early_stopping=False, epsilon=1e-08,
             hidden_layer_sizes=(100, 200, 300, 400, 500, 600, 700, 800, 700,
                                 600, 500, 400, 300, 200, 100),
             learning_rate='constant', learning_rate_init=0.001, max_iter=10000,
             momentum=0.9, n_iter_no_change=10, nesterovs_momentum=True,
             power_t=0.5, random_state=None, shuffle=True, solver='adam',
             tol=0.0001, validation_fraction=0.1, verbose=False,
             warm_start=False)
The MSE for this Estimator is 5315716.9545163
The RMSE for this Estimator is 2305.5838641255928
```

2) **Between 2/1 8 am—8pm: 5-min window:**

```
GridSearchCV(cv=KFold(n_splits=5, random_state=None, shuffle=True),
             error_score='raise-deprecating',
             estimator=MLPRegressor(activation='relu', alpha=0.0001,
                                    batch_size='auto', beta_1=0.9, beta_2=0.999,
                                    early_stopping=False, epsilon=1e-08,
                                    hidden_layer_sizes=(100,),
                                    learning_rate='constant',
                                    learning_rate_init=0.001, max_iter=10000,
                                    momentum=0.9, n_iter_no_change=1...
                                    validation_fraction=0.1, verbose=False,
                                    warm_start=False),
             iid='warn', n_jobs=-1,
             param_grid={'hidden_layer_sizes': [(200, 200), (300, 300, 300),
                                                (400, 400, 400, 400),
                                                (500, 500, 500, 500, 500),
                                                (800, 800, 800, 800, 800, 800,
                                                 800, 800),
                                                (100, 200, 300, 400, 500, 600,
                                                 700, 800, 700, 600, 500, 400,
                                                 300, 200, 100)]},
             pre_dispatch='2*n_jobs', refit=True, return_train_score=False,
             scoring='neg_mean_squared_error', verbose=2)
```

```
The best MLPRegressor for Aggregate Data Piece 2 is
MLPRegressor(activation='relu', alpha=0.0001, batch_size='auto', beta_1=0.9,
             beta_2=0.999, early_stopping=False, epsilon=1e-08,
             hidden_layer_sizes=(800, 800, 800, 800, 800, 800, 800, 800),
             learning_rate='constant', learning_rate_init=0.001, max_iter=10000,
             momentum=0.9, n_iter_no_change=10, nesterovs_momentum=True,
             power_t=0.5, random_state=None, shuffle=True, solver='adam',
             tol=0.0001, validation_fraction=0.1, verbose=False,
             warm_start=False)
The MSE for this Estimator is 1445256.2941616771
The RMSE for this Estimator is 1202.18812760802
```

3) **After 2/1 8pm.: 1-hour window**

```
GridSearchCV(cv=KFold(n_splits=5, random_state=None, shuffle=True),
             error_score='raise-deprecating',
             estimator=MLPRegressor(activation='relu', alpha=0.0001,
                                    batch_size='auto', beta_1=0.9, beta_2=0.999,
                                    early_stopping=False, epsilon=1e-08,
                                    hidden_layer_sizes=(100,),
                                    learning_rate='constant',
                                    learning_rate_init=0.001, max_iter=10000,
                                    momentum=0.9, n_iter_no_change=1...
                                    validation_fraction=0.1, verbose=False,
                                    warm_start=False),
             iid='warn', n_jobs=1,
             param_grid={'hidden_layer_sizes': [(200, 200), (300, 300, 300),
                                                (400, 400, 400, 400),
                                                (500, 500, 500, 500, 500),
                                                (800, 800, 800, 800, 800, 800,
                                                 800, 800),
                                                (100, 200, 300, 400, 500, 600,
                                                 700, 800, 700, 600, 500, 400,
                                                 300, 200, 100)]},
             pre_dispatch='2*n_jobs', refit=True, return_train_score=False,
             scoring='neg_mean_squared_error', verbose=2)
```

```
The best MLPRegressor for Aggregate Data Piece 3 is
MLPRegressor(activation='relu', alpha=0.0001, batch_size='auto', beta_1=0.9,
             beta_2=0.999, early_stopping=False, epsilon=1e-08,
             hidden_layer_sizes=(800, 800, 800, 800, 800, 800, 800, 800),
             learning_rate='constant', learning_rate_init=0.001, max_iter=10000,
             momentum=0.9, n_iter_no_change=10, nesterovs_momentum=True,
             power_t=0.5, random_state=None, shuffle=True, solver='adam',
             tol=0.0001, validation_fraction=0.1, verbose=False,
             warm_start=False)
The MSE for this Estimator is 2672624.7881032173
The RMSE for this Estimator is 1634.8164386570186
```

Period	Neural Network Architecture	MSE	RMSE
Before Feb. 1, 8:00 a.m.	hidden_layer_sizes=(100, 200, 300, 400, 500, 600, 700, 800, 700, 600, 500, 400, 300, 200, 100)	5315716.9545163	2305.5838641255928
Between Feb. 1, 8:00 a.m. and 8:00 p.m.	hidden_layer_sizes=(800, 800, 800, 800, 800, 800, 800, 800),	1445256.2941616771	1202.18812760802
After Feb. 1, 8:00 p.m.	hidden_layer_sizes=(800, 800, 800, 800, 800, 800, 800, 800),	2672624.7881032173	1634.8164386570186

Question 14

Based on discussion in Piazza, we interpret the problem to be asking that we should construct new features based on a 6 windows, 1 current and 5 priori to predict the number of tweets for next window length. We should train and find some best methods for this new crafted feature and test them on the test files, with each as a single test datapoint and require one prediction on the number of tweets for the corresponding window length.

With the question being clarified, we report our results to be:

```
For the test file: sample0_period1.txt
the predicted tweets for the next hour are: 127.42883940120839

For the test file: sample0_period3.txt
the predicted tweets for the next hour are: 124.84705528765721

For the test file: sample1_period1.txt
the predicted tweets for the next hour are: 787.7368102875156

For the test file: sample1_period3.txt
the predicted tweets for the next hour are: 173.37582544704387

For the test file: sample2_period1.txt
the predicted tweets for the next hour are: 57.31901706317703

For the test file: sample2_period3.txt
the predicted tweets for the next hour are: -15.827906607588769

For the test file: sample0_period2.txt
the predicted tweets for the next 5 minutes are: 2261.747988145746

For the test file: sample1_period2.txt
the predicted tweets for the next 5 minutes are: 1537.280692807896

For the test file: sample2_period2.txt
the predicted tweets for the next 5 minutes are: 174.3086146579621
```

There should be no training on the test data, so we used the same method for each period on all sample files.
The methods are:

Period	Method	RMSE (5-fold CV)
1	Linear Regression	2375.03
2	Gradient Boosting Regressor	998.1103604990587
3	Linear Regression	565.87

The parameter for the best gradient boosting regressor is:

```
GradientBoostingRegressor(alpha=0.9, criterion='friedman_mse',
                  init=None, learning_rate=0.1, loss='ls',
                  max_depth=100, max_features='sqrt',
                  max_leaf_nodes=None, min_impurity_decrease=0.0,
                  min_impurity_split=None, min_samples_leaf=1,
                  min_samples_split=2, min_weight_fraction_leaf=0.0,
                  n_estimators=800, n_iter_no_change=None, random_state=None,
                  subsample=1.0, tol=0.0001, validation_fraction=0.1,
                  verbose=0, warm_start=False)
```

It should be mentioned that we did not include any time data in our final version of the crafted new feature. The one-hot encoded hour/minute data are shown to be rather weak features in previous sections and the inclusion of hour/minute data does not improve the test RMSE metric. While it does contribute to worsened overfitting, we therefore decided to remove the time data entirely.

Question 15

1). The method we used is captured in the helper function "def parse_locale(location)".

Since there are no coordinates data available, we have to rely on the location metadata. To parse the data, we chose to use regular expressions.

For Washington state, we match the string with patterns "Washington" and "WA" that are either non-first address as in " xx_city, Washington" or the last words of the string. We also have to exclude any matches that are further matched with patterns "DC" or "D.C.".

Though, by excluding "DC" and "D.C.", there is no longer confusion with other state names, we have to consider cities named after Washington as well, hence the restriction to be non-first address or being the last word.

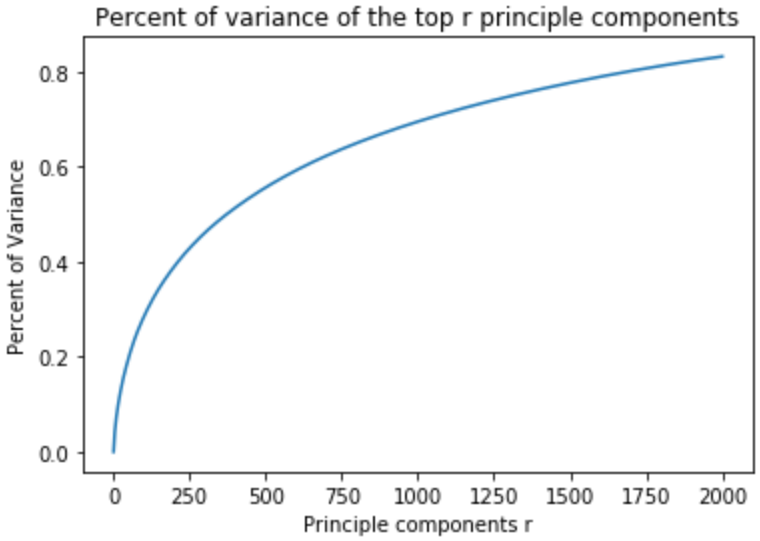
We did the same for Massachusetts, save for the "DC" part.

2). Between choosing the optimal number of principal components for dimensional reduction and fitting the best parameters for each classifier, we used 3 simpler models, Naive Bayesian, SVM and Logistic Regression to reduce runtime.

The tf-idf matrix extracted from the textual data corresponding to the located users has the shape:

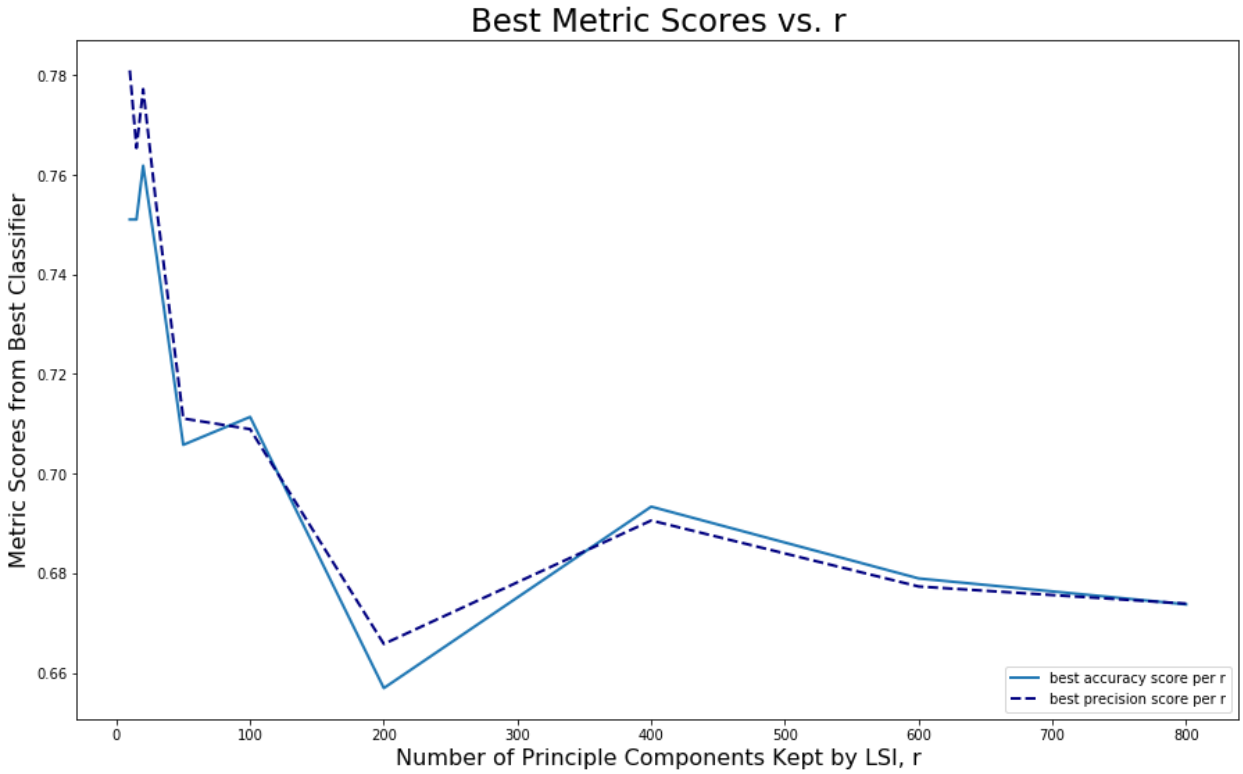
(24975, 5919)

In this part we used LSI for dimension reduction. The overall information retained is plotted in the graph below:



We started with Naive Bayesian, as the classifier itself does not require parameter fitting, and we can scan for a broad range of r, number of principal components.

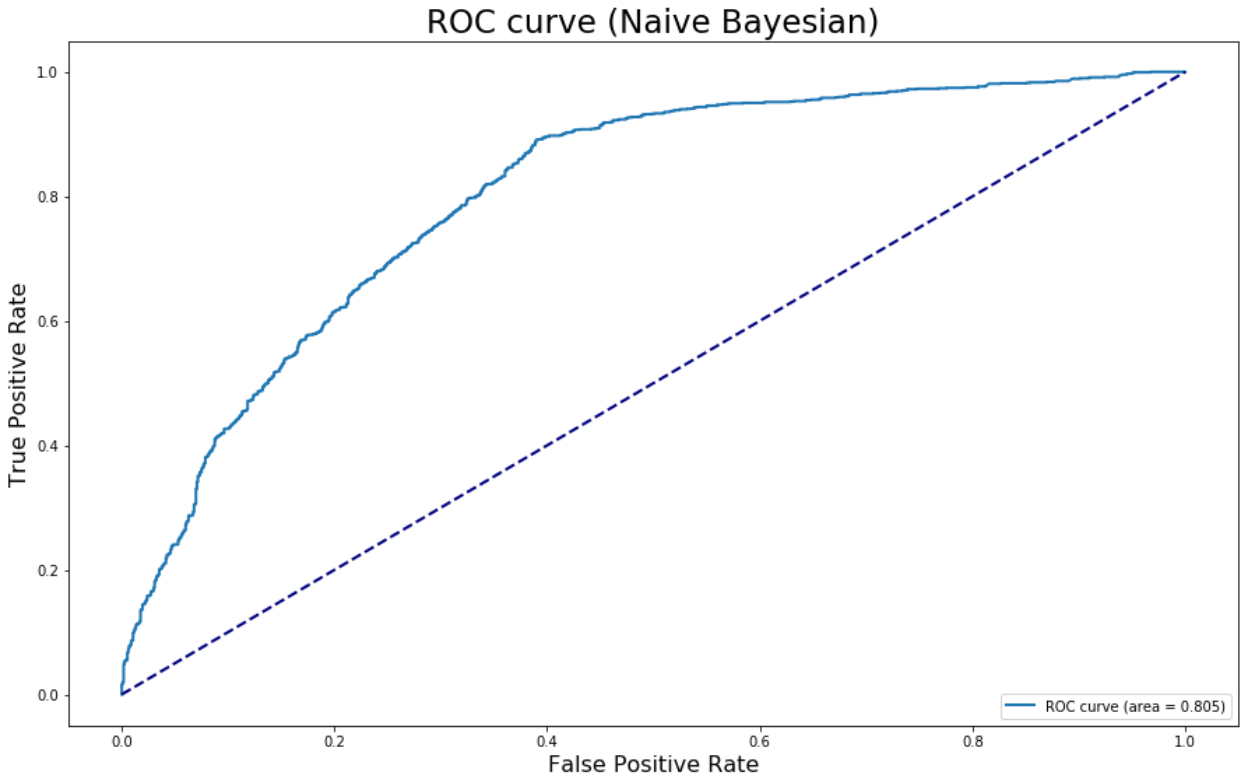
The performance of Bayesian with varying r is plotted below:



By this graph we recognize that r = 20 yields the best results, which report as:

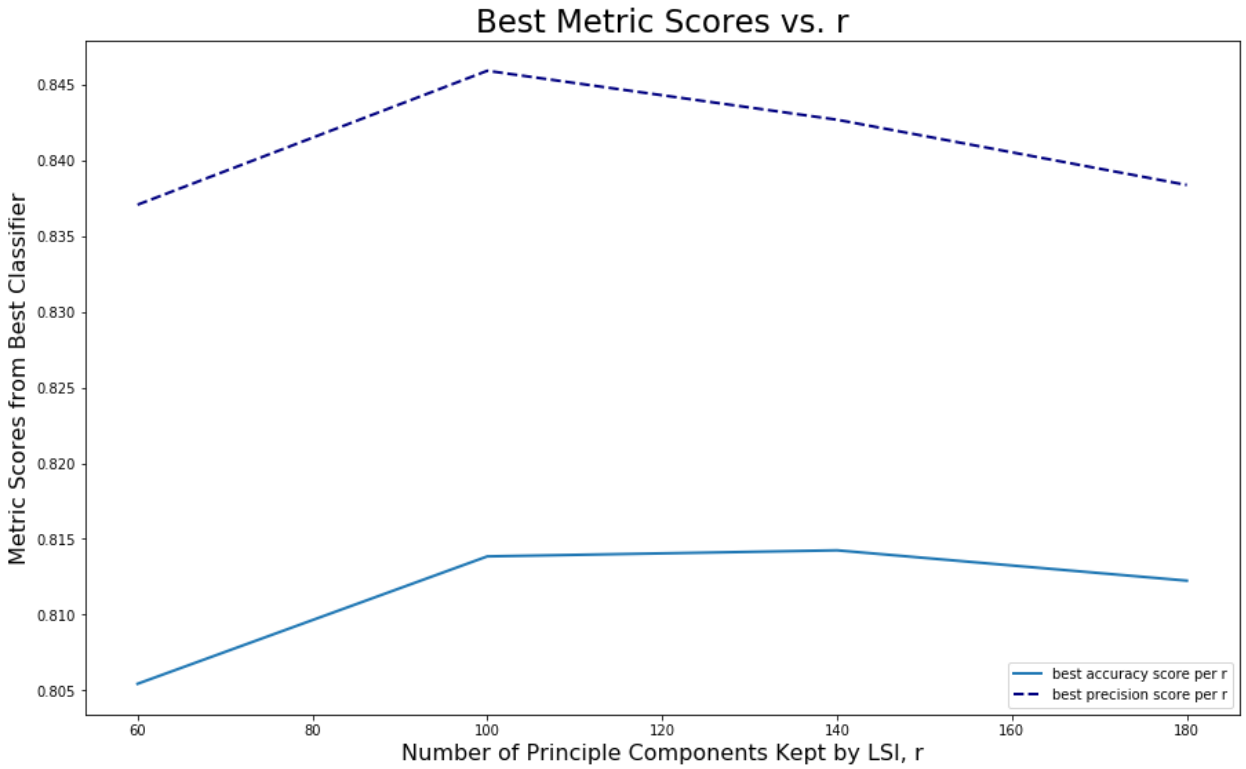
```
Using LSI(20 components) and Naive Bayesian:
{}
The accuracy is  0.7618094475580465
The precision is  0.7771792211954721
The recall is  0.7483134063500768
The confusion matrix is
[[ 684  448]
 [ 147 1219]]
```

and the associated ROC curve is plotted as:



Next we trained a Logistic Regression model and searched the best possible one using gridSearchCV() by varying the inverse regularization factor C.

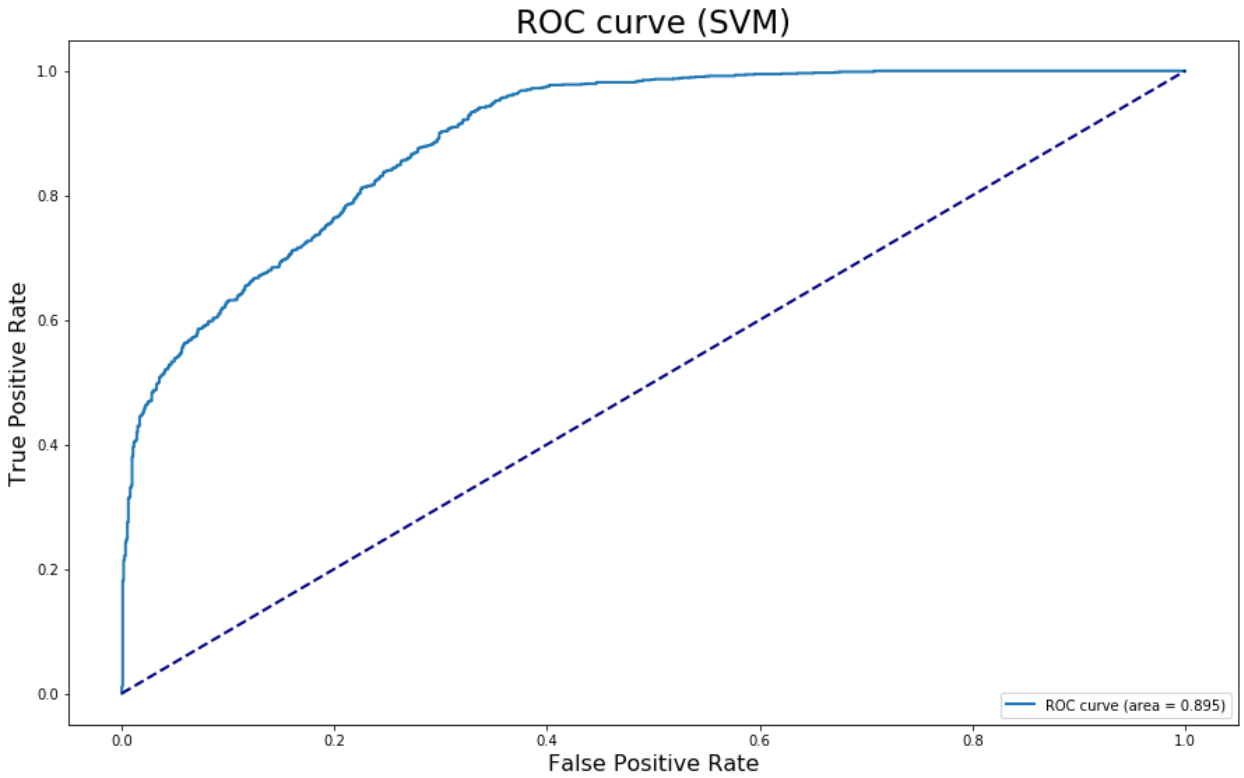
The performance of Logistic Regression with best C (per r) and varying r is plotted below:



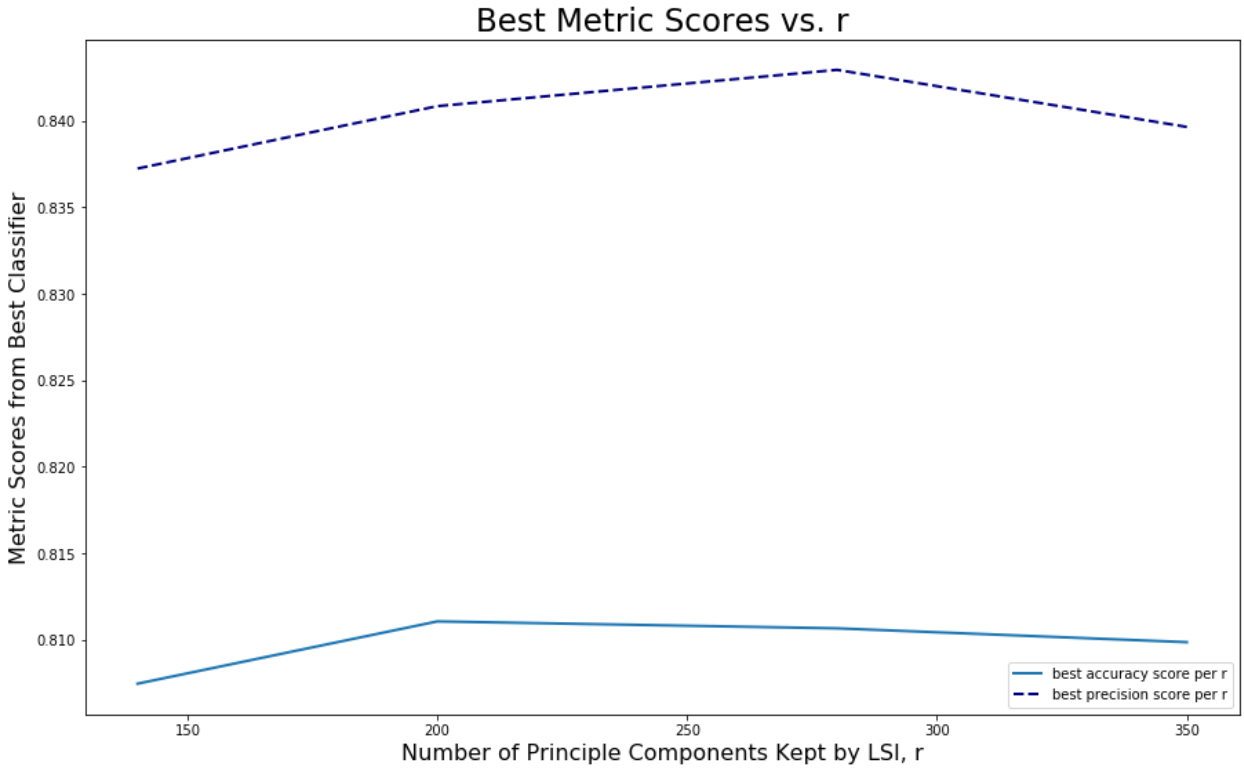
We choose 100 as the best r and the corresponding results are:

```
Using LSI(100 components) and SVM:
{'C': 1000}
The accuracy is  0.8138510808646917
The precision is  0.8458925729691273
The recall is  0.7986971581414359
The confusion matrix is
[[ 721  411]
 [  54 1312]]
```

C = 1000, is the best setup for r = 100. The associated ROC curve is plotted as:



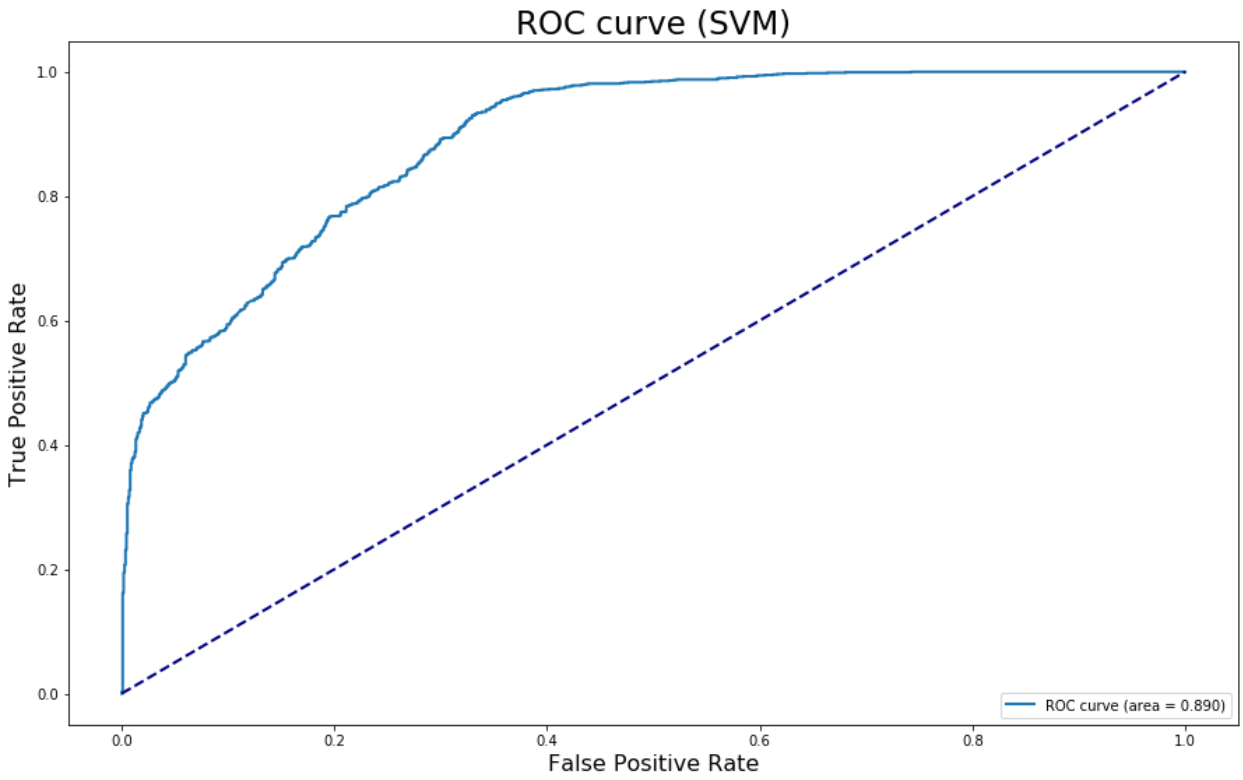
Finally we trained a SVM classifier, and the best r tested can be found by the following plot:



With r = 200, the results are reported as:

```
Using LSI(280 components) and SVM:
{'C': 5000}
The accuracy is  0.8106485188150521
The precision is  0.8429539390034074
The recall is  0.7953149170413216
The confusion matrix is
[[ 715  417]
 [  56 1310]]
```

the best inverse regularization factor,C, is 5000 and ROC curve is plotted below:



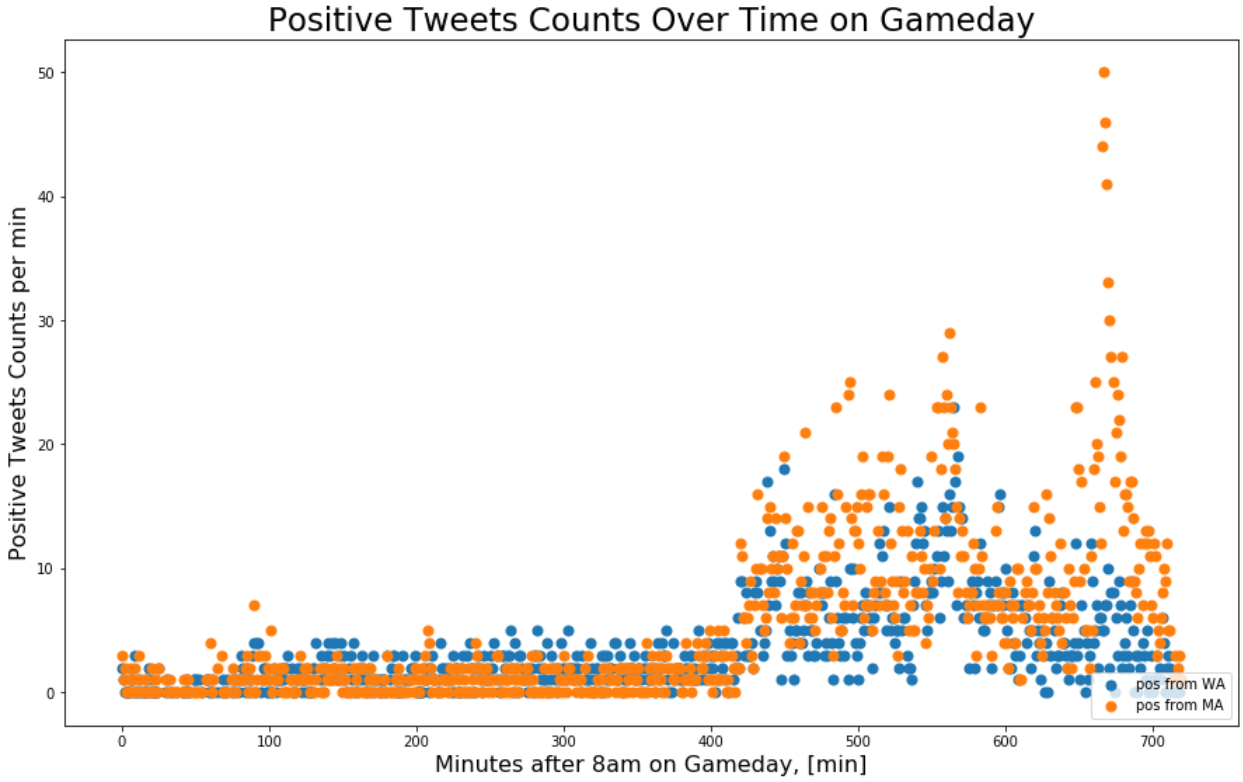
Question 16

In this part we start by analysing sentiments of tweets for the potential fans of the competing teams on game day (period 2 from Question 8) on the #superbowl dataset. Then try to see whether the sentiments are related to the popularity of the tweets'.

We expect to see the sentiment changes correspondingly as the match progresses and therefore confirming that indeed the fanbase of the competing teams can be inferred from their locations.

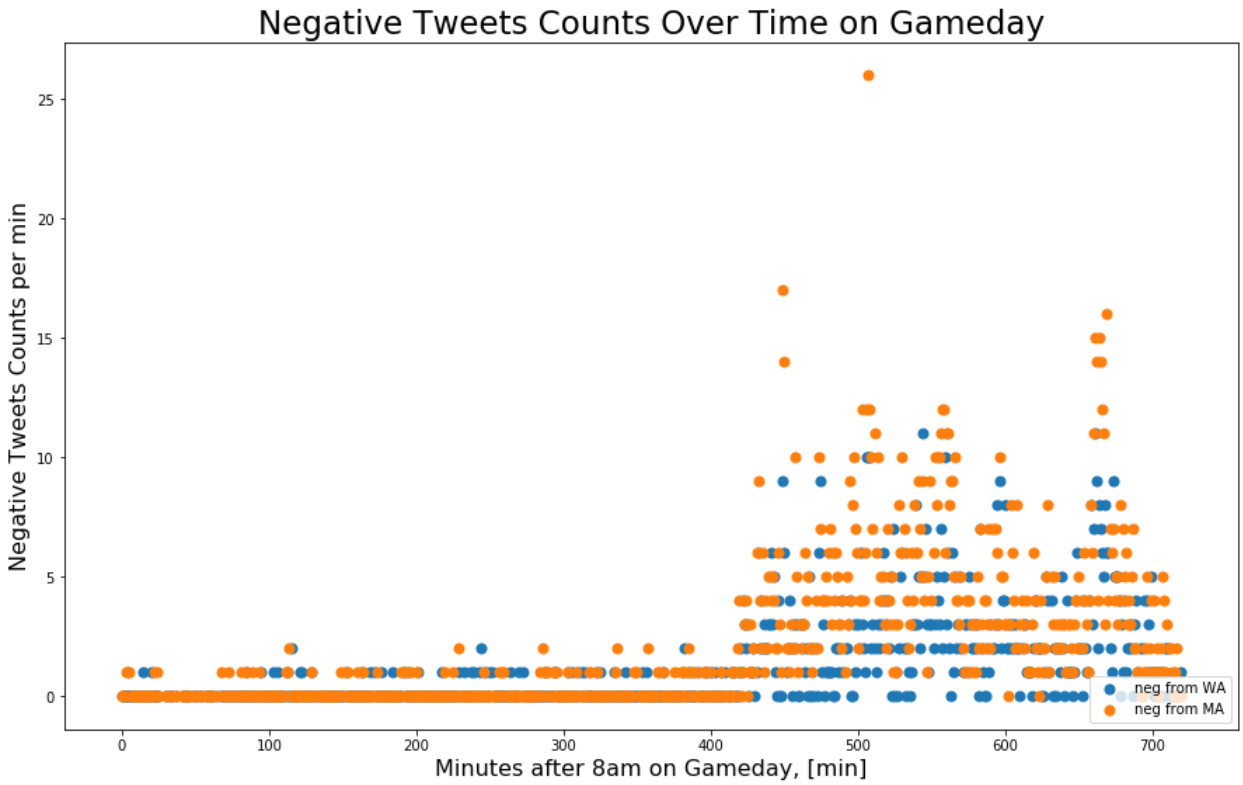
We extract all tweets on Feb. 1st from 8a.m to 8p.m by all users found in WA or MA in question 15. We then process and perform sentiment analysis on this dataset using vaderSentiment package.

We plotted number of positive and negative tweet per minute from users in WA and MA in the following two graphs:



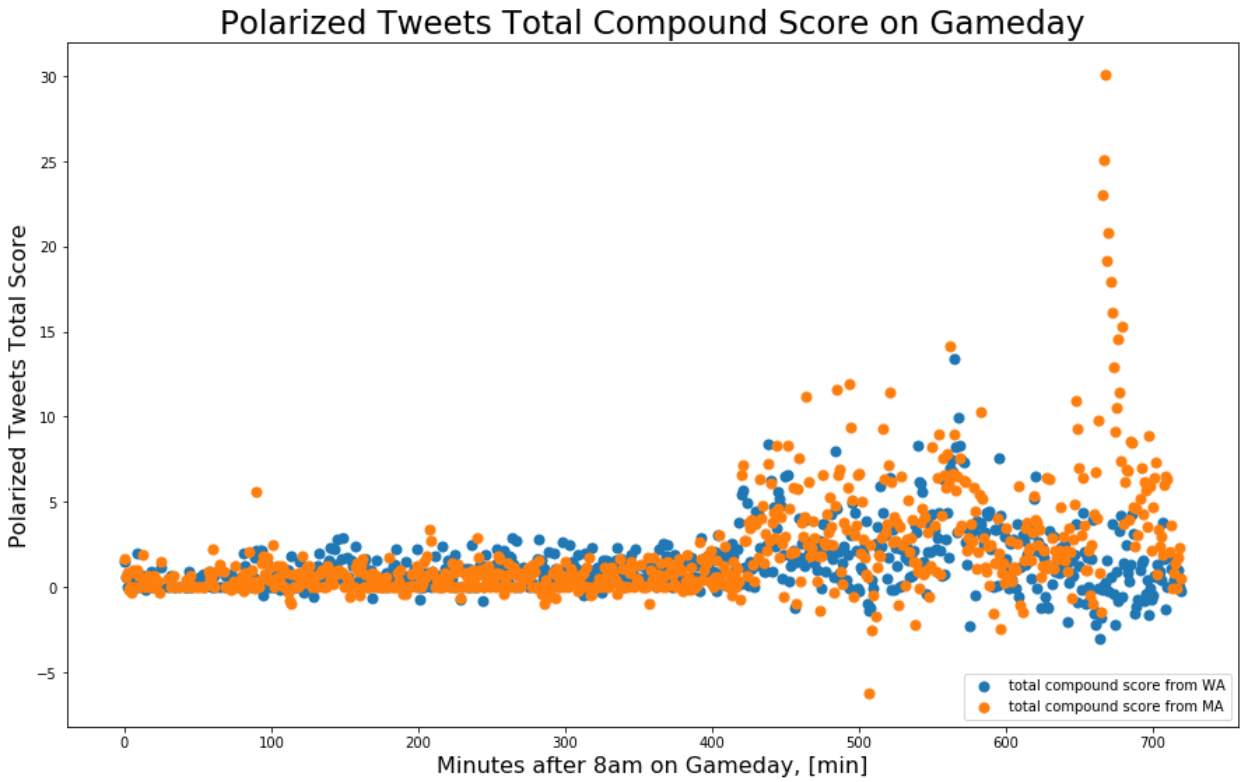
The vertical axis labels the number of positive tweets. The horizontal axis shows the number of minutes past after 8am (PST) on Feb. 1st, 2015. It should be noted that the game started at 3:30pm (PST), or 450 minutes after 8am, where we can see a very substantial rise in the number polarized tweets, positive ones, for the plot above.

The winning team for 2015 was from MA, shown in the orange dots. The fans for the winning sides have more positive tweets than those of the losing one and noticeably there was a burst of positive tweets from MA around 7:20pm (PST), probably due to the news spreading or the team returning from Arazona.



This graph here shows the number of negative tweets per minute on the game day. While we observe that the active time period overlaps with the burst of positive ones, the most interesting thing to notice is that fans from MA, the winning side, actually sent out more negative tweets as well. This can perhaps be understood by the observation that the peaks are actually acting in response to the previous one, as fans would argue with people from the opposing side.

The analyser we used rely on compound score to classify polarity as negative or positive. A score above 0.05 is classified as positive and below -0.05 is classified as negative. We then proceed to add up all compound scores on both sides per minute to observe the overall "mood" of the fanbase. The result is plotted below:

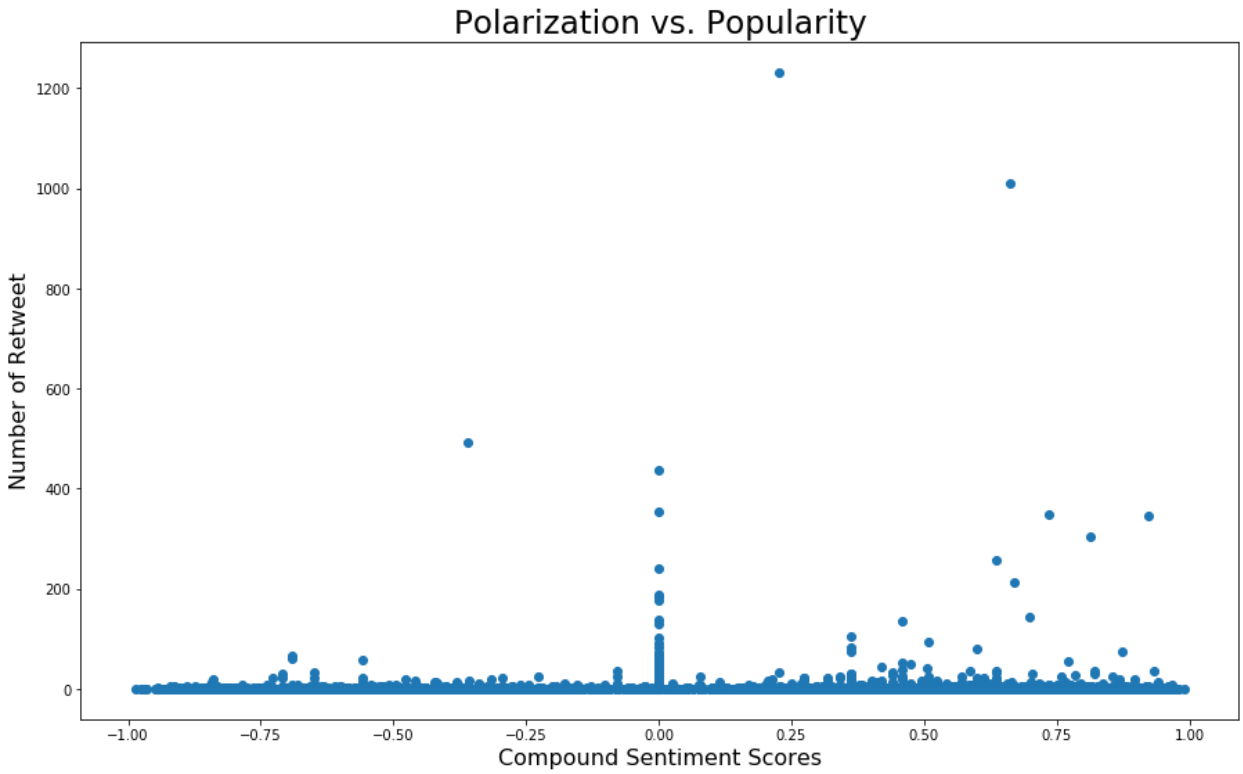


The vertical axis now shows the total compound score per minute. The more positive the total score is, the "happier" the fanbase is.

The total score plot better captures the sentiment changes as we can clearly observe the dramatic shifts in crowd mood corresponding to how the game developed. Most obviously at around the 510 minutes mark, or about 1hr since the game started, the fans from MA (orange) are at their most negative, which roughly corresponds to the largest score lead (24-14) by the opposing team.

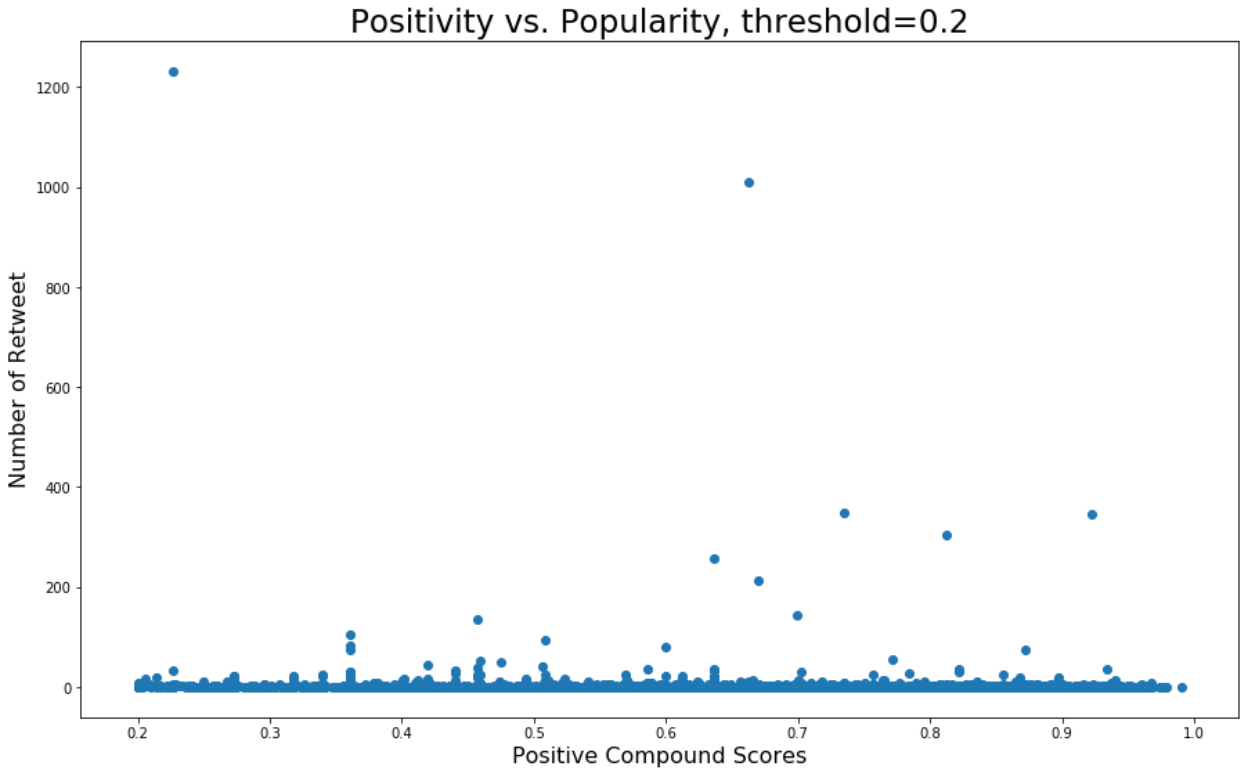
The clear shifts in the sentiments also serves to confirm that indeed the data we extracted for question 15 may well belong to the fans of these 2 teams.

Now onto popularity. First we plot a scatter graph for the number of retweets vs. compound scores to see if there are any correlations.



The graph suggests that a strongly positive tweet might be more likely to be retweeted.

At a closer look by the following graph:



This graph actually shows that though positivity might be related to popularity, due to the fact that most of the tweets were never retweeted, we cannot observe a decent correlation between these 2 metrics.

So instead of relying solely on positivity to predict popularity of a tweet, we now want to check whether positivity can be crafted with some intuitively popular indicator into a better feature for making predictions.

For this purpose, we only need to see improvements, so we chose the relatively simple Logistic Regression method. To avoid the dataset being overwhelmingly labeled as "unpopular", we set the threshold for being "popular" to 1 retweet, i.e. so long as a tweet is being retweeted, it will be categorized as "popular".

We used the data extracted from the fanbase from MA and WA on gameday.

The results are summarized below:

Features	Average Accuracy, Average Precision, CV-10
Follower Counts	0.82, 0.82
Follower Counts, Positivity (>0.5)	0.82, 0.81
Follower Counts, Friend Counts, Favorite Counts	0.82, 0.80
Follower Counts, Friend Counts, Favorite Counts Positivity (>0.5)	0.82, 0.80

Unfortunately, we were not able to establish correlation between polarity of a tweet and its popularity, nor can we use the polarity to improve upon natural popularity indicating features. But playing around with vaderSentiment analyzer, we did realize that sentiment analysis is a good method to obtain the opinion or "mood" of groups.