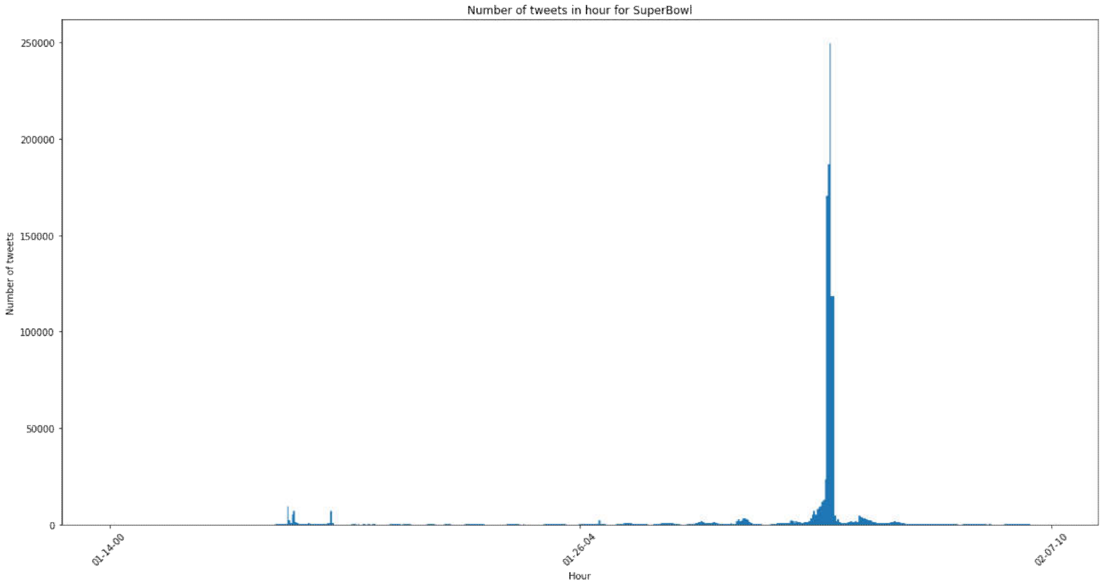
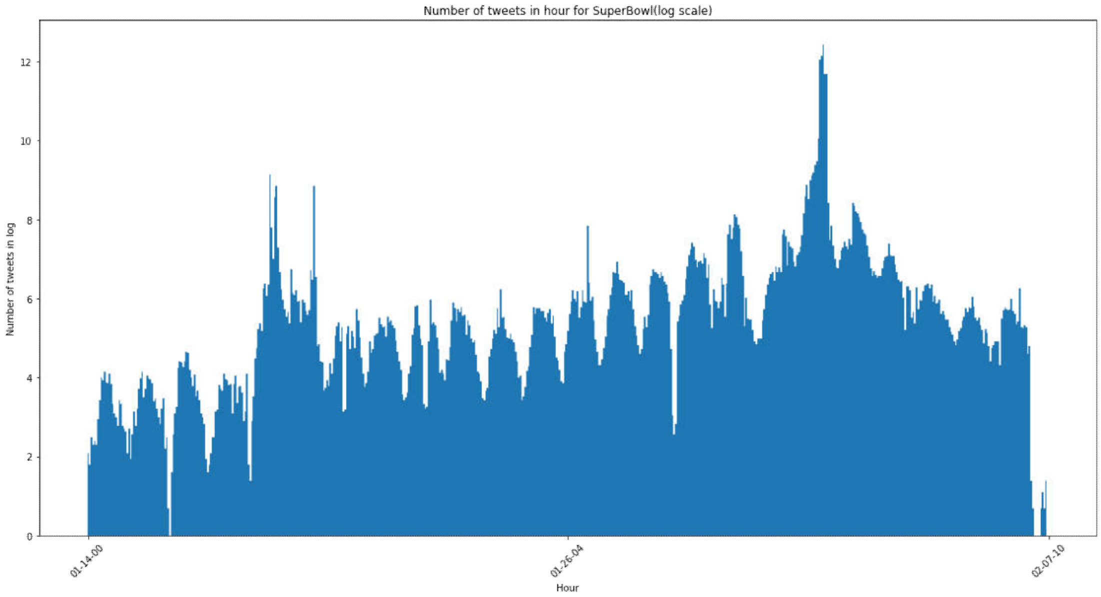
QUESTION 1: Report the following statistics for each hashtag

Hashtag	Average number of tweets per hour	Average number of followers of users posting the tweets per tweet	Average number of retweets per tweet
#gohawks	340.97	2217.92	2.01
#gopatriots	41.47	1427.25	1.41
#nfl	461.43	4662.38	1.53
#patriots	759.69	3280.46	1.79
#sb49	1275.56	10374.16	2.53
#superbowl	2343.27	8814.97	2.39

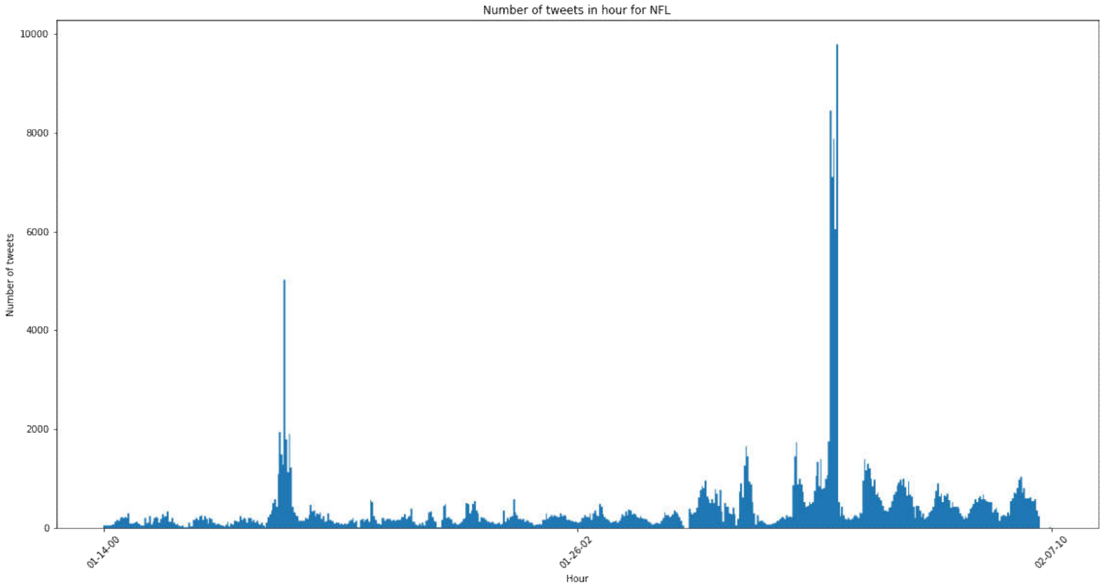
QUESTION 2: Plot "number of tweets in hour" over time for #SuperBowl and #NFL (a histogram with 1-hour bins).

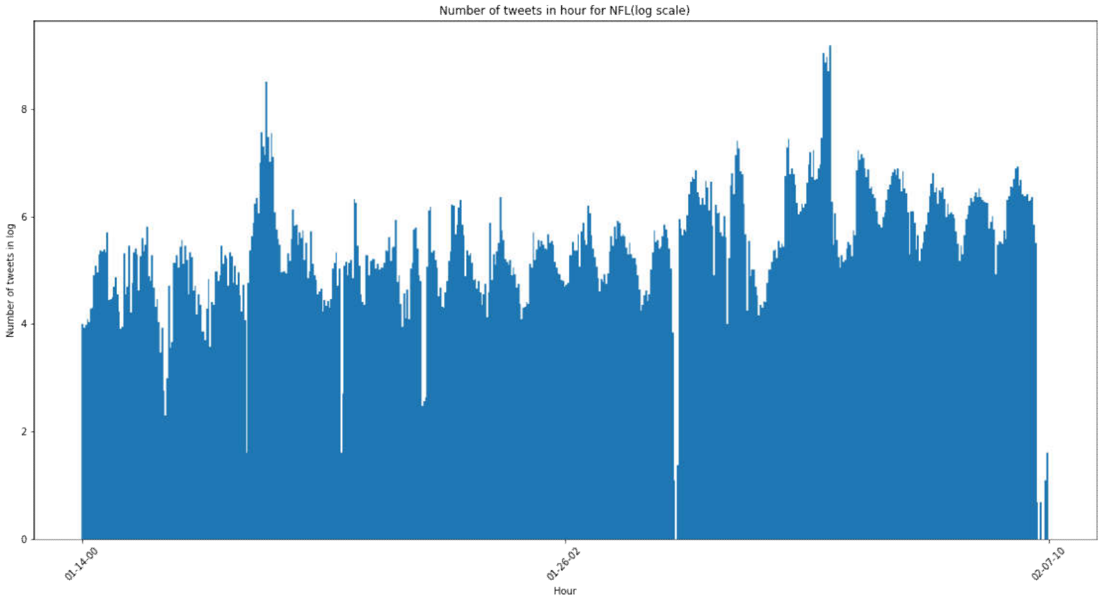
a) For #SuperBowl:




Number of tweets in hour for SuperBowl(log scale)
Number of tweets in log
0
2
4
6
8
10
12
01-14-00
01-26-04
02-07-10
Hour

b) For #NFL


Number of tweets in hour for NFL
Number of tweets
0
2000
4000
6000
8000
10000
01-14-00
01-26-02
02-07-10
Hour



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

a) #gohawks:

MSE value: 716101.690801841

R-squared measure: 0.5289248105728146

feature	t-test	P-value
Number of tweets	9.5294923	4.42086862e-20
Number of retweets	-4.44705288	1.04587450e-05
Sum of the number of followers of the users posting the hashtag	-3.8401074	1.36692130e-04
Maximum number of followers of the users posting the hashtag	2.14594155	3.22970250e-02
Time of the day	2.20091074	2.81404313e-02

b) #gopatriots:

MSE value: 30590.490704986703

R-squared measure: 0.6058234151577628

feature	t-test	P-value
Number of tweets	-0.47930696	0.63190443
Number of retweets	3.0076869	0.00274877

Sum of the number of followers of the users posting the hashtag	0.81210353	0.41707184
Maximum number of followers of the users posting the hashtag	-1.42904424	0.15354026
Time of the day	0.96439399	0.33525811

c) #nfl:
MSE value: 274001.9984320582
R-squared measure: 0.6517996737809222

feature	t-test	P-value
Number of tweets	4.71348753	3.05131530e-06
Number of retweets	-2.66286299	7.96262358e-03
Sum of the number of followers of the users posting the hashtag	4.08398895	5.04816949e-05
Maximum number of followers of the users posting the hashtag	-3.03572152	2.50706971e-03
Time of the day	4.01589862	6.69902841e-05

d) #patriots:
MSE value: 4579216.244725382
R-squared measure: 0.7146966270449413

feature	t-test	P-value
Number of tweets	15.41251794	3.47298678e-45
Number of retweets	-5.01275882	7.13446446e-07
Sum of the number of followers of the users posting the hashtag	1.57531078	1.15729057e-01
Maximum number of followers of the users posting the hashtag	0.70103572	4.83561393e-01
Time of the day	1.39257318	1.64281585e-01

e) #sb49:
MSE value: 13171504.39991799
R-squared measure: 0.8416361286654217

feature	t-test	P-value
Number of tweets	13.02630523	3.46364913e-34
Number of retweets	-2.22756426	2.62944491e-02
Sum of the number of followers of the users posting the hashtag	0.91650475	3.59785108e-01
Maximum number of followers of the users posting the hashtag	4.42783206	1.13873506e-05
Time of the day	-1.18988154	2.34582314e-01

f) #superbowl:
MSE value: 34209180.3254181
R-squared measure: 0.8699072245830363

feature	t-test	P-value
Number of tweets	24.00671388	5.91198600e-89
Number of retweets	-4.84396911	1.63518170e-06
Sum of the number of followers of the users posting the hashtag	-20.05965064	2.21115533e-68
Maximum number of followers of the users posting the hashtag	10.98490362	1.24170842e-25
Time of the day	-2.26364349	2.39646799e-02

QUESTION 4: Design a regression model using any features from the papers you find or other new features you may find useful for this problem. Fit your model on the data of each hashtag and report fitting MSE and significance of features.

In this part, we use 14 features, calculate their statistics in hours and train models on the data we get.
The features are:
['Number of tweets', 'Total number of retweets',
'Sum of the number of followers', 'Maximum number of followers',
'Time of the day', 'Total number of impressions',
'Total number of momentum', 'Total number of favorite count',
'Total number of ranking score', 'Total number of acceleration',
'Total number of replies', 'Total number of unique users',

```
 'Total number of unique authors', 'Total number of user mentions']
```

a) #gohawks:

```
MSE value:    414743.142968156
R-squared measure :    0.7271683517202944
p values :
   [4.32658730e-04 6.09384448e-05 1.34927098e-01 2.63677589e-01
   7.83523860e-03 1.73215741e-03 1.30150293e-16 2.57739273e-09
   2.55563154e-06 4.85694869e-01 9.24758854e-09 3.26926042e-02
   1.76026196e-01 1.01607789e-04]
t-test :
   [-3.54031072 -4.03969286 -1.49709824 -1.11885666 -2.66863321    3.1477661
   -8.53490438    6.05414237    4.7522881      0.69763197    5.83166609 -2.14111638
     1.3547922      3.91464638]
```

b) #gopatriots:

```
MSE value: 11176.642696427321
R-squared measure : 0.8559823413567706
p values :
  [5.77234745e-01 3.23316657e-13 2.10858207e-40 5.45714893e-20
  1.10254121e-01 9.85475249e-24 9.15501130e-01 7.21365262e-04
  7.49629580e-01 6.58658511e-01 9.24197599e-01 6.07378314e-19
  8.21335928e-20 5.10791411e-44]
t-test :
  [ 0.55775476    -7.46378187    14.4389285      -9.51251785    -1.59958749
  -10.51779325      0.10615028    -3.40019053     -0.31928486    -0.44200184
    -0.09519046    -9.21819072      9.46303665    15.21744533]
```

c) #nfl :

```
MSE value: 166528.59341512353
R-squared measure : 0.7883763224948585
p values :
  [3.01377080e-01 1.27090668e-02 2.33048083e-01 4.14838130e-03
  5.37258839e-01 1.23184554e-01 2.15657474e-01 4.64225986e-29
  7.73308910e-01 5.23172650e-01 1.23648251e-01 4.43465294e-02
  2.97779437e-02 6.97922515e-10]
t-test :
   [1.0344198      -2.49969416    -1.19380764      2.87824472    -0.61733593
   -1.54382177      1.23953033 -11.83515329    -0.2881858      -0.63885552
   -1.54191096      2.01524242    -2.17847925      6.27329779]
```

d) #patriots:

```
MSE value: 3568809.458762839
R-squared measure : 0.7776489858517225
```

```
p values :
 [1.73738836e-01 1.65083385e-02 9.24242568e-04 3.43381686e-09
  7.77782200e-01 9.16049970e-01 1.91389211e-03 3.35497701e-01
  1.65606833e-01 6.96285866e-01 5.80794184e-01 1.78965259e-04
  9.35407844e-05 5.41098419e-11]
t-test :
 [-1.3619847   -2.40456943   3.33004486 -6.00342856   0.28234358 -0.10545724
  -3.11775862   0.96391134   1.38823198 -0.3905357    0.55254014 -3.77177149
   3.9347083    6.6874702 ]
```

e) #sb49:

```
MSE value: 5532649.055765376
R-squared measure: 0.9334797532154293
p values :
 [1.63125863e-02 2.16129022e-01 1.63012674e-05 9.35502386e-04
  2.45265048e-01 4.16080948e-02 7.05345538e-03 4.46572739e-03
  1.47996269e-01 3.85667066e-04 2.91569431e-13 1.61150471e-12
  4.34859664e-46 6.78630862e-02]
t-test :
 [-2.40902177 -1.23826589   4.34775564   3.32672861 -1.1631303   -2.04204825
   2.70413776 -2.85465135   1.44861737   3.57106599 -7.47614165 -7.22589689
  15.63446821 -1.82940451]
```

f) #superbowl.txt :

```
MSE value: 17383904.794833045
R-squared measure: 0.9338914174256394
p values :
 [1.08823835e-04 1.59193948e-26 1.49322937e-03 4.42400693e-02
  6.49641228e-01 1.31890629e-01 1.52971535e-19 3.29515873e-13
  1.99682465e-02 5.90614579e-18 3.44271086e-02 5.81459719e-06
  1.35386323e-03 1.10551978e-01]
t-test :
 [-3.89726698 11.2165766    3.19148154 -2.01626455   0.45450068 -1.50884841
  -9.38056793 -7.45750738   2.33349687 -8.92812647   2.12011358   4.57614376
  -3.22018971 -1.59819858]
```