

Assignment 8

Q.1

We are creating a twitter app using <https://apps.twitter.com> . Once we create an app we make a file named auth.k which contains keys/tokens in the order Consumer_key, Consumer_secret, Access_token, Access_token_secret. These are copied from the twitter app which we created. Then we write script in Python to use the authentication keys to get data from twitter. We already have this python script which is tw_search.py and twitter_sentiments.py. We run tw_search.py python script using the configuration trump -c 180. This searches trump tweets and retrieve 180 tweets. The script writes the data in result.csv file having 6 columns – created time, retweet count, hashtag, followers count, friends count. We take different topics/politicians and re-run the python script to collect 180 tweets for each topics/politicians. We run twitter_sentiments.py python script for each topics/politicians. This script calculates the polarity and the subjectivity and adds in the data. The resulted file from this contains 10 columns: username, author id, created, text, retwc, hashtag, followers, friends, polarity, subjectivity. We change the query to different topics/politicians and re-run the python script to collect tweets and calculate polarity and subjectivity(sentiments) for each topics/politicians.

I have selected Steve Smith, Kim Jong, and Trump. So 180 tweets are collected for each personalities using tw_search.py and around 50 tweets using twitter_sentiments.py.

For Steve Smith :

We find the correlation between variables in the dataset created by twitter_sentiments.py

Correlation of Followers and Friends= **-0.178951077396**
 Correlation of Followers and Polarity= **0.0921856495584**
 Correlation of Followers and subjectivity= **-0.137120912875**
 Correlation of Friends and Polarity= **0.201494410746**
 Correlation of Friends and subjectivity= **0.159399656592**

So we first take the variables Friends and Polarity as it has the highest correlation among the variables (0.2014) which is low positive correlation. We try to predict Polarity using Friends.

OLS Regression Results						
Dep. Variable:	polarity	R-squared:	0.041			
Model:	OLS	Adj. R-squared:	0.018			
Method:	Least Squares	F-statistic:	1.820			
Date:	Fri, 30 Mar 2018	Prob (F-statistic):	0.184			
Time:	13:42:55	Log-Likelihood:	-2.3156			
No. Observations:	45	AIC:	8.631			
Df Residuals:	43	BIC:	12.24			
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	0.0010	0.041	0.024	0.981	-0.082	0.084
friends	1.091e-05	8.09e-06	1.349	0.184	-5.4e-06	2.72e-05
Omnibus:	8.661	Durbin-Watson:		1.800		
Prob(Omnibus):	0.013	Jarque-Bera (JB):		13.566		
Skew:	-0.419	Prob(JB):		0.00113		
Kurtosis:	5.556	Cond. No.		5.36e+03		

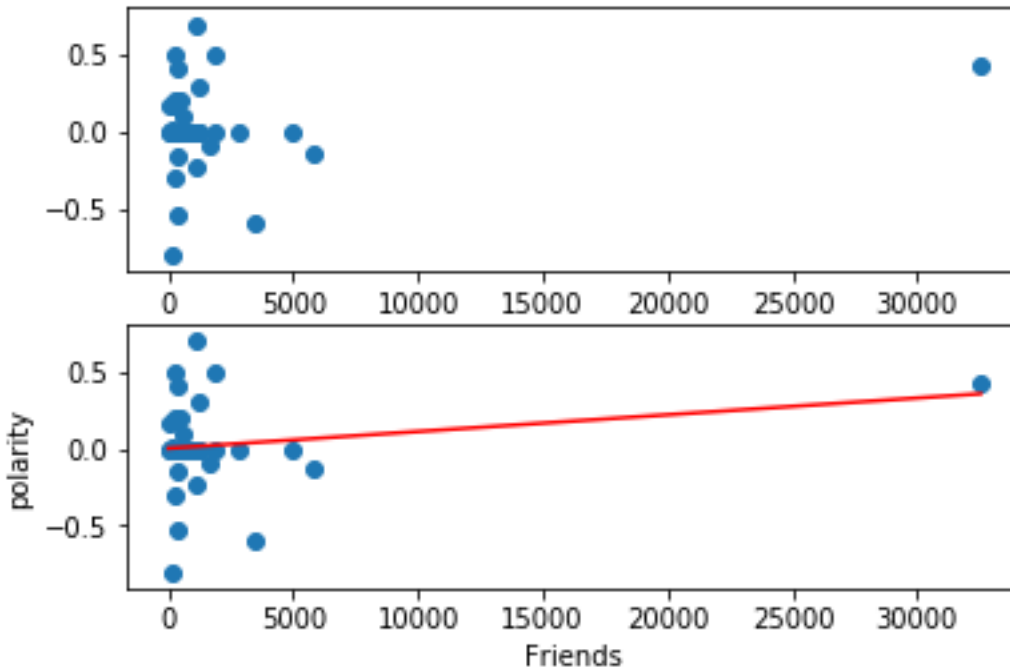
Warnings:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
 [2] The condition number is large, 5.36e+03. This might indicate that there are strong multicollinearity or other numerical problems.

```
const    0.000970
friends  0.000011
dtype: float64
```

We get a R-square value of 0.041 for friends and polarity. The equation of the regression line is
 $\text{Polarity} = 0.000011 * \text{Friends} + 0.000970$

Regression plot :



Hypothesis :

Based on the regression plot and model, We can say that More the friends, more the polarity is becoming from neutral (0) to positive.(0.5), so the tweets sentiments is becoming positive.

We then take the variables Friends and subjectivity which has (0.1593) low positive correlation. We try to predict subjectivity using Friends.

OLS Regression Results

Dep. Variable:	subjectivity	R-squared:	0.025
Model:	OLS	Adj. R-squared:	0.003
Method:	Least Squares	F-statistic:	1.121
Date:	Fri, 30 Mar 2018	Prob (F-statistic):	0.296
Time:	20:32:18	Log-Likelihood:	-15.429
No. Observations:	45	AIC:	34.86
Df Residuals:	43	BIC:	38.47
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
const	0.2826	0.055	5.146	0.000	0.172	0.393
friends	1.146e-05	1.08e-05	1.059	0.296	-1.04e-05	3.33e-05

Omnibus:	6.217	Durbin-Watson:	1.767
Prob(Omnibus):	0.045	Jarque-Bera (JB):	5.493
Skew:	0.771	Prob(JB):	0.0641
Kurtosis:	2.258	Cond. No.	5.36e+03

Warnings:

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

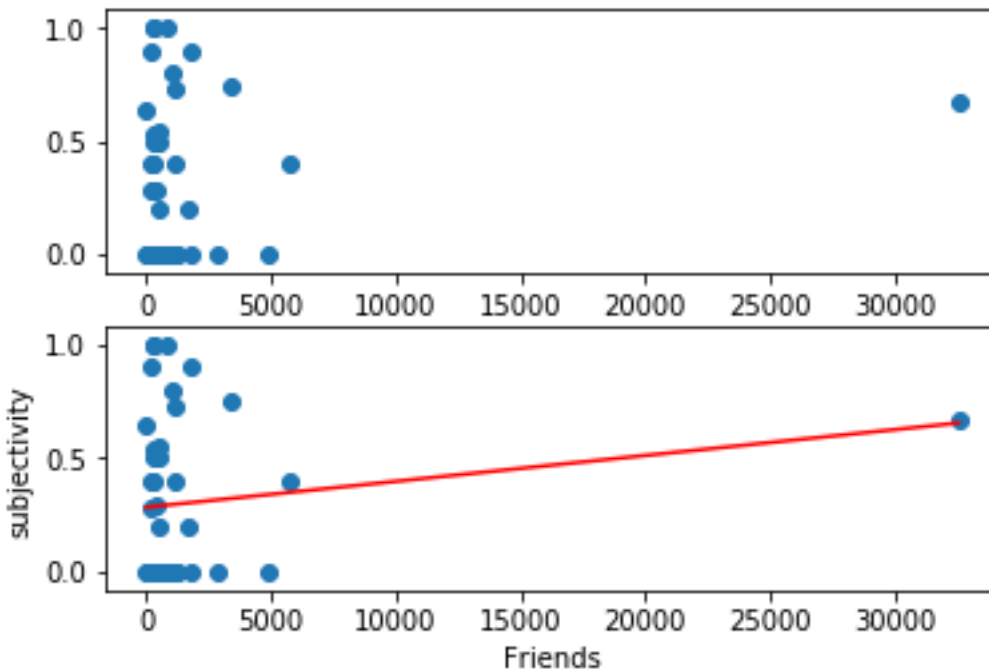
[2] The condition number is large, 5.36e+03. This might indicate that there are strong multicollinearity or other numerical problems.

const	0.282629
friends	0.000011

dtype: float64

We get a R-square value of 0.025 for friends and subjectivity. The equation of the regression line is $\text{Subjectivity} = 0.000011 * \text{Friends} + 0.282629$

Regression plot :



Hypothesis :

Based on the regression plot and model, We can say that More the friends, more the tweets sentiment is becoming subjective, as the subjectivity is becoming close to 1.

We then take the variables Followers and subjectivity which has (-0.1371) low negative correlation. We try to predict subjectivity using Followers.

```

=====
                        OLS Regression Results
=====
Dep. Variable:          subjectivity      R-squared:                0.019
Model:                  OLS              Adj. R-squared:          -0.004
Method:                 Least Squares    F-statistic:            0.8240
Date:                  Fri, 30 Mar 2018  Prob (F-statistic):      0.369
Time:                  20:32:18          Log-Likelihood:         -15.581
No. Observations:      45               AIC:                    35.16
Df Residuals:          43               BIC:                    38.77
Df Model:              1
Covariance Type:       nonrobust
=====
               coef      std err          t      P>|t|      [0.025      0.975]
-----
const          0.3366        0.065     5.177    0.000     0.205     0.468
followers -1.634e-08    1.8e-08    -0.908    0.369   -5.26e-08    2e-08
=====
Omnibus:            6.341    Durbin-Watson:           1.806
Prob(Omnibus):      0.042    Jarque-Bera (JB):         4.719
Skew:               0.662    Prob(JB):                 0.0945
Kurtosis:           2.125    Cond. No.                  4.50e+06
=====

```

Warnings:

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

[2] The condition number is large, 4.5e+06. This might indicate that there are strong multicollinearity or other numerical problems.

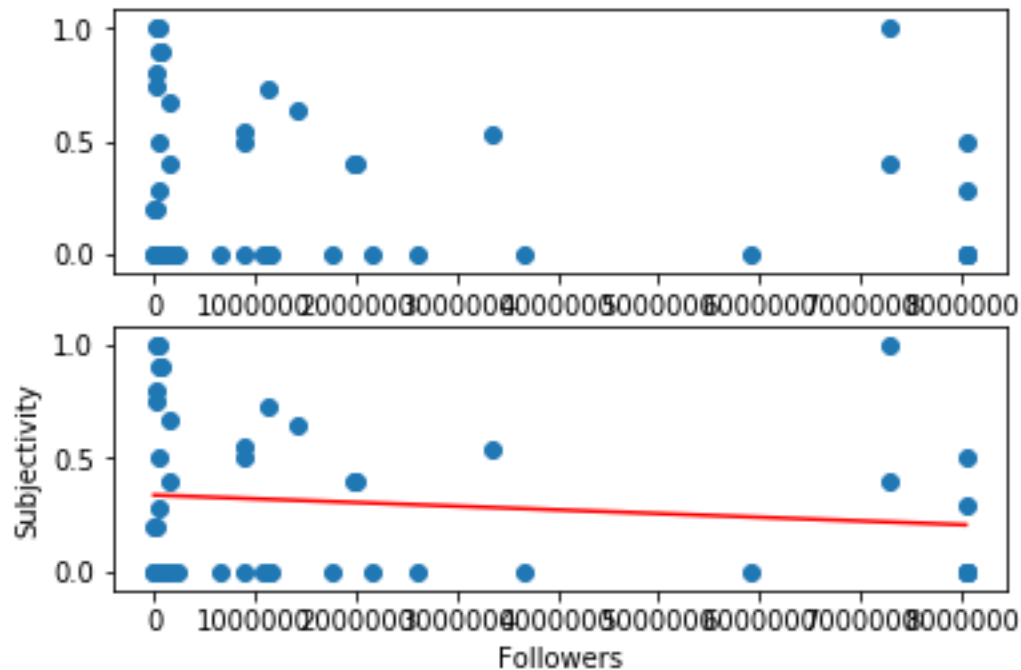
```

const      3.365781e-01
followers  -1.634069e-08
dtype: float64

```

We get a R-square value of 0.025 for friends and subjectivity. The equation of the regression line is Subjectivity = $-1.634069 \times 10^{-8} \times \text{Followers} + 3.365781 \times 10^{-1}$

Regression plot :



Hypothesis :

Based on the regression plot and model, We can say that More the followers, more the tweets sentiment is becoming objective, as the subjectivity is getting close to zero.

For Kim Jong :

We find the correlation between variables in the dataset using tw_search.py

Correlation of Followers and Friends= 0.132067668457
 Correlation of Followers and retweets= -0.0644412837818
 Correlation of Friends and retweets= -0.14146123666

OLS Regression Results

Dep. Variable:	followers	R-squared:	0.017
Model:	OLS	Adj. R-squared:	0.012
Method:	Least Squares	F-statistic:	3.160
Date:	Fri, 30 Mar 2018	Prob (F-statistic):	0.0772
Time:	21:28:44	Log-Likelihood:	-2340.4
No. Observations:	180	AIC:	4685.
Df Residuals:	178	BIC:	4691.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
const	6562.2557	8569.933	0.766	0.445	-1.03e+04	2.35e+04
friends	1.5252	0.858	1.778	0.077	-0.168	3.219

Omnibus:	395.486	Durbin-Watson:	2.009
Prob(Omnibus):	0.000	Jarque-Bera (JB):	232181.559
Skew:	13.244	Prob(JB):	0.00
Kurtosis:	176.942	Cond. No.	1.06e+04

Warnings:

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

[2] The condition number is large, 1.06e+04. This might indicate that there are strong multicollinearity or other numerical problems.

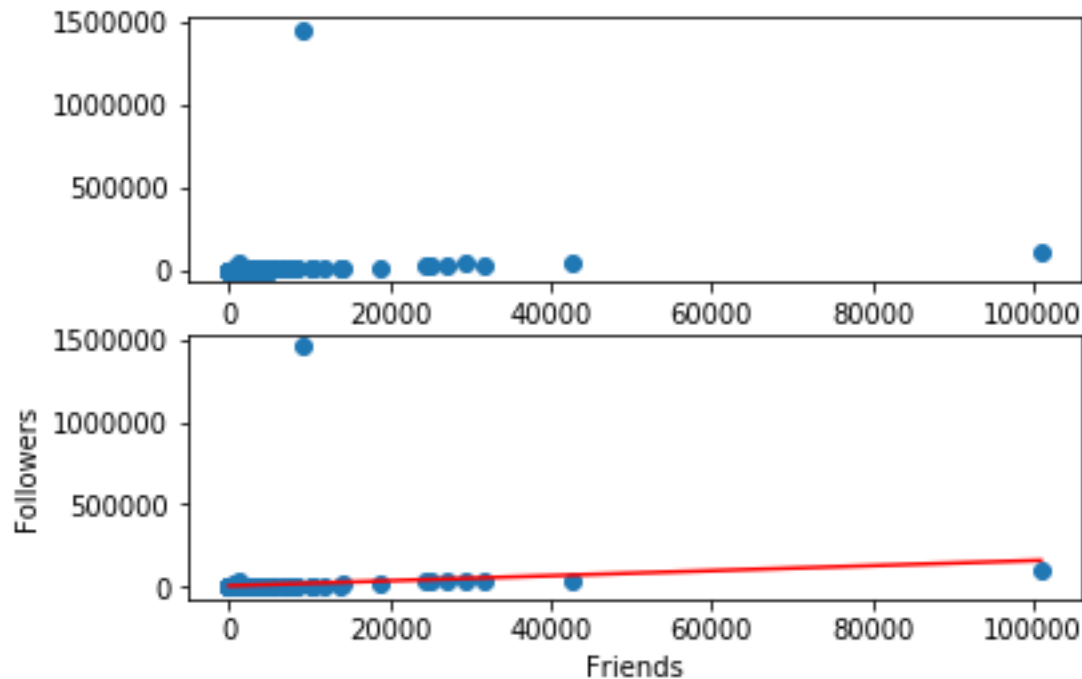
const 6562.255706
 friends 1.525247

dtype: float64

We then take the variables Followers and friends which has (0.32067) low negative correlation..
We try to predict Followers using Friends.

We get a R-square value of 0.017 for friends and followers. The equation of the regression line is
Followers = 1.5252*Friends + 6562.255706

Regression plot :



We find the correlation between variables in the dataset using twitter-sentiments.py

```
Correlation of Followers and Friends= -0.118811265005
Correlation of Followers and Polarity= -0.134251652549
Correlation of Followers and subjectivity= -0.00225257682909
Correlation of Friends and Polarity= -0.02922533351
Correlation of Friends and subjectivity= -0.125836955729
```

We then take Friends and subjectivity variables which has (-0.1258) low negative correlation and we try to predict subjectivity using Friends.

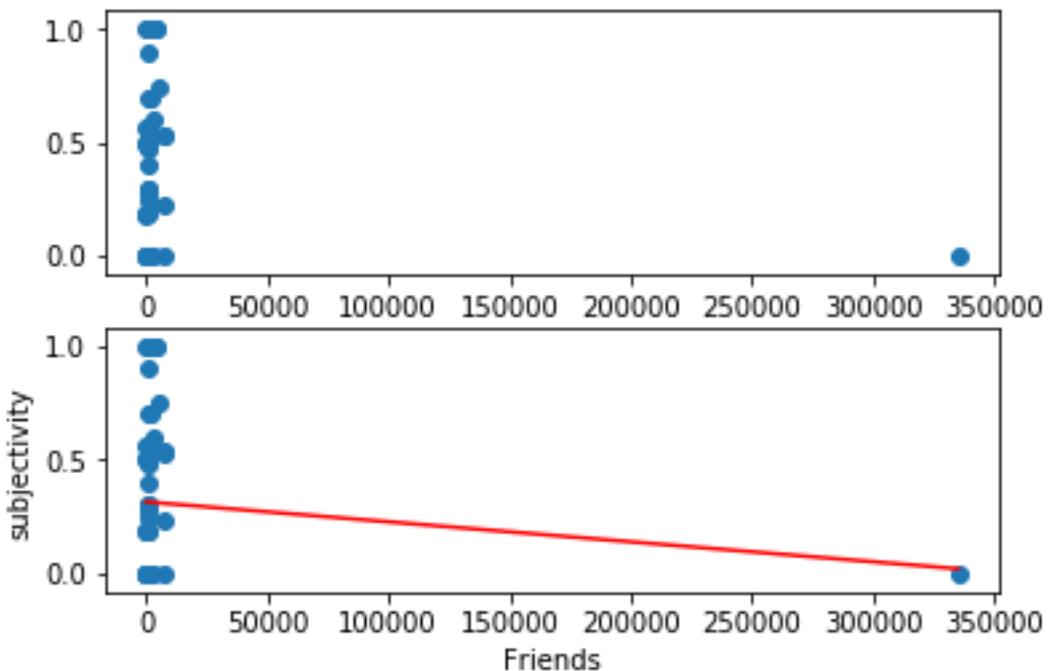
We get a R-square value of 0.016 for friends and subjectivity. The equation of the regression line is Subjectivity = -8.820022e-07*Friends + 3.142514e-01

```

=====
                        OLS Regression Results
=====
Dep. Variable:          subjectivity      R-squared:                0.016
Model:                  OLS              Adj. R-squared:           -0.005
Method:                 Least Squares     F-statistic:             0.7723
Date:                   Fri, 30 Mar 2018  Prob (F-statistic):      0.384
Time:                   21:51:13          Log-Likelihood:          -14.765
No. Observations:       50              AIC:                     33.53
Df Residuals:           48              BIC:                     37.35
Df Model:                1
Covariance Type:        nonrobust
=====
               coef      std err          t      P>|t|      [0.025      0.975]
-----
const          0.3143      0.048      6.598      0.000      0.218      0.410
friends      -8.82e-07      1e-06     -0.879      0.384     -2.9e-06     1.14e-06
=====
Omnibus:                 5.346    Durbin-Watson:           1.964
Prob(Omnibus):            0.069    Jarque-Bera (JB):         4.902
Skew:                     0.696    Prob(JB):                 0.0862
Kurtosis:                 2.354    Cond. No.                  4.82e+04
=====

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 4.82e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
const      3.142514e-01
friends    -8.820022e-07
dtype: float64

```



Hypothesis :

Based on the regression plot and model, We can say that More the Friends, more the tweets sentiment is becoming objective, as the subjectivity is getting close to zero.

For Trump:

We find the correlation between variables in the dataset using twitter_sentiments.py

Correlation of Followers and Friends= -0.130915199181

Correlation of Followers and Polarity= 0.024184757149

Correlation of Followers and subjectivity= -0.14211631085

Correlation of Friends and Polarity= -0.0170203212416

Correlation of Friends and subjectivity= -0.14998220444

So we first take the variables Friends and subjectivity as it has the highest correlation among the variables (-0.1499) which is low negative correlation. We try to predict subjectivity using Friends.

```

=====
                        OLS Regression Results
=====
Dep. Variable:          subjectivity    R-squared:                0.022
Model:                  OLS             Adj. R-squared:           0.002
Method:                 Least Squares    F-statistic:              1.105
Date:                  Fri, 30 Mar 2018  Prob (F-statistic):      0.299
Time:                  23:38:34          Log-Likelihood:          -8.9812
No. Observations:      50               AIC:                     21.96
Df Residuals:          48               BIC:                     25.79
Df Model:               1
Covariance Type:       nonrobust
=====

```

	coef	std err	t	P> t	[0.025	0.975]
const	0.3114	0.044	7.005	0.000	0.222	0.401
friends	-4.859e-06	4.62e-06	-1.051	0.299	-1.42e-05	4.44e-06

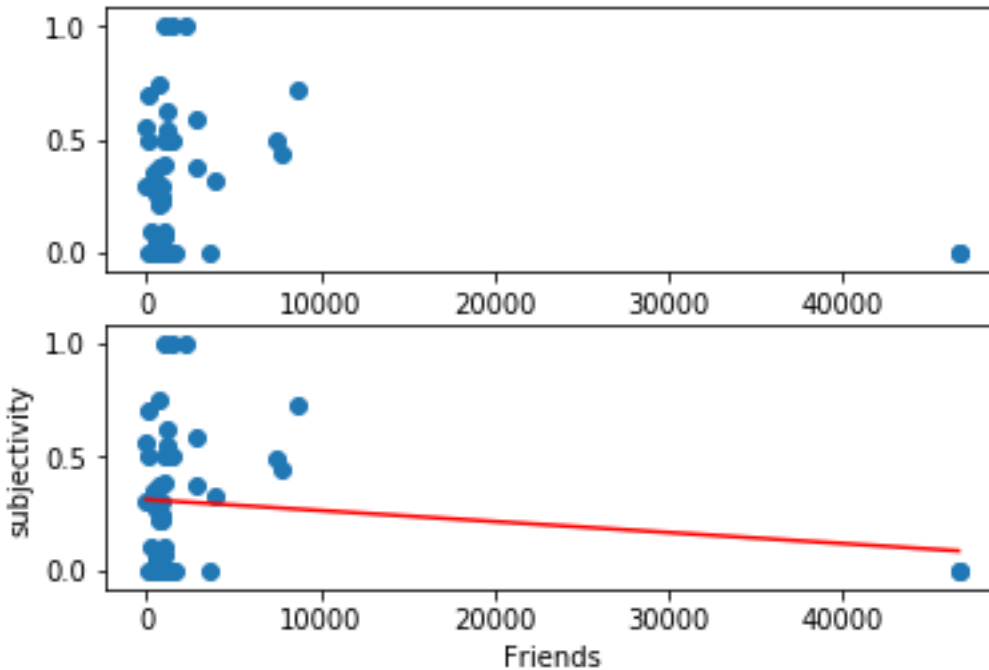
```

=====
Omnibus:                 5.017    Durbin-Watson:           2.664
Prob(Omnibus):           0.081    Jarque-Bera (JB):        4.818
Skew:                    0.754    Prob(JB):                0.0899
Kurtosis:                2.797    Cond. No.                1.02e+04
=====
Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 1.02e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
const      0.311353
friends    -0.000005
dtype: float64

```

We get a R-square value of 0.022 for friends and subjectivity. The equation of the regression line is Subjectivity = $-0.000005 \times \text{Friends} + 0.311353$

Regression plot:



Hypothesis :

Based on the regression plot and model, We can say that More the Friends, more the tweets sentiment is becoming objective, as the subjectivity is getting close to zero.

We take the variables Followers and subjectivity as it has the second highest correlation among the variables (-0.1421) which is low negative correlation. We try to predict subjectivity using Followers.

OLS Regression Results						
Dep. Variable:	subjectivity		R-squared:	0.020		
Model:	OLS		Adj. R-squared:	-0.000		
Method:	Least Squares		F-statistic:	0.9894		
Date:	Fri, 30 Mar 2018		Prob (F-statistic):	0.325		
Time:	23:38:34		Log-Likelihood:	-9.0399		
No. Observations:	50		AIC:	22.08		
Df Residuals:	48		BIC:	25.90		
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	0.3154	0.046	6.799	0.000	0.222	0.409
followers	-8.069e-09	8.11e-09	-0.995	0.325	-2.44e-08	8.24e-09
Omnibus:		4.295	Durbin-Watson:			2.591
Prob(Omnibus):		0.117	Jarque-Bera (JB):			4.156
Skew:		0.685	Prob(JB):			0.125
Kurtosis:		2.653	Cond. No.			6.34e+06

Warnings:

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

[2] The condition number is large, 6.34e+06. This might indicate that there are strong multicollinearity or other numerical problems.

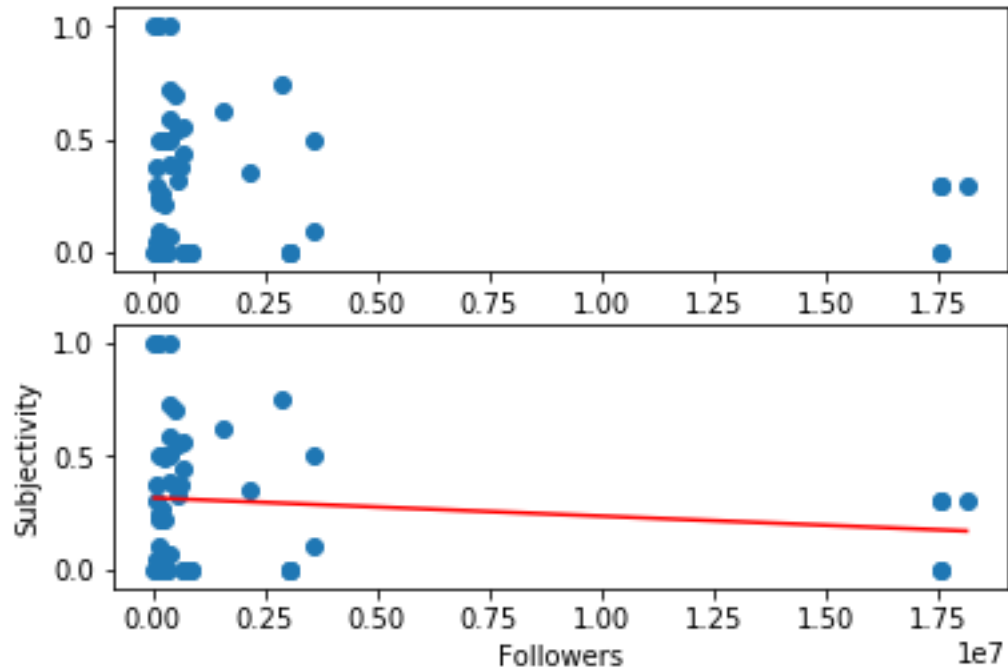
const 3.153665e-01

followers -8.069399e-09

dtype: float64

We get a R-square value of 0.020 for followers and subjectivity. The equation of the regression line is $\text{Subjectivity} = -8.069399\text{e-}09 * \text{Followers} + 3.153665\text{e-}01$

Regression plot:



Hypothesis :

Based on the regression plot and model, We can say that More the Followers, more the tweets sentiment is becoming objective, as the subjectivity is getting close to zero.