Problem Statement :

Creation of a regression model that can predict the number of followers of a user on twitter, given a twitter handle.

Model Used :

Multiple Linear Regression

A linear regression model that contains more than one predictor variable is called a multiple linear regression model. The following model is a multiple linear regression model with three predictor variables, X1 ,X2 ,X3

$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3$

Y is the value of the Dependent variable (Y), what is being predicted or explained

b_0 is the Constant or intercept

b_1 is the Slope (Beta coefficient) for X_1

$\mathbf{X_1}$ First independent variable that is explaining the variance in Y

b_2 is the Slope (Beta coefficient) for X_2

$\mathbf{X_2}$ Second independent variable that is explaining the variance in Y

$\mathbf{b_3}$ is the Slope (Beta coefficient) for X_3

$\mathbf{X_3}$ Third independent variable that is explaining the variance in Y

Technologies Used :

1) Python – For data extraction
2) R – For applying the model.

Note - Due to Library installation issues I was not able to continue with python and therefore I has to use R for applying the model.

Data Extraction :

Data was extracted using

1)Twitter API

2) Tweepy (open source library in python) - https://github.com/tweepy/tweepy

3) Twython (open source library in python) - https://twython.readthedocs.org/en/latest/

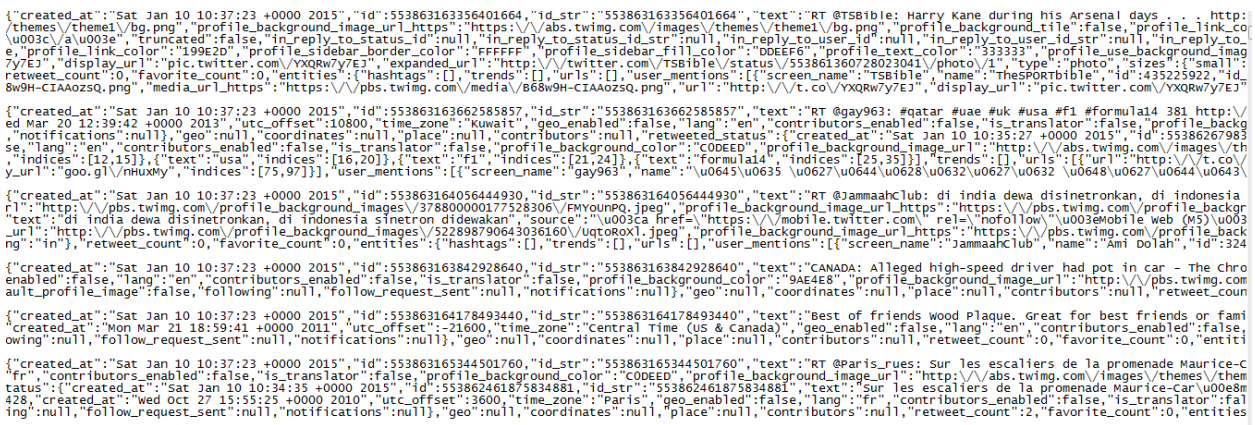
CODE for data extraction :

Tweet.py – Tweepy was used to access Twitter API.

Data consisted of Live tweets that were extracted using StreamListerner . A filter was applied to select tweets containing keywords from ['car' , 'modi' , 'obama' , 'india' , 'USA' ,'arsenal' ,'football'] . The date was saved in a .txt file.

```
from tweepy import Stream
from tweepy import OAuthHandler
from tweepy.streaming import StreamListener
import time

ckey = 'HjtRuaGIiU0HGfXE5sGiYj5BR'
csecret = 'DZNpecJr1bGcXaXWZwL07xF1bPOSBHfiyEb7jJqtzSY7RMflpX'
atoken = '264257074-fFtslg1CFtPVWXkU3A57hGSEu6SbwUVWBBBkwgZj'
asecret = 'xp92aCARZQY0R6RxSSV4JtNvldyC2noONzJcQ9lfg5tFE'

class listener(StreamListener):

    def on_data(self, data):
        try :
            print data
            saveFile = open ('twitter.txt','a')
            saveFile.write(data)
            saveFile.write('\n')
            saveFile.close()
            return True
        except BaseException, e:
            print 'failed'
            time.sleep(1)

    def on_error(self, status):
        print status
track = ['car' , 'modi' , 'obama' , 'india' , 'USA' ,'arsenal' ,'football']
auth = OAuthHandler(ckey, csecret)
auth.set_access_token(atoken, asecret)
twitterStream = Stream(auth, listener())
twitterStream.filter(track=track)
```

Twitter data saved in .txt format



As Twitter data is received in JSON format. A script was run to extract meaningful data from the file

TwitterJSON.py-

```
import json
from datetime import datetime
tweets_data_path = 'C:/Python27/twitter.txt'
tweets_file = open(tweets_data_path,'r')
count = 0
for line in tweets_file:
   try:
       tweet = json.loads(line)
       tweets_data.append(tweet)
       count = count + 1
        data = [tweet['user']['id'] , tweet['user']['name'] , tweet['user']['screen_name'] ,
        tweet['user']['followers_count'] , tweet['user']['friends_count'] ,
tweet['user']['listed_count'] ,
         tweet['user']['favourites_count'] ,  tweet['user']['statuses_count'] ,
tweet['user']['created_at']]
       date_object = datetime.strptime(data[8], "%a %b %d %H:%M:%S +0000 %Y")
       diff =  datetime.now()- date_object
       saveFile = open ('twitterDataF.csv','a')
       saveFile.write(str(data[0]) + ',' +str(data[1]) + ',' + str(data[2]) + ',' + str(data[3]) + ',' +
str(data[4]) +
       ',' + str(data[5]) + ',' + str(data[6]) + ',' + str(data[7]) + ',' + str(diff.days))
       saveFile.write('\n')
       saveFile.close()
   except:
```

continue

The data was saved in a .CSV file -

ID	NAME	HANDLE	FOLLOWERS	FOLLOWING	LISTED	FAVOURITES	STATUSES	DAYS
2882877664	Drashti Pooja	drashtipooja12	0	5	0	3	1799	53
2879932786	Angel Mind	angelsania44	1	4	0	7	3386	55
2901613928	Saira45	sairaprincess45	2	5	0	9	3269	55
2961595066	GlobalHosptlitySrvcs	J1andH2bUSAjob	3	51	0	1	39	6
2853416277	drashti234	drashti234	4	6	0	0	8599	71
53552190	Dean Sueck	dsueck	5	10	0	0	13339	2017
76345246	PB KRISHNAMURTHY	pibikay	6	30	0	82	44	1936
17832191	imhindu-gautam	gautamthp	7	7	0	4	7	2230
40659409	Eugine Goldstein	MONkGOLDsamu	8	21	0	13	3280	2064
2928424488	EC4J	ecourts4justice	9	5	0	6	93	29
52145687	Manjit Devgun	mira55	10	129	0	301	126	2021
805969	michael craig	michaelcraig	11	9	1	0	137	2871
2568568003	Rashid Khan	iam_rashidkhan	12	74	0	10	42	210
46224162	Jayne Whittam	whittam64	13	30	0	0	313	2040
2912079138	Toko Jersey Arsenal	TokJersyArsenal	14	1	0	0	10429	44
54895782	conkat33	conkat33	15	60	0	81	382	2012
44352381	Elizangela dos Reis	Leereis	16	41	0	9	5751	2047
32368443	Mary Baldwin	bmtbaldwin	17	103	1	16	215	2095
65064633	Chandresh Yadav	YADUVANSHI_CK	18	99	2	14	226	1977
78238380	Day Count	daycount	19	9	0	1	2288	1930
16173392	johnboy	dalosers	20	86	4	17	5663	2316

The final data is formatted as a CSV(Comma Separated Values) file with each row indicating a separate user and the columns as follows:

1. **handle** - twitter username | string
2. **name** - full name of the twitter user | string
3. **Days** - number of days the user has existed on twitter | number
4. **Statuses**- number of tweets this user has created (includes retweets) | number
5. **following** - number of other twitter users, this user is following | number
6. **favorites** - number of tweets the user has favorited | number
7. **lists** - number of public lists this user has been added to | number
8. **followers** - number of other users following this user | number

Moreover, The erroneous data, missing data was removed from the file

The data was split into 2 parts –

1) Training Data - Consisting of 6657 unique user

2) Testing Data - Consisting of 455 rows

Possible other parameters –

There are many other parameters that could have been useful in predicting the number of followers:

Refer [3] from the reference section -

1) Sentiment analysis of a tweet – Negative sentiments have a positive influence on follower gain and positive sentiment has the reverse effect.
2) More information about the twitter handler – For example : its been found that twitter account belonging to a group like NDTV , BCCI will have a large amount of followers than an individual.
3) A twitter account belonging to a famous person will have large number of followers.

Method Followed –

1) A Multiple linear regression model is created using all the dependent variables – FOLLOWING , LISTED, STATUSES ,DAYS and then the model is analyzed using R-Squared , P-VALUE for every variable . Those variables are removed which do not provide good P-VALUE i.e <0.05 . After removing the variables , A new model is created using the remaining variables and evaluated accordingly.

2) Correlation values between the variable provide a good evidence of the relationship between the FOLLOWERS and the dependent variables. Based on the value of correlation coefficient, variable are added into the model and analyzed using R-Square and P-Value.

STEPS -

Loading data into an object

```
>datavar <- read.csv('C:/Python27/TwitterTrainU.csv')
>st = data.frame(datavar)
>summary(st) - provides analysis of the dependent variable data eg – mean ,median , quartile.
```

NOTE – NAME and HANDLE are not used in REGRESSION ANALYSIS.

```
       ID                                  NAME                   HANDLE
 Min.   :1.117e+05   TLS Auto Recycling  :  32   toyotalexuspart:  32
 1st Qu.:1.843e+09   Antonea Roudis      :  19   AntoneaRoudis  :  19
 Median :2.229e+09   Diskon Arsenal      :  13   jessleal7      :  13
 Mean   :1.959e+09   I'm Enrique Iglesias:  13   realmadridnialI:   9
 3rd Qu.:2.359e+09   niall follow finn ! :   9   AAPTrends      :   7
 Max.   :2.970e+09   AAP Trends          :   7   ynv786         :   7
                     (Other)             :1512   (Other)        :1518
   FOLLOWERS          FOLLOWING          LISTED          FAVOURITES
 Min.   :      0   Min.   :    0   Min.   :   0.0   Min.   :     0
 1st Qu.:     62   1st Qu.:   61   1st Qu.:   0.0   1st Qu.:     0
 Median :    268   Median :  245   Median :   2.0   Median :    50
 Mean   :   3554   Mean   : 1452   Mean   :  16.5   Mean   :  2690
 3rd Qu.:   1019   3rd Qu.:  858   3rd Qu.:   9.0   3rd Qu.:   872
 Max.   :1726801   Max.   :90300   Max.   :5191.0   Max.   :304992

    STATUSES           DAYS
 Min.   :     0   Min.   :   1.0
 1st Qu.:  2414   1st Qu.: 320.0
 Median : 13845   Median : 398.0
 Mean   : 47173   Mean   : 521.1
 3rd Qu.: 50277   3rd Qu.: 489.0
 Max.   :826296   Max.   :2942.0
```

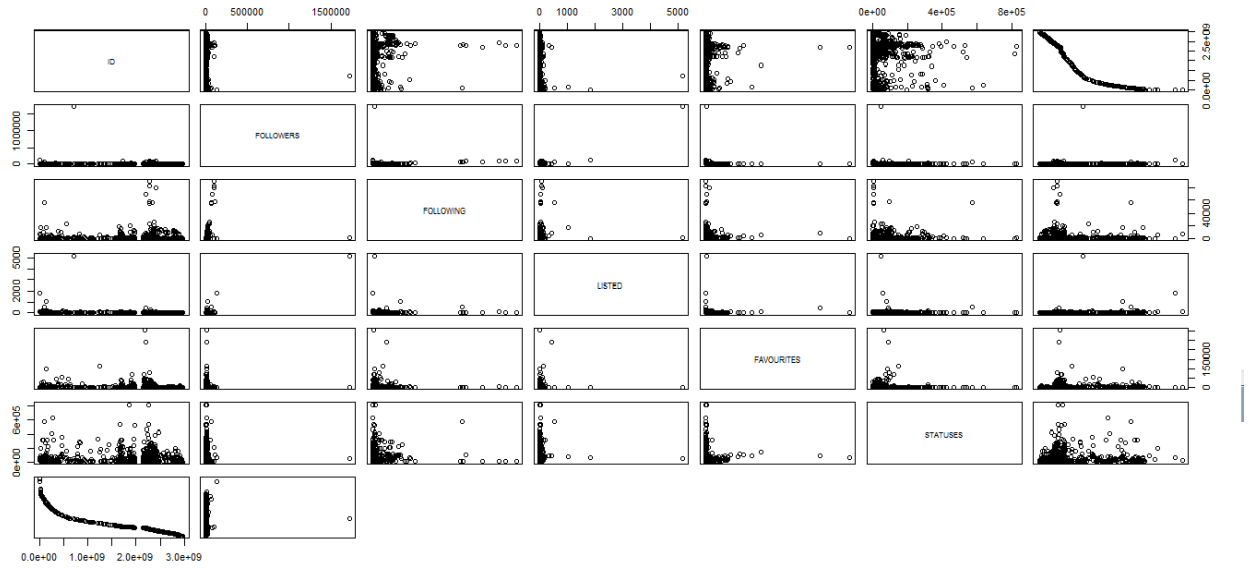
>Cor(st) – provides the correlation between the variables.

```
                ID    FOLLOWERS   FOLLOWING      LISTED   FAVOURITES    STATUSES        DAYS
ID          1.00000000 -0.067801595 -0.02511633 -0.13496523 -0.117783567 -0.28312230 -0.95267836
FOLLOWERS  -0.06780160  1.000000000  0.14689852  0.92879004  0.006383912  0.03865343  0.05943841
FOLLOWING  -0.02511633  0.146898523  1.00000000  0.08379166  0.037445152  0.09701489  0.03538801
LISTED     -0.13496523  0.928790042  0.08379166  1.00000000  0.054923791  0.08864715  0.14970350
FAVOURITES -0.11778357  0.006383912  0.03744515  0.05492379  1.000000000  0.03866931  0.11752781
STATUSES   -0.28312230  0.038653431  0.09701489  0.08864715  0.038669309  1.00000000  0.26916880
DAYS       -0.95267836  0.059438407  0.03538801  0.14970350  0.117527813  0.26916880  1.00000000
```

INFERENCE –

1) There is no correlation between the dependent variable pairs i.e FOLLOWING , LISTED , FAVOURTES , STATUSES ,DAYS – eliminating multicolinearity .
2) There is a strong correlation between FOLLOWRS and LISTED.

>Pairs(st) – Scatter plot with all the pairs of variables



Lm() function was used for applying to fit a multiple linear regression model.

1)

MODEL1 – Using all the dependent variable.

>model1 = lm(formula = FOLLOWERS ~ FOLLOWING + STATUSES + DAYS + LISTED, data = datavar)

>model1

Below are the values of the coefficient - b_0 , b_1 , b_3 , b_4

Coefficients:

(Intercept) FOLLOWING STATUSES DAYS LISTED

695.38825 1.06650 0.01124 -0.50824 113.05261

> summary(model1) – will show the analysis of the model providing various parameters to access the model.

```
Residuals:
     Min       1Q   Median       3Q      Max
-1660116    -1875     -713     -272  2661857

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 138.568075 873.076481   0.159    0.874
FOLLOWING     1.068542   0.146880   7.275 3.86e-13 ***
STATUSES      0.011494   0.009151   1.256    0.209
FAVOURITES   -0.033483   0.080041  -0.418    0.676
LISTED      113.009812   1.266060  89.261  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 62800 on 6669 degrees of freedom
  (1012 observations deleted due to missingness)
Multiple R-squared:  0.5519,    Adjusted R-squared:  0.5517
F-statistic:  2054 on 4 and 6669 DF,  p-value: < 2.2e-16
```

ANALYSIS -

Now we analyse this model using the R-squared , P value etc.

R-Squared value - 0.5519 i.e 55.19% of the points can be explained by this model.

As we are considering a confidence interval of 95%, the P-Value of the dependent variables should be <0.05. If it is > 0.05 - we shall remove the parameter from the model.

It can be seen that the P- value of Intercept , STATUSES ,FAVOURITES and DAYS > 0.05 – Therefore , we remove these variable from the model and now again apply the regression using FOLLOWING and LISTED dependent variables.

INFERENCE -

Therefore , Model1 rejected

2)

MODEL2 – Using only FOLLOWING and LISTED

> model2 = lm(formula = FOLLOWERS ~ FOLLOWING + LISTED, data = datavar)
> model2

Coefficients:

(Intercept) FOLLOWING LISTED

489.050 1.083 113.087

> summary(model2)

```
Residuals:
     Min       1Q   Median       3Q      Max
-1661234    -1899     -889     -557  2662744

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  489.050    792.240   0.617    0.537
FOLLOWING      1.083      0.146   7.420 1.32e-13 ***
LISTED       113.087      1.264  89.429  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 62790 on 6671 degrees of freedom
  (1012 observations deleted due to missingness)
Multiple R-squared:  0.5518,    Adjusted R-squared:  0.5517
F-statistic:  4107 on 2 and 6671 DF,  p-value: < 2.2e-16
```

ANALYSIS -

R-Squared - 0.5518 i.e 55.18% of the points can be explained by this model.

P-value of the intercept is >0.05 -therefore , we won use it in the final equation. P-value of FOLLOWING and LISTED are <0.05- therefore included in the equation.

>Confint(model2)

The confidence interval

	2.5 %	97.5 %
(Intercept)	-1063.9930065	2042.092287
FOLLOWING	0.7968102	1.369037
LISTED	110.6078802	115.565694

Therefore final equation is –

$Y = b_1 * FOLLOWING + b_2 * LISTED$

Where $b_1 = 1.083$ & $b_2 = 113.087$

PREDICTION –

Loading of Testfile

>datavar_pred <- read.csv('C:/Python27/TwitterD.csv')

>attach(datavar_pred)

Y will contain the prediction of the number of followers of a person with a given handle.

#Y is the predicted followers vector calculated from following and listed vector.

>Y = coef(model2)[2]*FOLLOWING + coef(model2)[3]*LISTED

ERROR –

Yis the predicted number of followers and datavar[4] contains the actual number of followers.

>e = Y-datavar[4]

The e i.e the error value was found to be very high.

3)

MODEL3 –

With only LISTED as the Dependent variable (As Inferred from the correlation data)

> model3 = lm(formula = FOLLOWERS ~ LISTED, data = datavar)

>model3

Coefficients:

(Intercept) LISTED

1841.4 113.8

>summary(model3)

```
Residuals:
     Min       1Q   Median       3Q      Max
-1679910    -2232    -1830    -1589  2656153

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1841.428    774.104   2.379   0.0174 *
LISTED       113.845      1.265  89.961   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 63050 on 6672 degrees of freedom
  (1012 observations deleted due to missingness)
Multiple R-squared:  0.5481,    Adjusted R-squared:  0.5481
F-statistic:  8093 on 1 and 6672 DF,  p-value: < 2.2e-16
```

The R-square value – 0.5481

And the P-value for both Intercept and LISTED is within 0.05 . Therefore , both are included in our final equation

$Y = b_0 + b_1 * LISTED$

Where b_0 = 1841.4 & b_1 = 113.8

Loading of Testfile
>datavar_pred <- read.csv('C:/Python27/TwitterD.csv')
>attach(datavar_pred)

Y will contain the prediction of the number of followers of a person with a given handle.

#Y is the predicted followers vector calculated from following and listed vector.

>Y = coef(model2)[1] + coef(model2)[3]*LISTED

ERROR –

Y is the predicted number of followers and datavar[4] contains the actual number of followers.

>e = Y-datavar[4]

The e i.e the error value was found to be very high

OVERALL CONCLUSION :

1) A Linear Regression Model is not appropriate for predicting the Number of Followers of a person given his twitter handle. There seem to be no linear relation between the various dependent variable and the number of followers as can also be seen from the scatter plot.

2)A non linear model may provide better results. The non linear regression analysis was not possible due to time constraints.

REFERENCES

[1] http://rstudio-pubs-static.s3.amazonaws.com/10578_319e4083c11341cfb8a7d79b65665f6f.html

[2] Video explaining regression https://www.youtube.com/user/BCFoltz

[3] A Longitudinal Study of Follow Predictors on Twitter - http://comp.social.gatech.edu/papers/follow_chi13_final.pdf