act_report

January 30, 2019

1 Analysis

```
In [1]: import pandas as pd
        import matplotlib.pyplot as plt

In [2]: twitr_clean_df=pd.read_csv('/home/workspace/image_predctns/twitter_archive_master.csv')

In [3]: twitr_clean_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1916 entries, 0 to 1915
Data columns (total 25 columns):
Unnamed: 0               1916 non-null int64
tweet_id                 1916 non-null int64
in_reply_to_status_id    1916 non-null float64
in_reply_to_user_id      1916 non-null float64
timestamp                1916 non-null object
source                   1916 non-null object
text                     1916 non-null object
expanded_urls            1916 non-null object
rating_numerator         1916 non-null int64
rating_denominator       1916 non-null int64
name                     1916 non-null object
retweets                 1916 non-null float64
favourites               1916 non-null float64
jpg_url                  1916 non-null object
img_num                  1916 non-null float64
p1                       1916 non-null object
p1_conf                  1916 non-null float64
p1_dog                   1916 non-null bool
p2                       1916 non-null object
p2_conf                  1916 non-null float64
p2_dog                   1916 non-null bool
p3                       1916 non-null object
p3_conf                  1916 non-null float64
p3_dog                   1916 non-null bool
dog_stage                1916 non-null object
```

```
dtypes: bool(3), float64(8), int64(4), object(10)
memory usage: 335.0+ KB
```

```
In [4]: twitr_ana_df=twitr_clean_df[['timestamp','rating_numerator','rating_denominator',
                                     'retweets','favourites']].copy()

        twitr_ana_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1916 entries, 0 to 1915
Data columns (total 5 columns):
timestamp             1916 non-null object
rating_numerator      1916 non-null int64
rating_denominator    1916 non-null int64
retweets              1916 non-null float64
favourites            1916 non-null float64
dtypes: float64(2), int64(2), object(1)
memory usage: 74.9+ KB
```

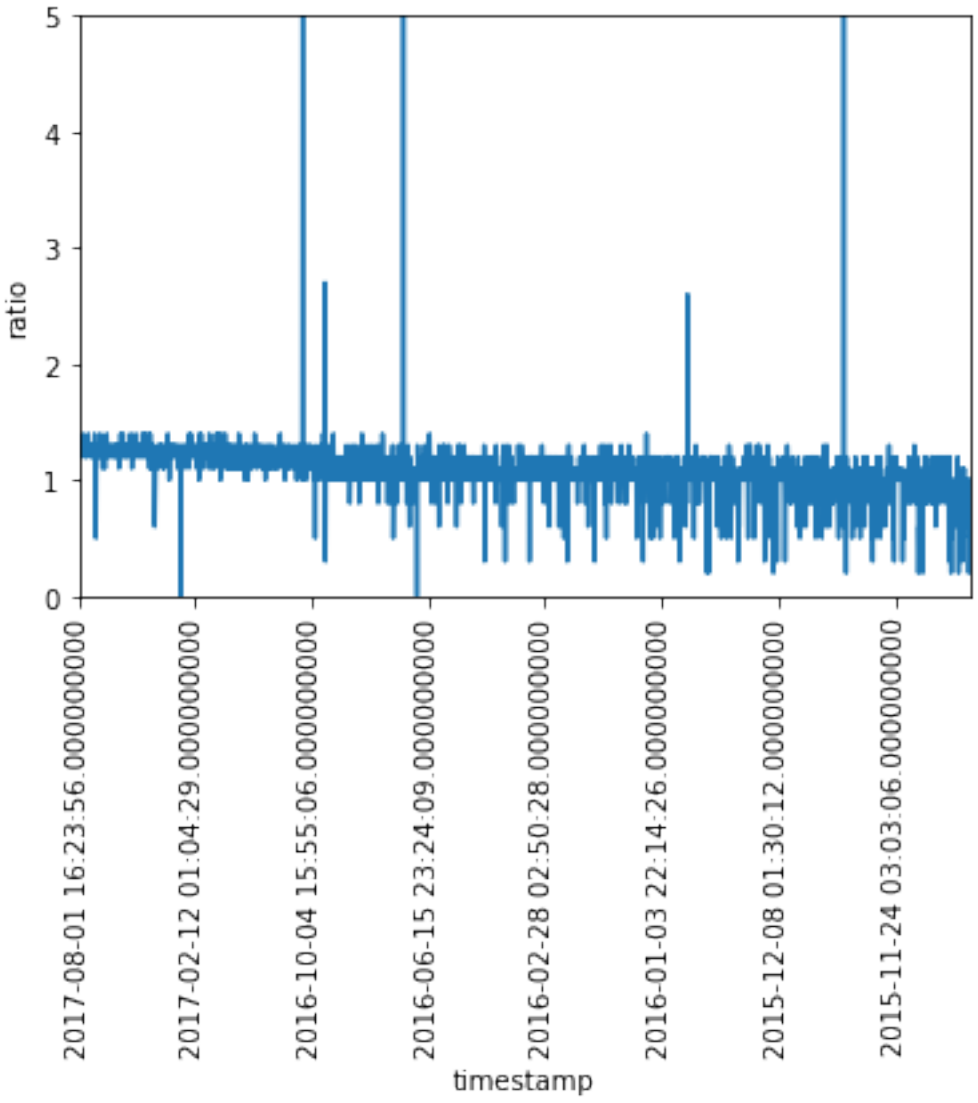
```
In [ ]:
```

```
In [5]: twitr_ana_df['rating_obtained']=twitr_ana_df['rating_numerator']/twitr_ana_df['rating_de


        twitr_ana_df.rating_obtained.value_counts()
```

```
Out[5]: 1.0      615
        1.2      437
        1.1      381
        1.3      257
        0.8       93
        1.4       34
        0.5       33
        0.6       32
        0.3       18
        0.2        9
        0.0        2
        2.6        1
        2.7        1
        177.6      1
        42.0       1
        7.5        1
        Name: rating_obtained, dtype: int64
```

```
In [6]: # set the timestamp as index to get the time display in plots against
        # the rating obtained
        twitr_ana_df.set_index(twitr_ana_df['timestamp'],inplace=True)
```

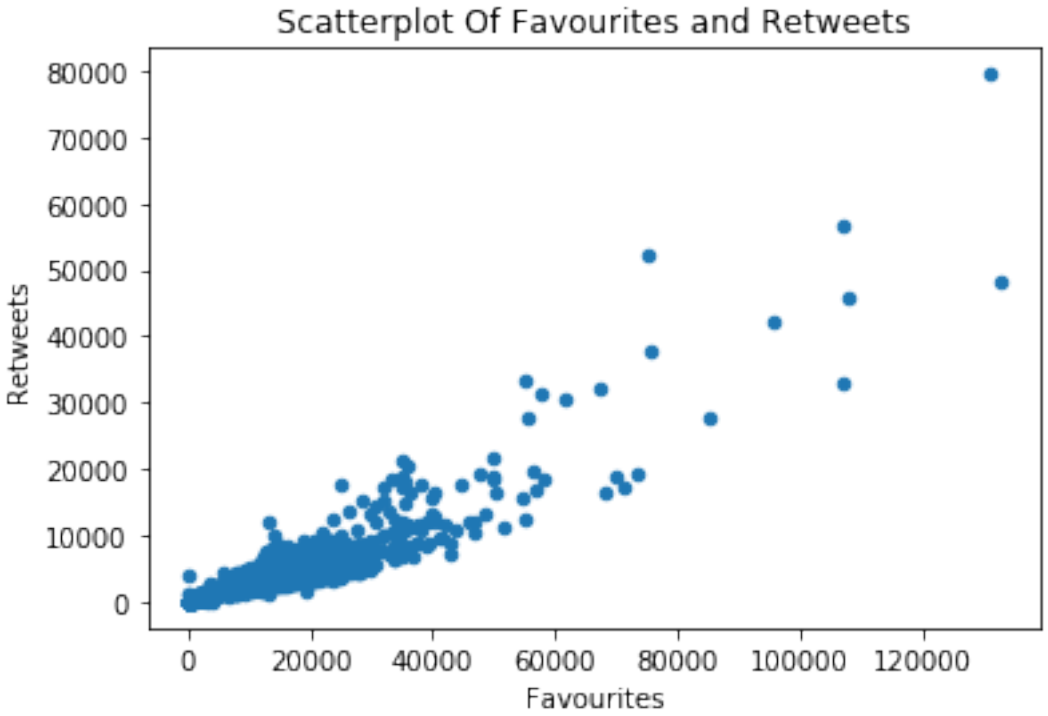
```
In [23]: twitr_ana_df['rating_obtained'].plot()
         plt.ylim(0,5)

         plt.ylabel('ratio')
         plt.xticks(rotation=90)
         plt.show()
```



So now we see that the ratings are mostly ranging around from 0 to 1 except for few occassions where the rating happened exorbitantly high by few users. Maybe they were extremely happy or so and did that.

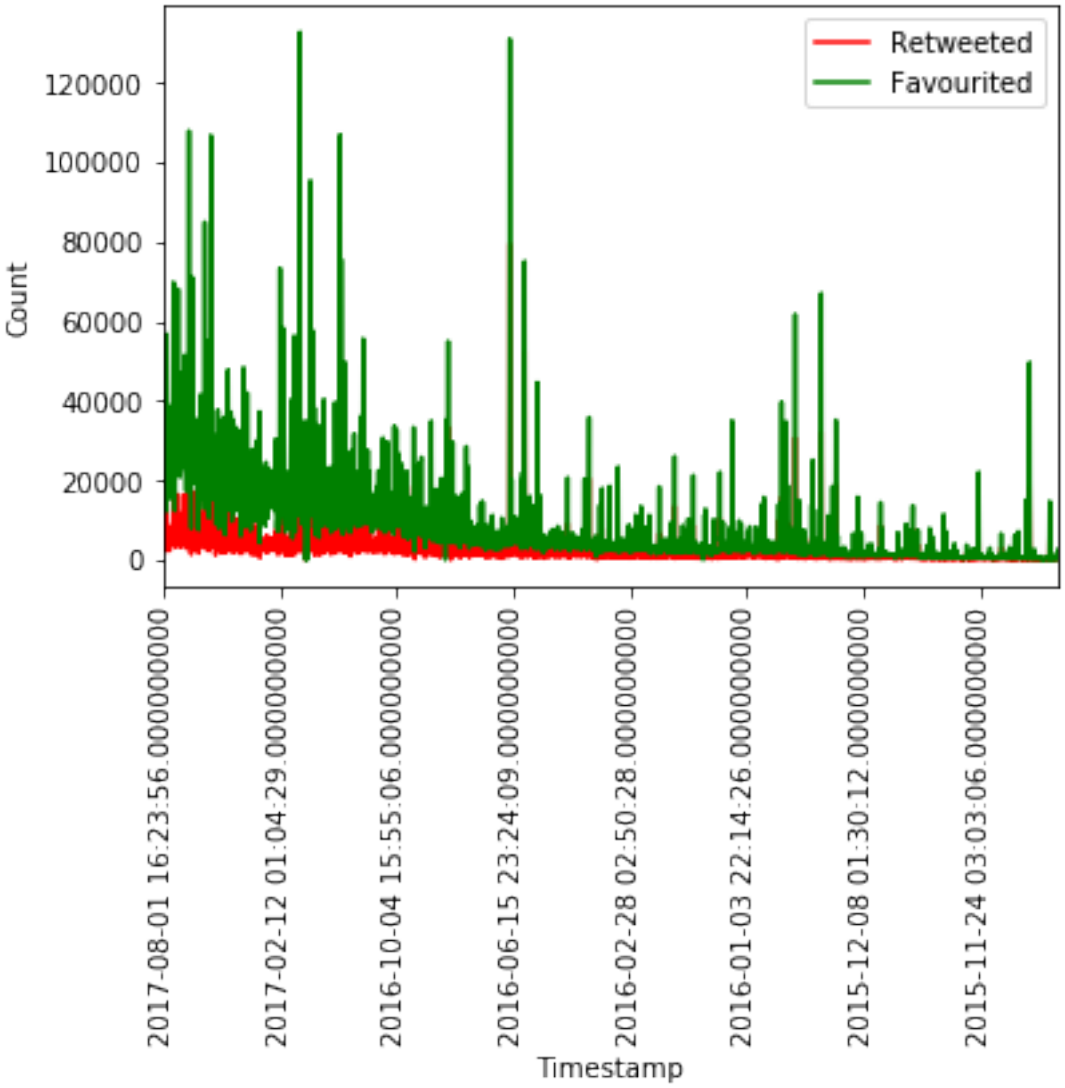
```
In [8]: twitr_ana_df.plot(kind='scatter',x='favourites',y='retweets')
        plt.xlabel('Favourites')
        plt.ylabel('Retweets')
```

```
plt.title('Scatterplot Of Favourites and Retweets')
plt.show()
```



The scatter plot shows that the retweets and favourites have strong positive correlation, meaning if a tweet is liked then it is most likely that it will be retweeted too.

```
In [24]: twitr_ana_df['retweets'].plot(color='red',label='Retweeted')
         twitr_ana_df['favourites'].plot(color='green',label='Favourited')
         plt.xlabel('Timestamp')
         plt.ylabel('Count')
         plt.xticks(rotation=90)
         plt.legend()
         plt.show()
```



We can see that the favourited numbers are generally more than the retweets.

In [ ]: