act_report

June 29, 2022

0.1 Report: act_report

- Create a **250-word-minimum written report** called "act_report.pdf" or "act_report.html" that communicates the insights and displays the visualization(s) produced from your wrangled data. This is to be framed as an external document, like a blog post or magazine article, for example.

REPORT ANALYSIS

The data used for this project consisted of three different datasets that were obtained as following:

1. Udacity provided the first dataset which is a csv file named twitter_archive_enhanced.csv.
2. The second dataset was a tsv file named image_prediction.tsv which was hosted on udacity server and I programmatically downloaded the file.
3. For the third dataset, I downloaded it from Udacity Server as an alternative to getting from api, "tweet_json_text" was

I combined the three datasets to generate 'twitter_archive_master.csv' dataset ,

I perfomed some programmatic methods as follows;

```
In [7]: ## For the purpose for the Report , I will name the dataframe  - "dataset2".
        import pandas as pd
        dataset2 = pd.read_csv("twitter_archive_master.csv")

In [8]: dataset2.describe()
```

Out[8]:

	tweet_id	rating_numerator	rating_denominator	img_num
count	2.073000e+03	2073.000000	2073.000000	2073.000000
mean	7.383634e+17	12.265798	10.511819	1.203570
std	6.780118e+16	40.699924	7.180517	0.561856
min	6.660209e+17	0.000000	2.000000	1.000000
25%	6.764706e+17	10.000000	10.000000	1.000000
50%	7.119681e+17	11.000000	10.000000	1.000000
75%	7.931959e+17	12.000000	10.000000	1.000000
max	8.924206e+17	1776.000000	170.000000	4.000000

```
           p1_conf       p2_conf       p3_conf  retweet_count  favorite_count  \
count  2073.000000  2.073000e+03  2.073000e+03    2073.000000     2073.000000
mean      0.594532  1.346665e-01  6.034005e-02    2976.089243     8556.718283
std       0.271234  1.006830e-01  5.092769e-02    5054.897526    12098.640994
min       0.044333  1.011300e-08  1.740170e-10      16.000000        0.000000
25%       0.364095  5.390140e-02  1.619920e-02     634.000000     1674.000000
50%       0.588230  1.186220e-01  4.947150e-02    1408.000000     3864.000000
75%       0.843911  1.955730e-01  9.193000e-02    3443.000000    10937.000000
max       1.000000  4.880140e-01  2.734190e-01   79515.000000   132810.000000

       followers_count  friends_count
count     2.073000e+03         2073.0
mean      3.200945e+06          104.0
std       4.497651e+01            0.0
min       3.200799e+06          104.0
25%       3.200900e+06          104.0
50%       3.200947e+06          104.0
75%       3.201002e+06          104.0
max       3.201018e+06          104.0
```

In [9]:
```
dataset2.head(5)
```

Out[9]:
```
             tweet_id            timestamp  \
0  892420643555336193  2017-08-01 16:23:56
1  892177421306343426  2017-08-01 00:17:27
2  891815181378084864  2017-07-31 00:18:03
3  891689557279858688  2017-07-30 15:58:51
4  891327558926688256  2017-07-29 16:00:24

                                              source  \
0  <a href="http://twitter.com/download/iphone" r...
1  <a href="http://twitter.com/download/iphone" r...
2  <a href="http://twitter.com/download/iphone" r...
3  <a href="http://twitter.com/download/iphone" r...
4  <a href="http://twitter.com/download/iphone" r...

                                                text  rating_numerator  \
0  0       This is Phineas. He's a mystical boy. ...                13
1  0       This is Phineas. He's a mystical boy. ...                13
2  0       This is Phineas. He's a mystical boy. ...                12
3  0       This is Phineas. He's a mystical boy. ...                13
4  0       This is Phineas. He's a mystical boy. ...                12

   rating_denominator      name stage_attribute  \
0                  10   Phineas             NaN
1                  10     Tilly             NaN
2                  10    Archie             NaN
3                  10     Darla             NaN
```

```
4                  10  Franklin             NaN

                                          jpg_url  img_num      ...        \
0  https://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jpg        1      ...
1  https://pbs.twimg.com/media/DGGmoV4XsAAUL6n.jpg        1      ...
2  https://pbs.twimg.com/media/DGBdLU1WsAANxJ9.jpg        1      ...
3  https://pbs.twimg.com/media/DF_q7IAWsAEuuN8.jpg        1      ...
4  https://pbs.twimg.com/media/DF6hr6BUMAAzZgT.jpg        2      ...

                   p2   p2_conf  p2_dog                         p3  \
0              bagel  0.085851   False                     banana
1           Pekinese  0.090647    True                   papillon
2           malamute  0.078253    True                     kelpie
3  Labrador_retriever  0.168086    True                    spatula
4    English_springer  0.225770    True  German_short-haired_pointer

    p3_conf  p3_dog retweet_count  favorite_count  followers_count  \
0  0.076110   False          8853           39467          3200889
1  0.068957    True          6514           33819          3200889
2  0.031379    True          4328           25461          3200889
3  0.040836   False          8964           42908          3200889
4  0.175219    True          9774           41048          3200889

   friends_count
0            104
1            104
2            104
3            104
4            104

[5 rows x 23 columns]
```

```
In [10]: dataset2.groupby(['stage_attribute'])['followers_count'].sum()
```

```
Out[10]: stage_attribute
         doggo            214460749
         doggo,floofer      3200892
         doggo,pupper      35210032
         doggo,puppo        3200891
         floofer           22406470
         pupper           672198936
         puppo             73620871
         Name: followers_count, dtype: int64
```

0.1.1 Insights:

1. The minimum retweet_count is 16 while the maximum is 79515
2. The minimum favorite_count is 0 while the maximum is 132810

3. Dogs in Pupper stage has more followers

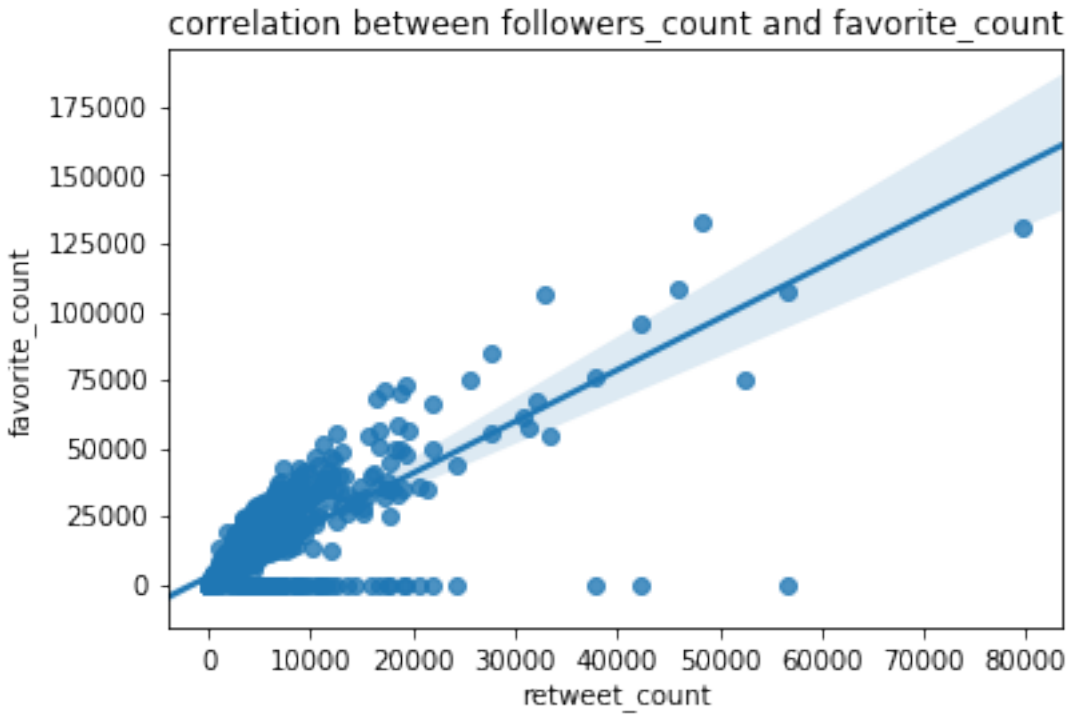
1 Does increase in followers_count has any effect in favorite_count?

```
In [15]: def scatter_plot(j, k, p):
             sns.regplot(x = dataset2[j], y = dataset2[k])
             plt.title(p)
             plt.xlabel(j)
             plt.ylabel(k)
             plt.show()
```

```
In [17]: import seaborn as sns
         import matplotlib.pyplot as plt
         %matplotlib inline

         scatter_plot('retweet_count','favorite_count','correlation between followers_count and

         # from the chart , it shows that there is a strong correlation between followers_count
```

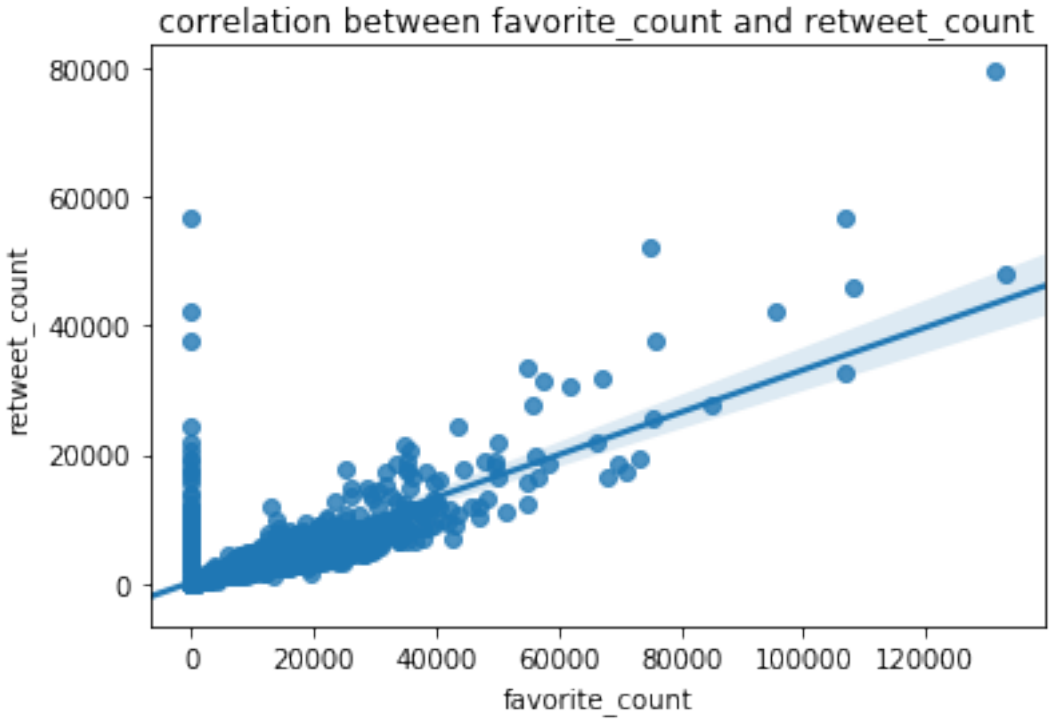


There is a strong relationship between followers_count and favorite_count.

2 Does increase in favorite_count has any effect in retweet_count?

```
In [19]: scatter_plot('favorite_count','retweet_count','correlation between favorite_count and r

         # from the chart , it shows that there is a strong correlation between favorite_count a
```



from the chart above, it shows that there is a strong correlation between favorite_count and retweet_count

```
In [ ]:
```