ACT REPORT

This is a report on the analysis and visualization of the twitter_archive_master.I used describe() to find the description of the different variables in twitter_archive_master, I used corr() to know how the different variables in twitter_archive_master correlate with each other and I finally use a scatterplot to visualize the relationship between favorite_count and retweet_count in twitter_archive_master.

Before starting the analysis I imported the following packages:

Import seaborn as sns
Import matplotlib.pyplot as plt
%matplotlib inline
from datetime import datetime as dt

To get a description of tweet_id, rating_numerator, rating_denominator, retweet_count, favorite_count, img_num, p1_conf, p2_conf and p3_conf in twitter_archive_master, I ran the following code
twitter_archive_master.describe ()

	tweet_id	rating_numerator	rating_denominator	retweet_count	favorite_count	img_num	p1_conf	p2_conf	p3_conf
count	2.097000e+03	2097.000000	2097.000000	2097.000000	2097.000000	1971.000000	1971.000000	1.971000e+03	1.971000e+03
mean	7.365594e+17	12.189318	10.448736	2841.667620	8980.902241	1.201928	0.594558	1.345850e-01	6.016556e-02
std	6.710178e+16	40.364996	6.645061	4709.238203	12199.636684	0.559020	0.272126	1.010527e-01	5.094156e-02
min	6.660209e+17	0.000000	2.000000	16.000000	81.000000	1.000000	0.044333	1.011300e-08	1.740170e-10
25%	6.768197e+17	10.000000	10.000000	644.000000	2046.000000	1.000000	0.363091	5.339800e-02	1.608055e-02
50%	7.098528e+17	11.000000	10.000000	1405.000000	4196.000000	1.000000	0.587764	1.173970e-01	4.944380e-02
75%	7.877176e+17	12.000000	10.000000	3285.000000	11416.000000	1.000000	0.847827	1.955655e-01	9.153815e-02
max	8.924206e+17	1776.000000	170.000000	79515.000000	132810.000000	4.000000	1.000000	4.880140e-01	2.734190e-01

*p1 has a higher mean, std, min, 25%, 50%, 75% and max than p2 and p3.

*The first prediction confidence is higher this shows that this prediction method works.

To know how tweet_id, rating_numerator, rating_denominator, retweet_count, favorite_count,img_num, p1_conf, p2_conf, p3_conf variables correlate with each other I used the following code:

twitter_archive_master.corr()

|	tweet_id	rating_numerator	rating_denominator	retweet_count	favorite_count	img_num	p1_conf	p2_conf	p3_conf
tweet_id	1.000000	0.024939	-0.017708	0.399295	0.650703	0.215396	0.104148	-0.001059	-0.046983
rating_numerator	0.024939	1.000000	0.185743	0.014683	0.011287	-0.003137	-0.006962	-0.019657	-0.004241
rating_denominator	-0.017708	0.185743	1.000000	-0.018487	-0.025235	-0.016962	0.008023	-0.031914	0.006861
retweet_count	0.399295	0.014683	-0.018487	1.000000	0.911201	0.108322	0.053674	-0.018573	-0.043431
favorite_count	0.650703	0.011287	-0.025235	0.911201	1.000000	0.138330	0.078291	-0.021465	-0.052545
img_num	0.215396	-0.003137	-0.016962	0.108322	0.138330	1.000000	0.202765	-0.156895	-0.137720
p1_conf	0.104148	-0.006962	0.008023	0.053674	0.078291	0.202765	1.000000	-0.510649	-0.707916
p2_conf	-0.001059	-0.019657	-0.031914	-0.018573	-0.021465	-0.156895	-0.510649	1.000000	0.480698
p3_conf	-0.046983	-0.004241	0.006861	-0.043431	-0.052545	-0.137720	-0.707916	0.480698	1.000000

I got the following insights from my correlation table:

- Retweet_count and favorite_count have a strong positive correlation of 0.911201 this means that as retweet_count increases also favorite_count increases.
- Favorite_count and tweet_id have also a strong positive correlation of 0.650703

To visualize the relationship between favorite_count and retweet_count in twitter_archive_master. I visualized the relationship using the following code. I plotted a scatter plot using the following code

plt.figure(figsize=(10,10))

plt.scatter(x=twitter_archive_master['favorite_count'], y=twitter_archive_master['retweet_count'])

plt.title('Favorite_count vs retweet_count')

plt.ylabel('retweet_count')

plt.xlabel('favorite_count')

plt.show()
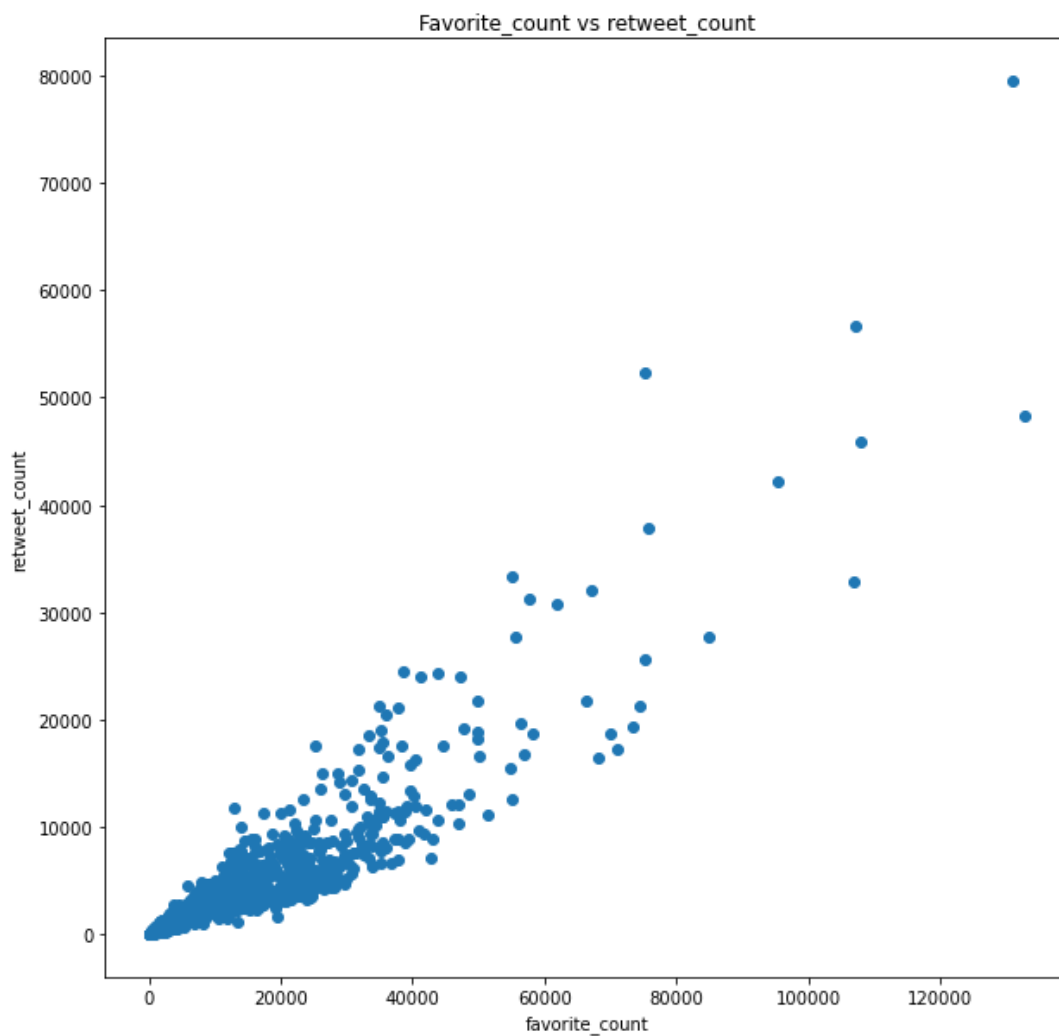
- This shows that favorite_count and retweet_count are linearly correlated an increase in favorite count leads to an increase in retweet_count