Report: act_report

- This act report documents contains my insights and displays the visualizations with detailed explanations, produced from the wrangled data.

The Analyzing and Visualizing Data section in the 'wrangle_act' notebook can also be called the Exploratory Data Analysis section, contained **three insights** inferred from the dataset after wrangling and **two Visualization** of the dataset after cleaning.

In [3]:
```
# importing the necessary libaries
import pandas as pd
%matplotlib inline
import seaborn as sns
import matplotlib.pyplot as plt
```

In [4]:
```
# using pandas to read the twitter_archive_master csv that was saved in the 'wrangle_act' notebook
df_clean = pd.read_csv("twitter_archive_master.csv")
```

First, I did a quick statistical analysis of the cleaned data by using the code below:`

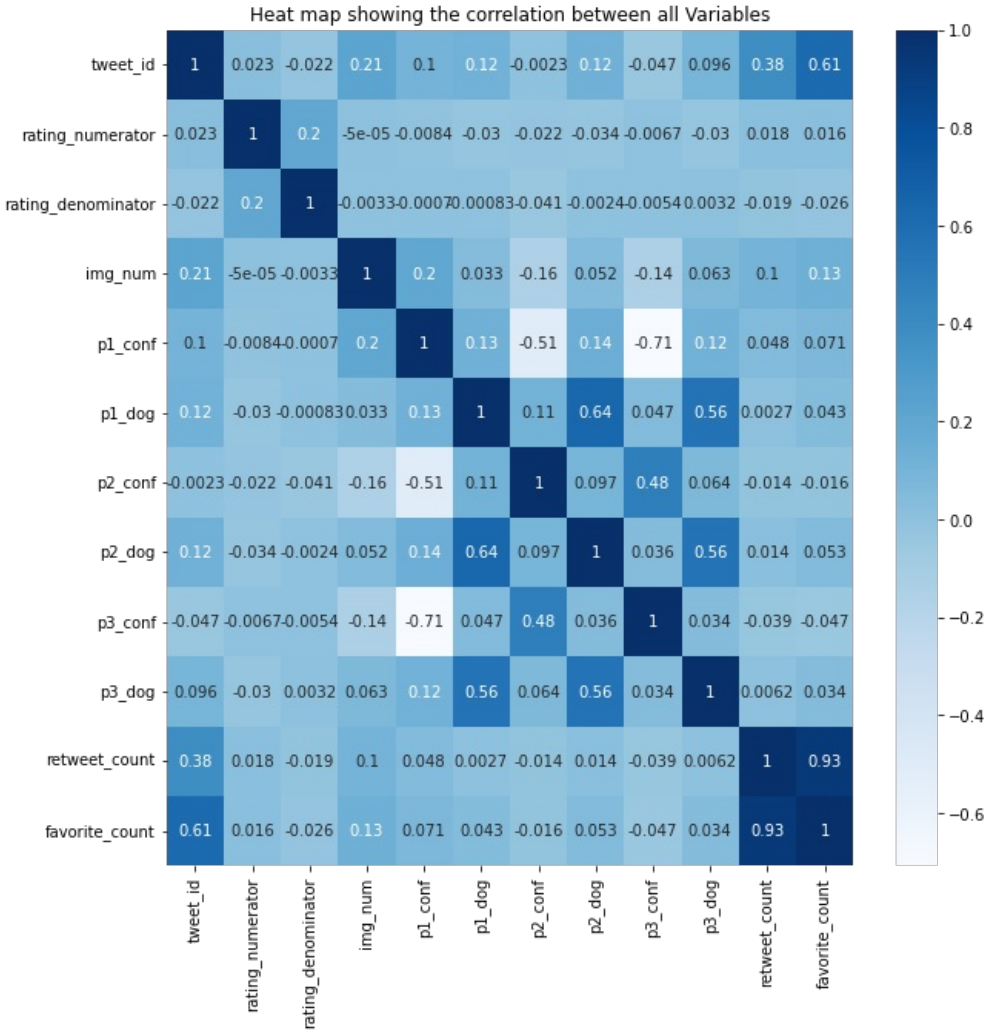
In [5]:
```
df_clean.describe()
```

Out[5]:

	tweet_id	rating_numerator	rating_denominator	img_num	p1_conf	p2_conf	p3_conf	retweet_count	favorite_co
count	1.986000e+03	1986.000000	1986.000000	1986.000000	1986.000000	1.986000e+03	1.986000e+03	1986.000000	1986.000
mean	7.356142e+17	12.231007	10.512085	1.203424	0.593452	1.344853e-01	6.034994e-02	2242.210977	7706.950
std	6.740686e+16	41.544680	7.276068	0.561492	0.271961	1.005944e-01	5.091948e-02	4016.627067	11370.339
min	6.660209e+17	0.000000	7.000000	1.000000	0.044333	1.011300e-08	1.740170e-10	11.000000	66.000
25%	6.758214e+17	10.000000	10.000000	1.000000	0.362656	5.407533e-02	1.624755e-02	494.500000	1636.250
50%	7.082494e+17	11.000000	10.000000	1.000000	0.587357	1.175370e-01	4.952715e-02	1079.000000	3463.000
75%	7.873791e+17	12.000000	10.000000	1.000000	0.844920	1.951377e-01	9.166433e-02	2556.750000	9556.250
max	8.924206e+17	1776.000000	170.000000	4.000000	1.000000	4.880140e-01	2.734190e-01	70689.000000	144829.000

Next, I plotted a heat map with the sole purpose of showing the correlation between all variables. The code was the heat map plot is:

In [6]:
```
plt.figure(figsize=(10,10))
sns.heatmap(df_clean.corr(),cbar=True,annot=True,cmap='Blues')
plt.title('Heat map showing the correlation between all Variables');
```



Then, I went ahead to state my findings.

The three insights are as follows :

1. The value of the minimum, first quartile, median, third quartile and the maximum favorite count is higher than that of the respective retweet count, I can infer that people generally like to favorite a tweet than retweet it.
2. From the heat map displayed above, we can see that there is a strong correlation between the retweet count and favorite count columns, with correlation value of 0.93
3. The standard deviation of the rating_numerator column is 41.544680.

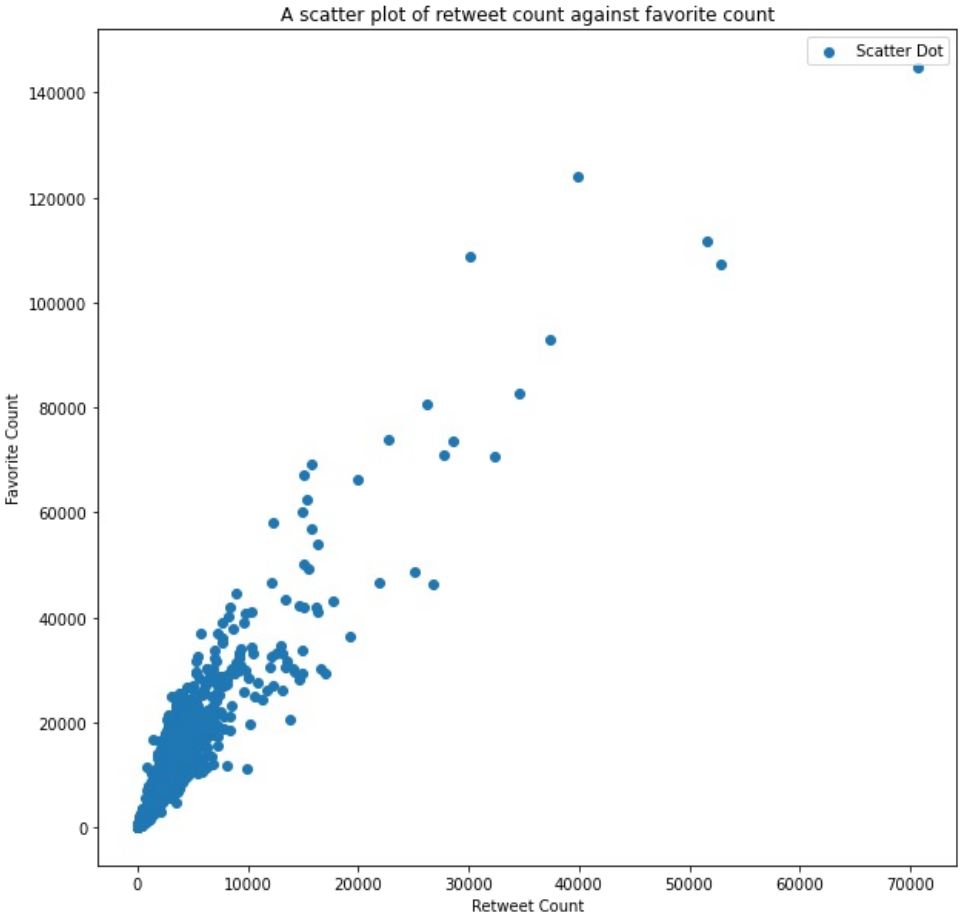
The first visualization displayed is the heat map shown above.

The second visualiation is a scatter plot that attempted to answer the question; "Is the relationship between the retweet_count and the favorite_count a positive or negative one?"

The code for the scatter plot is shown below:

In [7]:
```
# returns a scatter plot showing the relationship between the retweet_count and the favorite_count

plt.figure(figsize=(10,10))
plt.scatter(x='retweet_count',y='favorite_count',data=df_clean, label='Scatter Dot')
plt.title('A scatter plot of retweet count against favorite count')
plt.xlabel('Retweet Count')
plt.ylabel('Favorite Count')
plt.legend(loc='best');
```



The Retweet count is on the x-axis, and the Favorite count is on the y-axis.

From the distribution of the scatter plot shown above, it is evident that the relationship between both variables is a positive one. The reason for this relationship is not explored in this project, so therefore no conclusion will be made on it. But we can also see the value of the correlation between both variables from the heat map conducted earlier to be 0.93, that's a positive value and proves our positive relatonship theory. The value 0.93 equally tells us that it is a strong correlation.