act_report

September 16, 2019

1 Analysis on the Wrangled Data

```
In [2]: import requests
        import os
        import pandas as pd
        import json
        import tweepy
        import matplotlib.pyplot as plt
        %matplotlib inline
        pd.set_option('display.max_colwidth', -1)
        pd.set_option('display.max_rows', 500)
        plt.rcParams["figure.figsize"] = (10, 6) # (w, h)

        # Read data from 'twitter_archive_master.csv' into twitter_archive_master_df
        twitter_archive_master_df = pd.read_csv('twitter_archive_master.csv')
```

1.1 Insights

1.1.1 1. Top 10 Dog Breeds Tweeted (excluding the Unclassified Dog Breeds)

The Golden Retreiver is the most tweeted dog breed.

```
In [3]: dog_breeds_df = twitter_archive_master_df[twitter_archive_master_df['Dog_Breed'] != 'Unk
        dog_breeds_df.Dog_Breed.value_counts().nlargest(10)
```

```
Out[3]: Golden_retriever      156
        Labrador_retriever    106
        Pembroke              94
        Chihuahua             90
        Pug                   62
        Toy_poodle            50
        Chow                  48
        Samoyed               42
        Pomeranian            41
        Malamute              33
        Name: Dog_Breed, dtype: int64
```

1.1.2 2. Which Dog Stages are most tweeted? (excluding the ones which are not classified)

Pupper is the dog stage which is most tweeted. Followed by Doggo and Puppo. Floofer is the least tweeted dog stage

```
In [4]: dog_stages_df = twitter_archive_master_df[twitter_archive_master_df['Dog_Stage'] != 'Non
        dog_stages_df.Dog_Stage.value_counts()
```

```
Out[4]: pupper     222
        doggo      73
        puppo      24
        floofer    10
        Name: Dog_Stage, dtype: int64
```

1.1.3 3. Dogs identified with highest confidence by the Neural Network Program (excluding dogs without names)

The top twenty Dogs which were identified by the neural network program are listed below. The source of this data was the image_prediction.tsv file which was programatically downloaded from the provided URL

```
In [5]: dog_breed_pred_df = twitter_archive_master_df[['Dog_Name','Dog_Breed','Breed_Prediction_
        dog_breed_pred_df = dog_breed_pred_df[dog_breed_pred_df['Dog_Name'] != 'None']

        top20_count = dog_breed_pred_df.Breed_Prediction_Confidence.sort_values(ascending=[False
        dog_breed_pred_df[dog_breed_pred_df.index.isin(top20_count)]
```

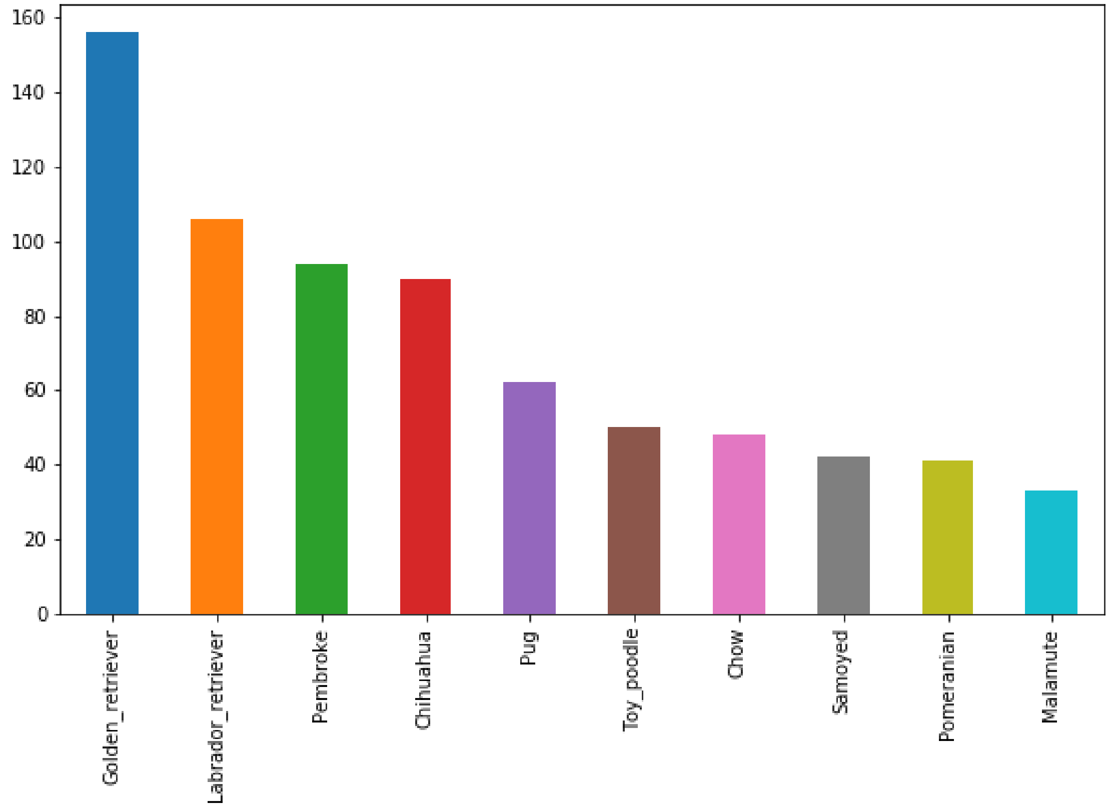
```
Out[5]:        Dog_Name             Dog_Breed  Breed_Prediction_Confidence
        254   Bob        Pug                   0.997445
        328   Sarge      Saint_bernard         0.998830
        338   Ulysses    Schipperke            0.997953
        369   Louis      Pomeranian            0.997210
        446   Olaf       Chow                  0.999837
        476   Panda      Pomeranian            0.997750
        542   Claude     French_bulldog        0.998544
        840   Kloey      Pomeranian            0.998275
        852   Ben        Blenheim_spaniel      0.998335
        916   Ozzy       Pug                   0.999365
        1198  Derek      Chow                  0.999823
        1284  Kyle       Pug                   0.996952
        1336  Oscar      Samoyed               0.998021
        1447  Stanley    Great_pyrenees        0.997692
        1496  Bell       Pug                   0.997310
        1730  Buddy      Chow                  0.999953
        1789  Pete       Old_english_sheepdog  0.999715
        1855  Cooper     Dalmatian             0.999828
        1911  Hubertson  Pug                   0.999044
        2047  Roscoe     French_bulldog        0.999201
```

1.2 Visualizations

1.2.1 1. Top 10 Dog Breeds Tweeted (excluding the Unclassified Dog Breeds)

The Golden Retreiver is the most Tweeted dog breed followed by Labrador Retreiver in second place and Pembroke just behind in third.

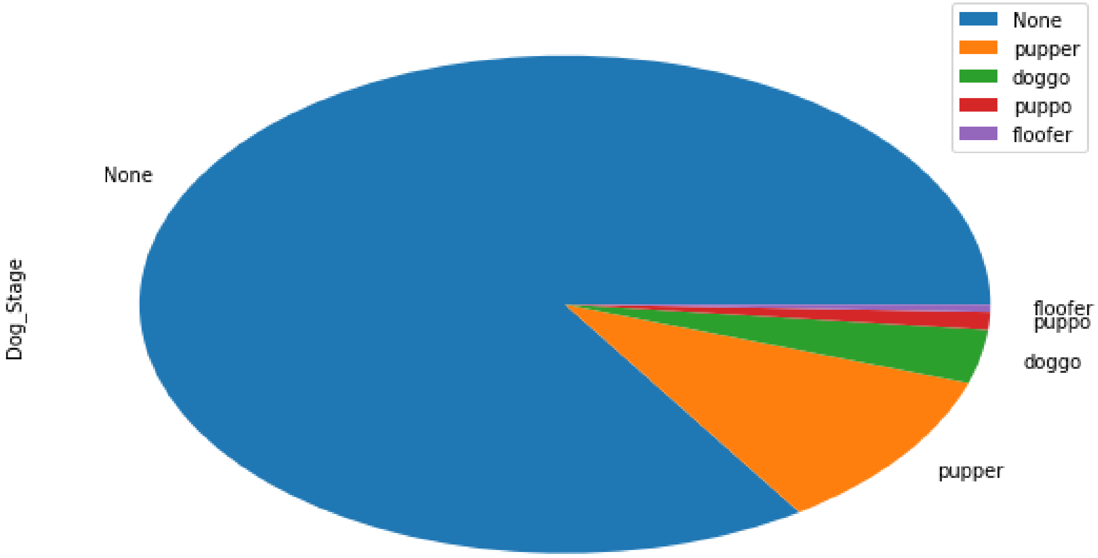
```
In [6]: plt.rcParams["figure.figsize"] = (10, 6) # (w, h)
        dog_breeds_df = twitter_archive_master_df[twitter_archive_master_df['Dog_Breed'] != 'Unk
        dog_breeds_df.Dog_Breed.value_counts().nlargest(10).plot(kind='bar');
```



1.2.2 2. Pie Chart Showing the distribution of Dog Stages

Majority of the tweets haven't been classified with 'Dog Stage'. For the ones which have been classified, Pupper forms the majority of the bunch

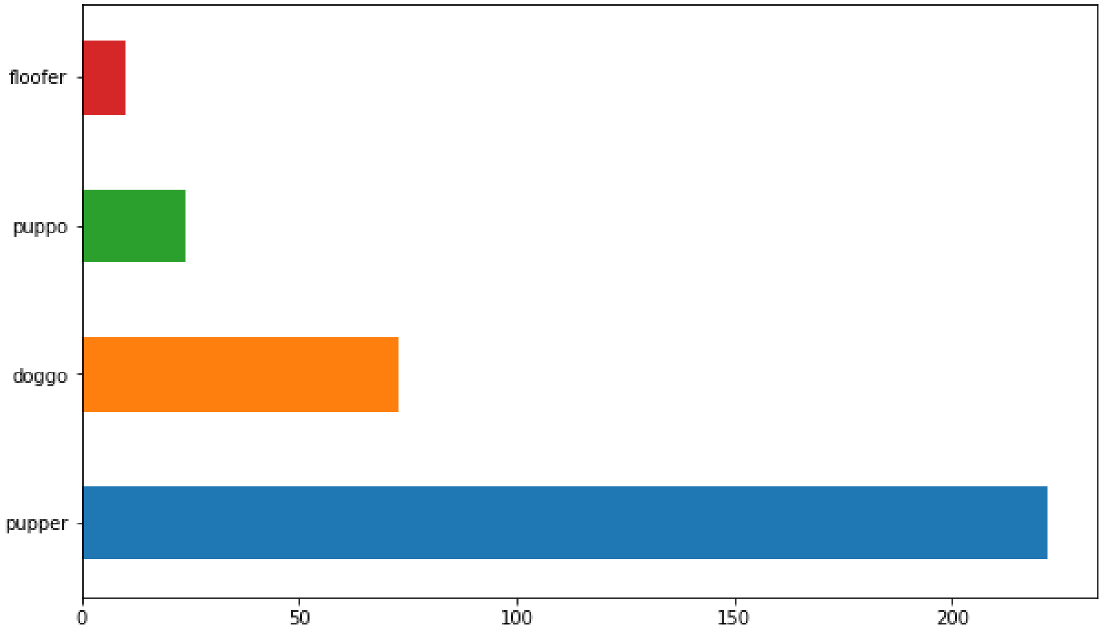
```
In [7]: dog_stage_df = twitter_archive_master_df[['Dog_Stage']]
        dog_stage_df.apply(pd.value_counts).plot.pie(subplots=True);
```



1.2.3 3. Popular Dog Stages (excluding the ones which are not classified)

Pupper is the dog stage which is most tweeted followed by doggo, puppo and floofer.

```
In [8]: plt.rcParams["figure.figsize"] = (10, 6) # (w, h)
        dog_stages_df = twitter_archive_master_df[twitter_archive_master_df['Dog_Stage'] != 'Non
        dog_stages_df.Dog_Stage.value_counts().plot(kind='barh');
```

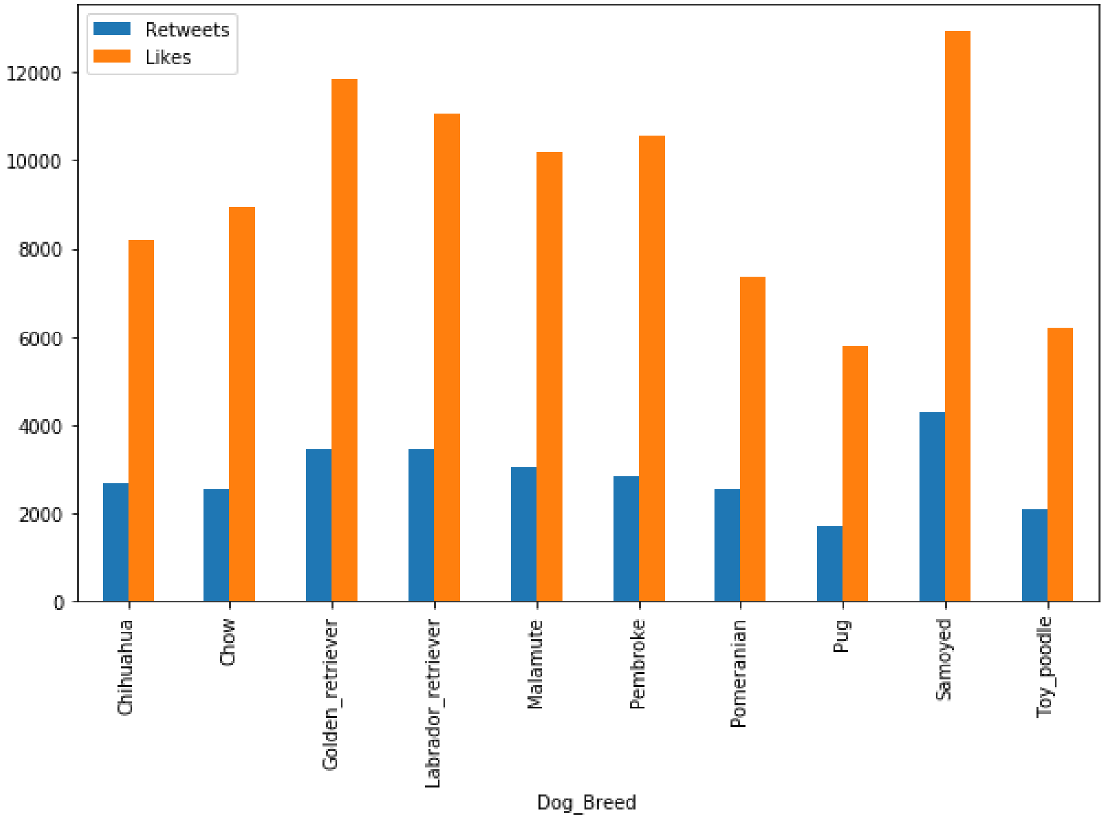


1.2.4 4. Average Retweets and Likes for the top 10 Dog Breeds

Following bar chart shows the average Retweets and average Likes for the top ten dog breeds. The dog breed Samoyed has the highest average number for Retweets and Likes.

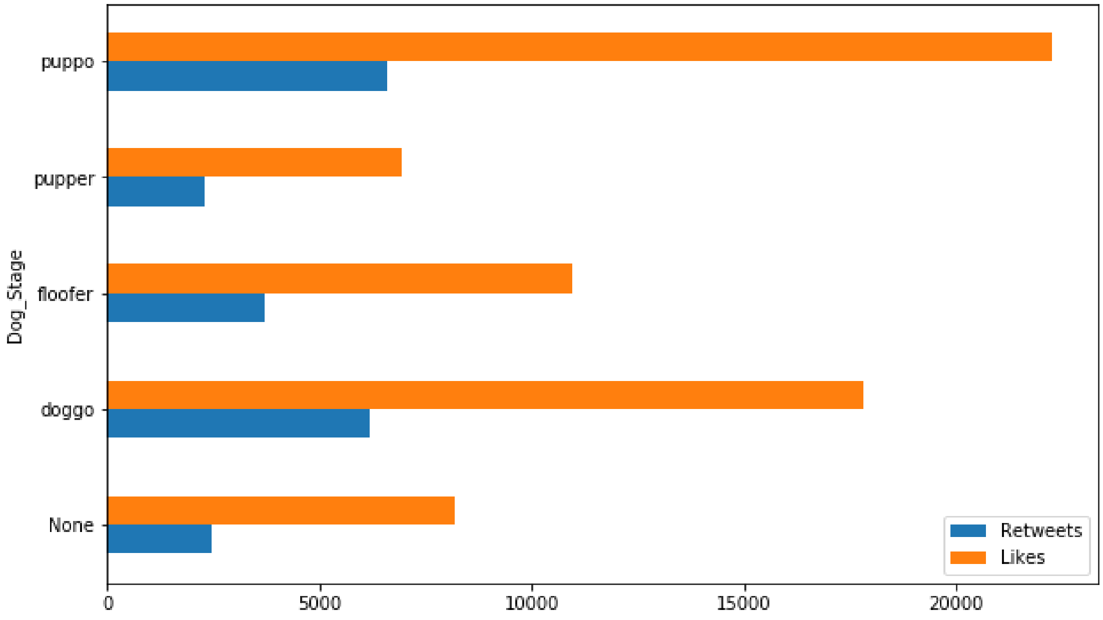
```
In [9]: dog_breed_retweet_df = twitter_archive_master_df[['Dog_Breed','Retweets','Likes']]
        dog_breed_retweet_df = dog_breed_retweet_df[dog_breed_retweet_df['Dog_Breed'] != 'Unknow

        top10_counts = dog_breed_retweet_df.Dog_Breed.value_counts()
        dog_breed_retweet_df[dog_breed_retweet_df.Dog_Breed.isin(top10_counts.nlargest(10).index
```



1.2.5 5. Retweets for a particular Dog_Stage

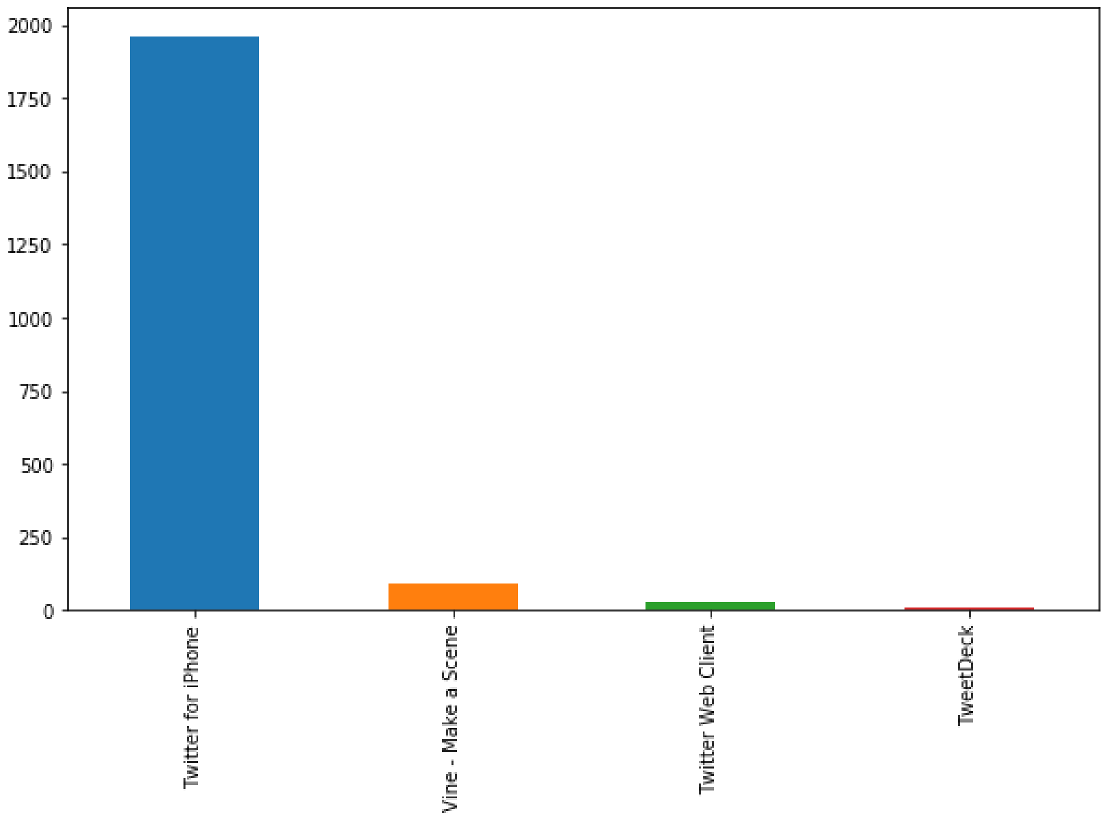
```
In [10]: dog_stage_retweet_df = twitter_archive_master_df[['Dog_Stage','Retweets','Likes']]
         dog_stage_retweet_df.groupby('Dog_Stage').mean().plot(kind='barh');
```



1.2.6 6. Visualize Number of Tweets from different Sources

The below bar chart show that 'Twitter for iPhone' is the platform which majority of the users tweet from. The number of tweets generated from iPhone twitter client is significantly more than all three Vine, Twitter Web Client and TweetDeck combined.

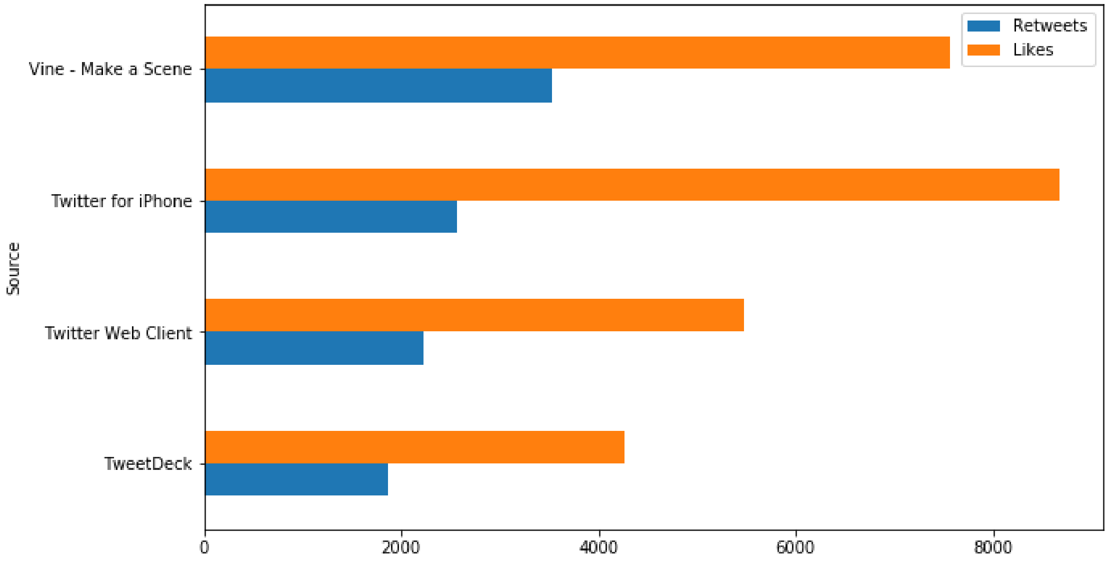
```
In [11]: twitter_archive_master_df.Source.value_counts().plot(kind='bar');
```



1.2.7 7. Visualize the average number of Retweets/Likes for tweets from different Devices(Sources)

The below horizontal bar chart shows the average number and Retweets and Likes on the tweets from the various platforms

```
In [12]: device_used_df = twitter_archive_master_df[['Source','Retweets','Likes']]
         device_used_df.groupby('Source').mean().plot(kind='barh');
```

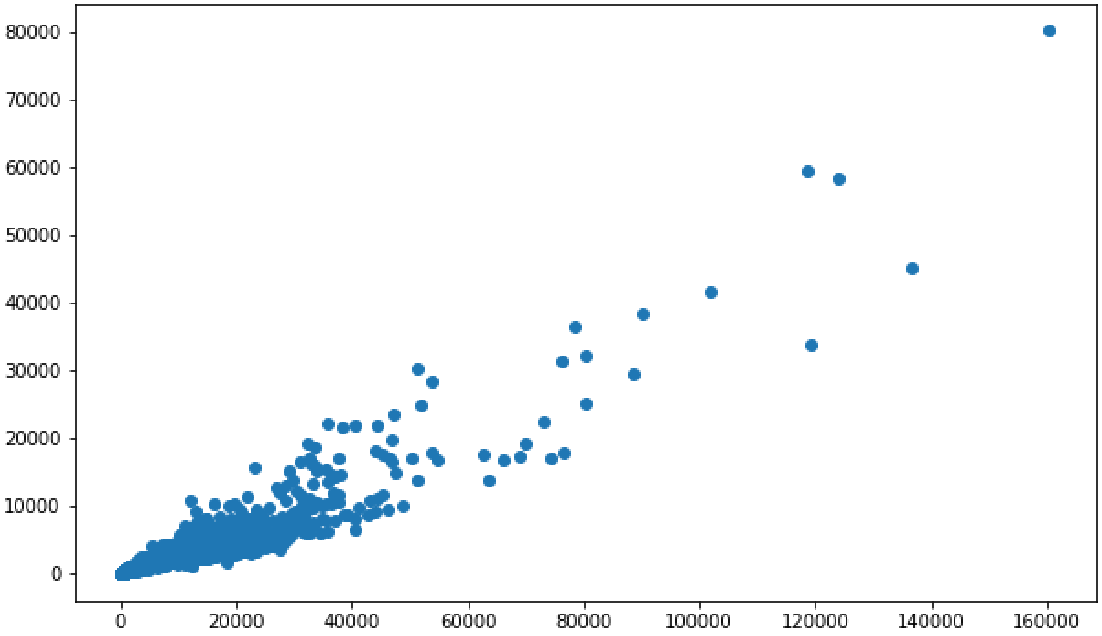


1.2.8 8. Visualize the correlation between the number of Retweets vs Likes

Below scatter plot shows the correlation between the number of Retweets vs the number of Likes. Both of these are directly propotional.

```
In [13]: x = twitter_archive_master_df.Likes
         y = twitter_archive_master_df.Retweets

         plt.scatter(x,y)
         plt.show()
```


80000
70000
60000
50000
40000
30000
20000
10000
0
0
20000
40000
60000
80000
100000
120000
140000
160000