act_report

December 16, 2022

0.1 Report: act_report

1 Report

After cleaning the data that was stored as "twitter_archive_master", the following questions were answered:

1. Top ten months with the highest favorite_count
2. Top ten months with the highest retweet_count
3. Influence of the day on the favorite_count
4. Finding the correlation between the Favorite and Retweet counts

2 Insights

1. favorite_count and retweet_count have been found to both reach their peaks in June. This can be rationally attributed to the fact that the dog festival normally occurrs during this period. Following this peak, January and December rank second and third for both favorite_count and retweet_count respectively. This may be due to an increase in festive activities during these periods.

2. Saturday usually has the highest favorite_count followed by Friday. This is probably due to most people not working on the weekends and having the time to scroll through Twitter.

3. Also, as expected, the correlation between favorite_count and retweet_count is positively very strong (0.86). Hence, favorited tweets are more likely to be retweeted.

4. On the other hand, the correlation between each feature (favorite_count and retweet_count) and both the numerator and denominator ratings show a very weak, positive relationship (for numerator_rating) and negative for denominator_rating.

3 Recommendations

1. It is prefferable that posts are posted on Fridays and Saturdays
2. Dog events should be hosted around June, December, or January
3. Another factor should be used in predicting the probability of retweeting as the numerator and denominator ratings are not effective

3.1 Write function for the visualization

```
In [1]: import pandas as pd
        import matplotlib.pyplot as plt
        df = pd.read_csv('twitter_archive_master.csv')
```

```
In [2]: def barhplot(x, y, xlabel, title):
            plt.figure(figsize=  (14, 8))
            plt.barh(x, y, align = 'center')
            plt.gca().invert_yaxis()
            plt.xlabel(xlabel, fontsize = 18)
            plt.title(title, fontsize = 18)
            plt.show();
```

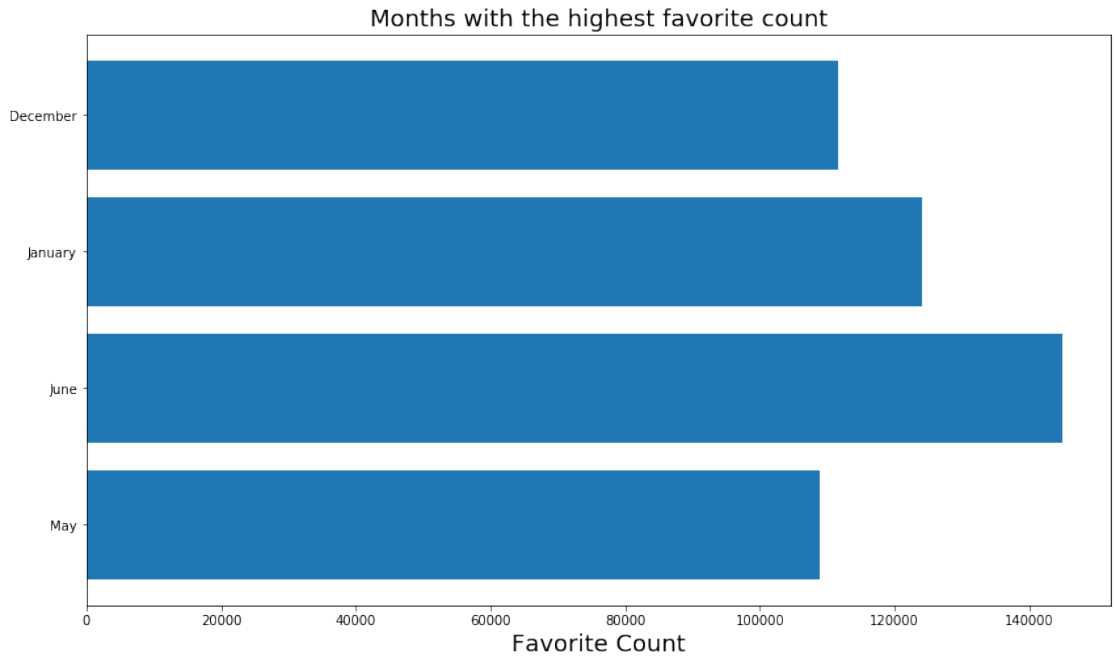
3.2 Top ten months with the highest favorite_count

```
In [3]: top = df.sort_values(by = 'favorite_count', ascending = False)
        top_10 = top[['tweet_id', 'source', 'favorite_count', 'month']].head(10)
        top_10
```

```
Out[3]:                tweet_id                                source  favorite_count  \
        836   744234799360020481  http://twitter.com/download/iphone          144897
        319   822872901745569793  http://twitter.com/download/iphone          124127
        422   807106840509214720  http://twitter.com/download/iphone          111710
        110   866450705531457537  http://twitter.com/download/iphone          108924
        871   739238157791694849  http://twitter.com/download/iphone          107253
        59    879415818425184262  http://twitter.com/download/iphone           92885
        348   819004803107983360  http://twitter.com/download/iphone           82714
        138   859196978902773760  http://twitter.com/download/iphone           80607
        94    870374049280663552  http://twitter.com/download/iphone           73941
        1485  678399652199309312  http://twitter.com/download/iphone           73528

                month
        836      June
        319   January
        422  December
        110       May
        871      June
        59       June
        348   January
        138       May
        94       June
        1485 December
```

```
In [4]: barhplot(top_10.month, top_10.favorite_count, "Favorite Count", "Months with the highest
```



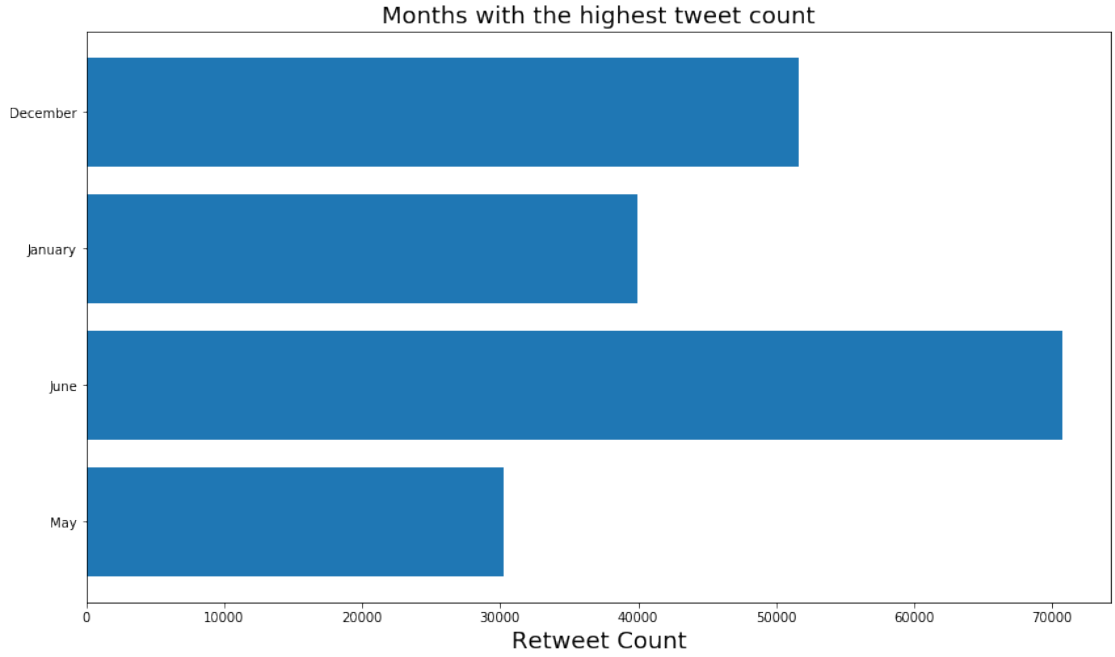
3.3 Top ten months with the highest retweet_count

```
In [5]: top = df.sort_values(by = 'retweet_count', ascending = False)
        top_10 = top[['tweet_id', 'source', 'text', 'retweet_count', 'month']].head(10)
        top_10.head()
```

Out[5]:

	tweet_id	source	text	retweet_count	month
836	744234799360020481	http://twitter.com/download/iphone	Here's a doggo realizing you can stand in a po...	70742	June
871	739238157791694849	http://twitter.com/download/iphone	Here's a doggo blowing bubbles. It's downright...	52908	June
422	807106840509214720	http://twitter.com/download/iphone	This is Stephan. He just wants to help. 13/10 ...	51687	December
319	822872901745569793	http://twitter.com/download/iphone	Here's a super supportive puppo participating ...	39926	January
59	879415818425184262	http://twitter.com/download/iphone	This is Duddles. He did an attempt. 13/10 some...	37457	June

```
In [6]: barhplot(top_10.month, top_10.retweet_count, "Retweet Count", "Months with the highest t
```



3.4 Influence of the day on favorite_count

```
In [7]: top = df.groupby(by = 'day')
        top = df.sort_values(by = 'favorite_count', ascending = False)
        top_10 = top[['tweet_id', 'source', 'text', 'favorite_count', 'day']].head(10)
        top_10.head()
```

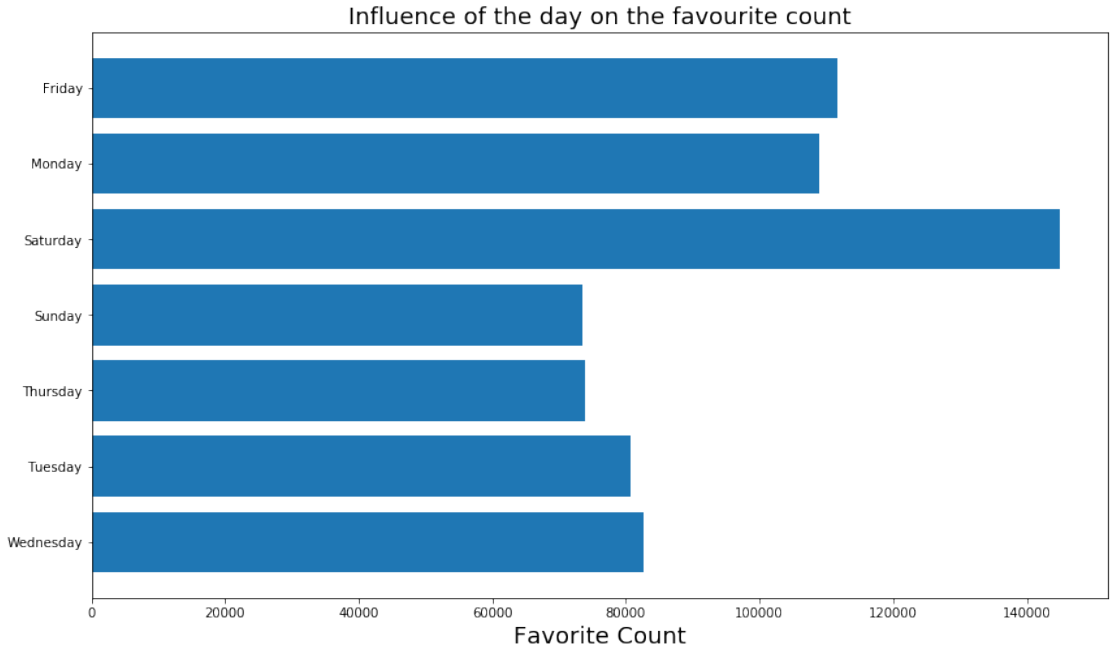
```
Out[7]:              tweet_id                                source  \
        836  744234799360020481  http://twitter.com/download/iphone
        319  822872901745569793  http://twitter.com/download/iphone
        422  807106840509214720  http://twitter.com/download/iphone
        110  866450705531457537  http://twitter.com/download/iphone
        871  739238157791694849  http://twitter.com/download/iphone

                                                    text  favorite_count  \
        836  Here's a doggo realizing you can stand in a po...          144897
        319  Here's a super supportive puppo participating ...          124127
        422  This is Stephan. He just wants to help. 13/10 ...          111710
        110  This is Jamesy. He gives a kiss to every other...          108924
        871  Here's a doggo blowing bubbles. It's downright...          107253

                 day
        836  Saturday
        319  Saturday
```

```
422    Friday
110    Monday
871  Saturday
```

```
In [8]: barhplot(top_10.day, top_10.favorite_count, "Favorite Count", "Influence of the day on t
```



3.5 Finding the correlation between the Favorite and Retweet counts

```
In [9]: df.favorite_count.corr(df.retweet_count)
```

```
Out[9]: 0.86102978252850915
```

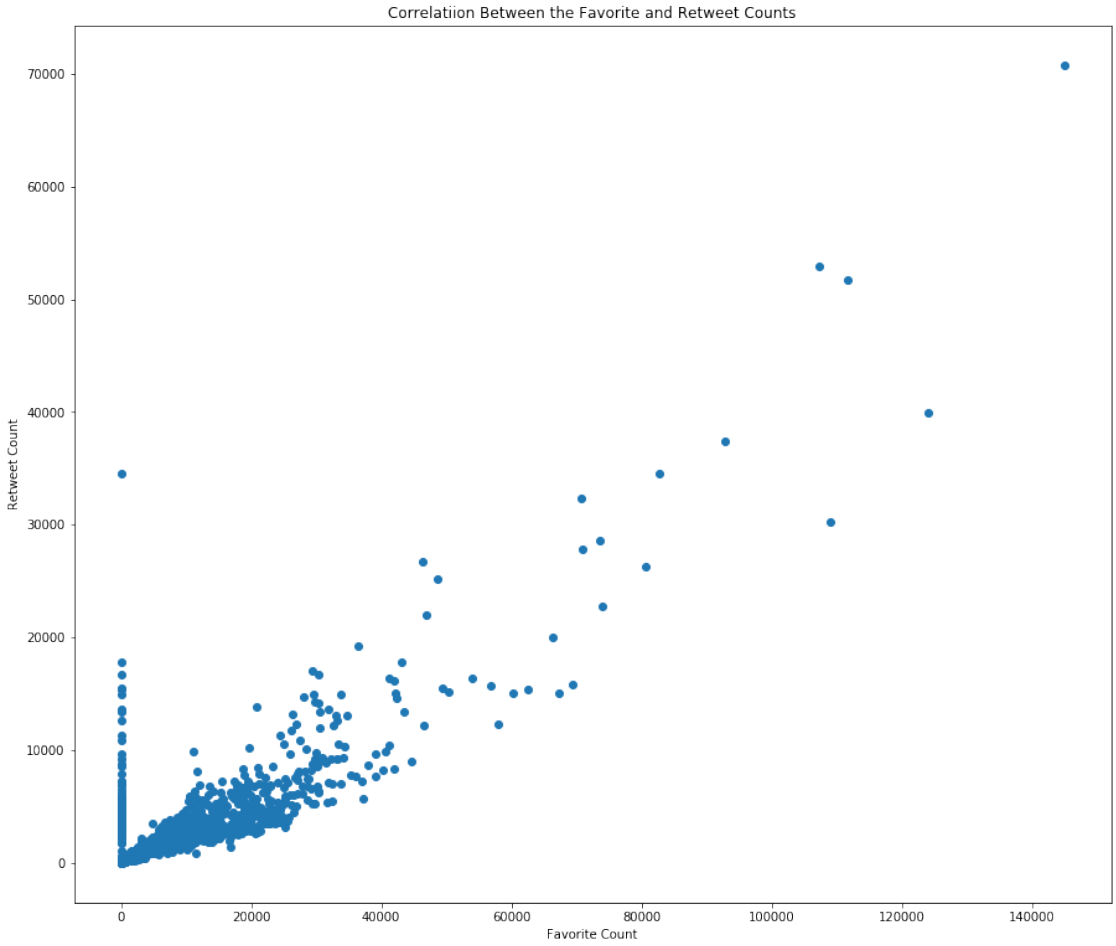
3.6 Visualization

```
In [10]: plt.figure(figsize=(15, 13))
         ax = plt.axes()
         ax.scatter(df.favorite_count, df.retweet_count)

         ax.set_xlabel('Favorite Count')
         ax.set_ylabel('Retweet Count')
         ax.set_title('Correlatiion Between the Favorite and Retweet Counts')

         ax.axis('tight')

         plt.show()
```




```
In [11]: df.favorite_count.corr(df.rating_numerator)
```

```
Out[11]: 0.01605740884435734
```

```
In [12]: df.favorite_count.corr(df.rating_denominator)
```

```
Out[12]: -0.025221199482836899
```

```
In [13]: df.retweet_count.corr(df.rating_numerator)
```

```
Out[13]: 0.01765920274621292
```

```
In [14]: df.retweet_count.corr(df.rating_denominator)
```

```
Out[14]: -0.021366610584886463
```