Sami Farooqui

For this project were asked to take data from a twitter account
WeRateDogs and first clean then analyze the data. After obtaining the data we
Were left with 3 untidy uncleaned datasets. After cleaning the data and combining it to one neat dataframe, I was able to perform some descriptive statistics.

	rating_numerator	rating_denominator	favorite_count	retweet_count	img_num	p1_conf	p2_conf	p3_conf
count	1299.000000	1299.000000	1299.000000	1299.000000	1299.000000	1299.000000	1.299000e+03	1.299000e+03
mean	12.843726	10.545804	8210.866051	2488.404157	1.187067	0.587034	1.370495e-01	6.144723e-02
std	51.147640	7.874498	11422.030839	4001.500373	0.540746	0.273638	1.018687e-01	5.202736e-02
min	1.000000	2.000000	80.000000	12.000000	1.000000	0.044333	1.011300e-08	1.740170e-10
25%	10.000000	10.000000	1697.500000	569.500000	1.000000	0.354703	5.438335e-02	1.649005e-02
50%	11.000000	10.000000	3794.000000	1238.000000	1.000000	0.578120	1.203580e-01	4.953060e-02
75%	12.000000	10.000000	10149.000000	2948.000000	1.000000	0.837040	1.982365e-01	9.470910e-02
max	1776.000000	170.000000	121434.000000	60038.000000	4.000000	1.000000	4.676780e-01	2.710420e-01

We can see from the chart that the mean rating was actually over 10, why was that? Upon further inspection we can also see that the max rating given was 1776/10. This could be one of the reasons why the average was so high, and we can see the rating was given because the dog was celebrating the fourth of July, and 1776 was the year The United States came into existence. I also decided to take a look at how many tweets got ratings over the 10 scale, and discovered to my surprise that 724 tweets were given ratings over 10. Ok, so that's why the mean rating was over 10. Now lets take a look at what rating was given most, and what names are most common.

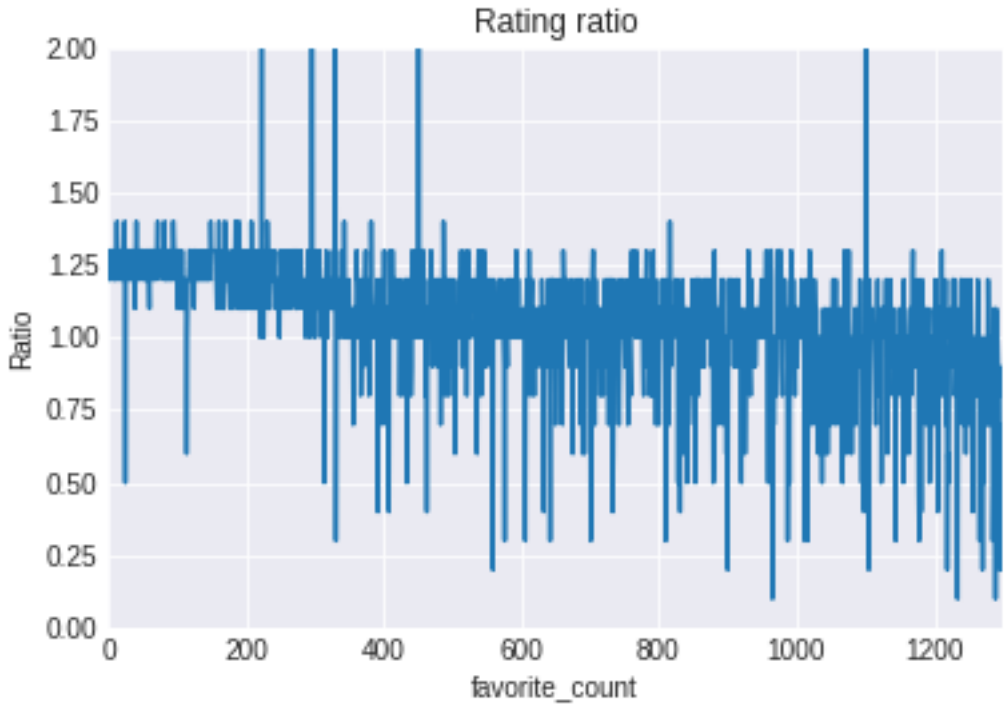
```
from collections import Counter
#Counting the most common rating
Counter(df_master['rating_numerator']).most_common(3)
```

```
[(10, 304), (12, 287), (11, 249)]
```

```
Counter(df_master['name']).most_common(3)
```

```
[('None', 434), ('Oliver', 8), ('Winston', 7)]
```

Now we can see that a rating of 10 was most often given, and in 434 of the tweets no names were given, names not beings given on tweets like these doesn't seem to uncommon, so that isn't too surprising, but we can also see that there were 8 dogs names Oliver on the twitter account. To me that is surprising, as I haven't heard of too many dogs named Oliver. Next I'd like to take a look at some graphs to see if there is any correlation between some of the variables in this data set. First we can divide the rating numerator by the denominator to obtain a rating ratio. This means that a ratio of 1 would be a full 10/10 and 0.5 would be 5/10. Now lets take look at the graph between ratings ratio and favorite tweets and see if there is any correlation.



Upon first inspection they don't seem to have too much of a correlation, however it seems to me that as the favorite count increased, the ratings ratio did not go over the 10/10 scale, I believe this is because more people liked to give honest ratings rather then something over the limit.
Next we can take a look at the correlation between favorites and retweets and see if there is any relationship. From the graph on page 4 we can see that there is a positive correlation between retweets and favorite count, as the retweets count increases so does the favorites count. This makes sense, as people will be more willing to retweet something they liked. Now we can conclude from this data that although many people try and give honest reviews of their dogs, outliers are always an issue and can skew statistical results often. We can also conclude that the more

liked tweets are retweeted more often, which in turn will garner more likes.