Wrangle Report

Phase 1: Gathering

I began this project by importing my packages and data sets. An enormous challenge for me at the start, was determining the proper code with which to access the twitter API. After several attempts and consultations with the internet, I found that a “for”, “try”, “except” loop was the only method that could extract real data from the source.

Phase 2: Assessing

After gathering the data, I began to assess it through a visual technique before utilizing the programmatic approach. On first glance, there were several inconsistencies in the twitter_df dataframe. I then explored these visual assessments using info(), duplicated(), describe(), an loc() functions.

In the end, I determined the eight quality and two tidiness issues that I would resolve in this project. In tweet_df, I would remove erroneous datatypes, change timestamp column to date time, find missing values for five columns, and remove the additional characters in front of the urls in the source column. In the image predictions dataframe, I needed to lowercase the p1_dog, p2_dog, and p3_dog columns, remove retweets, resolve confidence levels over 1 in the p2 and p3 columns, and strip away the duplicate images.

Phase 3: Cleaning

Upon resolving these issues in the cleaning steps, I was able to create an additional file with which to utilize the cleaned data to generate visualizations. I focused on favorite counts across time, favorite counts vs. retweet counts, representations of the different dog_type, as well as the rating numerators. From these visualizations came several key takeaways from this data set:

1. In fact, as the author tweeted more frequently (across time), the higher the numerator rating became. This could be due to the fact that a high numerator became more characteristic of the twitter user and thus became more of a signature.
2. Also, favorite count and retweet count are extremely positively correlated. Favorite and tweet counts are also positively correlated with tweet_id, suggesting that tweets increased in popularity with the passage of time/tweets.
3. Finally, the data shows that tweets are much more likely to receive a higher numerator rating at in the early and late hours of the evening: 0400, 2000, 2100. Perhaps due to an urge from the user to generate as much attention/shock from twitter followers as possible.

Project 3: Representations and Analysis ¶

)

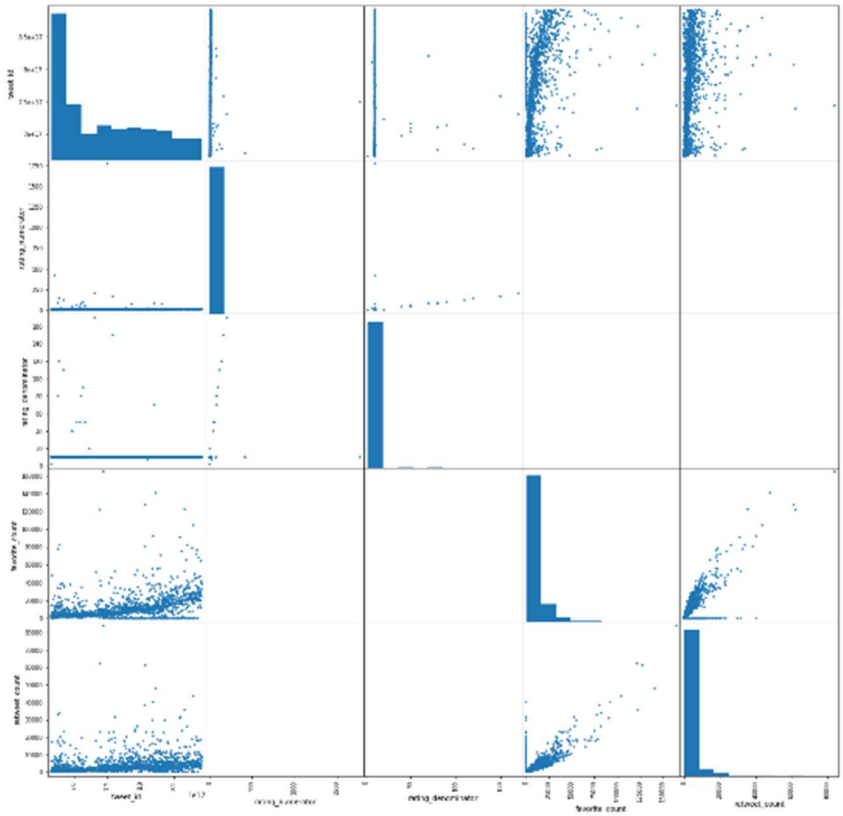
tweet_id	source	text	expanded_urls	rating_numerator	rating_denominator	name	timestamp	favo
892420643555336193	http://twitter.com/download/iphone	This is Phineas. He's a mystical boy. Only eve...	https://twitter.com/dog_rates/status/892420643...	13.0	10.0	Phineas	2017-08-01 16:23:56	
892177421306343426	http://twitter.com/download/iphone	This is Tilly. She's just checking pup on you....	https://twitter.com/dog_rates/status/892177421...	13.0	10.0	Tilly	2017-08-01 00:17:27	
891815181378084864	http://twitter.com/download/iphone	This is Archie. He is a rare Norwegian Pouncin...	https://twitter.com/dog_rates/status/891815181...	12.0	10.0	Archie	2017-07-31 00:18:03	
		This is Darla. She					2017-07-	

Out[4]:

	tweet_id	rating_numerator	rating_denominator	favorite_count	retweet_count
tweet_id	1.000000	0.024164	-0.022247	0.503767	0.378341
rating_numerator	0.024164	1.000000	0.185600	NaN	NaN
rating_denominator	-0.022247	0.185600	1.000000	NaN	NaN
favorite_count	0.503767	NaN	NaN	1.000000	0.797092
retweet_count	0.378341	NaN	NaN	0.797092	1.000000

As the above chart shows, tweet id and rating numerator are positively correlated. This indicates that, as the author tweeted more often, the higher the rating became. This could be due to the fact that a high numerator became more characteristic of the twitter user and thus became more of a signature

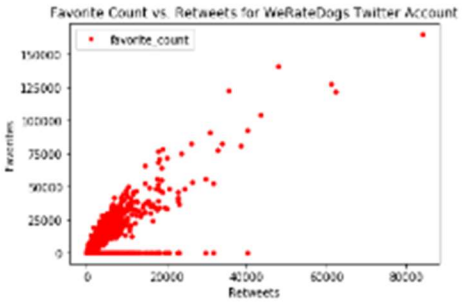
Also, favorite count and retweet count are extremely positively correlated. Favorite and tweet counts are also positively correlated with tweet_id, suggesting that tweets increased in popularity with the passage of time/tweets.



As the above chart shows, tweet id and rating numerator are positively correlated. This indicates that, as the author tweeted more often, the higher the rating became. This could be due to the fact that a high numerator became more characteristic of the twitter user and thus became more of a signature

Also, favorite count and retweet count are extremely positively correlated. Favorite and tweet counts are also positively correlated with tweet_id, suggesting that tweets increased in popularity with the passage of time/tweets.

```
Out[6]: Text(0.5,1,'Favorite Count vs. Retweets for WeRateDogs Twitter Account')
```



This plot demonstrates the obvious positive relationship between favorite count and retweet count, as they are both indicators of a tweet's popularity amongst users. What is interesting is the high amount of retweets for which there is little--to-no favorites, while the opposite is never true.