Data Wrangling Project Report — WeRateDogs®

As part of Udacity Data Analyst Nanodegree program, I will be working on a dog's dataset that is split in three different data-source formats.

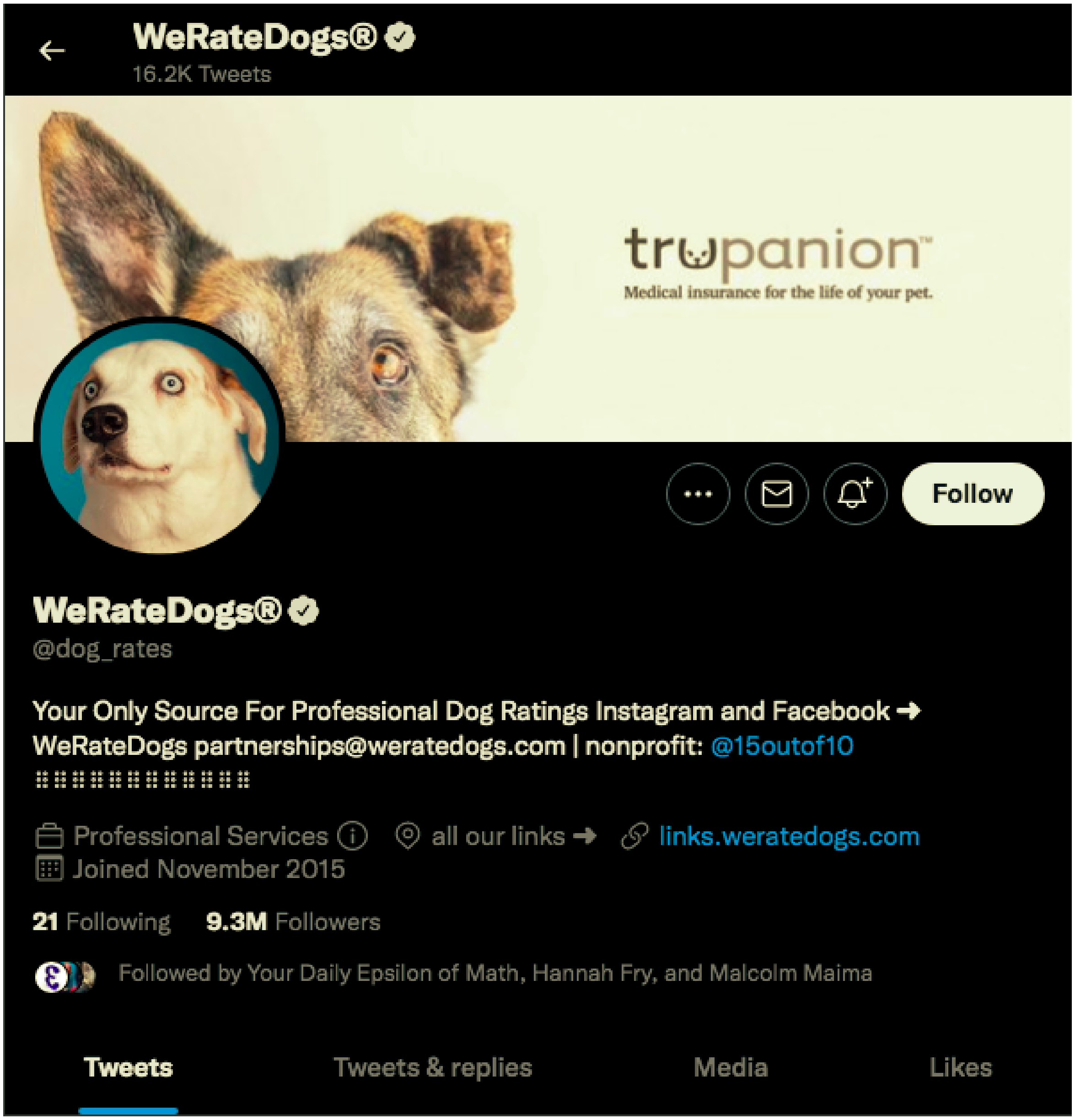
Working on this project will require gathering data from the three different data sources. After this, then go ahead and merge the three datasets for better and more comprehensive analysing of the data to answer a couple of questions later in the project.

The process is as follows:

1. Data Gathering: The goal is to gather all the three pieces of datasets. The provided dataset in csv format, twitter dataset through querying Twitter API and finally pragmatically through requests of the image prediction dataset.
2. Assess: After gathering the data, assess it to check its quality and tidiness. This will be done both visually and programmatically for replicability. Any data that falls into the low quality and messy category is highly unwanted and effort must be put to clean it.
3. Clean: After noting what needs to be cleaned in the assessment phase, effort is applied in this phase to remove the unwanted characteristics whether by transformation or dropping data points
4. Analyze: After cleaning, the data is analyzed with the goal of answering posed questions or observations through charting or calculations.
5. Understand the limitations
6. Conclusions: Report findings found after analysing the cleaned data.

For this report, I will include briefly my findings after analysing the assessed and cleaned data.

On page 2 and 3 respectively, is a pictorial view of WeRateDogs Twitter profile and sample post on the same account.



WeRateDogs®
16.2K Tweets

WeRateDogs®
@dog_rates

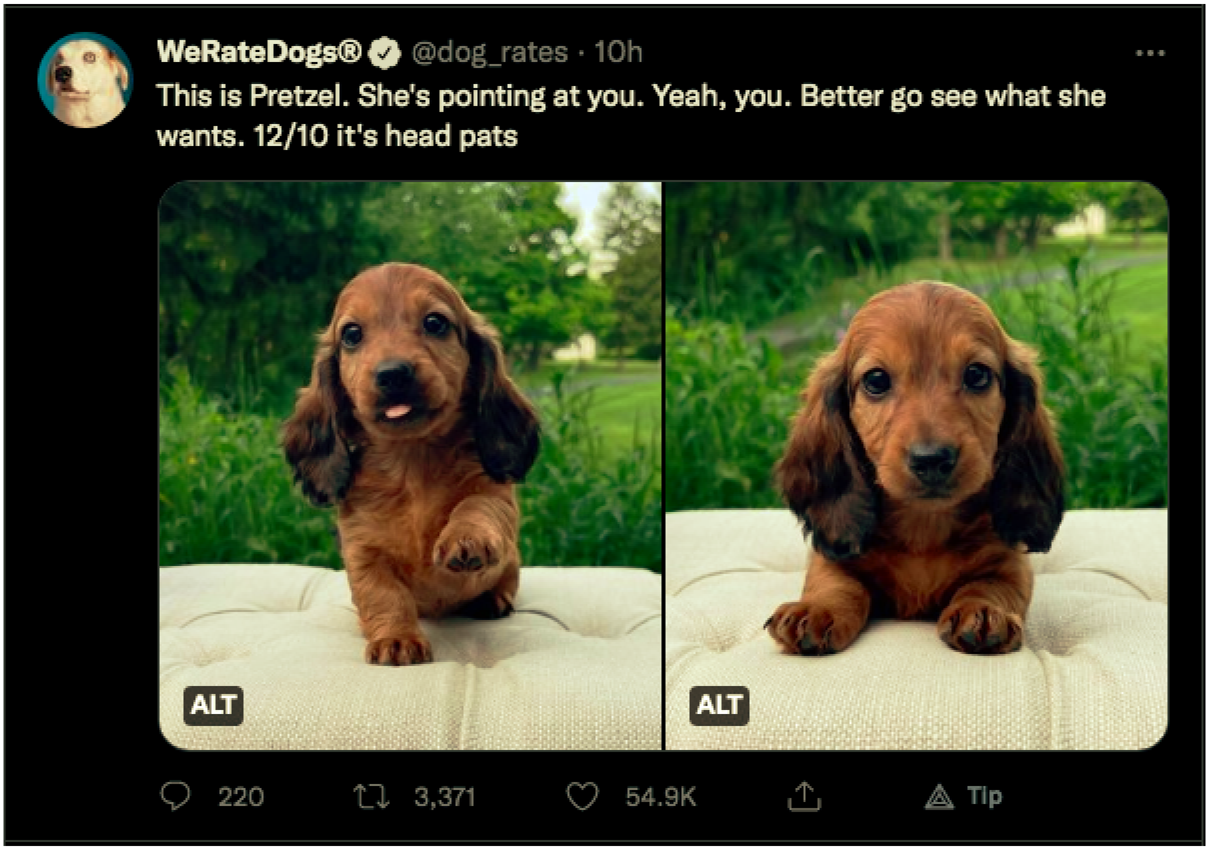
Your Only Source For Professional Dog Ratings Instagram and Facebook ➜ WeRateDogs partnerships@weratedogs.com | nonprofit: @15outof10

Professional Services ⓘ all our links ➜ links.weratedogs.com
Joined November 2015

21 Following 9.3M Followers

Followed by Your Daily Epsilon of Math, Hannah Fry, and Malcolm Maima

WeRateDogs is a Twitter account that rates people's dogs with humorous comments about the dogs. WeRateDogs has over 9 million followers and has received international media coverage as at the time of writing this.



WeRateDogs® @dog_rates · 10h

This is Pretzel. She's pointing at you. Yeah, you. Better go see what she wants. 12/10 it's head pats

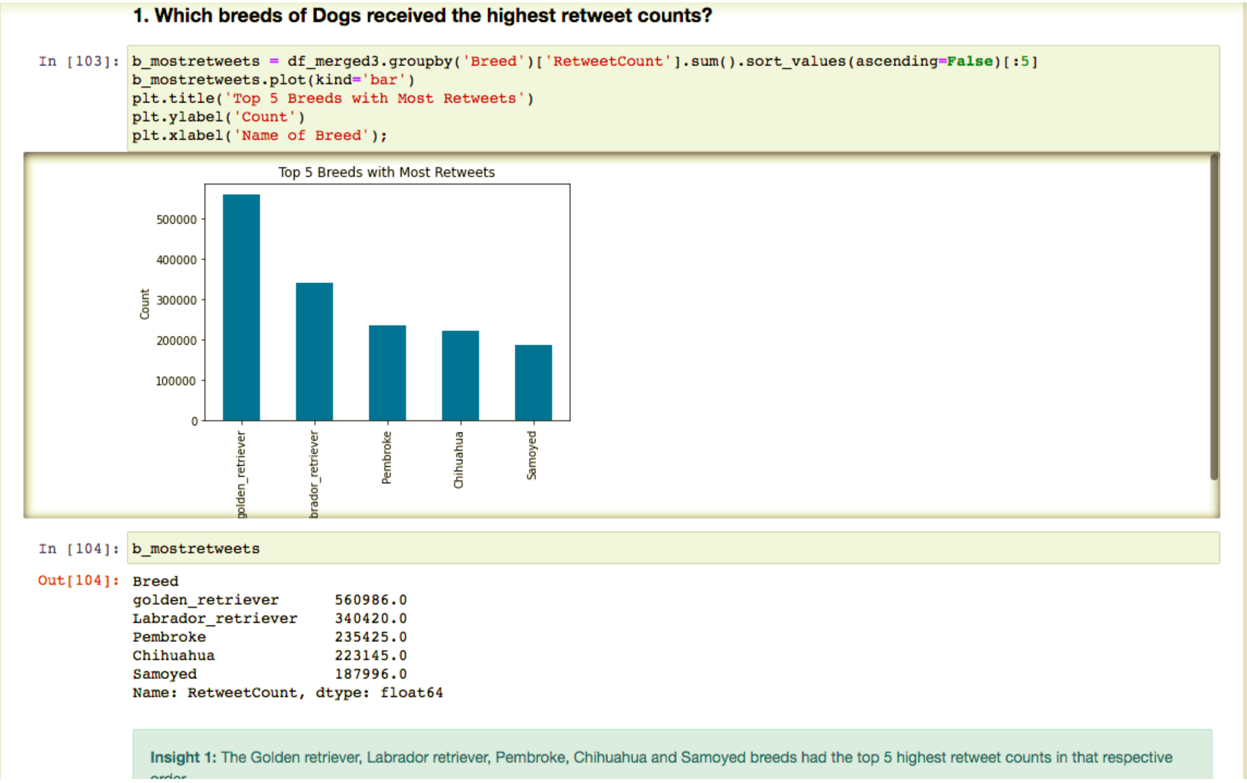
This is a sample of a tweet from the WeRateDogs Twitter account. As seen in the tweet, a few data points can be picked up. Here are a few of the data points:

1. Name of the dog: Pretzel
2. Rating: 12/10 with 12 as the numerator and 10 as the denominator
3. Image of the dog
4. Number of Retweets
5. Number of Likes
6. Number of Replies
7. Time of posting of the tweet

My findings and conclusions from analyzing the data are as follows and may include a pictorial view of the same.

Summaries

1. The Golden retriever, Labrador retriever, Pembroke, Chihuahua and Samoyed breeds had the top 5 highest retweet counts in that respective order.



1. Which breeds of Dogs received the highest retweet counts?

```
In [103]: b_mostretweets = df_merged3.groupby('Breed')['RetweetCount'].sum().sort_values(ascending=False)[:5]
          b_mostretweets.plot(kind='bar')
          plt.title('Top 5 Breeds with Most Retweets')
          plt.ylabel('Count')
          plt.xlabel('Name of Breed');
```

```
In [104]: b_mostretweets
```

```
Out[104]: Breed
          golden_retriever      560986.0
          Labrador_retriever    340420.0
          Pembroke              235425.0
          Chihuahua             223145.0
          Samoyed               187996.0
          Name: RetweetCount, dtype: float64
```

Insight 1: The Golden retriever, Labrador retriever, Pembroke, Chihuahua and Samoyed breeds had the top 5 highest retweet counts in that respective order

2. The Golden retriever, Labrador retriever, Pembroke, Chihuahua and French bulldog breeds had the top 5 highest favorite counts in that respective order.

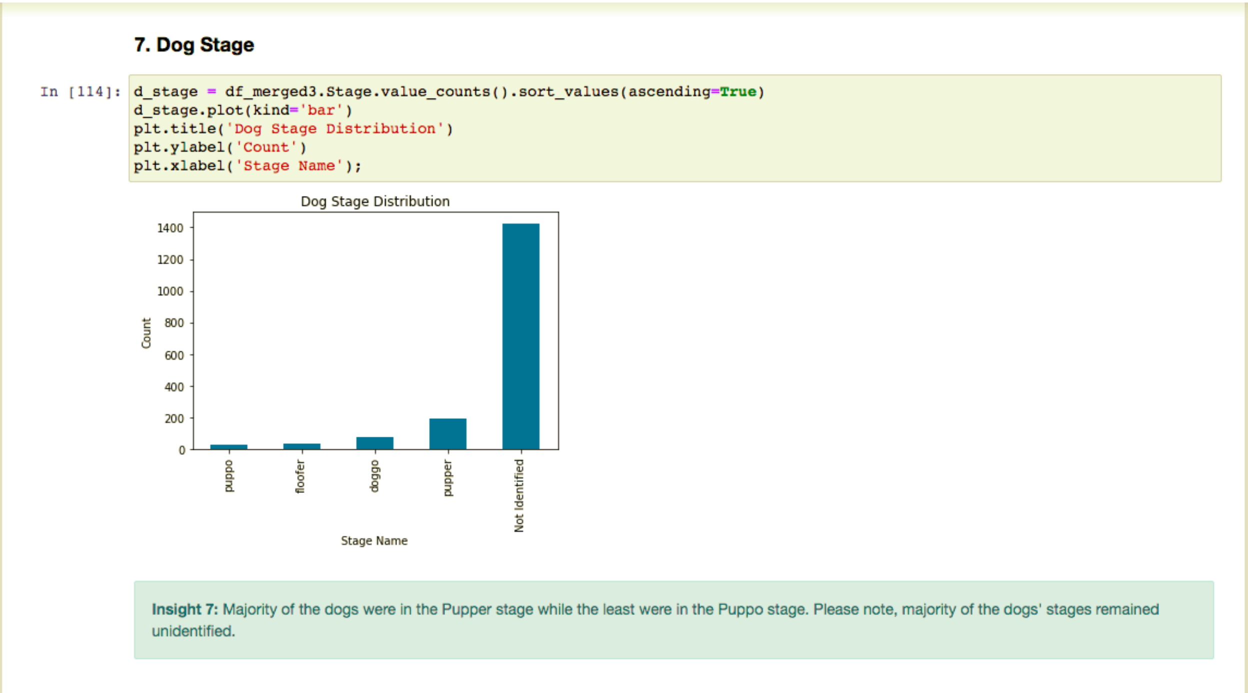
3. The Japanese Spaniel, Groenendael, EntleBucher, Irish Wolfhound and Brabancon Griffon breeds had the top 5 lowest retweet counts in that respective order.

4. The Brabancon Griffon, Groenendael, Standard Schnauzer,Irish Wolfhound and Japanese Spaniel breeds were the five least favorited breeds.

5. The Standard Poodle, Bedlington Terrier and Afghan Hound breeds were the three most averaged retweeted breeds in that respective order.

6. Tucker, Cooper, Lucy and Charlie were the most frequent names of the posted dogs.

7. Majority of the dogs were in the Pupper stage while the least were in the Puppo stage. Please note, the majority of the dogs' stages remained unidentified.

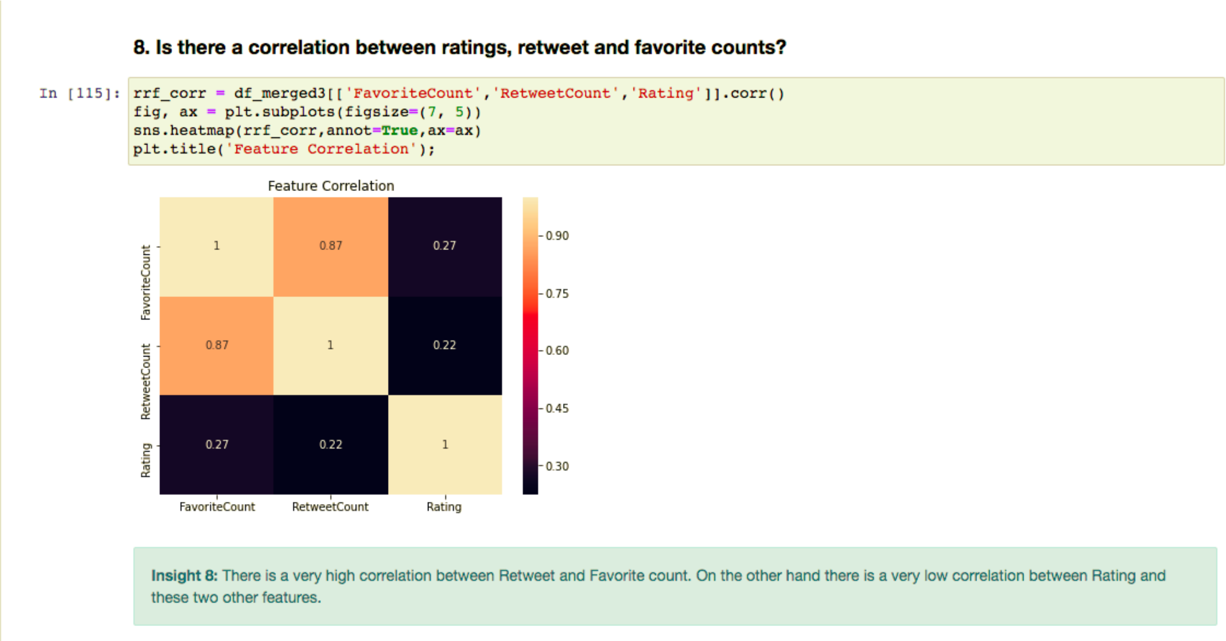


7. Dog Stage

```
In [114]: d_stage = df_merged3.Stage.value_counts().sort_values(ascending=True)
          d_stage.plot(kind='bar')
          plt.title('Dog Stage Distribution')
          plt.ylabel('Count')
          plt.xlabel('Stage Name');
```

Insight 7: Majority of the dogs were in the Pupper stage while the least were in the Puppo stage. Please note, majority of the dogs' stages remained unidentified.

8. There is a very high correlation between Retweet and Favorite count. On the other hand there is a very low correlation between Rating and these two other features.

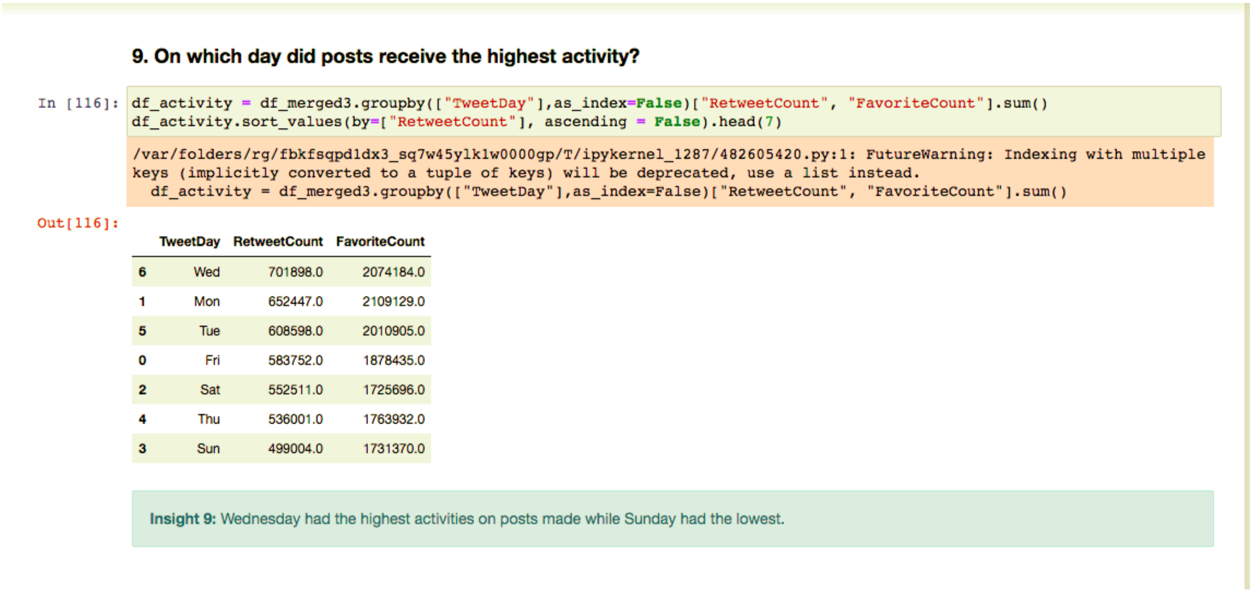


8. Is there a correlation between ratings, retweet and favorite counts?

```
In [115]: rrf_corr = df_merged3[['FavoriteCount','RetweetCount','Rating']].corr()
          fig, ax = plt.subplots(figsize=(7, 5))
          sns.heatmap(rrf_corr,annot=True,ax=ax)
          plt.title('Feature Correlation');
```

Insight 8: There is a very high correlation between Retweet and Favorite count. On the other hand there is a very low correlation between Rating and these two other features.

9. Wednesday had the highest activities on posts made while Sunday had the lowest.



9. On which day did posts receive the highest activity?

```
In [116]: df_activity = df_merged3.groupby(["TweetDay"],as_index=False)["RetweetCount", "FavoriteCount"].sum()
          df_activity.sort_values(by=["RetweetCount"], ascending = False).head(7)

/var/folders/rg/fbkfsqpdldx3_sq7w45ylklw0000gp/T/ipykernel_1287/482605420.py:1: FutureWarning: Indexing with multiple
keys (implicitly converted to a tuple of keys) will be deprecated, use a list instead.
  df_activity = df_merged3.groupby(["TweetDay"],as_index=False)["RetweetCount", "FavoriteCount"].sum()
```

Out[116]:

	TweetDay	RetweetCount	FavoriteCount
6	Wed	701898.0	2074184.0
1	Mon	652447.0	2109129.0
5	Tue	608598.0	2010905.0
0	Fri	583752.0	1878435.0
2	Sat	552511.0	1725696.0
4	Thu	536001.0	1763932.0
3	Sun	499004.0	1731370.0

Insight 9: Wednesday had the highest activities on posts made while Sunday had the lowest.

10. Majority of posts were made through Twitter for iPhone while the least were made through Vine.

Conclusions

- If you are going to make a dog post for purposes of getting high ratings in relation to WeRateDogs Twitter Account, you have a better chance if you post on Wednesday. There is a lot of retweeting and favoriting activity on this day.

- Majority of the posts were made through Twitter for iPhone, meaning the majority of the users were iPhone owners. This also brings an interesting question on why other mobile platforms are not visible on this list of sources.

- Looking at the correlations between retweet and favoriting, we can see there is a very high chance of a post favorited being retweeted and vice versa. Also to note, rating minimally affects the possibility of retweet or favorite of a post.

- The dog stage with the most favorites and retweets is the puppy.