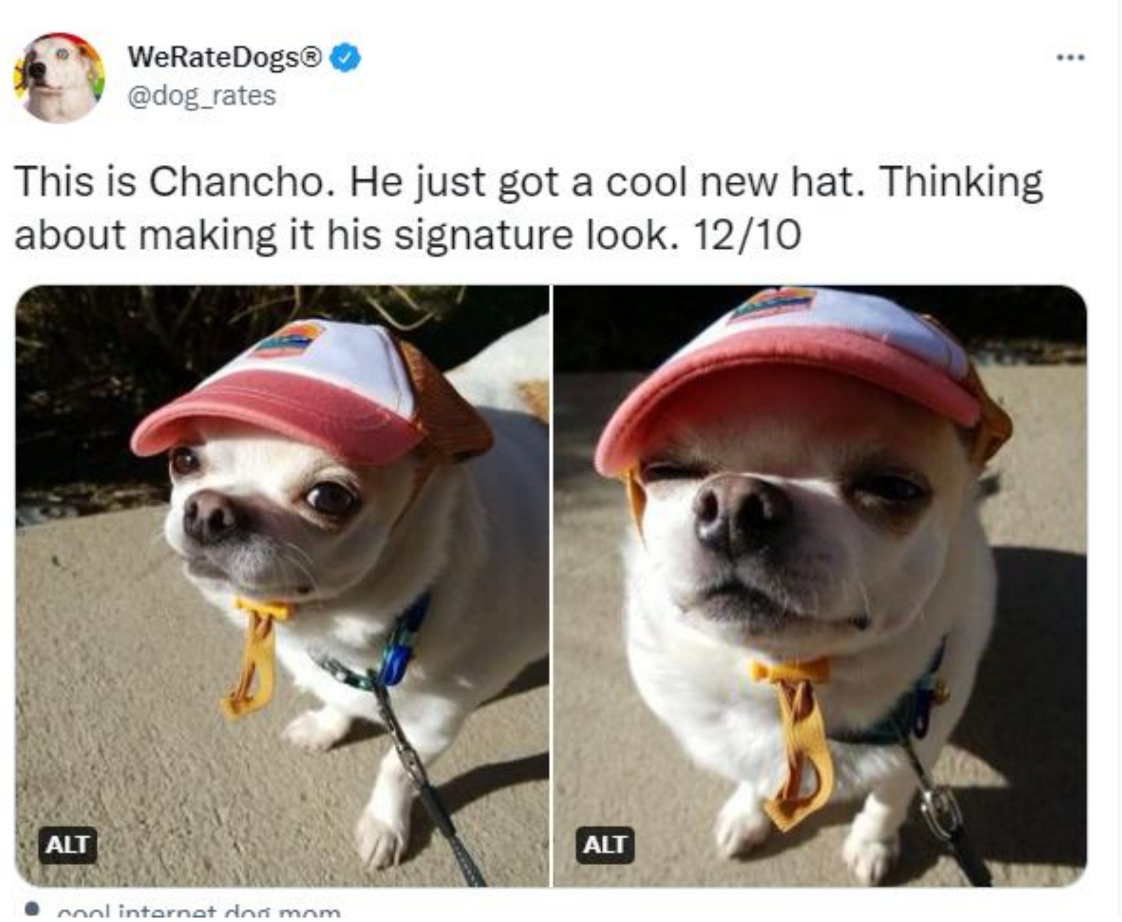
Report: act_report

An analysis of 5000+ tweets of dog images posted by @WeRateDogs user.



WeRateDogs®
@dog_rates

This is Chancho. He just got a cool new hat. Thinking about making it his signature look. 12/10

● cool internet dog mom

Image posted by @dog_rates on twitter

This analysis is the second project in my Udacity data analyst nanodegree program and involves wrangling and analyzing tweet archive of twitter user @dog_rates alias WeRateDogs. WeRateDogs is a Twitter user that post and rate dog images using funny descriptions text about the dog appearances. The ratings almost always have a denominator of 10 and a numerator which is almost always greater than 10. Say 11/10, 12,10, 13/10, 15/10 etc. this special rating system is hugely responsible for the popularity of this Twitter user (see more here: they-are-good-dogs-brent) with a following of over 9.2million. Since real-world data rarely comes clean, our major focus is to demonstrate data wrangling skills/techniques using python and its libraries and then analyze and visualize the clean data.

SCOPE

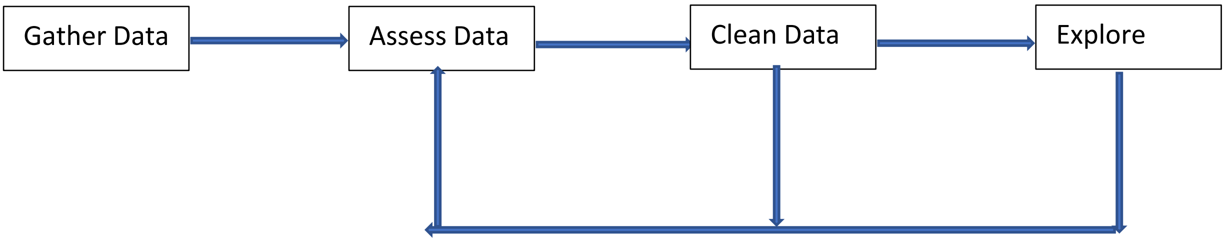
For this analysis, we utilize three datasets from different sources. these includes twitter archive data of @WeRateDogs user. The archive contains 5000+ tweet data and only tweets up until August 1, 2020, was considered for this analysis. The next dataset is an image prediction data containing dog breed predictions using neural network algorithm base on dog images contained in twitter archive data. The third and final dataset is tweet data queried from Twitter API containing additional tweet information like favorite and retweet counts not available in archive data.

We attempt to derive the following insights about dog images posted by WeRateDogs:

- Which dog breed is most popular.
- Which dog breeds are most likeable based on favorite counts.
- Which dog breads has the highest average ratings.
- Is likability (favorite counts) for a particular dog breed influenced by the number of retweets by followers?.

PROJECT STRATEGY/WORKFLOW

The project workflow includes the following phases. 1) data acquisition (both manual and programmatic). 2) assessing data to identify quality or tidiness issues. 3) cleaning task executed to improve the integrity and reliability of the datasets. 4) exploration and visualizations.



Project flowchart developed by Gabriel.

The python libraries used for this analysis include; Request library for downloading the image prediction file programmatically, Tweepy API client for querying Twitter API, Pandas for assessing, cleaning and exploring the data using various available methods, Json package for parsing JSON object responses from APIs, Matplotlib for visualisation.

DATA ACQUISITION

In this phase, we gather data from different sources. Details on data gathered and acquisition methods are as follows:

One of the datasets used for this analysis is the enhanced Twitter Archive provided as file to udacity by the WeRateDogs user. It contains basic tweet data including tweet id, text, rating, dog name and stages. We downloaded this file manually and loaded the data into pandas dataframe using it’s read_csv method.

The second dataset used in this analysis is the Image Predictions File hosted on Udacity servers. it contains predictions for dog breeds based on images in tweet data using a neural network algorithm. This file was downloaded programmatically using the Requests library and read into pandas.

The third and final dataset is tweet data. it contains additional tweet data such as favorite and retweet counts not available in our archive datasets and can be accessed via twitter API. For each tweet IDs in the WeRateDogs Twitter archive, we queried data from Twitter API using Tweepy API client library. each tweet's entire set of JSON data is written to a json-tweet.txt file and loaded into pandas dataframe.

DATA ASSESSMENT & CLEANING.

As with most raw dataset, the three datasets for this analysis were found to be dirty and untidy. Some of the issues identified during visual and programmatic assessing of the data include: extraneous data (archive data contains retweet, reply and non-rating tweets) not relevant to our analysis, incorrect values in name and rating numerator column of twitter archive data, missing values, invalid data types, inconsistent value format, duplicated columns in the datasets, tweet dataset not part of twitter archive data.

To resolve the aforementioned issues. We defined cleaning task (see below) and wrote code to execute them.

- remove records in twitter_archive dataset where record is a retweet, reply or a tweet that is not a dog rating. these records have missing dog name, retweet id or reply id.
- drop records where dog name is "a". few are as result of erroneous extraction due to inconsistent text format, but majority are non-dog rating tweet.
- replace all "None" in the dataset with NAN.
- convert timestamp column from str to datetime.
- replace outlier value of 1776 numerator ratings for Atticus with mean rating numerator.
- replace missing value for favorite and retweet count with mean of the respective column.
- capitalize all values in the breed column in image_prediction dataset.
- Convert data type for breed column from str to category.
- Define a column “stage” to hold dog stage names in twitter_archive using stage names in tweet text. Drop doggo, floofer, pupper, and puppo columns from twitter_archive dataset.
- Extract subset of image prediction that are dog breeds (p*dog = True) for each prediction and confidence level p1_conf, p2_conf, p3_conf to new datasets. Rename columns for all sub-datasets to common names. Then append all three confidence level sub-dataset to form a dataset with all dog name predictions.
  Sort appended datasets on tweet_id and prediction confidence value in descending order. Finally, Remove duplicate tweet_id to keep only predictions with the highest confidence value.

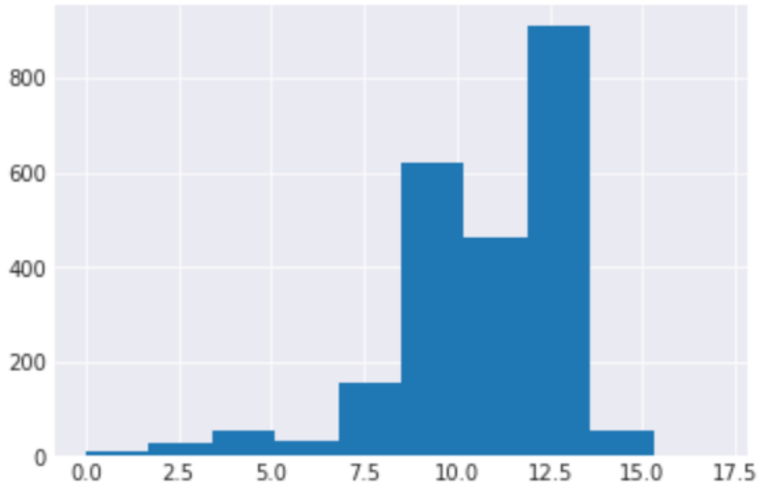
- Combine both twitter_archive and tweet_df into one observational type using tweet_id.

The result of the cleaning task is a high quality and tidy data for a more accurate and efficient analysis.

DATA EXPLORATION

In this section we analyze the cleaned data to gain insights using descriptive statistics and visualizations.

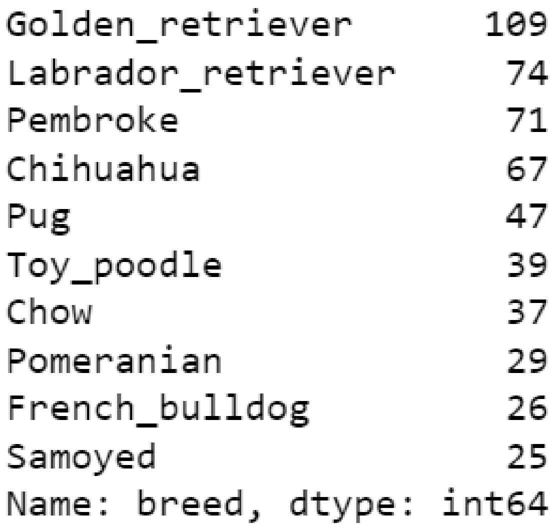
- Distribution of dog rating (numerator):



Distribution of rating numerator for dog rating tweets

From the above histogram, we see that most dogs posted received ratings of between 12 and 13 for numerator rating value.

- Which breed is the most popular?



```
Golden_retriever      109
Labrador_retriever     74
Pembroke               71
Chihuahua              67
Pug                    47
Toy_poodle             39
Chow                   37
Pomeranian             29
French_bulldog         26
Samoyed                25
Name: breed, dtype: int64
```

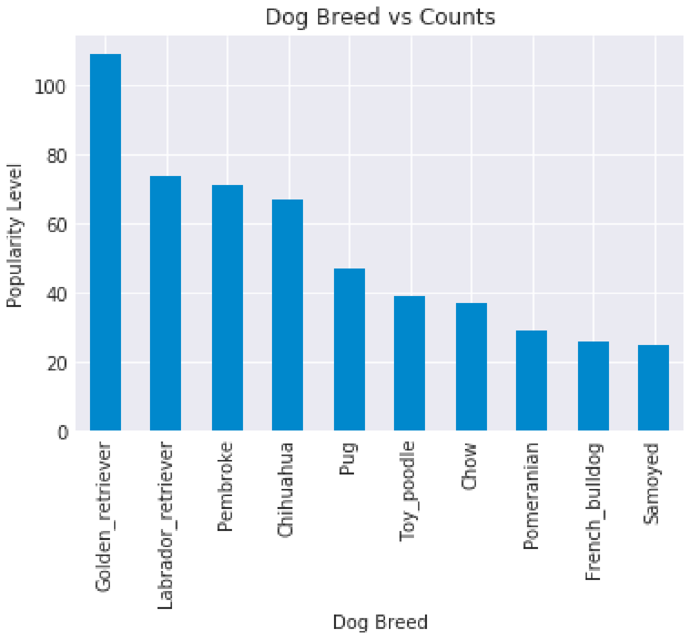


Chart of top 10 most popular dog breads in dog rating tweets

Golden retriever on a search and rescue mission (shot by sono_ilcane baloo/Mercury Press)

From the above chart, the Golden retriever breed is the most popular breed based on count of tweets that contains pictures of this breed. this may be due to their versatile utility in hunting, field work, as guides for the blind, and in search-and-rescue. A total of 109 record in our tweet archive contain the Golden retriever breed data.

- **Which dog breeds are the most likeable**

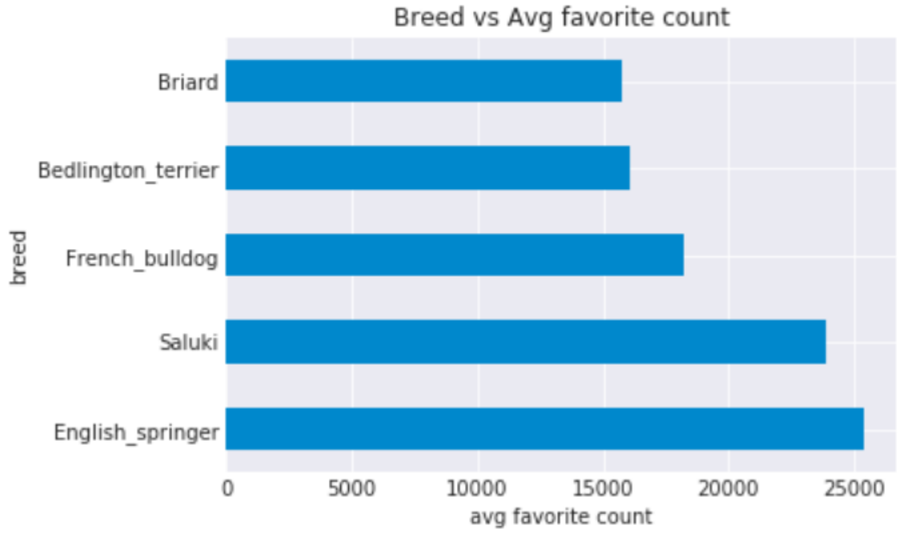
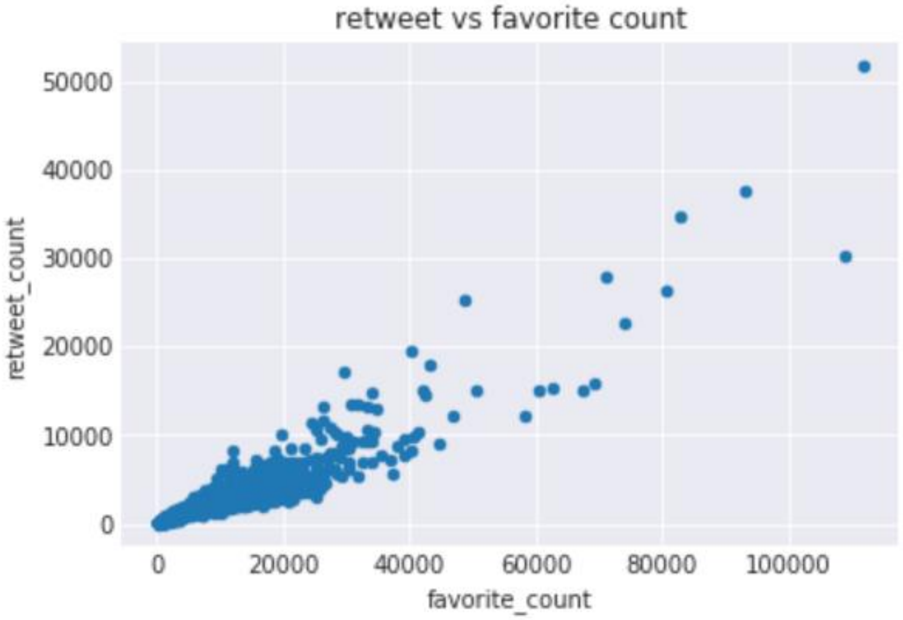


Chart of top 5 dog breads with most favorite count for a dog rating tweets

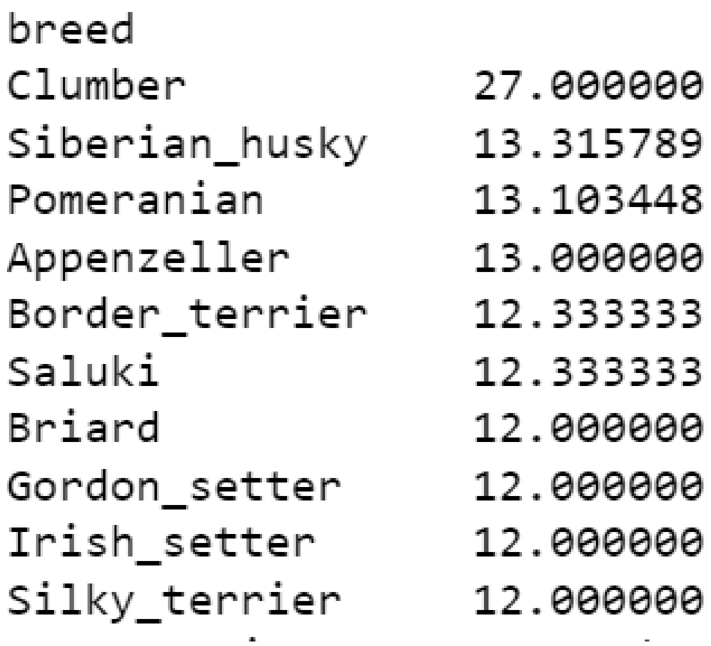
From the horizontal bar chart above, we see that English Springer spaniel dogs are the most adorable breed as seen by the highest average likes per tweet received for each posting of the English springer. one reason for this may be because spaniels are sport dogs that are activity driven. photos of these dogs engaging in sport activities will be an adorable sight. this is closely followed by the Saluki breed. the French bulldog, Bedlington terrier and Briard makes up the top five for breeds with most likes.

- Is the likability (favorite counts) for a particular dog breed influenced by the number of retweets by followers.

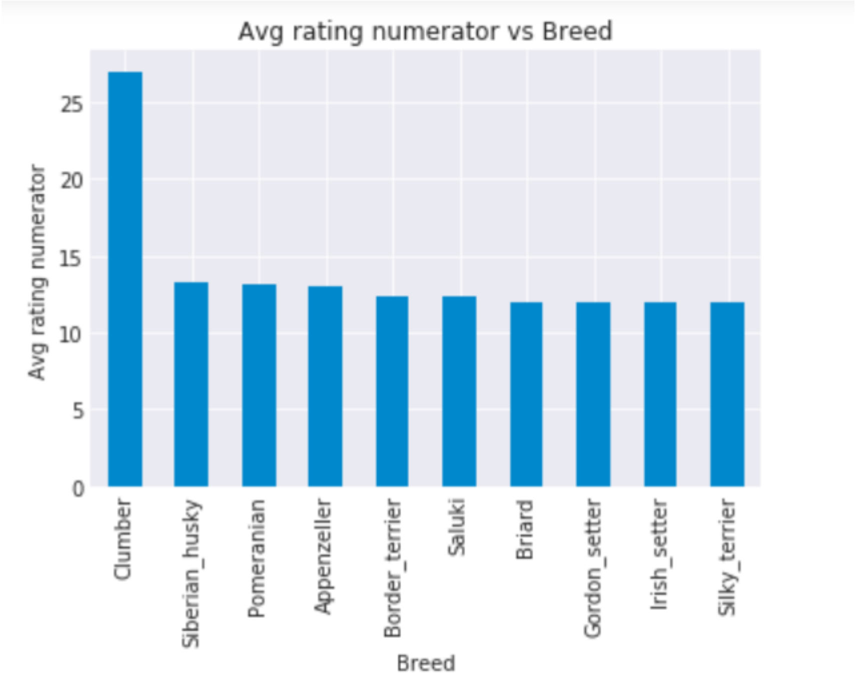


the scatter plot above shows a strong positive correlation between the likability (favorite count) of a dog tweet and the count of retweets it receives from other twitter users. the higher the retweets the higher the favorite count as more users can view and like the tweet due to increased exposure.

- **Which dog breeds have the highest average ratings**



```
breed
Clumber            27.000000
Siberian_husky     13.315789
Pomeranian         13.103448
Appenzeller        13.000000
Border_terrier     12.333333
Saluki             12.333333
Briard             12.000000
Gordon_setter      12.000000
Irish_setter       12.000000
Silky_terrier      12.000000
```



From the bar chart above, The Clumber breed has the highest average rating numerator of 27.0. The Siberian_husky, Pomeranian, Appenzeller, Border_terrier, Saluki, Briard, Gordon_setter, Irish_setter, Silky_terrier makes up the top 10 avg rating numerator with an avg of 12.5 ratings.

SUMMARY:

From the analysis and visualization of the wrangled @WeRateDogs twitter archive data, we derived the following insights.

- Golden retriever is the most popular breed based on the number of tweets that contains pictures of this breed. if you want a dog preferred by most people that you can engage in sport activities then this breed is for you.
- The English Springer spaniel dogs are the most adorable breed as seen by the highest average likes(favorite) per tweet.
- The Clumber breed has the highest average rating numerator of 27.0
- There is a positive correlation between the likability (favorite count) of a dog tweet and the count of retweets it receives from other twitter users.

it is important to note that these insights are tentative as we do not have enough data to draw the above conclusions.

Limitations:

- The dog stage wasn't considered in this analysis as this information is not available for a significant portion of our twitter archive data. considering this variable would not have provided accurate representation of our data.
- Due to the inconsistence tweet text format from which dog names were extracted, a few of records in the twitter archive had incorrect dog name and were dropped from our dataset since they were insignificant.

Thanks for taken out time to read this report. Hope you enjoyed reading it as much as I did writing it.