Data Wrangling Project from WeRateDogs Twitter account

Data Wrangling Project contents

Data wrangling, which consists of:

1. Gathering data (downloadable file and using tweetpy to collect the necessary data).
2. Assessing data
3. Cleaning data
4. Analyzing and visualizing

1. Gathering data

Data have been collected from three different tables

1. The WeRateDogs Twitter archive
2. The tweet image predictions, i.e., what breed of dog (or another object, animal, etc.) is present in each tweet according to a neural network
3. The tweet table, each tweet's retweet count and favorite ("like") count at hase been collected using tweetpy library.

2. Assessing data

Data has been assessed visually and programmatically for quality and tidiness issues.

3. Cleaning data

3.1 Quality

Data quality dimensions are completeness, validity, accuracy, consistency

1. twitter_archive_enhanced table

- Missing values (NaN) in expanded_urls
- Null represintade as (None) in name column
- in name [a, an, the]
- drop null raws
- outliers in rating_numerator and rating_denominator columns

2. image_predictions table

- rename columns name [p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3_conf, p3_dog]
- Only have 2075 tweet_id

3. de_tweet table

- I only have get 1777 tweets

4. de_final table

- Drop null rows

3.2 Tidiness

It is stractural issues

1. Each variable forms a column.

2. Each observation forms a raw.

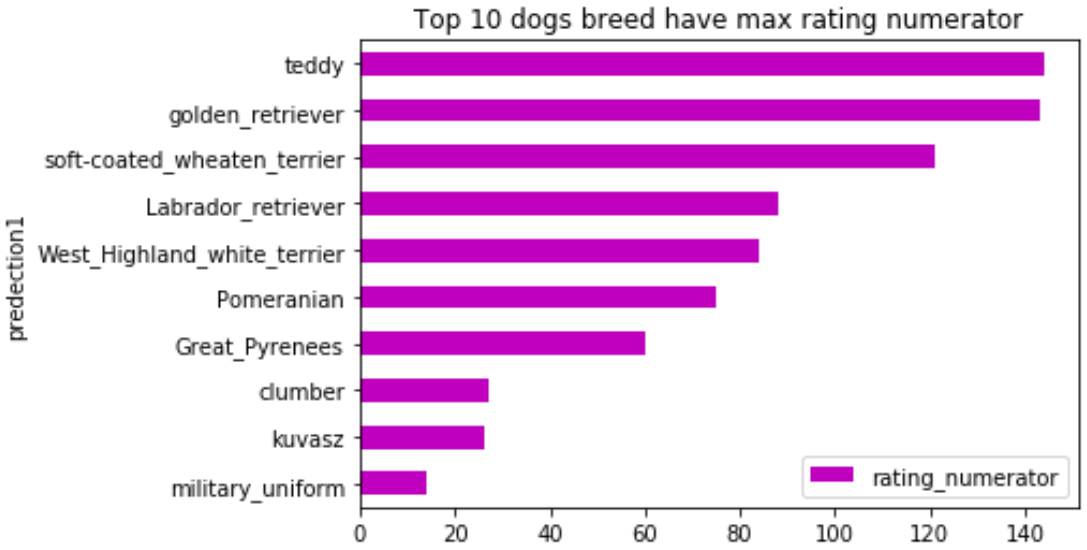
3. Each observation forms a table.

- Drop [doggo, floofer, pupper, puppo] columns
- Drop [in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp] columns
- Split date and time in timestamp column
- All data in one table (df_final), the final table should include [tweet_id, favorate_count, retweet_cont, dog_breed, time, date, rating_numerator, rating_denominator] in the same dataframe

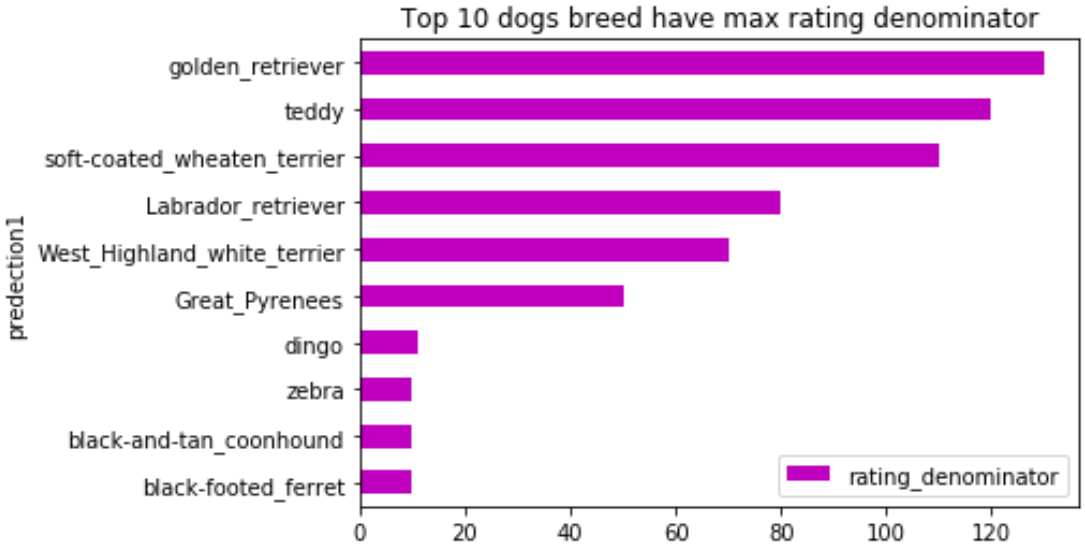
4. Analyzing and visualizing

The analysis shows

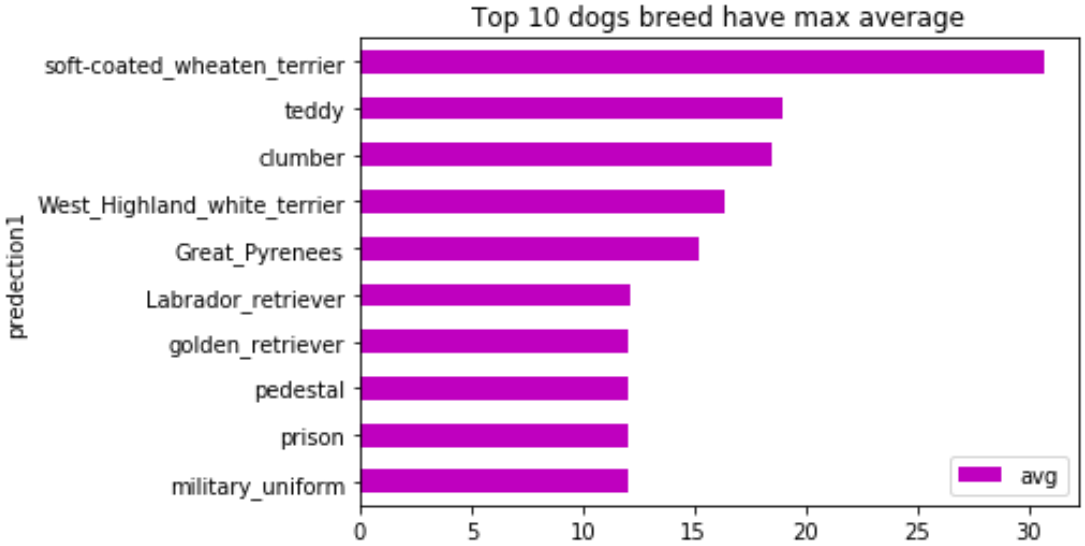
- Which dogs breed have max rating numerator?



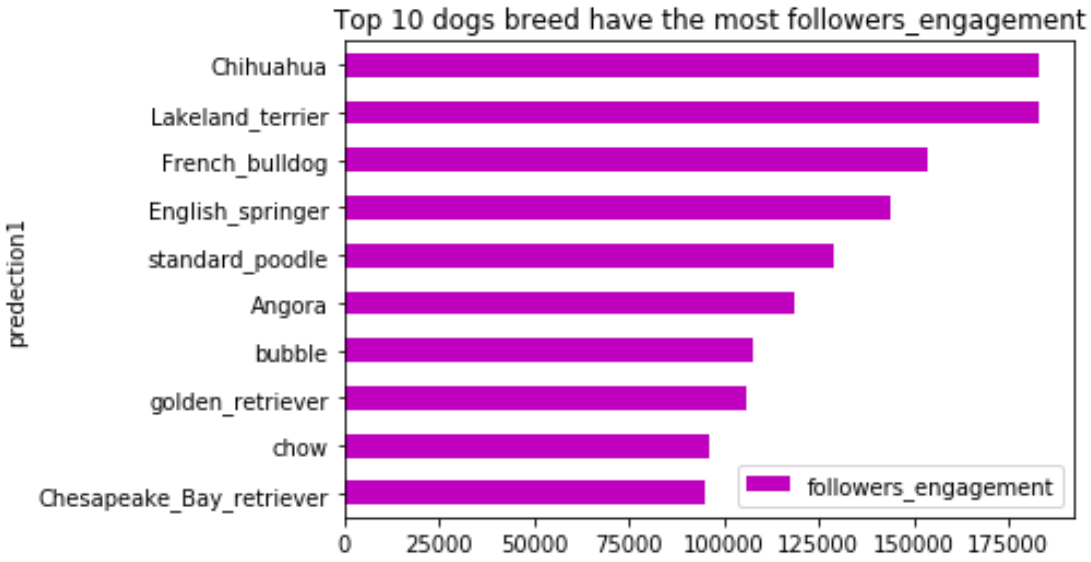
- Which dogs breed have max rating denominator?



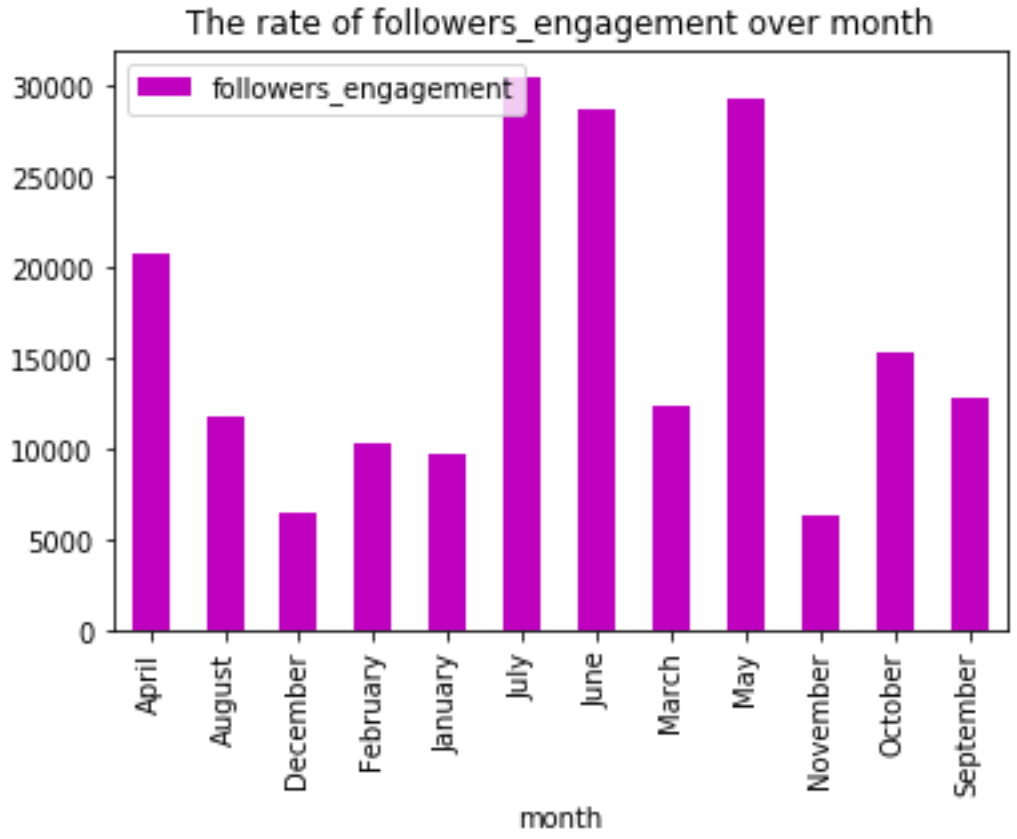
- Which mean dog breeds have highest rating numerator and rating denominator average?



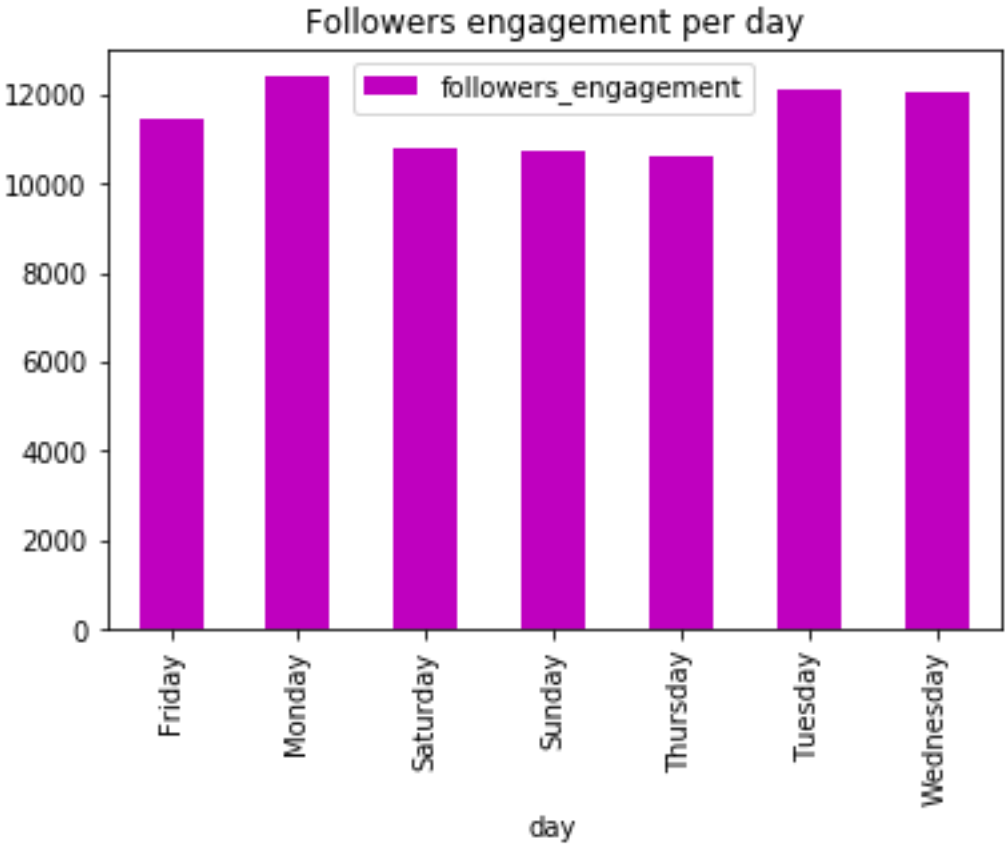
- Which dog breed have the most followers' engagement (favourite count and retweet count)?



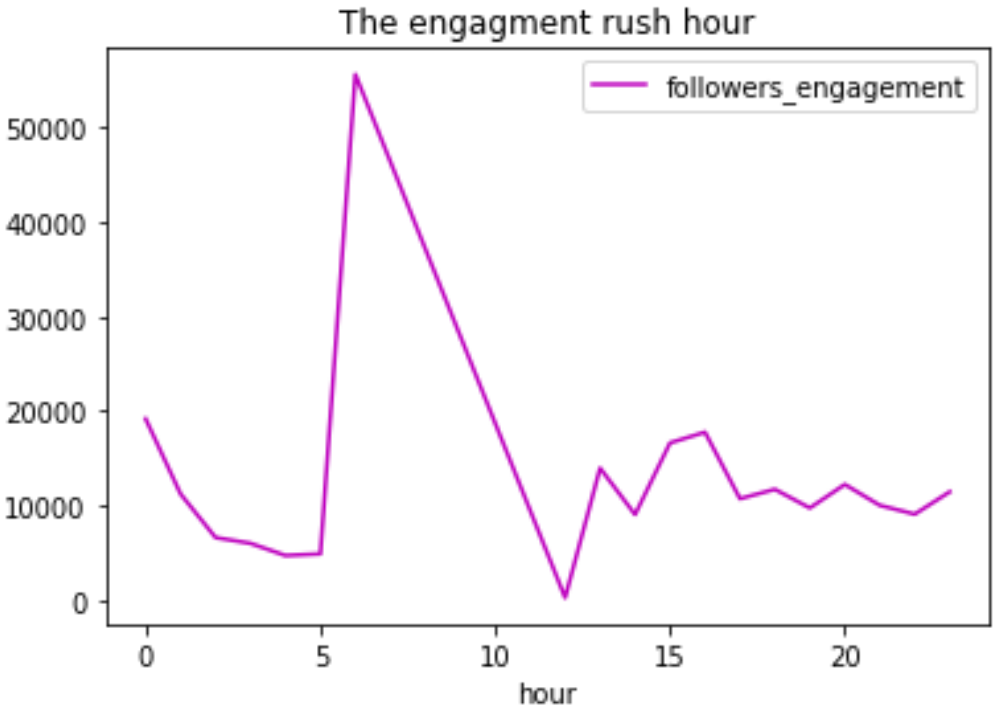
- What is the rate of followers' engagement (favorite_count and retweet_count average) over month?



- Which day has the highest followers' engagement?



- Which hour has the highest followers' engagement (favorite_count and retweet_count average)?



- What is the rate of followers_engagement over year?