Act

_report

September 7, 2022

0.1 Reporting:

0.1.1 Wrangle Report:

The dataset wrangled in the project is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous com ment about the dog.

The WeRateDogs Twitter project goals included: >- Wrangling the twitter data through the following processes: >- Gathering data >- Assessing data >- Cleaning data >- Storing, analyzing, and visualizing your wrangled data >- Reporting on the data wrangling efforts and data analyses and visualizations

0.2 Wrangle Report

> My wrangling efforts for the WeRateDogs Twitter project included gathering data from the following sources: - The WeRateDogs Twitter archive. The twit ter_archive_enhanced.csv file was provided to Udacity students. WeRateDogs down loaded their Twitter archive and sent it to Udacity via email exclusively for you to use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017. - The tweet image pre dictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file was provided to Udacity students. - Twitter API and Python's Tweepy library to gather each tweet's retweet count and favorite ("like") count at minimum, and any additional data I find interesting.

0.2.1 Assessing Report:

I extensively assessed the data and used the primary parameters to determine data quality:

0.2.2 Quality issues

Data quality is determined by the following parameters:

- Completeness: availability of missing data
- Validity : Different columns have right data input
- Accuracy : columns have right data types. No duplicates.
- Consistency

1. dog names: some dogs have 'None' as a name, or 'a', or 'an.'

2. A lot of missing data in, retweeted_status_timestamp, retweeted_status_id, retweeted_status_user_id

3. timestamp is an object
4. p1, p2, p3 are column duplicates of image dataframe
5. retweeted_status_timestamp is also an object instead of datetime
6. rating_numerator goes up to 1776

7. rating_denominator should be a standard 10, but there are many values above 10
8. the source column still has the HTML tags

0.2.3 Tidiness issues:

On tidy issues, I was able to identify the following:

1. The last 4 variables of twitter_archive are relatable therefore can form one column

2. p1, p2, p3 seems relatable therefore can be combined to form one column

0.2.4 Cleaning Report:

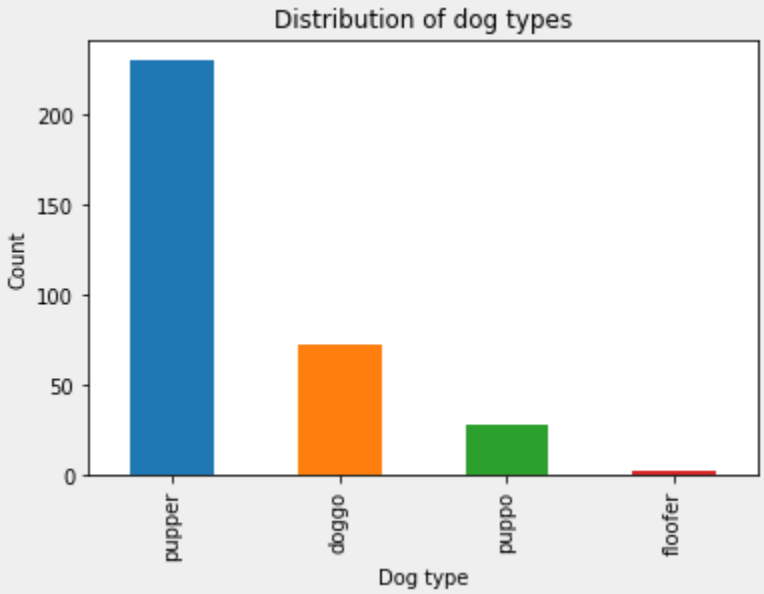
0.2.5 Issues:

I defined the following issues in my data cleaning process:

• Merge the clean versions of twitter_archive, image_df, and tweet_df dataframes. • Create one column for the various dog types: doggo, floofer, pupper, puppo then drop other unnecesary columns
• Converting rating_numerator and rating_denominator into float
• Remove unnecessary columns for easier manipulation
• Change the timestamp to correct datetime format from object
• Creating a rating column
• creating month column
• remove missing values

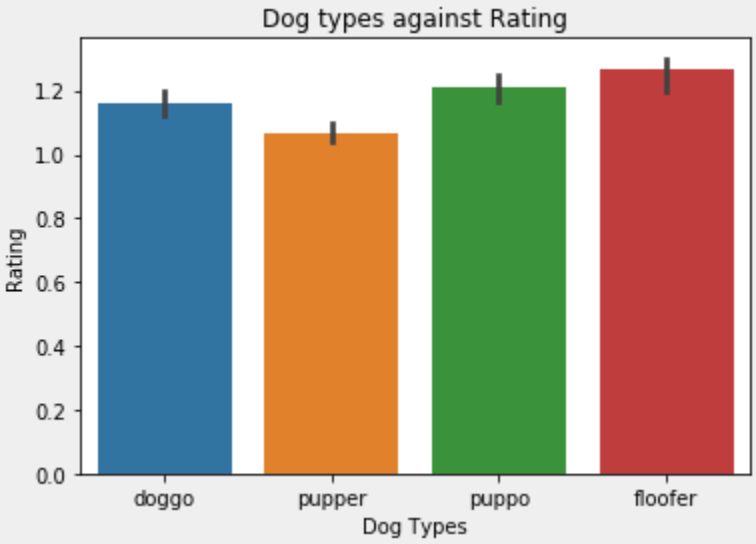
Visualization and Analysis:

1. Determining the most popular dog type



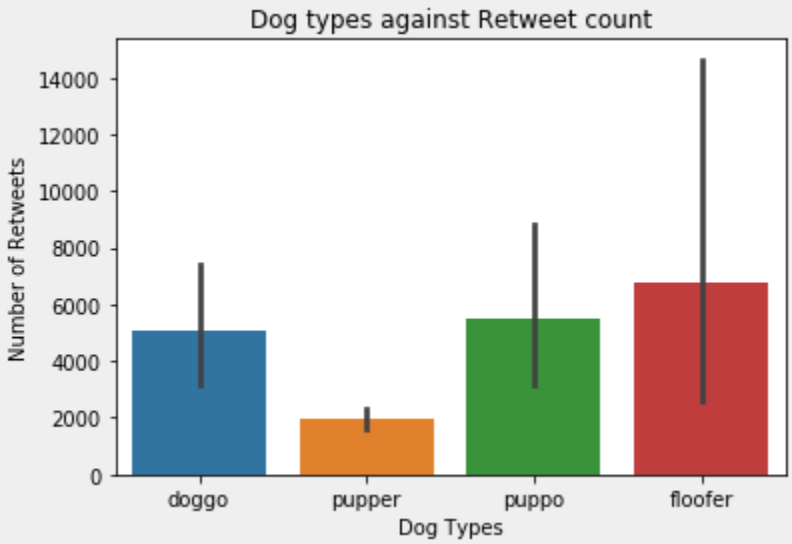
Observation: Pupper is the most popular dog type followed by doggo, puppo then the least popular dog is floofer.

2. Determining the dog type that has the highest rating.



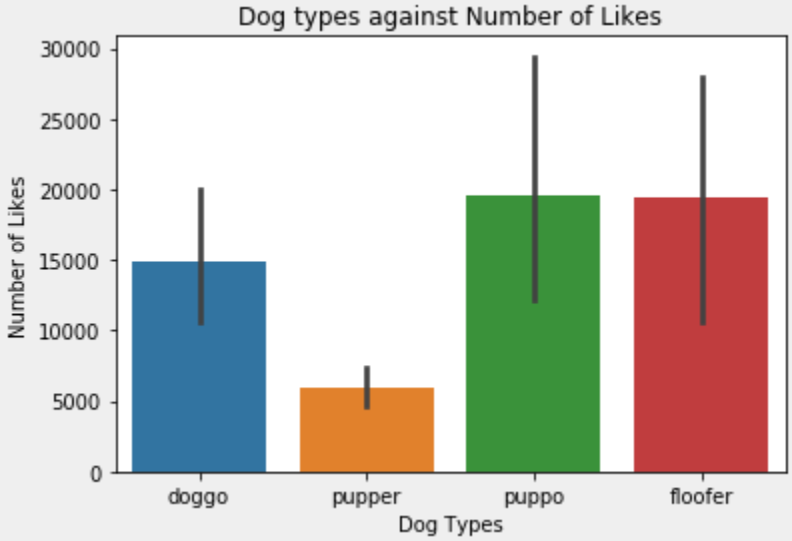
Observation: Floffer is the best rated dog type followed by the puppo, doggo while pupper is the least rated dog

3. Determining the dog breed with the highest number of retweets.



Observation: Floofer has the highest retweets, followed by Puppo, doggo while pupper has the least retweet

4. Determining dog breed that recorded the highest number of likes/favorites.



Observation: Floofer and puppo have almost the same amount of likes, and also the highest. Pupper has the least number of likes

0.2.6 Conclusion:

I did some visualizations to help me derive key insights from the data. The followingnare the key insights I came up with after visualization and analysis:

1. Floffer has the highest rating while pupper has the least rating

2. floofer and puppo have almost similar amount of likes while pupper has the least 3. floofer has the highest number of retweets while pupper has the least

4. pupper is the most popular dog breed while floofer is the least one