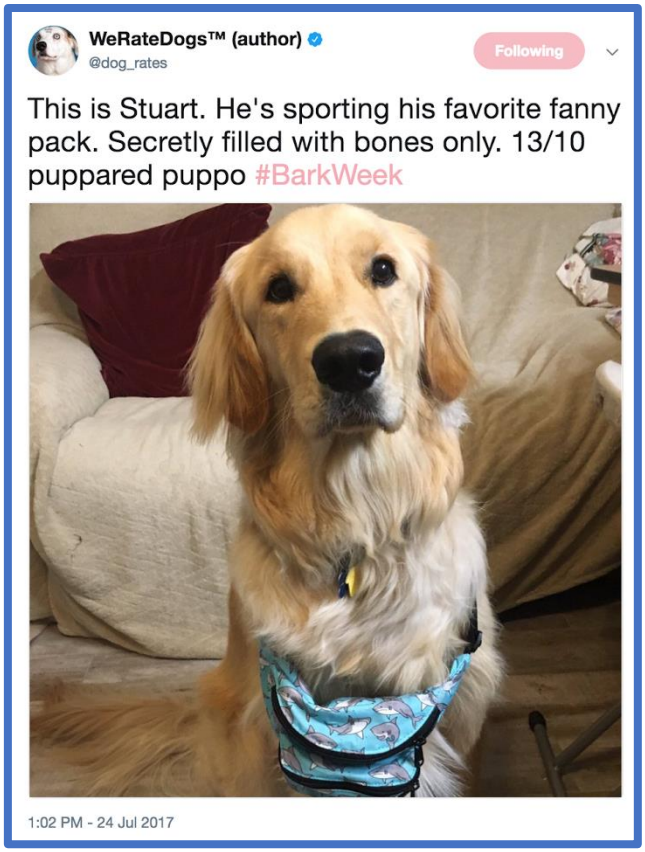
WeRateDogs Twitter Account Dataset Analysis



Data Wrangling

Gathering Data:

Data in this project was gathered from 3 sources with different file formats:

1- Enhanced Twitter Archive.
2- Image Predictions.
3- Additional Data via the Twitter API.

- Enhance Twitter Archive file `twitter_archive_enhanced.csv` data was extracted from Twitter API and given to me then I imported it into a DataFrame called "twitter_archive".

- Image Predictions File `image-predictions.tsv` a file was given containing predictions of tweet photos, whether it contained a dog or not and what is its breed and how confident the model is about its prediction. It's imported in another DataFrame called "image_predictions".

- Additional Data of retweet counts and favorite counts were extracted via Twitter API and loaded into a DataFrame called "twitter_queried".

Assessing:

In this part I started assessing data visually and programmatically with code.

After assessing data these were the issues encountered separated into 2 sections

(Quality – Tidiness).

Quality:

`twitter_archive` table:

1. ID fields: like tweet_id should be objects, not integers or floats because they are not numeric and aren't intended to perform calculations.
   implemented after cleaning tidiness issues

- wrong values in `rating_numerator` and `rating_numerator`columns .
  1. tweet_id `810984652412424192` has no rating but contains 24/7 in text which means the whole week.
  - tweet_id `775096608509886464` wrong rating, its rating value is 14/10 -visually inspected text and urls- .
  - tweet_id `682962037429899265` wrong rating, its rating value is 10/10 -visually inspected text and urls- .
  - tweet_id `666287406224695296` wrong rating, its rating value is 9/10 -visually inspected text and urls- .
  - tweet_id `786709082849828864` wrong rating of 9.75/10 as a joke on harry potter, it is found from comments that it is 13/10 instead.
  - tweet_id `670842764863651840` a man named dogg with 420/10 and not a dog.
  - tweet_id `778027034220126208` wrong rating by 27/10 but it is actually 11.27/10
  - some tweets have aggregated ratings for multible dogs.
  - tweet_id `749981277374128128` has a rating of 1776/10, not an error but it is an outlier in analysis phase.
  - there is no actual rating in any other tweet -except for those mentioned in previous points-
    that has a value of `rating_denominator` remainder by 10 != 0.

- tweets that are replies are not counted in this analysis
  - then no need for `in_reply_to_status_id`, `in_reply_to_user_id` columns.
- tweets with no `expanded_urls` and has no data in `image_predictions` table
- No need for are retweeted tweets
  - retweet columns should be removed
- `name` column string 'none' not special value `NaN`
- extra +0000 in `timestamp`
- `timestamp` is object datatype instead of timestamp datatype
- `source` column datatype is object instead of being categorial

- `source` is object instead of being categorial
- `none` string in `doggo`, `floofer`, `pupper`and `puppo` colums instead of Nan.
- dog stages has to be categorial data type after merging them
- some tweets have more than one dog stage.

- redirection link in `text` needs to be removed to have the plain text of the tweet.
    - If I'm to make a word-cloud i will clean it, otherwis it is not necessarily an issue. not implemented in this notebook submission

`twitter_queried` table:

- retweet_count and favorite_count should be integers, not floats. implemented after cleaning tidiness issues

Tidiness:

`twitter_archive` table

1. two columns for rating in table `twitter_archive`, should be in one column instead
2. dog staged should be in one column

`image_predictions` table

3. Multible columns define predictions.
   not implemented in this notebook submission

4. join all tables with `tweet_id` as primary key.

Cleaning:

Cleaning process was then conducted by Code in a Jupyter notebook with assessment and EDA.

https://github.com/IbrahimMansey/WeRateDogs_Dataset