WeRateDogs Data Wrangling Report

The WeRateDogs twitter data wrangling project consisted of gathering data from three different sources, assessing, merging the data into one data frame, cleaning, and finally analyzing the data.

Gathering

Sources -

- Downloaded csv twitter-archive-enhanced.csv as provided by udacity.
- Downloaded image-predictions.tsv from internet using requests
- Got JSON object of all the tweet_ids using Tweepy in tweet_json.txt

Assessing

I visually and programmatically assessed all three data frames, documenting all tidiness and quality issues.

Quality

Completeness, Validity, Accuracy, Consistency , i.e. content issues

twitter_archive dataset

- We don't want retweets.
- Erroneous datatypes (timestamp, tweet_id, in_reply_to_status_id, in_reply_to_user_id).
- The numerator and denominator columns have invalid values.
- Extra characters after '&'
- In several columns null objects are non-null (None to NaN).
- Name column have invalid names i.e 'None', 'a', 'an' etc.
- Should change columns type (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id and tweet_id) to string because we'll not be using them for any computation.
- Remove colums like 'retweeted_status_id' , 'retweeted_status_user_id and 'retweeted_status_timestamp' because we don't need them . Also, column date_time we imported from the API, it has the same values as timestamp column , so we don't need that either .
- Sources difficult to read.

image_predictions dataset

- Tweets with no images (2075 rows instead of 2356)
- Some tweet_ids have the same jpg_url
- Some tweets are have 2 different tweet_id one redirect to the other (Dataset contains retweets)

tweet_data dataset

- Remove column user_favourites

Tidiness

Untidy data , i.e. structural issues

- Dog "stage" variable in four columns: doggo, floofer, pupper, puppo
- Join 'tweet_data' and 'images' to ‘tweets'

Merging/Cleaning

I cleaned all documented tidiness issues followed by quality issues by defining, coding, and testing each issue.

Once the cleaning process was completed, I stored the final data frame and began analysis.