Wrangle Report

Gather

twitter-archive-enhanced.csv: File downloaded manually from Udacity's server, uploaded into Jupiter notebook and read using Pandas library -def: WeRateDogs Twitter archive

image-predictions.tsv: File downloaded programmatically using requests module from Udacity's server and read using Pandas library - def: tweet image predictions, i.e., what breed of dog (or other object, animal, etc.)

tweet_json.txt: File provided by Udacity gathered using twitter API **as Security of information and Privacy was a concern for me applying for the twitter API,** read using json module -def: file contains (followers count - retweets count - favorites count) for each tweet

Assess

- Started with an overall visual assessment using **Numbers**, then focused on programmatic assessment in a **Jupyter notebook** using **Pandas** library and its functions.

Issues found were divided into **Quality and Tidiness**

Quality issues

`twitter_archive_df`

Retweets, replies, quotes and self-status tweets (Original tweet with a photo):

- Drop replies, retweets by checking their respective columns' values (non-null)
- Check with image prediction df for tweets that have an image (intersection between arch df and img df)
- **timestamp** is an object: convert to datetime.

. Irrelevant columns to be dropped
retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, doggo, floofer, pupper,puppo, in_reply_to_status_id, in_reply_to_user_id, source, expanded_urls

expanded_urls

- Duplicated values: solved by dropping retweets, replies and check with image prediction file

- Null values: same solution as above

- Invalid urls: correct urls are repeated more than once could be solved using a regex pattern (https til https) yet column won't be of use

name

- None values: convert to null

- a/an names: extract names from text using regex pattern of (name/ named) + "name", other than that set to null

- lowercase names other than a/an: found that all names starting with lowercase letters are invalid, solved by setting to null

rating_numerator and rating_denominator

- Float value ratings extracted and set incorrectly: detect using a regex pattern, extract correct rating from text and manual fix

- Incorrect values extracted given multiple ratios in text (usually first ratio is taken) : detect, extract correct rating from text and manual fix

- Collective rating for multiple dogs in a photo (e.g.. 420/400 for 40 dogs): rating factor numerator/denominator

- Multiple ratings for multiple dogs in one tweet (e.g.. tweet_id = 676191832485810177): slight difference wouldn't cause an issue

image_prediction_df

- Retweets, replies and quotes (Original tweet with a photo): check with arch_df
- **Duplicated** values in **jpg_url**: check with arch_df
  img_num: many (3) values represent 4 photos (irrelevant column)
- Naming **Breeds** in p1, p2 and p3 is inconsistent. Some breeds start with uppercase letters, others with lower case letters: convert all breed names to lowercase letters using str.lower() and replace (_) underscore with space for proper display
- Columns names are not expressive of their values: **renaming** correctly

  p( ) to prediction( )

  p( )_conf to confidence_( )

  p( )_dog to dog_( )

count_df

Retweets, replies, quotes etc.. : inner merge with twitter archive (after drop and check with image prediction)

Tidiness issues

- **doggo, floofer, pupper and puppo** columns in twitter archive df to be reduced into one column **dog_stage**
- Each dataframe should be an independent observation units: create a master dataframe mainly of twitter archive and joining breed, favorite and retweets

Clean

followers_count column in count_df is nearly a constant and would be irrelevant, not considered in master_df

Breed column

Created in img_clean_df and master_df for dog breed using concept of programmatic overwriting concluded from image prediction df on basis:

Second prediction confidence greater than 0.1 and **Third prediction confidence greater than 0.03**.

Creating a **master data frame** structured mainly of twitter archive df and merging with counts df and dog_type column for dog breeds obtained from image prediction df