Wrangling Report

Gathering Data :

We had three data sets to gather data from

1) **twitter_archive_enhanced.csv**

   which was in a csv format , I downloaded it and read it using pandas read function (read_csv) and saved it as archive_df .

2) **Image-predictions.tsv**

   (The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network)

   which was in a tsv format , It is hosted on Udacity's servers . I downloaded it programmatically using the Requests ibrary and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

   Then read it using pandas read function (read_csv) and saved it as Image_predictions_df .

3) **Tweet_json**

   (It contains Each tweet's retweet count and favorite ("like") count at minimum)

   Which was in zip format , I extracted the json file from it and read it using pandas .

   Then I read this json file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count and saved the results as api_df

Accessing and Cleaning data

These are the quality and tidness problems found in the data sets and how I managed to clean them .

Quality issues

- **consistency issue :** unnessary data (retweet_status_id , retweet_status_user_id ,in_reply_to_status_id , in_reply_to_user_id , retweeted_status_timestamp)
- **Validity issue :** Some tweets are actually retweets and replies not original tweets

    - ✓ I deleted tweets which is originally a retweet or a reply
    - ✓ Then I drop (retweet_status_id , retweeted_status_user_id ,in_reply_to_status_id , in_reply_to_user_id , retweeted_status_timestam)

- **consistency issue :** timestamp is a string , not a date time object

    - ✓ I changed of the type of timestamp column to datetime

- **completeness issue :** Alot of the dogs are not classified

    - ☒ (couldn't fix)

- **completeness issue :** missing urls

    - ☒ (couldn't fix)

- **completeness issue :** discrepancy in the number of tweets between the archive_df dataset and the image_prediction_df.

    - ✓ I deleted tweets which has no images
    - ✓ Then matched the number of tweets in both tables and checked that they had the same tweet ids

- **Accuracy issue :** names are not extracted correctly from Name Column

    - ✓ I dropped "None" and "a" values
    - ☒ Couldn't fix rest of the names as it needs to be done manually and would take a very long time

- **Validity issue :** null values are written as string (None)

    - ✓ I replaced "None" values with np.nan

- **Accuracy issue :** Some rating numerators don't follow The unique rating system is of WeRateDogs.

    - ☒ (couldn't fix)

tidiness issues

- (doggo, floofer, pupper, puppo) these Column headers are values, not variable names , columns should be united under a classification column

    - ✓ I combined these columns under 1 column called dog_breeds

- Multiple types of observational units are stored in the same table (Image_predictions_df) ,jpg_url and img_num should be separated

    - ✓ I separated jpg_url and img_num to another table (images_df)