Data Wrangling Report

Observations

- There are 77 reply and 163 retweet and 26 quoted status which are unnecessary

- geo, id_str, contributors,coordinates are unnecessary

- some entries here aren't related to dogs at all

- some columns are repeated

- There are a lot of None and NaN values throughout the dataset

- some column names are not clear enough and hard to understand.

Cleaning Data

(1) Tidiness issues:

- delete any column that has "retweeted" and "reply" in archive_df

- delete quoted_status , quoted_status_id, quoted_status_id_str ,quoted_status_permalink, in_reply_to_screen_name , in_reply_to_status_id ,in_reply_to_status_id_str ,in_reply_to_user_id ,in_reply_to_user_id_str, contributors ,coordinates, id_str, geo in api_df

- change column name of : p1, p1_conf, p1_dog

(2) Quality issues:

in archive_df:

- convert the data type of timestamp from string to datatime & tweet_id to string
- Remove the duplicated expanded_urls
- Remove NaNs from expanded_urls
- Remove rating_num that are more than 15 and less than 10
- Remove rating_denominator that are not equal 10
- Remove the rows that are not related to dogs

in image_pred_df:

- delete NaN rows in jpg_url
- combine puppo, pupper, floofer, doggo into 1 column called dogs_stages

in api_df:

- covert id column type into string

Overall

- combining all the dataframes together
- Delete unnecessary columnns