Wrangle Report

The WeRateDogs Twitter Archive Data

In this project we wrangle the WeRateDogs Twitter Archive Data several times to get clean and more powerful for further analysis and visualization.
Starting by Gathering data for WeRateDogs Twitter using Enhanced Twitter Archive data as basic data and then get additional data from Twitter using Twitter API to include retweet count and favorite count for each tweet in the basic data which are “Enhanced Twitter Archive data”. Lastly we include the Image Predictions data set which for images prediction in each tweet.

For the assessing and cleaning (done programmatically as in wrangle_act.ipynb file) part of the dataset I detect 8 quality issues and 2 tidiness issues in this stage of wrangling, which are as below:

Tidiness issues:

- Combine/Merge the Enhanced Twitter Archive data "df_twitter_archive_enhanced" with Image Predictions File "df_image_predictions" through tweet_id column
- Concat retweet_count and favoirte_count with "df_twitter_archive_enhanced" through tweed_id

Quality issues:

- We want original ratings (no retweets) that have images (because not all are dog ratings and some are retweets), so we will drop all columns that not related on rating from "df_twitter_archive_enhanced" table, which are:
  - in_reply_to_status_id
  - in_reply_to_user_id
  - retweeted_status_id
  - retweeted_status_user_id
  - retweeted_status_timestamp

Also we will check the images for each tweet to make sure all rating have images, from "df_twitter_archive_enhanced" data "expanded_urls" column have missing values, by checking the "df_image_predictions" table we didn't have these pictures of same URl, so we will delete the row with missing "expanded_urls"

- drop rows of "expanded_urls" with null

- Change data type of 'timestamp' column to to_datetime

Check the following correction of text extraction in "df_twitter_archive_enhanced" data, and fix it if isn't not correct:

- The ratings probably aren't all correct "rating_numerator" column:
    1. tweet_id of 820690176645140481, 758467244762497024, 731156023742988288, 713900603437621249, 710658690886586372, 709198395643068416, 704054845121142784, 697463031882764288, 684225744407494656, 684222868335505415, 677716515794329600 and 675853064436391936, are not base of 10 so we will divide to be base on 10

    2. tweet_id of 810984652412424192 didn't rate, he/she means 24 hours 7 days, so we will drop this row

    3. tweet_id of 775096608509886464, 722974582966214656, 716439118184652801, 682962037429899265 and 666287406224695296 are wrongly assign the rating. so we get the rating manually through text then assign the write values

- probably dog stages "doggo", "floofer", "pupper" and "puppo" column:
    4. add missing doggo to doggo column
    5. fix dog stages columns (make it dummies) where None = 0 and other value to 1