Wrangle and Analyze Data (WeRateDogs)

Introduction:

In this project, I wrangled **WeRateDogs** Twitter data to create interesting and trustworthy analyses and visualizations. Since the Twitter archive only contains very basic tweet information, I additionaly gathered data using Tweeter API and combined with the WeRateDogs Twitter data. The combined data was assessed and cleaned to get insightful analyses and visualizations. Followings were list of data I have used.

WeRateDog Twitter Archive

The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets, but not everything. One column the archive does contain though: each tweet's text, which I used to extract rating, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and puppo) to make this Twitter archive "enhanced." Of the 5000+ tweets, I have filtered for tweets with ratings only (there are 2356).

Additional Data via the Twitter API

Retweet count and favorite count are very important information but these values are omitted. So I gathered these information through Twitter's API for all 5000+ tweet IDs within the enhanced tweetter archive file.

Twitter Image Predictions File

This file contains the dog breed classification results from a Nuerual Network model for every images in the WeRateDogs Twitter archive. This file has a table full of image predictions (the top three only) alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction (numbered 1 to 4 since tweets can have up to four images)

2. Procedure

The tasks of this wrangling process are as follows: Gathering, Assessing and Cleaning.

2.1 Gathering:

WeRateDog Twitter Archive *(twitter_archive_enhanced.csv),* Twitter Image Predictions File*(image_predictions.tsv)* were given by Udacity. However, additional data (missing data) was obtained by querying Twitter API for each Tweet IDs in twitter archive file. The queried dataset was saved in JSON data file called tweet_json.txt file. From this json file data for each tweet id was read and tweet id,

retweet_count, favorite_count, retweet_count, retweeted, followers_count, friend_count were saved in pandas dataframe. This dataframe is saved in *api_data.csv* file.

2.2 Assessing:

To assess the data I have done visual and programmatical assessment:

- Visual Asessement: I used both Excel and Jupyter notebook printing statements to check files.
- Programmatical Assessment: I used pandas functions (describe, info, value_counts, duplicated, groupby, query etc)

From this assessments I have listed out issues both in quality and tidiness. These are the findings from this assessments.

Quality Issues:

Twitter Archive data:

- Retweets should not be used for analysis.
- Twitter ID type should be same for each datasets
- Many columns that may not be used for analysis since a lot of data missing or duplicated in this columns.
- doggo, floofer,pupper,puppo columns are not True/False values. Actual doggo, floofer, pupper, puppo stage names are exist.
- Multiple labeling for dog stages since there are multiple dogs in a picture.
- Weird ratings observed in ratings_numerator & ratings_denominator.
    - No clues for actual ratings (666/10, 182/10, 1776/10, All time 24/7, Date 11/15/15, 20/10, snoop dog 420/10, 4/20(tweet id: 686035780142297088))
    - Only part of decimal numbers were extracted for numerator (11.27/10, 9.75/10, 11.26/10)
    - Ratings for Multiple dogs in a image get aggregated ratings (44/40,50/50, 165/150, 84/70,88/80, 144/120,143/130,45/50,99/90, 121/110, 204/170)
    - Extracted duplicated OO/OO format in text column (Current value --> Updated value) (Event 9/11--> 14/10,Size3 1/2 legged --> 9/10, 50/50 --> 11/10, 17/10 --> 13/10, 960/00 -->13/10, 4/20 --> 13/10)

Image Prediction data

- Duplicated image predictions (66 duplicates)
- Include confidence for multiple objects captured in a picture

Twitter API extracted data

- Friend count is not real data (Twitter limitation), needs to be dropped

Tidiness Issues:

- Month, Day, Year columns can be seperted from timestamp
- The dataframes need to be merged to get Dog labeling information from image prediction and retweet_count & favorite count from API extracted data.

2.3 Cleansing

After Assessing the dataset I have copied original dataset so that I can iterate my data cleansing process without re-extracting data from the source. Once I have copied, I start by dropping columns and retweeted rows so I don't have to handle unnecessary long list of columns or rows.

Then I have merged doggo, pupper, floofer, puppo columns into one column called stage allowing multiple category values.

I have put a lot of efforts cleansing rating values both numerator and denominator since these values would be used for main analysis. Original values were recaptured from text column which contains both comments and ratings about dogs.

Once I finished Twitter archive data, I have cleansed Image prediction data. First I have dropped duplicated rows. Then I tried to capture real prediction of the type dog since original data had three predictions and confidence levels. Using if functions I filtered out wrong predictions and confidence levels and make these columns into one prediction and one confidence level columns.

Finally, I have cleansed API Data simply dropping unused columns. Friend counts & retweeted (status) contained the same value for every row which later turned out to be twitter api permission limitations.

To make analysis easier, I tried merged all three data into one dataframe joining in with tweet_id. Tweet id types were all different in three dataframe so I first changed type to string before the joining these three datafraeme into one.

Before proceeding to visual analysis, I have saved the cleaned copy of data in csv file so that I can restart from this cleaned dataset anytime I want.