wrangle_report

January 5, 2023

0.1 Introduction :

In this report we document our wrangling efforts in the project. In this project we work with three datasets:

- **Enhanced Twitter Archive:** where WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets
- **Additional Data via the Twitter API :** where these data contain retweet count and favorite count for dogs
- **Image Predictions :** where these data contain dogs image and the prediction of neural network (where this neural classify breeds of dogs)

0.2 Gathering data

In this project we work with three data as we mentiond in the introduction so we need to gathering three different format.

- **Enhanced Twitter Archive:** these data is avaliable to us on udacity in csv format
- **Additional Data via the Twitter API :** As for these data we need to scraping it from twitter by twitter API but i download these data from udacity and these data is provide in JSON format
- **Image Predictions :** these data also is avaliable to us on udacity in tsv format

0.3 Assessing data

We assessing our data both visually and programmatically

First, we visually assessing the three data using `sample()` method

0.3.1 Quality issues

Enhanced Twitter Archive - Retweet-related columns - Unwanted columns - There are ratings numerators less than 10 - There are rating denominator not equal 10 - There are other expanded urls than Twitter's URL - Missing value for name of dogs - Dog names are invalid (such as a , an and the) - Data type of tweet_id is wrong

Image Predictions - Underscores used in multi-name in columns p1 , p2 and p3 - Invalid data type of tweet_id column

Tweet from Twitter API - Invalid data type of tweet_id column

0.3.2 Tidiness issues

- Dog is seperated into 4 columns (doggo , floofer , pupper , puppo) in twitter archive enhanced
- The data is divided into three tables

0.4 Cleaning data

Our goal in this project is to clean our data to make it suitable for analysis and modeling later. We discovered many issues in this data, but treating all these issuess will require a lot of time and effort, so we will suffice to solve some issues

Data cleaning programmatically contains three process, the first of which is defining the problem and developing a solution to it The second process is writing the code to solve this problem The third process is to test the code to see if the problem is resolved or not

- We remove retweet-related columns
- Fixing data type for tweet_id in each dataframe
- Modify all values that are not equal to 10 in `rating_denominator` and make them equal to 10
- Collecting invalid names of dogs and making them None values
- Remove underscore in multi-name in p1 , p2 and p3 columns of image prediction
- We merge the four columns (doggo , floofer , pupper , puppo) into one column called `dog_stage`
- And we deleted the useless columns that we will not need in the analysis, such as `in_reply_to_status_id` and `in_reply_to_user_id`... etc
- Finally, we combined the three datasets together into one dataset