Wrangle and Analyze WeRateDogs Data

1. Gathering Data

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. I gathered data from three sources first the WeRateDogs Twitter archive csv file is given by Udacity contains basic tweet data for 2356 of their tweets. The archive contain rating, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and puppo) . Second file is image predictions file is hosted on Udacity's servers I downloaded it programmatically.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. The third gathered by using Tweepy python library that query Twitter API for additional data beyond the data included WeRateDogs twitter archive file.

2. Assessing Data

Assessing data is the second step in data wrangling. I inspected the dataset for two things: data quality issues (i.e. content issues) and lack of tidiness (i.e. structural issues). By using many pandas functions for example info, describe, duplicated and unique…etc.

- **Quality Issues (content issues)**

Twitter archive data frame file:

1. timestamp and tweet_id columns types aren’t correct timestamp is string should be converted to date and tweet_id is int64 should be converted to string

2. Source column contains the link of the tweet source; extract only the text (name of the source) from the link.

3. Name column contains faulty names should be replaced with none value.

4. Delete retweeted tweets rows because they are not original tweet

5. rating_denominator column not only contains 10

Image predictions file:

1. p1 and p2 columns are string contain (- and _) rather than white spaces between words. 2. Capitalize the first letter of p1,p2 and p3 columns values. 3. tweet_id column type is int64 should be converted to string

- **Tidiness issues (structural issues)**

1. Remove all columns related to retweeted tweets also remove unwanted columns that will not provide useful information in the analysis of the project 2. The data frame contains separated columns for each dog stage ['doggo', 'floofer', 'pupper', 'puppo'], melt these multiple columns into one column called stage. 3. Merge image_predictions and twitter archive into one data frame.

3. Cleaning Data

Is the third step in data wrangling. in these phase I fixed all quality and tidiness issues I identified in the assess step. Then I merged the data into one data frame.