Wrangle Report

Wrangle and Analyze Data

Jiawei He

1. Project Details

The tasks of this project were as follows:

- Data wrangling, which includes:
  1) Gathering data
  2) Assessing data
  3) cleaning data
- Storing, analyzing and visualizing the cleaned data
- Writing reports:
  1) describing data wrangling efforts
  2) data analyze and visualization

2. Gathering Data

In this project, the following three datasets were worked on.

1) WeRateDogs Twitter Archive

The data was named as archive, which is provided by Udacity. The link provided allows you to download the file "twitter_archive_enhanced.csv" directly.

2) Prediction Data of Twitter Images

This data is the result of predicting the breed of dog (or other objects, animals, etc.) that appears in each tweet, based on a neural network, and was named as pred_image. This data is downloaded through the link provided by Udacity, and the downloaded file name is image_predictions.tsv.

3) Additional Data

The data was named as tweet_df, including tweet ID, retweet_count, favorite_count, which is taken from the "tweet_json.txt" file provided by Udacity.

3. Assessing Data

Quality Issue

1) Archive Data Issue

- The data type of tweet_id should be object
- The data type of timestamp, retweeted_status_timestamp should be datetime
- Some are retweets (We only want original ratings)
- Some columns are not necessary for the analysis
- Poor readability of the data in source column
- some names in name column are false

2) Pred_image Data Issue

- Some breeds of dogs have lowercase initials
- The type of tweet_id should be object

Tidiness Issue

- doggo, floofer, pupper, puppo these 4 variables shoule be combined into one categorical variable Dog Type
- 3 dataframe should merge to be one

4. Cleaning Data

1) Cleaning Archive Data

- Change the data type of tweet_id to object.
- Change the data type of timestamp, retweeted_status_timestamp to datatime.
- Select original ratings.
- Delete these columns ('in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp', 'expanded_urls') which are not necessary for the analysis.
- Extract the user source according to the source column and put it in the new column 'user_source'.
- Remove None and items with lowercase initials in name column.

- Combine columns for dog types and remove rows where one dog has two types.

2) Cleaning Pred_image Data

- Capitalize the first letter of each word in the columns p1, p2, p3.
- Change the data type of tweet_id to object.

3) Combine three tables

- Combine three tables according to tweet_id