Wrangle Report¶

Introduction

This projcet is for wrangling data from WeRateDogs Twitter account using Python and then save it to notebook titled (wrangle_act.ipynb)

The goal of this project is to wrangle the WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. The.

There are 3 Dataframes in this analysis as follows:

Enhanced Twitter Archive

The WeRateDogs Twitter archive as (twitter_archive_enhanced.csv)

Image Predictions File

The tweet image predictions as (image_predictions.tsv) hosted on udacity server and will imported programmatically

Twitter API

using Tweepy to query Twitter's API for additional data beyond the data included in the WeRateDogs Twitter archive Before we start, herea are few points to keep in mind when data wrangling for this project:

In assessing phase I selected these problems for each dataframe to solve

Quality issues: For twitter-archive-enhanced

- Missing data columns (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, etweeted_status_user_id, retweeted_status_timestamp, expanded_urls)
- Keep original ratings (no retweets) that have image as requested
- drop columns that's not important for analysis
- timestamp is object
- All of rating_denominator should be 10
- tweet_id to str
- tweet_id to object to merge

Tidiness issues:

- The for last columns need to be modified

Quality issues: For df_image_predictions

- Delete columns that won't be used for analysis
- Drop 66 jpg_url duplicated
- tweet_id to object to merge
- Renaming the column names
- tweet_id to object to merge

Tidiness issues: df_json

- create a coulmn for the dog type

Quality issues: For df_json

- Renaming the column names

Steps of wrangling

For twitter-archive-enhanced

- Kept the original ratings using isnull
- Drop columns that I don't need for my analysis
- Edit timestamp with pandas function
- Edit rating_denominato using replace function

    #### Tidiness

- modified the four last columns using if conditions

Quality issues: For df_image_predictions

- Delete columns that won't be used for analysis
- tweet_id to object to merge using astype fucntion

    #### Tidiness

- Created dog name column using if conditions and other functions

Quality issues: df_json

- tweet_id to object to merge using astype fucntion