Wrangle Report

Introduction

As a data analyst, our job is to collect, evaluate, and clean data to achieve analysis and decision-making through it.
WeRateDogs is a popular Twitter account that posts and rates dogs from photos submitted by their followers. They have 8,9 M of followers by March of 2021.
WeRateDogs has even developed it's own dog classification, base on the dog's. Puppies are "puppers", older puppers are "puppos", older dogs are called "doggos", hairy dogs are classified by their appearance and they are called "floofers".
Their archive is provided by Udacity in 3 parts, so we gather each of the three pieces of data as CSV file, a programmatically request hosted on an Udacity server and an API.
After gathering, the data is assessed to find quality and tidiness issues, then clean it to finally analyse it.

1. **Gathering data**

   - Twitter_archive_enhanced.csv is provided by Udacity and is loaded into a dataframe called “archive_df” with Pandas.
   - Image prediction is downloaded through an url programmatically into a dataframe called “image_prediction”.
   - Finally, JSON file is assigned to a dataframe called “df_api”.

2. **Assessing data**

   Quality

   - Replace all faulty names to none
   - Assign 10 to rating_denominator
   - Convert timestamp columns to date
   - Create dog_stage column combining 'doggo', 'floofer', 'pupper' and 'puppo'
   - "tweet_id" and "tweet_id" are numeric and not categorical (string)
   - 2075 tweet ids present (while the archive dataset has a total of 2356 ids, so 281 IDs are missing)
   - column names can be improved
   - p1, p2 and p3 contain underscores instead of spaces in the labels
   - ID variables are floats and integers
   - "in_reply_to..." and "retweeted_status..." variables are numeric
   - retweets are present in the data
   - some of the column names are not meaningful

- "timestamp" and "retweeted_status_timestamp" are not a datetime variable
- "source" values are not formatted as <a> href=url
- rating_numerators are not always correctly accounting for decimals
- the dog names are not standardized

Structure (tidyness)

- more than one stage is filled for a particular dog
- "source" and "expanded_urls" have several information inside them
- columns "doggo", "floofer", "pupper" and "puppo" refer to the same measurement unit, i.e., dog stage
- eliminate retweets

3. Cleaning data

- Make a copy of every dataframe.
- Remove retweets where the status contains “@” on text columns.
- Combine dog stage columns (doggo, floofer, pupper, puppo) into one dog_stage column.
- Convert timestamp to datetime.

4. Store data

- Save the new dataframe to a csv file.