Wrangle Report

Norah Alotaibi
June 30, 2019

Step 1: Gather Data

- Using the following import:
  import numpy as np
  import pandas as pd
  import requests
  import tweepy
  import json
  import matplotlib.pyplot as plt

- Data is gathered from 3 resources from file provided in the project by Use pd.read_csv()

1. df1 has file twitter-archive-enhanced-2.csv.
2. df2 has file image_prediction-3.tsv
3. Gather data from twitter API using Python's Tweepy library from twitter API. By using data in the text file tweet_json.txt, then read the file and store data in df3, and store the following columns as retweet_count and favorite_count.
4. Using info() function for all df1, df2, and df3

Step 2: Assess Data

Quality

- Some tweet_ID is missing, and the tweet_ID is not the right data type.
- Erroneous data types and values for in_reply_to_status_id,in_reply_to_user_id, and timestamp.
- Using only tweet_id with images.
- In twitter-archive-enhanced, which is df1, some ratings are wrong such as: rating_numerator column has values < 10 as well as some very large numbers like 1776. Also, rating_denominator column has values not equal to 10.
- Some dog names are not correct with lowercase characters.

Tidiness

- Columns in df1 `retweeted_status_id', 'retweeted_status_user_id' ,and 'retweeted_status_timestamp` are not needed, which can be dropped.
- The four columns in df 1 which are `doggo, floof, pupper and puppo` should be merged into one column named `stage`
- `retweet_count and favorite_count` columns from df3 table should be joined with df1.

- `rating_numerator and rating_denominator` should be merged into one column named `rating`.

Step 3: Clean Data

Copy df1 to df1_clean
Copy df2 to df2_clean
Copy df3 to df3_clean

Issue 1

- Using only tweet_id with images

Define

- Choose tweet_id with image in df1_clean using df2_clean has image_prediction-3.tsv

Issue 2

- Erroneous data types and values for in_reply_to_status_id,in_reply_to_user_id, and timestamp.

Define

- Convert the following: in_reply_to_status_id and in_reply_to_user_id to data type integer. And, Convert timestamp to datetime data type.

Issue 3

- Some dogs names are not correct with lowercase characters.

Define

- Set wrong names to the value 'None' and replace 'None' with np.nan.

Issue 4

- Columns in df1_clean retweeted_status_id', 'retweeted_status_user_id' ,and 'retweeted_status_timestamp are not needed, which can be dropped.

Define

- Delete the following columns: retweeted_status_id', 'retweeted_status_user_id' ,and 'retweeted_status_timestamp.

Issue 5

- The four columns in df 1 which are doggo, floof, pupper and puppo should be merged into one column named stage.

Define

- Create column 'stage' to show dog stage,THEN drop columns 'doggo','floofer','pupper','puppo'. Replace 'None' with np.nan.

Issue 6

- retweet_count and favorite_count columns from df3 should be joined with df1_claen.

Define

- *Join df3_clean into df1_clean using tweet_id.*

Issue 7

- rating_numerator and rating_denominator should be merged into one column named rating.

Define

- Create new column in df1_clean rating=rating_numerator/rating_denominator. And, Drop rating_numerator and rating_denominator.

Step 4: Store Data

- Store the clean data in df1_clean as DataFrame in a CSV file named 'twitter archive master.csv' as requested in the project.

Step 5: Analyze and Visualize Data

- In this section use graphs to show the relations between the following:
    1. Rating with dogs numbers.
    2. Favorite and Retweet with rating
    3. Dogs stage with dogs numbers.
    4. Dogs stage with rating
    5. Dogs stage with Favorite and Retweet
    6. Top 5 commons names of dogs