Date: July 24, 2022

Purpose

This document chronicles the data wrangling processes that led to the creation of the twitter_archive_master.csv dataset

Scope

This document is meant for the data analyst/scientist in the organization.

Definition and Acronyms

- Data wrangling – The process of gathering, assessing and cleaning data.

Procedures

- **Step 1: Gathering the data**
  - Pandas **read_csv** method was used to read in the provided "twitter_archive_enhanced.csv" dataset
  - Python requests library was used to download the second dataset from the link given. The download data was also read with pandas as image_predictions.
  - I could not geta twitter developer account so I downloaded the Twitter with the link provided. This data was stored as tweet_json.txt. Eventually the data was read as a pandas' Data Frame.
- **Step 2: Assessing the Data**
  - Visual inspection of the three datasets were done. This was done to identify quality and tidiness issues in the datasets.
  - Step 2.2: Programmatic inspection was also carried out using methods such as info, describe, sample etc. This enabled me to easily identify issues with datatypes as well as missing values. Upon visual and programmatic assessment, some of the identified issues included:

1. time stamp as string instead of datetime variable in arch dataset

2. 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id','retweeted_status_user_id', 'retweeted_status_timestamp' all have null values.

3. tweet-id in arc is 2356 while in image it is 2075 which means 281 ids are retweets that are not needed

4. tweet-id in image is an integer instead of a string

5. source in arc are formatted in html values which should be categorical

6. Typographical error in dog names (Jessiga, Fwed, Sampson)

7. multiple denominator rating values

8. invalid dog names like 'a' 'the' 'unacceptable etc. in arc

- **Step 3: Data Cleaning**

  Following a pattern – Define, Code and Test, each of the identified quality and tidiness issues were fixed in the following manner:

  - The rows containing the retweets were dropped using pandas Data Frame drop method.
  - The timestamp column was converted to a datetime using pandas to_datetime method.
  - Irrelevant columns (*in_reply_to_status_id*, *in_reply_to_user_id*, *retweeted_status_id* and *retweeted_status_user_id*) were dropped.
  - Illegal dog names were replaced with "Null" were applicable.
  - Typographical error in dog names (Jessiga, Fwed, Sampson) were fixed.
  - The 4 dog status columns *(doggo, floofer, pupper, puppo*) columns were combined into to a single *dog_status* column and afterwards, dropped.
  - The arc_copy dataset and the count_copy dataset was joined to the image_copy table on tweet id.
  - multiple denominator rating values were corrected to 10 and the rating _numerator table was renamed as rating.
  - The source in arc were formatted in html values which and categorized to ('iPhone', 'Web Client' etc.)

- **Step 4: Data Storage**

  Finally, the cleaned master dataset was saved to a CSV file named "twitter_archive_master.csv".