Date: 24th February, 2023

PURPOSE:

This document details the data wrangling steps that led to the creation of the twitter_archive_master.csv dataset.

SCOPE:

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

DEFINITION AND ACRONYMS:

- Data wrangling – This is the process of gathering, assessing and cleaning data.

PROCEDURES:

- Step 1: **Data Gathering**
  - Step 1.1. Pandas read_csv method was used to read in the provided "twitter_archive_enhanced.csv" dataset
  - Step 1.2. Python requests library was used to download the second dataset from this link. The download data was also read with pandas as image_predictions.
  - Step 1.3. The Tweepy library was used to query additional data via the Twitter API. This data was stored as tweet_json.txt. Eventually the data was read as a pandas' DataFrame.

- Step 2: **Assessing the Gathered Data**
  - Step 2.1: 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 data types as well as missing values. Upon visual and programmatic assessment, some of the identified issues included:
    1. Typographical errors in dog name
    2. Improper Dog names
    3. Wrong data type
    4. Irrelevant tweets (some tweets are not about dogs)
    5. One variable in different columns
    6. The same record in different data sets
    7. Rows containing retweets
    8. Rows with zero rating numerator and zero rating denominator

9. Incomplete dog names (Dog name at tweet with tweet_id 776201521193218049 is O instead of O'Malley)

- Step 3: **Data Cleaning**

  The data cleaning steps followed a define, code, and test pattern.

  Each of the identified quality and tidiness issue was fixed following the under-listed steps –

  - The 181 rows containing retweets were dropped using pandas DataFrame drop method.
  - Rows containing non-canine tweets were removed.
  - in_reply_to_status_id and in_reply_to_user_id rows that mostly contained null values were dropped.
  - The timestamp column was converted to a datetime using pandas to_datetime method.
  - Improper dog names were replaced with “None (nan)”
  - Dog name “O” was replaced with “O’Malley” in the name column of tweet with tweet_id 776201521193218049
  - Misspelt dog name “Shawwn” was replaced with “Shawn”
  - Rows where rating_numerator or rating_denominator is 0 were removed.
  - All the dog stages (4 columns) were combined into a single column dog_stage and the irrelevant columns were deleted
  - All the datasets were merged into a single table using pandas DataFrame merge method.

  Step 4: Data Storage

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