Date: May 1st, 2022

Purpose

This document highlights 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: 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 datatypes as well as missing values. Upon visual and programmatic assessment, some of the identified issues included:
        - records that did not contain original tweets from WeRateDogs.
        - Tweets of dogs with no images/tweets of other animals/items e.g., Mechanic, Penguin, Black bears etc.
        - The wrong data type (timestamp column).
        - Irrelevant columns.
        - Illegal dog names e.g., an, this, unacceptable, all, the, by, such, such etc.
        - Valid dog names in text column replaced (e.g., with 'a') in name column.
        - Incomplete names and Typographical error in dog names
        - One Variable (Dog Stage) in 4 columns
        - The same records in two different datasets

- **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 DataFrame drop method.
  - The duplicate tweet with dog name Canela was removed.
  - Tweets from “twitter_archived_enhanced” dataset that had no image or whose image did not match any dog were removed.
  - 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 “Nan” were applicable.
  - Some dog names were extracted from the text column and used to fill the missing dog name column while others were replaced with None.
  - 'O' and ‘his’ was replaced with O'Malley and Quizno respectively.
  - Potential 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 image_predictions dataset and the extra_tweets dataset were joined to the twitter_archive_enhanced table on tweet id. Then the redundant id column was dropped.

- **Step 4: Data Storage**

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

Related Resources

- Pandas API reference documentation.