Introduction

This report is on the data wrangling process that led to the creation of the twitter_archive_master.csv dataset file. Data wrangling is the process of gathering, cleaning and unifying messy and complex data sets for easy access and analysis.

Steps of Wrangling

1. Gathering data
    - In this step the "twitter_archive_enhanced.csv" dataset was loaded into a dataframe called tweet_archive which was given to us.
    - The image prediction file was obtained using the requests library with this link and was saved as a file using os library and loaded into an image dataframe.
    - The retweet count and favorite count of the tweets id were obtained using the tweepy library to access the twitter api which created a json file which was read in a file and the id, retweet count and favorite count into retweets dataframe
2. Accessing data
   The dataframes were assessed visually and problematically. Problematically assessment methods like info(), describe(), value_count() etc were used. Some of the quality and tidiness issues which were
    - The wrong data type in the timestamp column.
    - Irrelevant columns.
    - Null represented as None in name, pupper, doggo, floofer and puppo column
    - Text column shows evidence of gender
    - Typographical error in dog names
    - One Variable (Dog Stage) in 4 columns
    - The same records in two different datasets
    - Repetitive words in source column
3. Cleaning data
   The dataframe was cleaned using the define code and test method. The dataframe were cleaned the following way:
    - The timestamp column was changed from string to datetime datatype
    - Irrelevant columns such as expanded url, text etc
    - None values in name, pupper, doggo, floofer and puppo column were replaced with nan
    - The dog gender was extracted from the text column
    - The names of dogs in the names column with typographical errors were replaced with the right names
    - The four columns of the dog stages were placed in a column
4. Storing data
   The cleaned dataframe were stored in a csv file called twitter_archive_master.csv

.