Wrangle Report For WeRateDogs Twitter Archive

David Onwachukwu

DAND Project 2

May 2022

This report briefly describes the wrangling effort carried out on three datasets which were used for this project. I carried out four main processes, namely;

- Data Gathering
- Assessment
- Cleaning
- Storing

Data Gathering

I gathered the three datasets from different sources and stored them separately as follows:

1. WeRateDogs twitter enhanced archive, which was manually downloaded from the Udacity classroom. This dataset was in csv format and was eventually loaded into a pandas dataframe.
2. The image predictions file, which was programmatically downloaded using the requests library. This dataset was in tsv format and was eventually converted to a pandas dataframe
3. I used tweepy library to query the twitter API and gather each tweets' JSON data which had the retweet and favorite counts of majority the tweets. The data was then loaded into a dataframe.

The dataframes in order of gathering are; archived_tweets, pred_images and tweets_api.

Assessment and Cleaning

I carried out two forms of assessment, visual and programmatic to check for quality and tidiness issues. For the archived_tweets, quality issues assessed include;

- wrong names for some dogs e.g. ‘a’, which was changed to ‘np.nan’ to represent missing dog data.
- Numerators that were 0 were dropped.
- Denominators above ten should were dropped to ensure consistency and follow the unique rating system.
- Wrong data type in timestamp and tweet_id column. Timestamp was converted to datetime and tweet_id to string during cleaning
- Rows that had retweets were dropped and their columns were dropped afterwards.
- Source list should be in categories based on device or platform tweeted from. During cleaning this was carried out.

Tidiness Issues

- Dog stage which is one variable was spread out across four columns, it was put into one column during cleaning.
- Information about one type of observational unit (tweets) is spread across three different files/dataframes.

The image predictions table had two main issues, the img_num column which was I did not use for analysis at this time was dropped and the columns which had three false predictions were also dropped. The tweets_api table had three notable issues, the id column had to be renamed to tweet_id to ease merging, two columns had missing data in both retweet and favorite counts and were dropped, after these two issues were tackled, the tweets_api table was merged with the archived_tweets table on the tweet_id column.

The predicted images table was then merged with the already merged archived_tweets and tweets_api table and since for analysis, only tweets with images were needed, tweets without images were dropped.

Storing

The gathered, assessed and cleaned master dataset was then saved to a CSV file named twitter_archive_master.