Udacity Data Analysis Nanodegree Project Report

Project: Wrangle & Analyze Data from WeRateDogs Twitter

June 2022

WeRateDogs is a Twitter account that rates people's dogs with humorous comment about the dogs. These ratings almost always have a denominator of 10. WeRateDogs has over 4 million followers and has received international media coverage.

WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively to be used in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets.

This document records the data wrangling efforts made; data gathering, assessment, cleaning and merging the three datasets into one master document to be used in data analysis and visualization.

Data Gathering

Three sources of data were provided for this wrangling task.

- **Tweet Archive:** A CSV file named 'twitter_archive_enhanced' provided by Udacity and downloaded and stored to my local directory. The CSV files contains 5000+ Tweets content from WeRate Dogs.
- **Image Predictions:** A Tab separated value text who url was used to extract it's content from a website. The content was extracted programmatically using python requests library. Stored in a image-predictions.tsv file.
- **Tweet information extracted with TwitterAPI:** I used the text file provided by Udacity, and loaded the file content with JSON.

Assessing the Data

The data was assessed programatically and visually for Quality Issues.

Quality issues: This is when Data has quality issues has problems with its contents: missing, corrupted, inaccurate, duplicate or incorrect data.

Tidiness Issues: This is when the data has specific structural issues. It is also called "dirty data", "untidy" or "messy" data.

Quality Issues

1. Only original ratings (no retweets) are to be included, therefore retweets and reply will be identified and dropped.

2 The following columns have missing data and they will be dropped.

- in_reply_to_status_id
- in_reply_to_user_id
- retweeted_status_id
- retweeted_status_user_id
- retweeted_status_timestamp

3 Columns with Missing values in the dataset will be dropped.

4 Timestamp data type is not supposed to be object, datatype will be changed to datetime64.

5 Extract URL from HTML anchor tags in the source column.

6 The None in the four columns: [‘doggo’, ‘floofer’, ‘pupper’, ‘puppo’] will be dropped.

7 The correct rating_numerator will have to be extracted from 'text' column.

8 Remove extreme values from the numerator and denominator columns.

9 id column in the tweets dataframe will be renamed as tweet_id.

Tidiness Issues.

1 Dog "stage" is spread in four columns. [“doggo”, “flooter”, “pupper”, “puppo”], it will be combined into “stages” column.

2 The column text had multiple variables like a URL link, rating, and some tweets.

3 The three dataframe will be merged into one because the rows in each are all for the same observations.

Cleaning the Data

After all the Issues listed above were cleaned, the three datasets were merged into one dataframe called ‘df_master_dataset_copy’.

Storing the Data

The data was stored as a CSV file named **‘twitter_archive_master.csv’.**