WeRateDogs

Data Wrangling Project – Wrangle Report

Cyndi Morris
WGU D309
Udacity Data Analytics Nanodegree

Introduction

This is the wrangle report of the analysis of Twitter's WeRateDogs as part of the Wrange and Analyze Data project at Udacity. The report documents the wrangling performed during the gathering, assessment and cleaning steps.

Gathering Data

The project required gathering three different datasets with three different file formats.

- The WeRateDogs Twitter archive, twitter_archive_enhanced.csv, as provided by Udacity for this project.
- The tweet image predictions, image_predictions.tsv, hosted on Udacity's severs was downloaded programmatically using the Requests library.
- The Twitter API, tweet_json.txt. I chose to also download the file programmatically using the Requests library since I don't do social media.

Assessment – Data Quality Issues

The twitter_archive dataset:

- Not all of the cols were needed for the analysis.
- The data included retweets which were not needed for this analysis.
- The timestamp col included extra numbers and was a str dtype, not datetime.
- The timestamp col's date and time was all in one col.
- Some of the rating numerators and denominators were incorrect.

The twitter_json dataset:

- Not all of the cols are needed for this analysis.
- The twitter_id col was named as id in this set.

The image_predictions dataset:

- Not all of the cols were needed for this analysis.
- The p1 and p1_conf col names were not properly descriptive.

Assessment – Tidiness Issues

- The twitter_archive's doggo, floofer, pupper and puppo cols are separate, but should be a in a single col.
- Information from the three different tables should be in a single table.

Cleaning

The twitter_archive dataset:

- Removed unneeded cols for this analysis using df.drop().
- Removed retweet rows using .isnull() to identify rows with a retweeted_status_id of NaN and keep only those rows; removing retweets.
- Removed +0000 from timestamp and convert it to datetime using pd.to_datetime(). Split the timestamp into separate date and time cols using .dt.date, .dt.time the dropped timestamp col since it was no longer needed. Finally, the date and time cols were moved next to the tweet_id col for easier reading.
- Recalled all entries with a denominator !=10. Corrected the ratings based on earlier exploration. Created a function that used correct ratings using information manually collected from the text col. Created a function to remove non-ratings rows using tweet_ids manually collected for those rows lacking a rating.

The twitter_json dataset:

- Created a dataframe with only the id, favorite_count and retweet_count cols.
- Renamed the id col to tweet_id.

The image_predictions dataset:

- Creater a dataframe with only the tweet_id, p1, p1_conf, and p1_dog cols.
- Renamee the cols p1 and p1_conf to be more descriptive; p1 to breed, p1_conf to confidence level.
- While cleaning, it became apparent that some additional cleaning was needed to make sure the dataset info was relevant to the analysis. Removed rows with non-dog images to make sure the data was relevant to the analysis.

Tiddiness

- Merge the doggo, floofer, pupper, and puppo col values into a single col and remove the individual cols. For further clarity, remove any rows where dog stages are a combination of stages.
- Merged all three tables into one table and saved as master.csv file for analysis.