Project 2

Data Wrangling on WeRateDog

Ebenezer Acquah

Udacity Data Analysis

Project 2

Table of Contents

DATA WRANGLING PERFORMED ON WeRateDog DATASET

This report documents how I gathered, assessed, and cleaned WeRateDogs data from Twitter Archive using Twitter API for better visualization and analysis.

FIRST STAGE: GATHERING DATA

Here, I gathered 3 different datasets from 3 separate sources. They are:

1. WeRateDogs dataset from Twitter archive data (twitter_archive_enhanced.csv).
2. Used the request library to download the tweet image prediction (image_prediction.tsv)
3. Used the tweepy library to query additional data via the Twitter API (tweet_json.txt).

SECOND STAGE: ASSESSING DATA

I analyzed the dataset visually using Power Bi and programmatically using the pandas library. The list shown below is the issues I found with the dataset.

QUALITY ISSUES:

Twitter Archive Data

- There were missing data in in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls columns.
- The data type of timestamp column was string instead of datetime.
- Some ratings were incorrectly labeled.
- Ratings with decimal values in the numerator were incorrectly extracted.
- Some of the dog names were invalid: examples include an, as, space, etc.

Image Prediction Data

1. Some of the images were not dogs. Examples include data from Index 290 and 403.

- The names of the dogs are not standardized.

Tidiness Issues

1. Multiple columns for dog type in the Twitter archive data

- All 3 columns in the data sets refer to the same type of observation, however, it was contained in 3 separate data frames.

Data Cleaning

I cleaned the data based on the issues I found when investigating the dataset.

- Dropped retweet rows
- Dropped in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls columns.
- Change the timestamp column to datetime data type.
- Changed incorrect ratings for rows where the rating denominator was wrongly extracted.
- Replaced inappropriate values with “None”
- Dropped rows with non-dog images
- Standardized the dog names.
- Extracted ratings from the text where the rating numerator contained decimals and updated the ratings.
- Merged all the three dataset into one.

SAVING DATA

After merging all the dataset into one, I saved it as csv file.