Wrangle Report

Table of Contents

1. Introduction

The purpose of this project is to put in practice what I learned in data wrangling data section from Udacity Data Analysis Nanodegree program. The dataset that is wrangled is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs.

2. Gathering data

1) **Twitter archive file:** The WeRateDogs Twitter archive. I am giving this file to you, so imagine it as a file on hand. Download this file manually by clicking the following link: twitter_archive_enhanced.csv
2) **The tweet image predictions**, The tweet image predictions, i.e., what breed of dog (or another object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: (https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv)
3) **Twitter API & JSON:** Each tweet's retweet count and favorite ("like") count at minimum, and any additional data you find interesting. Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON

data in a file called tweet_json.txt file. Each tweet's JSON data should be written to its own line. Then read this .txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count. Note: do not include your Twitter API keys, secrets, and tokens in your project submission.

3. Assessing data

i. Quality

1) twitter_archive

- Keep original ratings (no retweets) that have images
- Delete columns that won't be used for analysis
- Convert data type in 'tweet_id' column from an integer to string.
- the timestamp should be date-time instead of an object (string).
- In several columns, null objects are non-null (None to NaN).
- The NA value in the name column is not inaccurate data format.
- There are many invalid names ('just, 'None', 'a', 'an', 'all'). - We only want original rating tweets, not retweets.
- Sources are not readable.
- Erroneous data types (doggo, floofer, pupper and puppo columns)
- The numerator and denominator columns have invalid values.
- Extract the rating numerator from the 'text' column because some values in the 'rating_numerator' column are wrong.
- Convert the data type in both 'rating_numerator' and 'rating_denominator' columns as a float.

2) image_predictions

- Some tweet_ids have the same jpg_url and drop it
- Simplify the table by keeping only one prediction, according to the odds priority order is as p1 > p2 > p3;
- Convert data type in 'tweet_id' column from a integer to string.

3) Api_tweet

- Convert data type in 'tweet_id' column from integer to string.

ii. Tidiness

- Merge dog stage column into a single column instead of 4 columns (doggo, floofer, pupper, puppo)
- Consolidate the 3 tables.

4. Cleaning data

- This part of the data wrangling was separated into three parts: Define, code, and test the code. These three steps were on each of the issues described in the assessment section.
- First create a copy of the three original data frames. I wrote the codes to manipulate the copies.
- To prepare for the analysis part, these three data sources/tables needed to be merged with each other. To create a better overview, I dropped unneeded columns.
- Whenever I made a mistake, I could create another copy of the data frames and continue working on the cleaning part.
- The original table had three predictions and confidence levels. I filtered this into one column for dog type and one column for a confidence level.
- Another interesting cleaning code was to melt the dog stages in one column instead of four columns as originally presented in the Twitter archive.

5. Storing

The final Data Frame called 'twitter_archive_clean' contains 1990 rows and 15 columns with the correct data types. It then stores the dataset in a CSV file called 'twitter_archive_master.csv'. The data was successfully wrangled and therefore ready for analysis and visualization.

6. Analysis & Visualization

Visualizations and insights are provided in ‘act_report.pdf’