Wrangling WeRateDogs Dataset Report

1 PROJECT DESCRIPTION

This project is part of the Udacity Data Analysis program, and the one which is about the wrangling part. In this project, I practiced what I learned throughout the nanodegree starting from Python programming ending to wrangling and visualizing the dataset.

In this report, I'm going to briefly describe my wrangling efforts which were in three main steps which are:

- Gathering the datasets.
  - Twitter archive CSV file.
  - Tweet image predictions TSV file.
  - Tweet-json TXT file.
- Assessing the datasets.
  - Quality Issues
  - Tidiness Issues
- Cleaning the datasets.
  - Define.
  - Code.
  - Test.

2 GATHERING

1. The WeRateDogs Twitter archive CSV file, downloaded manually.
2. The tweet image predictions TSV file downloaded programmatically using the **Requests** library.
3. The tweet-json TXT file, downloaded manually because I couldn't get the developer account from twitter.

3 ASSESSING

After gathering each of the above pieces of data, I assessed them visually and programmatically for quality and tidiness issues.

I detected and documented **(12) quality issues** and **(2) tidiness issues** which could be found in the jupyter notebook.

3.1 QUALITY ISSUES

3.1.1 twitter_archive_df Table

3.1.1.1 timestamp and retweeted_status_timestamp are string not datetime.

3.1.1.2 Some names are missed, such as 'a' and 'an'.

3.1.1.3 some rows does not contain dog ratings but retweets instead.

3.1.1.4 some columns are for retweets, such as retweeted_status_timestamp, retweeted_status_id, retweeted_status_user_id.

3.1.1.5 Some rating_numerator are decimal!

3.1.2 img_pred_df Table

3.1.2.1 Duplicated tweets by using the same jpg_url.

3.1.2.2 p2 contains inconsistency in algorithm's predictions names.

3.1.3 tweet_json_df Table

3.1.3.1 Non-related columns that contain only NaN values like contributors, coordinates, geo.

3.1.3.2 Non-related columns contain less than 30 values compared to 2354 entries in other columns.

3.1.3.3 created_at is string not datetime.

3.1.3.4 id, retweet_count, and favorite_count is from the most useful columns.

3.1.3.5 followers_count in user series would be good for visualization.

3.2 TIDINESS

3.2.1.1 All tables should be combined into one single table, since they are all describing the tweets themselves.

3.2.2 twitter_archive_df Table

3.2.2.1 - doggo, floofer, pupper, puppo is a single column.

4 CLEANING

I created a copy of the datasets and cleaned each of the issues that I documented while assessing, in the same notebook wrangle_act.ipynb as well. The result is a high quality and tidy master pandas DataFrame which is twitter_archive_master.

The cleaning part consisted of 3 main parts which were:

- Define
- Code
- Test

4.1 DEFINE

In this part, I defined the cleaning tasks that I would perform after that.

4.2 CODE

In this part, I performed the defined tasks above to clean the dataset quality & tidiness issues.

4.3 TEST

In this part, I tested the performed tasks to make sure that the dataset is cleaned as defined.