Wrangle Report for tweet archive of Twitter user @dog_rates

Introduction

The dataset that was wrangled is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. But the numerators are almost always greater than 10. 11/10, 12/10, 13/10, etc. WeRateDogs has over 4 million followers and has received international media coverage.

Using Python and its libraries, data has been gathered from a variety of sources and in a variety of formats, assessed for its quality and tidiness, then cleaned. This is called data wrangling. I have documented my wrangling efforts in this Jupyter notebook and performed analysis and visualizations using Python (and its libraries).

Below, each performed step of the wrangling process is discussed.

Gathering Data

The three pieces of data were gathered as mentioned below:

1. WeRateDogs Twitter archive: This file had to be downloaded manually by clicking on the provided link.
2. tweet image predictions file: This file hosted on Udacity servers, had to be downloaded programmatically using Requests library and the provided URL.
3. tweet_json.txt file: After signing up for Twitter developer account, I have setup my Twitter applicationUsing the tweet IDs in the WeRateDogs Twitter archive, I have queried the Twitter API using Tweepy for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire set of JSON data in a file called tweet_json.txt file. Each tweet's JSON data was written to its own line. Then this .txt file has been read line by line into a pandas DataFrame with tweet ID, retweet count, and favorite count. As mentioned, Twitter API keys, secrets, and tokens have been removed in the project submission.

Assessing

After gathering each of the above pieces of data, they were assessed visually and programmatically for quality and tidiness issues. The issues that satisfy the Project Motivation were also assessed.

Quality

archive_df table

- Records with non null values in 'retweeted_status_id' and 'retweeted_status_user_id' and 'retweeted_status_user_id' columns do not belong in the df.
- Records that have non null values for 'in_reply_to_status_id' and 'in_reply_to_user_id' should not be present in the df.
- All five columns mentioned in above two lines do not help with the analysis.
- 'timestamp' column not required in the df.
- 'expanded_urls' and 'source' columns not useful for analysis.
- Some rows have incorrect values for rating numerators and denominators.

breed_predict table

- Some of the images do not display dogs.
- p2 and p3 data and columns are not required for analysis.
- 'jpg_url' and 'img_num' columns are not required for analysis.

Tidiness

- Dog stage data breaks the 'Each variable forms a column' tidy rule.
- The 'name' column has values 'None', 'a', 'an', 'the', 'not', 'one' which need to be replaced by 'NA'.

Cleaning

The below mentioned operations were performed on copies of the datasets.

Quality Isssues

archive_df

- Delete records with non null values in 'retweeted_status_id' and 'retweeted_status_user_id' and 'retweeted_status_user_id' columns
- Delete records that have non null values for 'in_reply_to_status_id' and 'in_reply_to_user_id'.
- Drop the above mentioned columns for remaining records.
- Delete 'timestamp' column.
- Delete 'expanded_urls' and 'source' columns.
- Changed the rating numerator and denominator values for the incorrect values according to the text.

breed_predict

- Drop columns containing p2 and p3 data.
- Drop 'jpg_url' and 'img_num' columns.
- Delete records having 'p1_dog' value False.

Then, all three datasets merged into a single dataframe df_final on which the tidiness operations were performed.

Tidiness Issues

- A single column 'dog_stage' created using np.select() function. For ids with two dog_stages mentioned in the original df,I have mentioned both the stages under dog_stage column with a '/' in between them.
- The 'name' column has values 'None', 'a', 'an', 'the', 'not', 'one' which were replaced by 'NA' using replace().