Project: WeRateDogs tweets

In this project, I wrangled a dataset of tweets of the WeRateDogs Twitter account. I have used my best data wrangling efforts to gather, assess and clean the tweet data.

The data wrangling was part of my bigger goal to analyze and find out

- the most favored tweets according to the likes and retweets
- top breeds within the top 50 tweets utilizing the image prediction data
- most common breeds recognized in all the tweets
- most common dog names in all the tweets

I will be reporting the results separately in another project report.

Step 1: Gathering Data

I started by gathering data from three different sources.

- WeRateDogs Twitter archive
    - A csv file given in the beginning of the project.
- Twitter API
    - Using the tweet IDs in the WeRateDogs Twitter archive, querying the Twitter API for each tweet's JSON data using Python's Tweepy library and storing each tweet's entire set of JSON data in a file called tweet_json.txt file.
    - Each tweet's JSON data is written to its own line. Reading the .txt file line by line into a pandas DataFrame.
- Image predictions file
    - The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers.

Step 2: Assessing Data

Next, I assessed the data for its quality and tidiness.

- Completeness:
    - Are there missing data?
- Validity:
    - We have the records, but they're not valid, i.e., they don't conform to a defined schema. A schema is a defined set of rules for data. These rules can be real-world constraints (e.g. negative height is impossible) and table-specific constraints (e.g. unique key constraints in tables).
- Accuracy:
    - Inaccurate data is wrong data that is valid. It adheres to the defined schema, but it is still incorrect.
- Consistency:
    - Inconsistent data is both valid and accurate, but there are multiple correct ways of referring to the same thing. Consistency, i.e., a standard format, in columns that represent the same data across tables and/or within tables is desired.
- Tidy data requirements:
    - Each variable forms a column.
    - Each observation forms a row.
    - Each type of observational unit forms a table.

Quality Issues

Twitter archives file:

- missing data in expanded_urls column for some of the tweets
- some of the tweets are replies or retweets
- timestamp should have datetime as dtype
- name, doggo, floofer, pupper and puppo columns have missing values as a string 'None' which means these values don't show up as NaN or null values
- unnecessary columns

Tweet data from Twitter API:

- 2329 tweets, missing 27 tweets compared to the original archive file (df_1)
- missing data in following columns: extended_entities, geo, coordinates, place and contributors
- some of the tweets are replies or retweets
- favorited, retweeted, possibly_sensitive and possibly_sensitive_appealable columns contain only boolean values equaling false
- created_at should have datetime as dtype
- unnecessary columns

Image predictions file:

- 2075 tweets, missing 281 tweets compared to the original archive file (df_1)
- some breeds are written in lower case

Tidiness Issues

- Twitter archives file:
    - doggo, floofer, pupper and poppo columns have values as headers

- Tweet data from Twitter API:
    - entities and extended entities contain a table of extra information in one cell per tweet
    - retweeted_status and quoted_status contain a table of another tweet information in one cell per tweet

- All the data is divided in three separate dataframes.

Step 3: Cleaning Data

Finally, I fixed the quality and tidiness issues identified in the data assessment.

Data cleaning process:

- defining
    - missing data and completeness issues
    - tidiness issues
    - quality issues
- coding
- testing

I started by making a copy of the original dataframes to "_clean" and removing unnecessary columns. I merged all of the dataframes, checking missing data, cleaning the data as defined and removing unnecessary columns in the process. I continuously tested after coding and after the final dataset was clean it was stored into "twitter_archive_master.csv" file.