Wrangle Report

Project Intro

For this project, we were to use Twitter data from @dog_rates Twitter account otherwise known as WeRateDogs. WeRateDogs rates dog photo submissions and posts them for their followers to see. My job was to use the data to perform analysis and create interesting visualizations.

Gather Data

3 sources of data were used to collect the data needed for analysis. The first was in the form of a csv file (twitte-archive-enhanced.csv) that was provided. Data from the file was loaded into a Pandas dataframe called twitter_archive.

The second source of data was obtained from a cloud storage site using the Python response library. A connection was established to the URL and the data was written to a local tsv file called image-predictions.tsv.

The last set of data was to be pulled in directly from the Twitter API, but due to developer accounts needing to pay for the right, an alternate method was implemented. A file called tweet_json.txt was provided and specific columns from that json file were read into a dataframe called tweet_details.

Assess Data

The project required that we assess the datasets for at least 8 quality issues and 2 tidiness issues. There were far more issues that could be addressed, but these 10 would provide a good dataset to do analysis on.

Visual and programmatic assessments were performed on the 3 datasets and these are the issues I discovered and would clean in the Clean Data portion of the project.

Quality issues

twitter_archive dataframe

1. Timestamp column is not in timestamp format
2. Column floofer name should be floof not floofer, and all floofer values should be floof
3. Retweet messages should not be included, only original tweets. Columns observed: retweeted_status_id and in_reply_to_status_id
4. Columns contain missing values and are not needed for analysis. expanded_urls, in_reply_to_status_id, in_reply_to_user_id, retweeted_status_user_id, retweeted_status_id, retweeted_status_timestamp
5. Columns have inconsistent case, invalid names. Columns observed: name, doggo, floof, pupper, puppo. Also fix case p1, p2, p3 in image_predictions.

6. Values above and below 10 observed in rating_denominator column.

tweet_details dataframe

7. Rename 'id' column to tweet_id

image_predtictions dataframe

8. Non-Dog names in columns p1, p2, p3

Tidiness issues

1. Dog variables doggo, floofer (floof), pupper, and puppo are in individual columns in *twitter_archive* and should be categorical.
2. Columns from *image_predictions* and *tweet_details* should merge with *twitter_archive*.

Clean Data

The next stage was to clean the data by addressing each of the issues identified above. All 3 dataframes were copied into new dataframes for cleaning purposes. For each issue, a process of define, code test was performed. The define step commented on what cleaning process would take place. The code step is where the Python code was created and executed to do the cleaning. Finally, the test step was where the results of the cleaning process were validated, again using Python.

Store Data

Once all the steps had been performed, a final dataframe called *twitter_master* was created, and a copy of its contents were archived to a local csv file called twitter_archive_master.csv.

After storage of the cleaned data, analysis steps were performed to glean insights from the data, and to create visualizations.