Wrangle and Analyze Data

Wrangle Report

The dataset which we have wrangled is WeRateDog. It is a twitter account that rate dogs with tweets.

The WeRateDogs Twitter project goals included:

- Wrangling the twitter data through the following processes:
  - Gathering data
  - Assessing data
  - Cleaning data
- Storing, analyzing, and visualizing your wrangled data
- Reporting on the data wrangling efforts and data analyses and visualizations

Gathering Data:

Gather data from twitter archive master csv., image-predictions from given URL.

We have done the analysis through twitter API and with help of Tweepy Library and all the data we found is from tweet and retweet.

Assessing Data:

Once the data was gathered, I began to assess the data on both quality and tidiness issues.

Quality Issues
archive:

- Completeness:
  - missing data in the following columns: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls
  - tweet_id is an int
- Validity:
  - dog names: some dogs have 'None' as a name, or 'a', or 'an.'

  - this dataset includes retweets, which means there is duplicated data retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp
- Accuracy:
  - timestamp is an object
  - retweeted_status_timestamp is also an object
  - rating_numerator goes up to 1000+
- Consistency:
  - rating_denominator should be a standard 10, but there are a multitude of other values
  - the source column still has the HTML tags

images:

- Validity:
  -
  - Some dogs have invalid names which are having two alphabet.
  - p1, p2 and p3 columns have invalid data

- Consistency:
  - In three columns aren't consistent when it comes to capitalization: sometimes the dog breed listed is all lowercase, sometimes it is written in Sentence Case.
  - in p1, p2 and p3 columns there is an underscore for multi-word dog breeds

twitter_counts_df:

- Completeness:
  - missing some data

Tidiness Issues

1. I have merged 4 columns into 1 column.

2.. All tables should be part of one dataset

Cleaning Data:

After the assessment, I cleaned the data through the following means:

Define, Code and Test

1. Merge the clean versions of archive [df_tarchive], images [df_images], and tweets [df_tweet] dataframes Correct the dog types
2. Create one column for the various dog types: doggo, floofer, pupper, puppo,
3. Remove columns no longer needed: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp
4. Delete retweets
5. Remove columns no longer needed
6. Change tweet_id from an integer to a string
7. Change the timestamp to correct datetime format
8. Correct naming issues
9. Standardize dog ratings