WeRateDogs Wrangle Report

By Ekwunife Mmesoma

Objective: Wrangle the twitter data (@WeRateDogs) to create interesting and trustworthy analyses and visualizations.

Data Gathering

The datasets used for this analysis were gathered from three different sources.

- The twitter_archive_enhanced.csv file was manually downloaded and uploaded and read in using read_csv().

- The image_predictions.tsv file was downloaded from this URL using the requests library.

- The tweet_json.txt was gotten by querying the Twitter API for additional data using the python tweepy library.

Assessing Data

The datasets were visually and programmatically assessed using excel and some panda methods such as head(), tail(), info(), describe(), unique(), value_counts(), etc.

Quality issues

The following issues were observed after assessing the dataset.

Twitter archive table

- Only original tweets are needed therefore no retweets.

- Original tweets with images (image_urls).

- Descriptive column name (dog_name) instead of name.

- "O" instead of "O'Malley" as dog name.

- Texts not readable in source columns

- Erroneous datatypes (tweet id, timestamp, source, dog stage, rating numerator, and rating denominator)

- & instead of & in the text column.

- Some texts contain floof and still have 'None' as their dog stage values.

- The dog_name column contains inconsistencies like a, the, quite, an, such, not, very, mad etc.

- Tweet with incorrect rating of 24/7 which means another thing.

Image predictions table

- Erroneous datatypes (tweet_id)

Tweet count table

- tweet_id instead of id in order to merge with the other tables.

Tidiness issues

- One variable in four columns in twitter archive table (dog_stages)

- Irrelevant columns (retweeted status id, in reply to status id, retweeted status timestamp, retweeted status user id and in reply to user id

- Merge the Image predictions, tweet count and twitter archive tables to form a single table.

Data Cleaning

Copies of the original datasets were first made before cleaning the data

Twitter archive table

Define:

- Extracting out the original tweets by removing rows where the retweet features have values.

- Extracting only tweets with jpg_url.
- Change the column name 'name' to 'dog_name'.
- Replace "O" with "O'Malley" in the dog_name column.
- Replacing texts in source column with readable texts.
- Changing the data types of these features to their correct data types using astype() and to_datetime().
- Replacing values & with & in the text column.
- Replace floof in text to floofer as values for the dog stage feature.
- Extracting dog names from texts for values like a, the, quite, etc. in the dog_name column using regex function.
- Deleting the tweet with the 24/7 rating because 24/7 stands for something else in the text.

Image Predictions Table

Define:

- Changing the data types of the tweet_id feature to its correct data type using astype().

Tweet Count Table

Define:

- Changing the id column name to tweet_id in tweet count table using rename method

Tidiness

Define**:**

- Creating a new feature dog stage by extracting values from doggo, floofer, pupper and puppo features.
- Removing some irrelevant columns using the drop method.
- Joining the three tables to form a single table using the merge method.

Code

The astype(), merge(), replace(), str.extract(), copy(), drop(), rename(), re.findall(), etc. were used in cleaning the dataset.

Test

The isnull().sum(), value_counts(), info(), head(), notna().sum(), isnull(), str.contains(), etc. were used to test the codes.