Wrangle_report
September 24, 2023
Marcus Thompson

The complete data wrangling procedure for WeRateDogs involved gathering diverse data from multiple sources, evaluating the data for both quality and tidiness issues, and then proceeding to clean and integrate the datasets.

I began the data preparation process by gathering the necessary data for analysis. Initially, I imported the Twitter archive data, which was provided through the Udacity course. We could imagine that this data (in .csv) format had been provided by a Twitter user and shared with us through email. This data was then structured into a DataFrame, called 'df_archive'. Next, I utilized the requests library to fetch image prediction data, which I then organized into another DataFrame called 'df_image'. For managing the JSON data used to aggregate retweet and favorite counts, I read the file line by line and created a final DataFrame for this task, called 'df_json'.

Once the DataFrames were created, I then made working copies, df_archive_clean, df_image_clean, and df_json_clean. I performed my cleansing operations on these three DataFrames.

I began to visually assess the first DataFrame, 'df_archive_clean'. I used pandas methods such as 'df.shape', 'df.head()', 'df.value_counts()', and 'df.info()' to get a sense of what the data held. Then, I continued to use programmatic methods, such as 'df.duplicated()' and 'df.isnull()' to see what issues needed to be addressed.

Issues with df_archive_clean:

- Retweets needed to be removed.
- Several columns would not be needed for my analysis, and therefore should be removed.
- Some dog names are 'non-names', such as 'a', 'an', 'get', etc. These 'non-names' are not capitalized, and real names are capitalized.
- The 'tweet_id' column needed to be converted to the data type 'object' (string).
- The 'timestamp column needed to have the final five characters stripped from each value (+0000), and then needed to be converted to the 'datetime' data type.

I then used the same assessment techniques and applied them to the other DataFrames.

Issues with df_image_clean:

- Some rows existed where the values for 'p1_dog', 'p2_dog', and 'p3_dog' were false, suggesting that the photo associated with that row did not contain an image of a dog. These rows needed to be removed.

- I only needed to keep the dog breed value that was predicted with the highest confidence level. I then placed these values in a new column called ‘dog_breed’.
- Next, I needed to drop unnecessary columns from df_image_clean.
- The dog breed values were not displayed in a consistent, standardized format. Some were capitalized, others not. I made them all lowercase with an underscore between words if necessary (such as ‘labrador_retriever’).
- I then removed duplicates in the ‘jpg_url’ column, as they are likely to be retweets.

Issues with df_json_clean:

- The ‘id’ column should be renamed ‘tweet_id’
- the new ‘tweet_id’ column needed to be converted to ‘object’ data type from ‘int’.

Tidiness issues:

- in df_archive_clean, the four columns ‘doggo’, ‘floofer’, ‘puppo’ and ‘pupper’ were combined into one column, called ‘dog_stage’.
- The three DataFrames were then combined into one master DataFrame, ‘df_merged’, which I then used to perform my analysis.

After completing these steps, I then performed an analysis and searched for insights that could be drawn from the data.