Wrangle Report

Wrangling is simply gathering, assessing, cleaning and sometimes storing of data.

Gathering

For this project, 3 different datasets were gathered.

The first dataset was twitter_archive_enhanced.csv. The dataset is a comma separated value file and it was manually downloaded. This was read into a dataframe named tweet archive.

The second was image_predictions.tsv. The dataset is a tab separated value file and it was downloaded using code. This was also read into a dataframe named image predictions.

The third was gathering data from Twitter API. I was provided with two options of doing this. I went with the second option where the tweet-json.txt file for the data was already provided because Twitter up to this point has not approved my request for an elevated access on their developer's platform.

The tweet-json.txt text file was loaded and I extracted the data I needed for each tweet data into a dataframe named tweet json.

Assessing

Quality and Tidiness issues were checked for in the 3 dataframe above using visual and programmatic approach.

For the quality issues, missing data, duplicates and inaccurate were looked for while for the tidiness issues, I checked to ensure that every row in the dataframe is an observation, every column is a variable or feature and that every dataframe (observational unit) formed a table.

8 quality issues and 2 tidiness issues were identified and outlined below:

Quality Issues:

1. missing values in in_reply_to_status_id, in_reply_to_user_id, retweetwed_status_id, retweeted_status_user_id and retweeted_status_timestamp columns on visual check in the tweet archive dataframe.

2. timestamp column is of string data type in the tweet archive dataframe.
3. duplicates in the jpg_url column in the image predictions dataframe.
4. the number of rows in the tweet archive dataframe is not consistent with that of the image predictions dataframe.
5. name of the dogs in the p1, p2 and p3 columns of the image predictions dataframe are not consistent.
6. The number of rows in the tweet json dataframe is not consistent with that of the image predictions dataframe.
7. doubt over the accuracy of the rating_numerator column values in the tweet archive dataframe.
8. doubt over the accuracy of the rating_denominator column values in the tweet archive dataframe

Tidiness Issues:

1. doggo, floofer, pupper, puppo columns in the tweet archive dataframe should be in one column (dog_stage) rather than individual columns.
2. tweet json dataframe should be part of the tweet archive dataframe.

Cleaning

For each of the issue identified in the quality and tidiness category, the issue was cleaned using the rule of defining the issue, writing code to clean the issue and writing code to ensure that the issue has been resolved.

Columns in the tweet archive dataframe with large missing values were dropped, timestamp column in the tweet archive dataframe was converted to datetime and I ensured that the number of rows in the tweet archive dataframe was consistent with that of the image predictions dataframe.

Also, I ensured consistency in the name of dogs in the p1, p2 and p3 columns of the image predictions dataframe and remove rows whose jpg_url column are duplicates.

Furthermore, I checked for the accuracy of the rating_numerator and rating_denominator values columns of the tweet archive dataframe and ensured that the number of rows in the tweet json dataframe was consistent with that of image predictions dataframe.

Finally, for the tidiness issues were resolved as outlined.

Storing

The tweet archive dataframe was stored as a csv file named ‘twitter_archive_master.csv’ and the image predictions was stored as ‘image_predictions_master.csv’.

The wrangling effort then set the tone for the insights and visualization that was carried out.