Wrangling Efforts Report

Gathering the Data

This part was straight forward as the data were provided into three main datasets.

- First dataset for the archive of twitter was provided as a csv file. Thus, the importing was easy and straight forward via Pandas.
- Second dataset for image predictions was provided online. Thus, importing this dataset required using GET request.
- The third dataset was a bit challenging as it required creating a twitter developer account to get the access tokens. it was rejected. However, Udacity provided me with an alternative way to access tweets data.

Assessing the Data

In order to assess the data, two methods were used: the visual and programmatic. For instance, with the visual assessment, some issues were detected like having extreme values that needed to be cleaned not to affect the other data. In addition, to assess the data efficiently and spot issues to be resolved later at the cleaning stage, issues were categorized into Quality issues and Tidiness issues. These two categories were implemented on each one of the three datasets (Twitter_archive, image_predictions, and tweet_status) individually and the following organized report was noted.

Quality issues

Twitter_archive

- Too many values and might not be needed in this analysis.
- Delete the cells with numerator exceeding 20.
- For the column “rating_numirator” there are zero values.
- For the column “rating_numirator” there are outlier values.
- For the column “rating_denumirator” there is a zero value.
- For the column “rating_denumirator” there are extreme values.
- Some columns like (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id) are missing.
- Typo mistakes in the nomirator and the denominator.

image_predictions

- There are unneeded columns
- Cleaning irrelevant rows like “Egyptian_cat”.
- Cleaning irrelevant values in p1 like “school_bus, pillow, cartoon”.

- 473 rows (frequency less than 5) because they are less frequant and most likely not dogs.

tweet_status

- unneeded columns

Tidiness Issues

Twitter_archive

- unnessesarry columns
- Replacing 'None' with np.nan to indicate the missing values
- contain one column for dog classification instead of several columns

image_predictions

- not descriptive names (p1, p2 and p3).

tweet_status

- more convinient naming needed(id to be tweet_id) to ease merging
- tweet_status.created_at is the same as the column tw_archive.timestamp

Cleaning the Data

Using the assessment report, it was easy to **define** the issue to be fixed, **code** the needed code to get the job done, and **test** to check if the cleaning works as planned.

It is important to create a copy from the original datasets to apply all the cleaning on them.

For example, we need to exclude all the rows that has retweets as instructed in the analysis requirements. Thus, the columns that is related to the retweets were used to remove these rows.

Finally, it is important to gather all the cleaned datasets together in order to perform the visualizations and analysis on this master dataset.