Nishant Jha's

WRANGLE REPORT

1. GATHERING DATA

We use the following three pieces of data in a Jupyter Notebook titled `wrangle_act.ipynb`:

- The WeRateDogs Twitter archive file, which was manually downloaded. The file format was found to be `CSV`. The file name was found to be: `twitter_archive_enhanced.csv`
- The tweet image predictions file, named as `image_predictions.tsv`, which is in the format `tsv`. The file was hosted Udacity's servers and has been downloaded programmatically using the Requests library from the following URL, https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv.
- Using the IDs of the tweets, given in the Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called `tweet_json.txt` file. Then read this `.txt` file line by line into a pandas DataFrame with tweet ID, retweet count, and favorite count.

2. ASSESSING DATA

- For df_archive:
    - Some of the `expanded_urls` are missing & somewhere double links are present.
    - Keep original ratings (no retweets) that have images.
    - Delete columns that won't be used for analysis.
    - Erroneous datatypes (`doggo`, `floofer`, `pupper`, and `puppo` columns).
    - The `timestamp` has to be converted into DateTime format.
    - The `name` column is inconsistent (lower case and upper case).
    - The numerators that are with the decimal in the `text` don't match with `rating_numerator` column.
    - Drop rows where the `rating_denominator` is greater than the rating_numerator.
    - Drop rows with 'none' & 'a' present in `name` column.

- For df_prediction:
    - Drop 66 `jpg_url` that are duplicates.
    - Inconsistent writing for `p1`, `p2` & `p3`.
    - In 324 rows, no dogs are recognized.
- For df_json:
    - Change `id` to `tweet_id` for continuity of ID column among all three DataFrames.

- Globally:
    - Change `tweet_id` to string for all DataFrames.
    - All DataFrames should be part of one DataFrame & in-turn one Dataset.

3. CLEANING DATA

Making copies of all DataFrames is the first step.

- For df_archive:
    - Suppress double links and create a URL when missing.
    - retweets by filtering the NaN of retweeted_status_user_id.
    - Delete columns that are not needed.
    - Melt the doggo, floofer, pupper, and puppo columns to dogs and dogs_stage column.
    - Convert timestamp to DateTime dtype.
    - Change all names to lower case.
    - Replace numerators manually where the numerator is in decimals.
    - Select the rows only where the numerator is greater than or equal to the denominator.
    - Select the rows only where name isn't 'none' or 'a'.

- For df_prediction:
    - Drop rows with duplicated jpg_url.
    - Converting p1, p2 & p3 to lower case.
    - Create a new column with bool value for the dog (True) or not a dog (False).

- For df_json:
    - Using the 'rename' function to rename the column.

- Globally:
    - Using astype function to convert tweet_id into a string.
    - Merge df_archive & df_json by tweet_id with df_prediction.

4. STORING DATA

The master DataFrame is them stored in the CSV dataset named `twitter_archive_master.csv`.