Introduction

Using Python and its libraries, data was gathered from a variety of sources and in a variety of formats. Then, its quality and tidiness were assessed. Finally, data was cleaned to prepare it for analysis. The main dataset is the tweet archive of Twitter user @dog_rates, also known as 'WeRateDogs.' WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog; it has over 4 million followers and has received international media coverage.

Data gathering

For this project, three main pieces of data were collected. They were all downloaded and stored in the same directory as the project's Jupyter notebook. They are described below.

WeRateDogs Twitter archive

This is a CSV file that was manually downloaded from Udacity's website and is considered to be a file on hand. It contains basic tweet data for all 5000+ of WeRateDogs's tweets, but not everything. More specifically, it contains each tweet's text, rating, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and puppo). Of the 5000+ tweets, this file only contains tweets with ratings (a total of 2356 tweets).

Tweet image predictions

This is a TSV file and is hosted on Udacity's servers. It was downloaded programmatically using the requests library. This file contains what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network that can classify breeds of dogs. It contains a table full of image predictions (the top three only) alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction.

Twitter API file

This is a TXT file and was manually downloaded from Udacity's website. Usually, this data is obtained using Twitter's API; however, to make sure that the project could be completed with the exact required data, Udacity's provided JSON file was used instead. More specifically, the retweet count and favourite count for each tweet will be extracted from this file.

Summary of gathered data

By the end of the gathering stage there were 3 different dataframes downloaded:

1. *t_arc* -> the "twitter-archive-enhanced.csv" dataset containing information on basic tweet data
2. *i_pred* -> the "image-predictions.tsv" dataset containing the breed of dog prediction according to a neural network
3. *t_api* -> the twitter api dataset containing additional information about retweet and favourite count

Data Assessment

The goal of the project was to assess the datasets visually and programmatically for quality and tidiness issues detecting and documenting at least eight (8) quality issues and two (2) tidiness issues. For better clarity, each dataset was assessed individually. Quality and tidiness issues were also documented for each dataset.

t_arc_dataset

For the Twitter archive dataset, the following **quality** issues were identified:

- t_arc dataset contains retweets, while only the original tweets should be considered for analysis. The retweets should be deleted.
- tweet_id is wrong data type. It is int64, and it should be string.
- timestamp is wrong data type. It is string (object), and it should be DateTime.
- rating_denominator contains one value of 0 and other inconsistent values with the text of the tweet. They are shown in the wrong_denominator table.
- tweet with tweet_id:835246439529840640 has wrong rating_numerator. It should be 13.
- There are non-dog names in name column such as 'a', 'such', 'the', 'just', 'getting' and etc. These should be changed to Null.
- Records with no dog names in name column should store Null values instead of the 'None' string.
- Records with no dog stages doggo, floofer, pupper and puppo should store Null values instead of the 'None' string.
- Some tweets mention 2 dog stages instead of 1. Despite some of the tweets mentioning two dogs, a single-stage should be assigned for ease of analysis

Additionally, the following **tidiness** issues were identified:

- While the dataset contains rating_denominator and rating_numeratorcolumns it does not actually contain a rating column. This column should be added.
- Dog 'stage' classification is 4 distinct columns: doggo, floofer, pupper or puppo. This should be only one column.

i_pred dataset

For the image predictions dataset, the following **quality** issues were identified:

- tweet_id is wrong data type. It is int64, and it should be string
- Formatting of the predicted dog breed in prediction columns (p1, p2, p3) varies. A single lowercase formatting should be used.
- Since some of the algorithm's prediction are not even a dog breed, only the breed with the highest confidence should be kept.
- Prediction column names are confusing. They should be renamed.
- Only 2075 tweets have predictions. This is less than the number of tweets in the Twitter archive dataset.

Additionally, the following **tidiness** issue was identified:

- This dataset should be merged with the main Twitter archive dataset.

t_api dataset

For the Twitter API dataset, the following **quality** issues were identified:

- tweet_id is wrong data type. It is int64, and it should be string.
- The dataset contains 2354 records compared to 2356 records in the archive.

Additionally, the following **tidiness** issue was identified:

- This dataset should be merged with the main Twitter archive dataset.

Data Cleaning

All the documented issues above, were cleaned. Once again, the three datasets were cleaned individually before merging them together.

t_arc_dataset

For the Twitter archive dataset, the following data cleaning operations were performed:

- The datatype of 'timestamp' column was converted into DateTime data type.
- The values of denominators for tweets in the wrong_denominator were manually changed table by checking the associated tweet text.
- The value of the rating numerator of tweet_id:835246439529840640 was manually changed to 13.
- A new column was created by dividing the rating_denominator column by rating_numerator column. Then those columns were dropped.
- All the lower case non-dog names were replaced with null values in the name column.
- For name and every dog stage columns the string 'None' was replaced with Numpy's NaN.
- A dog_stage for every record based on the contents of doggo, floofer, pupper and puppo columns was assigned.
- The clean dataframe only included records that do not have retweeted_status_id.

i_pred_dataset

For the image predictions dataset, the following data cleaning operations were performed:

- The formatting for all records in all three prediction columns was changed to lower case using .lower()
- A new column to store the dog breed that was predicted with the most confidence was created. If all 3 predictions were False, a Null value was stored.
- The p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3_conf, p3_dog columns were renamed

Note: all the documented issues for t_api dataset were handled when the datasets were merged

Merging the three datasets

- The differences in dataframe shapes were addressed by only using the left merger. This way, only the data for the tweets in the main dataframe (t_arc) was stored in the new dataframe.
- 'tweet_id' column was convereted into string using .astype()

Storing Clean DataFrame

The clean dataframe was stored as a separate CSV file for future use. The index was not carried over, and 'utf-8' encoding was used.