Project: WeRateDogs Twitter Data Wrangling and Analysis

Introduction:

The project's objective is to wrangle, analyze, and visualize the tweet archive of WeRateDogs using three different sources: twitter_archive_enhanced.csv, image_predictions.tsv, and tweet_json.txt.

Gather:

1. Twitter_archive_enhanced.csv was downloaded manually using Pandas.
2. A url was provided to download programmatically the file, Image_predictions.tsv, by using the Requests library.
3. Attempted to use the tweet IDs in the WeRateDogs Twitter archive to 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. However, I was unable to obtain a Twitter developer account. Therefore, I used the dataset that was provided by the project and read the file using json package.

Assess:

The datasets were assessed visually by calling up each dataframe. They were programmatically assessed using functions such as .info(), to look at what the data types are and how many rows there are, and .duplicated(), to see whether there is any duplicated data. After assessing the datasets both visually and programmatically, I found the following quality and tidiness issues:

Quality:

1. In the *archive* table:
   - There are 181 retweets and 78 replies
   - Erroneous data types(timestamp, retweet_status_timestamp, tweet_id)
   - Erroneous names in the name column
   - Ratings are extracted incorrectly
   - Source column contains extra characters that are not needed
2. In the *image* table:
   - tweet_id is int not string
   - some images are not dogs
   - column names are not clear (p1, p1_conf, p1_dog)
   - inconsistent displays for image prediction output
3. In the *tweets* table:
   - id is int not string

Clean:

The datasets were then copied to perform data cleaning.

Tidiness -

1. The dog stages were merged into one column as they are the variables for the different dog stages. The ones that had None as a value were converted into null values then dropped, since the only information needed is the actual dog stage the users picked.
2. The datasets were merged into one dataframe with tweet_id being the common denominator. With three datasets merged into one, it made analyzing the data easier.

Quality -

1. Since only original tweets with images are needed, retweets and replies were removed.
2. Some columns had erroneous data type. They were changed as follow:
    a. *Timestamp*, from object to datetime
    b. *Tweet_id*, from int to object
    c. *Dog_stages*, from object to category
3. The entries in *name* contained random words that were in lower case and ‘None’ as a value. They were converted to ‘nan’.
4. There were ratings that were extracted incorrectly. The correct ratings were extracted from *text*.
5. *Source* contained tags that were not needed to show what utility was used post the tweet. The tags were removed. The data type was converted to category after removing the tags.
6. In order to analyze what dog breeds were most commonly predicted, predictions that did not contain dogs were removed.
7. The column names for image predictions were not intuitive, in that they did not explain what the columns were. New names were created to better explain what those columns were.
8. The output from image prediction had inconsistent name displays, where some breeds started with an uppercase while others were all lowercase. All the entries were converted to lowercase.

Summary:
The three datasets were merged into one, as each contained vital information for data analysis. There were columns that had consistency issues, wrong data types, incorrect information extracted, or information that was not needed. After the data was cleaned, it was saved as twitter_archive_master.csv.