WeRateDogs - Data Wrangling 2.0.0 - Wrangle Report

In this project, our goal was to gather data from three different sources that were in three different file formats. The data that was collected during the gathering process came from the following:

- WeRateDogs Twitter archive – this data was provided to us by Udacity in a .CSV format and titled: `twitter_archive_enhanced.csv`
- Tweet Image predictions – this file was also provided to us by Udacity, but it was in a .TSV format and titled: `image_predictions.tsv`
- Twitter API additional data – this data was provided to us as JSON data housed in a .TXT format titled `tweet_json.txt` and was compiled by using the tweet IDs from the WeRateDogs Twitter archive data.

Once the data was collected from the three sources, the next step was to assign each dataset to a pandas DataFrame and start assessing the quality and tidiness issues with each dataset. During the assessment, 9 quality issues were identified as well as 2 tidiness issues:

Quality Issues:

1. Data type issue #1 - Integers and floats that should be strings/objects:
   - The tweet_id column in the predictions table should be a string instead of an integer.
   - The tweet_id column in the archive table should be a string instead of an integer.
   - The id column in the tweet_data table should be a string and not a float.
2. Data type issue #2 - Columns that contain timestamp data as string/objects:
   - The timestamp column in the archive table should be datetime instead of object.
3. There are missing values for expanded urls column.
4. We need to remove tweets that are replies.
5. We need to remove the entries that are retweets.
6. We need to rename columns to be more descriptive or to mirror the other datasets:
   - Rename columns p1, p2, and p3 in the predictions table to be more descriptive.
   - Rename id column in tweet_data table to tweet_id to match other datasets.
7. We need to extract source information from HTML link in source column.
8. There are entries with a rating denominator and rating numerator higher than 10.

Tidiness Issues:

1.The doggo, floofer, pupper, and puppo columns contain data that should be stored in one variable rather than four.
2.Join the 3 datasets as one dataset.

After compiling a list of quality and tidiness issues, the next step was to go through each issue we identified and address each issue by cleaning the data to complete the data wrangling process. First, we made copies of original DataFrames to start the cleaning process. Each issue was then addressed individually and cleaned. Once cleaned, the three datasets were then merged into one dataset and saved to a .CSV file - `twitter_archive_master.csv`. After saving the file, we then began our analysis and visualization process on the newly cleaned dataset.