Report on Data Wrangling act

This report will describe the whole data wrangling process conducted for this project, "WeRateDogs."

1. Data Gathering
2. Data Assessing
3. Data Cleaning

1. Data Gathering

1.1. Data for this project was gathered from three different sources. The first one (`twitter archive enhanced.csv`)was downloaded manually.The second dataset (`image_predictions.tsv`) was about the tweet image predictions downloaded programmatically and the third file contains retweet_counts and favourite_count given by Udacity was in JSON format.

2. Data Assessing

Data was assessed both visually and programmatically. After assessing the data, the following quality and tidiness issues were found.

2.1. Quality Issues

2.1.1. Timestamp in archived_data needs three separate columns for months, days, and year
2.1.2. Remove all the columns that include data associated with retweets
2.1.3. Remove the unnecessary parts from at the end of each row of the 'text' columns
2.1.4. It is not necessary to have separate columns of doggo,floofer, pupper,puppo
2.1.5. In the 'name' column of the archived_data table, some invalid names. I have removed those names and 'None' entries.
2.1.6. The denominator should not be more than 10, and Numerator should be <= 10.
2.1.7. It is pertinent to remove the HTML from the 'source' column in the archived_data table
2.1.8. In the image_data table, we don't need the 'img_num' column.
2.1.9. Remove duplicated rows from the jpg_url column in the 'image_data' table

2.2. Tidiness Issues:

2.2.1. twitter_id column in the json_data dataset is in object format; convert it to int
2.2.2. Merging all the tables to make our effort of storing the data frame easy
2.2.3. In the image_data data frame, columns: p1,p2, and p3 include the name of the dogs in lower and uppercase letters.
2.2.4. Renaming the twitter_id to tweet_id in the jason_data table to make make the merging swift

3. Data Cleaning:

The following steps were taken to address the quality and tidiness issues mentioned above:

3.1. Quality Issues:

3.1.1. Creating four separate columns for 'year', 'time','month','day' from the 'timestamp' column. Then drop the 'timestamp' column.

3.1.2. Using the drop method on the 'archived_data' table to delete all the following columns (in_reply_to_user_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id,retweeted_status_timestamp,expanded_urls)
3.1.3. The entire text column was not visible in the workspace, so it was necessary to make it visible using the set_options method using the str.split() method
3.1.4. A single column for ('doggo,' 'pupper','floofer','puppo') was created
3.1.5. using the min() and max() on these two columns (Denominator and Numerator) and then summing up the number where the above condition was not followed.
3.1.6. Use the str.replace() method on the 4 HTML to replace them
3.1.7. use the drop_duplicates() method to drop the duplicated rows from the 'jpg_url' column
3.1.8. Using the drop method to drop the img_num column from the table

3.2. Tidiness Issues:
3.2.1. Use the 'astype('int64')' method on 'twitter_id' column of 'json_data' table
3.2.2. Using the df.rename() function to rename twitter_id column to tweet_id
3.2.3. Selecting all the rows (p1_dog, p2_dog, and p3_dog) where their values were true
3.2.4. Using the df_rename function on these columns (p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3_conf, p3_dog) to have descriptive column names.

Endnote:

No data is clean. Every data analyst should give their effort to make data clean to make inferences or decisions based on data. To do that, data wrangling is a skill all data analysts should have. After these three steps, it is also essential to analyze and visualize data to communicate the insights to the stakeholders. Otherwise, the data will be meaningless and won't be used to get the maximum benefits it can offer.