Report for data wrangling process in Project 4 of Udacity’s Data Analyst Nanodegree

1. Data gathering

- Download the twitter_archive_enhanced.csv file from the classroom and open it in the Jupyter notebook

- Download the image_predictions.tsv using the requests library.

- For the Twitter API data, I use the provided data by Udacity, then paste it into tweet-json.text file. Then read the file line by line and create them into a list of dictionaries, and convert the dictionary list into DataFrame.

2. Data assessing and cleaning

- To begin, I check the general information of 3 tables: all columns, duplicated rows, statistics results,... Next, I identify 8 quality issues and 3 tidiness issues as below.

Quality issues	Solutions
1. Only keep original ratings (no retweets) that have images.	1. Delete retweets which appear NaN in retweeted_status_user_id
2. Drop some unneeded columns.	2. Drop unneeded columns for this project
3. Wrong datatype in some columns: tweet_id, timestamp, source.	3. Convert tweet_id to string in all 3 tables, convert 'timestamp' to datetime, convert 'source' to category
4. Wrong decimal identification in rating_numerator.	4. Correct values that show wrong decimal numbers
5. Wrong dog names like 'a', 'an'.	5. Change the wrong name to 'None'
6. Some records have multiple dog types.	6. Create 1 column to display the dog type, multiple types or none of 4 types

7. Column 'text' includes hyperlinks	7. Remove hyperlinks in all tweets
8. Some rows have missing images	8. Drop tweets with no images
Tidiness issues	
9. Twitter_archive table: 4 columns 'doggo', 'floofer', 'pupper', and 'puppo' should be merged into 1 column	9. Create 1 column to display the dog type, multiple types or none of 4 types
10. Image_predictions table: This table should be added to twitter_archive table, as they share the same observational unit	10. Move image_predictions table to twitter_archive table
11. Twitter_data table: This table should be added to twitter_archive table, as they share the same observational unit	11. Move twitter_data table to twitter_archive table

3. Data storing

- Save gathered, assessed, and cleaned master dataset to a CSV file named "twitter_archive_master.csv".