In this project, I went through the data wrangling process, which consist of three steps: Gathering, assessing, and cleaning. Three datasets were gathered through different sources. Then, the datasets were visually and programmatically assessed to find quality and tidiness issues. Last step was cleaning the datasets to prepare it for further analysis and visualization. My effort in each step of the wrangling process is briefly described below.

1. Gathering data:

The data required to be gathered for this project were 3 different datasets, which were gathered as follow:

- **twitter_archive_enhanced.csv:** The file was manually downloaded to my local drive via Udacity website. Then the file was simply imported into pandas DataFrame by using the read_csv function.
- **image_predictions.tsv:** The file is hosted on Udacity's server and I programmatically downloaded it to my local drive by using Requests library. The content was written as a file in my local drive, to be imported into pandas DataFrame.
- **tweet_json.txt:** Due to difficulties in getting approval for a Twitter developer account to actually gather data through Twitter API, I had to download the JSON file directly from Udacity website. I wrote a code to read the JSON file line by line, and save it into a list to be converted and imported into a pandas DataFrame.

Note: each dataset was imported into separate pandas DataFrames.

2. Assessing data:

- Visually: The three datasets were first visually assessed through the Jupyter Notebook by printing each pandas DataFrame.
- Programmatically: The three datasets were programmatically assessed by using pandas' functions, and methods.

List of data quality/tidiness issues:

Quality:

Twitter Archive Table (twitter_archive_df):

- Some of the tweets are retweets, or replies
- Some tweets have a rating_denominator less than 10
- Some dogs have wrong names such as (a, an, the)
- Source column readability
- Timestamp column data type

Image Predictions Table (img_pred_df):

- Duplicated records (retweets)
- Some predictions start with capital letter and some small letter

Twitter API Table (fav_rt_df):

- Some tweets have low retweet/favorite counts (not original tweet)

Tidiness:

- The dog stage columns (doggo, floofer, pupper, and puppo)
- All tables can be merged into one table

3. Cleaning data:

I cleaned the datasets by following the programmatic data cleaning process: 1. Define 2. Code 3. Test. I went through each data issue found when assessed the datasets, and defined the action needed to fix the issue, and wrote code to do the needed cleaning. Lastly, I tested each code by using the assert statement and other method. Finally, the three DataFrames were merged into one DataFrame, and stored in a CSV file.