Data Gathering:

To conduct the analysis of the WeRateDogs Twitter Archive, I collected data from three distinct sources, each stored in separate files:

- Twitter Enhanced Archive: This file was manually downloaded from the Udacity servers. It provides comprehensive information about the WeRateDogs tweets.

- Image Predictions File: Programmatically downloaded from the Udacity servers, this file contains predictions generated by a neural network for each image associated with the WeRateDogs tweets.

- JSON Data for Each Tweet: By utilizing the Tweepy library to query the Twitter API, I downloaded the complete JSON data for each tweet. From this file, I extracted the favorite_count and retweet_count programmatically.

I loaded the three raw data files into separate tables named 'twitter_archive,' 'image_predictions,' and 'twitter' to facilitate subsequent analysis.

Quality Issues:

During the data wrangling process, I encountered several quality issues:

Twitter Enhanced Archive File

- There is inaccurate data in the columns 'doggo, 'floofer,' 'pupper,' and 'puppo'. The none entries will be replaced with null values.
- There is inaccurate data in the 'name' column that has entries of none and will be replaced with null values.
- The 'tweet_id' column datatype in all the files is incorrect and will be altered to string datatype.
- The 'timestamp' column datatype is incorrect and needs to be altered to a datetime datatype.
- Fixing incorrect data involving improper dog names and converting them to null values.
- Detecting and deleting duplicate data in the 'expanded_url' column and splitting data in the 'source' column.

Image Predictions File

- Detecting and deleting duplicate data in the 'jpg_url' column.
- Organizing the case formats of 'P’ columns and creating a 'breed' column.

Tidiness Issues:

In addition to the quality issues, I identified two tidiness issues that required attention:

- Converting 'doggo,' 'floofer,' 'pupper,' and 'puppo' columns into a single 'size' column, providing a consolidated view of the dog sizes.
- Retweets, replies, and the following columns will be dropped due to lack of value:
  'retweet_status_id,'
  'retweet_status_user_id,'
  'retweeted_status_timestamp,'
  'data_retweet,'
  'replies_data.'

By addressing these quality and tidiness issues, I aimed to ensure the data was well-structured, accurate, and ready for comprehensive analysis.