Data Gathering

The project begins with the collection of data from three distinct sources:

1. **Twitter Archive File**: A CSV file (**twitter_archive_enhanced.csv**) directly downloaded, containing basic tweet data.
2. **Image Predictions File**: Utilized the Requests library to download the tweet image predictions (**image_predictions.tsv**) hosted on Udacity's servers.
3. **Additional Tweet Data**: Gathered via the Tweepy library by querying the Twitter API for each tweet's JSON data, stored in a text file (**tweet_json.txt**).

Data Assessing

The assessment phase involved both visual and programmatic techniques to identify quality and tidiness issues in the datasets. The goal was to detect at least eight quality issues and two tidiness issues. Key considerations included focusing on original tweets with images, acknowledging the unique rating system of WeRateDogs where numerators can exceed denominators, and limiting the scope to tweets up to August 1st, 2017.

Data Cleaning

The cleaning process addressed the identified issues through a series of steps:

- Normalizing rating numerators and denominators.
- Extracting client information from the source.
- Handling missing values, particularly in the 'name' column by replacing NaNs with 'unknown'.
- Removing tweets beyond the specified date and those without images.
- Merging datasets to consolidate tweet information, image predictions, and additional data into a single, clean DataFrame.

Data Analysis and Visualization

The final phase involved analyzing the cleaned data to uncover insights. Key analyses included:

- Investigating the distribution of tweet sources to understand the most common platforms used for tweeting.
- Analyzing the relationship between tweet properties (like counts, retweet counts) and the ratings given.
- Visualizing the development of likes and retweets over time, highlighting user engagement trends.
- Examining the distribution of rating numerators to understand the commonality of different ratings.

Throughout the project, the focus was on applying best practices in data wrangling, including maintaining code efficiency, readability, and ensuring the reproducibility of the analysis. The project culminates in the creation of a **twitter_archive_master.csv** file, encapsulating the clean, merged dataset ready for analysis and visualization efforts, such as tracking the popularity of WeRateDogs over time and understanding user engagement through likes and retweets.

This documentation provides a 360-degree view of the data wrangling efforts undertaken in the "wrangle_act" project, demonstrating a structured approach to managing and analyzing data from multiple sources to derive meaningful insights.