Introduction:

This project revolves around the very important concept in the data analysis stage i.e. Data Wrangling (a stage that took most of the time in the data analysis process). The dataset used in this project is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. This report will briefly discuss the wrangling done to clean this dataset.

Data Wrangling:

Data Wrangling stage consists of three steps:

- Gathering
- Assessing
- Cleaning

Gathering:

This project requires to gather data from the following three different sources in three different file formats.

- **WeRateDogs Twitter archive:**

  This file was provided by Udacity and downloaded manually by the given link. This file contains the data in .csv format.

- **Tweet Image Predictions:**

  This file was hosted on Udacity's Server and downloaded programmatically using the Requests library in Python. This file contains the data in .tsv format.

- **Querying Twitter API:**

  To get each tweet's retweet count and favorite count, I have queried the Twitter API using Python's Tweepy library then stored each tweet's returned JSON as a new line in a .txt file. Next, I read the file named tweet_json.txt line by line to load data in panda's data frame.

Assessing:

After gathering the data from the above three resources, I have assessed them visually and programmatically for quality (content) and tidiness (structural) issues.

Visual assessments include displaying the data and scrolling through it to find any issue. head() and tail() method are good for visual assessments.

Programmatic assessments include methods such as shape, info(), .isna(), .duplicated() or any other programmatic way to disclose hidden issues within the data.

Cleaning:

The cleaning process is divided into three sub-steps.

- Define
- Code
- Test

I started the cleaning process by making copies of original data frames so that at any point if I make a mistake I could go back to the original data.

Most of the data quality issues are with the twitter archive data set. I dropped irrelevant columns, changed the data type of the column named timestamp, addressed invalid values in the name column, replaced all Nones by NaNs, and extracted correct denominator rating from the column text.

For the image predictions data frame, I have changed the non-descriptive column names to more descriptive and reasonable column names and then dropped columns that are not useful for analysis.

Finally, for tidiness issues, I have merged the columns doggo, floofer, pupper and puppo to a single column named dog_stage and also merged tweet json data frame with twitter archive data frame because tweet json contains only two columns (favorite_count, retweet_count) so, there was no need of keeping only two columns in the separate data frame.