Introduction:

This project sought insights into the tweet archive of Twitter user, user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. WeRateDogs has over 4 million followers and boasts of considerable international media coverage. The Twitter Archive contains over 5,000 tweet data which was made available for analysis by WeRateDogs via Udacity. This project was aimed at Wrangling

WeRateDogs Twitter data in combination with two other datasets to create interesting and trustworthy analyses and visualizations. Below are the steps taken in carrying out this project.

STEP 1: Data Gathering:

To analyze the dataset, I gathered the following datasets from multiple sources:

- WeRate Data: This dataset was read into Jupiter notebook using pandas via a URL provided by Udacity The werate_df data contains several interesting information such as tweet_id, timestamp, rating_numerator, rating_denominator, name. Also, there are some column headers which has unclear meaning such as doggo, floofer, pupper, puppo. However, this data doesn't seem to contain certain important information such as those who liked/favorited the tweets.
- Image Prediction Data: This file (image_predictions.tsv) hosted on Udacity's server was downloaded programmatically using the Python Requests library via the URL: [https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv]
- Twitter API Data: This contains key information such as retweet counts, favorite counts, etc., from the Twitter API Database that are missing in the original WeRate Data. The API will be queried using Python's Tweepy Library. Afterwards, I saved each tweet's entire set of JSON data in a file called tweet_json.txt file. Each tweet's JSON data was written to its own line, after which this .txt file was read line-by-line into a pandas dataframe with to provide this key retweet and favorite information for each specific tweet ID in the WeRate Data.

STEP 2: Assessing Data

After gathering the three data, the three datasets were painstakingly assessed using two major assessment methods – Visual and Programmatic Assessments(using code) to check for Data quality and Data Tidyness issues.

- Data Quality Issues: These refer to issues such as missing, duplicate, or incorrect data. A data with quality issues is popularly referred to as a 'dirty data.
- Data Tidyness Issues: These are primarily structural issues that impede effective cleaning analyzing, visualizing, or modeling of data.

The following Quality and Tidyness Issues were found after the datasets were visually and programmatically assessed:

Tidyness Issues:

1. Information about one type of observation unit (tweets) is spread across three different datasets. Given that these three datasets are part of the same observational unit, they need to be merged for effective analysis.
2. Dog stages in the WeRate data should be a single column rather than having the dog stage values (i.e., doggo, pupper, floofer and puppo) as columns.

Quality Issues:

1. p1, p2, and p3 are columns from the Image Prediction Data that contain variables that describe the three types of dog breeds for the dogs in the dataset. Dog breed should be a single column that matches each dog breed variable to individual dogs in the dataset.
2. **Too many missing values**: Some columns contain too many missing values e.g., retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, in_reply_to_status_id_str, in_reply_to_user_id_str, in_reply_to_screen_name, geo, coordinates, place, contributors, retweeted_status, quoted_status_id, etc. Many of these columns may have to be dropped for good analysis.
3. **Duplicate columns**: Some columns are duplicates. e.g., source_x and source_y, text and full_text, created_at and timestamp, etc.
4. **Incorrect data type**: Timestamp is wrongly assigned as a string datatype. It needs to be converted to Datetime.
5. **Poor column names**: Some columns need to be renamed to properly reflect their meanings e.g., source_basic needs to be renamed as source, timestamp to tweet_date, expanded_urls to tweet_urls, name to dog_name, id to tweet_id, etc.
6. **Duplicate rows**: There are cases of duplicate rows in the WeRate dataset. Duplicate rows may also occur after merging the three datasets. One of the duplicates needs to be dropped to prevent redundancy during analysis.
7. **Wrong values**: The name column contains some wrong dog name values such as ‘such’, ‘a’, ‘quite’, ‘one’, ‘my’, ‘his’, ‘not’, ‘getting’, ‘unacceptable’, ‘an’, ‘very’, ‘just’, ‘all’, ‘officially’, etc. These names need to be dropped.

8. Unique tweet sources need to be extracted from the source column.
9. tweet_id should be a string and not an integer.

STEP 3: Data Cleaning

The three datasets were properly cleaned based on the quality and tidyness issues earlier documented to improve the quality of analysis and visualization. Each cleaning process followed a defined outline:

i. **Define**: The specific issue was defined and properly stated.

ii. **Code**: The issue was solved using the right python codes.

iii. **Test**: A post-mortem sort of testing was carried out to check if the specific issue persists.

Conclusion

The data cleaning stage was hectic, challenging, and time-consuming. During the process of cleaning some issues, other issues arose which had to solved. However, it was an important learning process for me, and I am super proud to have partaken in the project.