**Wrangle Report:**

The dataset wrangled in this project is the the tweet archive of Twitter user [**@dog\_rates**](https://twitter.com/dog_rates), also known as [**WeRateDogs**](https://en.wikipedia.org/wiki/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 has received international media coverage.

The goals of this project were:

1. Gathering Data from three different sources
2. Assessing the Data both Visually and Programmatically
3. Cleaning the Data programmatically
4. Storing the Analyzed the data into a file in .csv format
5. Finally using the cleaned data to draw insights through visualization

**# Gathering Data**

The first step in the data wrangling process is Gathering data. Here I gathered data by three different means:

1. Manually uploading the file twitter\_archive\_enhanced.csv
2. Programmatically downloading the file image\_predictions.tsv from the Udacity Servers
3. Using the Tweepy library to gather tweet data in JSON format tweet\_json.txt, finally reading this text file data into a pandas dataframe

**# Assessing Data:**

The Quality Issues and Tidiness Issues identified through assessment are:

**Quality issues**

1. Data contains retweets --> retweeted\_status\_id for some rows is not NAN, hence duplicate data is present.
2. No of records in img\_predictions table does not match with df\_tweets table
3. No of records in df\_tweets table does not match with twitter\_archive table
4. No of records in img\_predictions table does not match with twitter\_archive table
5. Datatype issues in all three tables, the timestamps are not in format datetime
6. source column in twitter archive contains unncessary html tags.
7. Values in the name column - some lowercase, some start with uppercase, there are some invalid names like 'a' or are they valid names?
8. The denominator ratings in twitter\_archive table are >10 and the numerator ratings seem errenous
9. Multiple columns in image predictions table about the right breed name for the dog in the image

**Tidiness**

1. Removing columns -> empty values in retweet\_status\_id and retweet\_status\_userid hence, these columns to be dropped
2. doggo, pupper, puppo should all be merged into one single column 'growth\_stage'
3. Need to drop doggo, pupper, puppo, floofer columns
4. Need to drop the columns p1,p2,p3,p1\_dog,p2\_dog,p3\_dog,p1\_conf,p2\_conf,p3\_conf
5. Merge all three tables together to create a single dataset twitter\_archive table, image\_predictions tables and df\_tweets table

**# Cleaning Data**

**Firstly, a copy of all the datasets were made.**

**Addressing Quality Issues:**

1. Cleaned data in arc\_clean to only have original tweets - no retweets
2. (2,3,4) Cleaned to have consistent data across all three tables - arc\_new, df\_new, img\_pred\_new by removing mismatched records, keeping only those present in all three tables
3. Corrected datatype issues in all three tables, converting to datetime format
4. Removed the html tags from the source column
5. Replace all inaccurate names like ‘a’,’an’,’him’ etc., as None
6. Created a new column rating to correct errenous values in rating\_numerator and rating denominator
7. Created a new image predicted column based on the confidence levels of the images given

**Addressing Tidiness Issues**

1. Removed the unnecessary columns in arc\_clean table -- in\_reply\_to\_status\_id,in\_reply\_to\_user\_id,retweeted\_status\_id,retweeted\_status\_user\_id,retweeted\_status\_timestamp
2. Merged the columns, doggo, pupper, floofer and puppo into a single column which is growth\_stage
3. Dropped the extra columns - doggo, floofer, pupper, puppo
4. Dropped columns rating\_numerator and rating\_denominator as it is not necessary anymore
5. Removed the columns p1,p2,p3,p1\_conf,p2\_conf,p3\_conf,p1\_dog,p2\_dog,p3\_dog
6. Merged all datasets to create one single dataset

Finally, the merged dataset was written into a new .csv file called twitter\_master\_archive.csv

**# Analyzing & Visualizing Data**

In Analyzing and Visualizing data, the following observations were made:

1. The degrees of correlation between retweet\_count, followers\_count, favorite\_count
2. We can find the source from which users have tweeted the most
3. We can find what is the popular dog stage of our dog lovers
4. We can find which is the most popular breed of dog amongst our dog lovers
5. Using tweet\_create\_date we can seewhen or what time of day our dog lovers are most active at.