# Data Wrangling Report

# “WeRateDogs” tweets dataset

## Project objectives

The project’s main objectives were:

- Gathering data from csv, tsv, and Twitter API
- Analyze and clean
- Store and indicate some insights on the datasets

## Step 1: Data Gathering

In this phase, the three pieces of data were gathered and represented as pandas data frames:

- The WeRateDogs Twitter archive (file on hand, manual download of 'twitter-archive-enhanced.csv')
- The tweet image predictions ('image-predictions.tsv'). This file was downloaded programmatically using the Requests library from a provided URL.
- Each tweet's entire set of JSON data (with at minimum tweet ID, retweet count, favorite count, created date) in a file called 'tweet_json.txt' were stored using. Each tweet's JSON data was written to its own line.

## Step 2: Assessing Data and Cleaning Data

Data assessment was conducted against Quality and Tidiness issues. Standard python methods and functions were used to carry out visual and programmatical data assessment such as .head(), .value_counts(), .sample(), .describe(), .info() and etc. The following problems were identified, and steps were taken to solve these problems:

## Quality issues

| Dataset | Observation | Solution |
|---|---|---|
| **Twitter Archive** | Inaccurate data values in `name` column: a, an. the, mad etc.(are all in lower case) | Replace inaccurate names with none. The names starting with lower case are not the real names |
| | ‘tweet_id` column data type should be str not int | Change `tweet_id` datatype from int to object type |
| | ‘source` column data has unrequired HTML code | Remove the HTML from the `source` column using ‘Beautiful Soup’ function, leaving only the clear name of the source. |
| | `timestamp` datatype is str instead of datetime format | Change `timestamp` datatype from object to datatime type |
| | Some tweets are retweets or replies (which are not needed) | Drop these variables: `in_reply_to_status_id`, `in_reply_to_user_id`, `retweeted_status_id`, `retweeted_status_user_id`, and `retweeted_status_timestamp` in arc_clean dataframe as they are no needed |
| **Image Prediction** | `tweet_id` column data type should be str not int | Change `tweet_id` datatype from int to object type |
| | Some dog breed name starts with upper case. others lower case ( for `p1, p2, and p3` columns) | Convert all dog breeds name to start with and Upper case using string.capwords() function |
| | Underscore '_' present in some dog breed names instead of white space ' ' | Replace '_' with white space in `p1`, `p2`, `p3` column |
| **Extra Data** | `tweet_id` column datatype is int. it should be str | Change `tweet_id` datatype from int to object type |
| | `create_date` column datatype is incorrect | Change `create_date` datatype from object to datatime type |

## Tidiness issues

| Dataset | Observation | Solution |
|---|---|---|

| | | |
|---|---|---|
| **Twitter Archive** | Stages of dog category have seprate columns : `doggo`, `floofer`, `pupper`, `puppo`. | Convert [`doggo`, `flooter`, `pupper`, `puppo`] columns into one column called "dog_stage", then drop the four columns. |
| **Extra Data** | `retweet_count`, and `favorite_count` column is separated from the main dataset | Merge `retweet_count` and `favorite_count` to the main dataframe (`arch_clean`) and call it `tweet_master` |

## Step 3: Storing Data and Visualization

All datasets cleaned were saved to separate csv files and the main dataset was saved to the ‘twitter_archive_master.csv’ file as instructed.

EDA was performed on the master dataset where few insights were looked into.

List of Insights worked on:

- Which source has the highest tweet
- Relationship between retweet_count and favorite_count
- What dog stage has the highest retweet count and favorite count