## **Reporting: wragle\_report**

**Introduction:**

In this project, I performed data wrangling on WeRateDogs Twitter data. My goal was to perform data wrangling which involves the process of gathering, assessing, and cleaning data in order to prepare it for analysis & visualization.

**Body:**

*Gathering Data:*

I gathered the "twitter\_archive\_enhanced.csv" file by directly downloading it from Udacity. Then, I used the request library to download the "image\_predictions.tsv" file from the provided HTTP link. I also used the request library to download the "tweet-json.txt" file and extracted the required data into a dataframe. I merged all three dataframes into the "final\_df" dataframe and saved it as the "final\_tweet\_data.csv" file for convenience, this solved a tidiness issue.

*Assessing Data:*

I read the "final\_tweet\_data.csv" file into the "df" dataframe for assessing the data. Before proceeding, I did a visual assessment in Google Sheets and identified several tidiness and quality issues. I also checked for null values, head, data types using the ".info()" method, duplicates, value counts and made a list of identified issues.

*Cleaning Data:*

I created a copy of the "df" dataframe to preserve the original data in ”df\_copy”. I cleaned the second tidiness issue by creating a new column called "dog\_stages" by combining the four dog stage columns and deleting the unnecessary columns.

Next, I addressed the quality issues. I addressed the issue of retweets by removing all related columns, this solved the 1st quality issue (QI). I solved QI 2 by using regular expressions to remove HTTPS links from the "text" column. QI 3 was resolved by removing the missing data from rows which were related to retweets indicated by “RT @” and records containing no images. QI 4 included converting the "tweet\_id" from int to object, "favorites" from float to int, "retweets" from float to int, and "img\_num" from float to int using the "astype()" method. I also converted the "timestamp" column from object to datetime using the "to\_datetime()" method. QI 5 was resolved by converting the "p1\_dog", "p2\_dog", and "p3\_dog" columns from object to bool. QI 6 involved changing the "None" values in the dog stage columns to "pd.NA" values. QI 7 involved changing names in the "name" column that were not proper names to "pd.NA" values. QI 8 involved changing the "None" values in the "name" column to "pd.NA" values. QI 9 involved replacing underscores with spaces in the "p1", "p2", and "p3" columns. QI 10 involved capitalizing the names in the "p1", "p2", and "p3" columns. Furthermore, QI 11 involved extracting text from the "source" column using regular expressions.

*Analysis & Visualisation:*

After completing the data wrangling process, I saved the cleaned data in the "twitter\_archive\_master.csv" file and performed further analysis and visualizations. By using “value\_counts()” method on different columns, I found 3 insights as follows -

* Most tweets were done on iPhone as you can see from the above bar plot.
* TOP DOG STAGES are pupper & doggo, while puppo & floofer are the least dog stage type dogs tweeted.
* The most common dog names were 'Charlie', 'Lucy', 'Cooper', 'Oliver'.

One visualization created was on the top dog stages using pie plot.

**Conclusion:**

The data wrangling process is an important step in preparing data for analysis, and by addressing the 3 tidiness and 8 quality issues in the data, I was able to ensure that the data was in a usable and trustworthy state for further analysis & visualization.