Wrangle Report

1. **Gather data**

- Query Twitter API to get retweet and favorite count of each tweet_id
- Download image prediction programatically
- ***Importing favorite count and retweet count*****:** I imported 'tweet_json.text' into workspace, extracting favorite count and retweet count by tweet_id with json
- ***Merge with the main data:*** I merged with 'twitter_archive_enhance.csv' file with key "tweet_id".

2. **Assessing & Cleaning data**

- **Tidiness:**
    - The data has all information in one table. However, based on the purpose of the data, I think it should be separated into two tables: tweet table with all the information of tweets and dog table which contain the information of the dog and its rating
    - "dog" breed is currently in 3 different columns → It should be gather in one column
- **Quality:**
    - Some columns doesn't have correct data type → Actions: change data type
    - favorite_count and retweet_count have missing variable → Fill NA with 0, change data type
    - dog_breed has many missing variable → Actions: I took advantage of dog breed prediction from image_prediction.csv file. I merged image_prediction file with 'dog' table. If the dog already has breed information, then I keep them as it is. If the breed is missing, then I take the first breed prediction 'p1' (because p1's prediction probability is highest among p1, p2 and p3) as the dog breed.