This report records the wrangling effort of this project. Data wrangling consists of three main steps: gathering, assessing, and cleaning.

I. **Gathering**:

There are three data sources for this project: one csv, one tsv and one plain text file. For csv and tsv files, we use pandas' read_csv function to read and load them into DataFrame. Afterwards, we parse the text file first into a list of python's dictionaries. By using this data structure, we transform this data source into a DataFrame easily. Now we have the three base DataFrames.

II. **Assessing**:

Each DataFrame is assessed using two approaches: visually and programmatically.

- twitter(from 'twitter-archive-enhanced.csv') DataFrame represents tweets. It has quality and tidiness issues.
- predict(from 'image-predictions.tsv') DataFrame represents tweet image predictions. This DataFrame is high quality and structural in general.
- count(from 'tweet-json.txt') DataFrame contains tweet's retweet count and favorite count. These count observations are not complete without the original tweet content. Thus it has a tidiness issue.

III. **Cleaning**:

Before we clean, we clone the original DataFrames in case we overwrite them accidently. The cloned DataFrames will be used during this cleaning process.

1. Tidiness
    a. Generalize doggo, floofer, pupper, puppo columns in twitter_clean into a 'dog_type' column. Drop the original 4 columns.

    b. Merge DataFrame count_clean into twitter_clean based on tweet_id and id.
2. Quality:
    a. Extract value between tags in twitter_clean's 'source' column. Now the source of each tweet is clearer to observe.
    b. Convert 'a', 'an', 'the' in 'name' of twitter_clean to None. Name of each tweet is not polluted by pronouns.
    c. Convert 'timestamp' and 'retweeted_status_timestamp' in twitter_clean from Object(string) to datetime. Correct datetime type can be used in advanced analysis(e.g. sorting by datetime).
    d. Convert 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id' in twitter from float to int. Fill NA/NaN values with 0. This corrects the missing value and data type problems.

After data wrangling, we refine the three DataFrame into two DataFrame ‘twitter_clean’ and ‘predict_clean’ to analyze. They are exported out as csv to backup.