Quality issues:

1. Missing values in dataset **“twitter_archive_enhanced.csv”** were handled by filling numeric columns with 0 and dropping the columns that have more than 90% of records with missing values.

2. Missing values in dataset **“tweets_json.txt”** were handled by filling numeric columns with 0 and dropping the columns that have more than 90% of records with missing values.

3. The rest of rows that have missing values were deleted.

4. Data types of **“created_at”** column and **“timestamp”** column in datasets **“tweets_json.txt”** and **“twitter_archive_enhanced.csv”** respectively were changed to datetime.

5. Data types of columns (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, etc.) were changed from float to int.

6. Data types of columns (possibly_sensitive, possibly_sensitive_appealable) in dataset **“tweets_json.txt”** were changed to boolean.

7. Data inconsistency were found in columns **“retweeted”** and **“favorited”** in dataset **“tweets_json.txt”**. When the value of column **“retweet_count”** is bigger than 0 in any record, the corresponding value of column **“retweeted”** holds False indicating that there is no tweets and the same goes for columns **“favorited”** and **“favorite_count”**. Hence, the issue was fixed by replacing False values with True when the count in columns **“retweet_count”** and **“favorite_count”** is bigger than 0.

8. Columns such as ("truncated", "is_quote_status", "possibly_sensitive", "possibly_sensitive_appealable") that have only one value (no unique values) were dropped.

9. In dataset **“image_predictions.tsv”**, some predictions are non-dogs. Hence, only records that have the value True of column **“p1_dog”** were kept for data analysis and the rest were removed.

Tidiness Issues:

1. In dataset **“twitter_archive_enhanced.csv”**, there are 4 columns (doggo, floofer, pupper, puppo) that represent only one variable which is the dog stage, so they were deleted and combined into one new column called **“stage”**.

2. When merging two datasets or more, all datasets must have a column with the same name and the same data type in order to be merged, so the column **“id”** in dataset **“tweets_json.txt”** was renamed to **“tweet_id”** to match the other datasets.

3. After merging all datasets, the column **“created_at”** was found to be duplicated with another column **“timestamp”**, so the column timestamp was deleted.

4. To explore some statistics such as the most common dog species per year or the most common dog stage per year, a new column called **“year”** was extracted from column **“created_at”**.