Wrangling Report

Wrangling Objectives:

1. To gather three datasets using different methods
2. To assess the dataset for quality and tidiness issues
3. To clean the datasets
4. To store the wrangled data as a single master dataset

Step 1: Gathering data

This step involves gathering the three different dataset from three different methods

Dataset	**Method of gathering**
Twitter_archived_enhanced.csv	Downloaded manually
Image_prediction.tsv	Downloaded programmatically using requests library and url provided
Tweet_json.txt	By querying the twitter archive using API and tweepy to retrieve retweet_count, favorite_count and tweet_id

Step 2: Reading the dataset into DataFrame

The table show the dataset and the dataframe which each dataset is read into

S/N	Dataset	DataFrame
1	Twitter_archive_enhanced.csv	Twitter_archive
2	Image_prediction.tsv	Image_prediction
3	Tweet_json.txt	Tweet_extra_data

Step 3 and 4: Assessment & Cleaning

Visual and Programmatic Assessment were carried out on each of the DataFrame and ***Quality*** and ***tidiness*** issues observed are recorded and the Cleaning steps taken for each assessment are describe in the table below

Quality Issues		
dataframe	**Assessment**	**Cleaning**
Twitter_archive	▪ Some tweets are retweets as indicated by RT @ starting the text	▪ Filter out only those records where text does not start with RT @
	▪ Some tweet are not original tweet, as shown by 'in_reply_to_status_id' column	▪ Filter out only those in which the in_reply_to_status_id is

	having original tweet_id in it	NaN(null)
	▪ Missing values on columns (retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp)	▪ Drop the columns retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp
	▪ Missing values on columns (in_reply_to_status_id, in_reply_to_user_id)	▪ Drop the columns in_reply_to_status_id, in_reply_to_user_id
	▪ Incorrect datatypes on columns (timestamp, retweeted_status_timestamp)	▪ convert timestamp to datetime datatype ▪ retweeted_status_timestamp has been drop, so no need to change the datatype
	▪ Inaccurate dog names (some names are: a, an, the)	▪ Drop the column "name", since it is not required for analyis
	▪ Inconsistent data in expanded_url columns(some url are from twitter, gofundme) and missing urls	▪ Drop the expanded_urls column
	▪ Inaccurate data, some rating_numerator and rating_denominator are greater than 14 and 10 respectively	▪ Since this a unique rating system for @We_rate_dogs, we will not make any changes
Image_prediction	▪ Non-descriptive column name (p1, p1_conf, p1_dog, p2, p2_conf, p2_dog)	▪ Rename the columns to a descriptive names
Tidiness Issues		
Twitter_archive	▪ There are four columns of dog stages(doggo, floofer, puppo, pupper) in twitter_archive table when it should be one variable named dog stage	1. First replace None in stage columns with empty string 2. Then combine stage columns 3. Format entries with multiple dog stages and empty string 4. Drop the four columns(doggo, floofer, pupper and puppo)
Tweet_extra_data	▪ The columns (retweet_count, favorite_count) are attributes of the twitter-archive table	▪ Join the retweet_count and favorite_count to the twitter_archive_clean table
	▪ There should be one table whereas there are three(3) tables	▪ Merge the image_prediction_clean table to the twitter_archive_clean table

Step 5: Result

The result of the cleaning step is one single table "**twitter_archive_clean**"

Step 6: Storing Data

After the cleaning process, the cleaned dataset was stored as "***twitter_archive_master.csv***"