a. Introduction

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog, which has more than 8M followers till the date, the aim of the project is to analyze the tweet archive of Twitter user @dog_rates, using python libraries, so to analyze the data as accurate as possible we need to have it first by gathering it then assessing and cleaning (Data Wrangling Process).

In this project we will go deep in each of gathering, assessing and cleaning steps.

b. Data Gathering

For the sake of this project we had there different sources in order to gather our data which are:

- twitter-archive-enhanced – csv file
- image-predictions – tsv file
- tweet_json.txt

twitter_archive_enhanced.csv :

it's a csv file that has many details for every single tweet , it consists of 2356 Rows and 17 columns , the columns are (tweet_id, in_reply_to_status_id, in_reply_to_user_id, timestamp, source, text, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls, rating_numerator, rating_denominator, name, doggo, floofer, pupper, puppo)

image-predictions.tsv:

This is the second source of data what breed of dog is present in each tweet according to a neural network and which consists of 2075 rows and 12 columns (tweet_id, jpg_url, img_num, p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3_conf, p3_dog)

tweet_json.txt:

It's a file that has been created using Twitter API & Python's Tweepy library to store each tweet's entire set of JSON data that has three columns (tweet_id, retweet_count, favorite_count).

c. Data Assessing

In order to assess the data I have used both visual assessment using (Excel), and programmatic assessment using pandas functions, which were divided into Tidiness issues and quality issues as below

Quality Issues:

1. timestamp dtype here is incorrect.
2. tweet_id is int should be string as no mathematical calculation will be made with it.
3. Columns like("in_reply_to_status_id","in_reply_to_user_id","retweeted_status_id","retweeted_status_user_id" and "retweeted_status_timestamp") as we have been advised to not consider retweets nor replies we will drop its columns & rows.

4. some dogs name are incorrect like (a,an,this and the) I found out that by scrolling in Excel using (filter) as it's hard to see them all here in jupyter notebook.
5. expanded_urls has missing values.
6. rating_numerator has some wrong values like min value (zero) and max value (1776).
7. rating_denominator has some wrong values like min (zero) and max value (170).
8. Source column has the opening & closing HTML tags which make it hard to read.
9. Dogs stages has to many none values.
10. Dog breed's names some are capitalized some are lower cases – inconsistency.

Tidiness Issues:

1. The most obvious one here is dog stages which are separated into four columns which should be under one column, every dog stage with its value.
2. join all dataframes under one dataframe instead of having three dataframes.

d. Data Cleaning

Quality Issues:

1- Change dtype to datetime dtype.
2- Change tweet_id to str dtypr.
3- Drop extra columns for that related to retweet and replies.
4- Replace wrong names by None then by np.nan
5- Drop missing expanded_urls rows.
6- Check the right value of the rating_numerator and replace it wherever possible otherwise drop the row with incorrect values.
7- Check the right value of the rating_denominator and replace it wherever possible otherwise drop the row with incorrect values especially zero.
8- Make the source column more readable.
9- Replace missing dog stages with NaN values wherever we can’t obtain the right stage.
10- Lowercase all dog breed to ensure data consistency.

Tidiness Issues:

1- Merge all dog stage under one column and drop the four stages columns.
2- Merge all three dataframes into one dataframe.