1. Gathering:

Here In this part mainly, the data frames ae prepared to be used in the following phase's assessment and cleaning. There are three files. Each file is downloaded differently as follows:

- twitter_archive_enhanced.csv is downloaded manually by clicking the following link: twitter_archive_enhanced.csv.

- image_predictions.tsv is downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv.

- Then read tweet_json.txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count. But, it not downloaded from twitter since I don't have an account developer. Also, I sent for them to create a developer account but unfortunately, I am still in the process. I include the twitter API code as required but I didn't use it. I just downloaded the file from your server and then I parsed it line by line to be read into data frame.

Suggestion: *Could you please in the beginning of the wrangling data course announce for twitter development account creation.*

2. Assessment

There are two types of assessment for each table:

- Visually where I visualize the data frame for each table by showing the data frame itself, or use the df.sample() or df.head(), or df.tail().

- Programmatically: where I used the methods to explore the data frame size, column names, column data type, the null values number for each column, duplicated values, values counts in each column. These methods are such as df.info(), df.dulicated(), df.value_counts(), df.column.isnull()….ect.

The results of this assessment are two types of issues for each table

- Qualities which may be related to datatypes for each column, expressive name for each column, correcting values in columns, solving the missing values problem…ect.

- Tidiness which is related to merging some columns or separating them, merging tables or separating them,.ect.

The issues which I worked on them as follows:

- *Teitter_archive_df (Quality)*
  - Incorrect data type in the tweet_id column.
  - Retweet columns are retweeted_status_user_id, retweeted_status_id, retweeted_status_timestamp, retweeted_status_user_id.
  - Mismatch between the tweets ids of image_predictions_df and teitterDf_copy.
  - Including HTML url formats in the source column.
  - Incorrect values in rating_denominator and rating_numerator: denominator has some values less than 10 such as 0 and very large such as 170. The numerator has some large values such as 1776 and very small values such as 0. I didn't work on this issue. Because of the time, but I left in the nootebook my trial on this issue. I tried to correct values from the tweet text, but unfortunately, this didn't improve anything. If I have a time, I would fill all the incorrect values with medians. I will consider them nominals.
  - Dropping the reply useless columns which are : in_reply_to_user_id, in_reply_to_status_id
  - Missing values in the in_reply_to_status_id, in_reply_to_user_id columns.
  - Incorrect data type in the timestamp Column.
  - Incorrect and None values in the column name such as a.
- *Teitter_archive_df (Tidiness)*
  - doggo, floofer, pupper and puppo are stages for the dog, so they can be represented in one column.
- *Image_predictions_df (Quality)*
  - Incorrect data type in the tweet_id column.
  - Inexpressive header names of p1, p2, p3, pf p1_conf, p2_conf, p3_conf.
- *Tweets_json_df (Qualitiness)*
  - Incorrect data tye in the tweet_id columns. It should be string not integer since there is no required calculation on this column.
- *Tweets_json_df (Tidiness)*
  - Merging the retweet_count, favorite_count to the teitter_archive_df since this information is related to the tweet itself.

3. Cleaning:

In this phase, all above issues are solved per table and issues type. The issue cleaning is divided into three parts as follows:

- Define: which includes defection for fixing the issue.
- Code: which includes the code to fix the issue.
- Test: which includes testing issue fixing way.