Gathering Data:

1. Starter of the gathering process by downloading the ‘twitter-archieve-enchanced-2.csv file. This process was pretty straight-forward
2. Then with the help of the URL provided and the help of the ‘request’ library, I loaded the ‘image-predictions.tsv’ file with the ‘with open’ command and later to the memory.
3. The main part was to get the ‘retweet_count’ and the ‘favorite_count’ data to the system. The following are the steps that I took:
    a. With the help of the ‘os’ and the ‘tweepy’ library, I setup an access to the Twitter data set via API key to extract data.
    b. I created an empty ‘tweet_json.txt’ file and then with the help of the ‘.get_status’ function, I downloaded each ‘twitter id’ page info into the file with the encoding ‘utf-8’. Each page was separated by a ‘\n’ i.e. a new line.
    c. I created an empty list ‘tweets_detail’ to start off. I then opened the above file and with the help of the ‘json’ library, stored each page info as a dictionary and later append each of those dictionaries into the empty ‘list’ by running the entire code in a loop.
    d. Next was creating an empty Data frame ‘retweet_favorite’. I wrote few functions that would extract the required data from the ‘tweets_detail’. I stored the new columns into the empty data frame with the required values. I used the ‘map’ function to iterate a defined function over the entire ‘list’.
    e. I even mentioned a different way of extracting this data, if we were not saving the extract data into a ‘.txt. file first.
4. Lastly ,I merged all the data into 1 Data set. I know that this would be the part of the Access and Cleaning stage, but I just wanted to deal with ONE data frame to start the next step.

Accessing :

Quality Issue:

- 'timestamp' needs to be inthe "datetime" format. Currently it is not.
- 'tweet_id' should be in the 'string' format than in the current 'integer' format.
- Some rating are wrong in the dataset. Need to look at the 'rating_numerator' and the 'rating_denominator' as there are very few ratings above 15 in the former.
- 'in_reply_to_status_id, in_reply_to_user_id, and retweeted_status_timestamp' have a lot of missing values.
- Some of the dog's name in the 'name' column are not Name. They need to be replaced with 'NaN'.
- There are three columns that provide predictions to the breed of the dog. Need to find a way to make all of these columns provide better quality data.

- The column 'timestamp' provides no particular information that can be used for analysis. We will extract few data from it in the cleaning stage
- There are retweets presentin the data set. We only need the original ratings.

Tidiness issue:

- There are 4 columns for the Dog Stages. This violates the rule of Tidy Data. So 'floofer, pupper, puppo, and doggo' need to be combined in to one column.
- Rating numerator and denominator columns should be just one column.
- Needed, but already done. Combining all the 3 datset into one. I have already done that.

Cleaning:

The next stage was to clean them. The following is the process of going about this stage:

1. I looked into the summary of the dataset to find which columns had the missing values and dropped them.
2. I made changes to few columns including the ‘tweet_id’ and the ‘timestamp’for the data type
3. I added few new columns such as the ‘month’ and the ‘year’ columns for future analysis.
4. I made sure that the names of the Breeds and that of the Dogs were all in the ‘Title’ format. I cleaned up all the records that had weird names for the Dogs to ‘None’.
5. I also filtered out the ‘denominator_rating’ and the ‘numerator_ratings’ that made no sense or were in the decimal number and later merge them into one column ‘overall_rating’ as a ratio of the two.
6. I merged the ‘floofer, pupper, puppo, and doggo’ columns into one as they were under violation of Tidyness Rule 1.
7. I also merged all the prediction columns into ONE by classifying them into ‘Dog, NotADOg, and MaybeDog’ and then eliminated the three prediction columns.
8. I removed any retweets from the dataset as they were not need.