Data Gathering: I gathered data from different sources.

1. `twitter-archive-enhenced.csv` was given with the project and was manually download. This is loaded into `df1` dataframe
2. Request library was use to download tweet image prediction(image_prediction.tsv). This is called `df2` dataframe
3. I had problem with verification from twitter API and had to use the provided `tweet_json.txt. This becomes `df3` dataframe

Accessing Data and Cleaning Data: I accessed the 3 dataframes visually and programmatically. I found out the following to be issues with the dataframes and they were resolved too:

1. Accessing `twitter archive` (df1), I found out that there are retweets and tweets without images. These were removed
2. On `twitter archive` (df1), I found out that the timestamp column datatype is object instead of datetime. This was also converted to datetime.
3. The columns `doggo`, `floofer`, `pupper` and `puppo` have values that are `None`. This makes the dataframe info when checked shpwed that those columns have required values. These were changed to `NaN` using Numpy.
4. In `twitter archive`, `in_reply_to_status_id`, `in_reply_to_user_id`, `retweeted_status_id`, `retweeted_status_user_id`, `retweeted_status_timestamp` contain very little and most cases not needed information for the analysis. Also, in only `id_str`, `favorite_count` and `retweet_count` are needed
5. The content of the source column in `twitter archive` has only four unique values. These contents are in URL-like for but not valid URL. The real unique values of the `source` column are [iPhone, vine, web and tweet-deck]. They were changed to something more readable and clearer.
6. The datatype of `tweet_id` in ` image_prediction`(df2) dataframe is integer but ought to be string. This was changed to string for easy match-up with IDs in other dataframes.
7. Looking at `p1`, `p2`, `p3` columns in `image-prediction`, I noticed some wrong vales like `Shopping cart, laptop and the rest. Using the values of `p1_dog`, `p2_dog` and `p3_dog` I got rid of names that are not dog breeds. When there are more than one dog breed vale present in a row, I use the highest of `p1_conf`, `p2_conf` and `p3_conf` to select the best represented dog
8. The datatype for `id_str` in `tweet_json`(df3) is integer but should be a String. It was converted to String using astype() function.
9. `twitter archive` has 4 columns that ought to be one column. They are `doggo`, `floofer`, `pupper` and `puppo`. They were collapsed to one column `age_grade`
10. Also `p1`, `p2` and `p3` were collapsed to one column `breed`