Gathering Data

1- I downloaded the “twitter_archive_enhanced.csv” file manually which was provided by Udacity, then I uploaded this file manually to my online workspace and I read this file and called it “df_1”
2- I downloaded programmatically “image_predictions.tsv” file, which was hosted on Udacity's servers, then I saved it programmatically in the main directory and I called it “df_2”
3- I was not able to get a developer account in spite of me sending them all the data they needed but sadly they refused my request to get a developer account, so I read and understood the provided python script which reads tweeter's API and saves it to “tweet_json.txt” file, I then copied this code and pasted it in my notebook.
4- I then read the .txt file line by line into a pandas DataFrame with tweet ID, retweet count, and favorite count as the selected values required to read and then I pasted all this data in a DataFrame called “df_3”

Assessing Data

- I then ran a lot of commands (such as: “.head()”, “.tail()”, “.duplicated()”, “.info()” …etc.) on each dataframe to assess them and find error in those tables and found the following ones

Quality issues

in df_1:

1- "expanded_urls" column has alot of missing values
2- "doggo", "floofer", "pupper", "puppo" columns have alot of missing values
3- "name" column has alot of missing values or wrong ones such as: "None" or "a" or "one" "Bo" or "an" or "by" or "my" or "the" or "just" while "Al" and "O" should be done manually
4- a lot of tweets with no images url
5- "retweeted_status_id", "retweeted_status_user_id", "in_reply_to_status_id", "in_reply_to_user_id", "retweeted_status_timestamp" columns are present on this data although they shouldn't and any row that contains data in those columns shouldn't be in the table as well

6- "tweet_id" should be string
7- "retweeted_status_timestamp" should be datetype not string
8- "source" written as HTML while it should contain the name of the device which was the source of the tweet only

in df_2:

1- "tweet_id" should be string

in df_3:

1- a lot of tweets with no images
2- "tweet_id" should be string

Tidiness issues

1- "doggo", "floofer", "pupper", "puppo" should be one column
2- As "df_3" table shows data which is very related to the data shown in "df_1" and "df_2" table, they can be merged

Cleaning Data

- Then I started cleaning each of the issues I have documented while assessing as follows:

1- I first copied the data that I should assess, I then replaced "None" values with "" values in the columns "doggo", "floofer", "pupper", "puppo"
2- I then merged the data which was in those 4 columns and droped those columns then I formatted the data properly
3- I then merged "df_1", "df_2" and "df_3" tabled in a new table called “merged_df1”
4- I then droped any row which contained any data in any one of those columns as the data shouldn’t contain any retweets or any replies as metioned in the project specifications "retweeted_status_id", "retweeted_status_user_id", "in_reply_to_status_id", "in_reply_to_user_id", "retweeted_status_timestamp"
5- Then I removed those columns
6- I then converted "tweet_id" column data type to string
7- Then converted " retweeted_status_timestamp " column data type to datetime

8- Then I corrected the wrong data in the “name” column by searching for the word coming after any of the following formulas "name is", "named ", "this is ", "meet " and "called " and put NaN in the rows of the tweets containing no name
9- I then removed any excess data in the “source” column and keept only the name of the source platform