Wrangling part of the project

Gathering :
the gathering part was divided into three sections :

1 - downloading manually the " twitter-archive-enhanced.csv " file

2 – downloading programaticlly the "image_predictions.tsv" file from udacity servers using "Request" library in python , and putting it's content in a file contain the same name, and create "image_predictions._df" from

3- in this part i created twitter account to use thier API , looped through tweets id to get tweets data in a form of JSON files ,saved it locally , then extracted retweet and favourite counts to create "addiontal_info_df" data frame that holds their values.

Assessing :

i used visual asessing to know :

1- tidiness issues of (doggo, floofer, pupper, puppo) in "twitter-archive-enhanced.csv" , prediction ,confidence and breed by 3 different algrothims in "image_predictions.tsv" which don't follow the first role of tidy data. final tidiness issue was in merging the "image_predictions._df" dataframe with "twitter-archive-enhanced.csv"

2- knowing that name, doggo, floofer, pupper, puppo columns have value 'None' instead of 'NaN'

3 - many name column's values have the value of 'a', 'the', or 'an' .

4 - 'source' column's data was in HTML tag

5- nomiantor ratings values were integers instead of floats , so the friction part was missed

and i used Programmtic assessment to know :

1- there were denominators ratings that not equal to 10

2 – by inspecting nominators values that have weird high or low value like 420, or 1, i found a funny tweet that give a rating to the singer 'snop dogg ', found two other tweet that were wrongly extracted , one tweet wasn't about rating a dog,found an outlier tweet

3- by using .info() method i found tweet_id, timestamp, source ,rating_nominator in "twitter-archive-enhanced.csv" have wrong data types

4- columns that contain retweet and replays data has values that need to be dropped

Cleaning :

first i cleaned the **tidiness** issues by :

1- merging “addiontal_info_df” and “image_predictions.tsv” using merge() method in a data frame called “twitter_all_data_df”

2- melting (doggo, floofer, pupper, puppo) columns’ values to a new column called “stage” and assigned its data type to “catrgory”

3 – in “image_prediction.csv” and by using wide_to_long() method i created “prediction_level” column that holds the prediction level of the algorthim (1,2, or 3), so the alogrthim columns changed to be just one for confidence, one for breed and one for prediction

4- i merged “twitter_all_data_df” with ”image_prediction_df” to perform integrity by include dog’s breed

and in **Quality** issues :

1- i used .astype() method to change data type of stage, source columns to ‘category’, “tweet_id” to string in “twitter_all_data_df” dataframe, and ‘tweet_id ‘, ‘img_num’, and ‘prediction_level, columns to string in ‘image_prediction.csv’

2- changed all ‘None’ value that exit in some columns to NaN

3- droping all retweets, replays, and original tweets that don’t have images, by filter dataframe with values in “retweeted_status_id” , “in_reply_to_status_id” , and ‘NaN’ value in ‘expanded_urls’ columns respecetivly and then dropping all columns headers related to retweets or replays by using .query() and .drop() methods

4- extracing informtaion from the HTML tag in source column by .str.extract() method

5- extracing the ratings with frictions part using .str.extract(), then inspecting nominators values less than denominator and fix if it was wrong , and finally normalize all denomiantors to 10 and adjust corespoding nominators

6- solving 27 wrong extraced name , thorugh str.extracted() method by grab the word after the word ‘named’ or ‘name is’ in texts

7- replace ‘a’ ,’the’ ,’an’ value in name column to NaN

8- after merging “twitter_all_data_df” with ”image_prediction_df” i dropped all values that contain false in “breed” column