data wrangling steps for analyzing Twitter dataset

Gathering :

First I download **twitter_archive_enhanced.csv directly from my computer by using to_csv function.**

Then I download image_predictions.tsv from a server to my computer programmatically by using request library and then save it in dataframe .

After this I face some problems with twitter api it take so long time for each request ,therefore
I decide to download it directly from tweet_json.txt line by line and then deal With json format by extracting retweet and faivorate and tweet id columns then save it in dataframe

Assessing :

For Assessing I identify more than 13 **qualty** ***issues and 2*** **tidness issues** ***by visually and programmatically as follow :***

qualty issues:

- 1- we have many missing values in multaple attrebuts in df_1 table
  - 2-in_reply_to_status_id 3-in_reply_to_user_id 4-retweeted_status_id 5-retweeted_status_user_id 6-retweeted_status_timestamp 6-expanded_urls
- 7- all id's in tables should be in string format
- 8- change from df_1 in timestamp attribute type
- 9- in df_1 the source attrbute we have to remove href= word to make it only contine the url
- 10- in table df_1 in rating_denominator columns there 20 rows exceeding number 10 rating
- 11- there are 281 records missing in table df_2
- 12- there are missing values named by none in df_1 in name attribute
- 13- in df_2 change true/false values to make it 1/0 in p1_dog,p2_dog,p3_dog
- 14-in df_1 with tweet id number '890729181411237888' there are two expanded_urls

tidness issues:

- 1- there are 4 columns (doggo,pupper,puppo,floofer) we can make it as one column
- 2- there are three tables all belonges to each others we need to merge it into one table

Cleaning :

After identify the issues I clean them all by redefine the problems and fix it programmatically and then test them all.