Wrangling Report:

I did a wrangling process on the twitter data. The data come from three sources.

- First file was given called twitter_archived_enhanced.csv. What I did was just loaded into pandas using pd.read_csv function.
- Second file is a tsv file, I need to download it using request module, and write into a new file call image-predictions.tsv file, and I will do the same for twitter_archived_enhanced.csv file.
- The last one is a bit tricky, I need to actually use twitter API and load into a json file and capture favorite_count and retweet_count attributes and form a dataframe called df_tweet.

I detect 10 issue total, incudes tidiness issues and qualities issues. I will list all the issue below and explain how I fixed those issues.

1. The first issue is that data type for tweet_id column is not correct, it should be string instead of integer, so I use some function to convert integer to string for all three dataframe.
2. The second issue I detect is, in the image dataframe, p1,p2,p3 dog breed name has first letter either in lower case and upper case, this is a consistent issue, so I fixed by convert all into lowercase letters.
3. The third issue I detect is that retweet and reply tweet should not be included in the archive dataframe while calculating ratings in the analysis stage, so I filtered them out, leave only the original tweets in the dataframe.

4. The fourth issue I detect is that timestamps column in the archive dataframe should be datetime type instead of panda series, so I converted the data type.
5. The fifth issue I detect is that I find out that some of the rating numerator and denominator are able to capture the rating with decimals, so I use regular expression to extract the str with ratings and convert to floats.

6. The sixth issue I detect is that there are some missing value from the expanded_url column, so I use fillna method to fill it with something indicate the URL is unknown.
7. The seventh issue I detect is in the name column in archive dataframe, some name are called ‘a’. This is obviously not a name. For this issue I am not able to fix it, because there could be so many names that start with ‘a’. I know this may not be an issue that intervene with my analysis, but I do want to point out this is a issue with the data.
8. The eighth issue I detect is that there are some tweet in the df_tweet dataframe has 0 favorite_count, this is definitely not right, I checked online for some of those tweet with their id, they do have favorite_count. I think this is a issue with the API, I can’t really fix it unless I do it manually. And in my analysis, it also confirms my theory.
9. The ninth issue I detect is in the archive dataframe, each dog stage form a column, this is a tidiness issue. What I did is I convert the input of each of these column into a binary input, and use if-else statement to construct a new column and drop the other four.

10. The tenth issue I detect is also a tidiness issue, I think favorite_count and retweet_count are all part of the twitter information, it should be included in the archive dataframe, so I merged into one dataframe.