Data Wrangling effort.

Firstly, the gathering part. I wasn't able to gather data from the 3 sources at once. I only gathered and wrangled data from the first two sources and postponed gathering data from the third source until my application was accepted from twitter. I did not face any problems while downloading or scraping ata from the url. However, when I was gathering data from the API the code would not run and it was unclear why. Turns out that I was writing binary 'wb' in the json file while I should be just writing 'w'.

In the assessment part, problems that I faced in terms of quality:

1. Null values in the image
2. Duplicated tweet id
3. Inaccurate names like (is and a)
4. Timestamp was object instead of datetime
5. Some image predictions weren't even of a dog
6. Some predictions were very weak
7. Tweet id were integers which made a problem when joining the two dfs
8. Column names in image prediction were so confusing

And in terms of tidiness:

1. There were more than one obervational unit in the same df(tweets and retweets)
2. Values were variables/column names(doggo, floofer,…)

For the image prediction file, I generally didn't ask myself what I want to remove as much as I asked myself what I wanted to keep. I just cleaned the data from weak and irrelevant predictions then took the more accurate results to define the dog breed in the tweet archive dataframe. As for the API data I did not find any problems that might disrupt my analysis.

After, cleaning the detected problems, I merged the three data frames in order to answer four questions:

1. What are the most popular dog names?
2. What are the most favored dogs?
3. What are the most featured dog breeds?
4. What's the relation between retweeting and favorites?