Report

Gather:

I gathered data programmatically from a website and twitter API and saved them under the folder files. I gather data from twitter by calling the API using the tweepy and stored the data in json in a text file.

Asses :

By using pandas I assessed all these files. During the assessment I find the following tidiness and Quality issues

Tidiness

- Join 'tweet_info' and 'image_predictions' Dataframes to 'twitter_archive' DataFrame
- Combine the four columns: doggo, floofer, pupper, puppo into one column 'dog_stage'

Quality

- Remove rows where there are no images
- Remove 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp'
- Change incorrect dog names.
- Missing values in 'name' and dog stages showing as 'None'
- Rating numerators with decimals not showing full float
- Tweet with more than one #/# sometimes have the first occurence erroneously used for the rating numerators and denominators
- Tweet ID# 810984652412424192 doesn't contain a rating
- Extra characters after '&'
- Change sources to more readable categories.
- Erroneous datatypes (timestamp, source, dog stages, tweet_id, in_reply_to_status_id, in_reply_to_user_id)

Clean:

By using the above observations I performed data cleaning. After performing cleaning I store this dataframe with twitter_archive_master.csv.

Finally , I did my analysis between tweet_count and favourite_count