Gather

1. The image predictions file was downloaded programmatically using the requests library.
2. The twitter archive dataframe was provided by the instructors.
3. Favorite count and retweet count was retrieved using the twitter API through the tweepy library provided by Python.

Assess

There were 10 quality and 3 tidiness issues identified while assessing the datasets. These issues were all in the twitter archive dataframe since the other two were customized datasets (image predictions and retweet count and favorite count).

Quality:

1. Missing name of dogs. Name of dogs are missing when the text does not start with "This is".
2. None values in names to be changed with NaN
3. Name of a dog 'O' is actually O'Malley
4. Stages of dogs are missing. Where there is no stage mentioned nothing could be done.
5. Extract only the url from source column
6. Convert timestamp to datetime format from string
7. Wrong ratings for some tweets. e.g denominator = 11 and 2
8. Decimal ratings are wrongly extracted
9. Enteries with retweet ids to be removed because they are retweets
10. Enteries with no expanded urls don't have images so they will be removed

Tidiness

- Dog stages column to be created instead of 4 separate columns
- These columns will be removed since this info is not necessary: in_reply_to_status_id, in_reply_to_user_id retweeted_status_id, retweeted_status_user_id, non-null object □ 1 master dataframe to be created to combine all these 3 tables

Clean

The cleaning steps include the define, code and test parts. Copies of all dataframes were made before performing the cleaning steps. The cleaning steps of define, code and test steps are well documented in the wrangle_act.ipynb file.

Store

The master dataset was made after merging the 3 data frames into 1 and stored in a .csv file. This file was then used to perform the analysis and 4 insights were produced that are documented in the act_report.pdf file.