Gathering Data

For the data collection, data was retrieved from 3 sources:

1. twitter_archive_enhanced.csv file
2. the images predicition file, which was downloaded automatically with the requests package and saved locally
3. Twitter's api to retrieve the favorite count and retweet count for every tweet

Assessing Data

tData was assessed visually and programatically using pandas. The following quality and tidiness issues were identified:

Data quality issues:

1. dataframe includes retweets
2. dataframe includes reply tweets
3. 680 tweets without names and 55 with an 'a' name or other incorrect 1 or 2 character names
4. 22 tweets have rating denominators higher than 10
5. a few rating numerators seem to be incorrect
6. incorrect datetime column data type
7. dataframe includes tweets with no images
8. Timestamp column includes (+0000) at the end of the string
9. Date column datatype should be datetime ## Data tidiness issues:
10. Dog stage are over 4 different columns instead of 1
11. timestamp column includes date and time and should be split over 2 columns
12. Drop columns we don't need (including retweets and replies)

Cleaning Data

Each quality and tidiness issue identified in the assessment phase was either cleaned then or cleaned during the cleaning phase.

Issues cleaned in the assessment phase:

1. dataframe includes retweets

all tweets including a retweet_id were dropped from the original dataframe seeing how we're only interested in original tweets

2. dataframe includes replies

all tweets including a reply_id were dropped from the original dataframe seeing how we're only interested in original tweets

All other issues were cleaned in the cleaning phase and followed the define, code, test flow