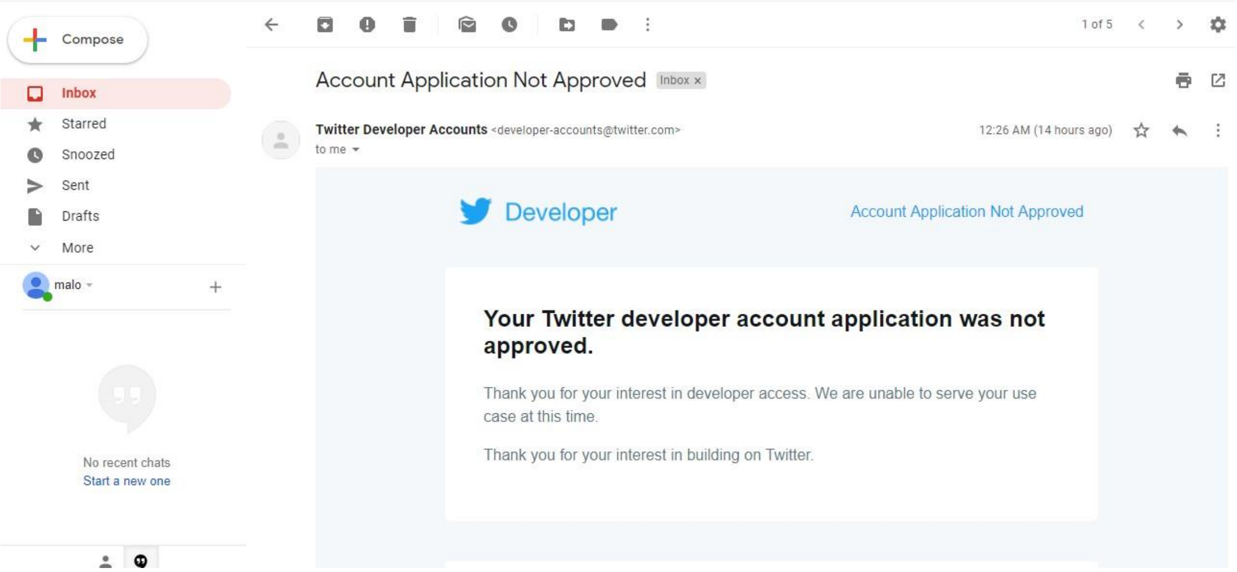
WeRateDogs provided their Twitter archive (which included tweets through August 1, 2017) of basic tweet data (tweet ID, timestamp, text, etc.) for use with this project. The "enhanced" csv file provided by Udacity (twitter_archive_enhanced.csv) also contains columns, which were extracted programatically: the rating numerator, rating denominator, dog's name, and dog stages (doggo, floofer, pupper, and puppo). These columns needed to be assessed and cleaned, as the extraction process was not perfect.

The provided Twitter archive lacked some useful information: retweet count and favorite count. I used tweet's entire set of JSON data in a file called tweet_json.txt. I then read the txt file line by line into a pandas DataFrame only including the desired variables; retweet count and favorite count.

Gathering

I applied for a twitter developer account, and my request was rejected, I don't know why by the way, but here is a snapshot to their email.



Account Application Not Approved

Twitter Developer Accounts <developer-accounts@twitter.com>
to me
12:26 AM (14 hours ago)

Developer
Account Application Not Approved

Your Twitter developer account application was not approved.

Thank you for your interest in developer access. We are unable to serve your use case at this time.

Thank you for your interest in building on Twitter.

So I decided to get the tweets from the tweet_json.txt file provided.

Assessing

Data Quality issues

Quality issue #1

Define: remove tweet that has been retweet as its not original.

Quality issue #2

Define: remove columns that are not needed for analysis

Quality issue #3

Define: Change timestamp from string to date time and make separate columns for date and time.

Quality issue #4

Define:p1,p2 and p3 have inconsisitent capital words.

Quality issue #5

Define: Drop duplicate jpg_url.

Quality issue #6

Define: p1,p2 and p3 have unnessary underscore instead of space

Quality issue #7

Define: rename id to tweet_id so can merge later

Quality issue #8

Define: change tweet_id from number to string.

Quality issue #9

Define:Remove incorrect dogs name

Tidiness Issues

Tidiness Issue #1

Define: combining dog stages to one column.

Tidiness Issue #2

Define: Newly created Date and time column needed to change from object(string) to date time format.

Tidiness Issue #4

Define: Merge df_clean, image_clean and twitter_clean dataframes

Cleaning

Used basic python function like duplicates, drop, sort, value count, and describe.