# Wrangle Report

In this project there were three datasets that needed to be gathered, combined, and cleaned. Of the three, two were provided and one had to be gathered using Twitter's API calls, creating tweet_json.txt. Not all tweets from twitter-archive-enhanced.csv were found through Twitter's API, and Twitter's API did not accurately provide favorite counts to all tweets. According to Twitter's API, about 174 tweets in our dataset had 0 favorite counts. However, I find 0 favorite counts to be highly unlikely since WeRateDogs is a very popular twitter account. I think not all favorite count information were collectible through API. For example, tweet 886053434075471873 had favorite_count = 0 in our dataframe, but had 3,364 likes when checked on twitter. Since 174 is less than 10% of 2349 of the tweets retrieved through Twitter API, I decided to drop these data points and focus on points that were more likely to be accurate.

The neural network produced image-predictions.tsv also had some accuracy problems. For example, tweet 666268910803644416 did not locate a dog in the image when there is a dog. However, the image predictions were fairly accurate in locating dogs and their species, so I decided to focus on the data points were the neural network was able to predict a dog and its species.

Denominator and numerator ratings were provided in twitter-archive-enhanced.csv. The rating data seems to be scraped from the tweet and is not always accurate. For example, if the tweet contained two ratios like "24/7… 13/10," the scraper picked 24/7 as the rating. Denominator ratings are usually multiples of 10 since tweeted images may contain multiple dogs. For my analysis, I analyzed only tweets with denominator of 10. I've also narrowed the range of numerators to be between 8 and 14. Most tweets below a rating of 8 did not pertain to dogs. Ratings above 14 were anomalies such as Satan's dog with 666/10 rating.