Wrangling Efforts.

Gather

The first step to wrangling is downloading all the necessary packages. We were then tasked with gather three datasets from three different locations and types. The first dataset was from WeRateDogs and had archived tweets in the form of csv. The second dataset was the image prediction dataset from Udacity's server This file had image prediction for each tweet. The third dataset was from Twitter directly in the form of JSON, which was accessed via Tweepy through Twitter's API.

Assess

For each of the datasets, we explore using jupyter notebook using functions in .info(), head(), and tail() to analyze what the steps we would need to go through in order to clean and consolidate the datasets into one clean dataframe.

Clean

We the applied the following cleaning steps:

1. df_twitter_arch_clean: (Data from WeRateDogs Archive)

1. Remove shortened URL
2. Remove rows from retweets or responses
3. Remove useless columns (dog stage)

df_breeds_pred_clean: (Image Prediction Dataframe from CNN)

1. Remove column "img_num"
2. Remove rows where: p1_dog, p2_dog, p3_dog are all False
3. Remove rows where: p2_dog and p3_dog are lower than 0.2

all dataframes:

1. Convert "tweet_id" in df_info_clean and df_breeds_clean as well as "id" in df_response_clean from integers to strings
2. For ID consistency: Removed ids from df_breeds_pred_clean need to be removed from the other dataframes
3. For ID consistency: Removed all ids from the failed_ids dataframe
4. Format the numerator and denominator using regular expressions

tidyness

1. Remove the pN_dog and pN_conf that we don't need, we will consolidate the probability column
2. Rename p1 to the predicted breed
3. Clean up the formatting of the predicted*_breed column, convert* to " " and A to a