Wrangle Report

The gather phase of the project was done in the following steps: I read the file " twitter-archive-enhanced.csv" into a dataframe called "t_archive". I programatically downloaded and extracted the file "image-predictions.tsv from the following link: 'https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv'. I read that file into a dataframe called "img_predict_df". I used the "tweepy" API to download a set of tweets matching the tweet ids found in "t_archive". I read that into a text file called "tweet_json.text", then from there back into a dataframe called "tweet_df". I then filtered out colums from "tweet_df" that were not needed.

The assess and clean phases of the project were done in the following steps: For "t_archive": I found 109 names in the "name" column that did not start with a capital letter and found them to not be names at all. I wrote a function called "not_name" to replace all names that start in lower case with "None". I found 181 retweets in this dataset and removed them by writing a query that removed all tweets that did not have a null in the "retweet_status_id" column. The datatype for the "timestamp" column was wrong and I switched it over to datetime. The "expanded_url" column had 59 instances of missing data, which i replaced with "None". There was leftover HTML language in the "source" column, so I wrote a function called "slice_source" to extract only the link itself. The "retweeted_status_id", "retweeted_status_user_id", and "retweeted_status_timestamp" columns were not needed so I removed them. There were 4 different columns which all described the type of dog, so I wrote a function called "stage_melt" to put all that information into one column called "dog_stage" For "img_predict_df": 66 of the images in the "jpg_url" column were duplicates, so I removed them. I changed the "jpg_url" comumn name to "picture_url" because the datatype for 2 of the pictures was not JPG. I then joined all three dataframes into a column called "clean_merge". The following issues were discovered: 7 rows were missing information for the "retweet_count", "favorite_count", and "lang" columns, so I repaced them with "None". Several of the tweets did not have images, so I wrote a query to filter out rows that had a null in the "jpg_url" column. During the merge, the datatype of the "img_num" column had been converted to float. I switched it back to "int" I also created a column to display the gender of the dog by creating a function called "gender" that searched the contents of the "text" column for keywords pertaining to gender..