wrangling efforts

gathering data

- twitter-archive-enhanced.csv was available for download
- image tsv used request librabry to get the file and saved it in image.tsv
- access the data using Twitter API these data will include retweets counts and favourites counts and save them in API_df

Assesing Archive df

Quality issues

- assessing the data there are a lot of quality issues:
  - name colume has some wrong entries
  - nmae column has missing values that already exist in the tweets text
  - timestamp column format is wrong
  - timestamp not date data type
  - dominator is recorded wrong in 3 observations
  - numerator is recorded wrong in 3 observations
  - drop records with no images
  - drop records that are retweets
  - drop records that is has no mattch in the API_df
  - convert the expanded_urls to links

tidiness issues

- the doggo floofer pupper puppo columns are values not variables
- the observations about the tweets are spread across 3 tables

Assessing image df

Quality issues

- assessing the data there are a lot of quality issues:
- tweet_id is wrong data type
- drop records that have a false prediction in the 3 predctions
- drop records that are retweets and not orginal tweets
- drop records that are not in the API_df

Assessing API df

Quality issues

- tweet_id is wrong data type

cleaning

Archive df

- extract the missing names from the tweet text using str.extract the mssing values format is "(named\s+[A-Z]+[a-z]+)" and "(name\s+is\s+[A-Z]+[a-z]+)"
- covert the wrong values in name column such as "a", "an", etc. to None using .replace()
- fix the format of the timestamp using string slicing
- convert timestamp to datetime datatype using pd.to_datetime
- droping the retweets observations using indexing
- droping retweeted and reply columns using drop using pd.drop()
- fix the wrong numerator values manually only found 3 values by indexing
- fix the wrong denominator values only found 3 values by indexing
- join the arc df and API_df inner join
- filter the arc df to tweets with images only
- combine the data from the doggo floofer pupper puppo columns in one column
- change tweet id data type using astype()

image df

- droping all observations that are not dogs in all 3 predictions using indexing
- filter the img df to have only orginal tweets no retweets
- creating new dataframe dog_breed having the dogs breedsfrom the 3 prediction in one column from the img_df
- adding the dog name and stage to this dataframe
- change tweet id data type using astype()

APi df

- change tweet id data type using astype()