Data Wrangling Steps

For this project, three different types of data were gathered. One was provided by Udacity ("twitter_archived"), one file was retrieved by using requests through Udacity's server ("image_predictions"), and the last file was retrieved by using twitter's APi and the tweepy library.

Upon opening the file on hand provided by Udacity, it was evident that the file was both messy and dirty. The file on hand called twitter_archived was assessed first. The data was visually inspected using the head, tail, and sample methods of pandas. Using some code such as the info and describe method it was evident that the twitter_archived data had many issues. The retweet_denominator had a value of 0, columns using timestamps were not datetime objects, some names were not actual names, it included retweets, the source column had unnecessary information, and some columns had values that had to be changed into Null Values, and some columns had to be joined to one column.

The file retrieved from Udacity's server also had some issues, however, not as much. Some columns had to be renamed in order to better describe the variables, some columns were also inconsistent with some values being capital and others being in lowercase.

The file retrieved from twitter's API did not have issues. Which was expected.