Data Wrangling WeRateDogs Twitter Page

Data wrangling was done by going through the three steps process of gathering, assessing and cleaning:

The data was gathered through three different sources by different methods: first I uploaded WeRateDogs Twitter archive read from an already downloaded csv file 'twitter-archive-enhanced', second I programmatically downloaded Tweets image predictions from Udacity server as 'image_predictions.tsv' and last, the retweet and favorite count queried using twitter API and stored as JSON file 'tweet_json.txt'. Each was stored in a separate pandas dataframe to be assessed.

When assessing data, I went by each dataframe first visually looking at a random sample to spot quality issues like null values or inconsistent values and identifying change to be made. Then I tried to explore each dataframe thoroughly by first checking the general info() for datatype issues, then I tried looking for null and duplicated entries. Then I assessed numerical columns looking for abnormalities like out of boundary for probabilities or 0 denominators for ratings and wrong ratings that don't match original tweet. After that, I tried looking for issues to clean in the data portion I intend on analyzing. Last I checked key point mentioned in project details. I found 8 issues that can be cleaned for better analysis. Then I went back to my samples for each dataframe to identify tidiness issues relating to dataframe structure like combining categorical columns into one.

Last, for the cleaning process I tried to go though it in order after making copies for all my dataframes. First starting with quality issues, I redefined each assessment observation to a clear instruction, handled it with code then tested to check if it's working. Through this process, I found one new assessment observation that I can fix in order to handle another, I added it the data assessment part then went on to handle it. Last, for tidiness, after fixing the issues identified, I combined all my dataframes in one for the analysis process.