Wrangle_Report

There were eight quality issues and three tidy issues that I fixed. The first quality issue that I fixed was getting rid of the retweets so there was only the original tweets left. I did that by keeping the rows that did not have a retweeted_status_id. The second quality issue was getting rid of columns that were not needed. This included; in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp. The third quality issue was getting rid of random words in place for the name. I did this by manually replacing certain strings with 'None'. The fourth quality issue was fixing the oddly large rating_denomintor values by deleting tweets without expanded urls. The fifth issue was replacing all of the 'None' on the dataframe with 'Nan' so I can combine some columns later. The sixth issue was taking all of the prediction confidence columns and changing them into percentages so they are more readable. The seventh issue was deleting the img_num column because no matter what number there was there was only one jpg_url. The last quality issue was fixing all of the tweets that had no favorites by finding the normal proportion of favorites to retweets and adjusting accordingly. The first issue of tidiness was merging all three dataframes together which included the twitter dataframe, image dataframe, and the dataframe created with the twitter api. These were all merged by tweet_id using inner merge. The second tidiness issue was taking the rating_numerator and rating_denominator columns and combining them to make one rating column. Some tweets had oddly large ratings that did not correspond with the others so I deleted those rows. The last tidiness issue was combining the four columns of "dogtionary" to make one column. Since a tweet can only have one value in the four columns I just filled the Nan value with each one.