Wrangling Report

The data wrangling efforts were done through three main parts which started with gathering, assessing, and cleaning the datasets. The gathering process started by reading the twitter_archive using pandas and using the requests library to read through the image_predictions. The other process involved using the twitter API to gather data from the WeRateDogs system. I did not manage to use this method as my twitter account was essential and not elevated. Therefore, I resulted in the second option which entailed reading through the json-text and used json.loads and readlines() to query and read through each line. The next part was the cleaning process which involved assessing the data. I performed both visual and programmatic assessment with the former involving reading passing commands such as .head() to examine the data visually. The latter involved slightly technical functions such as .info() & describe() to examine the data. The next step involved cleaning the data and from the earlier visual assessment, I denoted 8 quality and 2 tidiness issues which needed to be corrected. Some of the issues quality issues that were examined and corrected included dropping unnecessray columns, converting the date to the correct format, converting the tweet_id in tweet_data to a float, removing none dog rating tweets and filling in the missing names. It also involved renaming columns and correcting consistency issues. Therefore, the process ensured that the data abided with the four dimensions of data which involved completeness, validty, accuracy, and consistency. The second aspect of cleaning was addressing tidiness issues where the data structure was examined and some of the duplicated columns in the three datasets were removed and the three datasets were merged. Since the data had varying number of rows the pd.concat and reset_index methods were applied to merge the threee datasets. The final step involved visualizations and involved examining three aspects using the names of the rated dog and the relationship between their retweet_count and favorite_count. It also examined the name and the average image number that showed the accuracy of whether the image was a dog or not. Lastly the unique rating system was examined which helped analyze the average_rating numerator and the name of the dog which could be seen on average most dogs received a rating of between 10.5 and 11.