WeRateDogs Wrangle Report

After downloading the twitter_archive_enhanced.CSV file. I visualized the file to check the data in the columns. I found some columns not needed such as retweeted_status_id, retweeted_status_user_id also I found other rows that most of them havens values such as in_reply_to_status_id, in_reply_to_user_id.
Next, I imported the needed libraries that will be used in the code. After uploaded the previous file, I downloaded the image predication file programmatically. Then I created the JSON file. This was the gathering stage I applied to get the data.
In the assessing stage, I checked each dataset of the three files, I found some issues and I fixed it. Here what I have done:
For Quality:

- in df_twt_arch , as we don't need the retweets columns, I deleted them.
- in df_twt_arch , I removed '+0000' from the end of time_stamp for each row.
- in df_twt_arch , I changed time_stamp column datatype.
- in df_twt_arch , I updated all values of ratings_denominator and made it equal to 10 for the purpose of consistency.
- in df_twt_arch , I removed the text column that contains double rates for the dogs.
- in df_twt_arch , I capitalized all names.
- in df_twt_arch , I replaced the source column data with its own types to be more meaningful - instead of the urls and type, I just removed the url and kept the source type.
- in df_twt_arch , I removed the url of each tweet in the 'text' column.
- in df_img_pred , I deleted non-dogs rows where p1, p2, and p3 are false.
- in df_img_pred , I capitalized all names.
- in df_img_pred , I changed column names and make it more descriptive.
- in df_img_pred , I rounded the decimals in the Prediction Conf. columns.

For Tidiness:

- in df_twt_arch , I combined the dog stages (doggo, puppo, pupper, floofer) into one column named 'dog_stages'.
- in df_img_pred ,the img_num column has no use so I removed it.
- I combine the three datasets df_twt_arch, df_img_pred, and df_json datasets into one dataset using 'tweet_id'.

In the cleaning stage, I started with copying the three data frames into new three datasets. Then, I fixed every step I mentioned above. I also applied the define-code-test steps for each single point of the above points and some times I reassessed and retest for some points in order to get the best version of the data which will be used in the analysis and visualization stage. But first, I saved the cleaned data into a new file named (twitter_archive_master.CSV). I also saved the same set into a database named save_pandas.db in a table named the same name of the CSV file.
In the analysis and visualization stage, I started my analysis with the basic function (describe()) and wrote some insights that helped me to decide which plots I will apply. For example, I found a positive correlation between favorite count and the retweet count. I also plotted a bar chart for the tweet per day and tweets per month and year. Moreover, I plotted a horizontal bar char to the source types. I concluded that the more retweets the more favorites the tweet gets and the twitter users prefer to use iPhone to tweet. In addition, the predictions deceases as we go for the next prediction.