Wrangling Report

1- Data Gathering

I gathered the data from three different sources. First, I was provided with a file (twitter-archive-enhanced.csv) that I saved into the first dataframe (df_twitter_archive). I downloaded the second source of data programmatically using requests library, then I saved that data in dataframe and named it “df_dog_breeds”. Lastly, I used tweepy library to extract the rest of the data using tweet_ids that were provided in the “twitter-archive-enhanced.csv” file. I saved the data in dataframe and called it “df_tweets”.

Now, I have three dataframes that I need to assess and clean.

2- Assessing Data

I assessed the three dataframes I created using different pandas functions. I was able to identify 10 quality issues and 3 tidiness issues. I listed all the issues below.

Data Quality Issues

- tweet_id in the three dataframe is int.It shold be string
- timestamp in the df_arhcive is string
- missing values in mutiple columns (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted, status_timestamp)
- some values in the rating_denominator are greater than 10
- retweets in the df_archive table
- there are values in the name column extracted wrong from the tweet's text
- outliers in the rating_numerator column
- at tweet_id = 666287406224695296, rating_numerator and rating_denominator are extrated wrong. they should be 9 and 10 instead of 1 and 2
- decimal ratings aren't extracted properly (tweet_id = 786709082849828864)
- ratings should be float ranging 0 to 1.4 since we have decimal values

Data Tidiness Issues¶

- dog stages are in four columns
- tweets info separated in two tables (df_archive_tweets and df_tweets)
- dog breed should be a column in the combined table

3- Data Cleaning

I started cleaning the first dataframe (df_twitter_archive). I started by removing retweets data from the dataframe, then I dealt with missing values that was mainly concentrated in the first dataframe. during the assessment stage, I identified three tidiness issues. The first issue was four columns represented the dog stages, so I decided to create a new column for the dog stages instead of four. In the df_dog_breeds, I created a column that

represents each dog breed, Then I merged all the three dataframes into one table (df_all_tweets).
After dealing with missing values and the tidiness issues, I focused on the other quality issues I identified. Some of the dogs names in the df_twitter_archive table weren't extracted properly, so I replaced those values with nan. Also, ratings should be out of 10, but some values in the rating_denominator column were greater than 10, so I set the all values to 10. Also, the rating_numerator shouldn't exceed 14, however, some values were so big, so I removed the rows with rating_numerator greater than 14. Lastly, I created a column "rating" which is equal to rating_numerator divided by rating_denominator to hold the overall rating of a dog.