REPORT ON DATA WRANGLING PROCESS

After accessing the three (3) datasets programmatically and virtually, three (3) tidiness and eight(8) quality issues were detected. In other to clean up these issues, tidiness issues were treated first to get our complete dataset, followed by cleaning of quality issues. Although one quality issue was treated in the tidiness section (dropping doggo, floofer, pupper and puppo column). A copy was made for each dataset before wrangling.

Tidiness Issues and Cleaning Approach.

1. Timestamp(df) containing date and time.

   I used the split function to separate the data in this column in two (date and time), then dropped the timestamp column.

2. Doggo, flooded, pupper and puppy should form one column(stage_name)

   Using concatenation method these columns were put in one column called stage_name, followed by dropping doggo,floofer,pupper and puppo columns were stage_name was extracted.

3. The three datasets should be merged by tweet_id.

   Using merge function the three datasets were merged on tweet_id to give a new dataset df3.

Quality Issues and Cleaning Apprroach.

Based the project requirement, only the original tweets are needed not retweets, so I started by dropping rows that are retweeted data.

1. Since favorite_count is important in this dataset and my intended analysis, I started by dropping missing data in this column since there is no way to retrieve it.

2. Using the Boolean values in p1,p2 and p3_dog, I dropped rows that contain false twice since these are prediction columns thus a false occurring twice means it is not a dog.

3. Next I dropped columns containing too many missing values since they can't be retrieved.

4. I removed the underscore (_) in p1,p2, and p3 and change the first letter of each row to lowercase to maintain consistency.

5. The data needed from the source column is just 'Twitter for web or kind of device'. So I stripped every other character in that column leaving only 'Twitter for web or device'.

6. Empty rows,incorrect datatype in stage_name. All empty rows in stage_name were converted to NaN to make it easy to drop, the rows were dropped and the datatype converted to category using astype function

7. The data types of date, time, retweet_id, and favorite_count were changed to the correct types.

8. To ensure that the rating_denominator is 10 in all rows, I assigned 10 to the column so the rows will have a constant value of 10.