Wrangle report

After getting the data from their various sources, a visual inspection was done to assess the quality and tidiness of the data.

This was done using methods like .info(), .head(), .sample(), .describe().

For the twitter archive dataset. The **quality issues** include:

1. Wrong datatype in timestamp
2. Missing values in in_reply_to_status_id,in_reply_to_user_id, retweeted_status_id,retweeted_status_user_id ,retweeted_status_timestamp, expanded_urls.
3. name column has several entries of "None"
4. High number of "None" entries in doggo, floofer,pupper, puppo
5. Rating denominator has inconsistent values including a denominator of 0
6. Name column has several "None" entries and "a" entries
7. Inconsistent values in numerator

My approach to these quality issues

1. Timestamp has wrong datatype of object. This is a should be converted to datetime datatype. Using pandas method pd.to_datetime().This was converted to datetime datatype.
2. There are several missing values in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id ,retweeted_status_timestamp, expanded_urls. These columns were dropped since they there was no means of getting the missing values
3. The name column has several “None” entries. Probably these entries were used to replace the “None” values where used to fill in the missing values. This could have been extracted from the text column but due to inconsistency in the writing format for text, I decided to leave it out. For example: “

   "This is Phineas. He's a mystical boy. Only ever appears in the hole of a donut. 13/10 https://t.co/MgUWQ76dJU" compared to

   'Here we have a Japanese Irish Setter. Lost eye in Vietnam (?). Big fan of relaxing on stai r. 8/10 would pet https://t.co/BLDqew2Ijj'

4. High number of None entries in the doggo, floofer, pupper and puppo. It maybe that the e mpty rows were filled with None. Also, these four variables should be in one column. I e xtracted the keywords floofer, pupper, doggo and puppo from the text columns and set th em in one column called dog_status.
5. There were inconsistent values in the denominator and numerator. This maybe due to typ ographical error from users or maybe some people didn’t know the rating system.

For example, the denominator contained 0, 110,120, 130, 150, and 170. Most of the deno minator were 10 (over 90%). Theses inconsistent values were removed from the dataset.

6. Inconsistent values in the numerator were also dropped. Especially the very high values.

For **tidiness issues**. They include

1. timestamp contains both date (year, month and day) and time
2. doggo, pupper, floofer and pupper should be in one column
3. Favourite tweets, retweets and favourites should be part of twitter archive
4. twitter_id and timestamp are duplicated in tweet_selected.

My approach to tidiness issues

1. I extracted the year and month from the timestamp column. This was later used for analysis
2. I extracted the keyword doggo, pupper, floofer and pupper and set them in one column
3. Twitter archive and tweet selected dataset had duplicate values in twitter_id and timestamp. I merged both datasets together on twitter_id and timestamp. Merging was also required because favorites and retweets in the tweet selected dataset were supposed to be in one dataset.
4. I also merged the image data to the twitter archive. These two datasets were merged on the twitter_id. Some of the tweets didn't have images.