Data wrangling efforts

Firstly, I gather all I needed data from handed file, related URL and through Twitter APIs, then I used visual and programing methods to look details of the dataset. I assessed the data to identify any problems in the data's quality or structure. And the problems are listed as follows:

Quality: (Completeness, Validity, Accuracy, Consistency)

archive table

- 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id' should be integers instead of float
- 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id' should be changed to string because we don't use them for any numerical calculation
- 'retweeted_status_timestamp', 'timestamp' should be datetime instead of object (string)
- The numerator and denominator columns have invalid values, the numerator may have outliers and the denominator minimum shows 0 which is invalid
- "name" column has invalid name such as "None, a, an, the, this"
- In 'doggo', 'floofer', 'pupper', 'puppo' columns, null objects are non-null (None to NaN)
- Column 'timestamp' in df_archive is same as column 'date_time' in tweet
- Make 'source' column clearer to read in image_predictions table
- Missing values from images dataset (2075 rows instead of 2356 compared to df_archive)
- We only want original ratings (no retweets) that have images
- Some tweet_ids have the same jpg_url

Tidiness (Untidy data)

- Drop column related to retweets, and unnecessary columns in images table
- Various stages of dogs in columns, it is better to use rows archives dataset
- Add a gender column from the text columns in archives dataset

Then I cleaned the data by modifying, replacing or removing data to ensure that my dataset was of the highest quality and as well-structured as possible. And finally store the cleaned data to a csv file named "master.csv".