The Process of Data Wrangling

I loaded 'twitter-archive-enhanced.csv' file using pandas' read_csv function and 'image_predictions.tsv' file using the library 'requests'. I also used tweepy to get additional information about provided tweets, which are retweet counts and favorite counts.

After gathering data, I assessed the data and defined eight quality issues and two tidiness issues.

At first, I eliminated '+0000" from timestamp column. Then I dealt with data types. I changed datatype of tweet_id from int64 to object, and time stamp to datetime object. Also, I changed rating_numerator and rating_denominator columns from int64 to float.

I dropped the rows which have wrong rating_denominator. Rating denominator should be 10, and not other values. I also dropped unnecessary columns to make data more concise.

Many dog names had errors. I checked all the unique dog names and found out wrong names are all written in lower cases, such as 'a' or 'the'. It turns out that these errors are made because it was programmed to recognize a word followed by phrases like "This is" and "Here is" as its name, which is not always the case. I replaced these incorrect names with 'None'. But some of them didn't change, so I dropped the rows that I can't change the wrong names.

I also dropped many columns in image_predictions data. I made new column named 'breed' and made this column filled with first prediction of the pictures. However, if the first prediction isn't a dog, it is likely that that data is not adequate for WeRateDogs, as WeRateDogs rate only dogs and not others. So I dropped all the rows whose p1_dog value was False.

I also merged type information. I made new column named 'type' and if the row said in 'doggo' column 'doggo', I made its type 'doggo'. If the row said 'None' in all of four columns, which are 'doggo', 'floofer', 'pupper' and 'puppo', I made its type 'None'. For the rows that have more than two types, I tried to fill that type column with the value 'multiple', but there were no case like that.

There were redundant tag information in 'source' column. So I removed that information. Also I removed retweets using 'retweeted_status_id' column. I also checked rating_numerator outliers. Three of them were mistakes and the others were real outliers. I didn't change the real outliers and just correct the errors.

Finally, I merged all these three dataframes into one dataframe based on column 'tweet_id'.