Wrangle Act Report

First thing first, today I will talk briefly about my efforts with those three datasets from assessing them to clean them then finally to merging them into one master dataset to start in finding insights and visualizing them into charts.

Let's talk first about assessing the data. I began with looking to heads of the three datasets then, I looked on the data types and look for missing values using the info method to start in the visual assessment, I found that there is incorrect data types in all the columns that have an 'id' in twitter archive table like 'tweet id' or 'in reply to status id' etc. Then looked on the 'text' column to see if there is anything special, and I found a link for the tweet in the end of some values in that column, then I moved to the image prediction table which I find in that there are some rows that doesn't contain dogs at all by previewing the ' jpg url ' column, I found something like donuts and turtles and a moose, really I don't know who put an image for a moose in 'WeRateDogs' page. Then I found that the 'display text range' column must be integer column, but pandas recognize it as string (object) because it contains the lower bound, upper bound of the text count and the square brackets, and of course the 'doggo', 'floofer', 'pupper' and 'puppo' needed to be in one column called 'dog stage'. There are more problems in these datasets but let's move to the cleaning process

Now after I have done the assessing stage, I will talk about the cleaning process for these datasets, but first I had to make sure that I won't corrupt the original data so I made a copy for the three data sets to clean them, then I renamed the 'id str' column in tweet_jason table to 'tweet id' to make it easy to me to merge the three tables in the end of cleaning, let's talk about the first problem I solved. There was a URLs in the 'source' column of 'tweet_jason' table that indicate for the way that tweet was written with like iPhone or the webpage, etc, so I used replace function to remove the URLs and put in its place the word that the URL indicate for, then I fixed all the datatypes and checked after that using info method, then I filled the missing values in in 'retweeted status id' and 'retweeted status user id' with the string "No retweet", then I found that 'rating denominator' column isn't necessary since all it must contain is 10 values so I drop it and renamed the 'rating numerator' column to only 'rate', then to remove the rows that doesn't contain dogs from all the tables I merged the 'p1_dog', 'p2_dog' and 'p3_dog' from image prediction table with the two other tables so remove them from the two other tables by masking and this what I done using merge function There were more problems that I've cleaned up but let's move to the next stage of the data analysis process. The visualizations

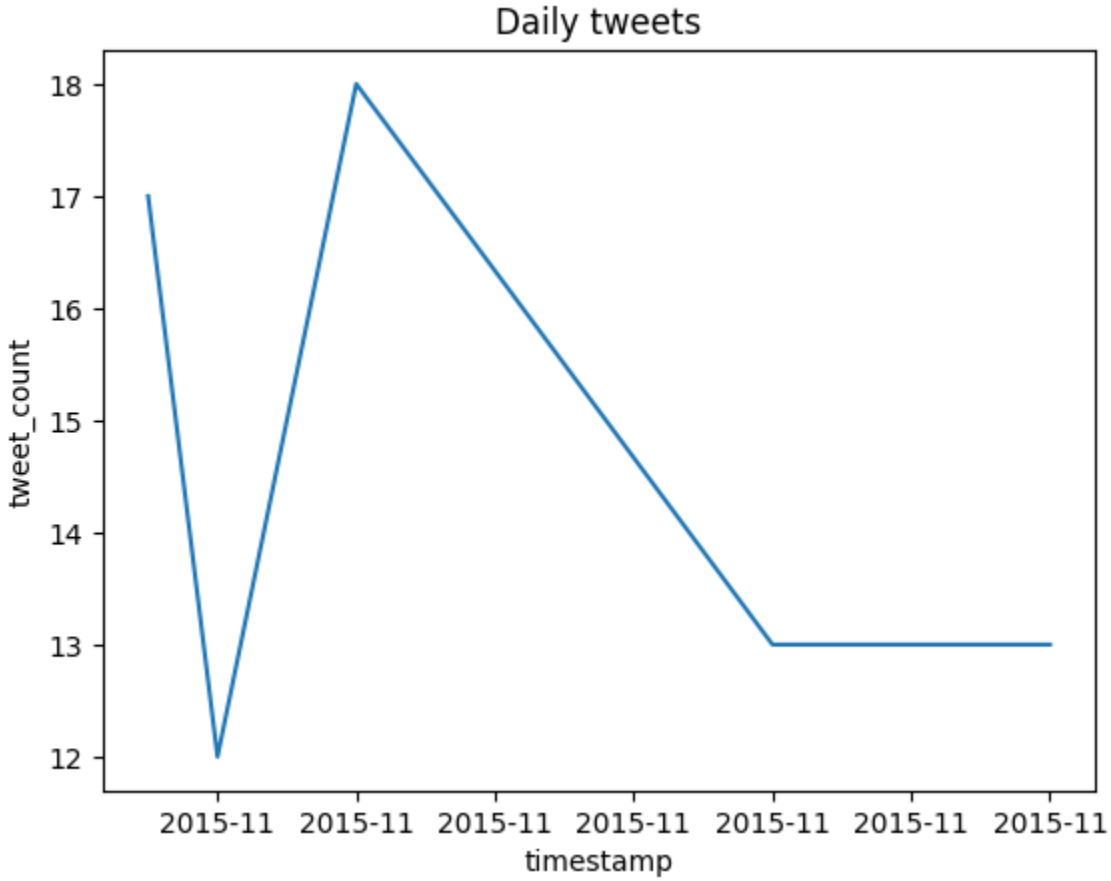
Wrangle Act Report

After I merged the only columns that I wanted from the three tables to the 'twitter archive master' table, I started in the Analyzing and visualizing the data, I will talk about the first two insight that I found from this dataset

The first insight was 'Which dog breed get the most rating ?' so first I used groupby method so i can get the average rating for each breed and picked the most 10 breeds then I ploted them as this bar chart

That shows us that the clumber dog is the breed which gets the highest ratings.

The second insight is 'What is the time that people shared their dog pictures the most ?'. Like the first insight I used groupby method to get how much the people post their images in each time on average, and created a dataframe that hold two columns, the timestamp and 'tweet count' on that time then I picked the highest 5 times and ploted them as this line chart shows



That shows us that people used to post their dog images the most the November 2015.