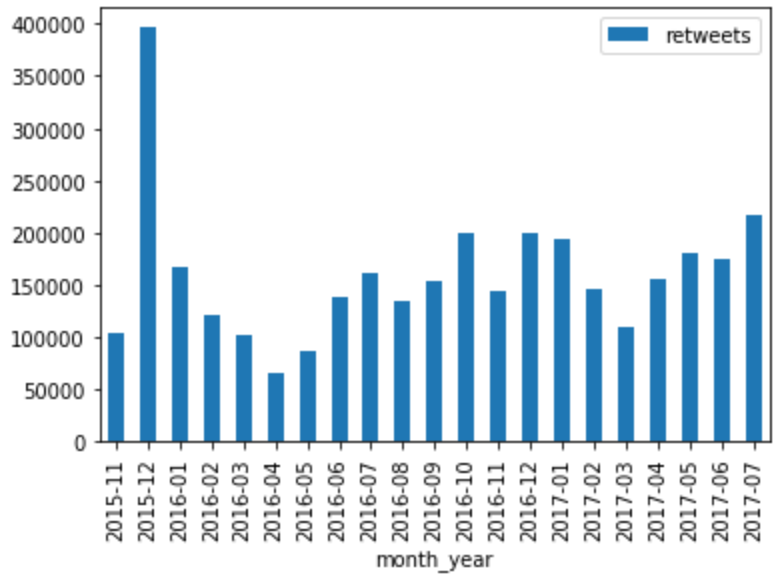
My analysis and visual insight:

After carefully cleaning the three dataframes and combining them into one, it was time to do some analysis. First, I wanted to see how accurate the p1_dog column was so I made a new dataframe with only two columns, one of them being p1_dog. I then did a group by and did a count to see how many were true and false. The breakdown was 340 False and 959 True, totalling 1299. To further explain, it was roughly 74% True and 26% False.

Next, I wanted to see the most common dog breeds that were predicted. The top 5 are Golden Retriever, Pembroke, Labrador Retriever, Chihuahua, and Pug. Doing this breakdown made it clear that there were still some tweets that did not contain dogs such as “ocarina”, “motor_scooter”, and “jellyfish”.

For the final analysis, I wanted to find the mean number of the ratings. So I found the rating_numerator and rating_denominator means which were 12.84 and 10.55 respectively. I also did a describe on the two columns to figure out why the numerator was above 12, because most of the tweets I saw were 12. According to the describe, 75% of the numerators were 12 with the max being 1776. The denominator had a similar trend that 75% were 10 with a max of 170.



For the visualization, I wanted to use the retweet column because as a fan of social media, that was the column I was most interested in seeing. So I created a new column named month-year so I would be able to group by the month and year and I totalled the number of retweets per month and put it on a bar graph. They were all pretty similar besides for the month of December in 2015

where the total reached nearly 400,000 retweets and the month of April of 2016 reaching as low as around 50,000 retweets for that month.