## BRIEF ANALYSIS AND GENERATION OF INSIGHTS INTO OUR DATA

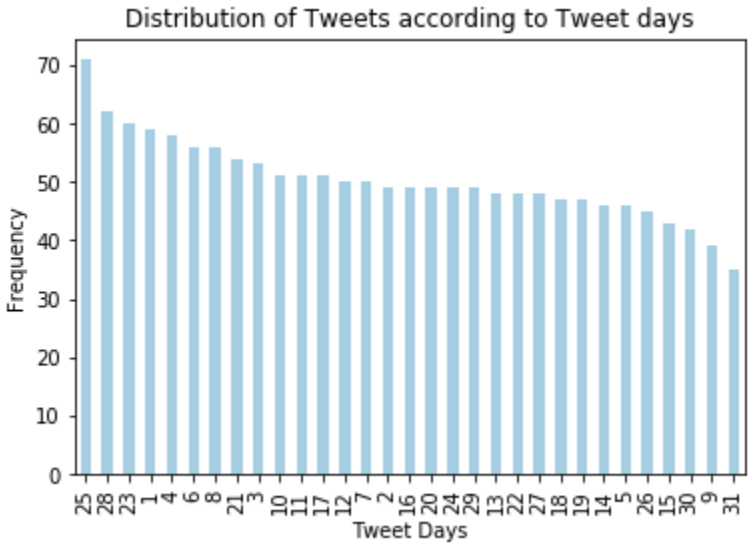
After Cleaning up the data and selecting important variables or parameters for our analysis, I joined the three data frames from three different sources into a master data frame called df_master.

The columns contained in the df_master are tweet_id, text, rating_numerator, tweet_day, tweet_month, tweet_year, p1, p1_conf, p1_dog, favorite_count and retweet_count.

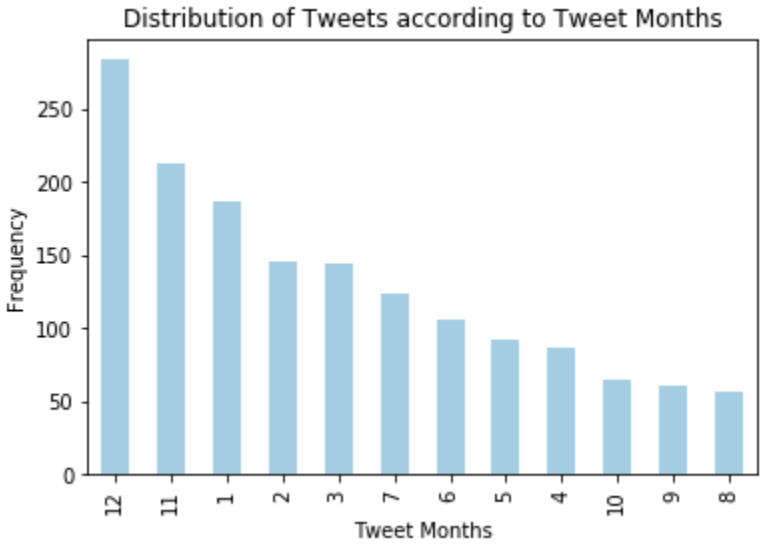
Diving into my data to get some relevant statistics, First, I was curious to know the tweet with the highest number of retweets. It turned out that the tweet with the highest number of retweets had a retweet_count of 79515. It was a tweet about a Labrador_retriever and it was made on the 18th June, 2016. The dog was given a rating of 13.

Next, I wanted to know the most liked tweet. The most liked tweet has a count of 132810 and it was made on the 21st January, 2017.

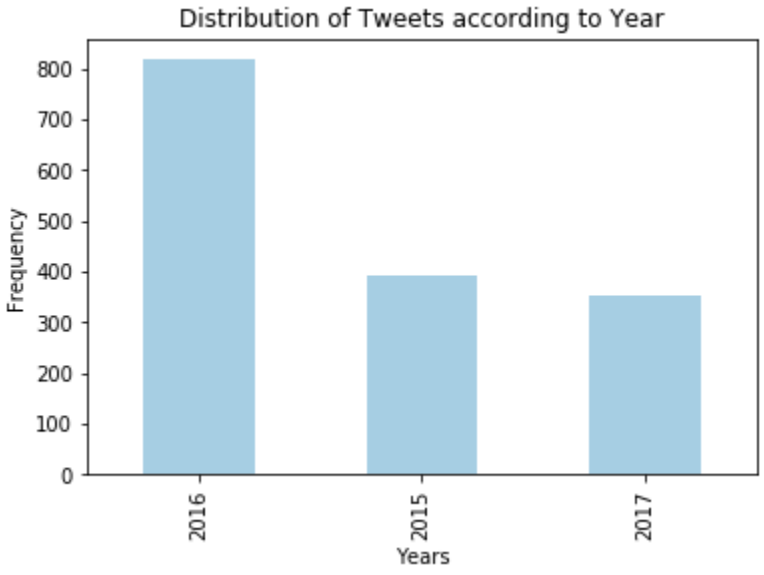
Diving more into our data, I wanted to see the distribution of tweets among tweet days(from 1st to 31st). I discovered that 28th, 25th and 23rd were the top three days with the most tweets. This is worth digging deep into. Perhaps, a clearer way would be to group data tweet days by day names(Monday-Friday).



The next thing I did was to check for the distribution of tweets according to the months of the year. I saw that the months of November and December have the highest number of tweets. Could this be related to tweet days? Perhaps. This would be worth further exploration.




Next, I looked at how tweets vary across the different years. As we can see from the graph beneath, there is a significant difference between the frequency of tweets in 2016 and (2017,2018). What caused these changes? This is worth further exploration



Next, I look at a word cloud of the most used dog names in our neural network dog predictor

**Word Cloud of most used dog names**