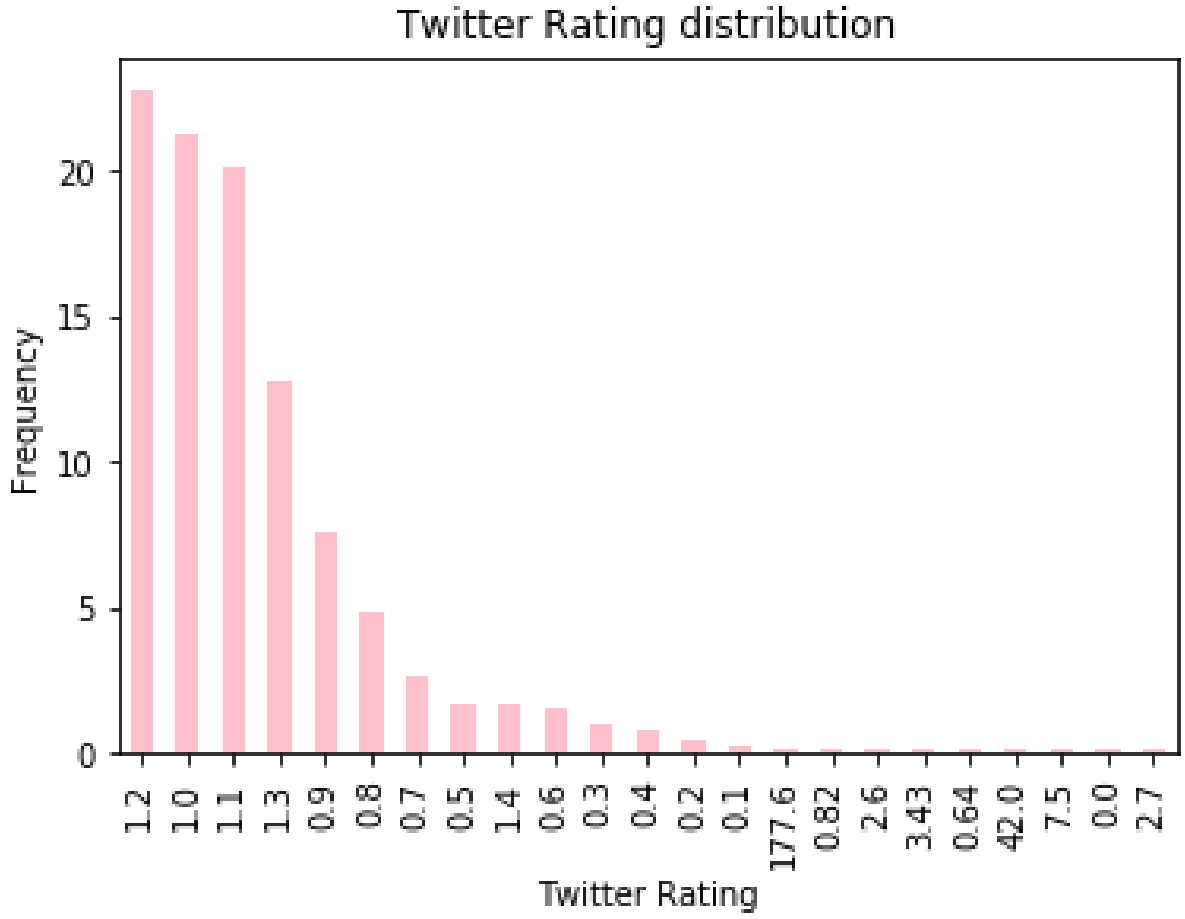
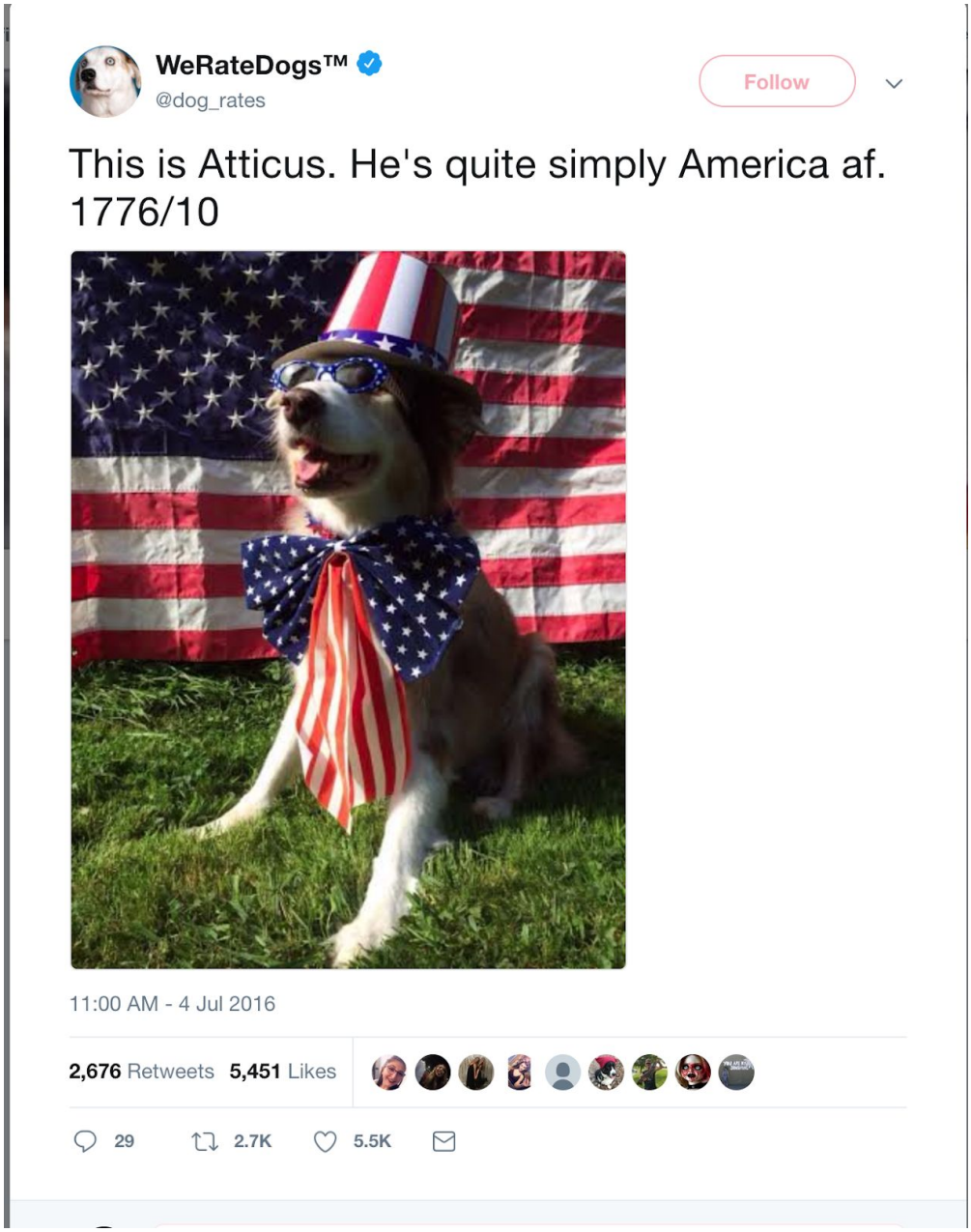
WeRateDogs - Twitter Data Analysis

Twitter data gathered from three different sources , cleaned and merged into a single file was analysed to explore the following questions.

1. Which twitter rating has the most occurence in the merged dataset?

   The aim was to find which twitter rating has the highest occurrence . From the analysis we see the max rating is 1.4 (14/10) , only 1.69% of dogs have this rating . The minimum rating is 0.0 a very small percentage of dogs have this rating as well (0.05%). About 22.8 %dogs have a rating of 1.2 (12/10) which makes it the most common rating followed by 21.36 % of dogs with a rating of 1.0(10/10).





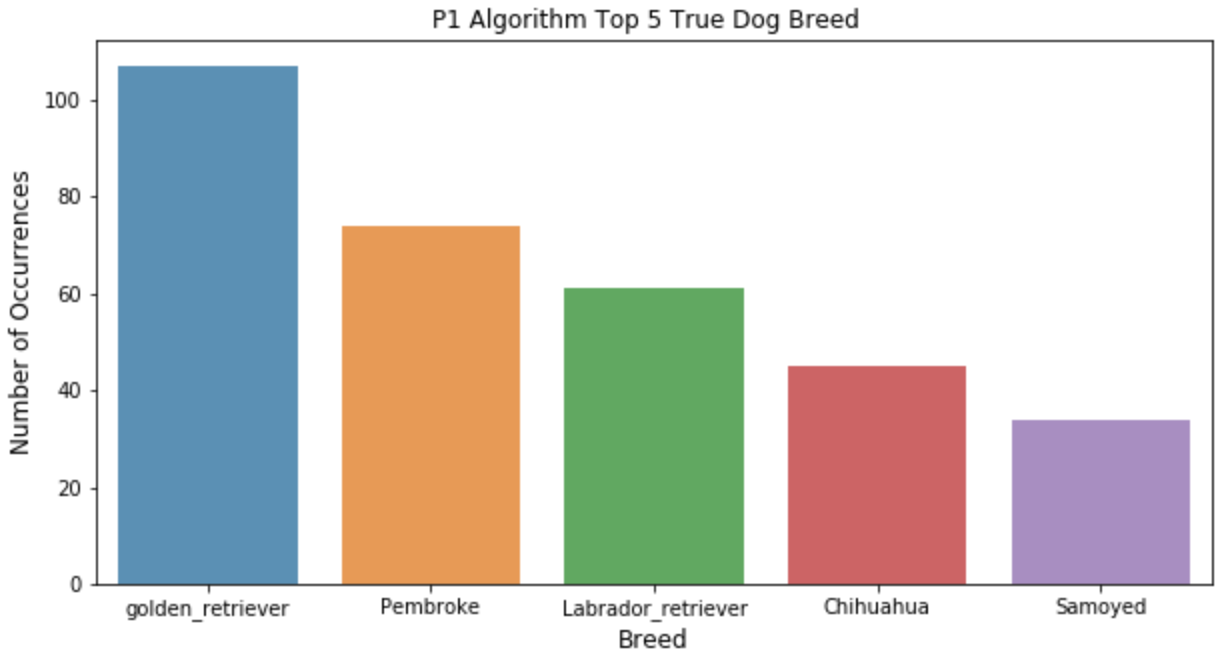
From the above distribution we notice very high twitter rating. On further analysis by checking the tweet online it is concluded that indeed these tweets have a very high rating.

2. Which source does the most twitter rating originate from?

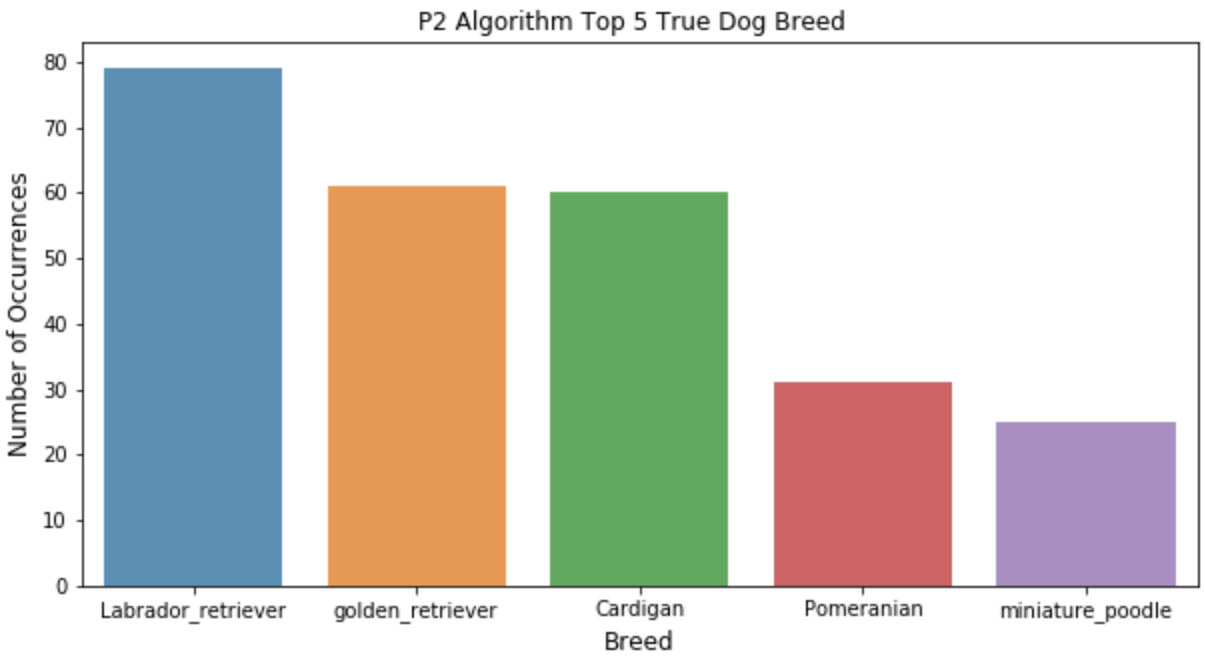
From the above tweet rating analysis we see there are very few dogs rating less than 1.0 (10/10) . Hence for future analysis i only consider tweets with rating greater than rating of 1.0. On analyzing this subset of data it can be seen that most rated dog tweets originate from iphone of 1101

Twitter for iphone 1101
Twitter Web Client 15
TweetDeck 7
Name: source, dtype: int64

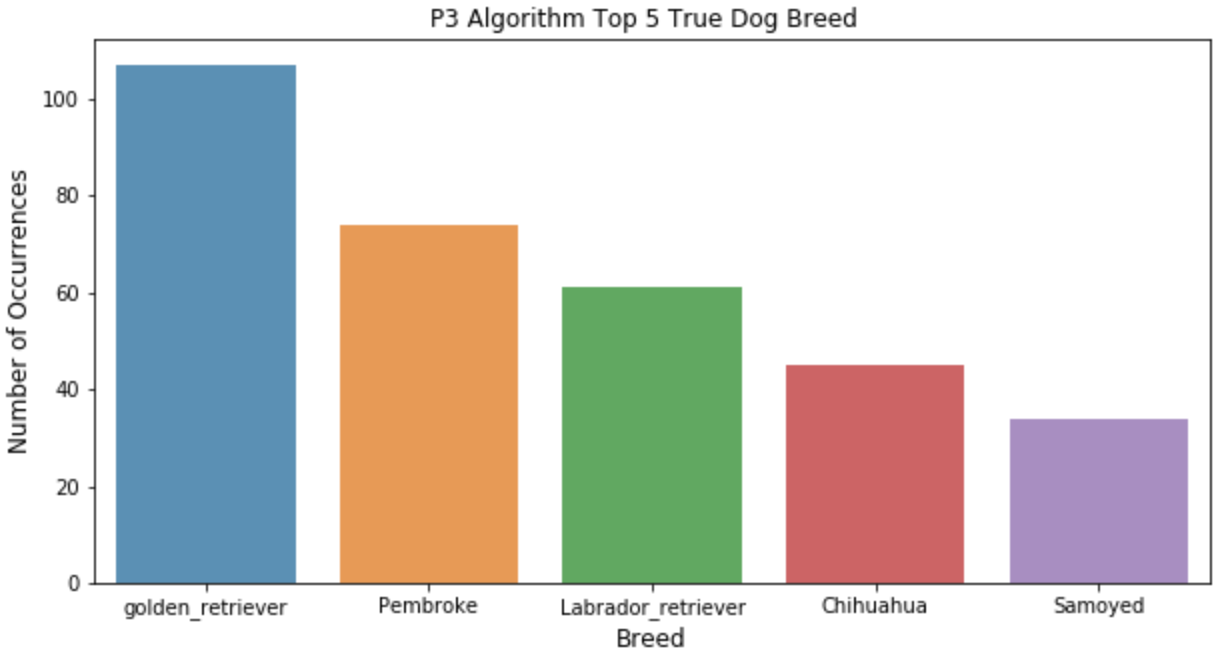
3. Most popular dog breed predicted by the three algorithms p1, p2 and p3 ?



For p1 algorithm the most popular dog breed is Golden Retriever.



For p2 algorithm the most popular dog breed is Labrador Retriever.



For p3 algorithm the most popular dog breed is Golden Retriever.
From the three graphs we can see the top three popular breed are the same for both p1 and p3 algorithm.

4) How is the retweet count and favourite count correlated?

The scatter plot below shows a positive correlation between retweet count and favourite count. That is the retweet count increases with increase in the favourite count .Furthermore the correlation coefficient also supports the graph.
Correlation coefficient = 0.9220422633043