INSIGHTS :

Now that we've got the master dataframe, let us try to find some informative insights from it. Most interesting columns to gain information from are : source, dog_breed, retweet_count, fav_count. So let us see what derivations can we make from these.

SOURCE:

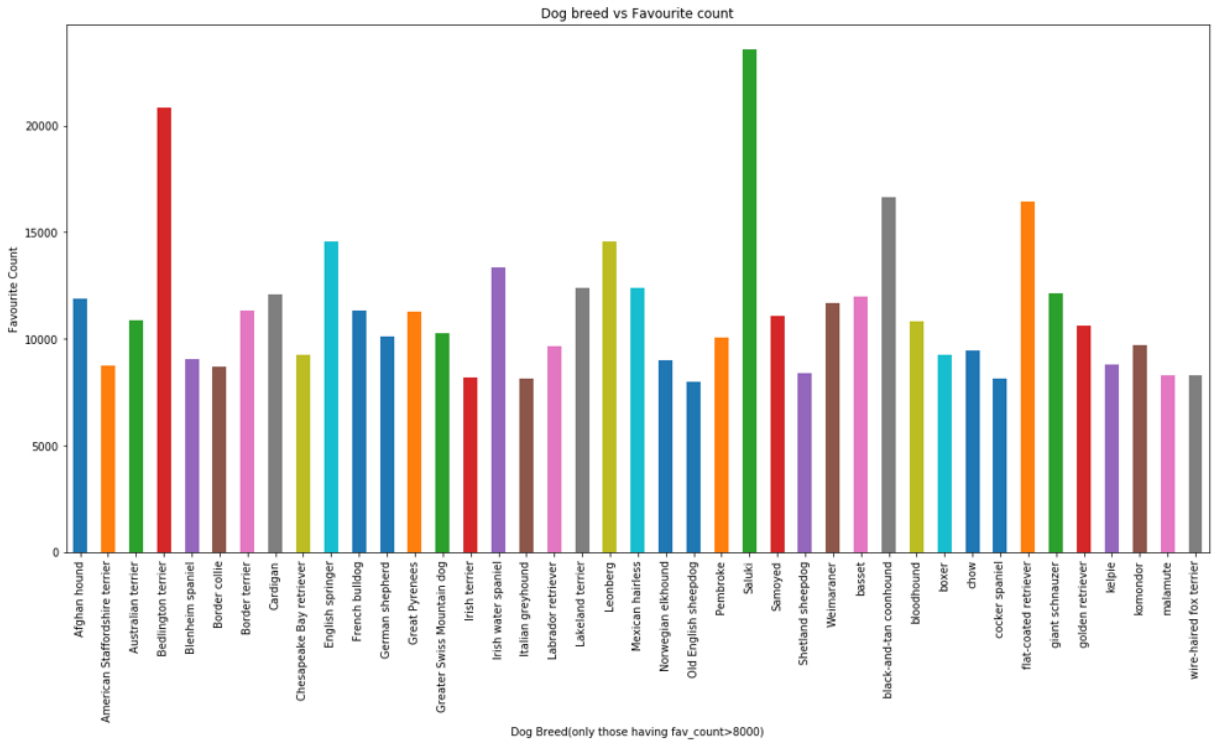
How do most people have access to these tweets? From where do most people make their uploads? This is what we'll try to answer by looking at the stats from the source column. By running value_counts() function of pandas on this column of the dataframe, we can see how most people make their uploads.

SOURCE	COUNT
Twitter for iPhone	1813
Twitter Web Client	26
TweetDeck	10

As we can see, most people of our dataset make their uploads from their iPhone's(1813 of 1849). A few people make uploads from the Twitter web client, while a very less number of people prefer to use TweetDeck for this purpose. TweetDeck is a social media dashboard application for management of Twitter accounts.

DOG BREED VS FAVOURITE COUNT:

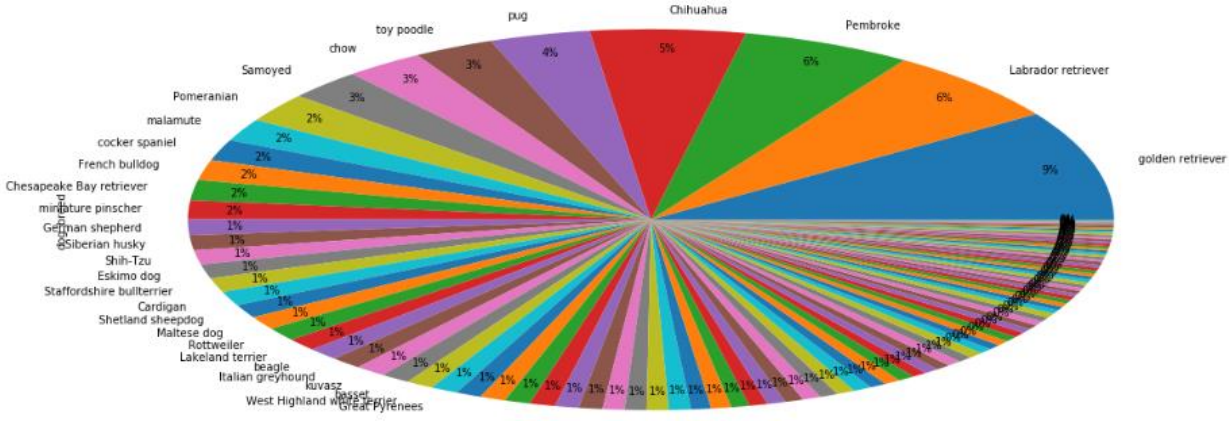
Is there any preference of people for any kind of dog breed ? Which breeds get the most number of favourites? We look to find the answer to this, by grouping our dataset by the dog breed, and finding the mean number of favourite count for each of them.



Very few dog breeds actually manage to get favourite counts above 15000. Bedlington Terrier, Saluki, Black-and-tan coonhound and flat-coated retriever are the breeds which make it. The mean favourite count for a dog however, is just a shade below 7600.

WHICH DOG BREED CONSTITUES HOW MUCH PERCENTAGE OF UPLOADS?

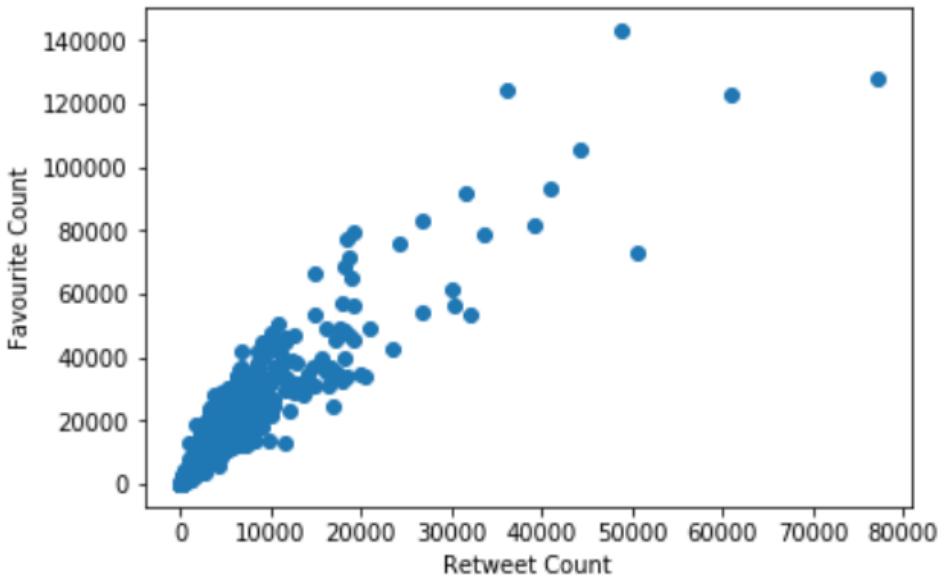
Which dog breed has the most number of uploads? Which all breeds are the most tweeted ones? This can be answered by the following visualisation:



As we can see, almost 9% of uploads (147) belong to golden retrievers. So, the title for most tweeted dog breed goes to Golden Retrievers. They are closely followed by Labrador retrievers(6%), and Pembroke(6%).

RETWEET COUNT VS FAVOURITE COUNT:

It seems logical that there will be some kind of correlation between these 2 features, right? After all, we only retweet those which we like a lot, and consider favourite. A scatter plot will help us visualise the 2 features, and we'll know if there is any kind of correlation between the 2.



As we thought, there does seem to be a strong correlation between the 2. There is a strongly positive relation between them. This is confirmed by their Pearson correlation coefficient, which is 0.91 .