Act Report

The current project contains Twitter archived dataset contains tweets related to the variety of dogs and their ratings. The data is at individual tweet id level containing:

- Date and time of tweet
- Source
- Retweet and inreply tweet id and date time
- Urls
- Dog names
- Text of the tweet
- Ratings
- Type of dog referred in the tweet

At the initial phase certain checks like missing values, formats of variables etc., were conducted to identify Data Quality issue.

The following Data Quality issues were identified and fixed like:

- Expanded urls were have missing values and thus we dropped those missing records
- Columns like timestamp and retweeted timestamp were in strings, so we convert those to datetime format
- Typo errors like vine.co in source column is replaced by vine.com
- ‘<a’ and ‘</a>’ are HTML codes present in source column, which has been removed
- Variables like tweet_id, retweeted_status_id, retweeted_status_user_id etc. were in integer format and they had been converted to string
- Column expanded_urls have comma separated repetitive values, which had been fixed by removing characters after the comma
- Rating_numerator and rating_denominator were in integer format but ratings could be float thus they are converted into floats
- It is noticed that the dog name column starting with lower case is having values like ‘a’, ‘an’, ‘by’ etc. which are not valid dog names. So we removed those records

Apart from this, Tidiness issues were also noticed in the dataset like:

- Timestamp variable had both date and time integrated into variable. But if one needs to analyze the monthly pattern or hourly patterns of the tweets, such analysis becomes difficult with the current data structure. Thus this timestamp column is separated into Date and Time respectively
- Having column names corresponding to the types of dogs, ‘doggo’, ‘floofer’, ‘pupper’ and ‘puppo’ makes the dataset heavy, thus we combine all the information from all the 4 columns and segregated into a single column

Once all the data cleaning process is done, we rechecked for data types, missing values etc. to ensure the data is ready for carrying out analysis.

The key questions which we looked for answering are as follows:

1. **Which Year has more number of tweets?**

We extracted the Year information corresponding to newly created Date column, following is the frequency of number of tweets across years:

Year	**Number of tweets**
2015	619
2016	1116
2017	453

Here, we see the number of tweets increases from 2015 to 2016 and then further decreases in 2017.

2. **What percentage of ratings for dogs are good?**

The ratings are quantified as Good, where numerator_rating is greater than denominator_rating. The percentages for Good and Bad ratings are as follows:

Type	**Percentages**
Good	62%
Bad	38%

Thus the above results, shows that the 62% of the total number of tweets has a Good review for dogs.

3. **What proportion of tweets contain the breed names?**

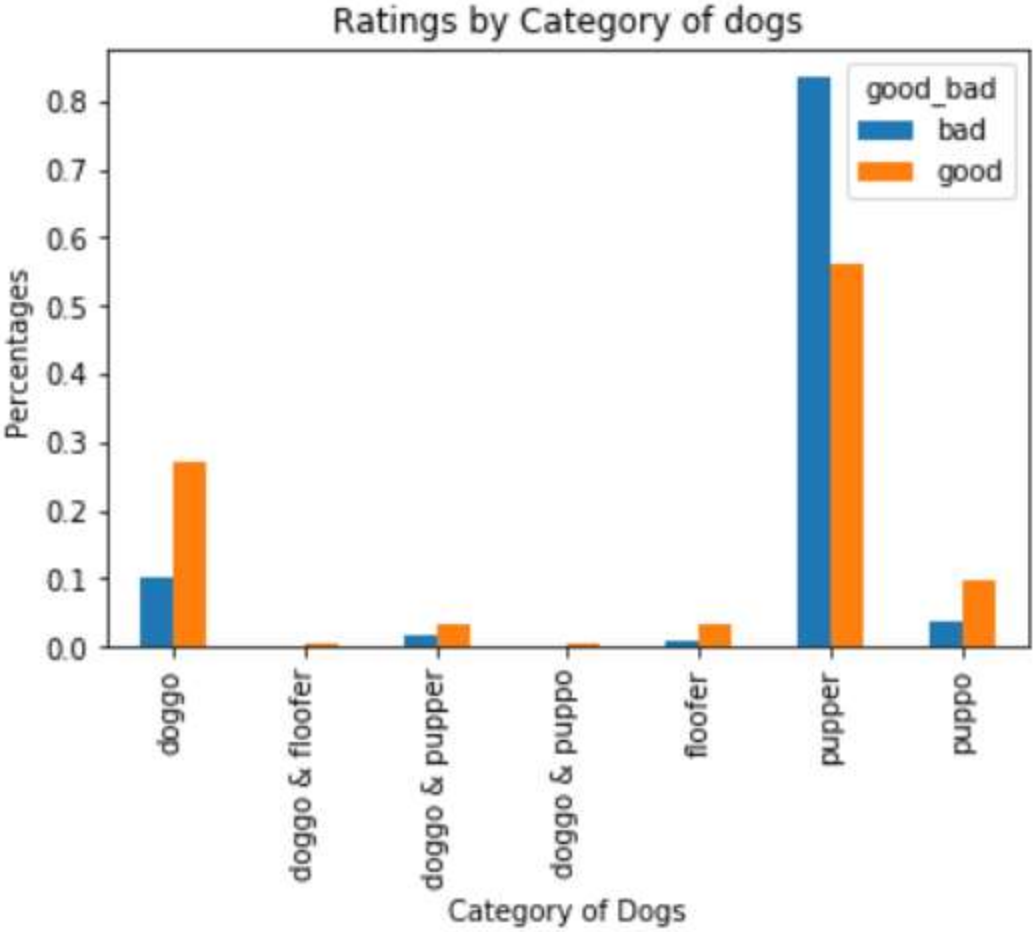
In order to get the popular breeds of dogs across tweets, we look for the counts across breeds:

Breed name	**Percentages**
None	83.64%
Pupper	10.55%
Doggo	3.56%
Puppo	1.27%
Doggo and Pupper	0.45%
Floofer	0.41%
Doggo and Puppo	0.05%
Doggo and Floofer	0.05%

Thus we see that almost 84% of the posts, doesn’t contain the breed of the dogs. Out of rest 16% of post which contains breed of dogs, 10.5% of them are Pupper, 3.5% by Doggo followed by the rest.

4. **Is there any relationship between ratings and the breeds of dog?**

For this analysis, the tweets which only breeds of dogs are considered and we plot the percentages of Good and Bad ratings across the breeds of dogs



Thus from the dog categories, we see that only Pupper is having more number of Bad ratings while among the rest, Doggo is having best rating among Floofer and Puppo. Puppo is having more rating as compared Floofer. Thus, above analysis suggests if we run a Machine Learning Algorithm, most likely the algorithm will predict a bad rating for dogs corresponding to breeds Pupper. While breeds like Doggo, Floofer and Puppo is most likely to have a good rating.