INTRODUCTION

The data set is using in this project is gotten from tweet archive of Twitter user @dog_rates. The data is based on twitter from different people and the dogs are rated based on the comments about the dog.

DATA GATHERING

In gathering the data, three separate datasets were used in the process of perform this analysis which includes 1. Enchanced Twitter Archive which was provided by weratedogs for analysis 2. The second data which is an image prediction file tsv file was gotten using programmatically with the url: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv 3. The third data tweet_json.txt was gotten using tweety via and API after an acceptance from twitter to become a twitter developer.

ACCESSING THE DATA

The data was accessed using python based jupyter notebook. While accessing the data 8 data qualities issues were identified and documented for data cleaning. Also, two tidiness issues were also identified and documented.

The data issues include:

1. The timestamp and retweeted_status_timestamp is not recognized as datetime
2. Some record in the twitter-archive are retweeted
3. some of the missing records are not descriptive enough
4. The doggo, floofer, pupper and poppo columns in the twitter-archive can lead to redundancy
5. Some ratings are not properly extracted and correctly formatted
6. Correcting the tweetid data type
7. Extracting URL address from source
8. Some of the dog's name are not correctly extracted for instance there are names like Goose, General, Mo, this, unacceptable etc.

Tidiness issues:

1. In the image-predictions we have more than one model generating prediction

2. The data are not merged together which can result in data loss

CLEANING DATA

Following the above defined issues, it is necessary to clean the data to make it ready for analysis all data cleaning to place in jupyter notebook. Most cleaning was performed on the twitter_archive_enhanced.csv as image_prediction.tsv and tweet-json.txt where already in a clean format. After which all the three data set were merge into one and saved as twitter_archived_master.csv