Project Report

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10.

This report document the steps taken to gather, clean and store the data

Gathering the Data
Three source was given for the data gathering exercise, the first source was a CSV provided by Udacity, which was downloaded from the project page called twitter_archive_enhanced, the second source was a tab seperated value text extracted from a page from Udacity website, while the third source was to be from twitter API, but I resulted to using the text file provided by Udacity, after using request to get the TSV file with requests and loading the text file with JSON, all the files was read into a DataFrame with Pandas

Accessing the Data
The data was access programtically and visually, and the following issues were indentified;

Quality Issues

1. Retweets and replies were identified and dropped.

2. The following columns had missing values and they were be dropped as they will not impact our analysis

- in_reply_status_id
- in_reply_to_user_id
- retweeted_status_id
- retweeted_status_user_id
- retweeted_status_timestamp

3. The timestamp column is in the wrong data type and was converted to datetime
4. The source column had text in an anchor tag and we need to extract only the text
5.There were tweets that are not about dogs in the text column, this will be dropped
6. The expanded urls column is not needed for our analysis, it will be dropped
7. The none value in doggo, floofer, pupper, will be dropped
8. dog rating to be roperly extracted
9. rename id in the tweets df as tweets_id

Tidying Issues
1. These columns doggo, floofer, pupper and puppo will be combined into one
2. These columns p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3_conf, p3_dog will be combine into two columns, dog breed and confidence
3. The three datasets will be combined as one – combined_dataset

Cleaning the Data
After all the above issues were clean, the three datasets was merged in one DataFrame called combined_dataset and the rating_numerator column was tested for outliers, which were also dropped.

Storing the Data
The data was store as a table called twitter_archive_master in a sqlite database as called WeRateDogs, as well as a CSV file