**Udacity Project II Report**

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog, and these ratings almost always have a denominator of 10. This report documents 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 were 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 was stored in a CSV format