WeRateDogs Tweet Analysis

Data Wrangling Report

Prepared By – Harsh Navin Gupta

1. About the Datasets

Three datasets were used for the project. The details for each of the dataset are as follows:

- WeRateDogs Tweet Archive: This dataset was given as an on-hand file, which was downloaded and stored on the local machine. This dataset was a collection of all the tweets from the user @WeRateDogs and contains data for 2356 tweets.
- Image Predictions Dataset: This dataset is hosted on the Udacity Cloud and was downloaded and stored onto the Local System using Python **Requests** library. This dataset contains the links to images attached with a particular tweet, and the top three predictions of the breed of the dog in that image, with their confidence levels and whether the prediction is a Dog Breed or not.
- Additional Information about Tweets: This dataset was created during the implementation of the Project, by making use of the Twitter API **tweepy**. The dataset contains the number of retweets and number of times the tweet has been marked favorite, for every Tweet ID in the Tweet Archive.

2. Data Quality Issues

The following Data Quality Issues were identified in the above datasets:

- The datatype of Timestamp is incorrect, should be **datetime** datatype.
- Remove All Tweets from Tweet Archive Dataframe, which have incorrect Tweets IDs.

- Replace Missing Name values in **name** represented by *None* with NaN.
- Replace the illogical values **a, by, the, an, all, just, very** in the **name** column with NaN, to represent Invalid values.

  It can be observed that the names a, by, the, an, all, just, very, appear in the name column of the Tweet Archives. Although the names are strings are correct according to the domain rules, they are illogical to be considered as the names of pets, and thus maybe an error.

- Replace Missing Values in *doggo, floofer, pupper and puppo* with NaN.
- Remove All Tweets which are **Retweets**.
- Remove All Tweets which are **Reply Tweets**.
- Remove All Tweets from Tweet Archive which do not have an image.
- Remove Tweets which have Numerators = 420, 1776, 960, 666, and Denominator = 7, as they are not rating tweets, and thus invalid.
- Incorrect Numerator and Denominator values in tweet with (1,2). Replace them by correct rating of (9,10).
- Extraction of Short URLs from Tweet *text* column, and placing into new column called *short_url*.

3. Data Tidiness Issues

- Melt the columns *doggo, floofer, pupper and puppo* into a single column.
- Multiple Columns in Image Prediciton dataframe, to represent the dog breed, change to representation in a single column.
- Merge the Tweet Archive, Image Prediction and Additional Information for every unique tweet into a single dataframe.