***Wrangling Report***

**Introduction:**

I have chosen the WeRateDogs account to analyse and this report consists of the wrangling efforts done on the datasets from different sources.

The wrangling is done in the jupyter note book (wrangle\_act.ipynb) and the steps involved in wrangling is explained below:

1. Gathering of the data from 3 different sources.
   1. Already there was CSV file which contains archive twitter data related to posts. I had to read this file into a data frame (df\_enhanced)
   2. An online TSV file which contains the image predictions, regarding dog breeds, generated by a neural network, based on the picture of the dogs. This was read into a data frame (df\_breed).
   3. For gathering the third data set used the Twitter API (Tweepy) to gather twitter data about the number of likes and favourites. For each tweet ID I downloaded the data and stored it into a JSON file text (tweet\_json). I then created needed columns and then extracted the information required into a python data frame (df\_tweets)
2. Assessing Data:
   1. As part of my analysis work, I have identified the issues as tidiness and quality issues for individual datasets. Below are the details for them:

Tidiness:

1. There are 3 datasets instead of a single master dataset with all required details.
2. In df\_enhanced and df\_breed there are multiple stage columns which could have been represented in single column.

Quality:

df\_enhanced-

1. the dog\_stage column is having string datatype
2. The rating column having two different for numerator and denominator without proper scaling.
3. Incorrect names,missing names in name column, such as a, an, the... (all names starting with lower case letters)
4. retweeted records: retweeted\_status\_id, retweeted\_user\_id, retweeted\_status\_timestamp
5. missing data in name and stages columns showing as 'None'
6. Time stamp is string datatype, which is not correct
7. in\_reply\_to\_\_status\_id, in\_reply\_to\_user\_id should be changed from scientific float to string data type
8. Some rating numerators were extracted from text column
9. There are few columns whose readability has to be improved, ex source column.

df\_breed:

1. The breed names starting with lower case letter

2. non-dog records in the dataset (ex:- desktop computer, box turtle etc)

3. the appropriate data type for columns p1., p2, p3 should be categorical.

4. Few columns can be removed which are not required.

**Cleaning: ( I have made a copy of each data frame into other for safer side)**

1. I have removed inconsistencies in data for names column, by replacing the incorrect and missing names. ( this was achieved by using regex in text columns)
2. As my requirement was to analyse, so queried the dataset without retweets.
3. There were incorrect data types, time stamp was string which is changed to datetime.
4. Columns in\_reply\_to\_status\_id, in\_reply\_to\_user\_id was changed from scientific float to a string. ( used lambda function to achieve this referring to code in stack overflow).
5. Source Column was not readable, hence I created a new user\_source column based on the source column and dropped the old one.
6. There were non dog records such as desktop computer, box turtle. I observed that there are lot of other things in the dataset, identified as dog by the neural network. As the predictions was not always consistent, I dropped the columns where none of the predictions were dogs. With this step I could exclude many incorrect data records at same time had number of good records.
7. The p1, p2, p3 columns are supposed to be categorical variables, for better analyzation I have changed the data type for these columns.
8. The dog\_stage is string data type. I also changed dog\_stage to categorical as the stages represent different categories.
9. There are incorrect numerators in the rating\_numerator column without header scale. For which I have removed the rating\_denominator column and added as title to rating\_numerator column.
10. There were unnecessary category columns such as doggo, floofer, pupper, puppo. I merged these columns into single column into one column named dog\_stage. I have also dropped the duplicates by querying the number of duplicates.
11. Final one but most important is merging the 3 datasets into a single data set better understanding and analysis. Initially I merged df\_enhanced and df\_tweets data frames on tweet\_id. Then I completed the cleaning for 3 data set df\_breed and merged with the resultant dataset previously. I did not use inner join because didn’t want to lose any data.