**Data Wrangling Report**

1. **Introduction**

This document records the Data Wrangling efforts done on the three pieces of twitter data the focus on Gathering, Assessing, Cleaning.

1. **Data Gathering**

Data for this project is gathered from three different sources.

1. *twitter\_archive – (from a csv* file) The WeRateDogs people are generous enough to provide us with a bunch of their tweets data to Udacity Course purposes. The data in this csv contained some basic data such tweet\_id, name, rating\_denominator, rating\_numerator, dog stages etc.
2. *image predictions* – (*from web*) This file is download from Udacity’s server using the provided url with request library. The file contained information such as dog breed, different algorithm data and their confidence values.
3. *tweet\_data*- (*from an api*) This file is downloaded and constructed using the python twitter API Tweepy. The resultant file contained information such as tweet\_id, retweet\_count, favorite\_count.

From these three files, three data frames have been created.

* **df\_twitter\_archive**
* **df\_images**
* **df\_tweet\_data**

1. **Data Assessing**

All the three data frames are assessed to identity quality and tidiness issues.

1. **Tidiness issues**

**df\_twitter\_archive**

* There seems to be some tweets that are retweets, we shall be deleting them for tidiness purpose.
* Dog stages are separated and stored in four different columns “doggo”, “floofer”, “pupper” and” puppo”. We shall melt them into one column “dog\_stage”
* We shall insert a new column "Rating" which is equivalent to numerator/denominator for our calculations later on in the analysis.
* We shall delete unnecessary columns for tidiness.

**df\_images**

* We shall melt down the algorithms columns (p1, p2, p3) and algorithms confidence columns (p1\_conf, p2\_conf, p3\_conf) to have only "algo", "algo\_conf" columns.
* We shall delete all unnecessary columns.

**df\_tweet\_data**

* This tweet\_data data frame seems to be tidy, no action needed here.

Overall, we shall merge the three data frames into one master data frame.

1. **Quality**

**df\_twitter\_archive**

* We shall change the timestamp to a standard format data type
* The name column has some weird names as dog names, we shall inspect and clean them up.
* The “ratings\_denominator” also has 0 values, we shall clean them up.
* The "ratings\_numerator" have some very high values. We shall delete any rows with numerator values greater than 20.
* Will change the dog names to all so that every names start with upper case.
* Will delete expanded urls row with missing data.
* Change the "source" text to something that is human readable.

**df\_images**

* There are rows where all three algorithms did not predict the image to be a dog. We shall delete such rows.
* Will change the dog breed columns to be consistent case (some are lower and some are upper)

**df\_tweet\_data**

* This tweet\_data data frame seems to be of quality, no action needed here.

1. **Data Cleaning**

Before cleaning the data, a copy of each these data frame has been saved.

All the three data frames are assessed in the previous step and ready to be cleaned.

* Some of the key tidiness issues that will be cleaned up are:

**df\_twitter\_archive**

* All the rows that are considered retweets have been deleted for analysis.
* Have removed the different dog stage columns ("doggo", "floofer", "Pupper", Puppo) and added a new column "dog\_stage".
* Have inserted a new column "Rating" which is numerator/denominator.
* Have deleted unnecessary columns from df\_twitter\_archive – *["in\_reply\_to\_status\_id","in\_reply\_to\_user\_id", "retweeted\_status\_id", "retweeted\_status\_user\_id", "retweeted\_status\_timestamp", "rating\_denominator", "rating\_numerator"]*

**df\_images**

* Have reduced the algorithms columns (p1,p2,p3) and algorithms confidence columns(p1\_conf, p2\_conf, p3\_conf) to have only "algo", "algo\_conf" and best predicted breed "dog\_breed".
* Have deleted unnecessary columns.

**df\_tweet\_data**

* There were no tidiness issues in df\_tweet\_data
* Some of the key quality issues that will be cleaned up are:

**df\_twitter\_archive**

* "*timestamp*" column datatype is changed to *“datetime*” data type
* "name" column with dog names "a", “an”, and “the” have been replaced by "None."
* Rows with ratings\_denominator = 0 values have been deleted.
* Rows with "ratings\_numerator" very high values (greater than 20) have been deleted.
* Dog names in the “name” column are capitalized.
* Have deleted rows where there are null values for “expanded\_url” columns
* Have changed the “source” column text to something that is more meaningful and readable.

**df\_images**

* Have deleted the rows where all three algorithms are predicting the image not be a dog.
* Dog breed columns p1, p2, p3 are capitalized.

**df\_tweet\_data**

* There were no quality issues in df\_tweet\_data

After the cleanup is done, the all the three data frames are merged into one dataframe “df\_master\_twitter\_data”. The final master twitter data is then stored into a csv called 'twitter\_archive\_master.csv.