# 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.

## 2. 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**
-

## 3. Data Assessing

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

### a. Tidiness issues

**df_twitter_archive**

- Dog stages are separated and stored in four different columns "doggo", "floofer", "pupper" and" puppo". We shall melt them into one column "dog_stage"

- There seems to be some tweets that are retweets, we shall be deleting them for tidiness purpose.
- We shall delete unnecessary columns for tidiness.
- We shall insert a new column "Rating" which is equivalent to numerator/denominator for our calculations later on in the analysis.

**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.

**b. 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 and some very high 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.

**df_images**

- There are rows where all three algorithms did not predict the image to be a dog. We shall delete such rows.

**df_tweet_data**

- This tweet_data data frame seems to be of quality, no action needed here.

1. **Data Cleaning**

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**

- Will delete the retweets.
- Will remove the different dog stage columns ("doggo", "floofer", "Pupper", Puppo) and just have a new cloumn "dog_stage".
- Will delete unnecessary columns from df_twitter_archive.
- Will insert a new column "Rating" which is numerator/denominator.

**df_images**

- Will reduce 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".
- Will delete 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 will be changed to *"datetime"* data type
- "name" column with dog names "a" will be replaced by "None."
- "name" column with the dog names "an" will be replaced by "None."
- "name" column with the dog names "the" will be replaced by "None."
- ratings_denominator with 0 value will be deleted.
- will clean up ratings_denominator values of 11 and make them 10.
- will clean up ratings_denominator values of 50 and make them 10.
- will clean up ratings_denominator values of 80 and make them 10.
- will clean up ratings_denominator values of 20 and make them 10.

- will delete anything other than 10 as the rating_denominator value (after changing the above rows).
- "ratings_numerator" have some very high values. Will delete any rows with numerator values greater than 20.

**df_images**

- will delete the rows where all three algorithms are predicating the image not be a dog.

**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.