REPORTING: DATA WRANGLING

This report contains explicit information about the data wrangling processes carried out while working on this project.

The data that was wrangled for this project is a tweet archive of Twitter user @dog_rates also knowns as WeRateDogs. WeRateDogs is a Twitter account that shares dog images and writes a brief panegyric about the dog, then they let their followers rate it by favoriting it. By asking WeRateDogs to share with us some of their tweets, they did. They have shared 5000+ of their tweets which contain some basic data. Sometimes in their brief panegyric, they mention the breed of the dog, and some others they don't. But thanks to Udacity, they have performed some neutral network procedures to classify the dogs based on their images which are shared with the tweets.

The wrangling process was conducted on the Udacity workspace. The wrangling process was carried out in 3 stages

1. Gathering of data
2. Assessing data
3. Cleaning data

GATHERING DATA

The data used for wrangling processes were gathered from 3 different sources

1. Enhanced Twitter Archive: This data was downloaded manually from the following link address: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/59a4e958_twitter-archive-enhanced/twitter-archive-enhanced.csv
2. Image Prediction tsv File: The data for this file was downloaded programmatically by importing requesting library

2. Use the Requests library to download the tweet image prediction (image_predictions.tsv)

```
: url = " https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-
response = requests.get(url)

with open('image-predictions.tsv', mode = 'wb') as file:
    file.write(response.content)
```

3. tweet_json.text: Additional data file was downloaded from Twitter by using the tweepy library to query additional data via the Twitter API. the resulting data generated was a JSON, which was later stored in a text file.

ASSESSING DATA

After gathering from 3 different sources, each data was assessed visually with code editors, and spreadsheets. Furthermore, after assessing the data visually, the data was programmatically assessed using the panda's inbuilt functions like head, tail, info, sample, and describe to gain more insights into the data. 8 different quality issues were highlighted and 2 tidiness issues were highlighted after assessing the data visually and programmatically.

CLEANING DATA

The highlighted quality and tidiness issues that were documented during the assessing stage were thoroughly cleaned during this phase. Before cleaning the data a copy of all the data to be cleaned was generated first.