# Project Reporting: Dog Ratings Data Wrangling and Analyzing

Below are the four steps taken in this Wrangling

1. Introduction
2. Data gathering
3. Data Assessment
4. Data cleasing

## INTRODUCTION

This is a project to gather data about Dog ratings from 3 different sources namely: WeRateDogs Twitter archive data, the tweet image prediction using the Requests library and from the Twitter API. The data is to be read into 3 different dataframes using differents methods that are stated in the project rubik

## DATA GATHERING

The first step i took in data gathering is to import all the libraries that i needed for the whole wrangling and analysis process.

```
import pandas as pd

import numpy as np

import wptools

import os

import requests

from PIL import Image

from io import BytesIO

import matplotlib.pyplot as plt

import seaborn as sns

%matplotlib inline
```

The first set of the data i gathered is from the WeRateDogs Twitter archive. This data is provided in the Udacity class in a csv file called twitter-archive-enhanced.csv. I read the data into a dataframe df = pd.read_csv('twitter-archive-enhanced.csv')

I downloaded the second set using request library then read it into a dataframe using: df_image = pd.read_csv('tweet_image_prediction/image-predictions.tsv', sep = '\t') taking note of the type of file

Using tweepy to query the Twitter API and save it into a json file, I used json to read the json file into a dictionary then into a dataframe:

with open('tweet-json.txt', 'r') as f:

dictlist = [json.loads(x) for x in f]

df_json=pd.DataFrame(dictlist)

# DATA ASSESSMENT

The assessment of the data was done using visual assessment and programatic assessment. Some of the assessment codes are: df.head

df_image.head

df_json.head

df.info

df_image.info

df_json.info

df_image.img_num.value_counts()

df_image[df_image.tweet_id.duplicated()]

all_columns = pd.Series(list(df) + list(df_image) + list(df_json))
all_columns[all_columns.duplicated()] among others

Using the different codes i was able to list some of the data quality issues and tidyness issues that i was going to clean

## Data Quality Issues

1. df'timestamp' and 'source' is in wrong datatype
2. df'source' still has the html residues
3. df'timestamp, text, name, source', df_image'p1, p2, p3 and their conf' do not have a clear descriptive heading
4. Reduce the decimal points in the p1_conf, p2_conf and p3_conf

5. df_json[id] does not have the correct naming. it has the tweet id so the heading needs to be changed
6. df_json columns that are not needed should be dropped
7. df dataframe has retweets. We only need the original tweets
8. df dataframe columns that are not needed should be dropped

### Tidiness issues

1. The information in the df_json dataframe and df_image dataframe is needed in the df dataframe
2. The stages of dog is spread out into four columns instead of one column

## DATA CLEANSING

The data cleansing format took the process of

1. Making a copy of all the original data
2. Highlight each issue, Define the mode of cleaning, Write the code to clean and test after cleaning is done to see the effect.

### STORING

After the cleaning process is done the cleaned data, i stored the data in the twitter_archive_master.csv

In [ ]: