# *Data Wrangling Report*

## Introduction

This project is part of Udacity Data Analyst Nano Degree course, data wrangling section. This project is primarily focused on wrangling data using Python. Gathered data from different sources. The archive contains basic tweet data. Each tweet image was run through a neural network . The prediction were downloaded programmatically. Queried data using Twitter API

## Data Gathering

Gathering data using Twitter API was the biggest pain.

The first dataset that was meant to be downloaded manually was relatively easy to import. This was done by clicking on the hyperlink provided to me, moving the data into the relevant directory and importing using Python's Pandas package using the pandas.read_csv() function.

The next dataset was meant to be downloaded programatically using the Requests package. The package takes a URL and saves the response to a variable which can then be saved or written to a relevant file before being imported again withPandas.

The final dataset was not trivial to download. I used some code from another repository.

## Data Valuation

Imported data programmatically and assessed it. Found few issues related to data quality and data tidiness. Many of the columns have invalid names, made changes appropriately. Had to change the data types for few of the columns to string. There were missing values which I cleaned in the data. I have cleaned data both manually and programmatically. As part of the visual evaluation, had removed few images which are not related to dogs. Cleaned few None sting values..

The provided Twitter archive lacked some useful information: retweet count and favorite count. I used the tweet IDs to query the Twitter API for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire set of JSON data in a file called tweet_json.txt. I then read the txt file line by line into a pandas DataFrame only including the desired variables; retweet count and favorite count.

## Project Details

For this project, we took original ratings from WeRateDogs (no retweets) that have images. Not all of the original tweets in the dataset are dog ratings and few are retweets.

Only a subset of issues (eight quality issues and two tidiness issues at minimum) has been assessed and cleaned.

**Project Tasks:**
- Data wrangling, which consisted of:
  - Gathering data
  - Assessing data
  - Cleaning data
- Storing, analyzing, and visualizing the wrangled data
- Reporting on my data analyses and visualizations (act_report.pdf)

**Conclusion**

Once I finished data cleansing, created csv files using pandas and sqlalchemy. The CSV creation was much less interesting, that was just done using `pandas.DataFrame.to_csv()`. The Project all together gave me a very good understanding of data wrangling, using API and JSON data. I also understood importance of data wrangling before we actually go for visualizations.

**Author**
Pallavi Bodepudi