# Wrangle Report

Mzoon Alkadi – Jan 2021

## Introduction

In this project, I learned and practiced data wrangling and analyzing as part of the Udacity Data Analysis Nanodegree program. The wrangled and analyzed dataset that is the tweet archive of the Twitter account WeRateDogs that rates people's dogs with a humorous comment about the dog. I used Python Prgramming Language and documented it using a Jupyter Notebook (wrangle_act.ipynb).

This report briefly describes my wrangling efforts in this project.

## Project Details

The wrangling process covers four main tasks:

- Gathering Data
- Assessing Data
- Cleaning Data
- Storing, Analyzing, and Visualizing Data

### Gathering Data

The data of this project consists of three different datasets obtained from different sources which are:

- **WeRateDogs Twitter Archive file** that contains data such as id, text, and stage of more than 5000 tweets. I downloaded the enhanced version provided by Udacity manually.
- **Tweet Image Prediction file** that is hosted on Udacity's servers and I downloaded it programmatically using the Requests library via URL information.
- **Twitter API JSON file** that includes additional data of tweets e.g. retweets and favorite counts.

## Assessing Data

I assessed the data visually and programmatically to understand the data and check for any issues.

### *Visual Assessment*

I used two ways to assess the data visually. The first one is by opening the CSV file in excel to scroll through data and observe. Secondly, I printed a sample of five rows from each dataset using python in jupyter notebook.

### *Programmatic Assessment*

I used several predefined functions in python to help me explore and check the datasets for example: info() , describe() , value_counts().

### *Quality Issues*

After exploring and checking the datasets, I observed nine quality issues that needs to be handled and fixed then I categorized them based on the dataset.

**Twitter Archive:**

1- Retweets are included.

2- The Source column Values should be more readable and classified.

3- The name column includes wrong values such as 'a' and 'not'.

4- Many None Values in the name, doggo, floofer, pupper, and puppo columns.

5- Unnecessary columns exist.

**Tweet Image Prediction:**

6- Duplicate values for jpg_url column.

7- Unnecessary columns exist.

**Twitter API Data:**

8- Retweets are included.

9- Unnecessary columns exist.

### *Tidiness Issues*

1- Merge the doggo, floofer, pupper, and puppo columns into a single column.

2- Merge the three datasets into a single column.

## Cleaning Data

In this step I created three copies of the datasets to clean them. Then I began the cleaning process by handling all of the nine quality issues in addition to the tidiness issues. For each issue I used a clear format by firstly defining the issue then the code and lastly testing to make sure that it has been fixed properly.

## Storing, Analyzing, and Visualizing Data

Finally, I stored the resulted master dataset into an external CSV file that is called 'twitter_master_ds.csv'. Then I analyzed the data and obtained a total of four insights with visualization of their results which I document in the 'act_report.pdf' file.