Wrangle Report

1. Introduction:

This report contains a description of my efforts and challenges faced while wrangling WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations.

Steps involved in Data Wrangling:

- Gathering data
- Assessing data
- Cleaning data

2. Gathering:

We gathered three pieces of data as described below in a Jupyter Notebook titled wrangle_act.ipynb:

- `twitter_archive_enhanced.csv`: We manually downloaded this file by clicking the link provided.
- `image_predictions.tsv:` This file is hosted on Udacity's servers and was downloaded programmatically using the Requests library.
- `tweet_json.txt:` Each tweet's retweet count and favorite ("like") count at minimum, and any additional data was downloaded using Python's Tweepy library and stored each tweet's entire set of JSON data in this file.

Querying the Twitter API was a challenge task. I looked through the support documentation for the Twitter API and it had many helpful functions, especially for people who are trying to learn how an API works for the first time.

3. Assessing :

After gathering each of the above pieces of data, I assess them visually and programmatically for quality and tidiness issues and cleaned them by following data wrangling template provided by Udacity.
I audited the dataset by checking data types, value counts, number of null entries, and numeric summaries. Data for few tweets was missing. All 3 sources were scattered. Dataset had multiple columns for dog stages which was unnecessary. Some of the names of dogs were wrong. Also the ratings of some dogs were recorded incorrectly due to presence of other numerical data in text column.

I also noticed some extreme outliers but by further investigation, those were just humorous ratings and trust me they were actually quite funny and deserved that rating.

4. Cleaning:

It is where I fixed all the issues reported in assessing part. I used two types of cleaning, the manual and programmatic. Before making any changes I was recommended that I should keep a copy of original dataset first so did just that. I followed the template process i.e. Define the issue; write the code to clean the issue followed by testing the issue. I iterated a lot of times to spot exact issues and added them in assessment notes.

5. Conclusion:

Real-world data rarely comes clean, so Data Wrangling is a core skill that everyone Data Analyst should know about. Clean data provides better visualizations and in process of cleaning you also learn about the data.