Project Data Wrangling of WeRateDogs Twitter Archives: Wrangle Report (Internal Document)

1. **Goal**. For this project, my goal was to wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations.

2. **Background Information**: WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

3. **Scope of Data**. To achieve this goal, I performed data wrangling on three separate sets of data, namely:

    a. A condensed set of the WeRateDogs Twitter archive, curated by Udacity. This contains a summary of about 5,000 tweets from 2015 to 2017.

    b. Image predictions archive. This contains a set of results arising from injecting the photos in the WeRateDogs Twitter archive into a neural network. The network's algorithm specializes in producing a set of predictions (confidence) as to whether the photo contains / does not contain an image of a dog.

    c. Additional data obtained by web-scraping the WeRateDogs Twitter website. This was achieved through querying the Twitter API for each tweet's JSON data using Python's Tweepy library.

4. **Tasks**. I completed the following tasks in this project as part of data wrangling:

    a. Data Gathering. The initial set of WeRateDogs Twitter data was provided by Udacity as a downloadable csv file. Following that, I obtained the image predictions archive through downloading it programmatically (via the Python Requests library) from Udacity's servers. Lastly, I gathered additional data by web-scraping the WeRateDogs Twitter website. This was achieved through querying the Twitter API for each tweet's JSON data using Python's Tweepy library.

    b. Data Assessing. After gathering each of the above pieces of data, I assessed them visually and programmatically for quality and tidiness issues. I detected and documented about eight quality issues and two tidiness issues. Some of the Pandas' library methods used were DataFrame.head(), DataFrame.info() and DataFrame.describe().

    c. Data Cleaning. Next, I cleaned the issues I documented while assessing the data. The result is a high quality and tidy master Pandas DataFrame of about 2000+ rows. I also stored the cleaned DataFrame in a CSV file locally.

    d. Data Visualisation. Lastly I produced five visualisations from the cleaned DataFrame.