# WRANGLING REPORT

20 May 2019
Project Reviewer
Wrangle and Analyze Data
Data Analyst NanoDegree
Udacity

Dear Project Reviewer,

This project was focused on gathering and cleaning data obtained from a number of sources, including CSV and TSV files, the Twitter API and JSON-formatted flat files. The original dataset had over 5,000 records of twitter posts by the twitter channel specialized in rating dogs in a fun and relaxed way: WeRateDogs (@dog_rates).

In order to gather, assess and clean the data programmatically and visually, Python programming language was our main tool, along with a few of its data-centered open-source libraries and some auxiliary ones (pandas, numpy, requests, matplotlib, and tweepy, to name a few). The main wrangling and cleaning file was created using Jupyter notebook, which allow for blocks of code, text and charts to be combined in one single document.

Although the tweets' archive and dog breed predictions by a neural network were provided in TSV and CSV files, we had no information on the popularity of the tweets. In order to allow for a more thorough analysis in the future, such data would be important, and that is where the Twitter API played a part.

After authenticating and obtaining the necessary developer approvals by Twitter, the company's API (Application Programming Interface) would allow us to programmatically extract useful data related to each of the tweets listed in the existing CSV and TSV files provided.

We proceeded with the registration of our application and, once approved and authenticated, we were able to develop code (thanks to Tweepy library) to query each tweet's id and write its JSON data to a text file, taking special care to write each tweet's data on its own line.

Following that, we developed code to read the file line-by-line and create an enhanced table that will allow us to perform various analysis in the future, called pandas DataFrame.

When extracting data from one or more sources, it is considered essential to assess the quality and structure of the obtained data before any useful analysis can be performed. In this case, we noticed over 13 issues related to quality and two regarding structure (or tidiness).

As an example, the quality issues included inconsistency of data format, the presence of outliers and unidentified null values and variables that needed rearranging in a number of columns. All of the issues were duly addressed and tested through code, resulting in three clean tables which were then combined (*merge* pandas function), in accordance with the widely-accepted rules of tidy data, so that we could extract initial insights and assess whether a more complete analysis should be performed.

Yours sincerely,

Alex Paz, Udacity student