# wrangle_report

May 30, 2018

## 1 Wrangle Report - Udacity Project: Wrangle and Analyze Data

by Shawn Gong

---

### 1.1 Introduction

---

The project Wrangle and Analyze data requires me to gather, access and clean the data based on datasets about dog ratings using Python language and external libraries including `pandas`, `NumPy` etc.

### 1.2 Gathering Data

---

Although it is the very first step of data wrangling, it sometimes could be the most tricky part of data wrangling in my opinion. Analysts often need to obtain data from different sources using various methods, and the quality of these datasets can be widely vary. The project requires us gather data using three different methods: manual download, download programmatically and using API to access the data on Twitter.

The first dataset is the general information of the tweets I downloaded manually. Although the data was previously cleaned, there are many visible quality and tidiness issues that can significantly affect further analysis process. Loading the dataset is pretty easy: After putting the dataset into the working directory, I simply used the function `pd.read_csv` to load it into a dataframe.

The second dataset is locately online, which I need to retrieve using `requests` library. After retrieving it from the url, I then write it to a `tsv` file, which stands for "tab seperated value". I then load the data into a dataframe using pandas.

I found the last dataset being the most tricky one. The thrid dataset requires me to retrieve data using Twitter API. `tweepy` library provides a pretty intuitive solution, but I have to spend a bit more time to understand the mechanic of the library and how to export a csv based on the library.

## 1.3 Assessing Data

---

I utilized visual and programmatic assessment in the assessment process. I mainly conducted the assessment in jupyter notebook. I divide the issues by quality issues and tidiness issues. Quality issues is about content and tidiness issue is about structual problem.

### 1.3.1 Visual Assessment

I simply used jupyter notebook for visual assessment. There's not much function involved because it is visual assessment. I directly observed the dataset themselves, which revealed most of the problems and interesting facts about the dataset, there are also things cannot be revealed by programmatic assessment. For example, most of the observations have empty dog_stage, but in the programmatic assessment, the info of the dataset suggests that there is no missing value for these observations. I would not know that these values are actually empty without assessing the dataframe by eye sight.

### 1.3.2 Programmatic Assessment

The power of programmatic assessment is that programmatic assessment can reveal hidden facts that are not presented directly. I used programmatic assessment to find the wrong data type of each variable by using `.info()` method. I also uilized programmatic assessment to filter data based on specific values of interest. I filtered out the retweet and response using the `DataFrame.loc()` method.

## 1.4 Cleaning Data

---

From the Udacity course, I learned to conduct data cleaning based on a specific order: Deal with completeness issue first, then the tidiness issue, finally dealing with rest of the issues. In practice, I find that this rule is a general guidance, there's more logic behind data cleaning however.

Take this project as an example, the data type of `tweet_id` variable needs to be fixed before merging the `df_info` and `df_stat` together, mainly because the `tweet_id` is the key parameter for merging the dataframe. I also have to replace all the 'None' to null value before merging the four dog stage variables, as the 'None' values are not null, which will make the merging fail. The previous two examples are all "Dealing with other quality issue before fixing the tidiness issue", despite they are not very significant at first glimpse, the problems really become crucial afterwards. This kind of situation is not explained by the general rules, but can be easily identified with the accumulation of experience.