Wrangle Report

By Huy Ngo

The Data project assigned by Udacity is a great resource for me to learn about the data gathering process, visualizing and using Twitter API. I am very thankful my Udacity mentor for his effort in helping me noticing my errors in every part of the project.

The project started by collecting data from three different sources. Twitter page WeRateDog has provided Udacity student with exclusive access to their tweet database, which has many data columns to work on including tweet_id, timestamp, text… However, most of the columns seem to have their own problem in quality, accuracy, and tidiness. It is the reason why I need to clean the database after gathering it from the project source.

At first, I discover that many columns are set with the wrong type, which will have a bad effect on the visualization as they can not be read and calculate correctly. It is pretty simple to fix with a few commands line to set them back to the correct type.

Secondly, I start noticing that many data are having bad quality. They may have been inputted wrong value, missing value or having redundant character. It is not a big deal as I have a loop through some of the related columns to extract the right value for that. More ever, I also rename columns to make them easier to read for the audience.

Last but not least, I have joint all three different tables in order to make managing them more effectively. Some column is even jointed to one as they should display the same content with a less complicated table.

In conclusion, wrangling data is the most important process in Data Analysis as its goal is to provide a reliable data set that we can trust on as well as using for an accurate visualization.