Project report – Wrangling Effort

By Tu Duong

The project requires all wrangling efforts: gather, assess, clean. The final purpose of the project is to draw interesting insights via analysis and visualization.

1. Gathering

There are 3 different sources of data: WeRateDogs extended archive, data from a file from internet, and data from Tweeter API. While the first two source were easy to gather, the Tweeter API data was a challenge. The challenge come from two issues: the tweets reported in the WeRateDogs were no longer available from Twitter. Perhaps they have been deleted). After this error was treated, the second issue come up: the rate limit. After spending a long time of research and trying various method, I came back to the original solution – ignoring missing data error, and set wait_on_rate_limit=True and wait_on_rate_limit_notify=True). It turns out the solution was not that complicated and it worked well. I should have asked for help instead of spending hours trying different solutions. Documentation from Tweeter was not clear enough for me – both about various limit concept and how to overcome them.

Furthermore, I spent time to investigate what additional interesting information apart from favorite counts and retweet counts that I can get from Twitter API. I also explored what are the additional information I can get if I gather extended Twitter information. I think the information about the source of the tweets could be interesting and worth further investigation.

2 Accessing

In accessing, I inspected the data visually by looking at the data frame to identify data problems.

I also use info, describe, and value counts to have an overview of the data frame. By this way I can see the outliers like the minimum or maximum value.

Next, I assess the data tidiness. Please refer to the Jupyter note book for the details.

3 Cleaning

Cleaning step followed the assessing steps. There are many cleaning activity, but most importantly I dropped the dog stage group because the not many observations have a dog stage value. Another important cleaning step was to join all three tables and create a master dataset ready for analysis.