Wrangling Effort

In the wrangling process, I chose the divide and conquer method and finished the gathering, assessing, and cleaning with iterations stepwise.

First of all, the gathering process took a lot of efforts. There are three files needed to be wrangled, twitter-archived-enhanced, image_predictions, tweet_json. The twitter archived was given by Udacity, so I didn’t need to worry. Then I used Python request library to save image-prediction.tsv and read it into Pandas dataframe. Then came the hard part. It took me about an hour to register for the developer account and download the whole dataset. Then I only fetched certain columns from the status file and converted them into json.

The assessing part is a lot of fun. I first visually assessed the three tables. When I saw too many None or NaN values on a column, I would mark and decided to delete them later since they didn’t contain much information. Then I saw some crazy rating with denominator not equal to ten. This is when I switched to programmatic assessment. I used describe and info to see the variables and distribution of numeric features. I also saw some tidiness issues like a variable is divided into multiple columns, which make future analysis more difficult. I marked all the issues and now came the cleaning stage.

Cleaning is a lot of headache. I managed to do fix some quality and tidiness issues stepwise. I looked at the tweets with dogs with no names and assessed them. If I could find name in the text, I would modify the name. Otherwise I would delete them. I also deleted entries where the denominator of rating was bigger than 10, along with some ratings wrongly fetched like 24/7. Then I tackled the tidiness issues in the prediction table by melting the table three times and join them by the prediction rank. This is the last step of the wrangling process.