wrangle_report

June 5, 2020

This project is to do wrangle and analyze on the tweet archive known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. Data wrangling was held in 3 steps gathering, assessing, and cleaning.

At the gathering step, I collect 3 datasets. The WeRateDogs Twitter archive by manual download, tweet image prediction using python requests library, and tweet's retweet count and favorite("like") count collected using Twitter API. Each dataset was stored as CSV, TSV, JSON format included txt.

At assessing step, I used python pandas. I changed each dataset to pandas DataFrame data structure and assess by visually and programmatically both. In the visual step, I looked at the records and try to find out some problems and also tried to understand the dataset properties. At programmatic step, I used useful pandas commands like info, head, sample, describe, and look at the dataset in more detail. For example, the info command helps us to look at a high-view of the dataset also shows us the data type of each column so in a short time we could skim fast.

I looked at the dataset from 2 perspectives. Tidiness and quality. There were 3 tidy issues and 13 quality issues. I didn't dive too deep to make a perfect observation but checked the issue that seems critical. Multiple variables are stored in one column, variables are stored in both rows and columns and a single observational unit is stored in multiple tables were tidiness issues and most quality issues converge to the wrong datatype, wrong data, missing data. In case of quality issues, I investigated the basis of 4 data quality dimensions completeness, validity, accuracy, and consistency.

As a cleaning step, I had to clean up the tidiness issues and quality issues found in the assessment step. First, I cleaned up the missing data because it is the most important issue. If we don't clean up the missing data it will be hard to handle further problems. And meanwhile, because the tweet_id was the most important column that will act as a key to join every table not only I did clean missing data I corrected the tweet_id using expanded_url. There were also duplicated rows of tweet_id after correcting so cleaned it up also.

After missing data was cleaned up I handle the tidiness issue. Because tidy datasets with data quality issues are almost always easier to clean than non-tidy as they have data quality issues. When we fix the quality issues first we have to do it again because when we fix the tidy issues which are structural issues we get another quality issue! When we handle tidiness issues pd.melt, pd.merge is very handy.

For quality issues, a Nullable integer type has mainly occurred. There was a problem that integer type cannot handle NaN so it automatically converted to float. So I had to use some tricks using pandas integer array. Also, erroneous datatype was an issue. Many columns that should use the bool type or int type was using object type. Datetime should not be an object and some columns like source could be much better with category datatype. Removing columns that have the same purpose as other columns were also important. For others, I had to fix an inconsistent format of name or predictions. Pandas str method was a very powerful overall handling quality issue. With regex expression str method give us tools to replace, extract, strip and indexing, etc.

After the data wrangling, only one master dataset left and for the datatype issue, it was stored to pickle format.