Wrangling Project Wrap-Up

August 31, 2022

The data wrangling process is made up of three stages, each of which presents a unique challenge. This project entailed working through the whole process.

For this project, the gathering part was straightforward. The data was prepackaged, mostly in the form of everyday, flat files. One file had values separated by commas (CSV), and the other had values separated by tabs (TSV). Both were read with `pd.read_csv()`—the latter with the `sep('\t')` parameter added in. There was also a JSON, which was likewise read with `pd.read_json()`.

The assessing part was a bit more involved. Across the three tables, there were 32 columns, which were parsed using both visual and programmatic assessments. Conclusion? Most of the columns would need cleaning or tidying. But addressing all of them would have been outside the scope of the assignment.

As such, the issues needed to be prioritized: before the data could be validated, it had to be checked for completeness—if missing values were to be imputed afterwards, the data as a whole would have to be re-validated; the data was also tidied before validating because tidy data is easier to validate.

The cleaning part was laborious. But using the Define-Code-Test (DCT) framework made things easier. Defining the issues in actionable terms transformed the list of problems into a blueprint for cleaning. Testing the code immediately allowed for a speedy, iterative approach to cleaning. If the code did not resolve the issue properly, then it could just be rewritten—no deliberation required! This approach proved especially useful when dealing with RegEx patterns. All in all, addressing the issues one by one made the assignment less daunting.

Data wrangling is a lot of work. Going about it blindly is not cost-effective. During the assessment phase, time should be spent not only assessing the quality of the data but also its relevance to the upcoming analysis. Additionally, the cleaning phase ought to be punctuated with periods of reassessment.