Review of *Tidy Data*

Grace Yi Chen

This paper talks about tidy data with detailed examples showing its importance and its construction. Data cleaning usually takes a large proportion of time in data analysis. One of the challenges is to structure datasets to be tidy so that they are easy to manipulate, model and visualize. A dataset is a collection of values and each value belongs to an observation and a variable. The structure of a tidy data is to make each variable as a column, each observation as a row and each type of observation as an independent table. The author uses real-world examples to demonstrate how to format messy datasets to be tidy. In particular, the author introduces tidy tools, which are tools that could input and output tidy datasets. The author discusses how tidy data and tidy tools together can make data manipulation, visualization and modelling easier using R packages like *plyr*.

I think the authors showed strong evidence that making data tidy is important for efficient data analysis. As the author mentioned, 80% of the data analysis time is spent on data cleaning. I have the same feeling in my own research experience. In many cases, my collaborators will provide spreadsheets that are formatted in every imaginable way. These messy formats may be easy for data entry, but they are not efficient for data analysis. I need to spend extensive efforts on data standardization so that these datasets are compatible with the functions I used in R. This article is very helpful as I learnt a lot from it about different tidy tools and how to make tidy data using the tools. The author and other tidyverse advocates also create a lot of packages and functions in R that makes data cleaning easier and more efficient. Also, the author mentioned that all the tidy data framework he proposed here are based on his real-world experience and there are few theoretical principles to guide the design of tidy data. I think this is an important part in data science research and we need to put more efforts in this.

Question:
Some people think “Tidyverse should be considered advanced R, not for beginners”. What should be a good way to start learning R now.