Review of ***Tidy Data***, submitted by Hadley Wickham to ***Journal of Statistical Software***

Overview: This paper makes a valuable contribution in an area that needs more attention than it has received: practical strategies for bringing data to the point where statistical explorations and analyses can begin. Wickham has thought hard for many years about the issues this paper addresses, and his insights and conceptualizations flow from his vast experience with a broad range of real-world data sets. "Tidy Data" is Wickham's name for what this reviewer calls a dataframe (as distinct from data table or data set). But whatever the name used, Wickham's definition and his systematic identification of the most common ways that data sets are "untidy" or "messy" serve as a foundation for restructuring a data set that make it amenable to statistical analyses in many forms. The author then introduces the concept of "tidy tools", tools that take tidy data as input and return tidy data sets as output. This conceptual framework is effective in placing the tools of Wickham's **plyr R**package in an intellectually cohesive context. The comparisons of commands that are functionally similar, across base **R** and **plyr** (and in one section across **SAS** and **SPSS**) are effective and instructive for the reader.

Exposition: At a global level, the exposition is first-rate. The organization seems very natural and nicely guides the reader who may be new to some of the terminology and the kind of thinking that underlies a high-level view of data structures. Few having any experience with the analysis of real data sets could fail to recognize at least a few of the challenges and difficulties that Wickham describes in this well-organized framework. The examples are well-chosen and illustrate well the general concepts associated with moving from messy data structures to tidy data. The use of **R** code from the base **R** package and from **plyr** or other supplemental packages, sometimes in parallel, provides concrete and convincing illustration of the practical value of Wickham's conceptual scheme.

Details of writing: There is opportunity for improvement and for strengthening the exposition at a local level – in sentences and phrases. This reviewer has just read Wickham's 2011 paper in the same journal on "Split-Apply-Combine", and the writing in that paper is good and can serve as a model for polishing the present paper. The differences between the two are in the details, and a thorough editing and tightening of the present manuscript would almost surely pay dividends. Without trying to be comprehensive, I will illustrate the kinds of small changes (including just fixing typos) which taken together would have a positive effect.

Abstract: "take in and take out" → "input and output"

p. 1 up 6: "subset" → "component"

p. 1 last full sentence: "The reorganization makes…because it conforms to a standard that facilitates well an initial exploration and analysis of the data; you don't…"

p. 2 line 2: Is "reformulating" an approximation for "munging"?

p. 2 line 9: → “an extension…”

p. 2, 5 up from subtitle: “…techniques with real examples.” -- could do nicely.

p. 2, 1 up from subtitle “…misses and what other approaches might be fruitful to pursue.”

{Tolstoy quote is effective and serves nice conceptual purpose!}

Section 2 is substantively excellent and is very readable. So is the lengthy Section 3!

p. 3 2/3 down “…were were…”

p. 3 bottom. Consider having a new bold subtitle as a guidepost for the reader:
Defining tidy data:

Examples in Section 3 are just right for this treatment! Good!

p. 10, up 2. Sentence is garbled.

p. 11 bottom: (left) and (right) may not be the apt descriptors…

p. 12, line 9: comma after file name?

p. 12: This reviewer would prefer to avoid “hopefully” about 13 lines down. More substantively, the sentence is garbled.

This reviewer greatly appreciates the “four fundamental verbs of data manipulation.”

p. 13 middle “…by the by preposition.” Should second “by’ be **bolded**? Or in quotes” Same question arises again.

p. 14 last paragraph. A difficulty is identified; does the author suggest a solution? Reader isn’t clear where we are left as we enter the Case Study.

p. 17 middle → “ disease**s** we work with have…”

p. 18 nice plots in an interesting context!

p. 19 last line “…seem like they should …” → “may”

p. 20 very effective use of exploratory data analysis

p. 21 down 13: might be clearer if “…an efficient equivalent to **join**.” – bolding or perhaps quotes.

Wickham's generous and gracious attitudes evident in the Discussion section are appreciated. This reviewer would be quite content if data analysts and statisticians could standardize on the useful conceptual framework that is provided in this nice paper!

This reviewer notes (in a spirit of collegiality) that the best writers about data analysis (from a previous generation) – Fred Mosteller, John Tukey, David Hoaglin, Paul Velleman, Frank Anscombe (as illustrations) – correctly used **data** as plural form (and occasionally **datum** as singular). From that perspective, "the data is …" grates more than a little … but regrettably this is a battle that is now lost. Sigh… ☺☺

Congratulations on a fine contribution!