Review of Tidy Data by Hadley Wickham

Summary

The manuscript touches upon an important subject that rarely gets the attention it deserves in the area of data analysis -- data cleansing, and the author does a great job of formalizing a set of rules for best practices. The author provides specific, real-world examples to help illustrate his points, which makes the discussion to-the-point and clearly relevant to any responsible data analyst.

My main issue with this manuscript is the lack of linguistic finesse. I would recommend the author reformulate some sentences and fix up grammar mistakes, where necessary, before resubmitting for publication.

Major Points

While the topic is important and the message is clear, I found the tone of the paper to be too conversational. The author overuses contractions, ambiguous pronouns ("It's often said..."), and casually addresses the reader in second-person form.

The author also switches between the use of italics, boldface, quotes, and the LaTeX verbatim environment constantly, which tends to be distracting at times.
- The author uses "colvars" in italics when first naming the terminology, and then subsequently mentions the Pew dataset contains one colvar, without italics.
- The author does not use formatting on the column names in the caption for Table 9 but does use \verbatim to reference column names in Table 10's caption.
- On page 13, middle of the page, it would be helpful to use quotes around "by" where applicable, as "modified by the by preposition" does not read well.

The author overuses colons. In some cases, I found hyphens to be more appropriate, and in others, I would think semicolons might be more suitable.

Some tables also have unreadable characters in them, such as the degree sign in Table 12's top panel (row 6's artist). Same issue with Table 3's 5th religion. Table 12's caption also describes there being a left and right dataset, whereas the subtables are clearly placed above and below.

The last sentence in the first paragraph of the discussion needs to be rephrased for clarity.

Line-By-Line

- Abstract: "variables are stored in columns, observations in rows, and a single type of..." -- this list isn't logically homogeneous.
- p.5, 3.1 -- "The Pew Center is an American ... that collects data on attitudes to topics RANGING from religion to the internet"

- "Billboard" should always be capitalized as it is a proper noun.
- p.13, section 4.2 - "Tidy visualization tools ONLY NEED to be..."
- p.16, 4-th line from bottom: "Next, we work out THE overall..."
- p.16, 2nd line from the bottom: "Then finally, WE join.."
- p. 17, "To ensure that the diseaseS we work with..."
- p.19, "The causes of death fall INTO three main groups: ... " There should be a hyphen between "transportation" and "related".
- p. 19, 2nd paragraph of discussion: "This makes it easy"; remove "is".
- p. 21, last sentence of first paragraph: "and A BETTER KNOWLEDGE OF how we can best design tools..."
- p. 21, last paragraph before Acknowledgements "verifying experimental design, AND filling in..."
- Author information, last page "Adjunct ASSISTANT Professor"