Response to reviewers

Submission ID 217815520

We thank the Editor and the three reviewers for their comments, which have helped substantially to improve the manuscript. In response, we have made considerable changes to the main manuscript as outlined below.

Major changes:

Reviewers suggested reframing the problem space so there is more emphasis on the case study.

Reviewer 1: How is this manuscript relevant for teachers and/or students in statistics and data science?

Reviewer 3: The authors prominently cite (Chatfield 1985) for the definition of the term "IDA". However, Chatfield fairly explicitly defines IDA to be a data summarizing and scrutinizing process; not a cleaning, scrubbing, or munging process. It seems curious to me that the whole paper is framed as being "IDA" based on the author's unsupported statement that data cleaning should be considered party of IDA.

Reviewer 3: My suggestion is that the authors reconfigure this paper to be presented as a pure case study. Ibelieve the work that has been done is valuable: the data cleaning process is often messy andunplanned, so a case study like this on a popular dataset would provide an excellent educational resource.

Supporting information for the lack of data cleaning in papers.

Reviewer 3: The authors also emphasize that "There are few research papers that document the data cleaning". While I agree that data cleaning is under-emphasized and under-documented, it's hardly true that there are not papers on the topic. A quick Google Scholar search for "data cleaning" unearths many scholarly papers and books on the principles of this process, none of which appear to be cited in this work.

Outlier detection

Reviewer 3: Section 3.2.1 goes into detail on the authors' approach to characterizing erroneous outliers in the data. Their approach is reasonable, but it is not (as far as I can tell) based on any established or tested procedure. This section once again presents a tension between the paper as a case study versus a topical commentary: if it is a case study only, then the "common sense" justification for the approach is appropriate, but if it is meant to establish future norms, these modeling choices must be more formally supported. The reference to a "reasonable degree of fluctuation" particularly struck me as a subjective or "ad hoc" claim. I am particularly a bit concerned by the statement on page 13: "The robust mixed model could be the best model to be employed in this case. However, this method is too computationally and memory expensive, especially for a large data set, like the NLSY79 data." Surely, fitting

a mixed-effect model to a dataset with only one predictor and 1188 individuals ought to be very feasible?

We thank the reviewer for pointing out to us that the data contains information on the work experience – this additional data is added in the extraction of the data in Figure 1. This information was compared with Singer & Willet's original data. Comparison shown in Figure X shows that the information does not match up, however, there is a high correlation between the variables.

Minor changes:

- We have corrected all minor grammatical errors pointed out by the reviewers.

- We have made the descriptions of the steps more explicit.

Reviewer 1: It's unclear from context what is being referred to from the Huebner et al 2020 citation.

- We have fixed the sentences and the reference.

Reviewer 1: What is dplyr used for? Are there tidyverse packages used but not mentioned? If not, splitting this sentence into two may make it clearer which package is used to what purpose. Also, I haven't tried it in code, but I imagine that the data could be tidied as described just using pivot_longer, at least for the job number, year, and wage data. In other words, I don't see why dplyr or stringr would be needed for the data described on page 6.

- We have split the sentences and described the use of each package mentioned.

Reviewer 1: please clarify. "If either the hourly wage or hours worked is missing, we do not tally this." I take "this" to mean "number_of_jobs". But in row 2, total_hours is missing, yet number_of_jobs is 1. The number of jobs (1) was tallied even though hours worked was missing.

- We have clarified this that we only tally the number of jobs if the hourly wage is not missing.

Reviewer 1: "to find the anomaly in the wages values" This comes as a surprise to the reader. What anomaly? The wage anomalies should be mentioned in the introduction so as not to be a surprise at this point in the manuscript.

Reviewer 1: I don't know that Table 1 is necessary as I don't have anything to compare it to. If the numbers check out, a simple statement to that effect would suffice. Also, what does "(?CHECK)" mean? This needs to be resolved.

- We have removed '(?CHECK)' in Table 1's caption. We also have stated in a sentences after the table that Table 1 reflects the consistency between the data we had and the database.

Reviewer 1: for example, ID 39 experienced an unusual wage only in one year.' That's one way to fix this sentence, which needs to be rewritten.

- We have rewritten this sentences.

Reviewer 1: How many individuals up to 8th grade are in the refreshed dataset? What are plausible reasons for these differences?

- There is no individuals up to 8th grade in the refreshed dataset as we only filtered the high school dropouts (started from 9th grade ,i.e., high school grade). The plausible reasons for the differences are we might have different definition of high school dropouts with the original data and the corresponding individual back to high school after they dropped out so that their highest grade completed is 12th grade.

Reviewer 1: How many individuals up to 8th grade are in the refreshed dataset? What are plausible reasons for these differences?

Reviewer 2: My primary comment is that many elements need to be more explicitly spelled out.