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 a major revision of the main manuscript as outlined below.

Major changes:

The editor and reviewers suggest reframing so that readers can help their students to experience the journal from wild to textbook data and either emphasize the case study, for the paper to be a regular submission for JSDSE, or to better explain the relevance for teachers and/or students in statistics and data science to be considered for the special issue. We have chosen to do the latter, because we hope that the paper fits into the goals of the special issue.

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

The introduction has been re-written with an emphasis on how this work should be interesting for teachers and students of data science and statistics.

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.

Yes, there is grey area in these three activities. Thanks for pointing out Chatfield's original viewpoint, which differs from Huebner et al's. We have revised this paragraph to de-emphasize cleaning as being part of IDA. We have kept the explanation of the terminology of IDA and EDA because we think this is important to clarify for teaching purposes.

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

We have kept the case study frame of mind but have re-cast the paper so that it focuses on the importance of this type of activity for teaching of statistics and data science.

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.

XXX We need to review this literature. It's not true that we have ignored it, the main reference is Dasu! But we need to decide whether there is more that we shoudl cite.

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. XXX Which one?

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: what does the natural log have to do with the y-axis?

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

Reviewer 2: I would like more discussion about the textbook selected as the exemplar. It sounds like this dataset is used in many books, why choose that one? The text should include the name of the book, so readers do not need to flip to the citations in order to learn what it was. From the name, I am guessing that the book focuses heavily on this dataset, which is why it makes sense to reproduce their subset. Be explicit about this.

Reviewer 2: When listing the files that come zipped from the data source, be more explicit about the file formats. Comma separated values instead of csv, etc.

- We have made it explicit.

Reviewer 2: Before the data cleaning section I would love a glimpse into the future to see what the target dataset would look like. What are the observations? What are the variables? This would allow me to follow the narrative of data wrangling more easily. Perhaps this could be put into section 2.2, which is titled Target data. Then, when starting to describe the raw data in 3,1, explain the same pieces and explicitly explain why the raw data it is not tidy to begin with.

Reviewer 2: Before the data cleaning section I would love a glimpse into the future to see what the target dataset would look like. What are the observations? What are the variables? This would

allow me to follow the narrative of data wrangling more easily. Perhaps this could be put into section 2.2, which is titled Target data. Then, when starting to describe the raw data in 3.1, explain the same pieces and explicitly explain why the raw data it is not tidy to begin with.

- We have added the explanation of the observations and the variable in Section 2.2.

Reviewer 2: With the IDA about wage data, I would appreciate a bit more context. Why sample 20 observations? I am not a longitudinal data expert, so I was a bit surprised when I flipped to the plot and found it to be small-multiple time series, although upon further consideration it makes sense. In that section it would be good to describe what the figure shows in a bit more detail. Perhaps "We randomly sample 36 respondents from the data, and plot their average wage as a time series. Looking at the patterns over time, we see a lot of variability in wages. For example, the people shown in panels 5, 7, and 11 have [explain what is interesting here, again I'm not an expert]." The next sentence here, "Some have had flat wages for years but had a sudden increase in one particular year, then it gone down again, while the others experienced an upsurge in their wage, for instance, the IDs in panel 9." does not seem to provide much additional information. What were you looking for here? Overall trends over time, increasing/decreasing? Any places where one year's wage looked really different?

Reviewer 2: I think there is a narrative step missing here, which is that looking at those samples led you to look at plots that show values for the wage variable over all participants and all years. The phrase "the summary plots" was not enough description for me to know what to expect when I flipped to Figure 3. Please be explicit about what the three plots were, why they were made, and what you were looking for. I think there is another narrative step missing about Figure 3C, which is that after you saw there were outliers you tried to identify which respondents had the extreme values. Again, please spell this out in the text. When you return to the plots for the entire dataset in Figure 5, please make parallel comments.

Reviewer 2: The next paragraph, which begins "The anomalies are also found" should perhaps be "Similar anomalies were found." Was it the same respondents with the extremely high wages that showed extremely high number of hours of work? I suspect not, but please be explicit.

Reviewer 2: Section 3.2.1 begins "As part of the IDA, which is the model formulation, we build a robust linear regression model to treat the extreme values in the data." This could use clarification. In 3.2, you said that part of IDA is " model formulation without any formal statistical inference." To me, that sounds like using modeling as a descriptive technique, and it's something I would consider to be more a part of EDA. But then in 3.2.1, while it is perhaps not using modeling in an inferential way, the model is being used for more than just description. I might rephrase the first sentence as follows: "In order to treat the extreme values in the data, we built a robust linear regression model."

Reviewer 2: At the bottom of page 11, where you describe the model formation, more detail would be appreciated. You are modeling the mean hourly wage based on year. Why? Is this standard practice for robust linear regression? What is a "slight" residual? Is it defined based on absolute

magnitude, or standard deviations, or something else? Later, you say "To minimize the risk of mistakenly identifying an outlier as an "erroneous outlier"." Is an erroneous outlier a technical term? If not, I think you are trying not to identify erroneous outliers, not mistakenly identifying outliers as erroneous outliers (that's a double negative). The value of 0.12 feels very specific and out of place in a paper that has had few numerals. Is this a number an expert in RLR would be able to understand? Again, is that a raw value? Is it a standard deviation? Is there literature about how to choose a threshold? Overall, the sentence "We find that 0.12 is the most reasonable value to be the threshold to minimize that drawback's risk because it still captures the sensible spikes in the data." is vague and should be expanded upon. What is "that drawback's risk"? What are "the sensible spikes in the data?" Be more explicit.

Reviewer 2: What does "It implies that the fluctuation can still be observed in the data after the treatment." mean? I suspect this sentence should be replaced with something along the lines of "The figure shows that fluctuations in wages still exist even in the imputed data."

Reviewer 2: In Section 4, you talk about the variables in the original data. "One would expect that there is a record of the day the individual first started a job, and this is used to adjust the year of collection." I'm not sure what this means. Please clarify.

- We removed this sentence as we already have the `yr_workforce` variable which is calculated from the year of the individual start working.

Reviewer 2: "The treatment of unlikely wages differ in the refreshed data." I think this means "The treatment of anomalous wages differs between the original and refreshed data sets." Then you say "We opted to use the weights from a robust linear regression to determine what should be set as missing value as described in Section 3.2.1." I don't think you were setting missing values though, were you? It was used to determine which values should be imputed, correct? Either way, please clarify.

- Yes, you are correct. We imputed the extreme wages with its predicted values instead of setting it as missing values. We have clarified this in the paper.

Reviewer 2: The bulleted difficulties in Section 5 are a bit jarring. Make them parallel (i.e. every bullet should start with a verb -ing so it is determining, calculating, etc) or turn them into a paragraph. I liked the specificity of $30,000/hour but it would be better up in the section where you discuss the anomalous values.

Reviewer 2: At the end of page 14, it would be good to explicitly say that if people don't like your decisions, they can grab the code and adjust the parameters themselves! It's implied, but should be made explicit.

Reviewer 2: Figure 2's caption should include something about how that data is the raw values from the survey, before inputing anomalous values so the figure and caption can stand alone.

Reviewer 2: For both Figure 3 and Figure 5, I would prefer the letter label before the description, but that's a matter of personal preference.

Reviewer 2: For Figure 7, why did you choose to show the entire dataset here? It would be easier to make a comparison if the plot was cut off at 1994 on the x-axis.

Reviewer 2: Table 1 includes the note ?CHECK which I assume was a note to self.

- We have removed it from the caption.

Reviewer 2: Table 2 seems to have more precision than is necessary. Certainly 100% does not need to be reported as 100.00%, and probably all the percentages could be rounded to the nearest 1 percent. The caption could explain this, as there might be some rounding errors.

Reviewer 2: What does "yearly mean hourly wages with years of workforce experience" mean? Is that one variable or two?

- They are two different variables. We have separated it in the revised version.

Reviewer 2: Was your goal to collect data from 1994-2018 or 1979-2018? Page 4 states 1994 onward, but from the figures later it appears that you collected data starting in 1979 (which makes more sense to me).

- The original data covers 1979-1994 period, while in the refreshed data, we extended the period covered until 2018 (so that the data is from 1979-2018). We have made this explicit in Section 2.2.

Reviewer 2: I was confused by the following section: "We choose to use this revised May data because it seemed to have less missing and presumably has been checked. However, there is no revised May data for 2012, 2014, 2016, and 2018, thus, we use the ordinary May data for these years." I was assuming the data was collected on a yearly basis, but this section suggests it was collected each month. Please clarify somewhere in the paper.

- Yes, the data is collected on a yearly basis. However, according to the database, the highest grade completed variable is revised as in May 1st of each survey year. Thus, we mentioned as the revised May data. In the revised version, we only mention it as revised data to avoid confusion.