Response to reviewers

Submission ID 217815520

We thank the Editor, Associate Editor and the three reviewers for their extensive 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 editors 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.

Editor and 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 1: I don't see the direct relevance of this manuscript with the readership of JSDSE. I think this could be a valuable paper to describe the author(s) R package and would have value for researchers using the NLSY79 dataset. I am not convinced that JSDSE is the appropriate venue for this. I do not see this manuscript being engaging enough to be used in the classroom as a case study in data cleaning.

- It is commonly said 80% of a data analysts/scientist's work is on data cleaning/preperation so we believe that shedding more light in this aspect of work is an important for students to be aware of. We believe that there are often subjective decisions made in the process (as we have done) which can be used as a conversation for transparency and reproducibility in class based on a real case study.

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: 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.

- We agree that there is substantial literature on data cleaning. We have referenced one of the main works, Dasu and Johnson (2003) but don't feel it is necessary to cite more than this. We have similarly only pointed to one main reference for IDA and EDA.

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?

- The approach is for the case study only and is not meant to be extrapolated for other cases. We suggest using a robust mixed model as a better alternative, however, our attempt to fit such a model via `robustlmm` resulted in computational issues.

Reviewer 3: The authors explain the complication of the "years of job experience" variable being unavailable in the raw data. This information is available in the NLSY79 data accessor; it seems to simply have been not selected in the initial data download by the authors. Given that the stated goal is to recreate the textbook data, this seems like a glaring oversight. It also leads to Fig 7 (Now Figure 8) being essentially meaningless: we are unsurprised that salaries are higher with more years of experience, and we are unsurprised that salaries go up over the years with both age of employee and inflation. The positive slope of these two different plots does not mean they are related to each other in any way.

- 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. We also add a Subsection dedicated to explain this variable, i.e., Subsection 3.2. Calculated variables: work experience. This information was compared with Singer & Willet's original data. Comparison shown in Figure 7 shows that the information does not match up, however, there is a high correlation between the variables.

- **NEED HELP in responding comments on Figure 8.**

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.

Minor changes:

- We have corrected all minor grammatical errors pointed out by the reviewers.
- We have made the descriptions of the steps more explicit as follows [LIST OF THE CHANGES]

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, it is reflected on the last paragraph in page 3.

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 did not go into great detail about the packages used as the whole code is available as a vignette within the package. Also, for the blind review purpose, we have made all of the code and the raw dataset available to reviewers and it is not depend on the R package we created. The reviewers can see the data processing code and run it to also see the reproducibility of the code. Besides, in the paper, 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.

- We have added in Introduction that "Section 3 presents the steps … to find and repair anomalies.".

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 are 108 individuals up to 8th grade in the refreshed dataset, compared to 366 individuals in the original dataset. The plausible reason for this difference is that some people have had GED after the period covered in the original dataset.

Reviewer 1: what does the natural log have to do with the y-axis?

- Sometimes, the natural log is used to rescale the variable in the graph to adjust with the skewness in the data. Singer and Willet (2003) use the natural log to store the wages data by default, although without mention their motivation of doing so. Thus, we also use the natural log in Figure 8 to compare the refreshed data with the original data.

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

- We have it more explicit as per the reviewers' comments.

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.

- We have mentioned the book's title in the Introduction. The motivation of refreshing this text book data and how it is used for teaching purpose are discussed in the Introduction of the paper.

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 the file format more 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.

- 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?

- We have changed Fig 2 to show 36 individuals, that display the range of wages experience, including some with questionable values. The figure caption reflects this and more details are provide din the text.

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.

- This section has been re-written, and additional details provided.

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.

- We have made this sentences more explicit that the extremely high values were also observed in the `total hours of work`.

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."

- Thank you for your suggetion. We have modified the description to "To treat the extreme values in the data, we build a robust linear regression model where we use the robustness weight to determine if a value should be replaced with the fitted value from the 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.

- Thank you for your enquiry. We have re-written this section to better describe our process and intent. In particular, the threshold values are chosen based on extensive eye-balling of the plots of wages over time for each individual. We try to maintain a balance of not overly smoothing the observations but clearly repairing unlikely wage values. As alluded in the section, we believe that an alternative approach such as robust linear mixed models is more appropriate but we did not proceed with this as we had issues fitting the model with `robustlmm` package. In addition, we thought that the use of robust linear models is easier to explain in the context of teaching and the use of non-optimal approach can generate a good discussion about alternative approaches in class.

Reviewer 3: The choice of 0.12 seems arbitrary at best. I don't see any justification for this approach (e.g., literature, simulations, etc). Why not use bi-square, instead of huber, and simply impute for observations with weight 0. While still arbitrary, it's decided by the bisquare model and isn't a guesstimate. Further, it would be sensible for any dataset and not just this one. (This would correspond to a smaller threshold, which may catch too many strange outcomes). In any case, it's hard to justify something so adhoc.

- Similar comment to this is addressed above.

Reviewer 3: You're using iteratively reweighted least squares with Huber weighting.

- Thank you, we have changed from "M estimation" to "iteratively reweighted least squares".

Reviewer 3: I'm not sure this is accurate. The robustlmm package in R fits these models rather quickly with thousands of observations. What are the authors using? Does it run out of memory (requires more RAM) or does it take too much time?

- The `robustlmm` model would not converge. We have added some text to the paper on this.

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."

- We have replaced this sentences as per the suggestion.

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 `stwork` and `yr_wforce` variables which are the year of individual start working and length of the years of individual entering the labor force, respectively.

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.

- We have added the specifity of the extremely high wages in Subsection 3.3. Initial Data Analysis. We have also edited every bullet so it started with verb -ing.

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.

- Done

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.

- We have edited the caption.

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.

- We have moved the letter label before the description.

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.

- We have filtered the x-axis of the refreshed data set until 1994, so they are comparable.

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.

- We agree that this level of precision was not necessary. We have modified Table 2 so percentages are to the nearest 1 percent and total percentage (=100%) is omitted.

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

- These are two different variables. We have made this easier to follow in the manuscript as we add explanation on what the target data is (See Subsection 2.2: Target Data).

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.

- The database has the variable of `the highest grade completed ever`. Thus, we use this variable for `hgc` instead of calculating it from the`hgc` the respondents answer for each round of the survey. Besides, we also add new variable `grade`, which is the highest grade completed for each round of survey, so the value changes corresponding to the year of the survey. According to the database, the `hgc` 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.

Reviewer 3: Abstract: The abstract portrays this work as "re-creating a textbook example data set." This doesn't appear to be quite accurate because there are several notable differences

between the end result and the textbook data. Currently, it's more about demonstrating a procedure for "initial data analysis."

- The abstract has been re-written to reflect the re-framing of the paper.

Reviewer 3: Where is IDA undervalued and neglected? Research? Teaching? Both?

- We removed this sentences, and edited the explanation about IDA in Paragraph 6 of the Introduction.

Reviewer 3: The authors discuss refreshing an example from Singer and Willet (2003), but this isn't ever completed. They compare the data, but not the longitudinal model. Again, it's more about demonstrating a procedure for "initial data analysis" than it is reproducing Singer and Willet (2003).

- Our main aim in this paper is to refresh the data, which is used as one of examples of longitudinal data and its modelling in Singer and Willet (2003). In other words, we only aims to refresh the data until the most recent survey period, and not the longitudinal model, so that it is more relevant for today's context, especially, if its is used for teaching.

Reviewer 3: Tidy and tame data are discussed in passing and the utility of tidy and tame data are not obvious. The authors later describe tidy data rather clearly (page 7) but tame data never receives the same treatment.

- We put explanation on the difference between tidy and tame data in the Introduction (Paragraph 7). Tame data is sort of a result product, in which, in order to produce it, we need to tidy the data.

Reviewer 3: A historical context of NLSY79 would be helpful. Specifically, the dropped subjects are mentioned but there's no context of why. By the argument of the paper, these steps should be transparent and well-considered.

- We have added the reason of why some subject were dropped from the interview in Section 2.1.

Reviewer 3: The author switches between gender (man/woman/etc) and sex (male/female/etc.) throughout the manuscript. Further, the explanation here covers important ideas in teaching how to use data on gender/sex and race, but it's just the beginning of ideas. It might be better to discuss the questions on the survey and discuss possible shortcomings, rather than make a somewhat empty gesture towards today's standard.

- **NEED HELP** The paper has been fixed so that male/female is used consistently. We have also rewritten the paragraph to focus on the survey questions and shortcomings.

Reviewer 3: The discussion around highest grade completed is mechanical in what was done, but lacks substantive reasoning for those decisions and possible downstream consequences.

- The variable of `highest grade completed`, besides it is available in the original data, it is also use to filter the high school dropouts (the subset of NLSY79 cohort that is contained in the original data). We explain how we use this variable, along with a newly added variable, GED, and possible downstream consequences in Subsection 4.1 Filtering: Determining who is a dropout.

Reviewer 3: It appears the goal of the paper switches to just IDA and dealing with problematic points here. This is completed in the paper.

- Paper has been re-written to change focus.

Reviewer 3: Gender (man/woman/etc). Sex (male/female/etc).

- See explanation above

Reviewer 3: It is not clear why participants without a high school degree are not part of the analysis.

- In the current version of the manuscript, we have included all of the cohort, regardless of their school degree in a dataset called `wages`. However, we also subset the data to be high school dropouts only as we aims to refresh Singer and Willet's data (only contains male high school dropouts aged 14-17 years old when first interviewed). Hence, for comparability, we only analyse the high school dropouts in Section 4.

Reviewer 3: Figure 2 is helpful for observing the data there is a lot of variability in both hourly wage and trends. A discussion about censored data might be helpful here – this is not something you impute from the data.

- **NEED HELP**

Reviewer 3: It's unclear about what fluctuation or treatment the author is referring to.

- Fluctuation is the variability in the data (up and down) in the wages, while treatment is imputing the extremely high wages (anomaly) with their RLM's predicted value. The detail explanation of fluctation is discussed in Subsection 3.3 Initial Data Analysis, while the treatment is explained in Subsection 3.4 (Paragraph 3). We also have edited this sentences to avoid implicity.

Reviewer 3: If the goal was to reproduce the textbook data, then why weren't wages inflation-adjusted?

- This is not done for the refreshed data because we plan to keep refreshing it as the data from the newest survey period release. Thus, the 1990 prices might not be relevant anymore for the future context. Hence, we keep it as the original value to allow the users to treat it as their needs. In other words, the inflation-adjustment of the wages is better to do with each wave of new data added, so that it is relative to the last date in the data.

- This explanation exists in the manuscript in Section 4.2 and Section 5.

Reviewer 3: Figure 7 (Figure 8 in the current manuscript) is interesting, but it wasn't until my second time through that I realized this was the textbook example for replicating. There is a clear difference in analyses (different question really), but is there anything to justify a similarity, or improvement of the analysis?

- **NEED HELP** Figure 8 has been updated, and other comparisons of the original and refreshed data have been added. There are still some differences, but with the change in variables being created and better filtering to get the dropouts the data are now more similar.

Reviewer 3: (page 14, line 46) Thank you for this acknowledgment. I think this is true for most analyses. It doesn't mean that you shouldn't transparently and fully investigate your own choices.

- NEED HELP We are really not sure what this refers to.