A Journey from Wild to Textbook Data to Reproducibly Refresh the Wages Data from the National Longitudinal Survey of Youth Database

Abstract

Textbook data is essential for teaching statistics and data science methods because they are clean, allowing the instructor to focus on methodology. Ideally textbook data sets are refreshed regularly, especially when they are subsets taken from an ongoing data collection. It is also important to use contemporary data for teaching, to imbue the sense that the methodology is relevant today. This paper describes the trials and tribulations of refreshing a textbook data set on wages, extracted from the National Longitudinal Survey of Youth (NLSY79) in the early 1990s. The data is useful for teaching modeling and exploratory analysis of longitudinal data. Subsets of NLSY79, including the wages data, can be found in supplementary files from numerous textbooks and research articles. The NLSY79 database has been continuously updated through to 2018, so new records are available. Here we describe our journey to re-create the wages data, and document the process so that the data can be regularly updated into the future. Our journey was difficult because the steps and decisions taken to get from the raw data to the wages textbook subset, have not been clearly articulated. We have been diligent to provide a reproducible workflow for others to follow, which also hopefully inspires more attempts at refreshing data for teaching. Three new data sets and the code to produce them are provided in the open source R package, called [CENSORED].

Keywords: Data cleaning; Data tidying; Reproducible workflow; Longitudinal data; NLSY79; Initial data analysis;

1 Introduction

Statistics and data science education relies on cleaned and simplified data, suitably called textbook data, for clear examples about how to apply different techniques. An example of this, is the wages data made public by Singer and Willett (2003) which can be used to teach generalized linear models, in addition to hierarchical, mixed effects and multilevel. The data records hourly wages of a sample of high school dropouts, from 1979-1994, along with the demographic variables, such as education and race, taken from the National Longitudinal Survey of Youth (NLSY79) (Bureau of Labor Statistics, U.S. Department of Labor 2021a).

The story from modeling the data (and as reported by Singer and Willett) is that wages increase with the length of time in the workforce, higher level of education leads to higher wages, and that race makes a difference, on average. An exploratory analysis reveals however that the individual experience varies a great deal from the overall average. Some individuals experience a decline in wages the longer they are in the workforce, and many experience volatility in their wages. The wages data was used to illustrate exploratory longitudinal data analysis in Ilk (2004), and was further developed into a case study for use in the teaching of exploratory data analysis at CENSORED.

The story from modeling the data (and as reported by Singer and Willett) is that wages increase with the length of time in the workforce, higher level of education leads to higher wages, and that race makes a difference, on average. An exploratory analysis reveals however that the individual experience varies a great deal from the overall average. Some individuals experience a decline in wages the longer they are in the workforce, and many experience volatility in their wages.

This disparity between the average and the individual is a part of statistics, as a discipline, that require more attention. This particular data set is a prime example of discussing this disparity. Textbook data sets have longevity if there is a unresolved mystery. The iris data (Andereson 1935) is a prime example. It has withstood the test of time because the three species cannot be perfectly classified, and so it continues to challenge researchers and instructors to do better in the analysis. (A side note: the iris data is best replaced today with the penguins data (Horst, Hill, and Gorman 2020), which has similar qualities, is new and does not suffer from a connection with eugenics (Stodel 2020). We argue that the wages

data is in this class of textbook data, too, because it presents a challenge for longitudinal data analysis: how can we better summarize and explain the individual experience?

For the field of statistics, and data science by association, it is increasingly important to reach the individual. One might describe this as a divergence of purpose, statistics for public policy or statistics for the public. The two are not the same. As the world becomes more electronically connected combating misinformation and mitigating conspiracy theories require that statistics address the individual. For example, with the wages data, even though the message for public policy is that demographic profile is related to different wage patterns on average, the message for the individual is that you are more than your demographic. The majority of people in the study do not have a pattern that is similar to the average. If you have a bad experience, that your wages have declined over time, you are not alone, there are others like you, and more than you think. A similar tone is echoed occasionally in the public media, for example, an article published in the Sydney Morning Herald argues that there is no average Australian (Moncrief 2015).

As a textbook data set though, the wages data is outdated. The most recent year in the data is 1994, 10 years prior to when Singer and Willett (2003) was published. Teachers of statistics need contemporary data sets to show how techniques are relevant for today's students. Using tired old textbook data sets can imbue a misconception that the field is not current. The wages data is extracted from NSLY79, one of the best examples of open data, which is constantly being updated. It should be possible to continuously refresh the textbook data from the data repository. This paper describes our (non-glamorous) journey from open (Open Knowledge Foundation 2021) wild data to textbook data.

This paper demonstrates the steps of cleaning data, including subjective decisions made on dealing with anomalies, and documents the process, as recommended by Huebner, Vach, and Cessie (2016). They emphasize that making the data cleaning process accountable and transparent is imperative and essential for the integrity of downstream statistical analyses and model building. Clean data often then goes through an "initial data analysis" (IDA) (Chatfield 1985), where one would summarize and scrutinize the data, especially to check if the data is consistent with assumptions required for modeling. This stage is related to exploratory data analysis (EDA), coined by Tukey (1977) with a focus on learning from

data. EDA can be considered to encompass IDA. In practice the three stages of cleaning, summarizing and exploring are cyclical, that one often needs to do more cleaning after scrutinizing. Dasu and Johnson (2003) say that data cleaning and exploration is a difficult task and typically consumes a large percentage of the time spent in analyzing data.

Our process of cleaning builds heavily on the `tidyverse` approach (Wickham, Averick, et al. 2019). The data is first organized into "tidy data" (Wickham 2014) and then further wrangled using the data pipeline and split-apply-combine approach (Wickham 2011). Tidy data shouldn't be confused with "tame data" which Kim, Ismay, and Chunn (2018) coined to refer to textbook data sets suitable for teaching, particularly teaching statistics. The resulting (tame) data is provided in a new R package called `[CENSORED]` which includes the code so that the process is reproducible, and could be used to further refresh the data as new records are made available in the NLSY79 database.

This paper is structured in the following way. Section 2 describes the NLSY79 data source. Section 3 presents the steps of cleaning the data, including getting and tidying the data from the NLSY79 and IDA to find and repair anomalies. Our final subset is compared to the old textbook subset in Section 4. Finally, Section 5 summarizes the contribution and makes recommendations for the NLSY79 data curators.

2 The NLSY79

The book "Applied longitudinal data analysis" (Singer and Willett 2003) used the wages and other variables of high school dropouts from the NLSY79 data as an example data set to illustrate longitudinal data modeling of wages on workforce experience, with covariates education and race. This data has been playing an important role in research in various disciplines, including but not limited to economics, sociology, education, public policy, and public health for more than a quarter of the century (Pergamit et al. 2001). In addition, this is considered a carefully designed longitudinal survey with high retention rates, making it suitable for life course research (Pergamit et al. 2001; Cooksey 2017). According to Cooksey (2017), thousands of articles, and hundreds of book chapters and monographs have utilized this data. Moreover, the NLSY79 is considered the most widely used and most important cohort in the survey data (Pergamit et al. 2001). Our aim is to

refresh this textbook data and append it with data from 1994 through to the latest data reported in 2018, a purpose that is consistent with Grimshaw (2015)'s statistics education goal of embracing authentic data experiences. Here, we investigate the process of getting from the raw NLSY79 data to a textbook data set as similar as possible to that provided by Singer and Willett (2003). We should also note that race is a variable in the original data set, and for compatibility it is also provided with the refreshed data, for the *purposes of studying racism, not race* (Fullilove 1998).

2.1 Database

The NLSY79 is a longitudinal survey administered by the U.S Bureau of Labor Statistics that follows the lives of a sample of American youth born between 1957-1964 (Bureau of Labor Statistics, U.S. Department of Labor 2021a). The cohort originally included 12,686 respondents aged 14-22 when first interviewed in 1979. For a variety of reasons, some structural, the number of respondents dropped to 9,964 after 1990. The surveys were conducted annually from 1979 to 1994 and biennially thereafter. Data are currently available from Round 1 (1979 survey year) to Round 28 (2018 survey year).

Although the main focus area of the NLSY is labor and employment, the NLSY also covers several other topics, including education, training, achievement, household, geography, dating, marriage, cohabitation, sexual activity, pregnancy, fertility, children, income, assets, health, attitudes and expectations, crime, and substance use.

There are two ways to conduct the interview of the NLSY79, which are face-to-face or by telephone interviews. In recent survey years, more than 90 percent of respondents were interviewed by telephone (Cooksey 2017).

2.2 Target data

The NLSY79 data used in Singer and Willett (2003) contains the longitudinal records of male high school dropouts who first participated in the study at age 14-17 years from 1979 through to 1994. This dataset contains several variables as follows:

1. ID: the respondents' ID.

2. EXPER: temporal scale, i.e., the length of time (years) in the workforce, starting on the respondents' first day at work.

3. LNW: natural logarithm of wages, adjusted with 1990's inflation rate.

4. BLACK: binary variable, 1 indicates black and 0 otherwise.

5. HISPANIC: binary variable, 1 indicates hispanic and 0 otherwise.

6. HGC: the highest grade completed.

7. UERATE: unemployment rate of the year of survey. When missing, the variable is set to be 7.875 (the average rate).

We refresh this data by re-creating the full data with records from survey years 1979 through to 2018 (the most recent year published). We also modify some variables. For example, use a single categorical race variable instead of the two binary race variables. We also plan to include additional variables, some for the purpose of providing more options for data exploration in teaching examples: year of the survey, age of individual in 1979, whether the individual completed high school with diploma or with Graduate Equivalency Degree (GED), the number of jobs that the individual had in the corresponding year of survey, total number of hours the individual usually works per week, the year when individual started to work, and the number of years the individual worked. We do not attempt to re-create the unemployment rate variable.

The plan is to create three datasets as follows:

1. The wages data of the whole NLSY79 cohort, including females.

2. A separate table of the demographic data of the whole NLSY79 cohort.

3. The wages data of the high school dropouts, which is closest to a refreshed version of Singer and Willett (2003)'s data.

3 Data cleaning

van der Loo and de Jonge (2018) in the context of official statistics describe the "statistical value chain" which includes various production stages of the data cleaning process as raw data (data in the initial that it arrives), input data (data organised with correct type and identified variables) and valid data (data that has been cleaned and more accurately

represents the intent of variables). What we have colorfully named as wild data can be considered to be raw data, and valid data could be considered to be textbook data, in the above statistical value chain. In this section, we outline the steps to download the raw data (Section 3.1) and then tidy the raw data into input data, specifically for the demographic variables (Section 3.1.1) and the employment variables (Section 3.1.2), so that the resulting input data can be used downstream for validating the data as described in Section 3.3.

3.1 Getting the data

The NLSY79 data contains a large number of variables but for our purposes the scope required is limited to demographic profiles, wages data, and work experience. More specifically, we went to the NLSY79 database website at `https://www.nlsinfo.org/content/cohorts/nlsy79/get-data`, clicked on the direct link to NSLY79 data and navigated as described in Figure 1.

The downloaded data set comes as a zip file, containing the following set of files:

- `NLSY79.csv`: Comma Separated Value format of the response data,
- `NLSY79.dat`: .dat format of the response data,
- `NLSY79.NLSY79`: Tagset of variables that can be uploaded to the website to recreate the data set, and
- `NLSY79.R`: R script for reading the data into R and converting the variables' names and label into something more sensible.

We alter only the file path in `NLSY79.R` and run the script without any other alteration. This results in an initial processing of the raw data into two data sets, `categories_qnames` (where the observations are stored in categorical/interval values) and `new_data_qnames` (the observations are stored in integer form).

According to Wickham (2014), tidy data sets comply with three rules: (i) each variable forms a column, (ii) each observation forms a row, and (iii) each type of observational unit forms a table. The raw data, `new_data_qnames`, does not comply with these rules as it is organised such that each row corresponds to an individual. As respondents can have multiple jobs at specific years, the column names, such as `HRP1_1979`, `HRP2_1979`,

7

HRP1_1980 and HRP2_1980, contain the information about the job number up to 5 (HRP1 = job 1, HRP2 = job 2) and the year. The raw data consequently has a large number of columns (663 to be specific). The values in the cell under the variables that begin with HRP correspond to the hourly wage in dollars. A glimpse of this data shows:

```
#> 'data.frame':    12686 obs. of  663 variables:
#>  $ CASEID_1979                      : int  1 2 3 4 5 6 7 8 ...
#>  $ HRP1_1979                        : int  328 385 365 NA 310 NA NA NA ...
#>  $ HRP2_1979                        : int  NA NA NA NA 375 NA NA NA ...
#>  $ HRP3_1979                        : int  NA NA 275 NA NA NA NA NA ...
#>  $ HRP4_1979                        : int  NA NA NA NA NA 250 NA NA ...
#>  $ HRP5_1979                        : int  NA NA NA NA NA NA NA NA ...
#>  $ HRP1_1980                        : int  NA 457 397 NA 333 275 300 394 ..
#>  $ HRP2_1980                        : int  NA NA 367 NA NA NA NA NA ...
#>  $ HRP3_1980                        : int  NA NA 380 NA NA NA 290 NA ...
#>  $ HRP4_1980                        : int  NA NA NA NA NA NA NA NA ...
#>   [list output truncated]
```

Thus we re-arrange and wrangle the data into tidy data form, columns corresponding to individual ID, year, job number, wage in dollars and the demographic variables. This is done using the `tidyverse` suite of packages (Wickham, Averick, et al. 2019): `tidyr` (Wickham 2020) to pivot the data into long form, with `dplyr` (Wickham, François, et al. 2020) and `stringr` (Wickham 2019) to mutate new variables from the downloaded data from the database, and code levels of factors by text wrangling. The long form of the data makes it possible to do these data transformations efficiently, and it is an intermediate step towards the final target data. The code for tidying the data are demonstrated at `CENSORED/articles/raw-to-input-data.html` but also described in the subsequent subsections.

3.1.1 Tidying demographic variables

In our final target data, we wish to include the demographic variables with variable names specified in brackets: gender (`gender`), race (`race`), age (`age_1979`), highest grade com-

pleted (`hgc`), highest grade completed in terms of years, e.g. 9th grade = 9, 3rd year college = 15, (`hgc_i`), and whether the graduate equivalency diploma is obtained (`ged`).

For gender and race, we only rename the column names. It is worth noting that using these two variable needs special attention. Gender, as reported in the data, only has two categories, which is recognised today as inadequate. Gender is not binary. Further, race is as reported in the database. When doing analysis with this variable, one should keep in mind that the purpose is to study racism rather than race.

The raw data, `new_data_qnames`, contains the variables `Q1-3_A~Y_1979` and `Q1-3_A~Y_1981` which records two versions of the birth year of the respondent; this is also the case for the record of birth month (`Q1-3_A~M_1979` and `Q1-3_A~M_1981`). The record contains two versions of birth year and birth month as the survey recorded this in 1979 and 1981. We checked for consistency between the two versions and found no discrepancy where the responses were recorded in both 1979 and 1981. The age was then calculated using the birth year.

The next step is tidying to obtain `hgc` and `ged`. As the highest grade completed is already exist in `new_data_qnames` and stored in year unit (e.g., 10, 11, 12, 13, and so on), we only rename the column name to be `hgc_i` and recode it so we have the factor data type (e.g., 10th grade, 11th grade, and so on). For `ged`, we only separate the year of survey and the GED status of the individual. Later on, this variable is useful to determine the refreshed high school dropouts data.

Finally, we get all of the demographic profiles of the NLSY79 cohort. We then save this data as `demog_nlsy79`.

3.1.2 Tidying employment variables

Our target variables for the employment are to obtain respondent's mean hourly wage (`wage`), the number of jobs (`njobs`), the total hours of work per week (`hours`) for each survey year, the year when individual starting to work (`stwork`), the length of time (years) in workforce (`yr_wforce`), and work experience measured as the number of years worked (`exp`). As the data only reports up to 5 jobs for each respondent, the maximum number of jobs is capped at 5.

From 1979 to 1987, `new_data_qnames` only contains one version of hours worked per week for each job (in the variables with names starting with `QES-52A`). From 1988 onward, we selected the total hours worked per week, including hours working from home (`QES-52D`). However, in 1993, this variable was missing for the first and last job so we selected to use `QES-52A` instead. In addition, 2008 only had jobs 1-4 for the `QES-52D` variable, so we use only these.

The hourly wages are in the variables beginning with `HRP` in `new_data_qnames`. As a respondent may have multiple jobs, the `mean_hourly_wage` is computed as a weighted average of the hourly wage for each job with the number of hours worked for each job as weights (provided that the information on number of hours is available); if number of hours worked for any job is missing, then the `mean_hourly_wage` is computed as a simple average of all available hourly wages. Prior to computing the mean hourly wage, we undertook a number of steps to treat unusual observations as described below:

- If the hourly rate is recorded as 0, we set wage as missing.
- If the total hours of worked for the corresponding job is greater than 84 hours, we set the wage and hour worked as missing.

The number of jobs (`number_of_jobs`) for each respondent per year is computed from the number of non-missing values of hourly wage. In other words, even if the information of hours worked exists for a particular observation, we do not tally when the hourly wage is missing.

```
#> # A tibble: 10 x 6
#>       id  year mean_hourly_wage total_hours number_of_jobs is_wm
#>    <int> <dbl>            <dbl>       <int>          <dbl> <lgl>
#> 1     1  1979             3.28          38              1 FALSE
#> 2     1  1981             3.61          NA              1 FALSE
#> 3     2  1979             3.85          35              1 FALSE
#> 4     2  1980             4.57          NA              1 FALSE
#> 5     2  1981             5.14          NA              1 FALSE
#> 6     2  1982             5.71          35              1 FALSE
```

10

```
#>  7     2  1983         5.71        NA            1 FALSE
#>  8     2  1984         5.14        NA            1 FALSE
#>  9     2  1985         7.71        NA            1 FALSE
#> 10     2  1986         7.69        NA            1 FALSE
```

For `stwork` variable, we only rename the column names from `new_data_qnames`. This variable is then used to calculate the next variable, `yr_wforce` for each survey round, which is the year of survey (`year`) minus the year of individual start working (`stwork`). Finally, `exp` variable is derived from the number of weeks worked since the last interview indicated as variable started with `WKSWK` in `new_data_qnames`. To obtain the work experience since 1979, we calculate the cumulative value. As the measurement unit is in week, we convert this to year.

The employment and demographic variables are then joined. These data are further filtered to the cohort who participated in at least three rounds in the survey, the minimum observation recommended for longitudinal data in Singer and Willett (2003). We also opt to restrict the minimum number of observations so that the data can be used to demonstrate within-person variation when it is used to teach longitudinal data. However, it is worth noting that the original data does not restrict the number of observation for each individual, i.e, there are individuals with only one and two observations.

Finally, we save the resultant wage data on this cohort as `wages`.

3.2 Calculated variables: work experience

Work experience is one of the most important variables in Singer and Willett (2003) as it indicates time, and makes longitudinal analysis possible. It is desirable to calculate this rather than using the survey year for time because it more accurately reflects a person's time in the workforce. Thus, in the spirit of refreshing the data to the newest round of the survey, this variable needs to be calculated from other variables provided. It is not straightforward. We start with the definition of experience in Singer and Willett (2003).

Experience is years after entering the labor force. It represents the difference between the day an individual enters the labor force (`EMPLOYERS_ALL_STARTDATE_ORIGINAL.01~Y_XRND`

11

in the database) relative to the date of the survey, which we call `yr_wforce`. However, using this calculation produced numbers which don't quite match with the original data.

Reading the section titled "Topical Guide to the Data" in the guide, suggests that it should be calculated based on the variable "number of weeks worked since last interview". This would remove periods of unemployment which makes sense when measuring experience while actually working. Since it is only measured since the last interview, this needs to be cumulated for each survey year. This produced results more similar to the original data (see Subsection 4.2).

3.3   Initial data analysis

According to Huebner, Vach, and Cessie (2016), initial data analysis (IDA) is the step of inspecting and screening the data after collection to ensure that the data is clean, valid, and ready to be deployed in the later analyses. This is supported by Chatfield (1985) who argues that the two main objectives of IDA are data description, which is to assess the structure and the quality of the data, and model formulation without any formal statistical inference.

In this paper, we conduct an IDA or a preliminary data analysis to assess the validity of the variable values in the cohort of data that the NLSY provides. A first step is validating numerical summaries of the raw data are the same as reported by NLSY79. This is followed with graphical summaries using methods available in `ggplot2` (Wickham 2016) and `brolgar` (Tierney, Cook, and Prvan 2020).

The respondents' ages ranged from 12 to 22 when first interviewed in 1979. Hence, we validate whether all of the respondents were in this range in the data we extracted. Additionally, the NLSY also provides the number of the survey cohort by their gender (6,403 males and 6,283 females) and race (7,510 Non-Black/Non-Hispanic; 3,174 Black; 2,002 Hispanic). To validate this, we used the `demog_nlsy79`, i.e., the data with the survey years 1979 sample. Tables 1 and 2 suggest that the demographic data we had is consistent with the sample information in the database.

In the next step, we explore the mean hourly wage data. In this case, we only explore the wages data of respondents that have the highest completed grade of up to 12th grade.

Table 1: Frequency table of the age at the start of the survey in NSLY79 cohort

Age	Number of individuals
15	1,265
16	1,550
17	1,600
18	1,530
19	1,662
20	1,722
21	1,677
22	1,680

We employ visualization techniques to perform the IDA as described next.

We randomly take 20 samples from the data and plot them, as shown in Figure 2. It shows that these respondents have a lot of variability in wages, for example, the IDs in panel numbers 5, 7, and 11. The plot also implies that the samples have a different pattern of mean hourly wages. 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. However, when checking the summary plots (Figure 3), we found that some observations had exceptionally high wages. Some of them, for example respondents in Figure 3 C had experienced unusual wages only in certain year.

The anomalies are also found in the total hours of work, where some observations reported as having worked for 420 hours a week in total. According to Pergamit et al. (2001), one of the flaws of the NLSY79 employment data is that the NLSY79 collects the information of the working hours since the last interview. Thus, it might be challenging for the respondents to track the within-job hours' changes between survey years, especially for the respondents with fluctuating working hours or seasonal jobs. It even has been more challenging since 1994, after which respondents were only surveyed every other year and thus had to recall two full years' job history. This shortcoming might also contribute to the fluctuation of one's wages data.

Table 2: Contingency table for gender and race for the full NLSY79 data. The percentage (rounded to closest 1%) is out of the total corresponding to row.

| Gender | Race | | | Total |
	Hispanic	Black	Non-Black, Non-Hispanic	
Male	1,000 (16%)	1,613 (25%)	3,790 (59%)	6,403
Female	1,002 (16%)	1,561 (25%)	3,720 (59%)	6,283
Total	2,002 (16%)	3,174 (25%)	7,510 (59%)	12,686

3.4 Replacing extreme values

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. We build the model for each ID utilizing the `nest` and `map` function from `tidyr` (Wickham 2020) and `purrr` (Henry and Wickham 2020), respectively. This fits a model separately for each individual, and a robust linear mixed model could be used as an alternative approach. We tried fitting a robust linear mixed model using `robustlmm` (Koller 2016), however perhaps because the data is large and imbalanced, the model was not converging. The full code for this is shown at `CENSORED/articles/input-to-valid-data.html` but also described in detail next.

We build the model using the `rlm` function from `MASS` package (Venables and Ripley 2002). We set the `mean_hourly_wage` and `year` as the dependent and predictor, respectively. Furthermore, we use M-Estimation with Huber weighting, where the observation with a slight residual gets a weight of 1, while the larger the residual, the smaller the weight (less than 1) (UCLA: Statistical Consulting Group 2021). The challenging part of detecting the anomaly using the robustness weight is determining the weight threshold in which the observations are considered outliers. Moreover, it should be noted that not all the outliers are due to an error. Instead, it might be that one had reasonably increasing or decreasing wages in a particular period.

To minimize the risk of mistakenly identifying an outlier as an "erroneous outlier", we

simulate some thresholds and study how they affect the data by eye balling the results. Based on this eye balling, we find that 0.12 is the most reasonable value to use as the threshold to minimize the risk of overly smoothing the wage by retaining sensible spikes in the data. In other words, we keep maintaining the natural variability of the wages while minimizing anomalies because of the error in the data recording. After deciding the threshold, we impute the observations whose weights are less than 0.12 with the models' predicted value. We then flag those observations in a new variable called `is_pred`.

Figure 4 shows the mean hourly wage before and after the extreme values are replaced. It implies that the fluctuation can still be observed in the data after the treatment. However, the large spikes, which are considered "erroneous outliers", are already eliminated from the data. Hence, the model produces a data set with a more reasonable degree of fluctuation.

Further, Figure 5 A shows that after eliminating the extreme values, the highest value has decreased to be around $350. The spikes are still observed but not as extreme as the original data set. In Figure 5 B, we plot the three features of mean hourly wages, namely the minimum, median, and maximum value. We still can see some extreme values in maximum wages, but consider it as a natural variability of the data. In Figure 5 C, we can see that after imputing the extreme value in ID=39, we can see how the the wages change over the years more clearly.

Finally, we save the imputed data and set the appropriate data type for the variables.

3.5   Recap

Figure 6 summarizes the steps taken to go from raw to input to valid data (van der Loo and de Jonge 2018) to create a refreshed wages data set.

4   Comparison of refreshed with the original data

```
#> # A tibble: 2 x 3
#>   hgc   hgc12     n
#>   <chr> <chr> <int>
#> 1 12TH  8TH       1
```

```
#> 2 8TH    12TH    243
```

The original set contains wages of high school dropouts (Singer and Willett 2003) from 1979 through to 1994. The refreshed data set for comparison is also wages on high school dropouts from 1979 to 2018. The original set is available in the R package `brolgar` and the refreshed data is available in the R package `[CENSORED]`.

4.1 Filtering: Determining who is a dropout

There is no explicit explanation of how the dropouts cohort is determined in the original data. Hence, we use the high school dropouts criteria from Wolpin (2005), which are:

1. An individual whose highest grade completed (`hgc`) is reported to be less than 12th grade, **or**

2. An individual whose highest grade completed (`hgc`) is reported to be at least 12th grade and have received a GED (`ged` is code to 2).

An additional criteria from Singer and Willett (2003) is to only include males aged between 14 and 17 years old in 1979. Filters using these criteria used re-create the dropouts data.

With this filtering, we obtained 670 individuals in the refreshed data compared to 888 individuals in the original data. To investigate the reason for the difference, individuals from the original and refreshed dataset were matched by id. This revealed several reasons for for the disparity:

1. 173 individuals were more than 17 years old in 1979. Thus, it looks like the description of the original data is not quite accurate, that there are people older than 17 in the subset. Our decision is to also include them in the refreshed data as the new data contains an age variable, so analyst could filter them later.

2. 79 individuals were less than or equal to 17 years old in 1979. However, they were not captured in the refreshed data because:

- 35 of them completed at least 12th grade with a diploma instead of GED (`ged` variable is coded to 1). This suggests that they are not dropouts, and so we excluded them from the refreshed data.

- The information about `ged` is missing in 38 individuals. We decided to include them in the refreshed data.

- 3 individuals have both diploma and GED (`ged` is coded to 3). These were kept in the refreshed data.

- 12 individuals do not exist in `wages` data because they were participated in less than 3 rounds of survey.

The filtering was re-applied using these decisions. The refreshed dropout data has 863 individuals.

4.2   The takeaways

There are two aspects of the original data that make the replication with exactly the same result is not obtained. The first one is that the calculation of the time variable provided, which is the "work experience", is not clearly articulated. Singer and Willett (2003) mention that the temporal variable is the years of experience since the first day of work. However, this exact same variable is not explicitly available in the database. From the NLSY79 topical guide (Bureau of Labor Statistics, U.S. Department of Labor 2021b), we find that several variables are tagged as work experience-related variables. One of them is the weeks of worked since the last interview. Hence, we use this to calculate the variable. Even though we got different result, the correlation between the original and refreshed work experience is reasonably high as shown in Figure 7.

Secondly, in the original dataset, wages were inflation-adjusted to 1990 prices. 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.

It is also worth noting that the treatment of unlikely wages differ in the refreshed data. In the original data by Singer and Willett (2003), wages greater than $75 are set to be missing. However, in recent context, this value is too low to be set as the maximum threshold. We opted to use the weights from a robust linear regression to determine what should be regarded as outliers and imputed them with their predicted values as as described

17

in Section 3.4.

Figure 8 shows a comparison of the two sets. (A direct ID matching is not possible.) There are 888 individuals in the original and 1,188 individuals in the refreshed data. This suggests some individuals were removed in the original data. In addition, in the original set the racial breakdown was 246 Black, 204 Hispanic and 438 White participants, while in the refreshed data there are 346 Black, 219 Hispanic, and 623 White participants, so the proportions have not been exactly reproduced. On education, in the refreshed data there are very few individuals with less than a 12th grade education, which is different from the original. In the original data there were 366 individuals with up to 8th grade and 522 9th-12th grade, as compared to 967 with 12th grade in the refreshed data.

5   Summary

This paper has illustrated the steps and decisions made to take a particular open data set and make it a textbook data set, ready for the classroom or research. In the first stage, we showed the steps performed to get the data from the NLSY79 database. The data format was converted to tidy format, for more flexibility in cleaning and exploring. An initial data analysis was conducted to investigate and screen the quality of the data. We found and provided a fix for many anomalous observations in wages using a robust linear regression model. The refreshed data is compared with the original set using a variety of numerical summaries and graphics. The current subset is made available in a new R package, called `[CENSORED]`.

The data cleaning process is documented and the code has been made available. These provide the opportunity to again refresh the textbook data as new data is published into the NLSY79 database. Determining an appropriate robustness weight from which to threshold unusual observations was conducted using a `shiny` (Chang et al. 2020) app, and the choices used in the refreshed data are documented.

Various difficulties were encountered in trying to refresh the data, which include:

- determining which records should be downloaded from the database.
- calculating experience in the workforce requires comparing the date of first job with

the first year the individual was recorded, both of which are available in the database.

- there are many unusually high hourly wages, e.g. greater than $30,000 per hour.
- there is no explicit variable in the database recording high school dropout, which means we needed to compare the date of 12th grade with the individual's age.
- matching id's from the original data with those in the refreshed data, do refer to the same person based on the demographic information available.

Ultimately, the refreshed data is reasonably similar to the original, but unsatisfactorily far from it. The last step required would be to inflation-adjust wages. This is better to do with each wave of new data added, so that it is relative to the last date in the data. Our decision was to provide the raw wages, and include code to do the adjustment as part of the package.

Some readers may disagree with our decisions made to produce the refreshed textbook data and may have better insight than us in producing a more appropriate textbook data. We do not assert that we have produced the best textbook data, but rather we describe our journey to provide a reasonable textbook data set. All code and documentation are provided for transparency. Readers could use this to make different decisions, or provide suggestions on through package for better choices. Future updates of the [CENSORED] package may contain additional variables, or filters of the full set, if it is deemed important.

For the data providers, we recommend that a better validation system with clear rules applied at data entry, and that alternative output formats, such as a tidy format would help users make better use of their resource. The problem with many of the wages records is that there are implausible values, or confusion on how to record wages for multiple jobs. These values can be validated with simple checks at data entry. Providing an open data resource also is accompanied with the responsibility that the data, especially data that is as valuable as this, is reliable. Users need to be able to trust the data.

Why is this exercise important for teachers and students? This work illustrates the steps in cleaning and processing data in the preparation for it to be a textbook data set. Choices made during the cleaning and processing can affect findings made with the data, and these should be transparent. Along with the textbook data, which is now provided as an R package, the documentation of the process provide some examples for teaching data

cleaning, initial data analysis and exploratory data analysis. The refreshed textbook data provides a resource for teaching longitudinal data analysis.

The wages data provides a good opportunity to discuss the difference between statistics to use for public policy and statistics that relate to the individual. Public policy is based on models, yielding averages that might vary across strata. Modeling the wages relative to workforce experience, with demographic covariates we learn that there are significantly different patterns. That more education leads to increasingly higher wages, which is a satisfying result for educators. It means that education makes a difference in the wage experience, and that public policy that encourages education is a data-supported action. We would also learn, although not presented here, that race matters and that being black leads to lower wages. This is a disturbing finding because there is no rational for such a difference in a fair society. This would provide support for action in public policy to remove this overall average effect. It also provides an example for educators to explain the use of data to support public policy action.

On an individual level, one needs to know where am I in this data, and does this data relate to me. To do this, the individual profiles need to be explored. Pre-dominantly, we would learn that the variation from one individual to another is far more than the variation between demographic strata. For example, many individuals with lower educational attainment earn very high wages. From a statistics and data science educator perspective, more focus and more methodology for this type of statistics needs to be included in the curriculum.

The above interpretations of results from analyzing the wages data, rely on trusting that the data provided is accurate and valid. The wages data is collected by a reputable organization, but we found that the data has obvious errors that should be corrected. This paper has illustrated procedures and guidelines to achieve valid data, and provides the code and details for it to be reproduced, and modified if deemed appropriate.

Having trustworthy data is imperative for statistics and data science education. Whenever one uses a textbook data set that is perceived as relating to the students lives, there will be interpretations made. The wages data is an example of this. Students will take away the interpretations from whatever is taught with the data, that wages increase with experi-

ence, education, race. If one uses data examples that are synthetic, instilled with our own inherent prejudices (for e.g. gender and race), or data that has been poorly processed containing errors, we as educators are being irresponsible because the societal message taught to students may be flawed. This paper demonstrates the process to produce a trustworthy data set for teaching.

6 Acknowledgements

7 Supplementary Materials

- **Codes**: R script to reproduce data tidying and cleaning is available in this page.

- **R Package**:`[CENSORED]` is a data container R package that contains 3 datasets, namely the high school mean hourly wage data, high school dropouts mean hourly wage data, and demographic data of the NLSY79 cohort. This package can be accessed here.

- **shiny app**: An interactive `shiny` web app to visualise the effect of selecting different weight threshold for substituting the wages data to its predicted value from a fit of the robust linear regression model. This app can be accessed here with the source code provided here.

8 Data Availability Statement

The authors confirm that the data supporting the findings of this study are available within the supplementary materials.

References

Andereson, Edgar. 1935. "The Irises of the Gaspe Peninsula." *Bulletin of the American Iris Society* 59: 2–5.

Bureau of Labor Statistics, U.S. Department of Labor. 2021a. "National Longitudinal Survey of Youth 1979 Cohort, 1979-2016 (Rounds 1-28)." Produced and distributed by the Center for Human Resource Research (CHRR), The Ohio State University. Columbus, OH, through `https://www.nlsinfo.org/bibliography-citing-nls-data`.

———. 2021b. "National Longitudinal Survey of Youth 1979 Cohort, Topical Guide to the Data." `https://www.nlsinfo.org/content/cohorts/nlsy79/topical-guide/employment/work-experience`.

Chang, Winston, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. 2020. *shiny: Web Application Framework for R.* `https://CRAN.R-project.org/package=shiny`.

Chatfield, C. 1985. "The Initial Examination of Data." *Journal of the Royal Statistical Society. Series A. General* 148 (3): 214–53.

Cooksey, Elizabeth C. 2017. "Using the National Longitudinal Surveys of Youth (Nlsy) to Conduct Life Course Analyses." In *Handbook of Life Course Health Development*, edited by Richard M. Lerner Neal Halfon Christoper B. Forrest, 561–77. Cham: Springer. `https://doi.org/https://doi.org/10.1007/978-3-319-47143-3_23`.

Dasu, Tamraparni, and Theodore Johnson. 2003. *Exploratory Data Mining and Data Cleaning.* Wiley Series in Probability and Statistics. Hoboken: WILEY.

Firke, Sam. 2020. *janitor: Simple Tools for Examining and Cleaning Dirty Data.* `https://CRAN.R-project.org/package=janitor`.

Fullilove, M. T. 1998. "Comment: Abandoning "Race" as a Variable in Public Health Research–an Idea Whose Time Has Come." *American Journal of Public Health* 88 (9): 1297–8.

Grimshaw, Scott D. 2015. "A Framework for Infusing Authentic Data Experiences Within Statistics Courses." *The American Statistician* 69 (4): 307–14. `https://doi.org/10.1080/00031305.2015.1081106`.

Henry, Lionel, and Hadley Wickham. 2020. *purrr: Functional Programming Tools.* `https://CRAN.R-project.org/package=purrr`.

Horst, Allison Marie, Alison Presmanes Hill, and Kristen B Gorman. 2020. *Palmerpenguins: Palmer Archipelago (Antarctica) Penguin Data.* `https://doi.org/10.5281/zenodo.3960218`.

Huebner, Marianne, Werner Vach, and Saskia le Cessie. 2016. "A Systematic Approach to Initial Data Analysis Is Good Research Practice." *The Journal of Thoracic and Cardiovascular Surgery* 151 (1): 25–27.

Ilk, Ozlem. 2004. "Exploratory Multivariate Longitudinal Data Analysis and Models for Multivariate Longitudinal Binary Data." PhD thesis, Iowa State University. `https://doi.org/10.31274/rtd-180813-11012`.

Kim, A. Y, C. Ismay, and J. Chunn. 2018. "The Fivethirtyeight R Package: "Tame Data" Principles for Introductory Statistics and Data Science Courses." *Technology Innovations in Statistics Education* 11 (1). `https://doi.org/10.5070/T511103589`.

Koller, Manuel. 2016. "robustlmm: An R Package for Robust Estimation of Linear Mixed-Effects Models." *Journal of Statistical Software* 75 (6): 1–24.

Moncrief, Marc. 2015. "By the Numbers - the Average Australian Doesn't Exist ... Not a Single One of Us Is 'Normal'." `https://bit.ly/smh-not-normal`.

Open Knowledge Foundation. 2021. "Open Definition. Defining Open in Open Data, Open Content, and Open Knowledge." 2021. `http://opendefinition.org/od/2.1/en/`.

Pedersen, Thomas Lin. 2020. *patchwork: The Composer of Plots.* `https://CRAN.R-project.org/package=patchwork`.

Pergamit, Michael R., Charles R. Pierret, Donna S. Rothstein, and Jonathan R. Veum. 2001. "Data Watch: The National Longitudinal Surveys." *The Journal of Economic Perspectives* 15 (2): 239–53.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. `https://www.R-project.org/`.

Singer, Judith D, and John B Willett. 2003. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence.* Oxford u.a: Oxford Univ. Pr.

Stodel, Megan. 2020. "Stop Using Iris." `https://www.meganstodel.com/posts/no-to-iris/`.

Tierney, Nicholas, Di Cook, and Tania Prvan. 2020. *brolgar: BRowse Over Longitudinal data Graphically and Analytically in R.* `https://github.com/njtierney/brolgar`.

Tukey, John W. (John Wilder). 1977. *Exploratory Data Analysis.* Addison-Wesley Series in Behavioral Science. Reading, Mass.: Addison-Wesley Pub. Co.

UCLA: Statistical Consulting Group. 2021. "Robust Regression | R Data Analysis Examples." February 2021. `https://stats.idre.ucla.edu/r/dae/robust-regression/`.

van der Loo, Mark, and Edwin de Jonge. 2018. *Statistical Data Cleaning with Applications in R.*

van der Loo, Mark P. J., and Edwin de Jonge. 2021. "Data Validation Infrastructure for R." *Journal of Statistical Software* 97 (10): 1–31. `https://doi.org/10.18637/jss.v097.i10`.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S.* Fourth. New York: Springer. `http://www.stats.ox.ac.uk/pub/MASS4`.

Wang, Earo, Dianne Cook, and Rob J Hyndman. 2020. "A New Tidy Data Structure to Support Exploration and Modeling of Temporal Data." *Journal of Computational and Graphical Statistics* 29 (3): 466–78. `https://doi.org/10.1080/10618600.2019.1695624`.

Wickham, Hadley. 2011. "The Split-Apply-Combine Strategy for Data Analysis." *Journal of Statistical Software, Articles* 40 (1): 1–29. `https://doi.org/10.18637/jss.v040.i01`.

———. 2014. "Tidy Data." *Journal of Statistical Software* 59 (10): 1–23.

———. 2016. *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. `https://ggplot2.tidyverse.org`.

———. 2019. *stringr: Simple, Consistent Wrappers for Common String Operations.* `https://CRAN.R-project.org/package=stringr`.

———. 2020. *tidyr: Tidy Messy Data.* `https://CRAN.R-project.org/package=`

`tidyr`.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. `https://doi.org/10.21105/joss.01686`.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2020. *dplyr: A Grammar of Data Manipulation*. `https://CRAN.R-project.org/package=dplyr`.

Wickham, Hadley, and Jim Hester. 2020. *readr: Read Rectangular Text Data*. `https://CRAN.R-project.org/package=readr`.

Wolpin, Kenneth I. 2005. "National Longitudinal Survey of Youth 1979 Cohort, 1979-2016 (Rounds 1-28)." Published by Bureau of Labor Statistics, U.S. Department of Labor. `https://www.bls.gov/opub/mlr/2005/02/art3full.pdf`.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. `http://www.crcpress.com/product/isbn/9781466561595`.

Xie, Yihui, Christophe Dervieux, and Emily Riederer. 2020. *R Markdown Cookbook*. Boca Raton, Florida: Chapman; Hall/CRC. `https://bookdown.org/yihui/rmarkdown-cookbook`.

Zhu, Hao. 2019. *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. `https://CRAN.R-project.org/package=kableExtra`.

▤ NLSY79 (`https://www.nlsinfo.org/investigator/pages/search?s=NLSY79`)

✔ The CASEID will be always be selected.

✔ The 3 recommended demographic variable (sample ID, race and sex) were selected.

For the remaining variables, we went to the "Variable Search" tab and select variables as follows

   ▷ Education, Training and Achievement Scores

      ▷ Education ▷ Summary measures ▷ All schools ▷ By year

         ▷ Dates of diploma or degree

            ✔ All variables named Q3-8A were selected.

         ▷ Highest grade completed

            ✔ Selected HGC_EVER, which is the highest grade respondent has ever reported completed and getting credit for.

   ▷ Employment

      ▷ Summary measures ▷ By job

         ▷ Hours worked

            ✔ All 447 primary variables in Hours worked were selected.

         ▷ Hourly wages

            ✔ All 156 variables in Hourly wages were selected.

      ▷ Summary measures ▷ Since date of last interview ▷ Weeks worked

         ✔ All 28 variables in Weeks worked were selected.

      ▷ Employer Roster ▷ Job dates ▷ Original start date

         ✔ Only selected the start date (Year) for the first job (E00101.02)

   ▷ Household, Geography & Demographics

      ▷ Demographics ▷ Basic demographics ▷ Date of birth

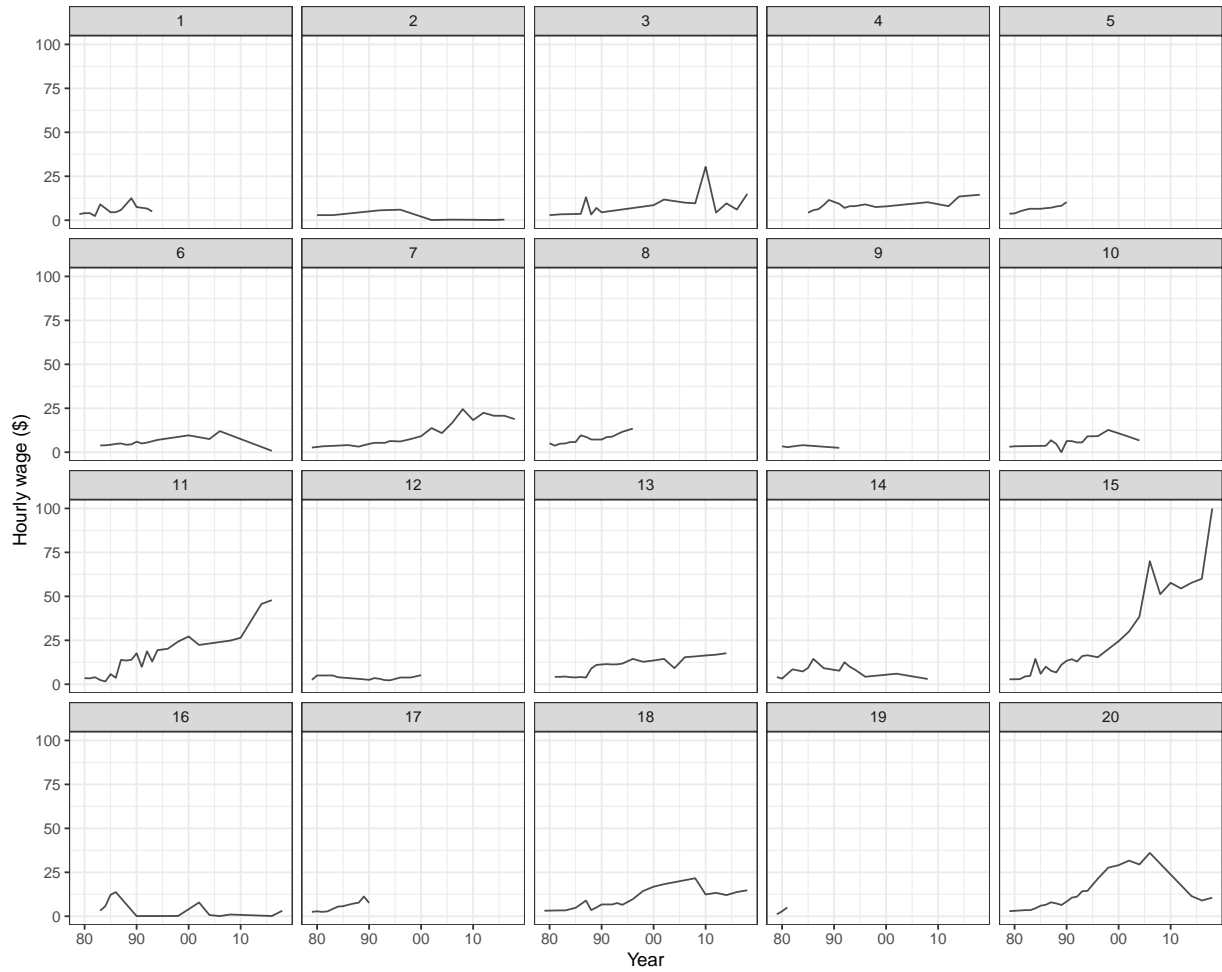   ✔ All 4 variables in Date of birth were selected.

Figure 2: Longitudinal profiles of wages for a random sample of 20 individuals in the refreshed data. Most, but not all, individuals here experienced increasing wages over time, and several have experienced considerable fluctuation in wages. Some individuals are only measured for a short period.
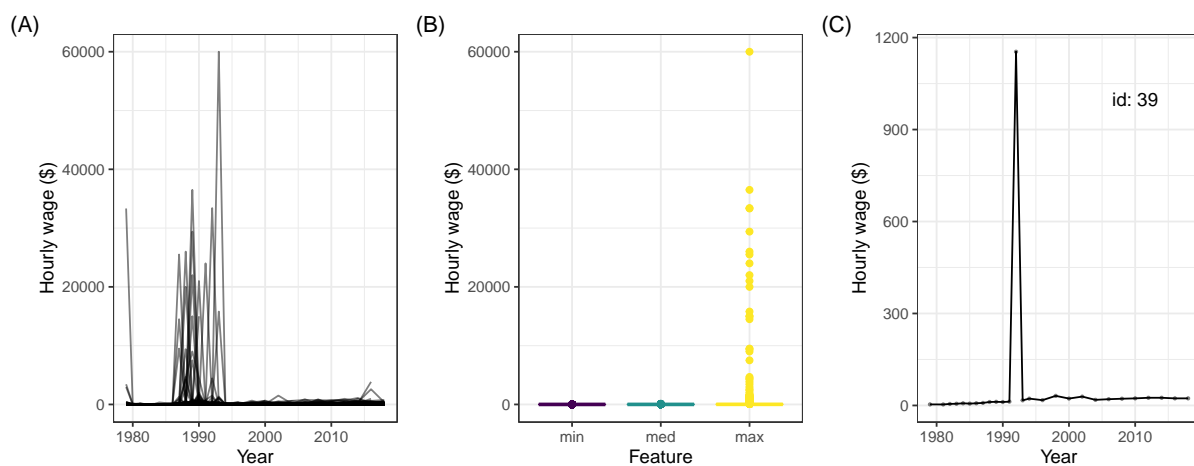
Figure 3: Summary plots to check the cleaned data reveal more cleaning is necessary: longitudinal profiles of wages for all individuals 1979-2018 (A), boxplots of minimum, median, and maximum wages of each individual (B), and one individual (id=39) with an unusual wage relative to their years of data (C). Some values of hourly wages are unbelievable, and some individuals have extemely unusual wages in some years.
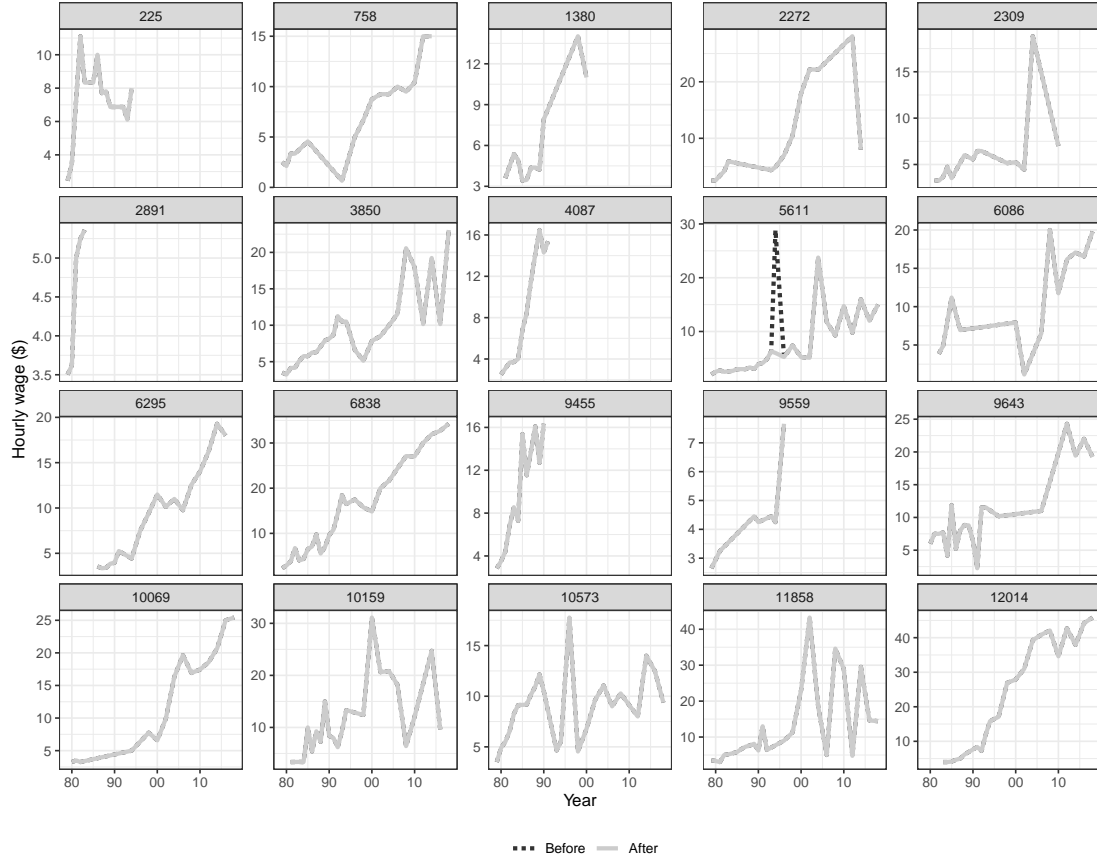
Figure 4: Comparison between the original (black dots) and the corrected (solid grey) mean hourly wage for a sample of individuals. A robust linear model prediction was used to correct mean hourly wages value. We can see that some extreme spikes, corresponding to implausible wages, have been replaced with values more similar to wages in neighboring years, but otherwise the profiles are not changed. Some spikes might remain when wage vaues are plausible.
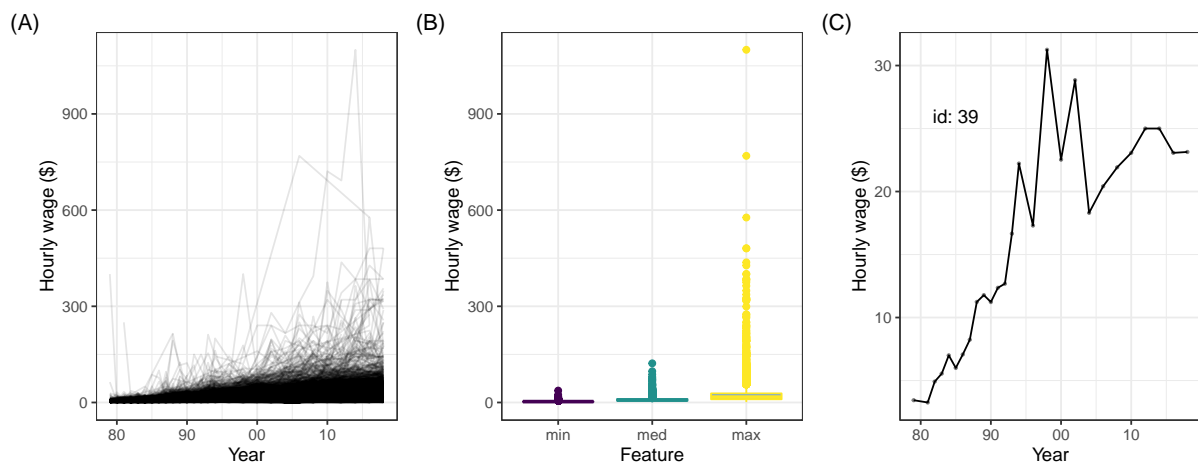
Figure 5: Re-make of the summary plots of the fully processed data suggest that it is now in a reasonable state: longitudinal profiles of wages for all individuals 1979-2018 (A), boxplots of minimum, median, and maximum wages of each individual (B), and one individual with an unusual wage relative to their years of data (C).
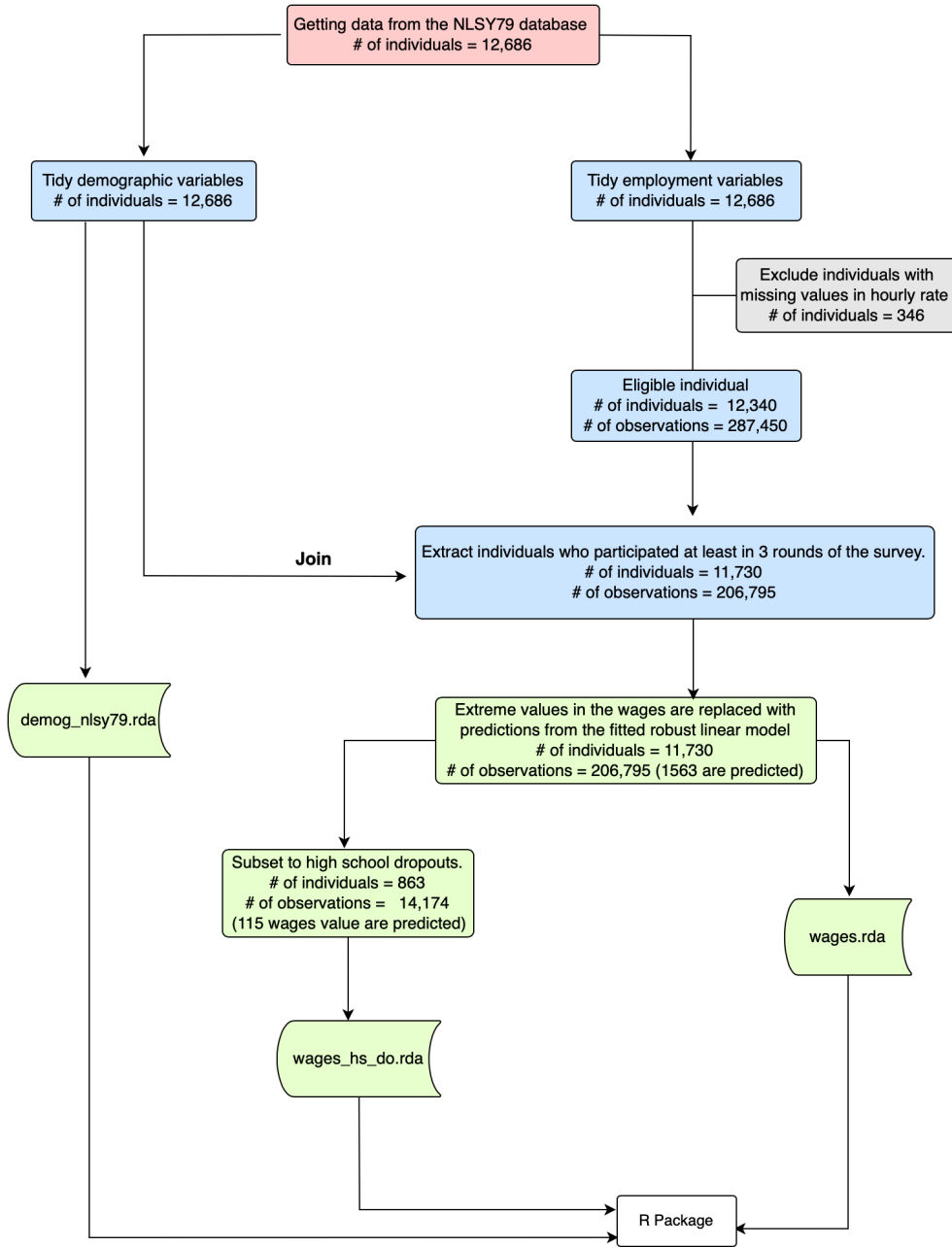
Figure 6: The stages of data cleaning from the raw data to get three datasets contained in [CENSORED]. "# of individuals" means the number of respondents included in each stage, while "# of observations" means the number of rows in the data. The color represents the stage of data cleaning in statistical value chain (van der Loo and de Jonge 2021). Pink, blue, and green represent the raw, input, and valid data, respectively.

Figure 7: Males dropouts' work experience in the original and refreshed data. Each point represents the work experience of individuals (identified by the ID) in the corresponding survey period. We can see that the correlation is reasonably high between them even there are different values for many individuals.
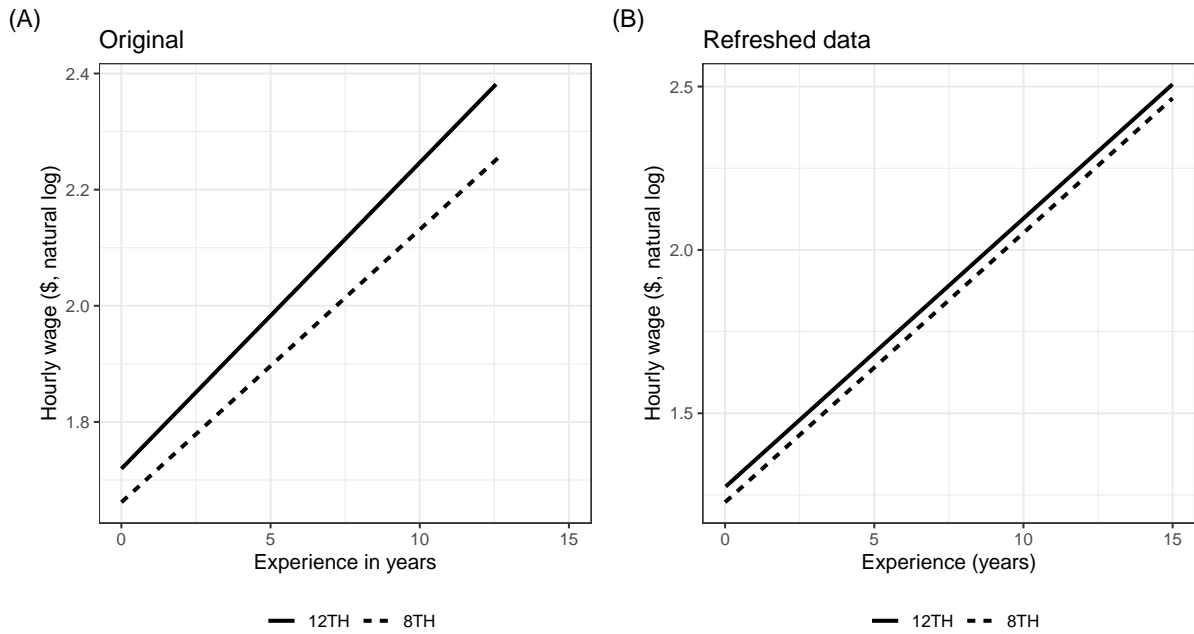
Figure 8: Comparison of original textbook example (A) with refreshed data (B). The original data was inflation-adjusted to 1990 prices and the individual's time of collection was converted to a length of experience in the workforce, which makes it difficult to precisely compare the two sets.