Tool 3.2 – Template for UI Wage Data Quality Control Memo

This template for a data quality control (QC) memo outlines key information that should be documented for a data analytics project that uses Unemployment Insurance (UI) wage data collected from a state or federal government source. The purposes of this type of QC memo are: (1) to document the quality of the data files for future reference, (2) to share information in an easy-to-digest format with other individuals you work with, and (3) to highlight potential issues you may want to ask the data provider about. There are several notes throughout the template and placeholders where information can be filled in.

The template assumes that those who use it already have a baseline knowledge of UI wage data. More information on UI wage data can be found in Expanding TANF Program Insights: A Toolkit for State and Local Agencies on How to Access, Link, and Analyze Unemployment Insurance Wage Data.[1]

Although this template is based on using UI wage data, many of these steps are relevant for checking the quality of any kind of data. Depending on your programming capacity, this kind of QC memo can often be autogenerated from statistical software, with the software producing the numbers presented in brackets.

[1] Yang et al. (2022).

[Programmer Name/Author]

[Date]

Project Background

[Project Name] is [description of project and purpose of using UI wage data].

This memo discusses the contents and quality of the UI wage data file received on [Date file received] that covers [First Quarter/Year on file] through [Last Quarter/Year on file]. This is the [number of files] received for the project.

File Locations

The data processing programs and datasets are located at the following paths shown in Table 3.2.1:

Table 3.2.1 File Locations

STAGE	TYPE	FILE PATH	PROGRAMMER
Checking Returned Data File	Program		
	Dataset		
Checking Data Updates	Program		
	Dataset		
Checking Person-Level File	Program		
	Dataset		

Key Decisions

Note: Below are examples of key data quality issues and decisions that teams may need to make depending on the quality of the data file. You can include decisions like these or any other decisions that you make here. This section provides a high-level summary of the quality of the data and the steps taken to address any issues. The decisions documented here will depend on the identified problems. Some examples are outlined below.

- [The file contained X number of exact duplicates (duplicates on all fields). For these duplicates, one of the records was dropped as they were thought to be a mistake.]

- [The file contained Y number of partial duplicates defined by having the same SSN, quarter, and earnings amount, but a different employer ID. For these duplicates, one record was dropped when it was thought the employer ID had changed over time.]

- [The file contained Z number of partial duplicates defined by having the same SSN, quarter, and employer ID, but a different earnings amount. For these duplicates, the mean of the two earnings amounts was taken and filled in on one of the records. The other record was dropped.]

- [The (number of outliers) records had earnings amounts greater than (threshold for outlier checks). At this time, no changes were made to these records. They were flagged so that the analysis could be run with and without them included, as a sensitivity check.]

Checking Raw Data File

The request file sent to the UI agency included [number of SSNs sent] SSNs. The returned file includes fields for [SSN, Quarter/Year, Employer ID, NAICS code, and earnings amount]. Data were returned at the [person-employer-quarter] level.

- **Record counts:**

  - There are [number of records] records and [number of unique SSNs] unique SSNs in the returned data file. [100*(number of missing SSNs)/(number of SSNs sent)] percent of the SSNs on the request file do not have a matched record in the returned file. [Insert text on whether this is expected, or unusually high/low.]

  - The earnings on this file total [sum of earnings on entire file].

- **Missing or invalid data:**

  - There are [number of missing earnings amounts] records with missing earnings amounts, and [number of invalid earnings amounts] records with negative or zero earnings amounts in the returned file. [If there are missing or invalid data, insert text on why and whether/what to do about it.]

    Note: In UI data, missing earnings amounts often simply mean that a person is not working in that quarter. Missing amounts do not necessarily indicate a problem if the incidence is within reasonable expectations given one's knowledge of the target population.

  - There are [number of missing quarters] records with missing quarters in the data. The missing quarters are [list of missing year/quarter variables].

  - There are [number of invalid quarters] records with quarters that fall outside of the date range expected on the returned file. These quarters account for [number or records with invalid quarters] records in the file. [Insert text on what to do about missing or invalid quarters.]

- **Duplicate records:**

  - There are [number of exact duplicates] exact duplicates in the file. [If there are exact duplicates, insert text on what was done to address them.]

Note: Exact duplicates are records that contain the same values for every field. For example, the same SSN, quarter, employer ID, and earnings amount.

Note: If the file has more than a few exact duplicates, you may want to check with the data provider for the source of these exact duplicates.

- There are [number of partial duplicates] partial duplicates by [fields used to check for partial duplicates] in the file.

  Note: Partial duplicates occur when there is more than one record that has the same value for some (but not all) of the fields. There may be different variations of partial duplicates, depending on what fields you have in the returned file. Look at each variation and insert text under this bullet about what you did for each type of partial duplicate. Examples of partial duplicates are:

  - Same SSN, Quarter, NAICS code, Employer ID, different earnings amount

  - Same SSN, Quarter, NAICS code, earnings, different Employer ID

  - Same SSN, Quarter, Employer ID, earnings amount, different NAICS code

  Note: Partial duplicates may appear when you receive multiple files from a UI agency that cover overlapping quarters (described more below).

- **Outliers:**

  - Per quarterly record: There are [number of earnings amounts ≥ $20,000] records with earnings amounts of $20,000 or more ([100*(number of high outliers)/(total number of records)] percent of all earnings records in the file). [Insert text about how you handled quarterly outliers.]

    Note: The definition of a potential high earnings outlier will vary depending on who is in your sample. You should determine what you would consider an outlier based on expectations for the study population.

  - There are [number of earnings amounts ≤ $20] earnings amounts of less than $20 in the file ([100*(number of high outliers)/(total number of records)] percent of all earnings records in the file). [Insert text about how you handled quarterly outliers.]

    Note: Teams often decide not to make any corrections to low earnings outliers, as they typically will not make a difference in the analysis and are considered plausible earnings amounts.

  - Per person: There are [number of summed earnings amounts across person/quarter ≥ $20,000] of individuals with more than $20,000 in total quarterly earnings, accounting for [100*(number of people with high outliers/total number of people on the returned file] of the individuals in the file. [Insert text about what to do about summed outliers by person/quarter.]

- **Consistency of record counts across quarters and SSN:**

  - The average number of individuals per quarter in all quarters is [average of counts of number of unique SSN over all quarters in the file]. [List of quarters with less than 95 percent of the average] have fewer than 95 percent of the average number of individuals per quarter. [List of quarters with more than 105 percent of the average] have more than 105 percent of the average number of individuals per quarter. [Insert text about whether these variations are expected or need more investigation.]

  - A more granular check of variation in earnings for the same individuals over time was also conducted. Earnings of individuals grouped by the first three digits of their SSN are [consistent or inconsistent, depending on variance of counts across quarters]. [Insert text about whether these variations are expected or need more investigation.]

- **Final record counts of cleaned returned data file:** After deleting invalid and duplicate records and handling outliers, the final returned data file has [number of records] earnings records that account for [total sum of earnings amounts] in earnings from [Start Year/Quarter] through [End Year/Quarter].

Checking Data Updates

Note: Teams often request and receive multiple shipments of UI wage data from UI agencies. This is often due to the availability of data (for example, some providers only maintain a certain number of quarters of data at a time). If possible, when requesting multiple files, it is good to request the same quarters of data on multiple files, as employers sometimes provide missing or updated records (earnings amounts or employer IDs) that will only be reflected in the latest file. For example, one file could cover Q1, 2000 to Q4, 2004 and a second file could cover Q1, 2002 to Q4, 2006. This section provides guidance on how to compare records from these overlapping time periods on a historical and current data file, as well as how to handle duplicates.

- **Partial duplicates:**

  Note: Count partial duplicates that are identical on every field except the file date. This is a check to make sure that earnings have not changed across files, so there should be a lot of these. For these sets of partial duplicates, remove record from the previous dataset.

  Note: Count partial duplicates that are identical on SSN and quarter but do not match on another field. The most common of these will be partial duplicates with different earnings amounts or different employer IDs. For the former, you may decide to keep the higher amount, the more recent amount, or take the mean of the two amounts. For the latter, you may decide to keep both records or only the more recent one.

- **Differences in earnings amounts:**

  Note: You will want to document differences in earnings amounts between the historical and current file. It is possible that employers corrected data they had previously submitted to the UI agency.

- **Record count:**

  Note: You should count the number of records after resolving duplicates and calculate the total sum of earnings in the merged, updated file. Then check that these amounts are consistent with what you expect.

Checking Person-Level File

- **Record counts:**

  - There are [number of records] records and [number of unique SSNs] individuals in the person-level file.

    Note: the number of records should match the number of individuals in the person-level file. These numbers should also match the number of valid SSNs that were sent to the UI agency in the request file.

  - The earnings on this file totals [sum of earnings on entire file].

    Note: the sum of earnings in this file should match the sum of earnings in the updated data file above.

- **Counts of outlier earnings:** About [100*(number of individuals with earnings outliers)/(number of total individuals on person-level file)] percent of all individuals have outlier earnings in the person-level file.

  Note: Table 3.2.2 shows an example of thresholds you could use to check for high earnings amounts. The threshold you use will depend on your study population.

Table 3.2.2 Counts of Outlier Earnings

EARNINGS AMOUNT (EQUAL TO OR GREATER THAN)	NUMBER OF OUTLIERS	NUMBER OF INDIVIDUALS WITH OUTLIER EARNINGS	INDIVIDUALS WITH OUTLIER EARNINGS (%)
$10,000	1695	334	7.38
$15,000	269	85	1.87
$20,000	93	26	0.57

- **Trends of average earnings and percent employed by quarter:** [Insert text to note changes in percent employed (with earnings) and average earnings (which include $0s for those who didn't earn in each quarter), and whether the variations from quarter to quarter are expected/reasonable.]

  Note: The two figures below, Figure 3.2.1 and Figure 3.2.2, are examples of how you can easily see trends in employment and earnings over time. The reasonableness of the trend you see will depend on your study sample and what you are analyzing. For example, when we use UI data to evaluate a training program, we often see an increase in employment and earnings in the quarters following participation.

  Note: The UI wage data file is a person-level file. Linking these data to a case-level file will require further data processing and quality control checks.

Figure 3.2.1 Percent with Earnings, 2005Q2 through 2009Q1

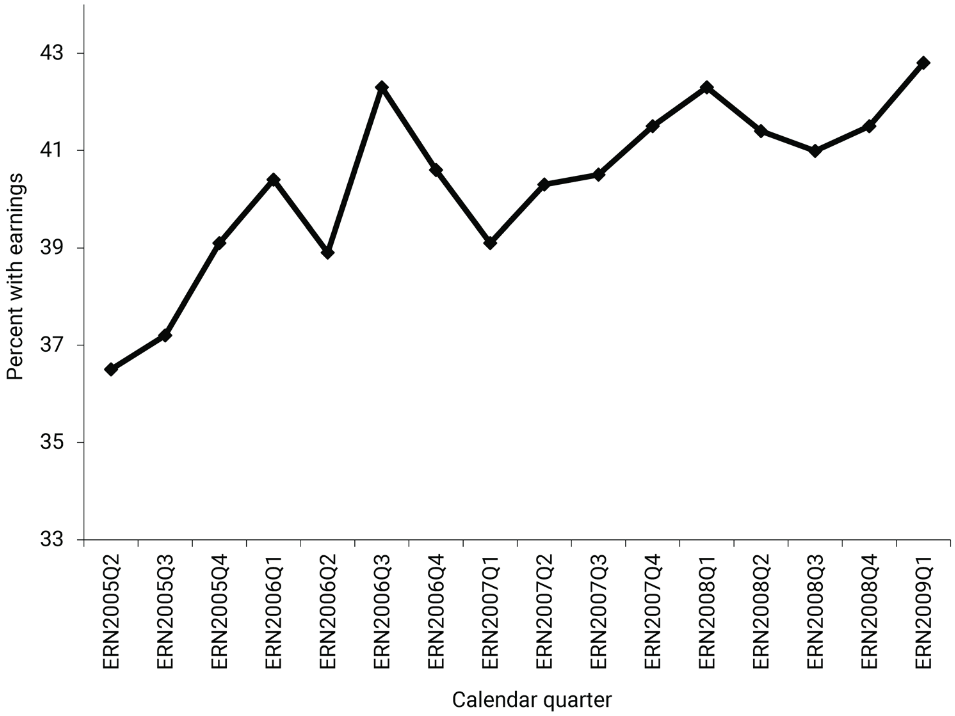
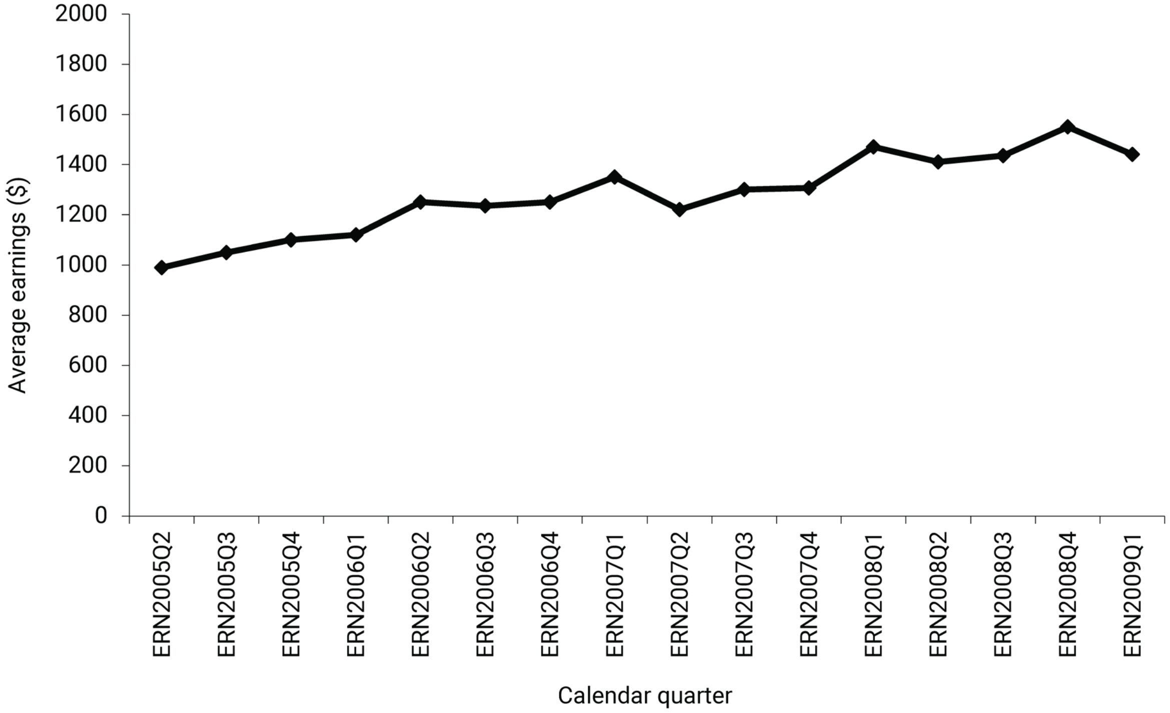


Figure 3.2.2 Average Earnings (in dollars), 2005Q2 through 2009Q1


Average earnings ($)
ERN2005Q2
ERN2005Q3
ERN2005Q4
ERN2006Q1
ERN2006Q2
ERN2006Q3
ERN2006Q4
ERN2007Q1
ERN2007Q2
ERN2007Q3
ERN2007Q4
ERN2008Q1
ERN2008Q2
ERN2008Q3
ERN2008Q4
ERN2009Q1
Calendar quarter