Data Extraction and Renaming (cl_00)

2025-06-14

Table of contents

Purpose 1
Setup 1
Load the data and rename it 1

Purpose

In this step, we load the raw dataset, convert it into a more flexible format, and rename it to make it easier to reference in later steps. Think of this as preparing your workspace before you start cleaning.

Setup

We'll start by loading the required libraries for reading Stata, Excel, and csv files:

```
# Load required libraries
library(haven)      # for reading/writing Stata files
library(openxlsx)   # for reading/writing xlsx files
library(readr)      # for reading/writing csv files
library(data.table) # good for large csv files
```

Load the data and rename it

Most datasets are available as .csv files, which we can use directly in R.

While R supports both .csv and .xlsx formats, .csv is generally preferred because it's lightweight, universally compatible, and easier to integrate with other tools and platforms[1].

```
# Load the source IPUMS data file; for example, if I am working with data in 2000, I will
data_2000 <- read_csv("data/source/your_file_name_here.csv")


write_csv(data_2000, "data/outcome/2000.csv")
```

If the files are large, you can also choose to use fread

```
data_2000 <- fread("data/source/your_file_name_here.csv")


fwrite(data_2000, "data/outcome/2000.csv")
```

If you have a list of files (I use BLS data as an example here), you can

```
years <- c(2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2

# Loop through each year
for(year in years) {

  # Define file paths
  input_file <- paste0("data/source/QCEW/fast_food sector/", year, "_Limited-service resta
  output_file <- paste0("data/outcome/", year, "_bls.csv")

  # Read CSV file
  data <- read_csv(input_file)

  # Save as CSV file with new name
  write_csv(data, output_file)
}
```

[1]For a comparison of `.csv` vs `.xlsx` in data workflows, see this LinkedIn article: CSV vs Excel: Pros and Cons