Creating Dataset with Labels

Pedro Henrique Brant

2025-07-13

Introduction

This file describes the brief process of merging the datasets available in the NHANES 2017-2018 website for usage in conducting the analysis in the PPCR 2025 Data Project by group 7.

All processes were done using R version 4.1

```
library(here)
library(tidyverse)
library(haven)
library(writexl)
```

Reading in the data

The raw NHANES data files were downloaded in SAS transport format (`.xpt`) and placed in the `Input` subfolder of our project directory. These files contain different domains such as demographics, laboratory results, and questionnaire responses.

We programmatically listed all `.xpt` files in the folder and read them into R as data frames. This approach ensures that any new `.xpt` files added to the folder in the future will automatically be included.

```
# list all .xpt files inside the "Input" subfolder
# If, in the future, you want to add more variables,
# download the .xpt file and add them in the Input folder

files <- list.files(here("Input"), pattern = "\\.xpt$", full.names = TRUE)

# read each .xpt file into a list of data frames
data_list <- lapply(files, read_xpt)
```

Merging the datasets

Each of the individual NHANES files includes a unique identifier variable called `SEQN`, which corresponds to the survey participant. We used this variable as the key to merge all datasets into a single analytic dataset.

We performed a series of full joins to ensure that all available data from each participant was retained, even if some variables were only present in certain files.

```
# start merging using full_join across all data frames in data_list
df <- reduce(data_list, full_join, by = "SEQN")
```

Selecting variables of interest

For this analysis, we prepared two distinct sets of variables:

1. **Default variable set:**
   Includes only the primary exposure, outcome, and essential covariates directly related to our main research question. This reduces the size of the dataset and focuses on the core variables needed for the primary analysis.

2. **Fregni + students variable set:**
   A broader list that combines variables suggested by the students with additional variables outlined by Professor Fregni in this document.

To keep the code flexible and maintainable, each list of variables is stored in an external text file. This makes it easy to update or expand the analysis in the future without modifying the underlying R code—new variables can simply be added to the corresponding text file.

The script then reads these lists, selects the relevant columns from the merged dataset, and checks that all requested variables are present. If any variables are missing, it issues a warning indicating which ones are absent so that they can be reviewed and addressed.

```
# ------------------------------
# Function to select variables and check for missing ones
# ------------------------------
# - Takes a dataset and a vector of variables to keep
# - Warns if any requested variables are missing from the dataset
# - Returns the dataset with only the selected variables
select_and_check <- function(data, vars_to_keep, label = NULL) {
  selected_data <- data %>% select(any_of(vars_to_keep))

  # Check for missing variables
  missing_vars <- setdiff(vars_to_keep, names(selected_data))
  if (length(missing_vars) > 0) {
    warning(
      "The following variables are missing",
      if (!is.null(label)) paste0(" in ", label) else "",
      ":\n",
      paste(missing_vars, collapse = ", "),
      "\n\nPlease check that the variable names are correct or that the appropriate NHANES datasets hav
    )
  }

  return(selected_data)
}

# ------------------------------
# Read in the lists of variables from external text files
# ------------------------------
# The first list contains only the main variables of interest
vars_to_keep <- read_lines(here("Input", "variable_list.txt"))

# The second list contains the variables suggested by the students plus
# those defined by Professor Fregni in his document
vars_to_keep_fregni_plus_students_suggestions <- read_lines(here("Input", "variable_list_fregni_plus_st
```

```r
# ------------------------------
# Create datasets with only the selected variables
# ------------------------------
df_selected <- select_and_check(df, vars_to_keep, label = "default variable set")
df_fregni_plus_students <- select_and_check(df, vars_to_keep_fregni_plus_students_suggestions, label =
```

Exporting the final datasets

Two sets of final merged datasets were created based on different lists of variables:

1. **Default variable set** — containing only the main exposure, outcome, and essential covariates as initially defined by the group.
2. **Fregni + students variable set** — a broader set including all variables proposed by Professor Fregni in his document, plus additional suggestions made by the students.

Each dataset was saved in three formats to facilitate different types of analyses:

- An Excel spreadsheet (`.xlsx`), which is convenient for manual inspection or sharing.
- A Stata file (`.dta`), allowing compatibility with Stata-based workflows.
- An R serialized file (`.rds`), which preserves data types and is efficient for direct loading into R.

All output files were written to the `Output` subfolder of the project directory.
The files follow a clear naming convention:

- `merged_dataset.*` for the default variable set.
- `merged_dataset_fregni_plus_students.*` for the expanded variable set.

```r
# This function takes a dataset and a base name,
# and writes it as .xlsx, .dta, and .rds files
# into the Output folder.
write_outputs <- function(data, base_name) {
  write_xlsx(data, here("Output", paste0(base_name, ".xlsx")))
  write_dta(data, here("Output", paste0(base_name, ".dta")))
  write_rds(data, here("Output", paste0(base_name, ".rds")))
}

# Write the default dataset outputs
write_outputs(df_selected, "merged_dataset")

# Write the extended dataset (Fregni + students) outputs
write_outputs(df_fregni_plus_students, "merged_dataset_fregni_plus_students")
```

Notes

This `.Rmd` file serves as a reproducible record of our data preparation process. Any updates to the input data or modifications to the processing steps should be documented here to maintain transparency and reproducibility.