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

```r
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.

```r
# 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.

```r
# 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 selected a subset of variables based on our research question and potential confounders. This reduces the size of the dataset and keeps only the relevant information.

The list of variables is stored in a single object, making it easy to update if future analyses require additional variables.

```r
# Define variables of interest
# If you want to add variables in the future, add them here
vars_to_keep <- c(
  "SEQN",       # key
  "DRQSDT1",
  "LBXHSCRP",
  "LBDHRPLC",
  "RIDAGEYR",
  "RIAGENDR",
  "RIDRETH1",
  "BMXBMI",
  "SMQ020",
  "PAQ605",
  "PAQ650",
  "SLD012"
)

# Select only these columns from the merged dataset
df <- df %>% select(any_of(vars_to_keep))

# Find which variables are missing
missing_vars <- setdiff(vars_to_keep, names(df))

# If any variables are missing, print a warning
if (length(missing_vars) > 0) {
  warning(
    "The following variables are missing from the final dataset:\n",
    paste(missing_vars, collapse = ", "),
    "\n\nPlease check that the variable names are correct or that the appropriate NHANES datasets contai
  )
}

rm(missing_vars)
```

Exporting the final dataset

The final merged dataset was saved in three formats to facilitate future analyses:

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

All output files were written to the `Output` subfolder of the project directory.

```r
# Write as .xlsx
write_xlsx(df, here("Output", "merged_dataset.xlsx"))

# Write as .dta (Stata)
write_dta(df, here("Output", "merged_dataset.dta"))

# Write as .rds (R)
write_rds(df, here("Output", "merged_dataset.rds"))
```

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.