EPID 674: Data Management in R

Homework 2

Kelly Bakulski, Lauren Middleton

Last compiled on December 20, 2021

Table of Contents

# Load Libraries

library(nhanesA)  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✓ ggplot2 3.3.5 ✓ purrr 0.3.4  
## ✓ tibble 3.1.5 ✓ dplyr 1.0.7  
## ✓ tidyr 1.1.4 ✓ stringr 1.4.0  
## ✓ readr 2.0.2 ✓ forcats 0.5.1

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(sjlabelled)

##   
## Attaching package: 'sjlabelled'

## The following object is masked from 'package:forcats':  
##   
## as\_factor

## The following object is masked from 'package:dplyr':  
##   
## as\_label

## The following object is masked from 'package:ggplot2':  
##   
## as\_label

library(here)

## here() starts at /cloud/project

# 1. Load data from 2013-2014 NHANES cycle and check the dimensions

# Load demographic data  
demog <- nhanes("DEMO\_H")

## Processing SAS dataset DEMO\_H ..

dim(demog)

## [1] 10175 47

# Load cognitive data  
cog <- nhanes("CFQ\_H")

## Processing SAS dataset CFQ\_H ..

dim(cog)

## [1] 1785 19

# Load cotinine data  
cotinine <- nhanes("COT\_H")

## Processing SAS dataset COT\_H ..

dim(cotinine)

## [1] 8913 5

# Load lead and cadmium data  
metals <- nhanes("PBCD\_H")

## Processing SAS dataset PBCD\_H ..

dim(metals)

## [1] 5932 17

The demographics dataset has 10175 rows and 47 columns. The cognitive dataset has 1785 rows and 19 columns. The cotinine dataset has 8913 rows and 5 columns. The metals dataset has 5932 rows and 17 columns.

# 2. Select columns of interest, join the datasets, and filter the participants.

## Select useful variables to keep

# For each dataset, select the variables of interest to keep  
demog\_clean <- demog %>%  
 select(SEQN,  
 RIAGENDR,  
 RIDAGEYR,  
 RIDRETH1,  
 INDFMPIR,  
 DMDEDUC2,  
 SDMVSTRA,  
 SDMVPSU  
 )   
  
cog\_clean <- cog %>%  
 select(SEQN,  
 CFDDS)  
  
cotinine\_clean <- cotinine %>%  
 select(SEQN,  
 LBXCOT)   
  
metals\_clean <- metals %>%  
 select(SEQN,  
 LBXBPB,  
 LBXBCD)

## Join the selected datasets together

# Join the four datasets  
nhanes\_join <- left\_join(demog\_clean, cog\_clean, by = "SEQN") %>%  
 left\_join(., cotinine\_clean, by = "SEQN") %>%  
 left\_join(., metals\_clean, by = "SEQN")  
  
  
dim(nhanes\_join)

## [1] 10175 12

#10175 12

## Filter the joined dataset to participants of interest

nhanes\_filter <- nhanes\_join %>%  
 filter(RIDAGEYR >= 60)  
  
dim(nhanes\_filter)

## [1] 1841 12

# 1841 12

The selected, joined, and filtered dataset has 1841 rows and 12 columns.

# 3. Create categorical variables from continuous variables

## 3a. Create a sex category variable from existing numeric variable

# Check the initial distribution  
table(nhanes\_filter$RIAGENDR, useNA = "always")

##   
## 1 2 <NA>   
## 874 967 0

nhanes\_homework <- nhanes\_filter %>%  
 mutate(sex = case\_when(RIAGENDR == 1 ~ "Male",  
 RIAGENDR == 2 ~ "Female"))  
  
# Optional, establish age\_groups as a factor variable and set reference level  
nhanes\_homework <- nhanes\_homework %>%   
 mutate(sex = relevel(factor(sex),  
 ref = "Male"))  
  
  
# Check the final distribution  
table(nhanes\_homework$sex, useNA = "always")

##   
## Male Female <NA>   
## 874 967 0

The number of participants in the “Male” category of the new sex variable is 874. The number of participants with the existing variable RIAGENDR value equal to 1 is 874. Yes, these values match.

## 3b. Create binary mild cognitive impairment variable based on continuous cognition values

# Calculate MCI cutoff based on: https://www.sciencedirect.com/science/article/pii/S0160412014003250?via%3Dihub  
  
# Check the initial distribution of the CFDDS variable  
summary(nhanes\_homework$CFDDS)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.00 33.00 46.00 45.95 58.00 105.00 249

#25th percentile is 33, min=5, median=46, max=93  
  
# Create the mild cognitive impairment variable  
nhanes\_homework <- nhanes\_homework %>%  
 mutate(MCI = case\_when(CFDDS > 33 ~ "No Impairment",  
 CFDDS <= 33 ~ "Mild Cognitive Impairment"),  
 MCI = relevel(factor(MCI),  
 ref = "No Impairment"))  
  
# Note, students were not asked to relevel this variable, no points are attributed to that step.  
  
# Optional, look at the top of these variables that were created.  
nhanes\_homework %>%  
 select(CFDDS, MCI) %>%  
 head()

## CFDDS MCI  
## 1 54 No Impairment  
## 2 63 No Impairment  
## 3 59 No Impairment  
## 4 79 No Impairment  
## 5 30 Mild Cognitive Impairment  
## 6 59 No Impairment

# Check the number of participants in each of these categories.  
table(nhanes\_homework$MCI, useNA = "always")

##   
## No Impairment Mild Cognitive Impairment <NA>   
## 1189 403 249

# 1189 No Impairment  
# 403 Mild Cognitive Impairment  
# 249 NA

The number of participants in the “No Impairment” category of the new variable is 1438. The number of participants in the “Mild Cognitive impairment” category of the new variable is 652.

## 3c. Make age category variable

# Optional, check the distribution of age in our dataset  
summary(nhanes\_homework$RIDAGEYR)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 60.00 64.00 69.00 70.01 76.00 80.00

# Make age groups of equal range  
nhanes\_homework <- nhanes\_homework %>%  
 mutate(age\_groups = cut\_interval(RIDAGEYR, n = 4))  
  
# Check how many partipants are in each group and what ages are in each group  
table(nhanes\_homework$age\_groups, useNA = "always")

##   
## [60,65] (65,70] (70,75] (75,80] <NA>   
## 617 398 319 507 0

# Optional, establish age\_groups as a factor variable and set reference level  
nhanes\_homework <- nhanes\_homework %>%  
 mutate(age\_groups = relevel(factor(age\_groups),  
 ref = "[60,65]"))

The lowest age group represents participants ages rlevels(nhanes\_homework$age\_groups)[1]`, which means greater than or equal to age 60 and less than or equal to age 65 years. The number of participants in the lowest age group is 617.

# Wrangle variables not included in the assignment for future use (not part of the homework assignment)

# Note, students were not asked to clean the education or race/ethnicity variables in this homework, so no points are attributed to these step. These steps are done to prepare the dataset for future assignments.  
  
# Create the education factor variable  
nhanes\_homework <- nhanes\_homework %>%  
 mutate(education = case\_when(DMDEDUC2 == 1 | DMDEDUC2 == 2 ~ "Less than high school",  
 DMDEDUC2 == 3 ~ "High school or GED",  
 DMDEDUC2 == 4 | DMDEDUC2 == 5 ~ "More than high school",  
 DMDEDUC2 == 7 | DMDEDUC2 == 9 ~ "Unknown"),  
 education = na\_if(education, "Unknown")) %>% #no missings  
 mutate(education = relevel(factor(education,  
 levels = c("Less than high school",  
 "High school or GED",  
 "More than high school")),  
 ref = "Less than high school"))  
  
# Create race/ethnicity factor variable  
nhanes\_homework <- nhanes\_homework %>%  
 mutate(race\_eth = case\_when(RIDRETH1 == 1 ~ "Mexican American",  
 RIDRETH1 == 2 ~ "Other Hispanic",  
 RIDRETH1 == 3 ~ "Non-Hispanic White",  
 RIDRETH1 == 4 ~ "Non-Hispanic Black",  
 RIDRETH1 == 5 ~ "Other Race")) %>%  
 mutate(race\_eth = relevel(factor(race\_eth),  
 ref = "Non-Hispanic White"))  
  
  
# Optional, relocate variable order in the dataset for future convenience  
colnames(nhanes\_homework)

## [1] "SEQN" "RIAGENDR" "RIDAGEYR" "RIDRETH1" "INDFMPIR"   
## [6] "DMDEDUC2" "SDMVSTRA" "SDMVPSU" "CFDDS" "LBXCOT"   
## [11] "LBXBPB" "LBXBCD" "sex" "MCI" "age\_groups"  
## [16] "education" "race\_eth"

nhanes\_columns <- c("SEQN", "RIAGENDR", "sex", "RIDAGEYR", "age\_groups", "RIDRETH1", "race\_eth", "INDFMPIR", "DMDEDUC2", "education", "SDMVSTRA", "SDMVPSU", "CFDDS", "MCI", "LBXCOT", "LBXBPB", "LBXBCD")   
nhanes\_homework <- relocate(nhanes\_homework, all\_of(nhanes\_columns))  
head(nhanes\_homework)

## SEQN RIAGENDR sex RIDAGEYR age\_groups RIDRETH1 race\_eth  
## 1 73557 1 Male 69 (65,70] 4 Non-Hispanic Black  
## 2 73559 1 Male 72 (70,75] 3 Non-Hispanic White  
## 3 73561 2 Female 73 (70,75] 3 Non-Hispanic White  
## 4 73564 2 Female 61 [60,65] 3 Non-Hispanic White  
## 5 73567 1 Male 65 [60,65] 3 Non-Hispanic White  
## 6 73571 1 Male 76 (75,80] 3 Non-Hispanic White  
## INDFMPIR DMDEDUC2 education SDMVSTRA SDMVPSU CFDDS  
## 1 0.84 3 High school or GED 112 1 54  
## 2 4.51 4 More than high school 109 1 63  
## 3 5.00 5 More than high school 116 2 59  
## 4 5.00 5 More than high school 114 1 79  
## 5 1.20 2 Less than high school 112 2 30  
## 6 5.00 5 More than high school 116 1 59  
## MCI LBXCOT LBXBPB LBXBCD  
## 1 No Impairment 6.510 NA NA  
## 2 No Impairment 0.052 1.45 0.22  
## 3 No Impairment 0.011 NA NA  
## 4 No Impairment 0.015 1.00 0.37  
## 5 Mild Cognitive Impairment 210.000 2.96 1.35  
## 6 No Impairment 0.011 NA NA

# 4. Save dataset

# Save dataset as an R object  
save(nhanes\_homework, file = here("nhanes\_homework.rda"))  
  
# Make sure you can load the dataset back into R  
load(file = here("nhanes\_homework.rda"))  
  
# Check the class of object that was loaded  
class(nhanes\_homework)

## [1] "data.frame"

The homework dataset is a data.frame class of object.