EPID674 Epidemiologic Data Analysis using R

Exploring Data with R

Kelly Bakulski

Last compiled on January 25, 2021

## Chapter 2, Exploring Data with R

# Identify base packages and install new packages

# Install packages. Do this only once.  
# Note, we already installed packages to this workspace. If working on your personal computer, will need to run this code  
install.packages("sas7bdat")  
install.packages("epiDisplay")  
install.packages("gmodels")  
install.packages("tidyverse")  
  
# To avoid installing every time: change set up in curly brackets to eval=FALSE

# Load packages. Load relevant packages every time you start a new R session  
library(sas7bdat)  
library(epiDisplay)

## Loading required package: foreign

## Loading required package: survival

## Loading required package: MASS

## Loading required package: nnet

library(gmodels)

##   
## Attaching package: 'gmodels'

## The following object is masked from 'package:epiDisplay':  
##   
## ci

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.0 ──

## ✓ ggplot2 3.3.3 ✓ purrr 0.3.4  
## ✓ tibble 3.0.5 ✓ dplyr 1.0.3  
## ✓ tidyr 1.1.2 ✓ stringr 1.4.0  
## ✓ readr 1.4.0 ✓ forcats 0.5.0

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## x ggplot2::alpha() masks epiDisplay::alpha()  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()  
## x dplyr::select() masks MASS::select()

sessionInfo() # check what versions of packages you just loaded

## R version 4.0.3 (2020-10-10)  
## Platform: x86\_64-pc-linux-gnu (64-bit)  
## Running under: Ubuntu 16.04.7 LTS  
##   
## Matrix products: default  
## BLAS: /usr/lib/atlas-base/atlas/libblas.so.3.0  
## LAPACK: /usr/lib/atlas-base/atlas/liblapack.so.3.0  
##   
## locale:  
## [1] LC\_CTYPE=C.UTF-8 LC\_NUMERIC=C LC\_TIME=C.UTF-8   
## [4] LC\_COLLATE=C.UTF-8 LC\_MONETARY=C.UTF-8 LC\_MESSAGES=C.UTF-8   
## [7] LC\_PAPER=C.UTF-8 LC\_NAME=C LC\_ADDRESS=C   
## [10] LC\_TELEPHONE=C LC\_MEASUREMENT=C.UTF-8 LC\_IDENTIFICATION=C   
##   
## attached base packages:  
## [1] stats graphics grDevices utils datasets methods base   
##   
## other attached packages:  
## [1] forcats\_0.5.0 stringr\_1.4.0 dplyr\_1.0.3 purrr\_0.3.4   
## [5] readr\_1.4.0 tidyr\_1.1.2 tibble\_3.0.5 ggplot2\_3.3.3   
## [9] tidyverse\_1.3.0 gmodels\_2.18.1 epiDisplay\_3.5.0.1 nnet\_7.3-14   
## [13] MASS\_7.3-53 survival\_3.2-7 foreign\_0.8-81 sas7bdat\_0.5   
##   
## loaded via a namespace (and not attached):  
## [1] gtools\_3.8.2 tidyselect\_1.1.0 xfun\_0.20 splines\_4.0.3   
## [5] haven\_2.3.1 lattice\_0.20-41 colorspace\_2.0-0 vctrs\_0.3.6   
## [9] generics\_0.1.0 htmltools\_0.5.1 yaml\_2.2.1 rlang\_0.4.10   
## [13] pillar\_1.4.7 withr\_2.4.0 glue\_1.4.2 DBI\_1.1.1   
## [17] dbplyr\_2.0.0 readxl\_1.3.1 modelr\_0.1.8 lifecycle\_0.2.0   
## [21] cellranger\_1.1.0 munsell\_0.5.0 gtable\_0.3.0 rvest\_0.3.6   
## [25] evaluate\_0.14 knitr\_1.30 fansi\_0.4.2 broom\_0.7.3   
## [29] Rcpp\_1.0.5 backports\_1.2.1 scales\_1.1.1 gdata\_2.18.0   
## [33] jsonlite\_1.7.2 fs\_1.5.0 hms\_1.0.0 digest\_0.6.27   
## [37] stringi\_1.5.3 grid\_4.0.3 cli\_2.2.0 tools\_4.0.3   
## [41] magrittr\_2.0.1 crayon\_1.3.4 pkgconfig\_2.0.3 ellipsis\_0.3.1   
## [45] Matrix\_1.2-18 xml2\_1.3.2 reprex\_0.3.0 lubridate\_1.7.9.2  
## [49] rstudioapi\_0.13 assertthat\_0.2.1 rmarkdown\_2.6 httr\_1.4.2   
## [53] R6\_2.5.0 compiler\_4.0.3

# Specify file directories

getwd() # Orient yourself to the default file path format on your computer

## [1] "/cloud/project"

directory <- "/cloud/project" #Class option when coding on RStudio Cloud  
  
# Three alternate options to customize and specify source data directory if programming on a personal computer (only need to do the one that works for you)  
# directory<-"M:/EPID674/Data" # Alternate option 1: location of the parent dataset in afs space of school computer  
# directory <- "/Volumes/GoogleDrive/My Drive/Teaching/EPID674/2020\_fall/EPID674\_Week2\_Class" # Alternate option 2: location of data on my Macbook  
# directory<-"C:/Users/bakulski/Documents/EPID674/Data" #Alternate option 3: location of data on a personal PC  
  
date <- format(Sys.Date(), "%Y%m%d")

# Load data from SAS, export data for R or excel

# Check file path  
file.path(directory, "nhanes3.sas7bdat")

## [1] "/cloud/project/nhanes3.sas7bdat"

# Load data from SAS  
nhanes <- read.sas7bdat(file.path(directory, "nhanes3.sas7bdat"))  
colnames(nhanes)

## [1] "strata" "seqn" "race" "sex" "age" "urban"   
## [7] "region" "pir" "psu" "wt\_mh" "sbp" "dbp"   
## [13] "bmi" "hematoc" "bpb" "chol" "trig" "scalc"   
## [19] "creat" "calc" "sodium" "potass" "educ" "smk"   
## [25] "packyrs" "diag\_ca" "diag\_dm" "diag\_ht" "alc" "phyact"   
## [31] "med\_ht" "htn" "d\_total" "pmon\_int" "pmon\_mec" "d\_cancer"  
## [37] "d\_cvd"

# Save data as an R object  
save(nhanes, file = file.path(directory, "nhanes3.rda"))  
load(file.path(directory, "nhanes3.rda")) # Try reading in the data that you just saved  
  
# Exporting data as a txt/csv file  
write.table(nhanes, file = file.path(directory, "nhanes3.txt"))  
write.csv(nhanes, file = file.path(directory, "nhanes3.csv"))  
  
# Try reading in the .csv that you just made  
nhanes.csv <- read.csv(file.path(directory, "nhanes3.csv"))  
names(nhanes.csv)

## [1] "X" "strata" "seqn" "race" "sex" "age"   
## [7] "urban" "region" "pir" "psu" "wt\_mh" "sbp"   
## [13] "dbp" "bmi" "hematoc" "bpb" "chol" "trig"   
## [19] "scalc" "creat" "calc" "sodium" "potass" "educ"   
## [25] "smk" "packyrs" "diag\_ca" "diag\_dm" "diag\_ht" "alc"   
## [31] "phyact" "med\_ht" "htn" "d\_total" "pmon\_int" "pmon\_mec"  
## [37] "d\_cancer" "d\_cvd"

# Clean up your workspace

## list all objects and remove the ones we do not want  
ls()

## [1] "date" "directory" "nhanes" "nhanes.csv"

rm(x, y, z)

## Warning in rm(x, y, z): object 'x' not found

## Warning in rm(x, y, z): object 'y' not found

## Warning in rm(x, y, z): object 'z' not found

# rm(list=ls()) #removes all objects  
ls()

## [1] "date" "directory" "nhanes" "nhanes.csv"

# Explore the NHANES dataset

# Explore the data set  
class(nhanes) # What type of object is it?

## [1] "data.frame"

dim(nhanes) # What are the dimensions?

## [1] 5074 37

names(nhanes) # What are the names?

## [1] "strata" "seqn" "race" "sex" "age" "urban"   
## [7] "region" "pir" "psu" "wt\_mh" "sbp" "dbp"   
## [13] "bmi" "hematoc" "bpb" "chol" "trig" "scalc"   
## [19] "creat" "calc" "sodium" "potass" "educ" "smk"   
## [25] "packyrs" "diag\_ca" "diag\_dm" "diag\_ht" "alc" "phyact"   
## [31] "med\_ht" "htn" "d\_total" "pmon\_int" "pmon\_mec" "d\_cancer"  
## [37] "d\_cvd"

colnames(nhanes) # Whate are the column names?

## [1] "strata" "seqn" "race" "sex" "age" "urban"   
## [7] "region" "pir" "psu" "wt\_mh" "sbp" "dbp"   
## [13] "bmi" "hematoc" "bpb" "chol" "trig" "scalc"   
## [19] "creat" "calc" "sodium" "potass" "educ" "smk"   
## [25] "packyrs" "diag\_ca" "diag\_dm" "diag\_ht" "alc" "phyact"   
## [31] "med\_ht" "htn" "d\_total" "pmon\_int" "pmon\_mec" "d\_cancer"  
## [37] "d\_cvd"

rownames(nhanes)[1:20] # What are the row names?

## [1] "1" "2" "3" "4" "5" "6" "7" "8" "9" "10" "11" "12" "13" "14" "15"  
## [16] "16" "17" "18" "19" "20"

head(nhanes) # What do the first 6 rows look like?

## strata seqn race sex age urban region pir psu wt\_mh sbp dbp bmi hematoc  
## 1 44 3 1 1 21 1 4 0.641 1 1735.14 120 67 25.5 43.80  
## 2 6 10 1 1 35 1 1 5.406 1 27769.56 130 82 29.4 46.70  
## 3 28 51 1 1 44 2 3 1.121 1 2730.56 130 86 44.4 49.80  
## 4 44 55 1 1 48 1 4 0.622 2 804.03 155 91 37.5 47.70  
## 5 48 63 1 2 66 1 2 NaN 2 4965.67 137 73 23.6 40.60  
## 6 9 70 2 1 63 1 2 0.702 1 2629.39 137 68 31.1 37.85  
## bpb chol trig scalc creat calc sodium potass educ smk packyrs diag\_ca  
## 1 5.0 268 174 9.5 1.1 70033 460 0.3099999 2 1 0.0 0  
## 2 2.0 225 109 9.4 1.2 70385 38591 0.5103998 3 1 0.0 0  
## 3 6.0 162 89 9.6 1.2 40040 3050 0.5599999 1 3 10.0 0  
## 4 15.5 212 479 10.6 1.0 10003 1321 0.8099999 1 3 2.0 0  
## 5 11.3 202 96 9.3 1.0 40021 5295 0.6499996 1 1 0.0 0  
## 6 7.2 186 300 8.8 1.3 80005 831 0.7099996 1 3 17.5 0  
## diag\_dm diag\_ht alc phyact med\_ht htn d\_total pmon\_int pmon\_mec d\_cancer  
## 1 0 1 1 1 0 1 0 203 203 0  
## 2 0 1 1 1 0 1 0 196 196 0  
## 3 0 1 NaN 1 0 1 0 216 215 0  
## 4 0 1 1 0 0 1 1 133 132 0  
## 5 1 0 0 1 0 0 0 185 184 0  
## 6 0 0 1 0 0 0 1 144 143 1  
## d\_cvd  
## 1 0  
## 2 0  
## 3 0  
## 4 1  
## 5 0  
## 6 0

nhanes[1:10, ] # What do the first 10 rows look like?

## strata seqn race sex age urban region pir psu wt\_mh sbp dbp bmi  
## 1 44 3 1 1 21 1 4 0.641 1 1735.14 120 67 25.5  
## 2 6 10 1 1 35 1 1 5.406 1 27769.56 130 82 29.4  
## 3 28 51 1 1 44 2 3 1.121 1 2730.56 130 86 44.4  
## 4 44 55 1 1 48 1 4 0.622 2 804.03 155 91 37.5  
## 5 48 63 1 2 66 1 2 NaN 2 4965.67 137 73 23.6  
## 6 9 70 2 1 63 1 2 0.702 1 2629.39 137 68 31.1  
## 7 43 71 1 2 37 1 4 NaN 1 1147.32 128 70 28.0  
## 8 14 84 2 2 23 2 3 3.340 1 2069.97 109 63 29.4  
## 9 26 90 1 1 28 2 4 0.764 1 3898.74 120 82 27.9  
## 10 39 97 1 2 86 1 3 0.761 1 583.88 143 85 26.4  
## hematoc bpb chol trig scalc creat calc sodium potass educ smk packyrs  
## 1 43.80 5.0 268 174 9.5 1.1 70033 460 0.30999994 2 1 0.0  
## 2 46.70 2.0 225 109 9.4 1.2 70385 38591 0.51039982 3 1 0.0  
## 3 49.80 6.0 162 89 9.6 1.2 40040 3050 0.55999994 1 3 10.0  
## 4 47.70 15.5 212 479 10.6 1.0 10003 1321 0.80999994 1 3 2.0  
## 5 40.60 11.3 202 96 9.3 1.0 40021 5295 0.64999962 1 1 0.0  
## 6 37.85 7.2 186 300 8.8 1.3 80005 831 0.70999956 1 3 17.5  
## 7 42.85 8.5 212 155 9.9 0.8 90024 182 5.50000000 1 3 5.0  
## 8 33.45 2.8 157 88 8.8 1.1 40035 77 0.50000000 2 1 0.0  
## 9 49.55 5.4 182 96 9.0 1.0 40048 2868 0.56999969 1 3 NaN  
## 10 NaN 4.1 194 89 9.0 1.2 30001 665 0.05999997 1 1 0.0  
## diag\_ca diag\_dm diag\_ht alc phyact med\_ht htn d\_total pmon\_int pmon\_mec  
## 1 0 0 1 1 1 0 1 0 203 203  
## 2 0 0 1 1 1 0 1 0 196 196  
## 3 0 0 1 NaN 1 0 1 0 216 215  
## 4 0 0 1 1 0 0 1 1 133 132  
## 5 0 1 0 0 1 0 0 0 185 184  
## 6 0 0 0 1 0 0 0 1 144 143  
## 7 0 0 0 0 0 0 0 0 202 202  
## 8 0 0 0 0 1 0 0 0 207 206  
## 9 0 0 0 0 0 0 0 0 188 187  
## 10 0 0 1 0 0 0 1 1 116 115  
## d\_cancer d\_cvd  
## 1 0 0  
## 2 0 0  
## 3 0 0  
## 4 0 1  
## 5 0 0  
## 6 1 0  
## 7 0 0  
## 8 0 0  
## 9 0 0  
## 10 0 1

# Explore individual variables in the NHANES dataset

# Explore the variables  
length(nhanes$age)

## [1] 5074

nhanes$age[1:10]

## [1] 21 35 44 48 66 63 37 23 28 86

length(nhanes$educ)

## [1] 5074

nhanes$educ[1:10]

## [1] 2 3 1 1 1 1 1 2 1 1

# Data description functions, numeric variables

mean(x = nhanes$age, trim = 0, na.rm = FALSE)

## [1] 48.74359

mean(na.rm = FALSE, x = nhanes$age, trim = 0)

## [1] 48.74359

mean(nhanes$age, 0, F)

## [1] 48.74359

mean(nhanes$age)

## [1] 48.74359

median(nhanes$age)

## [1] 46

quantile(nhanes$age)

## 0% 25% 50% 75% 100%   
## 20 32 46 65 90

quantile(nhanes$age, c(0.1, 0.9))

## 10% 90%   
## 24 77

sd(nhanes$age)

## [1] 19.29721

IQR(nhanes$age)

## [1] 33

summary(nhanes$age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 20.00 32.00 46.00 48.74 65.00 90.00

summary(nhanes)

## strata seqn race sex age   
## Min. : 1.00 Min. : 3 Min. :1.000 Min. :1.00 Min. :20.00   
## 1st Qu.:12.00 1st Qu.:11174 1st Qu.:1.000 1st Qu.:1.00 1st Qu.:32.00   
## Median :22.00 Median :32366 Median :1.000 Median :2.00 Median :46.00   
## Mean :23.26 Mean :27086 Mean :1.284 Mean :1.54 Mean :48.74   
## 3rd Qu.:33.00 3rd Qu.:42892 3rd Qu.:2.000 3rd Qu.:2.00 3rd Qu.:65.00   
## Max. :49.00 Max. :53594 Max. :2.000 Max. :2.00 Max. :90.00   
##   
## urban region pir psu   
## Min. :1.000 Min. :1.000 Min. : 0.000 Min. :1.000   
## 1st Qu.:1.000 1st Qu.:2.000 1st Qu.: 1.067 1st Qu.:1.000   
## Median :2.000 Median :3.000 Median : 2.011 Median :2.000   
## Mean :1.517 Mean :2.748 Mean : 2.461 Mean :1.511   
## 3rd Qu.:2.000 3rd Qu.:3.000 3rd Qu.: 3.442 3rd Qu.:2.000   
## Max. :2.000 Max. :4.000 Max. :11.889 Max. :2.000   
## NA's :487   
## wt\_mh sbp dbp bmi   
## Min. : 225.9 Min. : 80.0 Min. : 32.00 Min. :14.40   
## 1st Qu.: 2396.0 1st Qu.:112.0 1st Qu.: 68.00 1st Qu.:23.20   
## Median : 4495.9 Median :122.0 Median : 74.00 Median :26.40   
## Mean : 10416.3 Mean :126.3 Mean : 74.52 Mean :27.24   
## 3rd Qu.: 14750.3 3rd Qu.:137.0 3rd Qu.: 81.00 3rd Qu.:30.12   
## Max. :139744.9 Max. :237.0 Max. :134.00 Max. :68.50   
## NA's :10   
## hematoc bpb chol trig   
## Min. :19.10 Min. : 0.700 Min. : 83.0 Min. : 22.0   
## 1st Qu.:38.60 1st Qu.: 1.900 1st Qu.:175.0 1st Qu.: 80.0   
## Median :41.45 Median : 3.200 Median :204.0 Median : 115.0   
## Mean :41.43 Mean : 3.958 Mean :207.2 Mean : 147.2   
## 3rd Qu.:44.35 3rd Qu.: 5.000 3rd Qu.:234.0 3rd Qu.: 175.0   
## Max. :57.35 Max. :48.100 Max. :676.0 Max. :3616.0   
## NA's :61 NA's :58 NA's :68   
## scalc creat calc sodium   
## Min. : 7.300 Min. : 0.40 Min. : 6 Min. : 0   
## 1st Qu.: 9.000 1st Qu.: 0.90 1st Qu.:30013 1st Qu.: 968   
## Median : 9.300 Median : 1.00 Median :50049 Median : 3155   
## Mean : 9.271 Mean : 1.09 Mean :50852 Mean : 8121   
## 3rd Qu.: 9.500 3rd Qu.: 1.20 3rd Qu.:70241 3rd Qu.: 8626   
## Max. :13.900 Max. :13.90 Max. :90842 Max. :99712   
## NA's :132 NA's :132   
## potass educ smk packyrs   
## Min. : 0.01 Min. :1.000 Min. :1.00 Min. : 0.000   
## 1st Qu.: 0.27 1st Qu.:1.000 1st Qu.:1.00 1st Qu.: 0.000   
## Median : 0.55 Median :2.000 Median :1.00 Median : 0.000   
## Mean : 854.65 Mean :1.728 Mean :1.75 Mean : 9.579   
## 3rd Qu.: 0.83 3rd Qu.:2.000 3rd Qu.:3.00 3rd Qu.: 9.000   
## Max. :97800.00 Max. :3.000 Max. :3.00 Max. :456.000   
## NA's :33 NA's :218   
## diag\_ca diag\_dm diag\_ht alc   
## Min. :0.0000 Min. :0.00000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:0.00000 1st Qu.:0.0000 1st Qu.:0.0000   
## Median :0.0000 Median :0.00000 Median :0.0000 Median :0.0000   
## Mean :0.0745 Mean :0.09062 Mean :0.2814 Mean :0.4429   
## 3rd Qu.:0.0000 3rd Qu.:0.00000 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :1.0000 Max. :1.00000 Max. :1.0000 Max. :1.0000   
## NA's :9 NA's :39 NA's :132   
## phyact med\_ht htn d\_total   
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:1.0000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000   
## Median :1.0000 Median :0.0000 Median :0.0000 Median :0.0000   
## Mean :0.7704 Mean :0.1674 Mean :0.3821 Mean :0.2515   
## 3rd Qu.:1.0000 3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :1.0000 Max. :1.0000 Max. :1.0000 Max. :1.0000   
## NA's :45 NA's :5   
## pmon\_int pmon\_mec d\_cancer d\_cvd   
## Min. : 1.0 Min. : 0.0 Min. :0.00000 Min. :0.0000   
## 1st Qu.:151.0 1st Qu.:149.0 1st Qu.:0.00000 1st Qu.:0.0000   
## Median :171.0 Median :170.0 Median :0.00000 Median :0.0000   
## Mean :159.6 Mean :158.7 Mean :0.05031 Mean :0.1136   
## 3rd Qu.:194.0 3rd Qu.:193.0 3rd Qu.:0.00000 3rd Qu.:0.0000   
## Max. :218.0 Max. :217.0 Max. :1.00000 Max. :1.0000   
## NA's :5 NA's :5 NA's :5 NA's :5

# Create a factor variable (convert from a numeric variable)

class(nhanes$sex)

## [1] "numeric"

table(nhanes$sex)

##   
## 1 2   
## 2335 2739

sex1 <- factor(nhanes$sex, levels = c(1, 2), labels = c("male", "female")) ## creating sex1 from sex  
class(sex1)

## [1] "factor"

is.factor(nhanes$sex)

## [1] FALSE

is.factor(sex1)

## [1] TRUE

levels(sex1)

## [1] "male" "female"

nhanes$sex[1:10]

## [1] 1 1 1 1 2 1 2 2 1 2

sex1[1:10]

## [1] male male male male female male female female male female  
## Levels: male female

table(nhanes$sex)

##   
## 1 2   
## 2335 2739

table(sex1)

## sex1  
## male female   
## 2335 2739

nhanes <- cbind(nhanes, sex1) # if the new variable passes QC, you can bind it to the parent dataset  
  
### opposite way is as.numeric()  
sex2 <- as.numeric(sex1)  
sex2[1:10]

## [1] 1 1 1 1 2 1 2 2 1 2

table(sex2)

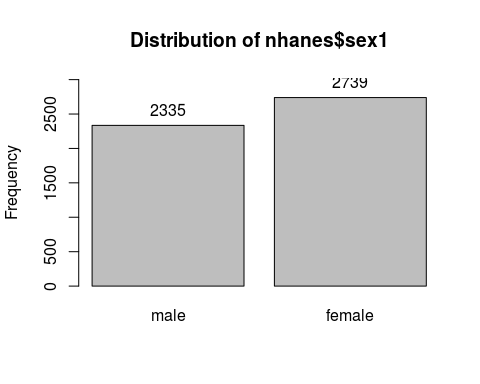
## sex2  
## 1 2   
## 2335 2739

# Calculate descriptive statistics on a categorical variable

table(nhanes$race)

##   
## 1 2   
## 3634 1440

tab1(nhanes$sex1)



## nhanes$sex1 :   
## Frequency Percent Cum. percent  
## male 2335 46 46  
## female 2739 54 100  
## Total 5074 100 100

table(nhanes$sex1, nhanes$race)

##   
## 1 2  
## male 1711 624  
## female 1923 816

CrossTable(nhanes$sex1, nhanes$race)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | Chi-square contribution |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 5074   
##   
##   
## | nhanes$race   
## nhanes$sex1 | 1 | 2 | Row Total |   
## -------------|-----------|-----------|-----------|  
## male | 1711 | 624 | 2335 |   
## | 0.894 | 2.257 | |   
## | 0.733 | 0.267 | 0.460 |   
## | 0.471 | 0.433 | |   
## | 0.337 | 0.123 | |   
## -------------|-----------|-----------|-----------|  
## female | 1923 | 816 | 2739 |   
## | 0.762 | 1.924 | |   
## | 0.702 | 0.298 | 0.540 |   
## | 0.529 | 0.567 | |   
## | 0.379 | 0.161 | |   
## -------------|-----------|-----------|-----------|  
## Column Total | 3634 | 1440 | 5074 |   
## | 0.716 | 0.284 | |   
## -------------|-----------|-----------|-----------|  
##   
##

# Do Exercise 2A

# Take a random sample of the dataset

## sample() to randomly sample  
samp <- sample(nhanes$seqn, 100)  
samp[1:20]

## [1] 14067 6244 16209 46204 9591 48062 50186 8424 45933 50387 18217 40457  
## [13] 4355 4348 16740 18255 16399 7700 3 12712

samp1 <- sample(nhanes$seqn, 100, replace = T)  
samp1[1:20]

## [1] 21371 6178 36728 47621 36441 38769 16996 14537 21821 21109 20288 51674  
## [13] 507 6651 32354 16655 4295 14244 38239 19172

# Create a subset of the dataset

# subset(): You can create a subset from a data frame using the subset() function.  
# For example, create a dataset for race==1 (White).  
table(nhanes$race)

##   
## 1 2   
## 3634 1440

nhanes.white <- subset(nhanes, nhanes$race == 1)  
dim(nhanes.white)

## [1] 3634 38

# To select only the 'age', 'bmi', and 'sex' variables among Black people  
nhanes.black <- subset(nhanes, nhanes$race == 2, select = c(age, bmi, sex))  
dim(nhanes.black)

## [1] 1440 3

# Use indexing []  
nhanes.young <- nhanes[nhanes$age < 30, ]  
dim(nhanes.young)

## [1] 1006 38

# Calculate descriptives on a subset of the dataset

summary(nhanes[, 1:5])

## strata seqn race sex age   
## Min. : 1.00 Min. : 3 Min. :1.000 Min. :1.00 Min. :20.00   
## 1st Qu.:12.00 1st Qu.:11174 1st Qu.:1.000 1st Qu.:1.00 1st Qu.:32.00   
## Median :22.00 Median :32366 Median :1.000 Median :2.00 Median :46.00   
## Mean :23.26 Mean :27086 Mean :1.284 Mean :1.54 Mean :48.74   
## 3rd Qu.:33.00 3rd Qu.:42892 3rd Qu.:2.000 3rd Qu.:2.00 3rd Qu.:65.00   
## Max. :49.00 Max. :53594 Max. :2.000 Max. :2.00 Max. :90.00

summary(nhanes[, c(1, 6, 11, 16, 21)])

## strata urban sbp chol   
## Min. : 1.00 Min. :1.000 Min. : 80.0 Min. : 83.0   
## 1st Qu.:12.00 1st Qu.:1.000 1st Qu.:112.0 1st Qu.:175.0   
## Median :22.00 Median :2.000 Median :122.0 Median :204.0   
## Mean :23.26 Mean :1.517 Mean :126.3 Mean :207.2   
## 3rd Qu.:33.00 3rd Qu.:2.000 3rd Qu.:137.0 3rd Qu.:234.0   
## Max. :49.00 Max. :2.000 Max. :237.0 Max. :676.0   
## NA's :58   
## sodium   
## Min. : 0   
## 1st Qu.: 968   
## Median : 3155   
## Mean : 8121   
## 3rd Qu.: 8626   
## Max. :99712   
##

summary(nhanes[, c("age", "sex", "race", "bpb", "sbp")])

## age sex race bpb   
## Min. :20.00 Min. :1.00 Min. :1.000 Min. : 0.700   
## 1st Qu.:32.00 1st Qu.:1.00 1st Qu.:1.000 1st Qu.: 1.900   
## Median :46.00 Median :2.00 Median :1.000 Median : 3.200   
## Mean :48.74 Mean :1.54 Mean :1.284 Mean : 3.958   
## 3rd Qu.:65.00 3rd Qu.:2.00 3rd Qu.:2.000 3rd Qu.: 5.000   
## Max. :90.00 Max. :2.00 Max. :2.000 Max. :48.100   
## sbp   
## Min. : 80.0   
## 1st Qu.:112.0   
## Median :122.0   
## Mean :126.3   
## 3rd Qu.:137.0   
## Max. :237.0

# Index within objects with [], create subsets, and you can run any R commands for subsets.  
mean(nhanes.white$age)

## [1] 50.57457

sd(nhanes.black$age)

## [1] 17.1088

table(nhanes$race, nhanes$sex)

##   
## 1 2  
## 1 1711 1923  
## 2 624 816

femalewhite <- nhanes[nhanes$sex == 2 & nhanes$race == 1, ]  
nrow(femalewhite)

## [1] 1923

summary(femalewhite$age)

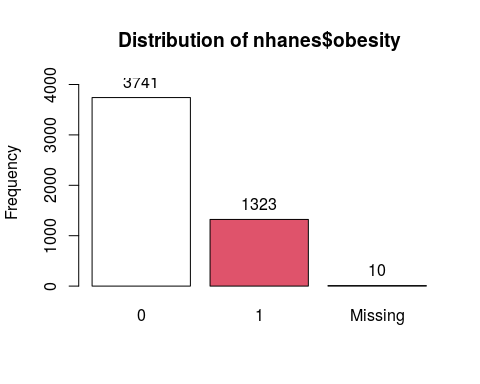
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 20.00 33.00 48.00 50.08 67.00 90.00

# Use the ifelse() function to create a categorical variable

# ifelse  
length(nhanes$bmi[nhanes$bmi >= 30])

## [1] 1333

nhanes$obesity <- ifelse(nhanes$bmi >= 30, 1, 0)  
tab1(nhanes$obesity)

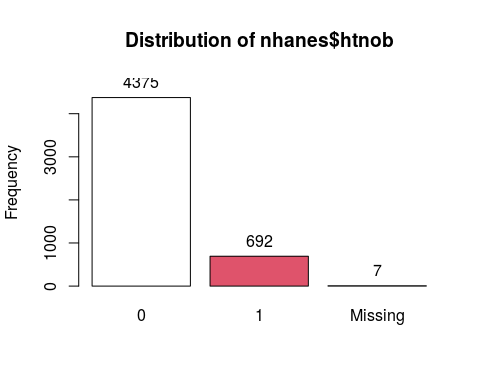


## nhanes$obesity :   
## Frequency %(NA+) %(NA-)  
## 0 3741 73.7 73.9  
## 1 1323 26.1 26.1  
## <NA> 10 0.2 0.0  
## Total 5074 100.0 100.0

nrow(nhanes[nhanes$htn == 1 & nhanes$bmi >= 30, ])

## [1] 699

nhanes$htnob <- ifelse(nhanes$htn == 1 & nhanes$bmi >= 30, 1, 0)  
tab1(nhanes$htnob)



## nhanes$htnob :   
## Frequency %(NA+) %(NA-)  
## 0 4375 86.2 86.3  
## 1 692 13.6 13.7  
## <NA> 7 0.1 0.0  
## Total 5074 100.0 100.0

# Create a categorical variable from a numeric variable

# Use the cut function to slice a numeric variable into groups

## create a categorical variable  
AGE5a <- cut(nhanes$age, 5) # Five equally spaced intervals (14 yrs)  
summary(AGE5a)

## (19.9,34] (34,48] (48,62] (62,76] (76,90.1]   
## 1541 1198 893 926 516

AGE5b <- cut(nhanes$age, quantile(nhanes$age, c(0, .2, .4, .6, .8, 1)), include.lowest = T) # Five groups with ~equal number of participants (quintiles)  
summary(AGE5b)

## [20,30] (30,40] (40,54] (54,69] (69,90]   
## 1130 962 1002 1030 950

AGE5c <- cut(nhanes$age, breaks = c(19, 40, 50, 60, 70, 90)) # Five groups based on pre-specified cut points  
summary(AGE5c)

## (19,40] (40,50] (50,60] (60,70] (70,90]   
## 2092 763 610 728 881

# Use the tidyverse to categorize a numeric variable

# Tidyverse: another way  
nhanes <- nhanes %>% # This is the pipe symbol  
 mutate(AGE5d = case\_when( # Use the mutate function to add a new variable  
 age < 30 ~ 1,  
 age >= 30 & age < 40 ~ 2,  
 age >= 40 & age < 50 ~ 3,  
 age >= 50 & age < 60 ~ 4,  
 age >= 60 ~ 5,  
 ))  
summary(nhanes$AGE5d)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 2.000 3.000 3.191 5.000 5.000

# Categorize a numeric variable the long way with indexing

# Long form: another way using brackets to index  
agecat <- nhanes$age  
agecat[nhanes$age < 30] <- "Twenties"  
agecat[nhanes$age >= 30 & nhanes$age < 40] <- "Thirties"  
agecat[nhanes$age >= 40 & nhanes$age < 50] <- "Fourties"  
agecat[nhanes$age >= 50 & nhanes$age < 60] <- "Fifties"  
agecat[nhanes$age >= 60 & nhanes$age < 70] <- "Sixties"  
agecat[nhanes$age >= 70 & nhanes$age < 80] <- "Seventies"  
agecat[nhanes$age >= 80] <- "Eighties"  
table(agecat)

## agecat  
## Eighties Fifties Fourties Seventies Sixties Thirties Twenties   
## 399 600 805 551 732 981 1006

# Convert categorical variables to numeric

age5c <- unclass(AGE5c) # Convert back into numeric  
summary(age5c)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 1.000 2.000 2.516 4.000 5.000

class(AGE5c)

## [1] "factor"

class(age5c)

## [1] "integer"

# Bind the variables that are useful to the nhanes dataframe for use later  
nhanes <- cbind(nhanes, AGE5b, AGE5c, age5c)

# Do Exercise 2B

# Remember to save your R script!

# Optional, save the nhanes object you modified  
save(nhanes, file = file.path(directory, "nhanes3.rda"))  
# Or you can save the workspace (includes all objects) with the name of your file  
# save.image(file="epid674.RData")

# To exit R

# q()  
## if you close R, you will be asked to save your workspace image