hw1Q1

黃章瑋

2025-10-11

#Q1. Among adults aged ≥20 years in the 2021–2023 NHANES, observe the association between BMI and mean systolic blood pressure (SBP) and does the association vary between sex ?

# ================= Q1: BMI & SBP Cleaning and Visualization =================  
# Load libraries -------------------------------------------------------------  
pkgs <- c("tidyverse","haven","janitor","stringr","scales","skimr","naniar")  
to\_install <- setdiff(pkgs, rownames(installed.packages()))  
if (length(to\_install)) install.packages(to\_install)  
invisible(lapply(pkgs, library, character.only = TRUE))

## Warning: package 'tidyverse' was built under R version 4.4.3

## Warning: package 'ggplot2' was built under R version 4.4.3

## Warning: package 'tibble' was built under R version 4.4.3

## Warning: package 'tidyr' was built under R version 4.4.3

## Warning: package 'readr' was built under R version 4.4.3

## Warning: package 'purrr' was built under R version 4.4.3

## Warning: package 'dplyr' was built under R version 4.4.3

## Warning: package 'stringr' was built under R version 4.4.3

## Warning: package 'forcats' was built under R version 4.4.3

## Warning: package 'lubridate' was built under R version 4.4.3

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.1 ✔ stringr 1.5.2  
## ✔ ggplot2 4.0.0 ✔ tibble 3.3.0  
## ✔ lubridate 1.9.4 ✔ tidyr 1.3.1  
## ✔ purrr 1.1.0   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

## Warning: package 'haven' was built under R version 4.4.3

## Warning: package 'janitor' was built under R version 4.4.3

##   
## Attaching package: 'janitor'  
##   
## The following objects are masked from 'package:stats':  
##   
## chisq.test, fisher.test

## Warning: package 'scales' was built under R version 4.4.3

##   
## Attaching package: 'scales'  
##   
## The following object is masked from 'package:purrr':  
##   
## discard  
##   
## The following object is masked from 'package:readr':  
##   
## col\_factor

## Warning: package 'skimr' was built under R version 4.4.3

## Warning: package 'naniar' was built under R version 4.4.3

##   
## Attaching package: 'naniar'  
##   
## The following object is masked from 'package:skimr':  
##   
## n\_complete

dir.create("outputs", showWarnings = FALSE)  
data\_dir <- "C:/Users/user/Desktop/raw data" # ← 改成你實際資料夾位置  
  
# Load data -----------------------------------------------------------------  
demo <- read\_xpt(file.path(data\_dir,"DEMO\_L.xpt")) %>% clean\_names()  
bpx <- read\_xpt(file.path(data\_dir,"BPXO\_L.xpt")) %>% clean\_names()  
bmx <- read\_xpt(file.path(data\_dir,"BMX\_L.xpt")) %>% clean\_names()  
  
skimr::skim(demo)

Data summary

|  |  |
| --- | --- |
| Name | demo |
| Number of rows | 11933 |
| Number of columns | 27 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| numeric | 27 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| seqn | 0 | 1.00 | 136344.00 | 3444.90 | 130378.00 | 133361.00 | 136344.00 | 139327.00 | 142310.0 | ▇▇▇▇▇ |
| sddsrvyr | 0 | 1.00 | 12.00 | 0.00 | 12.00 | 12.00 | 12.00 | 12.00 | 12.0 | ▁▁▇▁▁ |
| ridstatr | 0 | 1.00 | 1.74 | 0.44 | 1.00 | 1.00 | 2.00 | 2.00 | 2.0 | ▃▁▁▁▇ |
| riagendr | 0 | 1.00 | 1.53 | 0.50 | 1.00 | 1.00 | 2.00 | 2.00 | 2.0 | ▇▁▁▁▇ |
| ridageyr | 0 | 1.00 | 38.32 | 25.60 | 0.00 | 13.00 | 37.00 | 62.00 | 80.0 | ▇▃▃▅▆ |
| ridagemn | 11556 | 0.03 | 11.63 | 6.81 | 0.00 | 6.00 | 11.00 | 17.00 | 24.0 | ▆▇▇▆▆ |
| ridreth1 | 0 | 1.00 | 3.10 | 1.08 | 1.00 | 3.00 | 3.00 | 4.00 | 5.0 | ▂▂▇▂▂ |
| ridreth3 | 0 | 1.00 | 3.32 | 1.52 | 1.00 | 3.00 | 3.00 | 4.00 | 7.0 | ▃▇▂▁▂ |
| ridexmon | 3073 | 0.74 | 1.52 | 0.50 | 1.00 | 1.00 | 2.00 | 2.00 | 2.0 | ▇▁▁▁▇ |
| ridexagm | 9146 | 0.23 | 121.91 | 67.16 | 0.00 | 66.00 | 122.00 | 179.50 | 239.0 | ▆▇▇▇▇ |
| dmqmiliz | 3632 | 0.70 | 1.92 | 0.28 | 1.00 | 2.00 | 2.00 | 2.00 | 7.0 | ▇▁▁▁▁ |
| dmdborn4 | 19 | 1.00 | 1.16 | 0.36 | 1.00 | 1.00 | 1.00 | 1.00 | 2.0 | ▇▁▁▁▂ |
| dmdyrusr | 10058 | 0.16 | 7.33 | 15.83 | 1.00 | 3.00 | 6.00 | 6.00 | 99.0 | ▇▁▁▁▁ |
| dmdeduc2 | 4139 | 0.65 | 3.80 | 1.15 | 1.00 | 3.00 | 4.00 | 5.00 | 9.0 | ▂▇▅▁▁ |
| dmdmartz | 4141 | 0.65 | 1.78 | 3.10 | 1.00 | 1.00 | 1.00 | 2.00 | 99.0 | ▇▁▁▁▁ |
| ridexprg | 10430 | 0.13 | 2.24 | 0.49 | 1.00 | 2.00 | 2.00 | 3.00 | 3.0 | ▁▁▇▁▃ |
| dmdhhsiz | 0 | 1.00 | 3.24 | 1.70 | 1.00 | 2.00 | 3.00 | 4.00 | 7.0 | ▇▃▃▂▂ |
| dmdhrgnd | 7818 | 0.34 | 1.56 | 0.50 | 1.00 | 1.00 | 2.00 | 2.00 | 2.0 | ▆▁▁▁▇ |
| dmdhragz | 7809 | 0.35 | 2.54 | 0.64 | 1.00 | 2.00 | 2.00 | 3.00 | 4.0 | ▁▇▁▇▁ |
| dmdhredz | 8187 | 0.31 | 2.17 | 0.66 | 1.00 | 2.00 | 2.00 | 3.00 | 3.0 | ▂▁▇▁▅ |
| dmdhrmaz | 7913 | 0.34 | 1.38 | 0.68 | 1.00 | 1.00 | 1.00 | 2.00 | 3.0 | ▇▁▂▁▁ |
| dmdhsedz | 9806 | 0.18 | 2.28 | 0.69 | 1.00 | 2.00 | 2.00 | 3.00 | 3.0 | ▂▁▇▁▇ |
| wtint2yr | 0 | 1.00 | 27404.14 | 19449.16 | 4584.46 | 14331.75 | 21670.19 | 33831.33 | 170968.3 | ▇▂▁▁▁ |
| wtmec2yr | 0 | 1.00 | 27404.14 | 27962.96 | 0.00 | 0.00 | 21717.85 | 38341.15 | 227108.3 | ▇▂▁▁▁ |
| sdmvstra | 0 | 1.00 | 179.92 | 4.31 | 173.00 | 176.00 | 180.00 | 184.00 | 187.0 | ▇▇▇▇▇ |
| sdmvpsu | 0 | 1.00 | 1.49 | 0.50 | 1.00 | 1.00 | 1.00 | 2.00 | 2.0 | ▇▁▁▁▇ |
| indfmpir | 2041 | 0.83 | 2.71 | 1.67 | 0.00 | 1.18 | 2.50 | 4.50 | 5.0 | ▅▆▅▃▇ |

skimr::skim(bpx)

Data summary

|  |  |
| --- | --- |
| Name | bpx |
| Number of rows | 7801 |
| Number of columns | 12 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 1 |
| numeric | 11 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| bpaoarm | 0 | 1 | 0 | 1 | 147 | 3 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| seqn | 0 | 1.00 | 136349.49 | 3449.49 | 130378 | 133335 | 136382 | 139325 | 142310 | ▇▇▇▇▇ |
| bpaocsz | 190 | 0.98 | 3.52 | 0.67 | 2 | 3 | 4 | 4 | 5 | ▁▇▁▇▁ |
| bpxosy1 | 284 | 0.96 | 119.29 | 18.56 | 61 | 106 | 117 | 130 | 232 | ▁▇▂▁▁ |
| bpxodi1 | 284 | 0.96 | 72.75 | 11.90 | 33 | 64 | 72 | 80 | 142 | ▁▇▅▁▁ |
| bpxosy2 | 296 | 0.96 | 119.08 | 18.57 | 59 | 106 | 116 | 129 | 233 | ▁▇▃▁▁ |
| bpxodi2 | 296 | 0.96 | 72.09 | 11.85 | 32 | 64 | 71 | 79 | 139 | ▁▇▅▁▁ |
| bpxosy3 | 321 | 0.96 | 118.92 | 18.50 | 50 | 106 | 116 | 129 | 232 | ▁▇▅▁▁ |
| bpxodi3 | 321 | 0.96 | 71.81 | 11.77 | 24 | 64 | 71 | 79 | 136 | ▁▆▇▁▁ |
| bpxopls1 | 284 | 0.96 | 72.34 | 12.72 | 35 | 63 | 71 | 80 | 158 | ▂▇▂▁▁ |
| bpxopls2 | 296 | 0.96 | 73.09 | 12.78 | 32 | 64 | 72 | 81 | 141 | ▁▇▅▁▁ |
| bpxopls3 | 321 | 0.96 | 73.69 | 12.89 | 31 | 65 | 73 | 82 | 154 | ▁▇▃▁▁ |

skimr::skim(bmx)

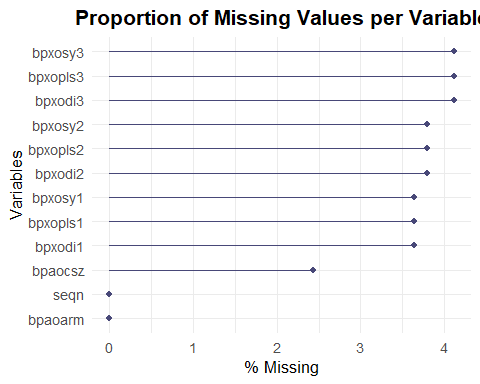
Data summary

|  |  |
| --- | --- |
| Name | bmx |
| Number of rows | 8860 |
| Number of columns | 22 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| numeric | 22 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| seqn | 0 | 1.00 | 136345.83 | 3453.78 | 130378.0 | 133319.75 | 136377.5 | 139336.2 | 142310.0 | ▇▇▇▇▇ |
| bmdstats | 0 | 1.00 | 1.13 | 0.50 | 1.0 | 1.00 | 1.0 | 1.0 | 4.0 | ▇▁▁▁▁ |
| bmxwt | 106 | 0.99 | 70.55 | 30.39 | 2.7 | 54.20 | 71.7 | 89.1 | 248.2 | ▃▇▂▁▁ |
| bmiwt | 8515 | 0.04 | 2.88 | 0.62 | 1.0 | 3.00 | 3.0 | 3.0 | 4.0 | ▁▁▁▇▁ |
| bmxrecum | 8406 | 0.05 | 84.33 | 14.06 | 48.5 | 73.48 | 84.7 | 96.1 | 118.8 | ▂▆▇▇▂ |
| bmirecum | 8842 | 0.00 | 1.00 | 0.00 | 1.0 | 1.00 | 1.0 | 1.0 | 1.0 | ▁▁▇▁▁ |
| bmxhead | 8790 | 0.01 | 41.93 | 2.80 | 34.4 | 40.20 | 42.4 | 44.0 | 46.5 | ▁▃▅▇▆ |
| bmihead | 8860 | 0.00 | NaN | NA | NA | NA | NA | NA | NA |  |
| bmxht | 361 | 0.96 | 159.66 | 19.86 | 79.1 | 154.40 | 163.6 | 172.1 | 200.7 | ▁▁▂▇▂ |
| bmiht | 8726 | 0.02 | 2.31 | 0.95 | 1.0 | 1.00 | 3.0 | 3.0 | 3.0 | ▅▁▁▁▇ |
| bmxbmi | 389 | 0.96 | 27.25 | 8.14 | 11.1 | 21.60 | 26.4 | 31.7 | 74.8 | ▆▇▂▁▁ |
| bmdbmic | 6368 | 0.28 | 2.56 | 0.88 | 1.0 | 2.00 | 2.0 | 3.0 | 4.0 | ▁▇▁▂▃ |
| bmxleg | 1525 | 0.83 | 38.13 | 3.86 | 24.9 | 35.50 | 38.1 | 40.8 | 51.6 | ▁▃▇▃▁ |
| bmileg | 8464 | 0.04 | 1.00 | 0.00 | 1.0 | 1.00 | 1.0 | 1.0 | 1.0 | ▁▁▇▁▁ |
| bmxarml | 292 | 0.97 | 35.11 | 6.18 | 10.0 | 33.60 | 36.5 | 39.0 | 49.2 | ▁▁▂▇▁ |
| bmiarml | 8660 | 0.02 | 1.00 | 0.00 | 1.0 | 1.00 | 1.0 | 1.0 | 1.0 | ▁▁▇▁▁ |
| bmxarmc | 298 | 0.97 | 30.56 | 7.37 | 12.0 | 26.40 | 31.2 | 35.4 | 63.3 | ▃▇▇▁▁ |
| bmiarmc | 8655 | 0.02 | 1.00 | 0.00 | 1.0 | 1.00 | 1.0 | 1.0 | 1.0 | ▁▁▇▁▁ |
| bmxwaist | 670 | 0.92 | 92.12 | 22.05 | 39.8 | 77.50 | 92.7 | 107.0 | 187.0 | ▃▇▆▁▁ |
| bmiwaist | 8513 | 0.04 | 1.00 | 0.00 | 1.0 | 1.00 | 1.0 | 1.0 | 1.0 | ▁▁▇▁▁ |
| bmxhip | 2084 | 0.76 | 106.26 | 14.66 | 69.9 | 96.40 | 103.7 | 113.5 | 187.1 | ▂▇▂▁▁ |
| bmihip | 8499 | 0.04 | 1.00 | 0.00 | 1.0 | 1.00 | 1.0 | 1.0 | 1.0 | ▁▁▇▁▁ |

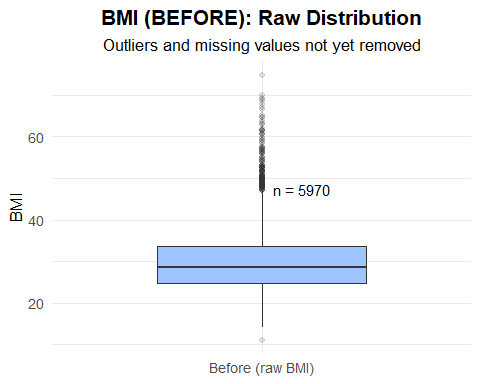
# Plot missing proportion visually  
gg\_miss\_var(bpx, show\_pct = TRUE) +  
 theme\_minimal(base\_size = 13) +  
 labs(title = "Proportion of Missing Values per Variable") +  
 theme(plot.title = element\_text(face = "bold", hjust = 0.5))



# Detect SBP/DBP columns ----------------------------------------------------  
sbp\_cols <- names(bpx)[str\_detect(names(bpx), "^bpxo?sy[1-3]$")]  
dbp\_cols <- names(bpx)[str\_detect(names(bpx), "^bpxo?di[1-3]$")]  
  
# Build raw BMI dataset -----------------------------------------------------  
bmi\_raw <- bmx %>% transmute(seqn, bmi\_raw = bmxbmi)  
demo <- demo %>%  
 mutate(riagendr = as.numeric(riagendr)) %>%  
 filter(is.na(riagendr) | riagendr %in% c(1, 2))  
demo\_sex <- demo %>%  
 transmute(seqn, age = ridageyr,  
 sex = factor(riagendr, levels=c(1,2), labels=c("Male","Female")))  
  
dat\_raw <- demo\_sex %>%  
 left\_join(bmi\_raw, by="seqn") %>%  
 filter(age >= 20) %>%  
 mutate(bmi\_raw = ifelse(is.nan(bmi\_raw), NA\_real\_, bmi\_raw))  
  
# BEFORE boxplot ------------------------------------------------------------  
bmi\_before\_df <- dat\_raw %>% transmute(stage = "Before (raw BMI)", value = bmi\_raw)  
x <- bmi\_before\_df$value  
qs <- quantile(x, c(.25,.75), na.rm = TRUE)  
iqr <- qs[2]-qs[1]  
upper\_whisker <- min(max(x, na.rm = TRUE), qs[2] + 1.5\*iqr)  
bmi\_before\_label\_y <- upper\_whisker + 0.05\*iqr  
bmi\_before\_N <- sum(!is.na(x))  
  
ggplot(bmi\_before\_df, aes(stage, value, fill = stage)) +  
 geom\_boxplot(width = 0.6, outlier.alpha = 0.15, fatten = 1.2) +  
 geom\_text(data = tibble(stage="Before (raw BMI)", y=bmi\_before\_label\_y, N=bmi\_before\_N),  
 aes(stage, y, label=paste0("n = ", N)), hjust = -0.2, size = 4) +  
 scale\_fill\_manual(values = c("Before (raw BMI)" = "#9EC5FE")) +  
 labs(title = "BMI (BEFORE): Raw Distribution",  
 subtitle = "Outliers and missing values not yet removed",  
 x = NULL, y = "BMI") +  
 theme\_minimal(base\_size = 13) +  
 theme(legend.position = "none",  
 plot.title = element\_text(face="bold", hjust=0.5),  
 plot.subtitle = element\_text(hjust=0.5))

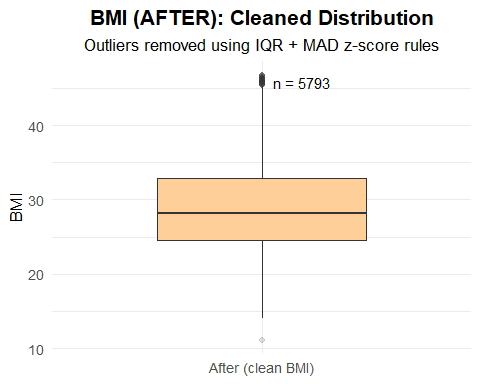
## Warning: Removed 1839 rows containing non-finite outside the scale range  
## (`stat\_boxplot()`).

## Warning: The `fatten` argument of `geom\_boxplot()` is deprecated as of ggplot2 4.0.0.  
## ℹ Please use the `median.linewidth` argument instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.

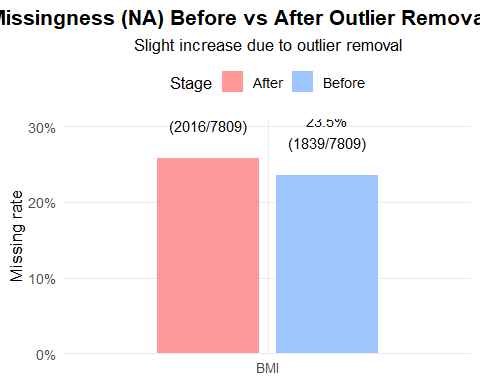


# Outlier cleaning ----------------------------------------------------------  
BMI\_LO <- 10; BMI\_HI <- 80  
bmi\_clean <- bmx %>%  
 transmute(seqn, bmxbmi) %>%  
 mutate(  
 q1 = quantile(bmxbmi, 0.25, na.rm=TRUE),  
 q3 = quantile(bmxbmi, 0.75, na.rm=TRUE),  
 iqr = q3 - q1,  
 lo\_iqr = q1 - 1.5\*iqr,  
 hi\_iqr = q3 + 1.5\*iqr,  
 med = median(bmxbmi, na.rm=TRUE),  
 madv = mad(bmxbmi, na.rm=TRUE),  
 z = ifelse(madv > 0, (bmxbmi - med)/(madv\*1.4826), 0),  
 flag = (bmxbmi < BMI\_LO | bmxbmi > BMI\_HI) |   
 (bmxbmi < lo\_iqr | bmxbmi > hi\_iqr) |   
 (abs(z) > 3.5),  
 bmxbmi\_clean = ifelse(flag, NA\_real\_, bmxbmi)  
 ) %>% select(seqn, bmxbmi\_clean)  
  
# Cleaned dataset -----------------------------------------------------------  
dat\_clean <- demo\_sex %>%  
 left\_join(bmi\_clean, by="seqn") %>%  
 filter(age >= 20) %>%  
 mutate(bmxbmi\_clean = ifelse(is.nan(bmxbmi\_clean), NA\_real\_, bmxbmi\_clean))  
  
# AFTER boxplot -------------------------------------------------------------  
bmi\_after\_df <- dat\_clean %>% transmute(stage = "After (clean BMI)", value = bmxbmi\_clean)  
x <- bmi\_after\_df$value  
qs <- quantile(x, c(.25,.75), na.rm = TRUE)  
iqr <- qs[2]-qs[1]  
upper\_whisker <- min(max(x, na.rm = TRUE), qs[2] + 1.5\*iqr)  
bmi\_after\_label\_y <- upper\_whisker + 0.05\*iqr  
bmi\_after\_N <- sum(!is.na(x))  
  
ggplot(bmi\_after\_df, aes(stage, value, fill = stage)) +  
 geom\_boxplot(width = 0.6, outlier.alpha = 0.15, fatten = 1.2) +  
 geom\_text(data = tibble(stage="After (clean BMI)", y=bmi\_after\_label\_y, N=bmi\_after\_N),  
 aes(stage, y, label=paste0("n = ", N)), hjust = -0.2, size = 4) +  
 scale\_fill\_manual(values = c("After (clean BMI)" = "#FFCF99")) +  
 labs(title = "BMI (AFTER): Cleaned Distribution",  
 subtitle = "Outliers removed using IQR + MAD z-score rules",  
 x = NULL, y = "BMI") +  
 theme\_minimal(base\_size = 13) +  
 theme(legend.position = "none",  
 plot.title = element\_text(face="bold", hjust=0.5),  
 plot.subtitle = element\_text(hjust=0.5))

## Warning: Removed 2016 rows containing non-finite outside the scale range  
## (`stat\_boxplot()`).

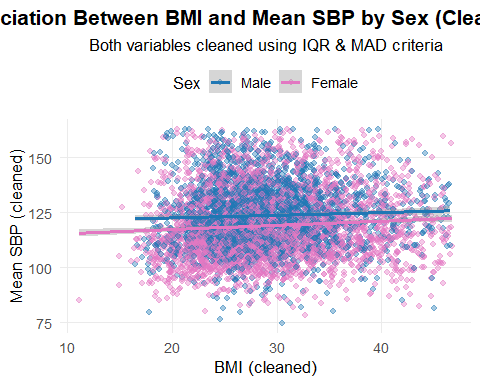


# Missingness comparison ----------------------------------------------------  
miss\_before <- tibble(  
 stage = "Before",  
 variable = "BMI",  
 n\_missing = sum(is.na(dat\_raw$bmi\_raw)),  
 n\_total = nrow(dat\_raw)  
) %>% mutate(p\_missing = n\_missing / n\_total)  
  
miss\_after <- tibble(  
 stage = "After",  
 variable = "BMI",  
 n\_missing = sum(is.na(dat\_clean$bmxbmi\_clean)),  
 n\_total = nrow(dat\_clean)  
) %>% mutate(p\_missing = n\_missing / n\_total)  
  
miss\_long <- bind\_rows(miss\_before, miss\_after)  
  
pos <- position\_dodge(width = 0.7)  
ggplot(miss\_long, aes(variable, p\_missing, fill = stage)) +  
 geom\_col(width = 0.6, position = pos) +  
 geom\_text(aes(label = paste0(scales::percent(p\_missing, 0.1),  
 "\n(", n\_missing, "/", n\_total, ")")),  
 position = pos, vjust = -0.8, size = 4) +  
 scale\_y\_continuous(labels = scales::percent, expand = expansion(mult = c(0, 0.2))) +  
 scale\_fill\_manual(values = c("Before" = "#9EC5FE", "After" = "#FF9999")) +  
 labs(title = "Missingness (NA) Before vs After Outlier Removal (BMI)",  
 subtitle = "Slight increase due to outlier removal",  
 x = NULL, y = "Missing rate", fill = "Stage") +  
 theme\_minimal(base\_size = 13) +  
 theme(panel.grid.minor = element\_blank(),  
 plot.title = element\_text(face = "bold", hjust = 0.5),  
 plot.subtitle = element\_text(hjust = 0.5),  
 legend.position = "top")



# SBP mean (before cleaning) ------------------------------------------------  
sbp\_raw <- bpx %>% transmute(seqn, sbp\_mean\_raw = rowMeans(select(., all\_of(sbp\_cols)), na.rm = TRUE))  
dat\_sbp\_raw <- dat\_raw %>%  
 left\_join(sbp\_raw, by = "seqn") %>%  
 filter(!is.na(bmi\_raw) & !is.na(sbp\_mean\_raw))  
  
# SBP outlier cleaning ------------------------------------------------------  
SBP\_LO <- 70; SBP\_HI <- 260  
sbp\_clean <- bpx %>%  
 transmute(seqn, across(all\_of(sbp\_cols))) %>%  
 mutate(  
 sbp\_all = pmap\_dbl(across(all\_of(sbp\_cols)), ~ mean(c(...), na.rm = TRUE)),  
 q1 = quantile(sbp\_all, 0.25, na.rm = TRUE),  
 q3 = quantile(sbp\_all, 0.75, na.rm = TRUE),  
 iqr = q3 - q1,  
 lo\_iqr = q1 - 1.5 \* iqr,  
 hi\_iqr = q3 + 1.5 \* iqr,  
 med = median(sbp\_all, na.rm = TRUE),  
 madv = mad(sbp\_all, na.rm = TRUE),  
 z = ifelse(madv > 0, (sbp\_all - med) / (madv \* 1.4826), 0),  
 flag = (sbp\_all < SBP\_LO | sbp\_all > SBP\_HI) |   
 (sbp\_all < lo\_iqr | sbp\_all > hi\_iqr) |   
 (abs(z) > 3.5),  
 sbp\_mean\_clean = ifelse(flag, NA\_real\_, sbp\_all)  
 ) %>% select(seqn, sbp\_mean\_clean)  
  
# Combine BMI + SBP clean dataset ------------------------------------------  
dat\_final <- dat\_clean %>%  
 left\_join(sbp\_clean, by = "seqn") %>%  
 filter(!is.na(bmxbmi\_clean) & !is.na(sbp\_mean\_clean))  
  
# Scatter plot: BMI vs SBP -------------------------------------------------  
ggplot(dat\_final, aes(x = bmxbmi\_clean, y = sbp\_mean\_clean, color = sex)) +  
 geom\_point(alpha = 0.4, size = 1.8) +  
 geom\_smooth(method = "lm", se = TRUE, lwd = 1.2) +  
 labs(title = "Association Between BMI and Mean SBP by Sex (Cleaned Data)",  
 subtitle = "Both variables cleaned using IQR & MAD criteria",  
 x = "BMI (cleaned)", y = "Mean SBP (cleaned)", color = "Sex") +  
 scale\_color\_manual(values = c("Male" = "#1F77B4", "Female" = "#E377C2")) +  
 theme\_minimal(base\_size = 13) +  
 theme(panel.grid.minor = element\_blank(),  
 plot.title = element\_text(face = "bold", hjust = 0.5),  
 plot.subtitle = element\_text(hjust = 0.5),  
 legend.position = "top")

## `geom\_smooth()` using formula = 'y ~ x'



#Q2. Among all the subjects in 2021-2023 NHANES dataset, observe the distribution of BMI in different races and education levels

library(knitr)

## Warning: package 'knitr' was built under R version 4.4.3

# 檢查原始教育變項  
demo %>% count(dmdeduc2)

## # A tibble: 7 × 2  
## dmdeduc2 n  
## <dbl> <int>  
## 1 1 373  
## 2 2 666  
## 3 3 1749  
## 4 4 2370  
## 5 5 2625  
## 6 9 11  
## 7 NA 4139

# 重編教育變項  
dat\_edu <- demo %>%  
 transmute(  
 seqn,  
 age = ridageyr,  
 EDU = case\_when(  
 dmdeduc2 %in% 1:5 ~ dmdeduc2,  
 TRUE ~ NA\_real\_  
 )  
 ) %>%  
 mutate(  
 EDU = factor(  
 EDU,  
 levels = 1:5,  
 labels = c("<9th grade", "9–11th grade", "High school/GED",  
 "Some college/AA", "College or above")  
 )  
 ) %>%  
 left\_join(dat\_clean %>% select(seqn, bmxbmi\_clean), by = "seqn") %>%  
 drop\_na(EDU, bmxbmi\_clean)  
  
# 教育分布表  
edu\_dist <- dat\_edu %>%  
 count(EDU) %>%  
 mutate(prop = n / sum(n)) %>%  
 rename(category = EDU)  
  
kable(edu\_dist, digits = 3, caption = "Distribution of Educational Attainment (EDU)")

Distribution of Educational Attainment (EDU)

| category | n | prop |
| --- | --- | --- |
| <9th grade | 278 | 0.048 |
| 9–11th grade | 457 | 0.079 |
| High school/GED | 1227 | 0.212 |
| Some college/AA | 1749 | 0.302 |
| College or above | 2079 | 0.359 |

# 檢查原始種族變項  
demo %>% count(ridreth3)

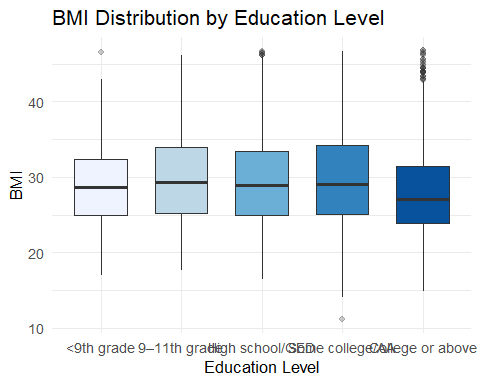
## # A tibble: 6 × 2  
## ridreth3 n  
## <dbl> <int>  
## 1 1 1117  
## 2 2 1373  
## 3 3 6217  
## 4 4 1597  
## 5 6 681  
## 6 7 948

# 重編種族變項  
dat\_race <- demo %>%  
 transmute(  
 seqn,  
 age = ridageyr,  
 Race = case\_when(  
 ridreth3 %in% 1:7 ~ ridreth3,  
 TRUE ~ NA\_real\_  
 )  
 ) %>%  
 mutate(  
 Race = factor(  
 Race,  
 levels = 1:7,  
 labels = c("Mexican American", "Other Hispanic", "Non-Hispanic White",  
 "Non-Hispanic Black", "Non-Hispanic Asian",  
 "Other Race", "Multi-Racial")  
 )  
 ) %>%  
 left\_join(dat\_clean %>% select(seqn, bmxbmi\_clean), by = "seqn") %>%  
 drop\_na(Race, bmxbmi\_clean)  
  
# 種族分布表  
race\_dist <- dat\_race %>%  
 count(Race) %>%  
 mutate(prop = n / sum(n)) %>%  
 rename(category = Race)  
  
  
kable(race\_dist, digits = 3, caption = "Distribution of Race Categories")

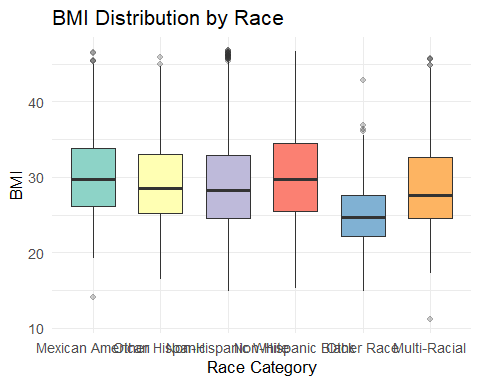
Distribution of Race Categories

| category | n | prop |
| --- | --- | --- |
| Mexican American | 390 | 0.067 |
| Other Hispanic | 593 | 0.102 |
| Non-Hispanic White | 3427 | 0.592 |
| Non-Hispanic Black | 689 | 0.119 |
| Other Race | 330 | 0.057 |
| Multi-Racial | 364 | 0.063 |

ggplot(dat\_edu, aes(x = EDU, y = bmxbmi\_clean, fill = EDU)) +  
 geom\_boxplot(outlier.alpha = 0.25, width = 0.65) +  
 scale\_fill\_brewer(palette = "Blues") +  
 labs(title = "BMI Distribution by Education Level",  
 x = "Education Level", y = "BMI") +  
 theme\_minimal(base\_size = 13) +  
 theme(legend.position = "none")



ggplot(dat\_race, aes(x = Race, y = bmxbmi\_clean, fill = Race)) +  
 geom\_boxplot(outlier.alpha = 0.25, width = 0.65) +  
 scale\_fill\_brewer(palette = "Set3") +  
 labs(title = "BMI Distribution by Race",  
 x = "Race Category", y = "BMI") +  
 theme\_minimal(base\_size = 13) +  
 theme(legend.position = "none")



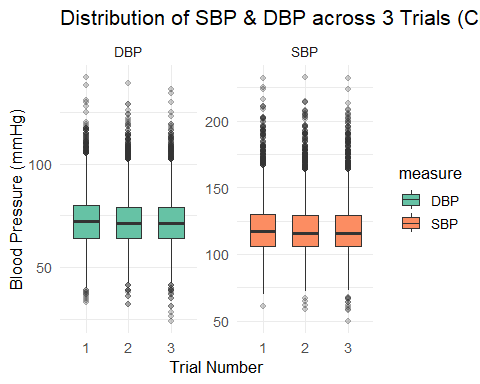
#Q3. Among all the subjects in 2021-2023 NHANES dataset, BPX is the data including three times of examination of blood pressure (SBP & DBP). The values were recorded in different columns (bpxosy1-3; bpxodi1-3) (Reminder: please use the “cleaned” BP data)

library(tidyverse)  
# 偵測 SBP 與 DBP 欄位名稱  
sbp\_cols <- names(bpx)[stringr::str\_detect(names(bpx), "^bpxo?sy[1-3]$")]  
dbp\_cols <- names(bpx)[stringr::str\_detect(names(bpx), "^bpxo?di[1-3]$")]  
  
# 轉換為長格式 (long format)  
bpx\_long\_clean <- bpx %>%  
 select(seqn, all\_of(c(sbp\_cols, dbp\_cols))) %>%  
 pivot\_longer(  
 cols = -seqn,  
 names\_to = c("measure", "trial"),  
 names\_pattern = "^bpxo?([sd]i|sy)([1-3])$",  
 values\_to = "value"  
 ) %>%  
 mutate(  
 measure = recode(measure, "sy" = "SBP", "di" = "DBP"),  
 trial = as.integer(trial)  
 )  
  
# 檢查轉換後的資料結構  
glimpse(bpx\_long\_clean)

## Rows: 46,806  
## Columns: 4  
## $ seqn <dbl> 130378, 130378, 130378, 130378, 130378, 130378, 130379, 130379…  
## $ measure <chr> "SBP", "SBP", "SBP", "DBP", "DBP", "DBP", "SBP", "SBP", "SBP",…  
## $ trial <int> 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3,…  
## $ value <dbl> 135, 131, 132, 98, 96, 94, 121, 117, 113, 84, 76, 76, 111, 112…

ggplot(bpx\_long\_clean, aes(x = factor(trial), y = value, fill = measure)) +  
 geom\_boxplot(outlier.alpha = 0.25, width = 0.6) +  
 facet\_wrap(~ measure, scales = "free\_y") +  
 scale\_fill\_brewer(palette = "Set2") +  
 labs(  
 title = "Distribution of SBP & DBP across 3 Trials (Cleaned Data)",  
 x = "Trial Number", y = "Blood Pressure (mmHg)"  
 ) +  
 theme\_minimal(base\_size = 13)

## Warning: Removed 1802 rows containing non-finite outside the scale range  
## (`stat\_boxplot()`).



# 計算每位受試者在三次測量中的最大差值 (SBP 和 DBP 分開)  
bpx\_diff <- bpx\_long\_clean %>%  
 group\_by(seqn, measure) %>%  
 summarise(  
 diff\_range = max(value, na.rm = TRUE) - min(value, na.rm = TRUE),  
 .groups = "drop"  
 )

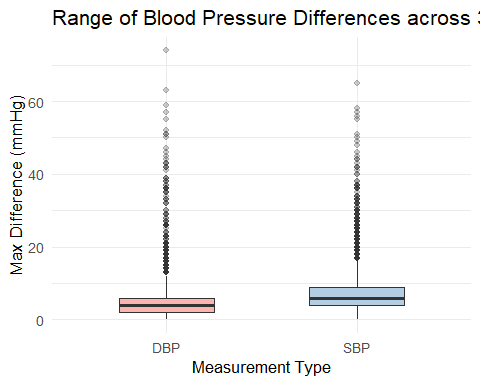
## Warning: There were 1132 warnings in `summarise()`.  
## The first warning was:  
## ℹ In argument: `diff\_range = max(value, na.rm = TRUE) - min(value, na.rm =  
## TRUE)`.  
## ℹ In group 37: `seqn = 130401` `measure = "DBP"`.  
## Caused by warning in `max()`:  
## ! no non-missing arguments to max; returning -Inf  
## ℹ Run `dplyr::last\_dplyr\_warnings()` to see the 1131 remaining warnings.

# 取出最大差異前 5 位作為示例 (方便檢查)  
head(bpx\_diff[order(-bpx\_diff$diff\_range), ], 5)

## # A tibble: 5 × 3  
## seqn measure diff\_range  
## <dbl> <chr> <dbl>  
## 1 141202 DBP 74  
## 2 136053 SBP 65  
## 3 135594 DBP 63  
## 4 131306 DBP 59  
## 5 141202 SBP 58

# 視覺化：不同測量型別的最大差值分布  
ggplot(bpx\_diff, aes(x = measure, y = diff\_range, fill = measure)) +  
 geom\_boxplot(outlier.alpha = 0.25, width = 0.5) +  
 scale\_fill\_brewer(palette = "Pastel1") +  
 labs(  
 title = "Range of Blood Pressure Differences across 3 Trials",  
 x = "Measurement Type", y = "Max Difference (mmHg)"  
 ) +  
 theme\_minimal(base\_size = 13)+  
 theme(legend.position = "none")

## Warning: Removed 566 rows containing non-finite outside the scale range  
## (`stat\_boxplot()`).



cat("  
### Observation & Interpretation  
  
- Both SBP and DBP show relatively small variations across the three trials, usually within ±10 mmHg.  
- The distributions of the 1st, 2nd, and 3rd readings are quite close, and there is no clear systematic shift.  
- This pattern indicates that the three measurements were likely taken \*\*on the same day\*\*, probably within a short interval, to ensure measurement reliability.  
- Larger outliers (e.g., >20 mmHg difference) may reflect temporary physiological fluctuations or measurement error rather than time gaps.  
")

##   
## ### Observation & Interpretation  
##   
## - Both SBP and DBP show relatively small variations across the three trials, usually within ±10 mmHg.  
## - The distributions of the 1st, 2nd, and 3rd readings are quite close, and there is no clear systematic shift.  
## - This pattern indicates that the three measurements were likely taken \*\*on the same day\*\*, probably within a short interval, to ensure measurement reliability.  
## - Larger outliers (e.g., >20 mmHg difference) may reflect temporary physiological fluctuations or measurement error rather than time gaps.

Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.