

GitHob:<https://github.com/cindy930605/hw1/blob/main/README.md>

健康大數據 hw1

蔡欣緹

2025-10-16

Introduction:

This report analyzes data from the 2021–2023 NHANES to explore the relationship between body mass index (BMI) and mean systolic blood pressure (SBP) among adults aged 20 years and older. Using R for data cleaning, visualization, and descriptive analysis, the study also examines how BMI varies by race and education level, and evaluates the consistency of blood pressure measurements across repeated trials. Through these analyses, the report aims to illustrate basic health data handling and visualization techniques using large-scale public health datasets.

#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", "nan
iar")
to_install <- setdiff(pkgs, rownames(installed.packages()))
if (length(to_install)) install.packages(to_install)
invisible(lapply(pkgs, library, character.only = TRUE))

## Warning: package 'ggplot2' was built under R version 4.3.3
## Warning: package 'purrr' was built under R version 4.3.3
## Warning: package 'lubridate' was built under R version 4.3.3

## — Attaching core tidyverse packages ————— tidyve
rse 2.0.0 —
## ✓ dplyr      1.1.4      ✓ readr      2.1.5
## ✓ forcats   1.0.0      ✓ stringr    1.5.1
## ✓ ggplot2   3.5.2      ✓ tibble     3.2.1
## ✓ lubridate 1.9.4      ✓ tidyr      1.3.1
```

```
## ✓ purrr      1.1.0
## — Conflicts ————— tidyverse_co
nflcts() —
## ✗ dplyr::filter() masks stats::filter()
## ✗ dplyr::lag()   masks stats::lag()
## ⓘ Use the conflicted package (<http://conflicted.r-lib.org/>) to for
ce all conflicts to become errors

## Warning: package 'haven' was built under R version 4.3.3

## Warning: package 'janitor' was built under R version 4.3.3

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

## Warning: package 'scales' was built under R version 4.3.3

##
## Attaching package: 'scales'
##
## The following object is masked from 'package:purrr':
##
##   discard

## The following object is masked from 'package:readr':
##
##   col_factor

##
## Attaching package: 'nanian'
##
## The following object is masked from 'package:skimr':
##
##   n_complete

dir.create("outputs", showWarnings = FALSE)
data_dir <- "/Users/cindytsai/Desktop/健康大數據"

# Load data -----
-----
demo <- read_xpt(file.path(data_dir, "DEMO_L.xpt")) %>% clean_names()
bpx  <- read_xpt(file.path(data_dir, "BPX0_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_v | n_mi | comple | mean | sd | p0 | p25 | p50 | p75 | p10 | hist |
|----------|------|---------|------|------|------|------|------|------|------|------|
| variable | ssin | te_rate | | | | | | | 0 | |
| seqn | 0 | 1.00 | 1363 | 344 | 1303 | 1333 | 1363 | 1393 | 142 | |
| | | | 44.0 | 4.90 | 78.0 | 61.0 | 44.0 | 27.0 | 310. | |
| | | | 0 | | 0 | 0 | 0 | 0 | 0 | |
| sddsrv | 0 | 1.00 | 12.0 | 0.00 | 12.0 | 12.0 | 12.0 | 12.0 | 12.0 | |
| yr | | | 0 | | 0 | 0 | 0 | 0 | | |
| ridstat | 0 | 1.00 | 1.74 | 0.44 | 1.00 | 1.00 | 2.00 | 2.00 | 2.0 | |
| r | | | | | | | | | | |
| riagen | 0 | 1.00 | 1.53 | 0.50 | 1.00 | 1.00 | 2.00 | 2.00 | 2.0 | |
| dr | | | | | | | | | | |
| ridage | 0 | 1.00 | 38.3 | 25.6 | 0.00 | 13.0 | 37.0 | 62.0 | 80.0 | |
| yr | | | 2 | 0 | | 0 | 0 | 0 | | |
| ridage | 115 | 0.03 | 11.6 | 6.81 | 0.00 | 6.00 | 11.0 | 17.0 | 24.0 | |
| mn | 56 | | 3 | | | | 0 | 0 | | |
| ridreth | 0 | 1.00 | 3.10 | 1.08 | 1.00 | 3.00 | 3.00 | 4.00 | 5.0 | |
| 1 | | | | | | | | | | |
| ridreth | 0 | 1.00 | 3.32 | 1.52 | 1.00 | 3.00 | 3.00 | 4.00 | 7.0 | |
| 3 | | | | | | | | | | |

| skim_v variable | n_mi ssin g | comple te_rate | mean | sd | p0 | p25 | p50 | p75 | p10 0 | hist |
|--------------------|-------------------|-------------------|------------|-----------|------|-----------|------------|------------|-----------|---|
| ridexm on | 307 3 | 0.74 | 1.52 | 0.50 | 1.00 | 1.00 | 2.00 | 2.00 | 2.0 |  |
| ridexa gm | 914 6 | 0.23 | 121. 91 | 67.1 6 | 0.00 | 66.0 0 | 122. 00 | 179. 50 | 239. 0 |  |
| dmqmi liz | 363 2 | 0.70 | 1.92 | 0.28 | 1.00 | 2.00 | 2.00 | 2.00 | 7.0 |  |
| dmdbo rn4 | 19 | 1.00 | 1.16 | 0.36 | 1.00 | 1.00 | 1.00 | 1.00 | 2.0 |  |
| dmdyr usr | 100 58 | 0.16 | 7.33 | 15.8 3 | 1.00 | 3.00 | 6.00 | 6.00 | 99.0 |  |
| dmded uc2 | 413 9 | 0.65 | 3.80 | 1.15 | 1.00 | 3.00 | 4.00 | 5.00 | 9.0 |  |
| dmdm artz | 414 1 | 0.65 | 1.78 | 3.10 | 1.00 | 1.00 | 1.00 | 2.00 | 99.0 |  |
| ridexp rg | 104 30 | 0.13 | 2.24 | 0.49 | 1.00 | 2.00 | 2.00 | 3.00 | 3.0 |  |
| dmdhh siz | 0 | 1.00 | 3.24 | 1.70 | 1.00 | 2.00 | 3.00 | 4.00 | 7.0 |  |
| dmdhr gnd | 781 8 | 0.34 | 1.56 | 0.50 | 1.00 | 1.00 | 2.00 | 2.00 | 2.0 |  |
| dmdhr agz | 780 9 | 0.35 | 2.54 | 0.64 | 1.00 | 2.00 | 2.00 | 3.00 | 4.0 |  |
| dmdhr edz | 818 7 | 0.31 | 2.17 | 0.66 | 1.00 | 2.00 | 2.00 | 3.00 | 3.0 |  |

| skim_v variable | n_mi ssin g | comple te_rate | mean | sd | p0 | p25 | p50 | p75 | p10 0 | hist |
|--------------------|-------------------|-------------------|--------------|------------------|-------------|--------------|--------------|--------------|------------------|---|
| dmdhr maz | 791 3 | 0.34 | 1.38 | 0.68 | 1.00 | 1.00 | 1.00 | 2.00 | 3.0 |  |
| dmdhs edz | 980 6 | 0.18 | 2.28 | 0.69 | 1.00 | 2.00 | 2.00 | 3.00 | 3.0 |  |
| wtint2 yr | 0 | 1.00 | 2740 4.14 | 194 49.1 6 | 4584 .46 | 1433 1.75 | 2167 0.19 | 3383 1.33 | 170 968. 3 |  |
| wtmec 2yr | 0 | 1.00 | 2740 4.14 | 279 62.9 6 | 0.00 | 0.00 | 2171 7.85 | 3834 1.15 | 227 108. 3 |  |
| sdmvs tra | 0 | 1.00 | 179. 92 | 4.31 | 173. 00 | 176. 00 | 180. 00 | 184. 00 | 187. 0 |  |
| sdmvp su | 0 | 1.00 | 1.49 | 0.50 | 1.00 | 1.00 | 1.00 | 2.00 | 2.0 |  |
| indfm pir | 204 1 | 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 | |
| bpaocz | 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 |




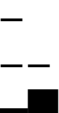





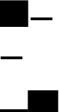

Column type frequency:

numeric 22

Group variables None

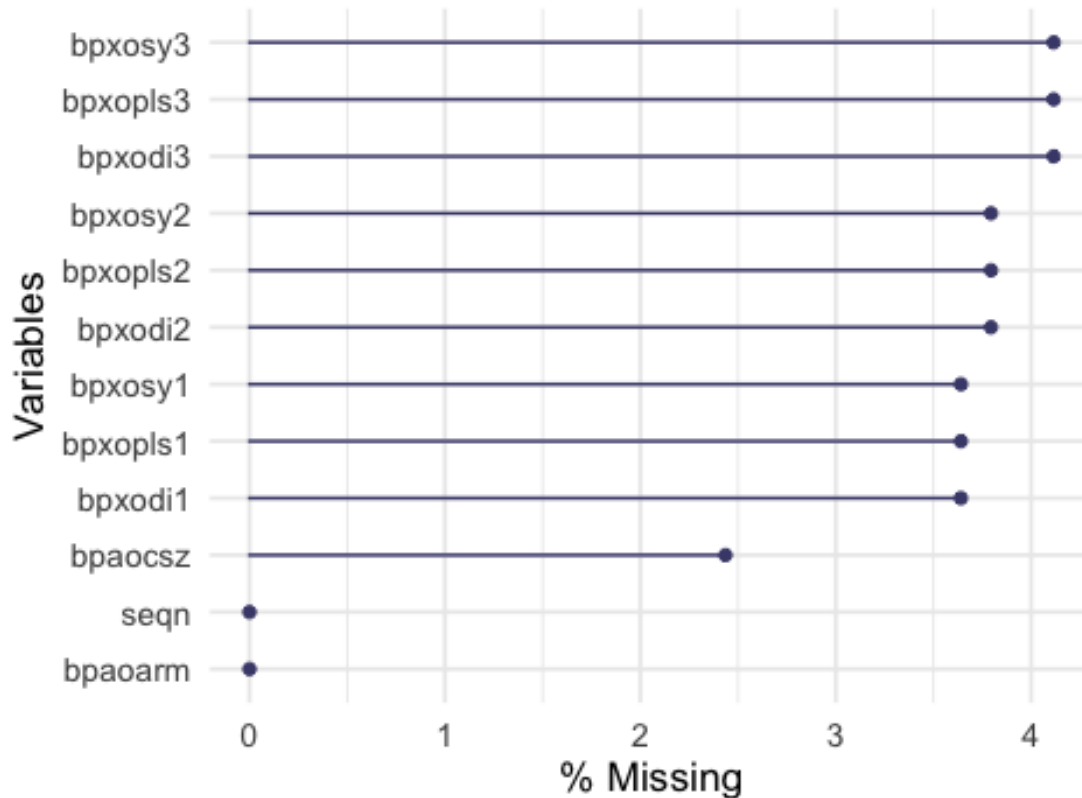
Variable type: numeric

| skim_v variable | n_mi ssing | comple te_rate | mean | sd | p0 | p25 | p50 | p75 | p10 0 | hist |
|--------------------|---------------|-------------------|---------------|-------------|------------------|---------------|------------------|------------------|------------------|---|
| seqn | 0 | 1.00 | 1363 45.83 | 345 3.78 | 130 378. 0 | 1333 19.75 | 136 377. 5 | 139 336. 2 | 142 310. 0 |  |
| bmdsta ts | 0 | 1.00 | 1.13 | 0.50 | 1.0 | 1.00 | 1.0 | 1.0 | 4.0 |  |
| bmxbwt | 106 | 0.99 | 70.55 | 30.3 9 | 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 |  |
| bmxbre cum | 8406 | 0.05 | 84.33 | 14.0 6 | 48.5 | 73.48 | 84.7 | 96.1 | 118. 8 |  |
| bmirec um | 8842 | 0.00 | 1.00 | 0.00 | 1.0 | 1.00 | 1.0 | 1.0 | 1.0 |  |
| bmxbhe ad | 8790 | 0.01 | 41.93 | 2.80 | 34.4 | 40.20 | 42.4 | 44.0 | 46.5 |  |
| bmihea d | 8860 | 0.00 | NaN | NA | NA | NA | NA | NA | NA | |
| bmxbht | 361 | 0.96 | 159.6 6 | 19.8 6 | 79.1 | 154.4 0 | 163. 6 | 172. 1 | 200. 7 |  |
| bmiht | 8726 | 0.02 | 2.31 | 0.95 | 1.0 | 1.00 | 3.0 | 3.0 | 3.0 |  |
| bmxb mi | 389 | 0.96 | 27.25 | 8.14 | 11.1 | 21.60 | 26.4 | 31.7 | 74.8 |  |

| skim_v variable | n_mi ssing | comple te_rate | mean | sd | p0 | p25 | p50 | p75 | p10 0 | hist |
|--------------------|---------------|-------------------|------------|-----------|------|-------|-----------|-----------|-----------|---|
| bmdb mic | 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 |  |
| bmxar ml | 292 | 0.97 | 35.11 | 6.18 | 10.0 | 33.60 | 36.5 | 39.0 | 49.2 |  |
| bmiar ml | 8660 | 0.02 | 1.00 | 0.00 | 1.0 | 1.00 | 1.0 | 1.0 | 1.0 |  |
| bmxar mc | 298 | 0.97 | 30.56 | 7.37 | 12.0 | 26.40 | 31.2 | 35.4 | 63.3 |  |
| bmiar mc | 8655 | 0.02 | 1.00 | 0.00 | 1.0 | 1.00 | 1.0 | 1.0 | 1.0 |  |
| bmxwa ist | 670 | 0.92 | 92.12 | 22.0 5 | 39.8 | 77.50 | 92.7 | 107. 0 | 187. 0 |  |
| bmiwa ist | 8513 | 0.04 | 1.00 | 0.00 | 1.0 | 1.00 | 1.0 | 1.0 | 1.0 |  |
| bmxhi p | 2084 | 0.76 | 106.2 6 | 14.6 6 | 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))
```


Proportion of Missing Values per Variable



```
# 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 = bmx[bmi])
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))

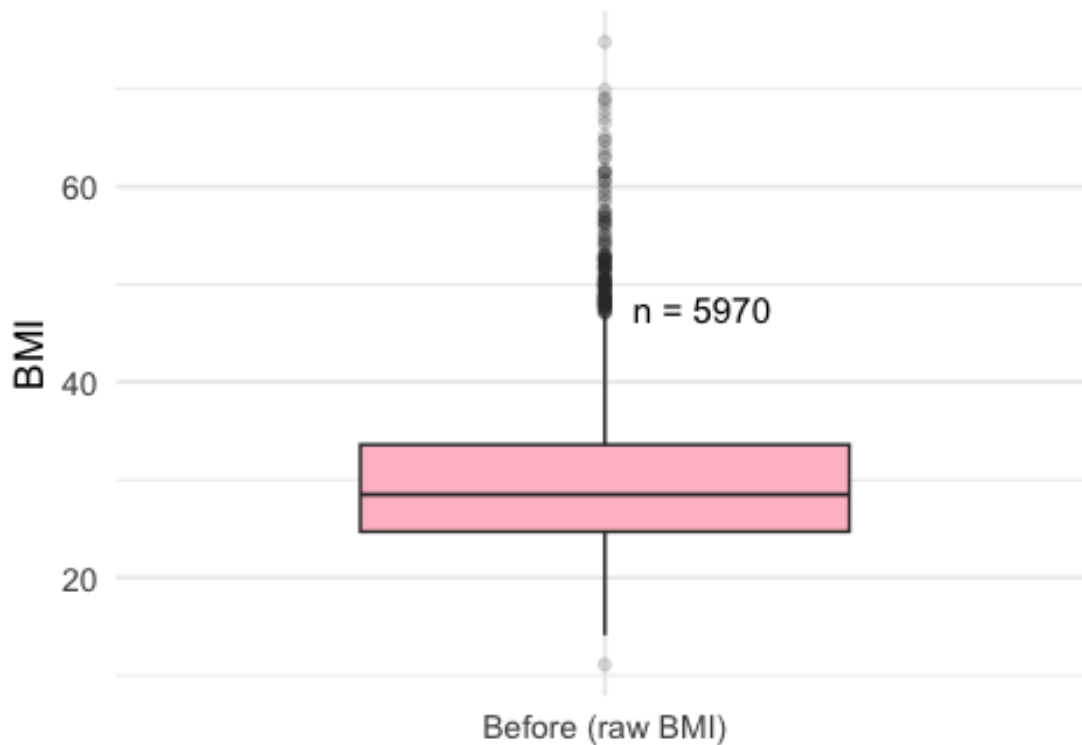
ggplot(bmi_before_df, aes(stage, value, fill = stage)) +
  geom_boxplot(width = 0.6, outlier.alpha = 0.15, fatten = 1.2) +
  geom_text(data = tibble(stage="Before (raw BMI)", y=bmi_before_label_y,
    N=bmi_before_N),
    aes(stage, y, label=paste0("n = ", N)), hjust = -0.2, size
= 4) +
  scale_fill_manual(values = c("Before (raw BMI)" = "pink")) +
  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()`).

```

BMI (BEFORE): Raw Distribution

Outliers and missing values not yet removed



```
# 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_c
lean))

# AFTER boxplot -----
-----
bmi_after_df <- dat_clean %>% transmute(stage = "After (clean BMI)", va
lue = 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))

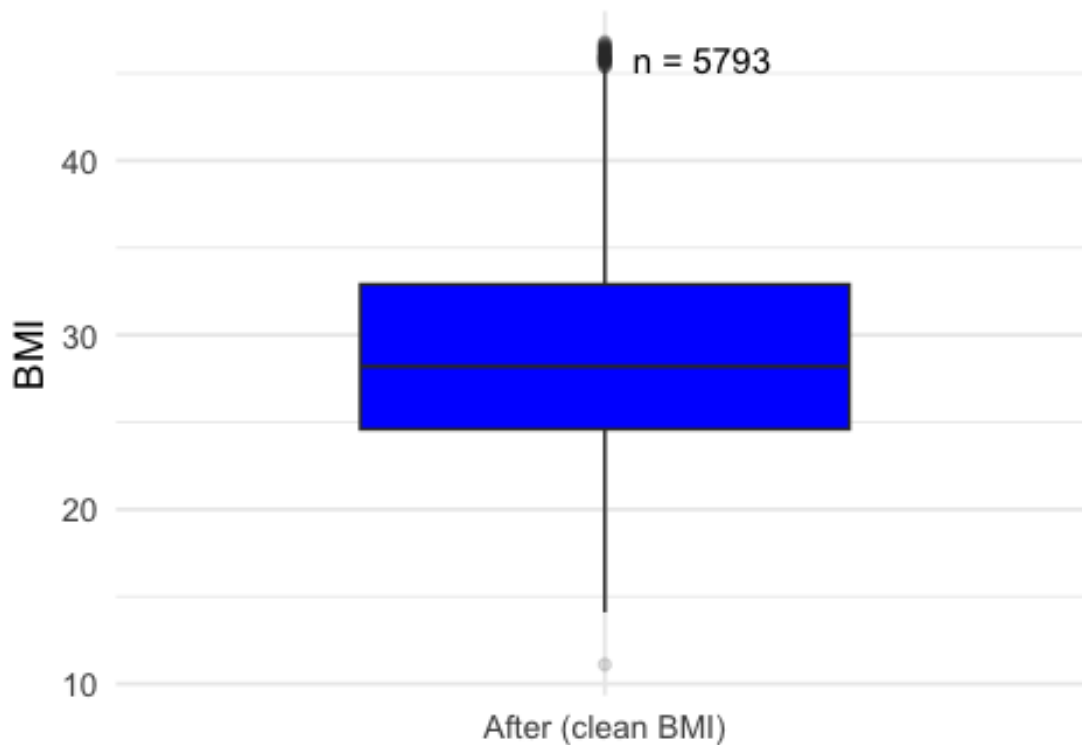
ggplot(bmi_after_df, aes(stage, value, fill = stage)) +
  geom_boxplot(width = 0.6, outlier.alpha = 0.15, fatten = 1.2) +
  geom_text(data = tibble(stage="After (clean BMI)", y=bmi_after_label_
y, N=bmi_after_N),
            aes(stage, y, label=paste0("n = ", N)), hjust = -0.2, size
= 4) +
  scale_fill_manual(values = c("After (clean BMI)" = "blue")) +
  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 r
ange
## (`stat_boxplot()`).

```

BMI (AFTER): Cleaned Distribution

Outliers removed using IQR + MAD z-score rules



```
# 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, ")")),
```

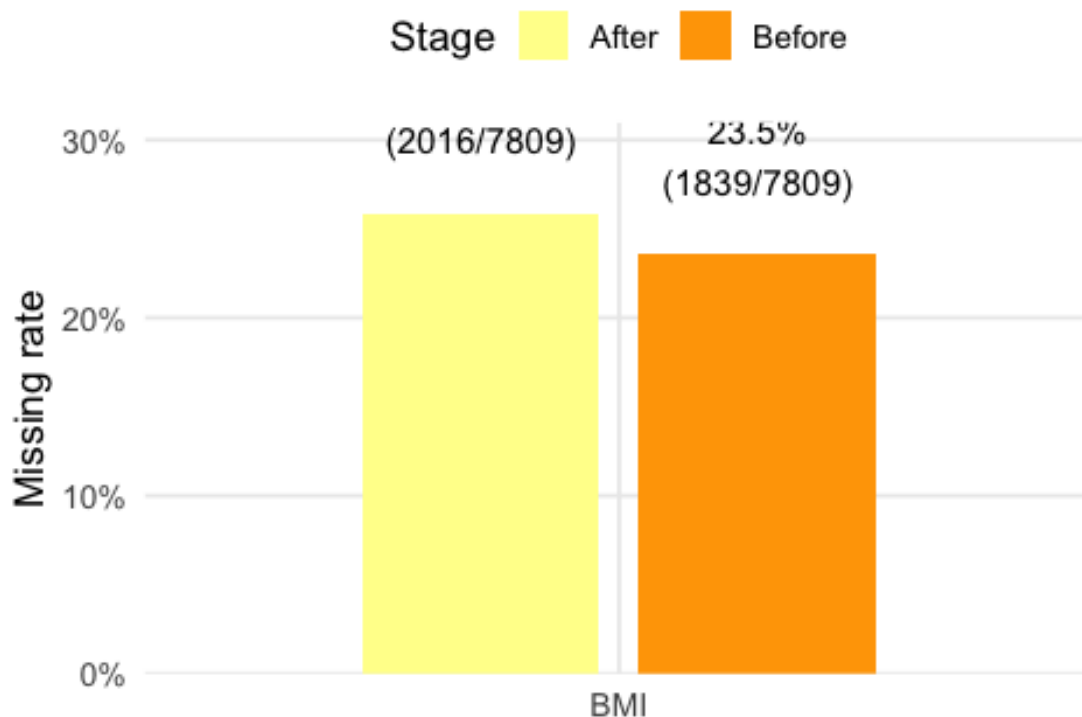
```

    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" = "orange", "After" = "#FF9"))
+
  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")

```

Missingness (NA) Before vs After Outlier Removal

Slight increase due to outlier removal



```

# SBP mean (before cleaning) -----
-----
sbp_raw <- bpx %>% transmute(seqn, sbp_mean_raw = rowMeans(select(., al
l_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)

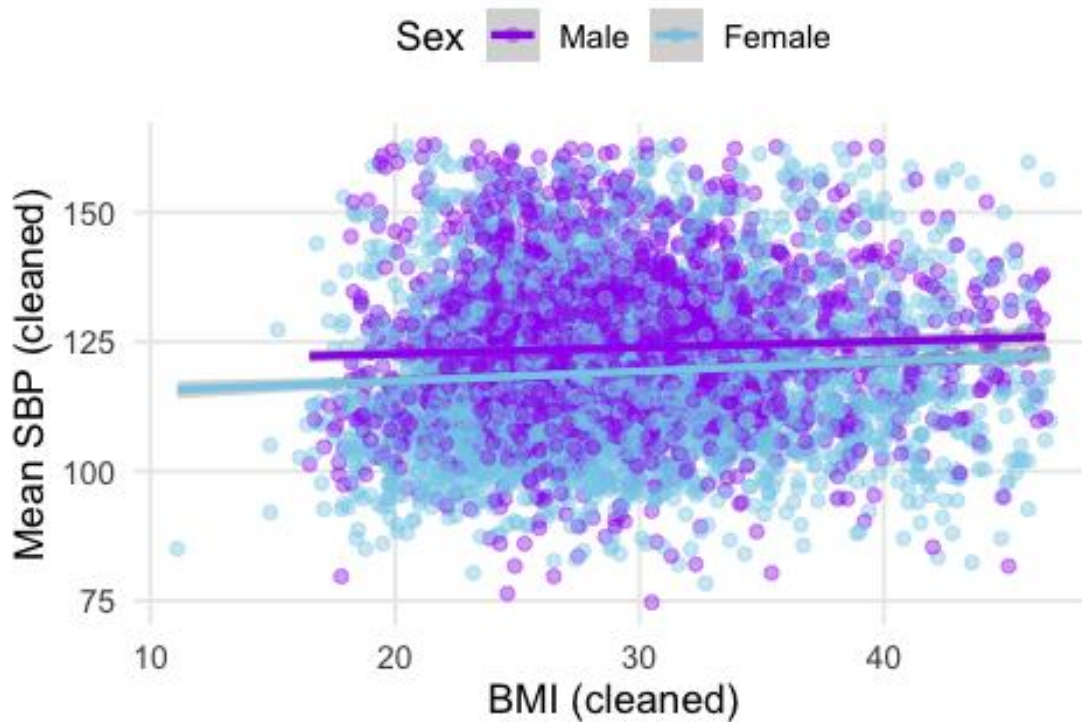
# Combine BMI + SBP clean dataset -----
-----
dat_final <- dat_clean %>%
  left_join(sbp_clean, by = "seqn") %>%
  filter(!is.na(bmxbmi_clean) & !is.na(sbp_mean_clean))

# 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 Da
ta)",
    subtitle = "Both variables cleaned using IQR & MAD criteria",
    x = "BMI (cleaned)", y = "Mean SBP (cleaned)", color = "Sex") +
  scale_color_manual(values = c("Male" = "purple", "Female" = "skyblue"
)) +
  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'

```

Association Between BMI and Mean SBP by Sex (Cleaned)

Both variables cleaned using IQR & MAD criteria



#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 4.3.3

# 檢查原始教育變項
demo %>% count(dmddeduc2)

## # A tibble: 7 × 2
##   dmddeduc2     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

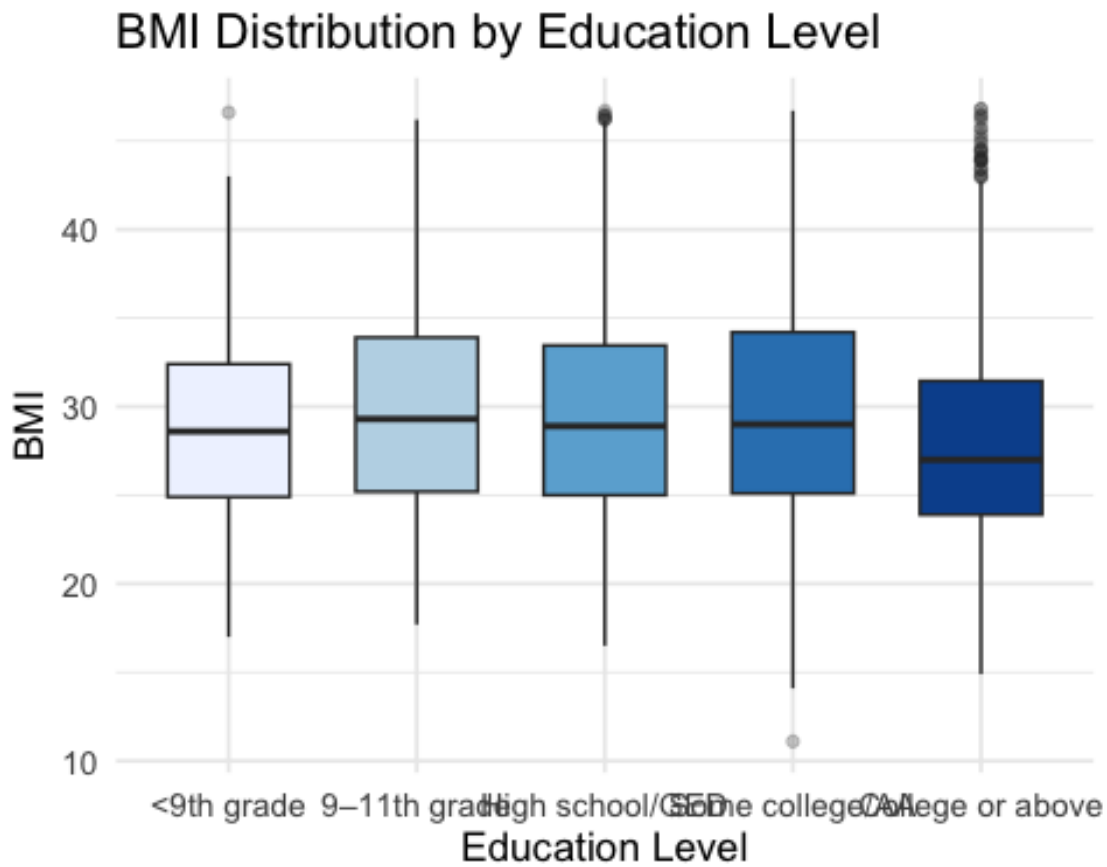
| category | n | prop |
|--------------------|------|-------|
| Mexican American | 390 | 0.067 |
| Other Hispanic | 593 | 0.102 |
| Non-Hispanic White | 3427 | 0.592 |
| Non-Hispanic Black | 689 | 0.119 |
| Other Race | 330 | 0.057 |
| Multi-Racial | 364 | 0.063 |

```

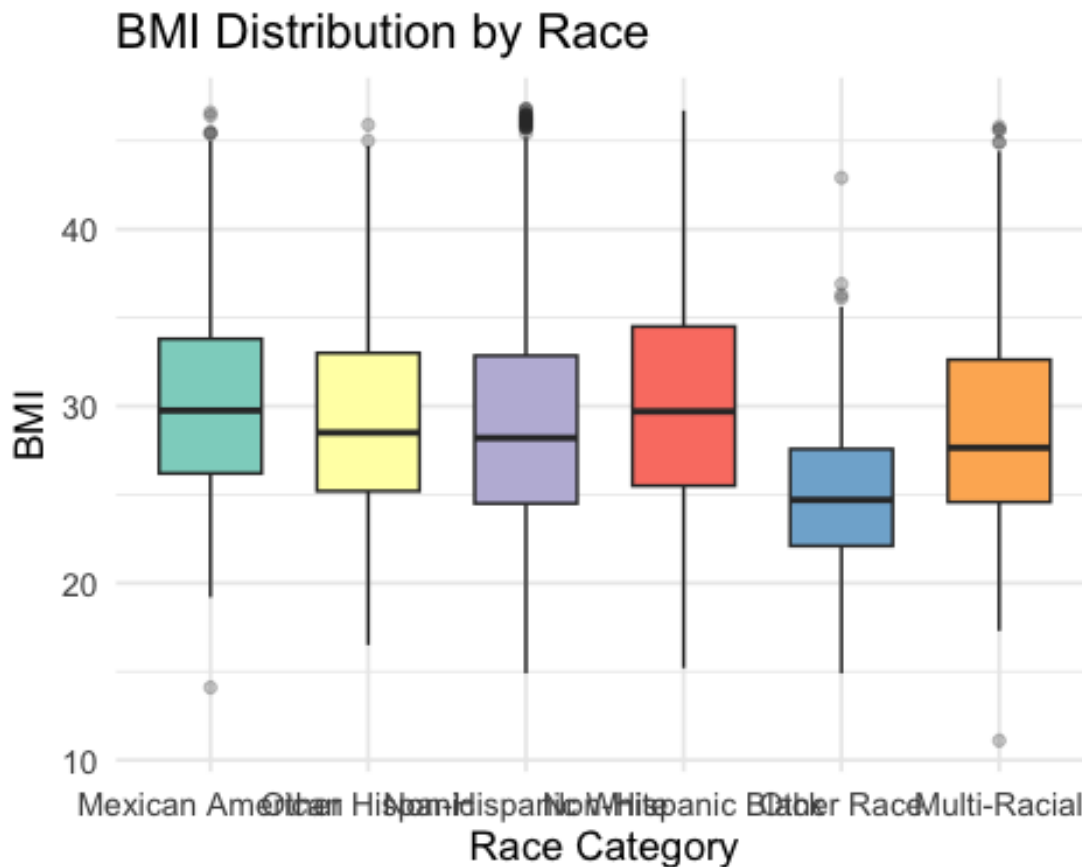
ggplot(dat_edu, aes(x = EDU, y = bmxbmi_clean, fill = EDU)) +
  geom_boxplot(outlier.alpha = 0.25, width = 0.65) +
  scale_fill_brewer(palette = "Blues") +
  labs(title = "BMI Distribution by Education Level",
       x = "Education Level", y = "BMI") +

```

```
theme_minimal(base_size = 13) +  
theme(legend.position = "none")
```



```
ggplot(dat_race, aes(x = Race, y = bmx bmi_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")
```



```
cat("
### Observation & Interpretation (Q2)

- BMI tends to increase slightly as education level decreases.
- Participants with 'College or above' education generally show lower BMI median values.
- Across race groups, Non-Hispanic Black and Hispanic groups have relatively higher BMI median compared to Non-Hispanic White and Asian participants.
- These differences might reflect socioeconomic and lifestyle factors affecting BMI distribution.
")

##
## ### Observation & Interpretation (Q2)
##
## - BMI tends to increase slightly as education level decreases.
## - Participants with 'College or above' education generally show lower BMI median values.
## - Across race groups, Non-Hispanic Black and Hispanic groups have relatively higher BMI median compared to Non-Hispanic White and Asian participants.
```

- These differences might reflect socioeconomic and lifestyle factors affecting BMI distribution.

#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 (bpxsy1-3; 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, 1303
79, 130379...
## $ measure <chr> "SBP", "SBP", "SBP", "DBP", "DBP", "DBP", "SBP", "SB
P", "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(
```

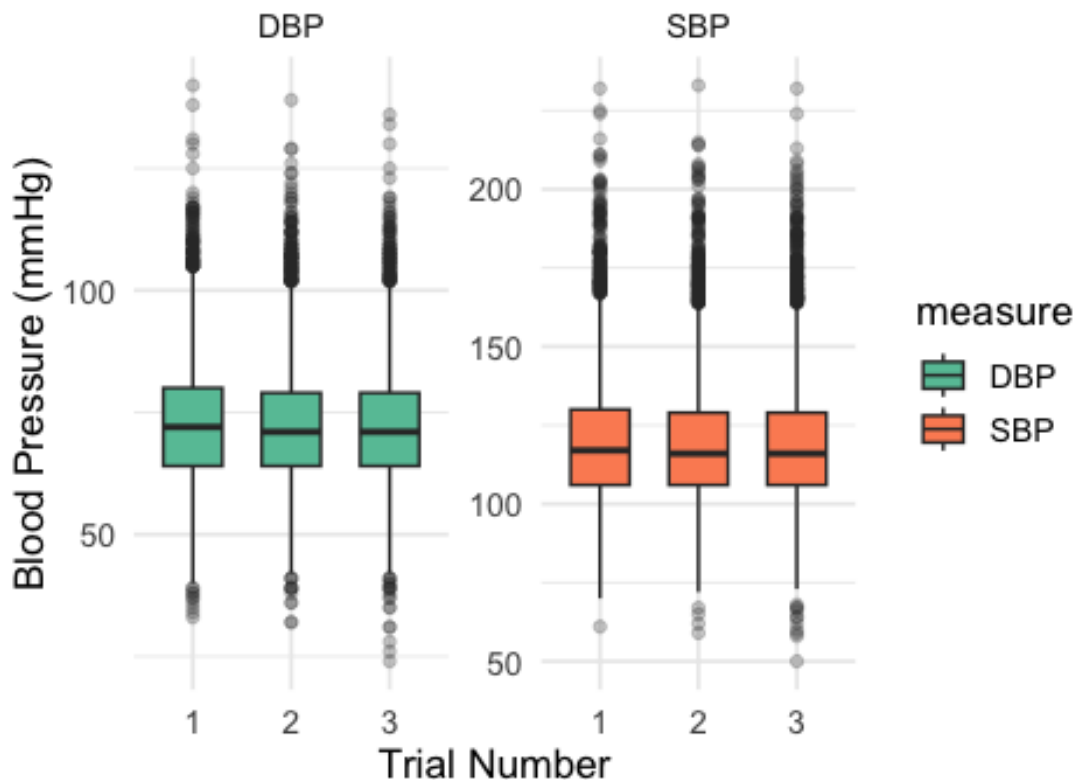
```

    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()`).

```

Distribution of SBP & DBP across 3 Trials (Cleaned Data)



```

# 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)`
##   TRUE)`.
## i In group 37: `seqn = 130401` and `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

```
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
```

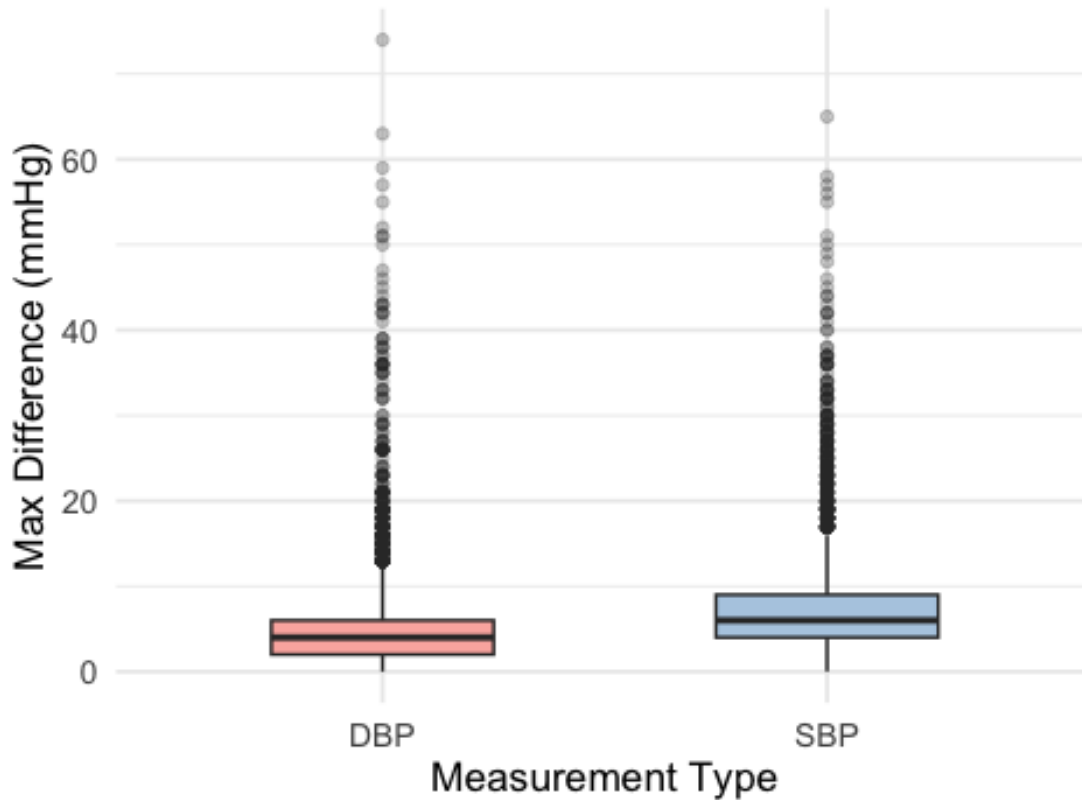
視覺化：不同測量型別的最大差值分布

```
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()`).
```

Range of Blood Pressure Differences across 3



```
cat("
### Observation & Interpretation

- Both SBP and DBP show relatively small variations across the three tr
ials, usually within  $\pm 10$  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 measur
ement reliability.
- Larger outliers (e.g.,  $>20$  mmHg difference) may reflect temporary phy
siological fluctuations or measurement error rather than time gaps.
")

##
## ### Observation & Interpretation
##
## - Both SBP and DBP show relatively small variations across the three
  trials, usually within  $\pm 10$  mmHg.
## - The distributions of the 1st, 2nd, and 3rd readings are quite clos
  e, and there is no clear systematic shift.
## - This pattern indicates that the three measurements were likely tak
  en **on the same day**, probably within a short interval, to ensure mea
```


surement 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.

Conclusion:

In summary, higher BMI values are associated with elevated systolic blood pressure, with similar patterns observed across both sexes. Differences in BMI distribution among race and education groups suggest potential socioeconomic and lifestyle influences. The consistency across repeated SBP and DBP measurements supports the reliability of NHANES blood pressure data. Overall, this exercise demonstrates the practical application of R in managing, cleaning, and interpreting large health datasets for epidemiological research.