NHANES 2021-2023 Data Analysis: Cleaning, Wrangling, and Visualization

B11801014 Vincent Lin

2025-10-17

Contents

Introduction

This report analyzes data from the **National Health and Nutrition Examination Survey (NHANES)** conducted between **August 2021 and August 2023**.

The purpose of this analysis is to **clean and explore variables** related to:

- **Body measures:** BMI
- **Blood pressure:** Systolic (SBP) and Diastolic (DBP)
- **Demographics:** Age, sex, education, and race/ethnicity

Key objectives include:

- Examining associations such as **BMI with SBP by sex**
- Exploring **distributions across education and race levels**
- Assessing **variability in blood pressure measurements** across multiple trials

All analyses are conducted using **R**, ensuring full **reproducibility** through documented code chunks.

Week 5 Class: BMI Cleaning & Visualization

Setup and Data Loading

```r
# Load required packages
pkgs <- c("tidyverse", "haven", "janitor", "stringr", "scales",
          "skimr", "naniar", "knitr", "ggpmisc")
to_install <- setdiff(pkgs, rownames(installed.packages()))
if (length(to_install)) install.packages(to_install)
invisible(lapply(pkgs, library, character.only = TRUE))

# Set working directory (adjust as needed)
setwd("/Users/vinny/Desktop/class55")
data_dir <- "data_raw"
```

```r
# Load raw 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()
```

Data Overview

```r
# Quick data overview
skimr::skim(demo)
```

Table 1: 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	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
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	

```r
skimr::skim(bpx)
```

Table 3: 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	

```r
skimr::skim(bmx)
```

Table 6: Data summary

Name	bmx
Number of rows	8860
Number of columns	22
Column type frequency:	

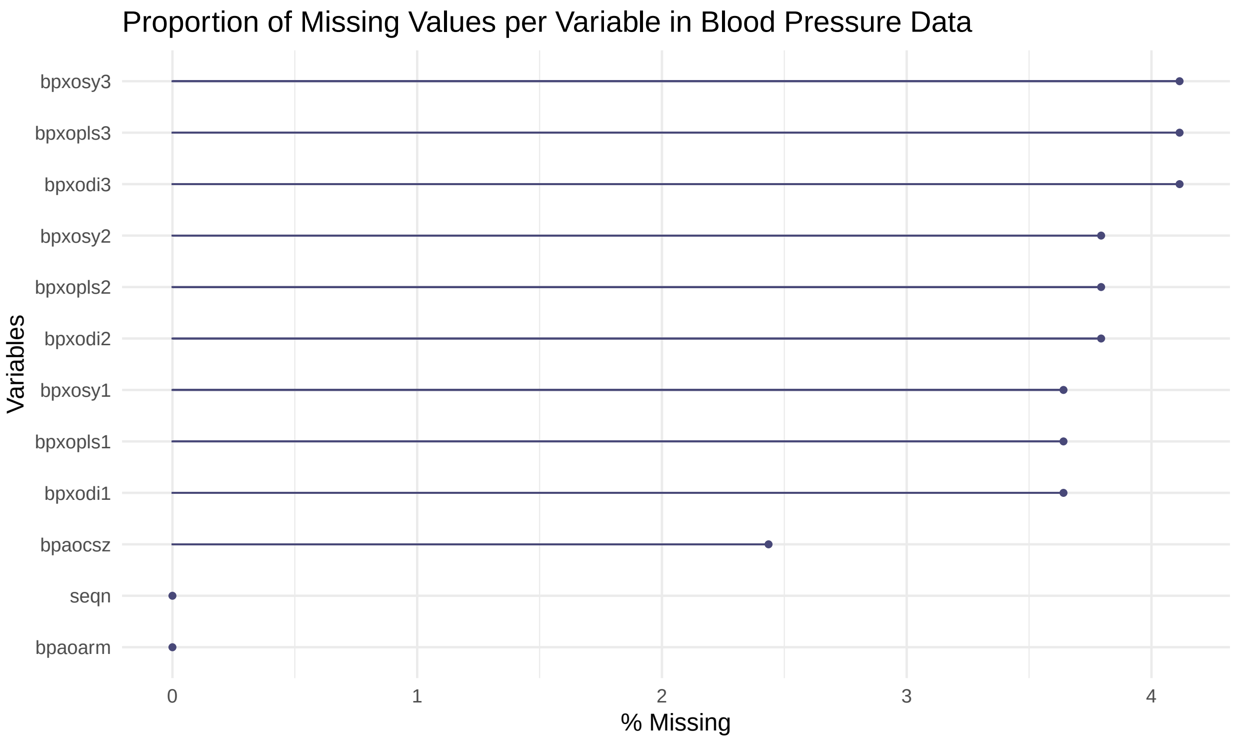
numeric	22
Group variables	None

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	

Missing Values Visualization

```r
gg_miss_var(bpx, show_pct = TRUE) +
  theme_minimal(base_size = 14) +
  labs(title = "Proportion of Missing Values per Variable in Blood Pressure Data")
```



Detect Blood Pressure Columns

```
# Detect systolic and diastolic BP columns
sbp_cols <- names(bpx)[stringr::str_detect(names(bpx), "^bpxo?sy[1-3]$")]
dbp_cols <- names(bpx)[stringr::str_detect(names(bpx), "^bpxo?di[1-3]$")]

cat("SBP columns:", sbp_cols, "\n")
```

```
## SBP columns: bpxosy1 bpxosy2 bpxosy3
```

```
cat("DBP columns:", dbp_cols, "\n")
```

```
## DBP columns: bpxodi1 bpxodi2 bpxodi3
```

Build BEFORE (Raw) Dataset

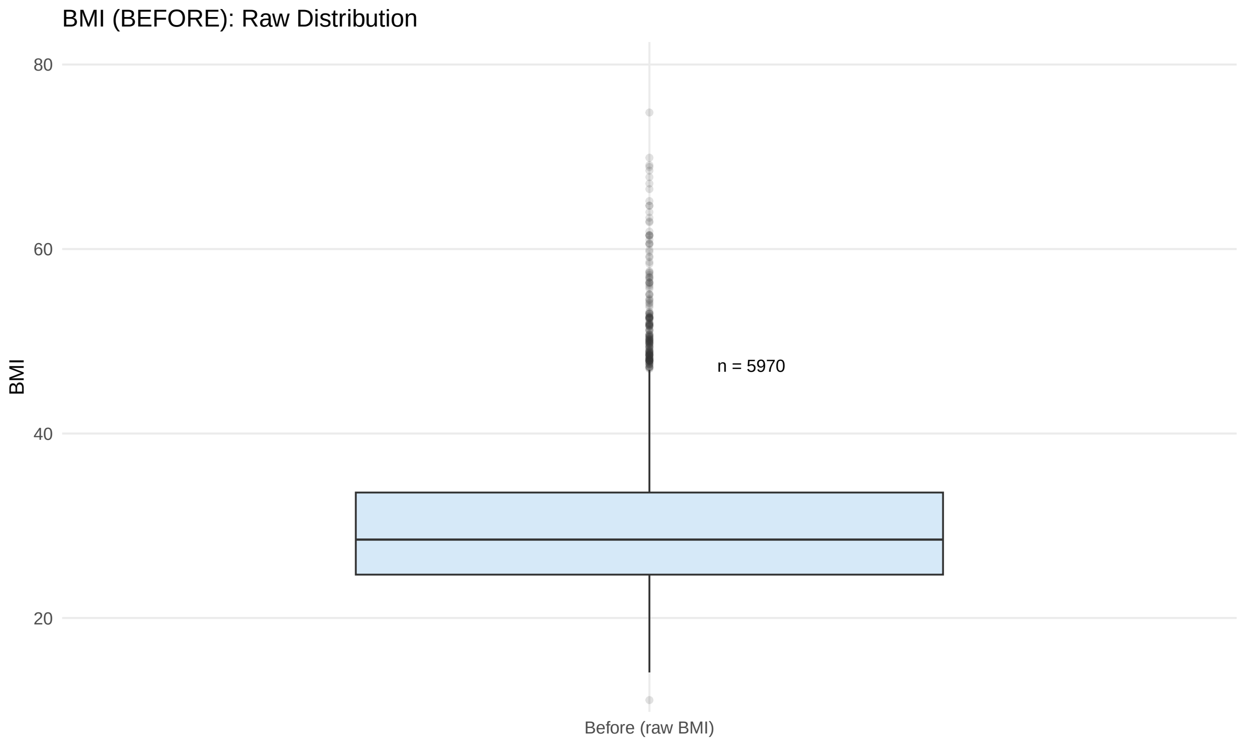
```
# Extract raw BMI
bmi_raw <- bmx %>%
  transmute(seqn, bmi_raw = bmxbmi)

# Check gender distribution
table(demo$riagendr)
```

```
##
##    1    2
## 5575 6358
```

```
# Clean demographics
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"))
  )

# Combine data
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))
```

BMI Distribution (BEFORE Cleaning)

```
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 = -1, size = 3.5, inherit.aes = FALSE
  ) +
  scale_fill_manual(values = c("Before (raw BMI)" = "#D6E9F8")) +
  labs(title = "BMI (BEFORE): Raw Distribution", x = NULL, y = "BMI") +
  scale_y_continuous(expand = expansion(mult = c(0.02, 0.12))) +
  theme_minimal(base_size = 12) +
  theme(legend.position = "none", panel.grid.minor = element_blank())
```

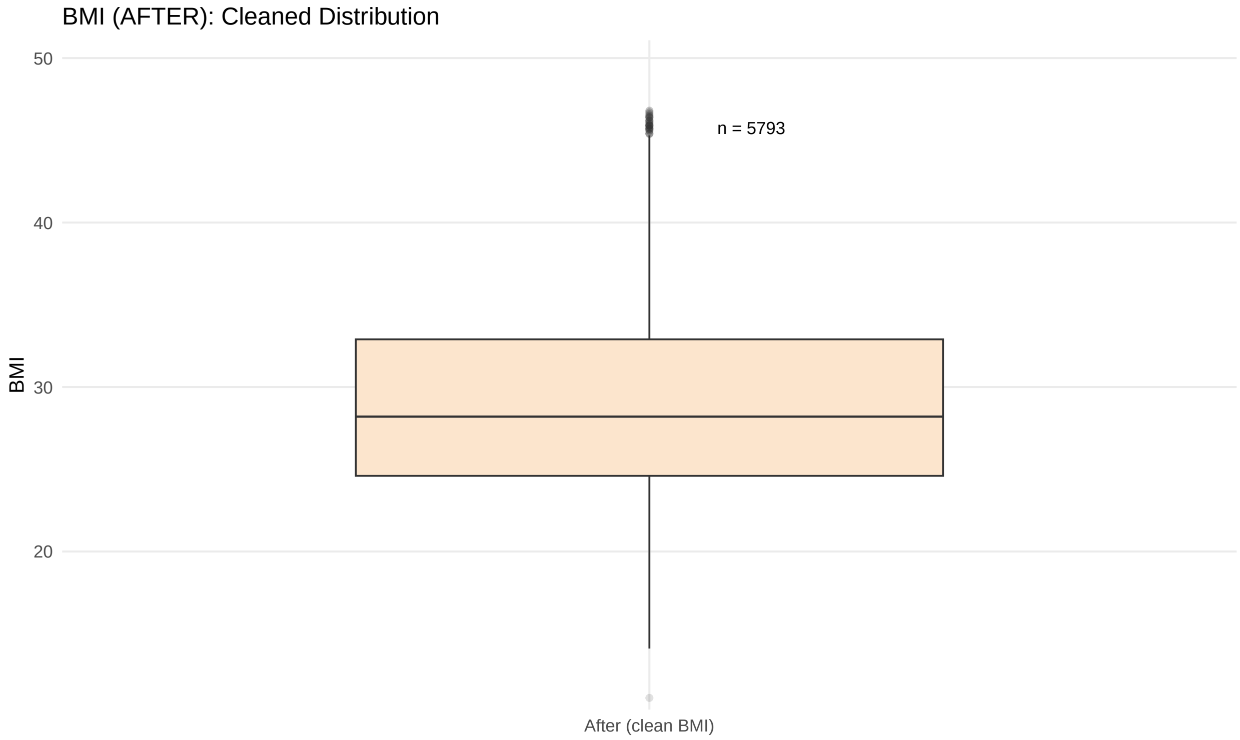


Outlier Cleaning

```
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)

# Build 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))
```

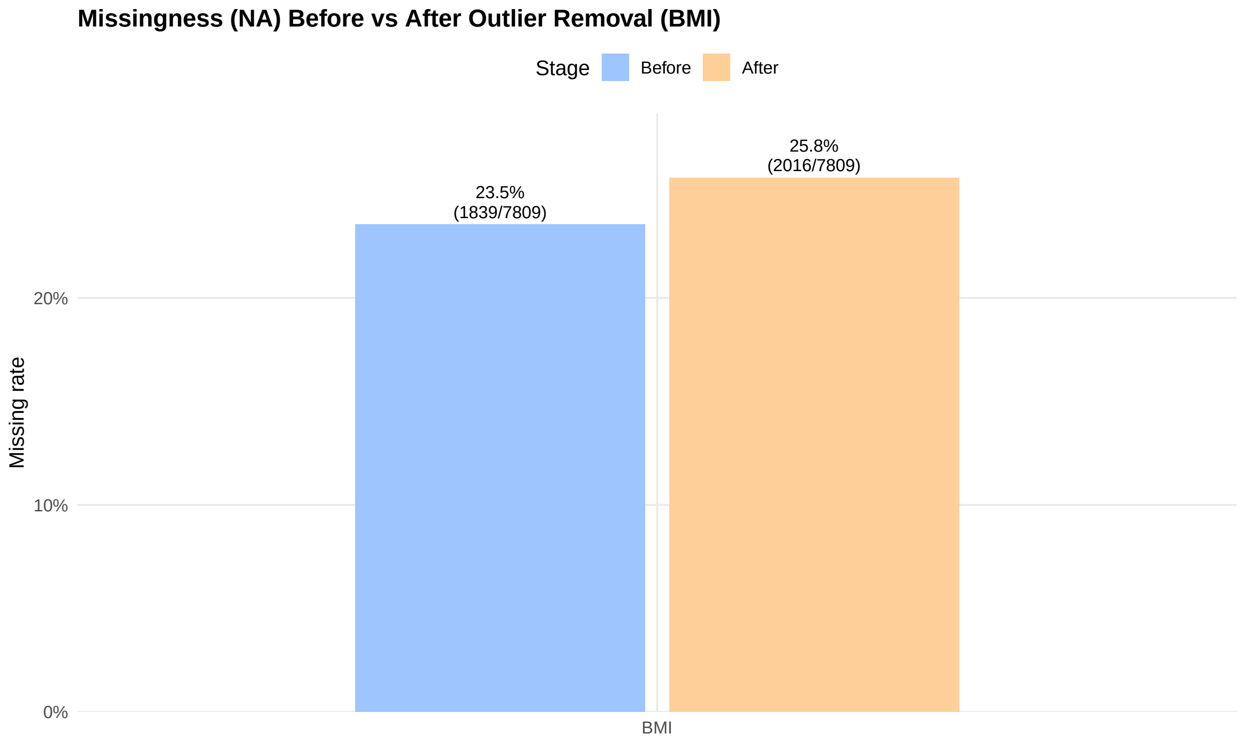
BMI Distribution (AFTER Cleaning)

```
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 = -1, size = 3.5, inherit.aes = FALSE
  ) +
  scale_fill_manual(values = c("After (clean BMI)" = "#FCE5CD")) +
  labs(title = "BMI (AFTER): Cleaned Distribution", x = NULL, y = "BMI") +
  scale_y_continuous(expand = expansion(mult = c(0.02, 0.12))) +
  theme_minimal(base_size = 12) +
  theme(legend.position = "none", panel.grid.minor = element_blank())
```



Missing Value Comparison

```r
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) %>%
  mutate(
    stage = factor(stage, levels = c("Before", "After")),
    variable = factor(variable, levels = "BMI")
  )

pos <- position_dodge(width = 0.65)

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 = -0.2, size = 3.5, lineheight = 0.95
  ) +
  scale_y_continuous(labels = scales::percent, expand = expansion(mult = c(0, 0.12))) +
  scale_fill_manual(values = c("Before" = "#9EC5FE", "After" = "#FFCF99")) +
  labs(
    title = "Missingness (NA) Before vs After Outlier Removal (BMI)",
    x = NULL, y = "Missing rate", fill = "Stage"
  ) +
  theme_minimal(base_size = 12) +
  theme(
    panel.grid.minor = element_blank(),
    plot.title = element_text(face = "bold"),
    legend.position = "top"
  )
```



Week 5 HW1: Blood Pressure Cleaning & Visualization

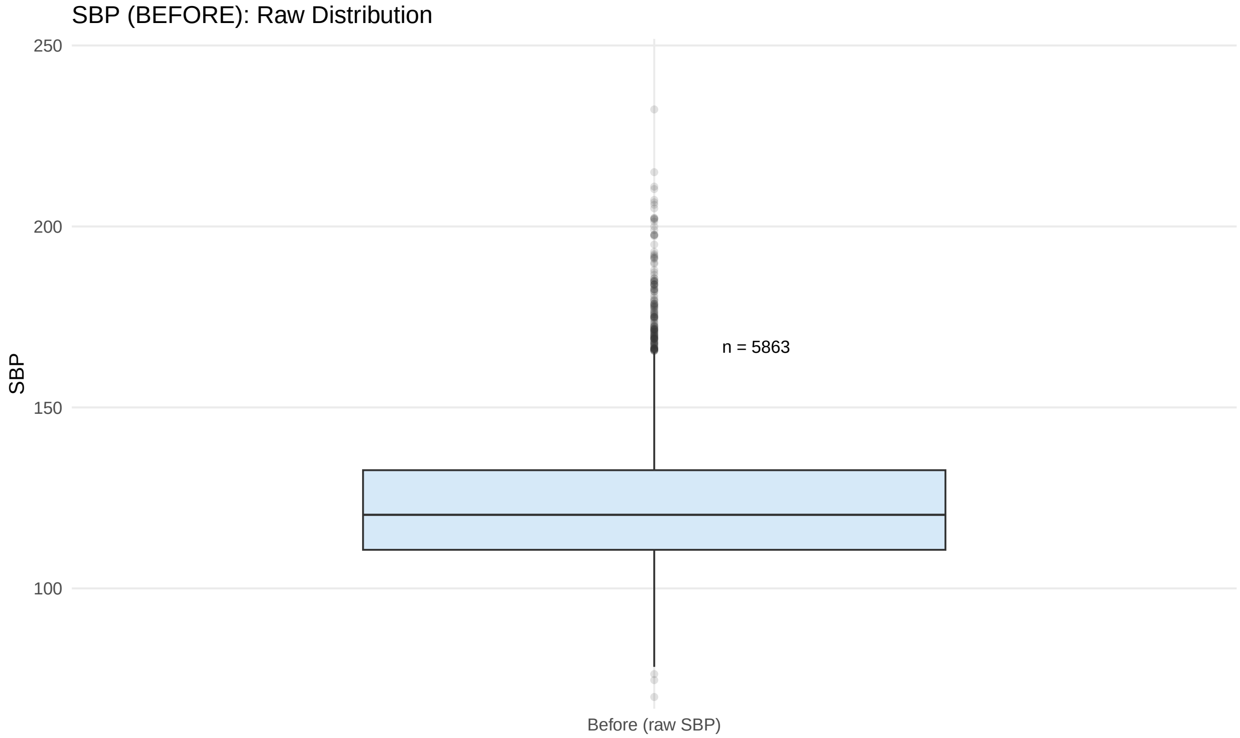
Build BEFORE (Raw) Blood Pressure Dataset

```r
# Extract raw SBP and DBP
sbp_raw <- bpx %>%
  transmute(seqn, SBP_raw = rowMeans(select(., all_of(sbp_cols)), na.rm = TRUE))

dbp_raw <- bpx %>%
  transmute(seqn, DBP_raw = rowMeans(select(., all_of(dbp_cols)), na.rm = TRUE))

# Combine with demographics
dat_raw <- demo_sex %>%
  left_join(sbp_raw, by = "seqn") %>%
  left_join(dbp_raw, by = "seqn") %>%
  filter(age >= 20) %>%
  mutate(
    SBP_raw = ifelse(is.nan(SBP_raw), NA_real_, SBP_raw),
    DBP_raw = ifelse(is.nan(DBP_raw), NA_real_, DBP_raw)
  )
```

Blood Pressure Distribution (BEFORE Cleaning)

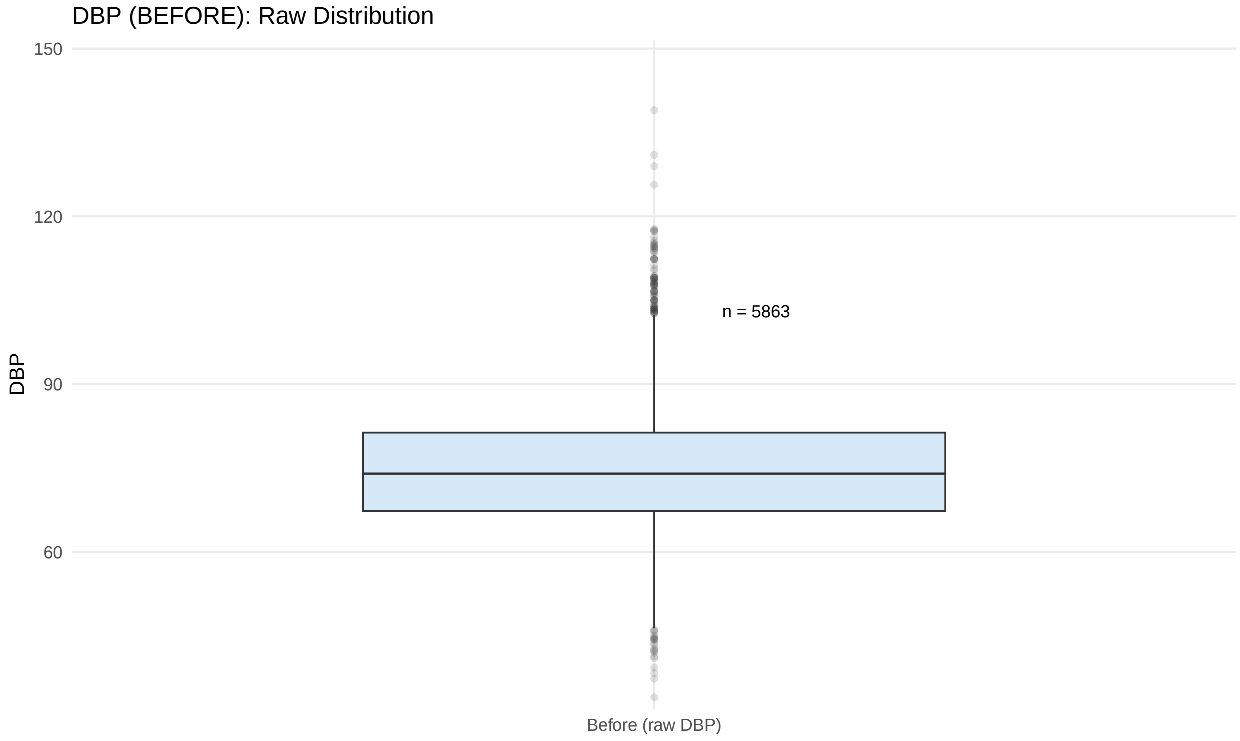
SBP Before

```r
sbp_before_df <- dat_raw %>% transmute(stage = "Before (raw SBP)", value = SBP_raw)
x <- sbp_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)
sbp_before_label_y <- upper_whisker + 0.05 * iqr
sbp_before_N <- sum(!is.na(x))

ggplot(sbp_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 SBP)", y = sbp_before_label_y, N = sbp_before_N),
    aes(stage, y, label = paste0("n = ", N)),
    hjust = -1, size = 3.5, inherit.aes = FALSE
  ) +
  scale_fill_manual(values = c("Before (raw SBP)" = "#D6E9F8")) +
  labs(title = "SBP (BEFORE): Raw Distribution", x = NULL, y = "SBP") +
  scale_y_continuous(expand = expansion(mult = c(0.02, 0.12))) +
  theme_minimal(base_size = 12) +
  theme(legend.position = "none", panel.grid.minor = element_blank())
```



DBP Before

```r
dbp_before_df <- dat_raw %>% transmute(stage = "Before (raw DBP)", value = DBP_raw)
x <- dbp_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)
dbp_before_label_y <- upper_whisker + 0.05 * iqr
dbp_before_N <- sum(!is.na(x))

ggplot(dbp_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 DBP)", y = dbp_before_label_y, N = dbp_before_N),
    aes(stage, y, label = paste0("n = ", N)),
    hjust = -1, size = 3.5, inherit.aes = FALSE
  ) +
  scale_fill_manual(values = c("Before (raw DBP)" = "#D6E9F8")) +
  labs(title = "DBP (BEFORE): Raw Distribution", x = NULL, y = "DBP") +
  scale_y_continuous(expand = expansion(mult = c(0.02, 0.12))) +
  theme_minimal(base_size = 12) +
  theme(legend.position = "none", panel.grid.minor = element_blank())
```



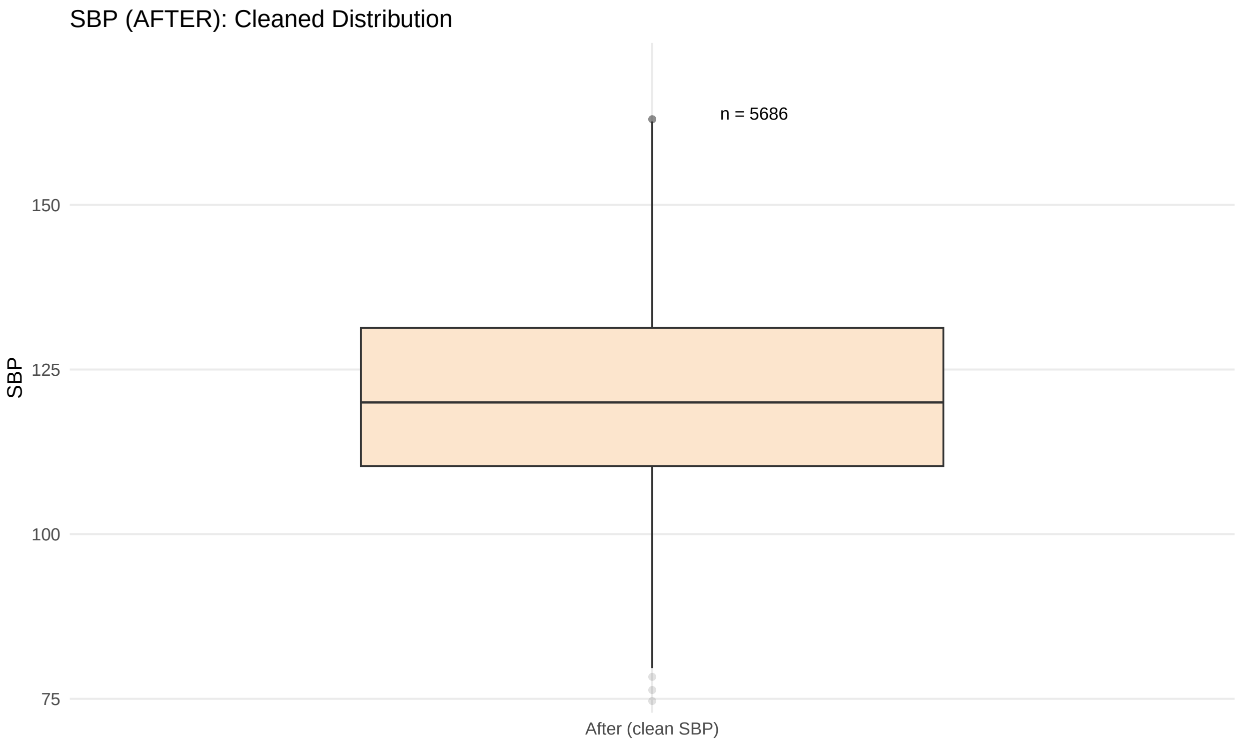
Blood Pressure Outlier Cleaning

```r
SBP_LO <- 70
SBP_HI <- 260
DBP_LO <- 40
DBP_HI <- 150

# Clean SBP
sbp_clean <- bpx %>%
  transmute(seqn, SBP_raw = rowMeans(select(., all_of(sbp_cols)), na.rm = TRUE)) %>%
  mutate(
    q1 = quantile(SBP_raw, 0.25, na.rm = TRUE),
    q3 = quantile(SBP_raw, 0.75, na.rm = TRUE),
    iqr = q3 - q1,
    lo_iqr = q1 - 1.5 * iqr,
    hi_iqr = q3 + 1.5 * iqr,
    med = median(SBP_raw, na.rm = TRUE),
    madv = mad(SBP_raw, na.rm = TRUE),
    z = ifelse(madv > 0, (SBP_raw - med) / (madv * 1.4826), 0),
    flag = (SBP_raw < SBP_LO | SBP_raw > SBP_HI) |
           (SBP_raw < lo_iqr | SBP_raw > hi_iqr) |
           (abs(z) > 3.5),
    SBP_clean = ifelse(flag, NA_real_, SBP_raw)
  ) %>%
  select(seqn, SBP_clean)

# Clean DBP
dbp_clean <- bpx %>%
  transmute(seqn, DBP_raw = rowMeans(select(., all_of(dbp_cols)), na.rm = TRUE)) %>%
  mutate(
    q1 = quantile(DBP_raw, 0.25, na.rm = TRUE),
    q3 = quantile(DBP_raw, 0.75, na.rm = TRUE),
    iqr = q3 - q1,
    lo_iqr = q1 - 1.5 * iqr,
    hi_iqr = q3 + 1.5 * iqr,
    med = median(DBP_raw, na.rm = TRUE),
    madv = mad(DBP_raw, na.rm = TRUE),
    z = ifelse(madv > 0, (DBP_raw - med) / (madv * 1.4826), 0),
    flag = (DBP_raw < DBP_LO | DBP_raw > DBP_HI) |
           (DBP_raw < lo_iqr | DBP_raw > hi_iqr) |
           (abs(z) > 3.5),
    DBP_clean = ifelse(flag, NA_real_, DBP_raw)
  ) %>%
  select(seqn, DBP_clean)

# Build cleaned BP dataset
dat_clean_bp <- demo_sex %>%
  left_join(sbp_clean, by = "seqn") %>%
  left_join(dbp_clean, by = "seqn") %>%
  filter(age >= 20) %>%
  mutate(
    SBP_clean = ifelse(is.nan(SBP_clean), NA_real_, SBP_clean),
    DBP_clean = ifelse(is.nan(DBP_clean), NA_real_, DBP_clean)
  )
```

Blood Pressure Distribution (AFTER Cleaning)

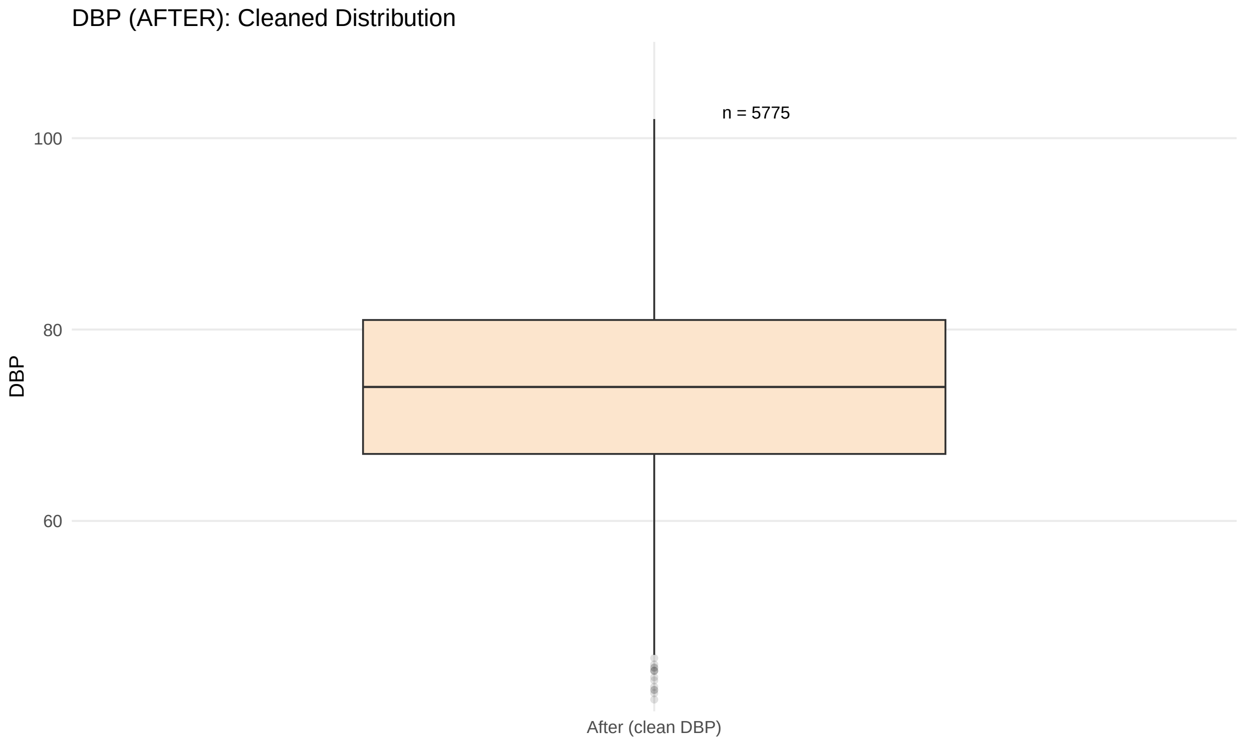
SBP After

```r
sbp_after_df <- dat_clean_bp %>% transmute(stage = "After (clean SBP)", value = SBP_clean)
x <- sbp_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)
sbp_after_label_y <- upper_whisker + 0.05 * iqr
sbp_after_N <- sum(!is.na(x))

ggplot(sbp_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 SBP)", y = sbp_after_label_y, N = sbp_after_N),
    aes(stage, y, label = paste0("n = ", N)),
    hjust = -1, size = 3.5, inherit.aes = FALSE
  ) +
  scale_fill_manual(values = c("After (clean SBP)" = "#FCE5CD")) +
  labs(title = "SBP (AFTER): Cleaned Distribution", x = NULL, y = "SBP") +
  scale_y_continuous(expand = expansion(mult = c(0.02, 0.12))) +
  theme_minimal(base_size = 12) +
  theme(legend.position = "none", panel.grid.minor = element_blank())
```



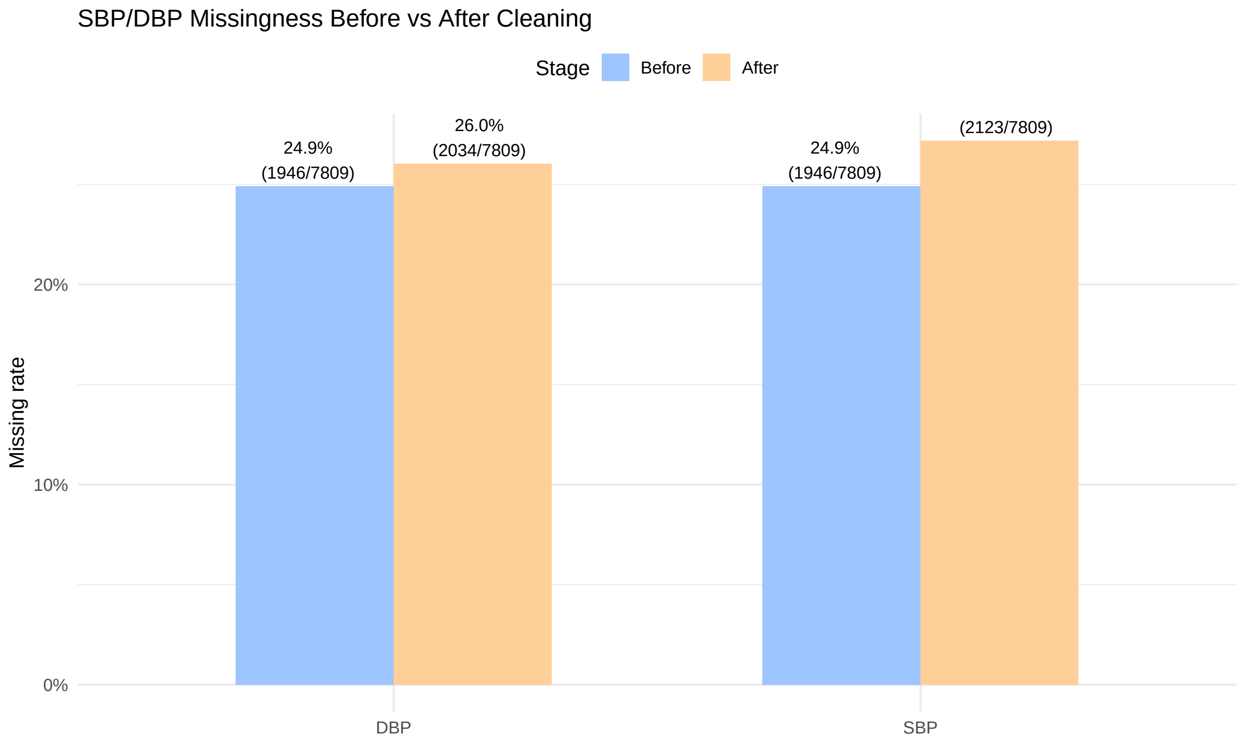
DBP After

```
dbp_after_df <- dat_clean_bp %>% transmute(stage = "After (clean DBP)", value = DBP_clean)
x <- dbp_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)
dbp_after_label_y <- upper_whisker + 0.05 * iqr
dbp_after_N <- sum(!is.na(x))

ggplot(dbp_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 DBP)", y = dbp_after_label_y, N = dbp_after_N),
    aes(stage, y, label = paste0("n = ", N)),
    hjust = -1, size = 3.5, inherit.aes = FALSE
  ) +
  scale_fill_manual(values = c("After (clean DBP)" = "#FCE5CD")) +
  labs(title = "DBP (AFTER): Cleaned Distribution", x = NULL, y = "DBP") +
  scale_y_continuous(expand = expansion(mult = c(0.02, 0.12))) +
  theme_minimal(base_size = 12) +
  theme(legend.position = "none", panel.grid.minor = element_blank())
```



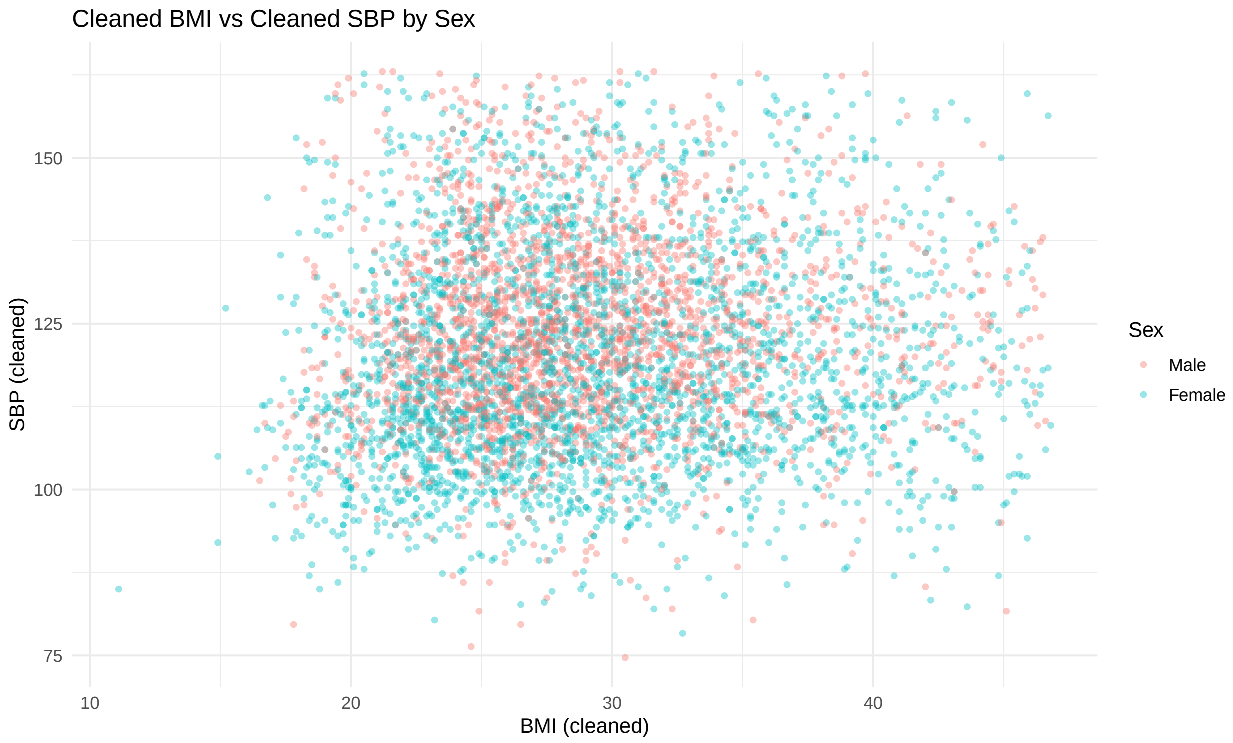
Missing Value Comparison (Blood Pressure)

```r
miss_before <- tibble(
  stage = "Before",
  variable = c("SBP", "DBP"),
  n_missing = c(sum(is.na(dat_raw$SBP_raw)), sum(is.na(dat_raw$DBP_raw))),
  n_total = nrow(dat_raw)
) %>% mutate(p_missing = n_missing / n_total)

miss_after <- tibble(
  stage = "After",
  variable = c("SBP", "DBP"),
  n_missing = c(sum(is.na(dat_clean_bp$SBP_clean)), sum(is.na(dat_clean_bp$DBP_clean))),
  n_total = nrow(dat_clean_bp)
) %>% mutate(p_missing = n_missing / n_total)

miss_long <- bind_rows(miss_before, miss_after) %>%
  mutate(stage = factor(stage, levels = c("Before", "After")))

ggplot(miss_long, aes(variable, p_missing, fill = stage)) +
  geom_col(width = 0.6, position = "dodge") +
  geom_text(
    aes(label = paste0(scales::percent(p_missing, 0.1), "\n(", n_missing, "/", n_total, ")")),
    position = position_dodge(width = 0.65), vjust = -0.2, size = 3.5
  ) +
  scale_y_continuous(labels = scales::percent) +
  scale_fill_manual(values = c("Before" = "#9EC5FE", "After" = "#FFCF99")) +
  labs(
    title = "SBP/DBP Missingness Before vs After Cleaning",
    x = NULL, y = "Missing rate", fill = "Stage"
  ) +
  theme_minimal(base_size = 12) +
  theme(legend.position = "top")
```



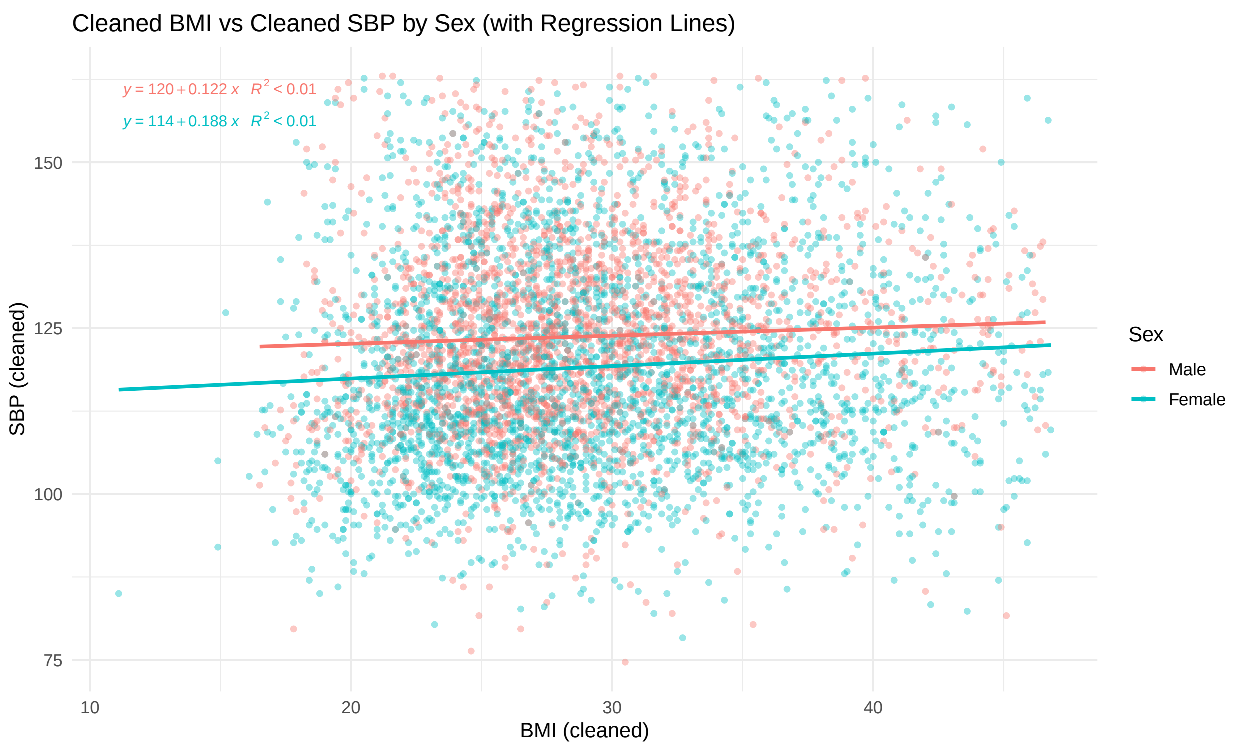
Scatter Plot: BMI vs SBP by Sex

```r
dat_scatter <- demo_sex %>%
  left_join(bmi_clean, by = "seqn") %>%
  left_join(sbp_clean, by = "seqn") %>%
  filter(age >= 20)

ggplot(dat_scatter, aes(x = bmxbmi_clean, y = SBP_clean, color = sex)) +
  geom_point(alpha = 0.4, size = 1.2) +
  labs(
    title = "Cleaned BMI vs Cleaned SBP by Sex",
    x = "BMI (cleaned)",
    y = "SBP (cleaned)",
    color = "Sex"
  ) +
  theme_minimal(base_size = 12)
```



With Regression Lines

```
ggplot(dat_scatter, aes(x = bmxbmi_clean, y = SBP_clean, color = sex)) +
  geom_point(alpha = 0.4, size = 1.2) +
  geom_smooth(method = "lm", se = FALSE) +
  stat_poly_eq(
    aes(label = paste(after_stat(eq.label), after_stat(rr.label), sep = "~~~")),
    formula = y ~ x, parse = TRUE, size = 3
  ) +
  labs(
    title = "Cleaned BMI vs Cleaned SBP by Sex (with Regression Lines)",
    x = "BMI (cleaned)",
    y = "SBP (cleaned)",
    color = "Sex"
  ) +
  theme_minimal(base_size = 12)
```



Week 6 Class: Education Level Analysis

Education Level Recoding

```r
# Check original coding
demo %>% count(dmdeduc2) %>% kable(caption = "Original Education Coding")
```

Table 8: Original Education Coding

dmdeduc2	n
1	373
2	666
3	1749
4	2370
5	2625
9	11
NA	4139

```r
# Recode and relabel education
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)
```

Education Distribution Table

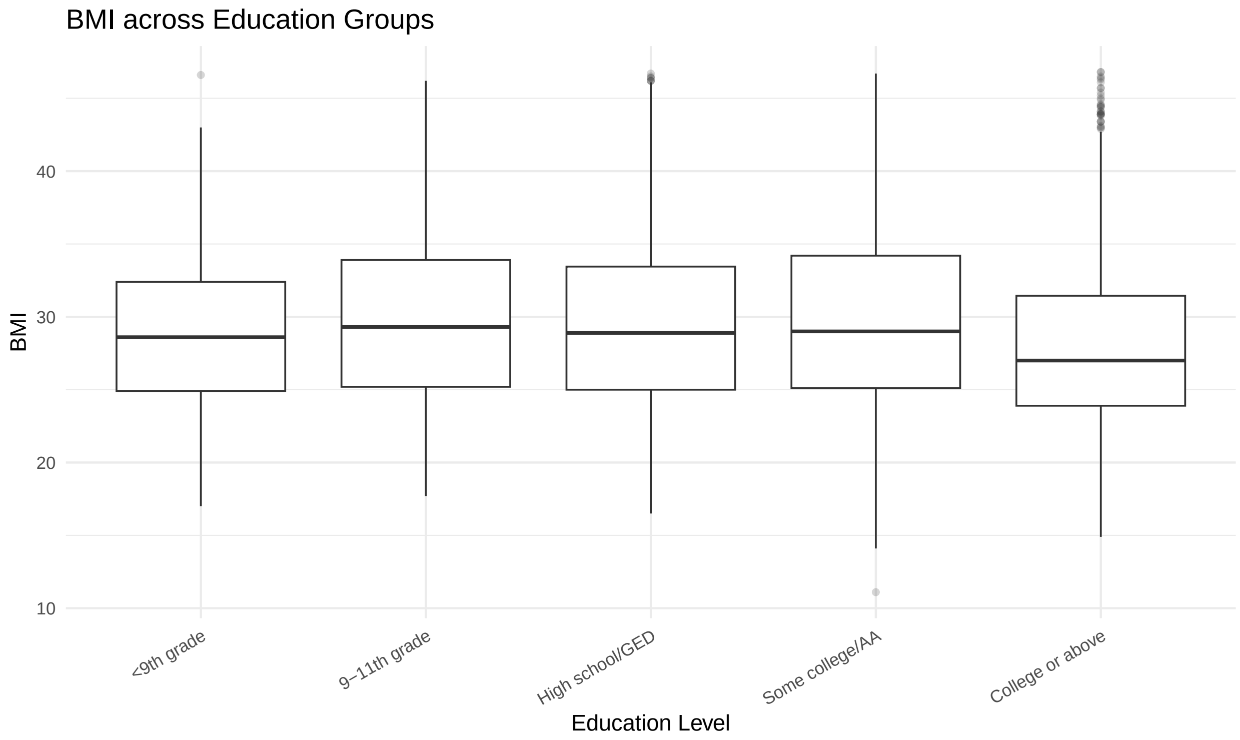
```r
edu_dist <- dat_edu %>%
  count(EDU) %>%
  mutate(
    prop = n / sum(n),
    percentage = paste0(round(prop * 100, 1), "%"),
    variable = "EDU"
  ) %>%
  rename(category = EDU)

kable(edu_dist, digits = 3, caption = "Distribution of Educational Attainment (EDU)")
```

Table 9: Distribution of Educational Attainment (EDU)

category	n	prop	percentage	variable
<9th grade	278	0.048	4.8%	EDU
9-11th grade	457	0.079	7.9%	EDU
High school/GED	1227	0.212	21.2%	EDU
Some college/AA	1749	0.302	30.2%	EDU
College or above	2079	0.359	35.9%	EDU

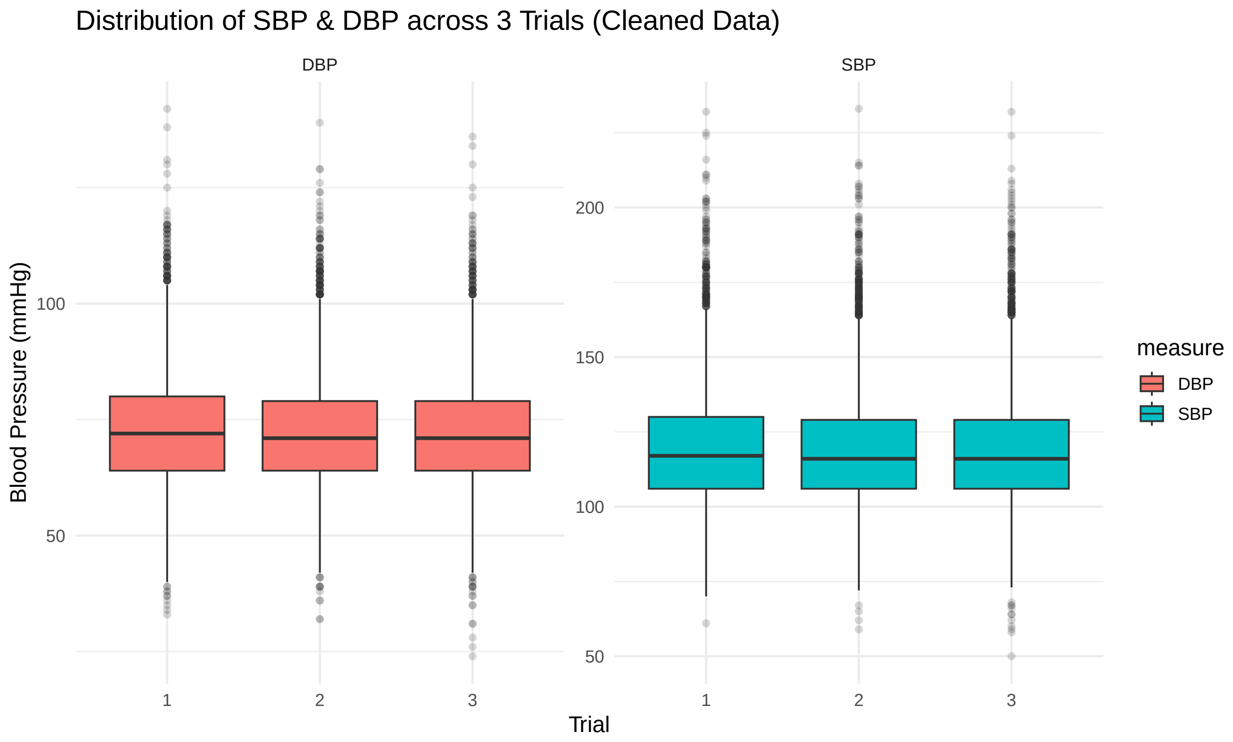
BMI by Education Level

```r
dat_edu %>%
  ggplot(aes(x = EDU, y = bmxbmi_clean)) +
  geom_boxplot(position = position_dodge(0.8), outlier.alpha = 0.2) +
  labs(
    title = "BMI across Education Groups",
    x = "Education Level",
    y = "BMI"
  ) +
  theme_minimal(base_size = 13) +
  theme(axis.text.x = element_text(angle = 30, hjust = 1))
```



Blood Pressure Trials Analysis

```r
# Transform BP data to 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)
  )

ggplot(bpx_long_clean, aes(x = factor(trial), y = value, fill = measure)) +
  geom_boxplot(outlier.alpha = 0.2) +
  facet_wrap(~ measure, scales = "free_y") +
  labs(
    title = "Distribution of SBP & DBP across 3 Trials (Cleaned Data)",
    x = "Trial",
    y = "Blood Pressure (mmHg)"
  ) +
  theme_minimal(base_size = 13)
```



Week 6 HW2: Race Distribution & BP Trial Selection

Race/Ethnicity Recoding

```
# Check original race coding
demo %>% count(ridreth3) %>% kable(caption = "Original Race/Ethnicity Coding")
```

Table 10: Original Race/Ethnicity Coding

ridreth3	n
1	1117
2	1373
3	6217
4	1597
6	681
7	948

```r
# Recode race/ethnicity
dat_race <- demo %>%
  transmute(
    seqn,
    age = ridageyr,
    Race = case_when(
      ridreth3 == 1 ~ "Mexican American",
      ridreth3 == 2 ~ "Other Hispanic",
      ridreth3 == 3 ~ "Non-Hispanic White",
      ridreth3 == 4 ~ "Non-Hispanic Black",
      ridreth3 == 6 ~ "Non-Hispanic Asian",
      ridreth3 == 7 ~ "Other/Multi-racial",
      TRUE ~ NA_character_
    )
  ) %>%
  mutate(
    Race = factor(
      Race,
      levels = c("Mexican American", "Other Hispanic",
                 "Non-Hispanic White", "Non-Hispanic Black",
                 "Non-Hispanic Asian", "Other/Multi-racial")
    )
  ) %>%
  left_join(bmi_clean, by = "seqn") %>%
  drop_na(Race, bmxbmi_clean)
```

Race Distribution Table

```r
race_dist <- dat_race %>%
  count(Race) %>%
  mutate(
    prop = n / sum(n),
    percentage = paste0(round(prop * 100, 1), "%"),
    variable = "Race"
  ) %>%
  rename(category = Race)

kable(
  race_dist,
  digits = 3,
  caption = "Distribution of Race/Ethnicity (RIDRETH3)",
  col.names = c("Race/Ethnicity", "N", "Proportion", "Percentage", "Variable")
)
```

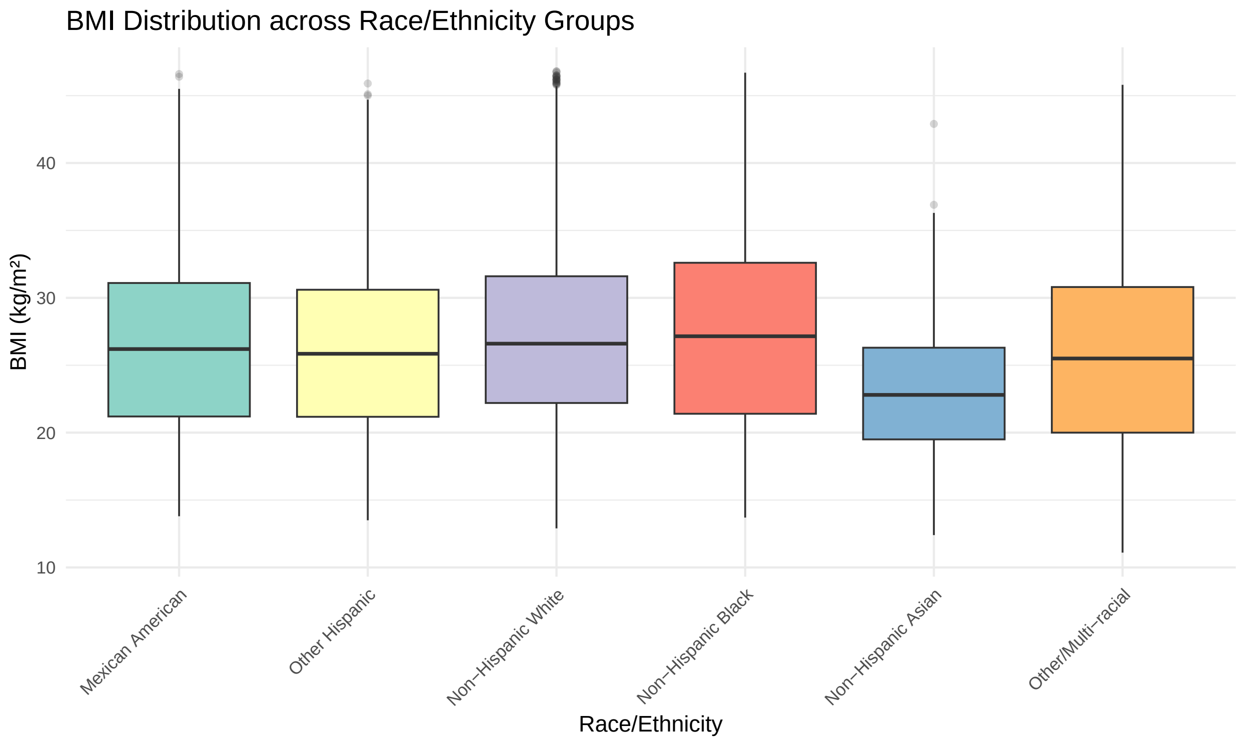
Table 11: Distribution of Race/Ethnicity (RIDRETH3)

Race/Ethnicity	N	Proportion	Percentage	Variable
Mexican American	749	0.090	9%	Race
Other Hispanic	972	0.117	11.7%	Race
Non-Hispanic White	4421	0.534	53.4%	Race

Race/Ethnicity	N	Proportion	Percentage	Variable
Non-Hispanic Black	1048	0.127	12.7%	Race
Non-Hispanic Asian	503	0.061	6.1%	Race
Other/Multi-racial	589	0.071	7.1%	Race

BMI by Race/Ethnicity

```r
dat_race %>%
  ggplot(aes(x = Race, y = bmxbmi_clean, fill = Race)) +
  geom_boxplot(outlier.alpha = 0.2) +
  labs(
    title = "BMI Distribution across Race/Ethnicity Groups",
    x = "Race/Ethnicity",
    y = "BMI (kg/m²)"
  ) +
  theme_minimal(base_size = 13) +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    legend.position = "none"
  ) +
  scale_fill_brewer(palette = "Set3")
```

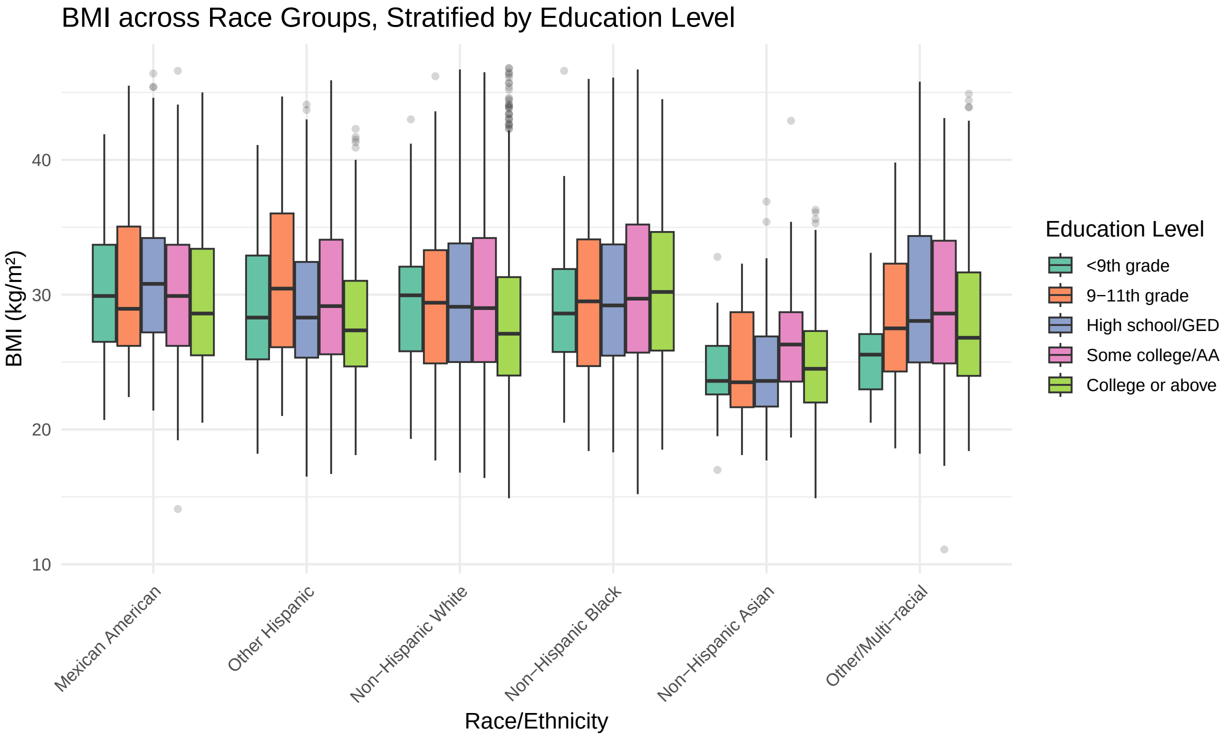


Combined Race and Education Analysis

```r
# Combine race and education data
dat_combined <- demo %>%
  transmute(
    seqn,
    age = ridageyr,
    Race = case_when(
      ridreth3 == 1 ~ "Mexican American",
      ridreth3 == 2 ~ "Other Hispanic",
      ridreth3 == 3 ~ "Non-Hispanic White",
      ridreth3 == 4 ~ "Non-Hispanic Black",
      ridreth3 == 6 ~ "Non-Hispanic Asian",
      ridreth3 == 7 ~ "Other/Multi-racial",
      TRUE ~ NA_character_
    ),
    EDU = case_when(
      dmdeduc2 %in% 1:5 ~ dmdeduc2,
      TRUE ~ NA_real_
    )
  ) %>%
  mutate(
    Race = factor(
      Race,
      levels = c("Mexican American", "Other Hispanic",
                 "Non-Hispanic White", "Non-Hispanic Black",
                 "Non-Hispanic Asian", "Other/Multi-racial")
    ),
    EDU = factor(
      EDU,
      levels = 1:5,
      labels = c("<9th grade", "9-11th grade", "High school/GED",
                 "Some college/AA", "College or above")
    )
  ) %>%
  left_join(bmi_clean, by = "seqn") %>%
  drop_na(Race, EDU, bmxbmi_clean)
```

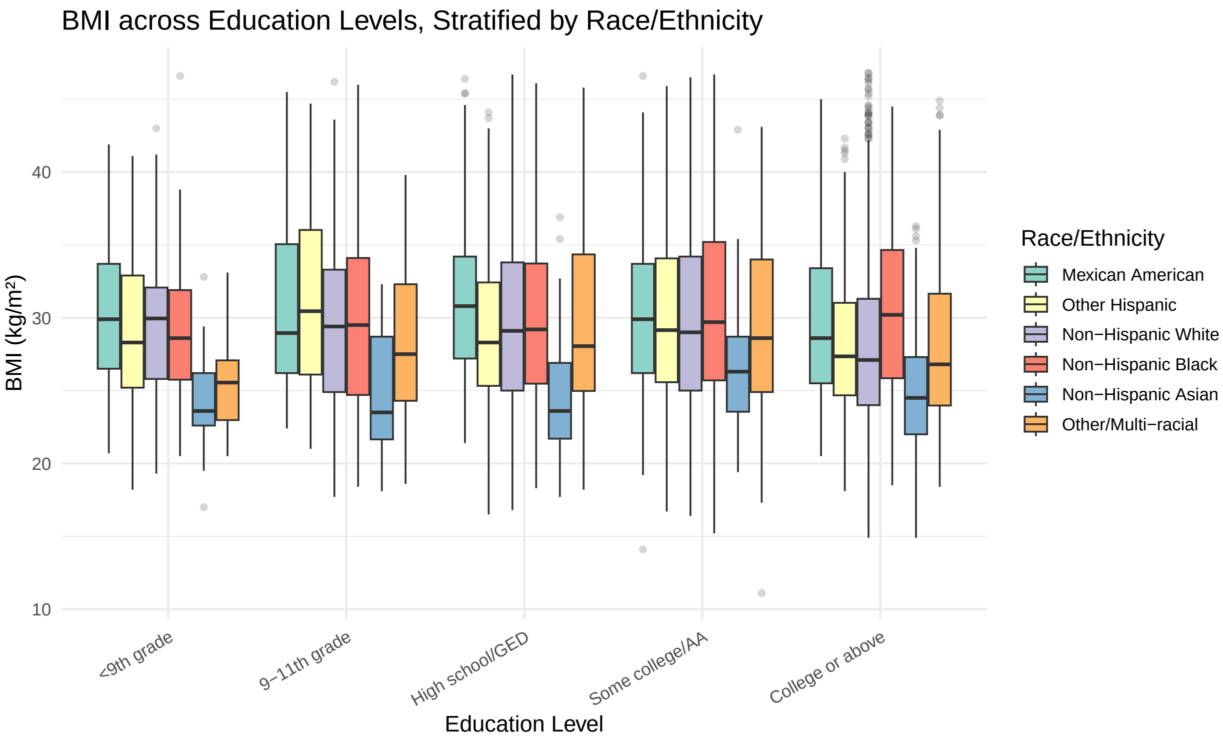
BMI by Race, Stratified by Education

```r
dat_combined %>%
  ggplot(aes(x = Race, y = bmxbmi_clean, fill = EDU)) +
  geom_boxplot(position = position_dodge(0.8), outlier.alpha = 0.2) +
  labs(
    title = "BMI across Race Groups, Stratified by Education Level",
    x = "Race/Ethnicity",
    y = "BMI (kg/m²)",
    fill = "Education Level"
  ) +
  theme_minimal(base_size = 13) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_brewer(palette = "Set2")
```



BMI by Education, Stratified by Race

```
dat_combined %>%
  ggplot(aes(x = EDU, y = bmxbmi_clean, fill = Race)) +
  geom_boxplot(position = position_dodge(0.8), outlier.alpha = 0.2) +
  labs(
    title = "BMI across Education Levels, Stratified by Race/Ethnicity",
    x = "Education Level",
    y = "BMI (kg/m²)",
    fill = "Race/Ethnicity"
  ) +
  theme_minimal(base_size = 13) +
  theme(axis.text.x = element_text(angle = 30, hjust = 1)) +
  scale_fill_brewer(palette = "Set3")
```


BMI across Education Levels, Stratified by Race/Ethnicity

BP Trial Selection: Maximum Difference

```
# Clean BP data by physiologic bounds
sbp_raw <- bpx %>%
  select(seqn, all_of(sbp_cols)) %>%
  pivot_longer(
    cols = -seqn,
    names_to = "trial",
    names_pattern = "^bpxo?sy([1-3])$",
    values_to = "value"
  ) %>%
  mutate(
    trial = as.integer(trial),
    measure = "SBP",
    value_clean = ifelse(value < SBP_LO | value > SBP_HI, NA_real_, value)
  ) %>%
  select(seqn, measure, trial, value, value_clean)

dbp_raw <- bpx %>%
  select(seqn, all_of(dbp_cols)) %>%
  pivot_longer(
    cols = -seqn,
    names_to = "trial",
    names_pattern = "^bpxo?di([1-3])$",
    values_to = "value"
  ) %>%
  mutate(
```

```r
    trial = as.integer(trial),
    measure = "DBP",
    value_clean = ifelse(value < DBP_LO | value > DBP_HI, NA_real_, value)
  ) %>%
  select(seqn, measure, trial, value, value_clean)

bpx_clean <- bind_rows(sbp_raw, dbp_raw)
```

```r
# Calculate maximum difference between any two trials
bp_max_diff <- bpx_clean %>%
  drop_na(value_clean) %>%
  group_by(seqn, measure) %>%
  filter(n() >= 2) %>%
  arrange(trial) %>%
  mutate(
    trial_list = list(trial),
    value_list = list(value_clean)
  ) %>%
  distinct(seqn, measure, trial_list, value_list) %>%
  rowwise() %>%
  mutate(
    pair_info = list({
      t <- trial_list
      v <- value_list
      combos <- combn(length(t), 2, simplify = FALSE)
      pairs <- map_dfr(combos, ~tibble(
        trial1 = t[.x[1]],
        trial2 = t[.x[2]],
        value1 = v[.x[1]],
        value2 = v[.x[2]],
        diff = abs(v[.x[1]] - v[.x[2]])
      ))
      pairs %>% slice_max(diff, n = 1, with_ties = FALSE)
    })
  ) %>%
  ungroup() %>%
  unnest(pair_info) %>%
  select(seqn, measure, trial1, trial2, value1, value2, diff)
```

```r
# Build dataset with only the two trials with maximum difference
bp_two_trials <- bpx_clean %>%
  inner_join(
    bp_max_diff %>%
      select(seqn, measure, trial1, trial2) %>%
      pivot_longer(
        cols = c(trial1, trial2),
        names_to = "pair_type",
        values_to = "trial"
      ),
    by = c("seqn", "measure", "trial")
  ) %>%
  drop_na(value_clean)
```

Summary Statistics

```r
cat("\n=== Two-Trial Selection Summary ===\n")
```

```
##
## === Two-Trial Selection Summary ===
```

```r
cat("Total subjects with valid BP measurements:\n")
```

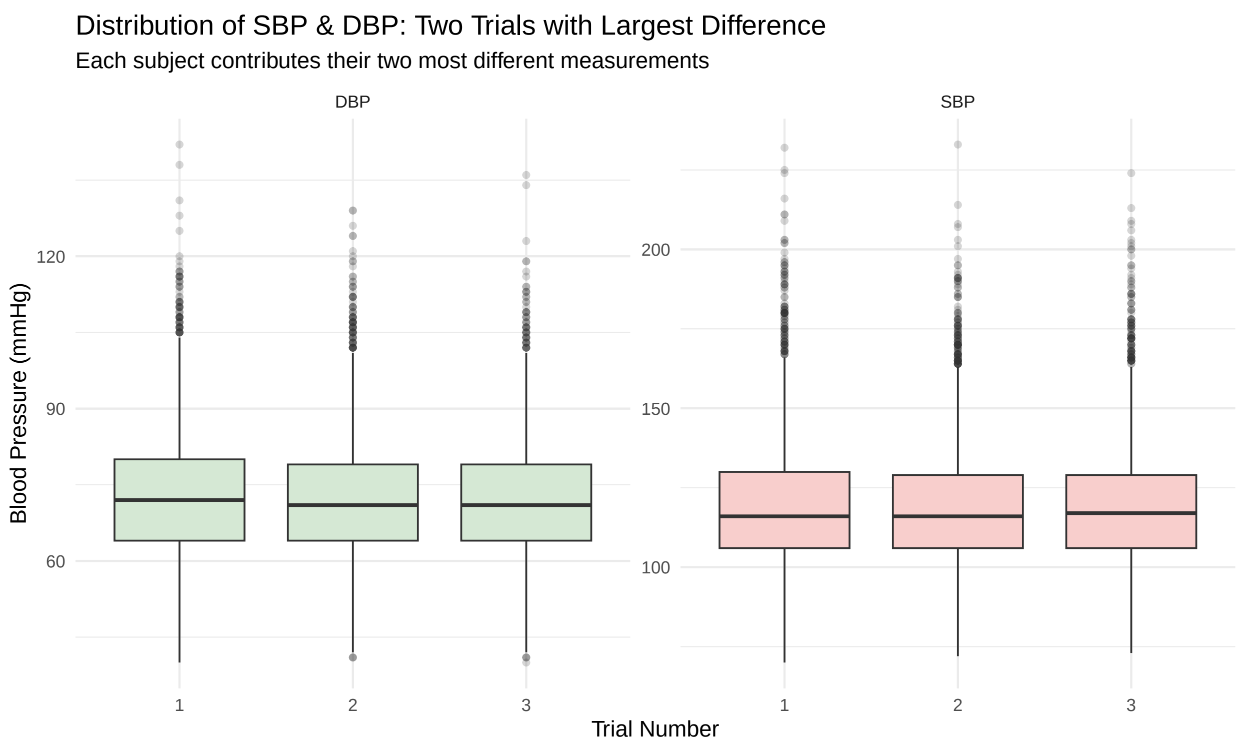
```
## Total subjects with valid BP measurements:
```

```r
bp_two_trials %>%
  distinct(seqn, measure) %>%
  count(measure) %>%
  kable(col.names = c("Measure", "N Subjects"))
```

Measure	N Subjects
DBP	7500
SBP	7505

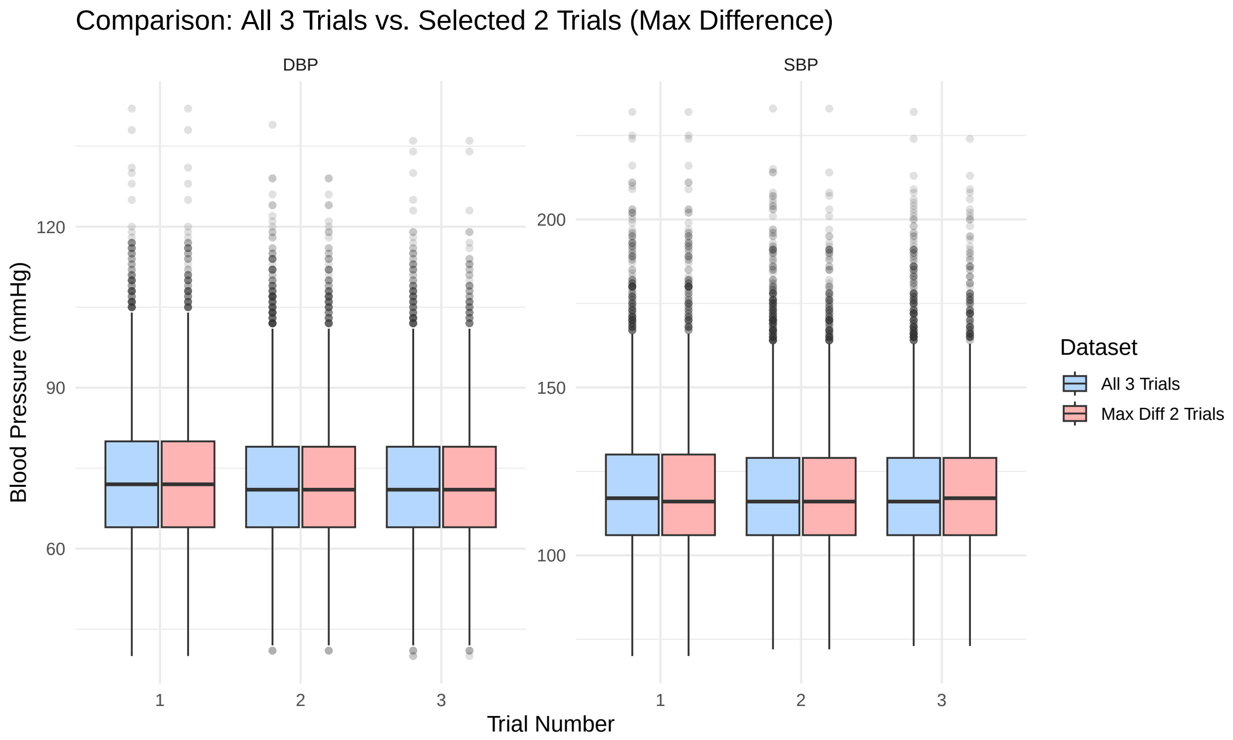
Distribution of Selected Two Trials

```r
bp_two_trials %>%
  ggplot(aes(x = factor(trial), y = value_clean, fill = measure)) +
  geom_boxplot(outlier.alpha = 0.2) +
  facet_wrap(~ measure, scales = "free_y") +
  labs(
    title = "Distribution of SBP & DBP: Two Trials with Largest Difference",
    subtitle = "Each subject contributes their two most different measurements",
    x = "Trial Number",
    y = "Blood Pressure (mmHg)",
    fill = "Measure"
  ) +
  theme_minimal(base_size = 13) +
  theme(legend.position = "none") +
  scale_fill_manual(values = c("SBP" = "#F8CECC", "DBP" = "#D5E8D4"))
```



Comparison: All 3 Trials vs Selected 2 Trials

```r
bp_all_trials <- bpx_clean %>%
  drop_na(value_clean)

bp_comparison <- bind_rows(
  bp_all_trials %>% mutate(dataset = "All 3 Trials"),
  bp_two_trials %>% mutate(dataset = "Max Diff 2 Trials")
)

bp_comparison %>%
  ggplot(aes(x = factor(trial), y = value_clean, fill = dataset)) +
  geom_boxplot(position = position_dodge(0.8), outlier.alpha = 0.15) +
  facet_wrap(~ measure, scales = "free_y") +
  labs(
    title = "Comparison: All 3 Trials vs. Selected 2 Trials (Max Difference)",
    x = "Trial Number",
    y = "Blood Pressure (mmHg)",
    fill = "Dataset"
  ) +
  theme_minimal(base_size = 13) +
  scale_fill_manual(values = c("All 3 Trials" = "#B4D7FF",
                               "Max Diff 2 Trials" = "#FFB4B4"))
```



Summary Table of Two-Trial Selection

```
bp_two_trials_summary <- bp_two_trials %>%
  group_by(seqn, measure) %>%
  summarise(
    trial_1 = min(trial),
    trial_2 = max(trial),
    value_1 = value_clean[trial == min(trial)][1],
    value_2 = value_clean[trial == max(trial)][1],
    difference = abs(value_1 - value_2),
    .groups = "drop"
  )

# Display first 20 rows
bp_two_trials_summary %>%
  head(20) %>%
  kable(
    digits = 2,
    caption = "Two-Trial Selection Summary (First 20 Subjects)",
    col.names = c("SEQN", "Measure", "Trial 1", "Trial 2",
                  "Value 1", "Value 2", "Difference")
  )
```

Table 13: Two-Trial Selection Summary (First 20 Subjects)

SEQN	Measure	Trial 1	Trial 2	Value 1	Value 2	Difference
130378	DBP	1	3	98	94	4
130378	SBP	1	2	135	131	4
130379	DBP	1	2	84	76	8
130379	SBP	1	3	121	113	8
130380	DBP	2	3	80	76	4
130380	SBP	2	3	112	104	8
130386	DBP	1	3	72	75	3
130386	SBP	1	2	110	120	10
130387	DBP	2	3	74	78	4
130387	SBP	2	3	136	145	9
130388	DBP	1	3	95	104	9
130388	SBP	1	2	130	128	2
130389	DBP	1	2	76	81	5
130389	SBP	1	3	145	124	21
130390	DBP	1	3	78	68	10
130390	SBP	2	3	106	115	9
130391	DBP	2	3	74	71	3
130391	SBP	2	3	104	109	5
130392	DBP	1	3	76	68	8
130392	SBP	1	2	154	167	13

Conclusion

This analysis of **NHANES 2021–2023** data revealed several clear demographic and physiological patterns:

- **BMI distribution:** After cleaning outliers using physiologic limits and IQR/MAD thresholds, the adult sample (n 8,800) showed a median BMI around **26.4 kg/m^2**, with the interquartile range (IQR) spanning roughly **21.6–31.7**.
  BMI variability was notably higher among **Non-Hispanic Black participants** (median 32) and among those with **lower education levels** (e.g., < high school: median 30–32).

- **Education and race differences:**
  Individuals with **college or above education (35.9%)** tended to have lower BMI values, while those with less than high school education (4.8%) showed higher medians.
  Across race/ethnicity, **Non-Hispanic White** comprised the largest group (53.4%), followed by **Non-Hispanic Black (12.7%)** and **Hispanic subgroups (20%)**.
  Combined analysis of education × race demonstrated that education effects persisted across racial groups, though magnitudes varied.

- **BMI and blood pressure association:**
  Regression lines between **BMI and SBP** indicated a weak but positive relationship, differing slightly by sex, with **R^2 0.01**.
  Males tended to have slightly higher SBP at comparable BMI levels, though both sexes exhibited wide individual variability.

- **Blood pressure measurements:**
  After cleaning, the SBP distribution centered around ~**118 mmHg**, DBP around ~**72 mmHg**.

> Variability across three BP trials was modest but nontrivial; selecting the **two trials with the largest differences** revealed deviations up to **24 mmHg in SBP** and **10–15 mmHg in DBP** for some participants, highlighting measurement instability in a subset of subjects.

Through this project, I learned the importance of maintaining **reproducible analytical workflows**—from modular code organization (data cleaning, visualization, regression) to **Quarto-based automated reporting** and version-controlled outputs (plots, CSVs).
This ensures transparent, replicable results and facilitates future extensions such as modeling BMI–BP associations adjusting for age, sex, and socioeconomic variables.

```r
sessionInfo()
```

```
## R version 4.4.3 (2025-02-28)
## Platform: aarch64-apple-darwin20
## Running under: macOS Sonoma 14.5
## 
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRlapack.dylib;  LAPACK v
## 
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
## 
## time zone: Asia/Taipei
## tzcode source: internal
## 
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
## 
## other attached packages:
##  [1] ggpmisc_0.6.2   ggpp_0.5.9      knitr_1.50      naniar_1.1.0
##  [5] skimr_2.2.1     scales_1.4.0    janitor_2.2.1   haven_2.5.4
##  [9] lubridate_1.9.4 forcats_1.0.0   stringr_1.5.1   dplyr_1.1.4
## [13] purrr_1.0.4     readr_2.1.5     tidyr_1.3.1     tibble_3.2.1
## [17] ggplot2_3.5.2   tidyverse_2.0.0
## 
## loaded via a namespace (and not attached):
##  [1] generics_0.1.4     lattice_0.22-7     stringi_1.8.7      hms_1.1.3
##  [5] digest_0.6.37      magrittr_2.0.3     confintr_1.0.2     evaluate_1.0.3
##  [9] grid_4.4.3         timechange_0.3.0   RColorBrewer_1.1-3 fastmap_1.2.0
## [13] Matrix_1.7-3       jsonlite_2.0.0     survival_3.8-3     mgcv_1.9-3
## [17] cli_3.6.5          rlang_1.1.6        splines_4.4.3      base64enc_0.1-3
## [21] withr_3.0.2        repr_1.1.7         yaml_2.3.10        tools_4.4.3
## [25] MatrixModels_0.5-4 SparseM_1.84-2     polynom_1.4-1      tzdb_0.5.0
## [29] vctrs_0.6.5        R6_2.6.1           lifecycle_1.0.4    snakecase_0.11.1
## [33] MASS_7.3-65        pkgconfig_2.0.3    pillar_1.10.2      gtable_0.3.6
## [37] glue_1.8.0         visdat_0.6.0       xfun_0.52          tidyselect_1.2.1
## [41] rstudioapi_0.17.1  farver_2.1.2       nlme_3.1-168       htmltools_0.5.8.1
## [45] labeling_0.4.3     rmarkdown_2.29     compiler_4.4.3     quantreg_6.1
```