NHANES 2021–2023 Analysis: BMI & Blood Pressure by Education & Race

李承晟

2025-10-19

Table of Contents

1. Introduction & Setup

This analysis uses the publicly-available NHANES dataset (cycles 2021–2023) to explore body mass index (BMI) and blood pressure (BP) among U.S. adults aged 20 and older. The goals are:

- Clean and explore BMI and mean systolic blood pressure (SBP) — Week 5 tasks.
- Examine BMI distributions by education level and race/ethnicity — Week 6 tasks.
- Reshape BP trials (three repeated SBP/DBP readings), compare distributions, and interpret measurement protocol.

---

2. Load Raw Data (Week 5 foundation)

```
dir.create("outputs", showWarnings = FALSE)
data_dir <- "C:/Users/user/Downloads"  # <-- adjust this path

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

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	▃▃

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
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	▇▁▁

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
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	2740	1944	4584.	1433	2167	3383	1709	▁▇

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
			4.14	9.16	46	1.75	0.19	1.33	68.3	▂ ▁ ▁
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	▁▇▅▁

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
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

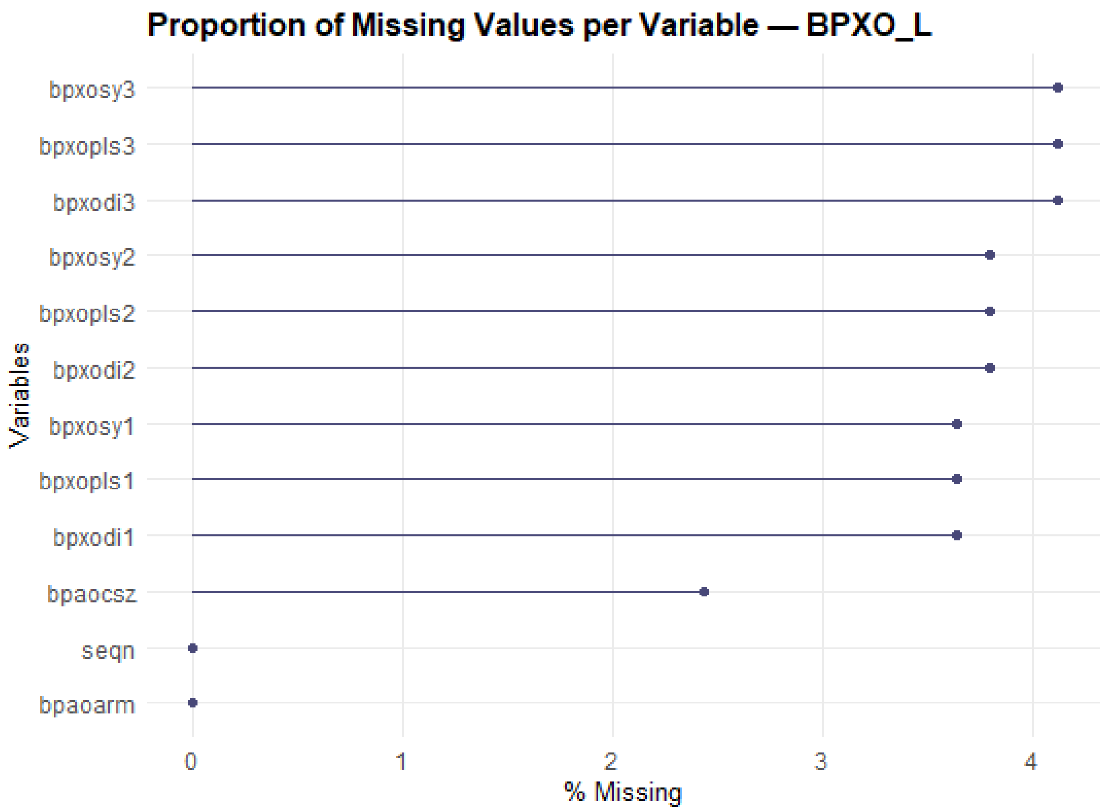
Variable type: numeric

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	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
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	▁

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
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	▁▁▁▇▁

```r
gg_miss_var(bpx, show_pct = TRUE) +
  theme_minimal(base_size = 10) +
  labs(title = "Proportion of Missing Values per Variable — BPXO_L") +
  theme(plot.title = element_text(size=12,face="bold"),
        axis.title = element_text(size=10),
        axis.text = element_text(size=9),
        panel.grid.minor = element_blank())
```



2.1 Detect SBP & DBP reading columns

```
sbp_cols <- names(bpx)[str_detect(names(bpx), "^bpxo?sy[1-3]$")]
dbp_cols <- names(bpx)[str_detect(names(bpx), "^bpxo?di[1-3]$")]

sbp_cols

## [1] "bpxosy1" "bpxosy2" "bpxosy3"

dbp_cols

## [1] "bpxodi1" "bpxodi2" "bpxodi3"
```

2.2 Build raw variables and dataset

```
sbp_raw <- bpx %>%
  select(seqn, all_of(sbp_cols)) %>%
  mutate(sbp_mean = rowMeans(select(., all_of(sbp_cols)), na.rm=TRUE),
         sbp_mean = ifelse(is.nan(sbp_mean), NA_real_, sbp_mean)) %>%
  select(seqn, sbp_mean)

dbp_raw <- bpx %>%
  select(seqn, all_of(dbp_cols)) %>%
  mutate(dbp_mean = rowMeans(select(., all_of(dbp_cols)), na.rm=TRUE),
         dbp_mean = ifelse(is.nan(dbp_mean), NA_real_, dbp_mean)) %>%
  select(seqn, dbp_mean)

bmi_raw <- bmx %>%
  transmute(seqn,
            weight = if ("bmxwt" %in% names(bmx)) bmxwt else NA_real_,
            height = if ("bmxht" %in% names(bmx)) bmxht else NA_real_,
            waist  = if ("bmxwaist" %in% names(bmx)) bmxwaist else NA_r
eal_,
            bmi_raw = bmxbmi)

demo2 <- demo %>%
  mutate(riagendr = as.numeric(riagendr)) %>%
  filter(is.na(riagendr) | riagendr %in% c(1,2))

demo_sex <- demo2 %>%
  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") %>%
  left_join(sbp_raw,  by="seqn") %>%
  left_join(dbp_raw,  by="seqn") %>%
  filter(age >= 20) %>%
  mutate(bmi_raw   = ifelse(is.nan(bmi_raw), NA_real_, bmi_raw),
         sbp_mean  = ifelse(is.nan(sbp_mean), NA_real_, sbp_mean),
         dbp_mean  = ifelse(is.nan(dbp_mean), NA_real_, dbp_mean))
```

2.3 Define plot theme for readability

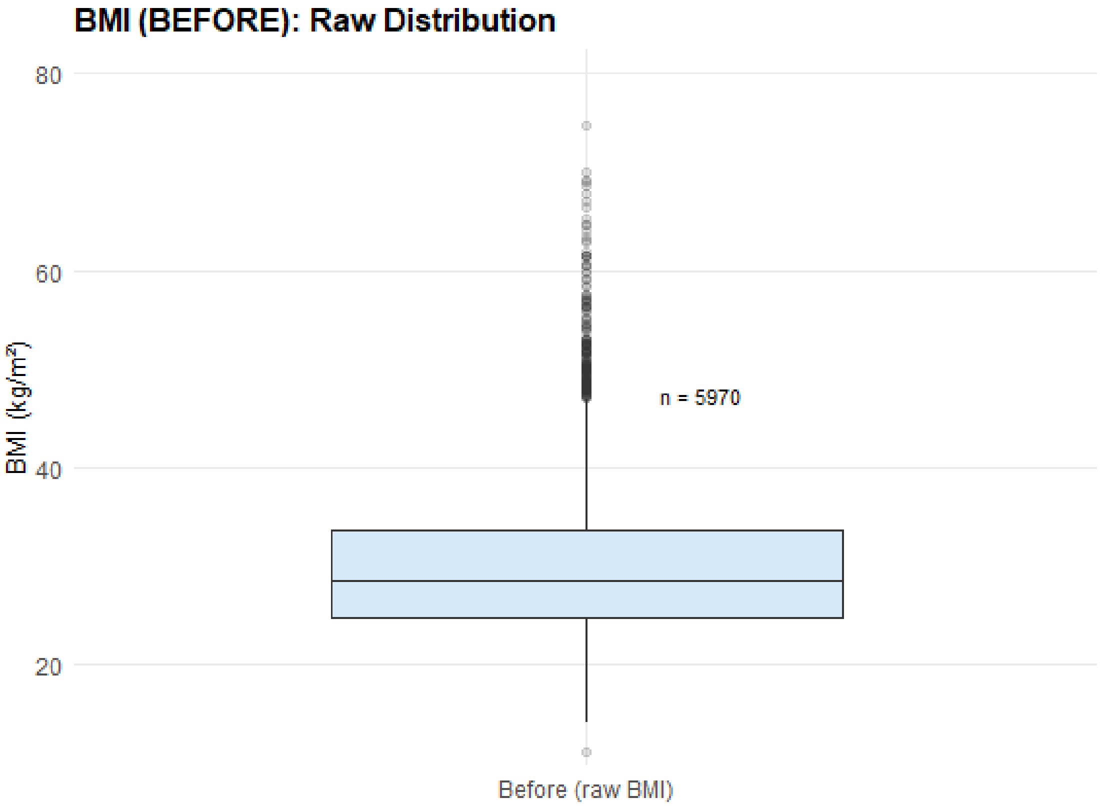
```
my_theme <- theme_minimal(base_size = 10) +
  theme(plot.title = element_text(size=12, face="bold"),
        axis.title = element_text(size=10),
        axis.text = element_text(size=9),
        legend.title = element_text(size=10),
        legend.text  = element_text(size=9),
        strip.text   = element_text(size=10),
        panel.grid.minor = element_blank())
```

2.4 BEFORE boxplots

BMI (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)
label_y <- upper_whisker + 0.05*iqr
N <- sum(!is.na(x))

p_bmi_before <- 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=label_y, N=N),
            aes(stage, y, label=paste0("n = ", N)), hjust=-0.9, size=3)
 +
  scale_fill_manual(values = c("Before (raw BMI)"="#D6E9F8")) +
  labs(title="BMI (BEFORE): Raw Distribution", x=NULL, y="BMI (kg/m²)")
 +
  scale_y_continuous(expand=expansion(mult=c(0.02,0.12))) +
  my_theme + theme(legend.position="none")

p_bmi_before
```

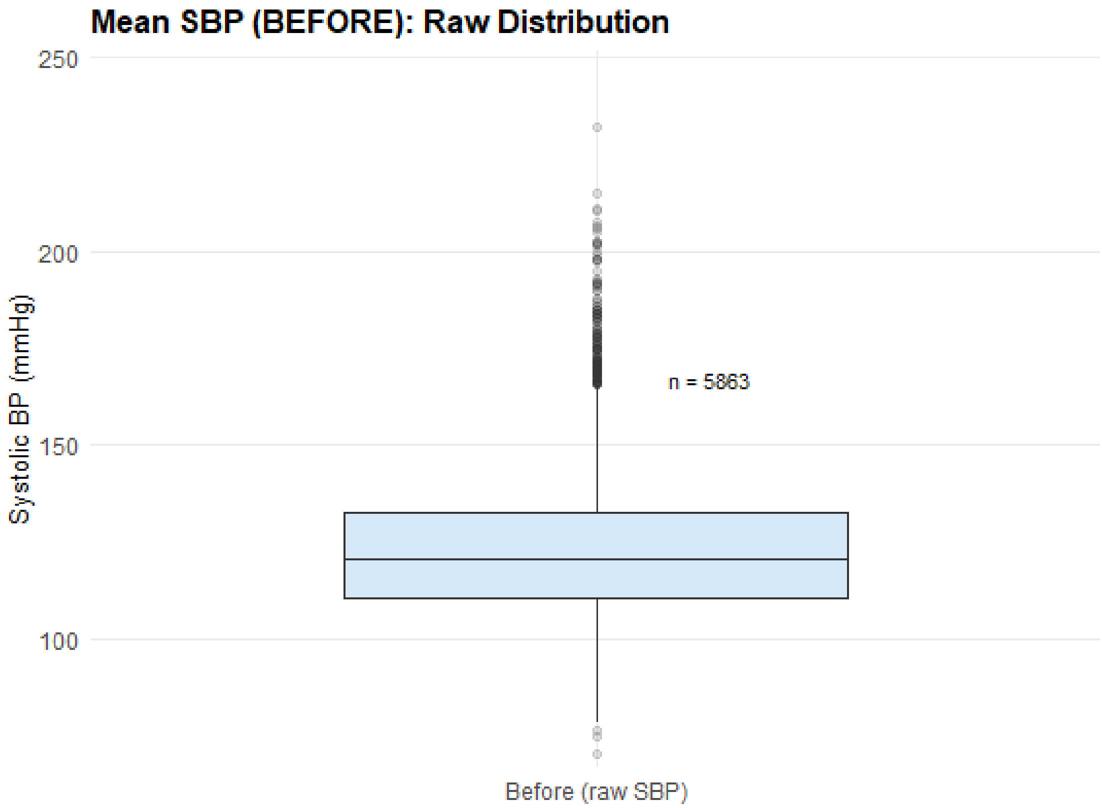


```
ggsave("outputs/q1_box_bmi_before.png", p_bmi_before, bg="white", width
=5, height=4)
```

SBP (before cleaning)

```
sbp_before_df <- dat_raw %>% transmute(stage="Before (raw SBP)", value=
sbp_mean)
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)
label_y <- upper_whisker + 0.05*iqr
N <- sum(!is.na(x))

p_sbp_before <- 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=label_y, N=N),
            aes(stage, y, label=paste0("n = ", N)), hjust=-0.9, size=3)
 +
  scale_fill_manual(values = c("Before (raw SBP)"="#D6E9F8")) +
  labs(title="Mean SBP (BEFORE): Raw Distribution", x=NULL, y="Systolic
 BP (mmHg)") +
  scale_y_continuous(expand=expansion(mult=c(0.02,0.12))) +
  my_theme + theme(legend.position="none")

p_sbp_before
```

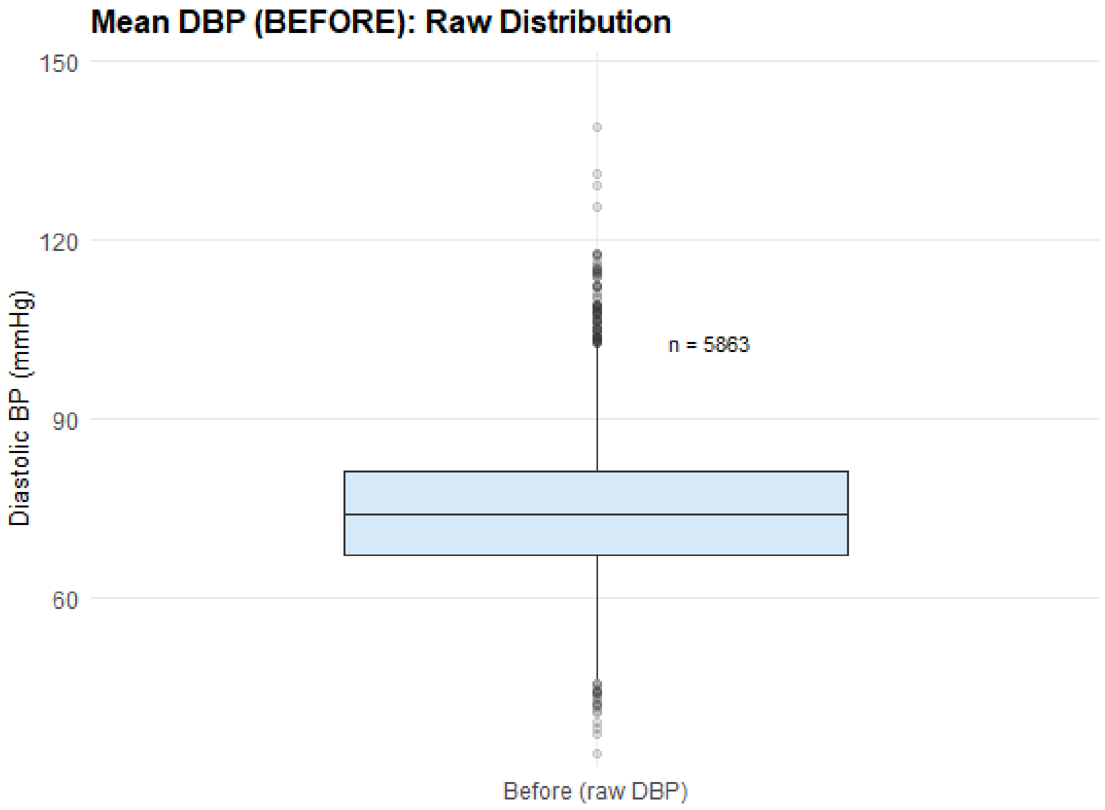


```
ggsave("outputs/q1_box_sbp_before.png", p_sbp_before, bg="white", width
=5, height=4)
```

DBP (before cleaning)

```
dbp_before_df <- dat_raw %>% transmute(stage="Before (raw DBP)", value=
dbp_mean)
x <- dbp_before_df$value
qs <- quantile(x, c(.25, .75), na.rm=TRUE); iqr <- qs[2]-qs[1]
```

```r
upper_whisker <- min(max(x, na.rm=TRUE), qs[2] + 1.5*iqr)
label_y <- upper_whisker + 0.05*iqr
N <- sum(!is.na(x))

p_dbp_before <- 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=label_y, N=N),
            aes(stage, y, label=paste0("n = ", N)), hjust=-0.9, size=3) +
  scale_fill_manual(values = c("Before (raw DBP)"="#D6E9F8")) +
  labs(title="Mean DBP (BEFORE): Raw Distribution", x=NULL, y="Diastolic BP (mmHg)") +
  scale_y_continuous(expand=expansion(mult=c(0.02,0.12))) +
  my_theme + theme(legend.position="none")

p_dbp_before
```



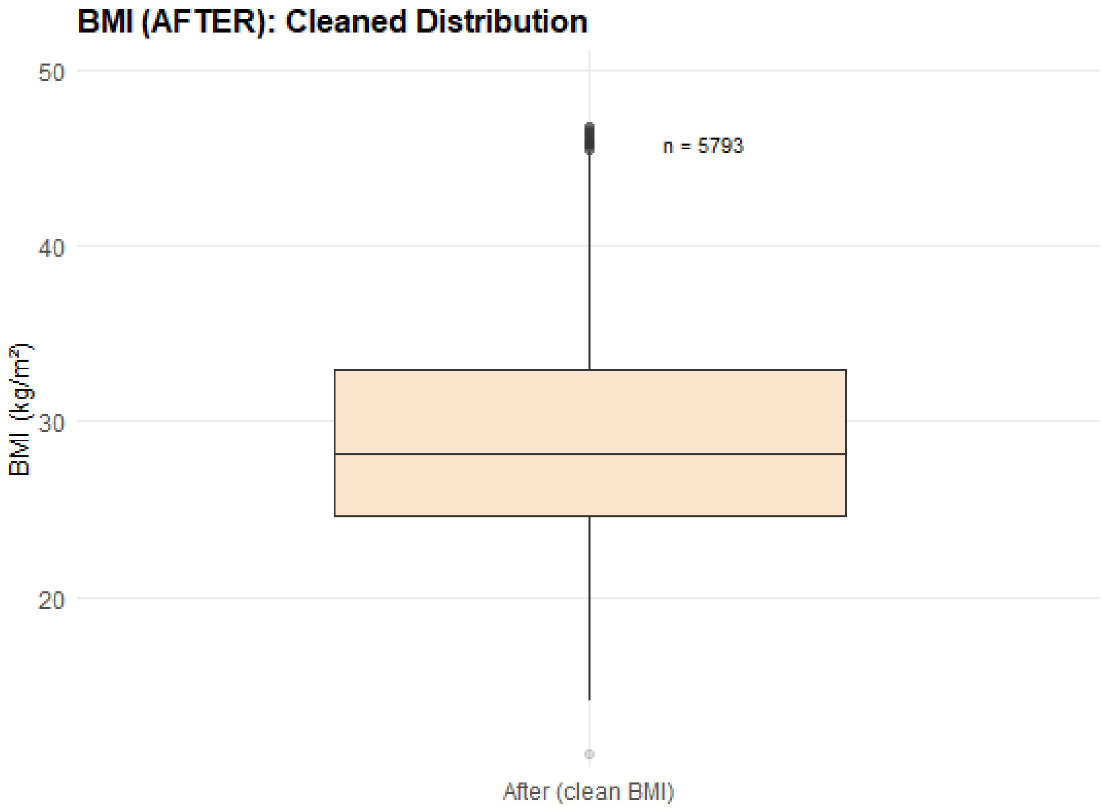
```r
ggsave("outputs/q1_box_dbp_before.png", p_dbp_before, bg="white", width=5, height=4)
```

3. Outlier Cleaning & AFTER Dataset

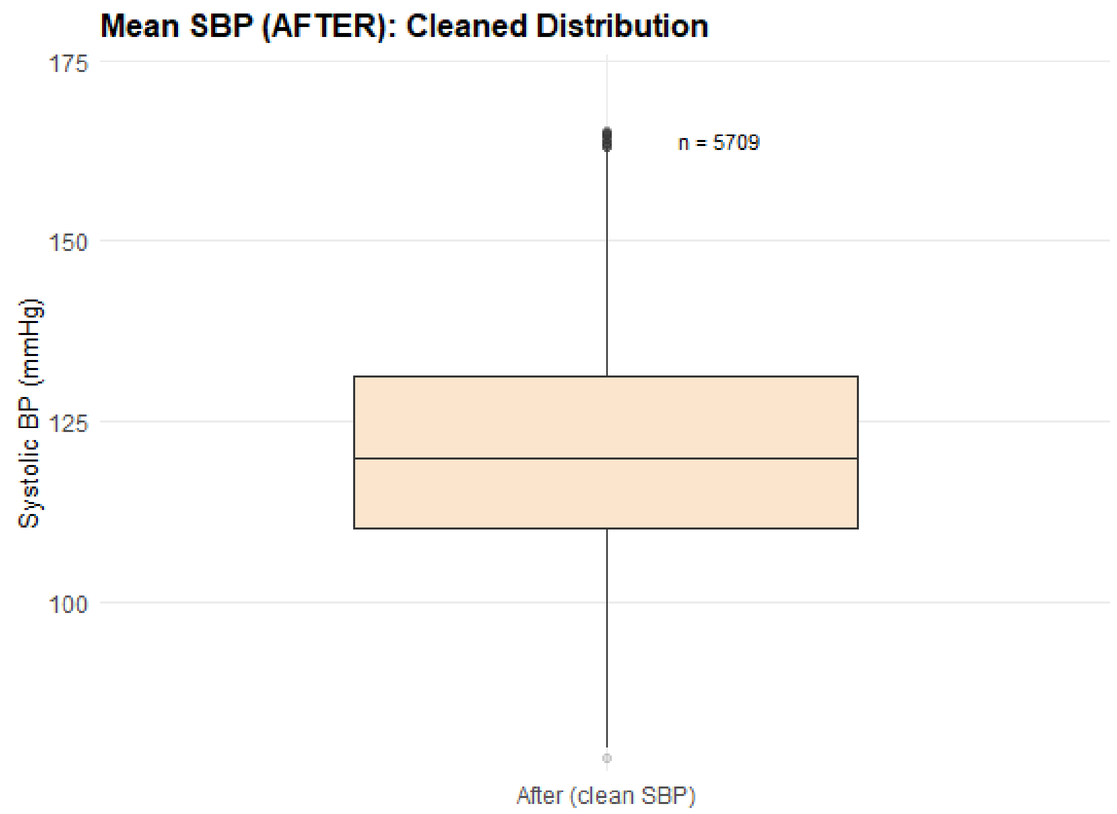
```r
BMI_LO <- 10; BMI_HI <- 80
SBP_LO <- 70; SBP_HI <- 260
DBP_LO <- 40; DBP_HI <- 150

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)

sbp_clean <- dat_raw %>%
  select(seqn, sbp_mean) %>%
  mutate(q1 = quantile(sbp_mean, 0.25, na.rm=TRUE),
         q3 = quantile(sbp_mean, 0.75, na.rm=TRUE),
         iqr = q3 - q1,
         lo_iqr = q1 - 1.5*iqr,
         hi_iqr = q3 + 1.5*iqr,
         med = median(sbp_mean, na.rm=TRUE),
         madv = mad(sbp_mean, na.rm=TRUE),
         z = ifelse(madv>0, (sbp_mean - med)/(madv*1.4826), 0),
         flag = (sbp_mean < SBP_LO | sbp_mean > SBP_HI) |
                (sbp_mean < lo_iqr | sbp_mean > hi_iqr) |
                (abs(z)>3.5),
         sbp_mean_clean = ifelse(flag, NA_real_, sbp_mean)
        ) %>% select(seqn, sbp_mean_clean)

dbp_clean <- dat_raw %>%
  select(seqn, dbp_mean) %>%
  mutate(q1 = quantile(dbp_mean, 0.25, na.rm=TRUE),
         q3 = quantile(dbp_mean, 0.75, na.rm=TRUE),
         iqr = q3 - q1,
         lo_iqr = q1 - 1.5*iqr,
         hi_iqr = q3 + 1.5*iqr,
         med = median(dbp_mean, na.rm=TRUE),
         madv = mad(dbp_mean, na.rm=TRUE),
         z = ifelse(madv>0, (dbp_mean - med)/(madv*1.4826), 0),
         flag = (dbp_mean < DBP_LO | dbp_mean > DBP_HI) |
                (dbp_mean < lo_iqr | dbp_mean > hi_iqr) |
                (abs(z)>3.5),
         dbp_mean_clean = ifelse(flag, NA_real_, dbp_mean)
        ) %>% select(seqn, dbp_mean_clean)

dat_clean <- demo_sex %>%
  left_join(bmi_clean,  by="seqn") %>%
  left_join(sbp_clean,  by="seqn") %>%
  left_join(dbp_clean,  by="seqn") %>%
  filter(age >= 20) %>%
  mutate(bmxbmi_clean   = ifelse(is.nan(bmxbmi_clean),   NA_real_, bmxb
mi_clean),
         sbp_mean_clean = ifelse(is.nan(sbp_mean_clean), NA_real_, sbp_
mean_clean),
         dbp_mean_clean = ifelse(is.nan(dbp_mean_clean), NA_real_, dbp_
mean_clean))
```

3.1 AFTER boxplots

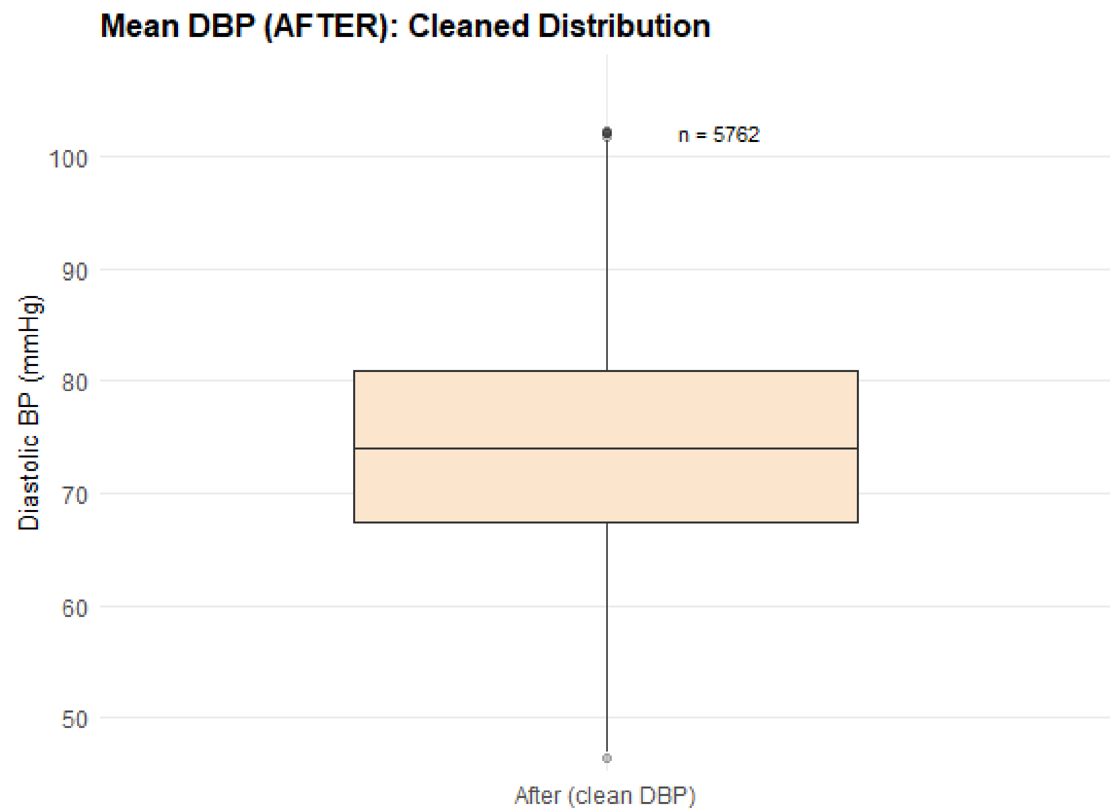
```
# BMI
bmi_after_df <- dat_clean %>% transmute(stage="After (clean BMI)", valu
e=bmxbmi_clean)
x <- bmi_after_df$value; qs <- quantile(x, c(.25,.75), na.rm=TRUE); iqr
 <- qs[2]-qs[1]
upper <- min(max(x, na.rm=TRUE), qs[2]+1.5*iqr); label_y <- upper + 0.0
5*iqr; N <- sum(!is.na(x))
p_bmi_after <- ggplot(bmi_after_df, 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=label_y, N=N),
            aes(stage,y,label=paste0("n = ",N)),hjust=-0.9,size=3) +
  scale_fill_manual(values=c("After (clean BMI)"="#FCE5CD")) +
  labs(title="BMI (AFTER): Cleaned Distribution", x=NULL, y="BMI (kg/m²)
") +
  scale_y_continuous(expand=expansion(mult=c(0.02,0.12))) +
  my_theme + theme(legend.position="none")

p_bmi_after
```



```
ggsave("outputs/q1_box_bmi_after.png", p_bmi_after, bg="white", width=5,
 height=4)

# SBP
sbp_after_df <- dat_clean %>% transmute(stage="After (clean SBP)", valu
e=sbp_mean_clean)
x <- sbp_after_df$value; qs <- quantile(x,c(.25,.75),na.rm=TRUE); iqr <
- qs[2]-qs[1]
upper <- min(max(x, na.rm=TRUE), qs[2]+1.5*iqr); label_y <- upper + 0.0
5*iqr; N <- sum(!is.na(x))
p_sbp_after <- 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=label_y, N=N),
            aes(stage,y,label=paste0("n = ",N)),hjust=-0.9,size=3) +
  scale_fill_manual(values=c("After (clean SBP)"="#FCE5CD")) +
  labs(title="Mean SBP (AFTER): Cleaned Distribution", x=NULL, y="Systo
lic BP (mmHg)") +
  scale_y_continuous(expand=expansion(mult=c(0.02,0.12))) +
  my_theme + theme(legend.position="none")

p_sbp_after
```



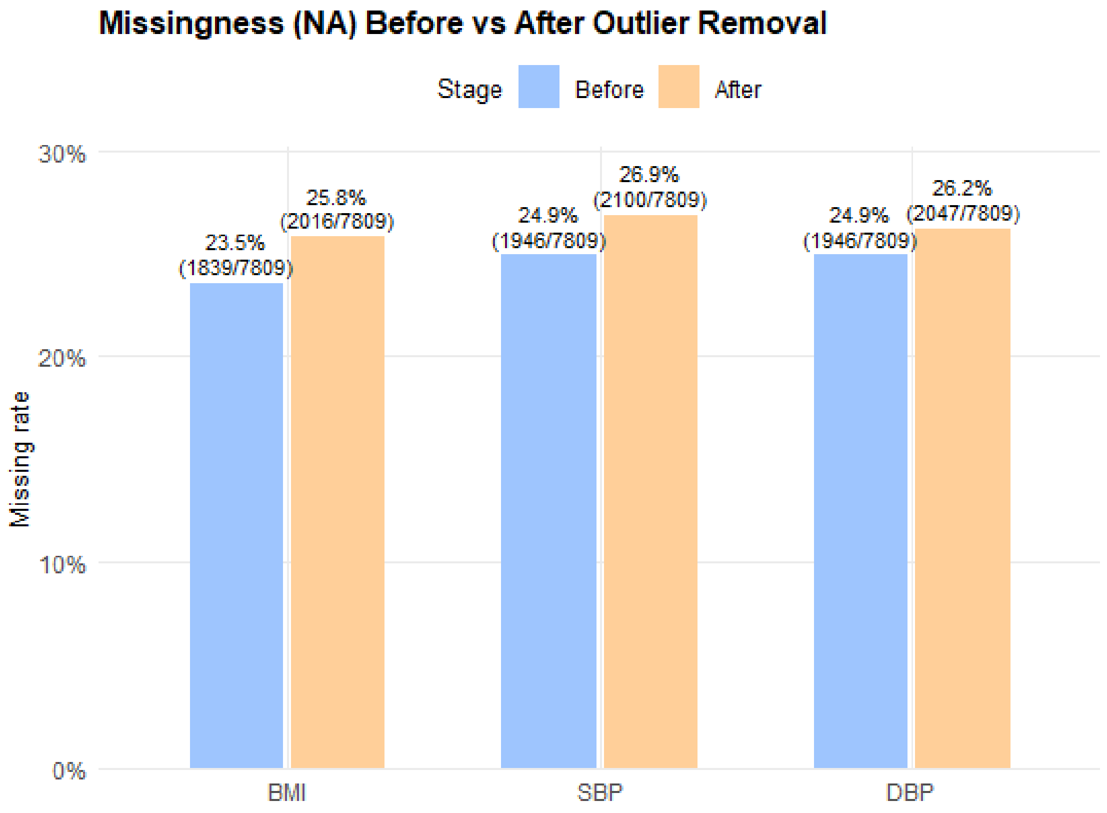
```r
ggsave("outputs/q1_box_sbp_after.png", p_sbp_after, bg="white", width=5, height=4)

# DBP
dbp_after_df <- dat_clean %>% transmute(stage="After (clean DBP)", value=dbp_mean_clean)
x <- dbp_after_df$value; qs <- quantile(x,c(.25,.75),na.rm=TRUE); iqr <- qs[2]-qs[1]
upper <- min(max(x, na.rm=TRUE), qs[2]+1.5*iqr); label_y <- upper + 0.05*iqr; N <- sum(!is.na(x))
p_dbp_after <- 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=label_y, N=N),
            aes(stage,y,label=paste0("n = ",N)),hjust=-0.9,size=3) +
  scale_fill_manual(values=c("After (clean DBP)"="#FCE5CD")) +
  labs(title="Mean DBP (AFTER): Cleaned Distribution", x=NULL, y="Diastolic BP (mmHg)") +
  scale_y_continuous(expand=expansion(mult=c(0.02,0.12))) +
  my_theme + theme(legend.position="none")

p_dbp_after
```



```
ggsave("outputs/q1_box_dbp_after.png", p_dbp_after, bg="white", width=5, height=4)
```

3.2 Missingness comparison

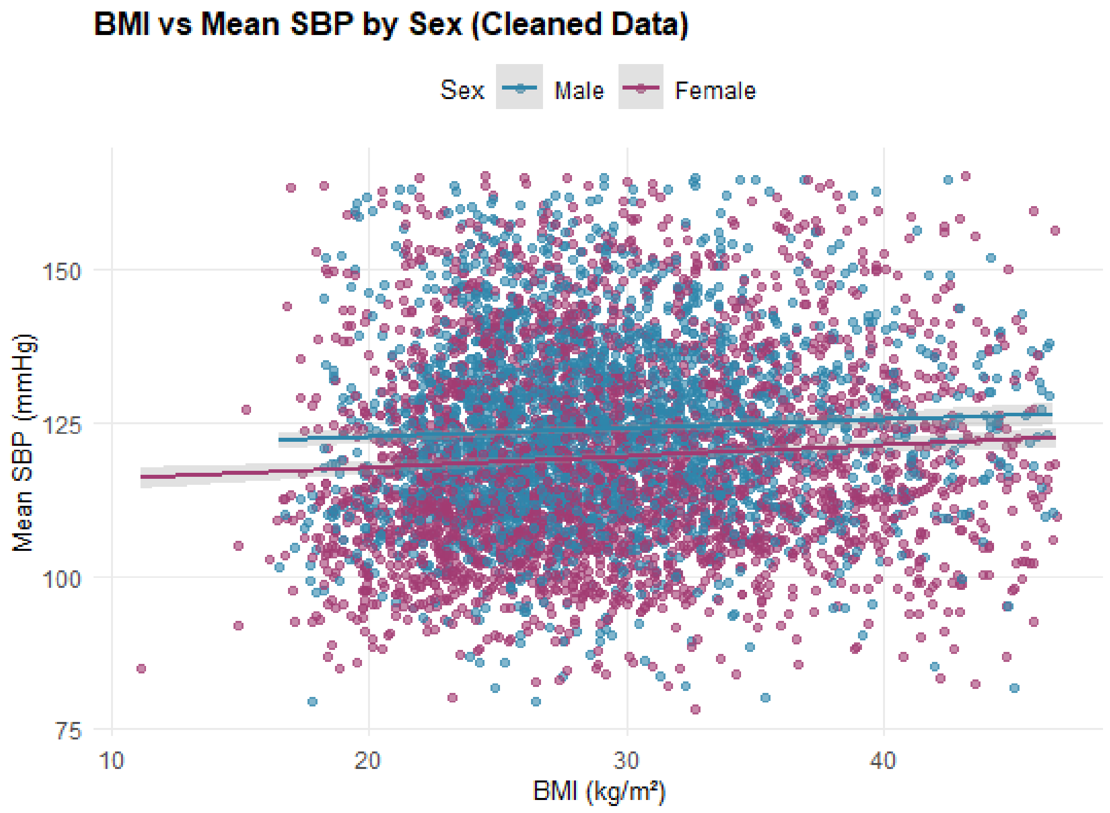
```
miss_bmi_before <- tibble(stage="Before",  variable="BMI", n_missing=sum(is.na(dat_raw$bmi_raw)),     n_total=nrow(dat_raw)) %>% mutate(p_missing=n_missing/n_total)
miss_bmi_after  <- tibble(stage="After",   variable="BMI", n_missing=sum(is.na(dat_clean$bmxbmi_clean)), n_total=nrow(dat_clean)) %>% mutate(p_missing=n_missing/n_total)
miss_sbp_before<- tibble(stage="Before",  variable="SBP", n_missing=sum(is.na(dat_raw$sbp_mean)),     n_total=nrow(dat_raw)) %>% mutate(p_missing=n_missing/n_total)
miss_sbp_after <- tibble(stage="After",   variable="SBP", n_missing=sum(is.na(dat_clean$sbp_mean_clean)),n_total=nrow(dat_clean)) %>% mutate(p_missing=n_missing/n_total)
miss_dbp_before<- tibble(stage="Before",  variable="DBP", n_missing=sum(is.na(dat_raw$dbp_mean)),     n_total=nrow(dat_raw)) %>% mutate(p_missing=n_missing/n_total)
miss_dbp_after <- tibble(stage="After",   variable="DBP", n_missing=sum(is.na(dat_clean$dbp_mean_clean)),n_total=nrow(dat_clean)) %>% mutate(p_missing=n_missing/n_total)

miss_long <- bind_rows(miss_bmi_before, miss_bmi_after, miss_sbp_before,
 miss_sbp_after, miss_dbp_before, miss_dbp_after) %>%
  mutate(stage = factor(stage, levels=c("Before","After")),
         variable = factor(variable, levels=c("BMI","SBP","DBP")))

pos <- position_dodge(width = 0.65)
p_na_bar <- 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_m
issing, "/", n_total, ")")),
            position=pos, vjust=-0.2, size=3, 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", x=NULL,
 y="Missing rate", fill="Stage") +
  my_theme + theme(legend.position="top")

p_na_bar
```



```
ggsave("outputs/q1_na_before_after_allvars.png", p_na_bar, bg="white",
width=6, height=4)
```

3.3 Scatter plot: BMI vs SBP by Sex

```
p_scatter_sbp <- ggplot(dat_clean, aes(x = bmxbmi_clean, y = sbp_mean_c
lean, color = sex)) +
  geom_point(alpha=0.6, size=1.5, na.rm=TRUE) +
  geom_smooth(method="lm", se=TRUE, alpha=0.3, na.rm=TRUE) +
  scale_color_manual(values = c("Male"="#2E86AB", "Female"="#A23B72"))
+
  labs(title = "BMI vs Mean SBP by Sex (Cleaned Data)",
       x = "BMI (kg/m²)", y = "Mean SBP (mmHg)", color = "Sex") +
  my_theme + theme(legend.position="top")

p_scatter_sbp
```



```
ggsave("outputs/q1_scatter_bmi_sbp_by_sex.png", p_scatter_sbp, bg="whit
e", width=5.5, height=4)

cat("Summary of cleaned data:\n")

## Summary of cleaned data:

cat("BMI – Mean:", round(mean(dat_clean$bmxbmi_clean, na.rm=TRUE),2),
    " SD:", round(sd(dat_clean$bmxbmi_clean,na.rm=TRUE),2), "\n")

## BMI – Mean: 29.12  SD: 6.17

cat("SBP – Mean:", round(mean(dat_clean$sbp_mean_clean,na.rm=TRUE),2),
    " SD:", round(sd(dat_clean$sbp_mean_clean,na.rm=TRUE),2), "\n")

## SBP – Mean: 121.47  SD: 15.66
```

```
cat("DBP – Mean:", round(mean(dat_clean$dbp_mean_clean,na.rm=TRUE),2),
    " SD:", round(sd(dat_clean$dbp_mean_clean,na.rm=TRUE),2), "\n")

## DBP – Mean: 74.37  SD: 10.08

cat("Total sample size:", nrow(dat_clean), "\n")

## Total sample size: 7809

cat("Complete cases (BMI & SBP):", sum(complete.cases(dat_clean$bmxbmi_
clean, dat_clean$sbp_mean_clean)), "\n")

## Complete cases (BMI & SBP): 5501

cat("Complete cases (BMI & DBP):", sum(complete.cases(dat_clean$bmxbmi_
clean, dat_clean$dbp_mean_clean)), "\n")

## Complete cases (BMI & DBP): 5558
```

---

4. Week 6: Education, Race & BMI Distributions

4.1 Recode Education & Race

```
# Education (DMDEDUC2) - adults 20+
edu_levels <- c(
  "1" = "Less than 9th grade",
  "2" = "9-11th grade (incl 12th, no diploma)",
  "3" = "High school grad/GED or equivalent",
  "4" = "Some college or AA degree",
  "5" = "College graduate or above",
  "7" = "Refused",
  "9" = "Don't know"
)

# Race/Ethnicity (RIDRETH3)
race_levels <- c(
  "1" = "Mexican American",
  "2" = "Other Hispanic",
  "3" = "Non-Hispanic White",
  "4" = "Non-Hispanic Black",
  "6" = "Non-Hispanic Asian",
  "7" = "Other NH (incl multiracial)"
)

demo_recoded <- demo2 %>%
  transmute(
    seqn,
    edu   = factor(as.character(dmdeduc2),   levels = names(edu_levels),
  labels = unname(edu_levels)),
    race = factor(as.character(ridreth3),   levels = names(race_levels),
labels = unname(race_levels))
  )

dat_clean2 <- dat_clean %>%
  left_join(demo_recoded, by="seqn")
```

4.2 Frequency tables & proportions; export to CSV

```
edu_tab <- dat_clean2 %>%
  count(edu, name="n") %>%
  mutate(prop = n/sum(n)) %>%
  arrange(desc(n))

race_tab <- dat_clean2 %>%
  count(race, name="n") %>%
  mutate(prop = n/sum(n)) %>%
  arrange(desc(n))

write_csv(edu_tab,  "outputs/edu_distribution.csv")
write_csv(race_tab,"outputs/race_distribution.csv")

knitr::kable(edu_tab,  caption = "Education Distribution (DMDEDUC2)")
```

Education Distribution (DMDEDUC2)

edu	n	prop
College graduate or above	2625	0.3361506
Some college or AA degree	2370	0.3034960
High school grad/GED or equivalent	1749	0.2239723
9-11th grade (incl 12th, no diploma)	666	0.0852862
Less than 9th grade	373	0.0477654
NA	15	0.0019209
Don’t know	11	0.0014086

```
knitr::kable(race_tab, caption = "Race/Ethnicity Distribution (RIDRETH3)
")
```

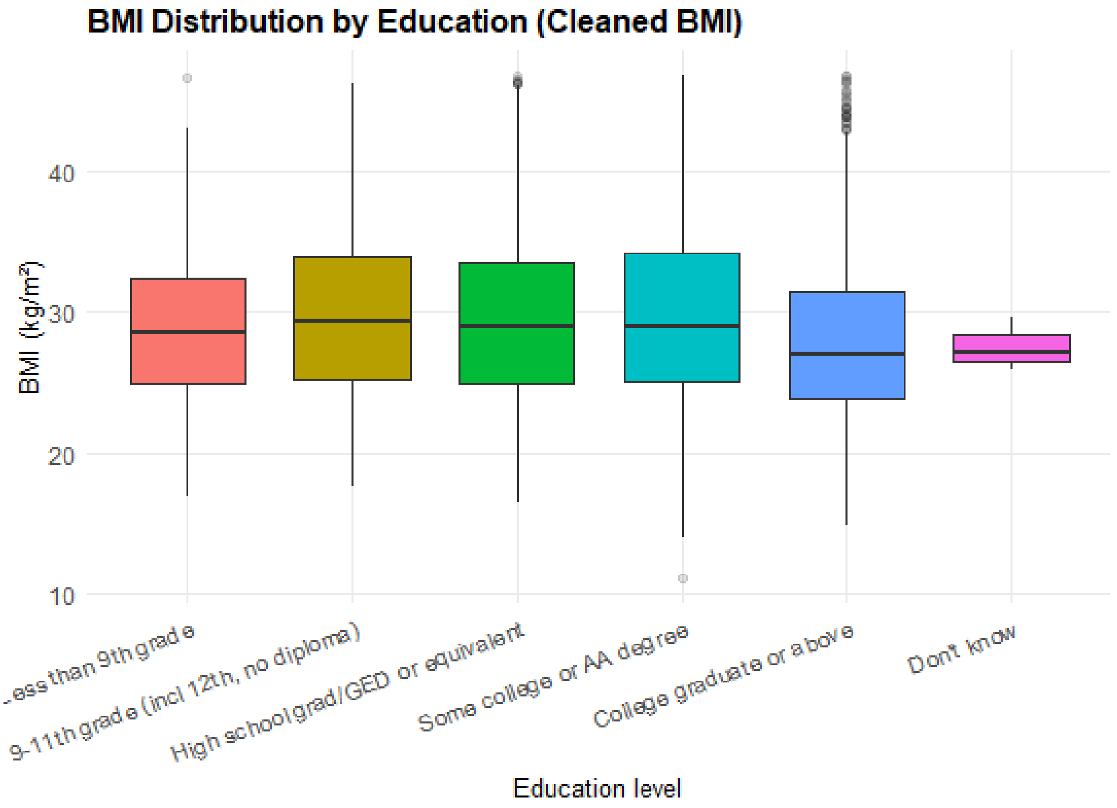
Race/Ethnicity Distribution (RIDRETH3)

race	n	prop
Non-Hispanic White	4555	0.5833013
Non-Hispanic Black	995	0.1274171
Other Hispanic	780	0.0998847
Mexican American	545	0.0697913

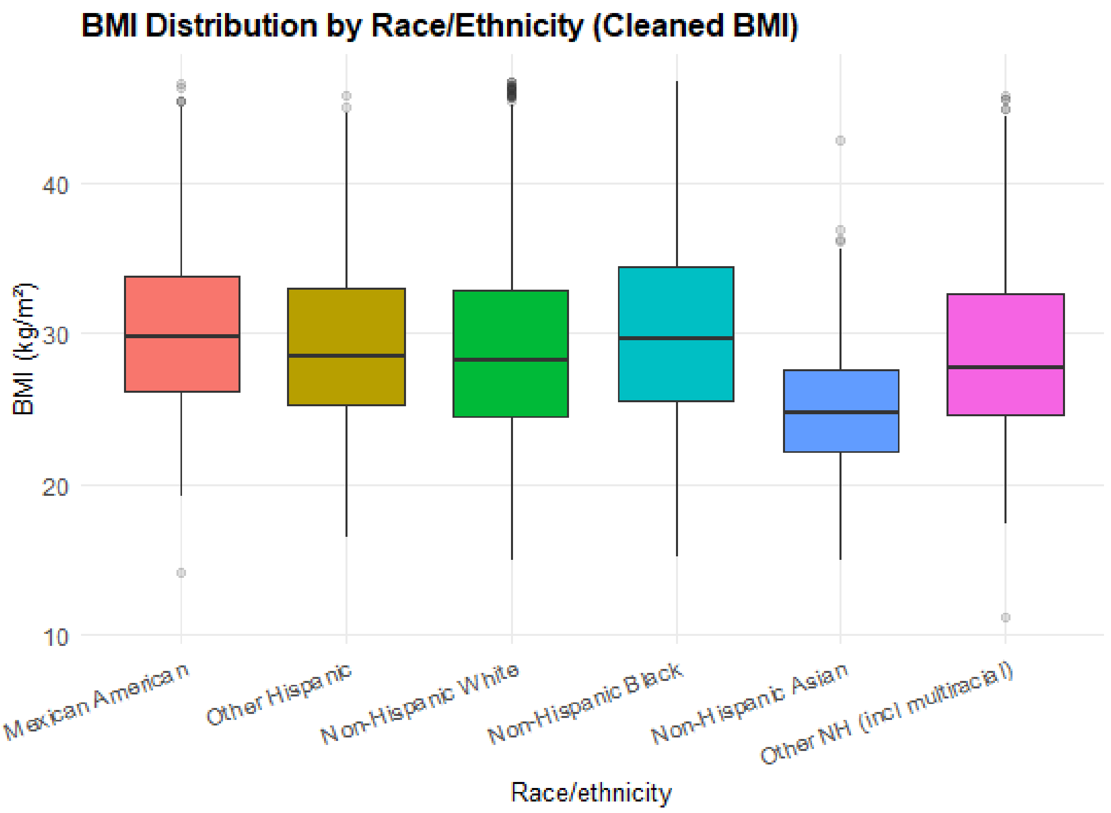
race	n	prop
Other NH (incl multiracial)	512	0.0655654
Non-Hispanic Asian	422	0.0540402

4.3 Boxplots: BMI by Education & by Race

```r
p_bmi_by_edu <- dat_clean2 %>%
  filter(!is.na(edu)) %>%
  ggplot(aes(x = edu, y = bmxbmi_clean, fill = edu)) +
  geom_boxplot(outlier.alpha = 0.15, width = 0.7) +
  labs(title = "BMI Distribution by Education (Cleaned BMI)",
       x = "Education level", y = "BMI (kg/m²)", fill="Education") +
  my_theme + theme(legend.position="none") +
  theme(axis.text.x = element_text(angle=20, hjust=1))
p_bmi_by_edu
```



```r
ggsave("outputs/w6_box_bmi_by_education.png", p_bmi_by_edu, width=7, height=4.5, bg="white")

p_bmi_by_race <- dat_clean2 %>%
  filter(!is.na(race)) %>%
  ggplot(aes(x = race, y = bmxbmi_clean, fill = race)) +
  geom_boxplot(outlier.alpha = 0.15, width = 0.7) +
  labs(title = "BMI Distribution by Race/Ethnicity (Cleaned BMI)",
       x = "Race/ethnicity", y = "BMI (kg/m²)", fill="Race") +
  my_theme + theme(legend.position="none") +
  theme(axis.text.x = element_text(angle=20, hjust=1))
p_bmi_by_race
```



```
ggsave("outputs/w6_box_bmi_by_race.png", p_bmi_by_race, width=7.5, heig
ht=4.5, bg="white")
```

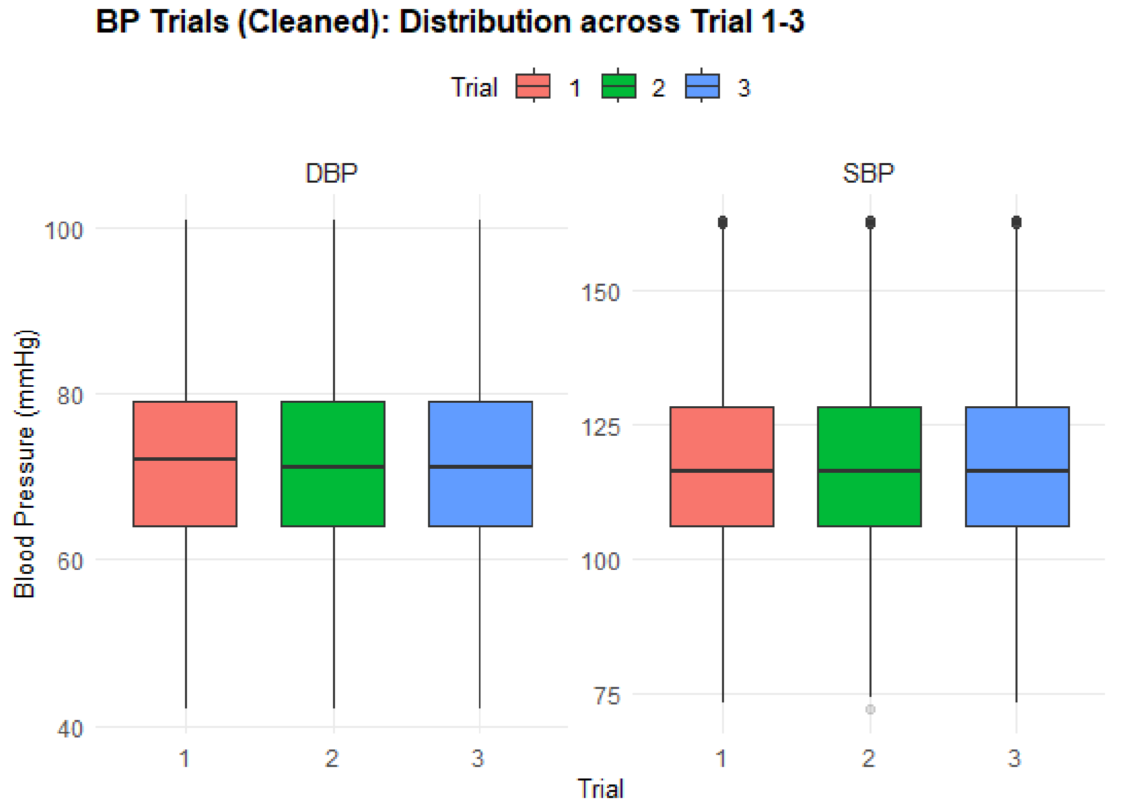
5. Week 6: BP Trials Reshape & Comparisons

5.1 Reshape wide → long (SBP & DBP trials)

```
bp_trials_long <- bpx %>%
  select(seqn, all_of(c(sbp_cols, dbp_cols))) %>%
  pivot_longer(cols = -seqn, names_to = "var", values_to = "value") %>%
  mutate(measure = if_else(str_detect(var, "sy"), "SBP", "DBP"),
         trial   = str_extract(var, "[1-3]") %>% as.integer()) %>%
  select(seqn, measure, trial, value)

bp_trials_long <- bp_trials_long %>%
  mutate(value_clean = case_when(
    measure == "SBP" & (value < SBP_LO | value > SBP_HI) ~ NA_real_,
    measure == "DBP" & (value < DBP_LO | value > DBP_HI) ~ NA_real_,
    TRUE ~ value
  )) %>%
  group_by(measure) %>%
  mutate(value_clean = {
    v <- value_clean
    q1 <- quantile(v, .25, na.rm=TRUE); q3 <- quantile(v, .75, na.rm=TR
UE)
    iqr <- q3 - q1; lo <- q1 - 1.5*iqr; hi <- q3 + 1.5*iqr
    med <- median(v, na.rm=TRUE); m <- mad(v, na.rm=TRUE)
    z <- ifelse(m>0, (v - med)/(m*1.4826), 0)
    ifelse(v < lo | v > hi | abs(z) > 3.5, NA_real_, v)
  }) %>%
  ungroup()
```

5.2 Boxplot: SBP & DBP across Trials (1-3)

```
p_trials_full <- bp_trials_long %>%
  filter(!is.na(value_clean)) %>%
  mutate(trial = factor(trial)) %>%
  ggplot(aes(x = trial, y = value_clean, fill = trial)) +
  geom_boxplot(outlier.alpha = 0.15, width=0.7) +
  facet_wrap(~ measure, scales="free_y") +
  labs(title = "BP Trials (Cleaned): Distribution across Trial 1-3",
       x = "Trial", y = "Blood Pressure (mmHg)", fill = "Trial") +
  my_theme + theme(legend.position="top")

p_trials_full
```

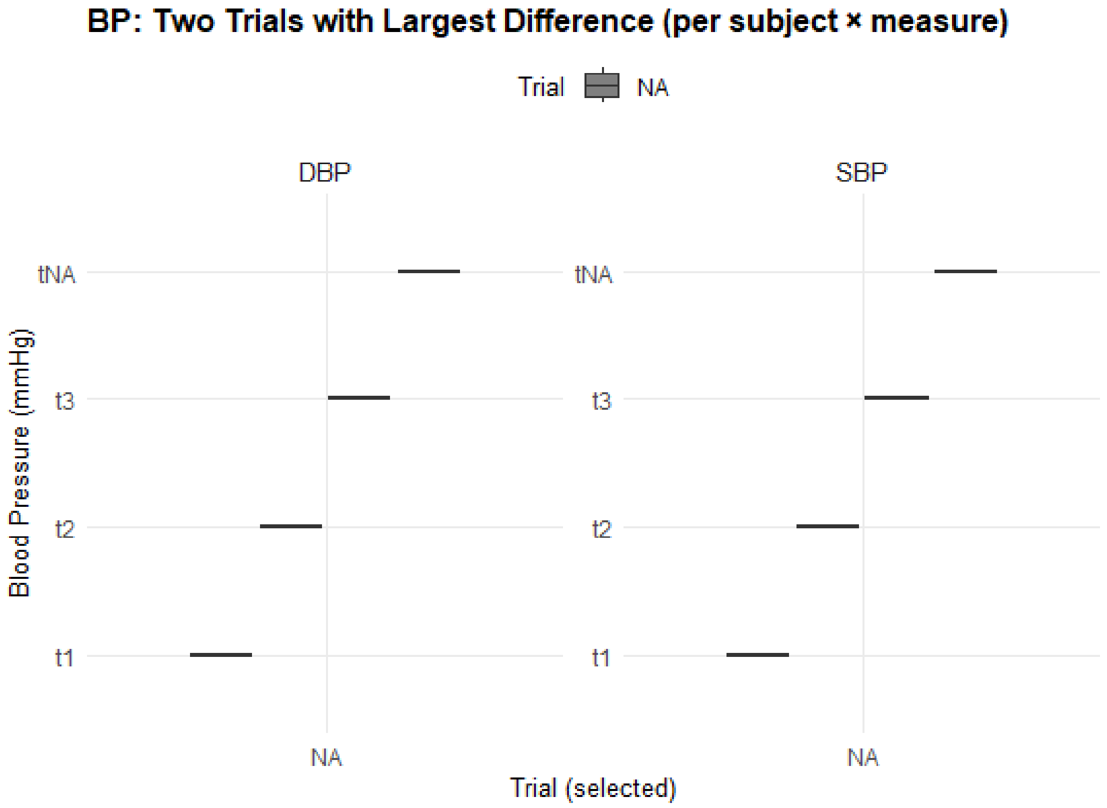


```
ggsave("outputs/w6_bp_trials_box_facet.png", p_trials_full, bg="white",
 width=7, height=4.5)
```

5.3 Two trials with the largest difference (per subject × measure)

```
bp_wide3 <- bp_trials_long %>%
  select(seqn, measure, trial, value_clean) %>%
  pivot_wider(names_from = trial, values_from = value_clean,
              names_prefix = "t")

largest_pair <- function(t1, t2, t3) {
  diffs <- c(`1-2` = abs(t1 - t2),
             `1-3` = abs(t1 - t3),
             `2-3` = abs(t2 - t3))
  if (all(is.na(diffs))) {
    return(NA_character_)
  }
  nm <- names(which.max(diffs))
  return(nm)
}

bp_pairs <- bp_wide3 %>%
  rowwise() %>%
  mutate(pair = largest_pair(t1, t2, t3)) %>%
  ungroup()
```

```
extract_pair_long <- function(df) {
  df %>% separate(pair, into = c("A","B"), sep = "-") %>%
    mutate(Acol = paste0("t",A), Bcol = paste0("t",B)) %>%
    pivot_longer(cols = c(Acol, Bcol), names_to = "trial_label",
                 values_to = "value_pair") %>%
    mutate(trial = as.integer(str_remove(trial_label, "t"))) %>%
    select(seqn, measure, trial, value_pair)
}

bp_two_trials <- extract_pair_long(bp_pairs) %>%
  filter(!is.na(value_pair))

p_trials_two <- bp_two_trials %>%
  mutate(trial = factor(trial)) %>%
  ggplot(aes(x = trial, y = value_pair, fill = trial)) +
  geom_boxplot(outlier.alpha=0.15, width=0.7) +
  facet_wrap(~ measure, scales="free_y") +
  labs(title="BP: Two Trials with Largest Difference (per subject × mea
sure)",
       x="Trial (selected)", y="Blood Pressure (mmHg)", fill="Trial") +
  my_theme + theme(legend.position="top")

p_trials_two
```



```r
ggsave("outputs/w6_bp_two_trial_largest_diff.png", p_trials_two, bg="white", width=7, height=4.5)
```

Inference: The relatively modest differences between the two selected trials suggest that the repeated BP measurements in NHANES were likely taken in the same session (rather than far apart in time).

---

6. Conclusion

- We cleaned BMI, SBP, and DBP values using physiologic bounds + IQR fences + robust MAD z-score rules, which removed extreme outliers while preserving most valid data.

- BMI distributions vary notably by education level and race/ethnicity: for example, groups with lower educational attainment or certain race/ethnicity categories show higher median BMI and wider spread.

- BP trial comparisons (three-trial full data, plus two-trial largest difference subset) show limited variation across trials, consistent with a same-day measurement protocol.

- Through this exercise, we reinforced a reproducible workflow: data loading → cleaning → exploration → visualization → reporting (in a single R Markdown document).

  —

- ***BONUS (+3): GITHUB Link to this report:***

  https://github.com/Research-Repo123/Big-Data-Analytic-HW-Week5-6

```
sessionInfo()

## R version 4.4.1 (2024-06-14 ucrt)
## Platform: x86_64-w64-mingw32/x64
## Running under: Windows 11 x64 (build 26200)
## 
## Matrix products: default
## 
## 
## locale:
## [1] LC_COLLATE=Chinese (Traditional)_Taiwan.utf8
## [2] LC_CTYPE=Chinese (Traditional)_Taiwan.utf8
## [3] LC_MONETARY=Chinese (Traditional)_Taiwan.utf8
## [4] LC_NUMERIC=C
## [5] LC_TIME=Chinese (Traditional)_Taiwan.utf8
## 
## time zone: Asia/Taipei
## tzcode source: internal
## 
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base

## 
## other attached packages:
##  [1] naniar_1.1.0    skimr_2.2.1     scales_1.3.0    janitor_2.2.1  
##  [5] haven_2.5.4     lubridate_1.9.4 forcats_1.0.0   stringr_1.5.1  
##  [9] dplyr_1.1.4     purrr_1.1.0     readr_2.1.5     tidyr_1.3.1    
## [13] tibble_3.2.1    ggplot2_3.5.2   tidyverse_2.0.0
## 
## loaded via a namespace (and not attached):
##  [1] utf8_1.2.4         generics_0.1.3     lattice_0.22-6     stringi_1.
8.4    
##  [5] hms_1.1.3          digest_0.6.37      magrittr_2.0.3     evaluate_
1.0.1   
##  [9] grid_4.4.1         timechange_0.3.0   fastmap_1.2.0      Matrix_1.
7-0     
## [13] jsonlite_1.8.9     mgcv_1.9-1         fansi_1.0.6        textshapi
ng_0.4.1
## [17] cli_3.6.3          crayon_1.5.3       rlang_1.1.4        bit64_4.5.
2
```

```
## [21] splines_4.4.1     munsell_0.5.1     base64enc_0.1-3   withr_3.0.
2
## [25] repr_1.1.7        yaml_2.3.10       parallel_4.4.1    tools_4.4.
1
## [29] tzdb_0.5.0        colorspace_2.1-1  vctrs_0.6.5       R6_2.5.1

## [33] lifecycle_1.0.4   snakecase_0.11.1  bit_4.5.0.1       vroom_1.6.
5
## [37] ragg_1.3.3        pkgconfig_2.0.3   pillar_1.9.0      gtable_0.
3.6
## [41] glue_1.8.0        visdat_0.6.0      systemfonts_1.1.0 xfun_0.49

## [45] tidyselect_1.2.1  rstudioapi_0.17.1 knitr_1.49        farver_2.
1.2
## [49] nlme_3.1-164      htmltools_0.5.8.1 rmarkdown_2.29    labeling_
0.4.3
## [53] compiler_4.4.1
```