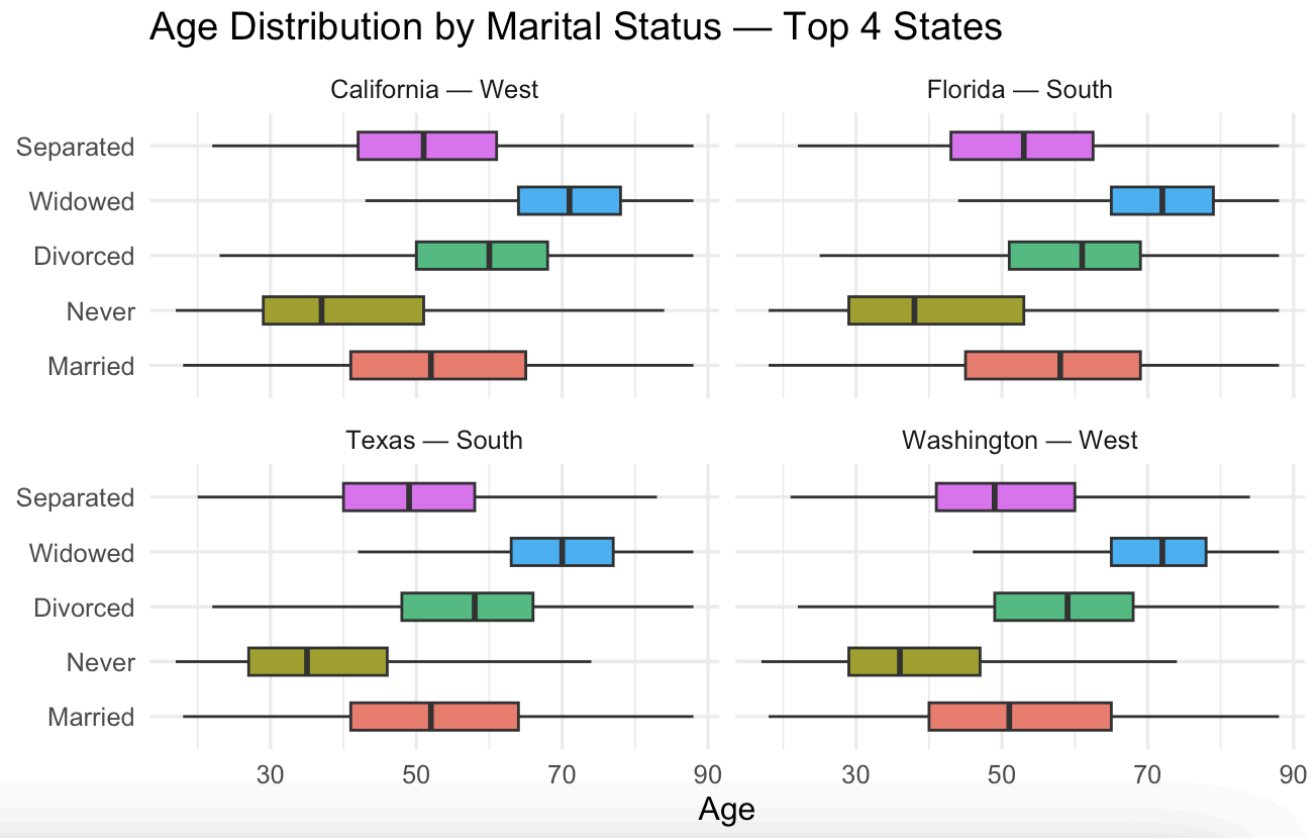
9/24/2025
B2000
Homework 4

Group members: Aqeel Choudhury, Keyla Pereira, Marwan kenawy, Michael Stewart

Keyla: Age Distribution by Marital Status

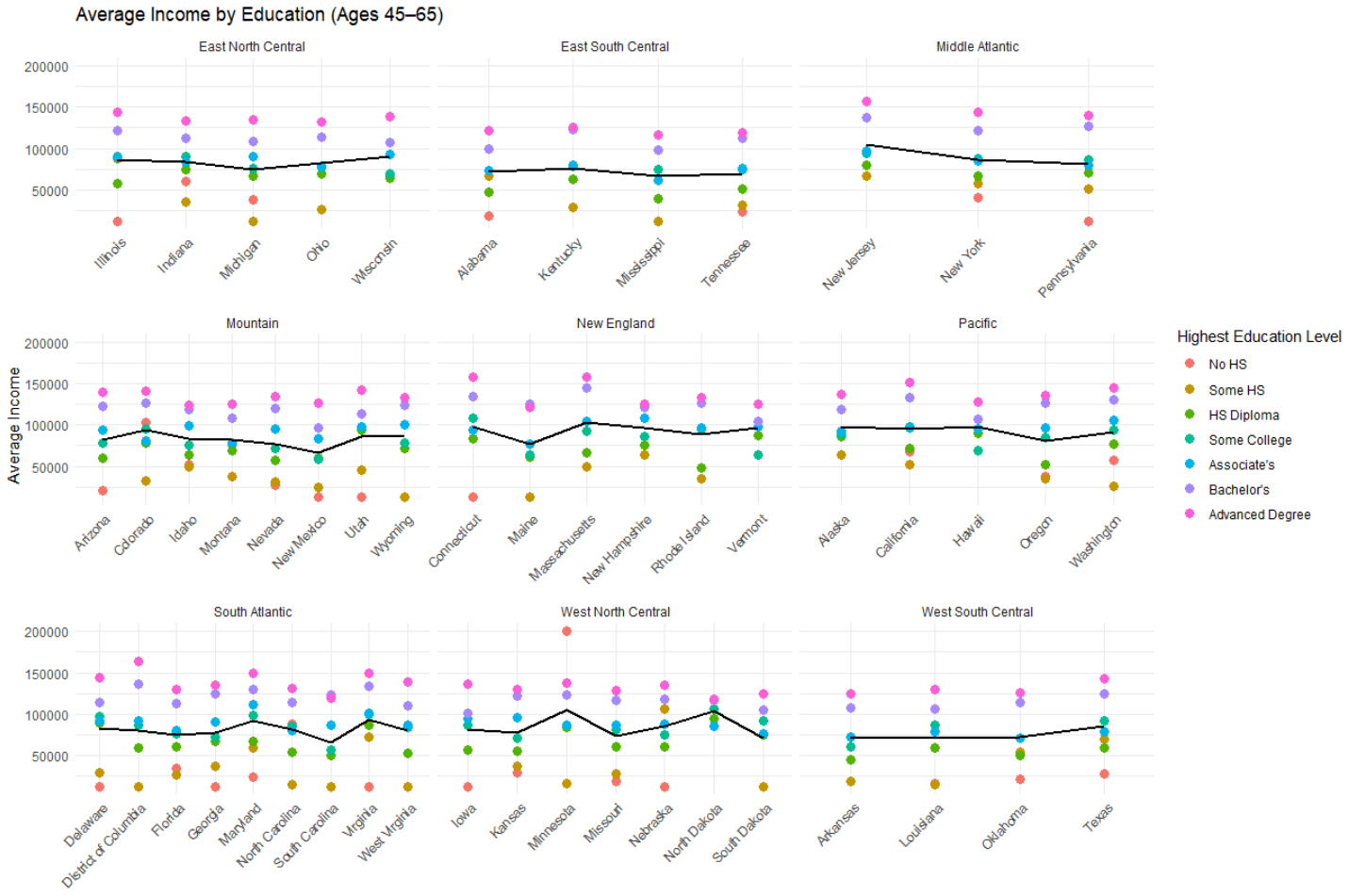


Across California, Texas, Florida, and Washington, age rises in a stable sequence from never married to married to divorced or separated to widowed; Florida skews older overall, Texas skews younger, and California and Washington fall between, each box shows the middle 50 percent of ages with the central line as the median. The gaps between marital-status groups are larger than the gaps between states, so marital status is the stronger driver of age patterns in this sample.

```
p_box <- ggplot(plot_data, aes(x = Mar_Stat, y = Age, fill = Mar_Stat)) +
geom_boxplot(width = 0.5, outlier.shape = 16, outlier.alpha = 0.35) +
facet_wrap(~ StateLab, nrow = 2) +
coord_flip() +
labs(
  title    = "Age Distribution by Marital Status — Top 4 States",
  x = NULL, y = "Age"
) +
```

```
  theme_minimal(base_size = 12) +
  theme(legend.position = "none")
p_box
```

Michael: average income by education level for those aged 45-65



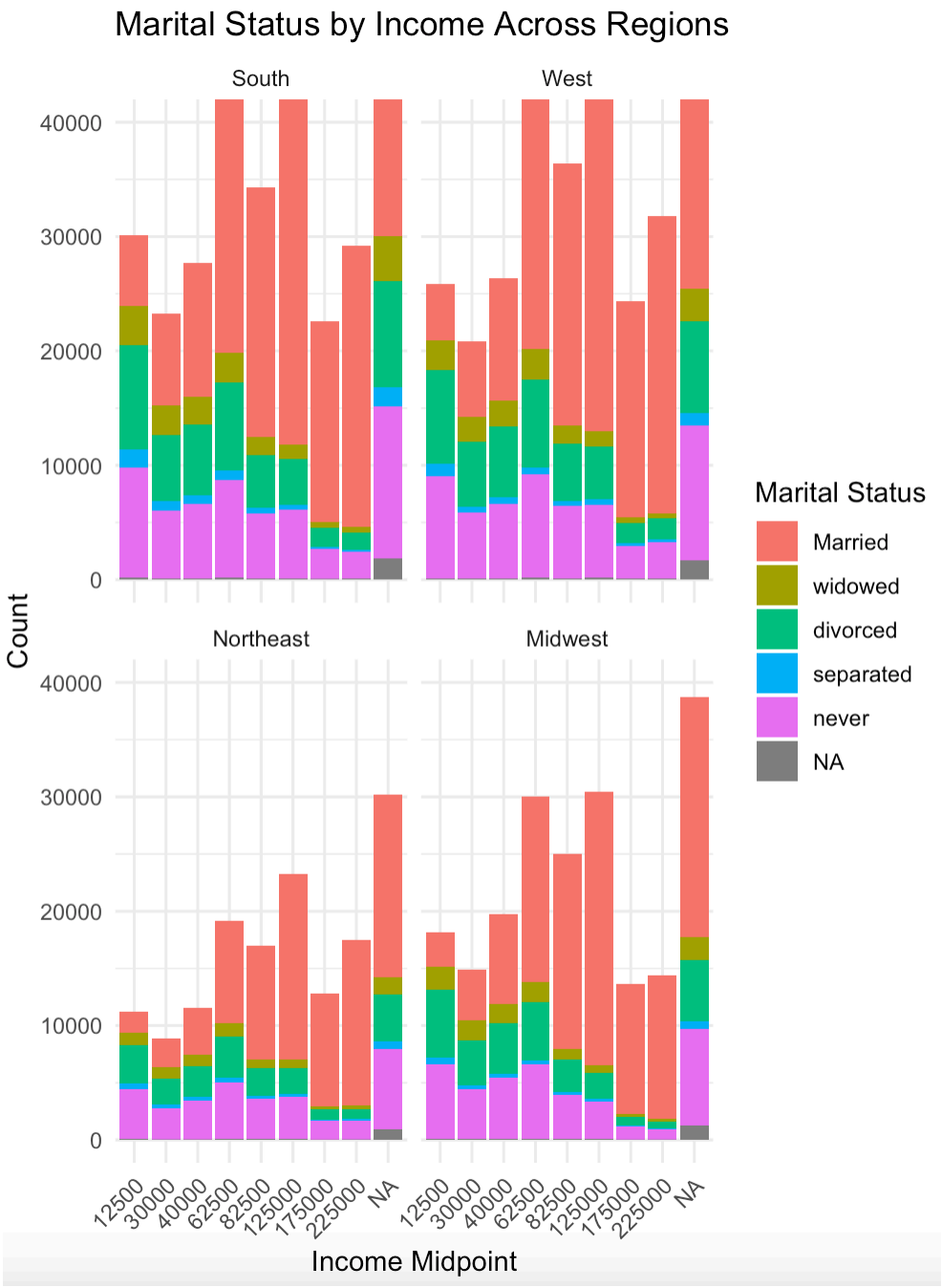
Key Observations:

- In some states, the difference in income between a bachelor's and an advanced degree is minimal.
- In Minnesota, it appears the wealthiest are those with no high school! (Perhaps an error in the data, or something specific to how Minnesota codes it)
- Across regions, those with "Some College" or higher education levels have higher than average incomes (average denoted by the line graph).

```
# Load ggplot2
install.packages("ggplot2")
library(ggplot2)

# Filter and prepare data
df_summary <- Household_Pulse_data

# Map income and compute age
df_summary$INCOME_NUM <- income_map[as.character(df_summary$INCOME)]
df_summary$Age <- 2025 - df_summary$TBIRTH_YEAR
```

```r
# Keep only age 45–65 and non-missing income
df_summary <- df_summary[df_summary$Age >= 45 & df_summary$Age <= 65 &
!is.na(df_summary$INCOME_NUM), ]

# Compute averages
# Mean income by state and education
df_summary_agg <- aggregate(INCOME_NUM ~ REGION + EST_ST + EEDUC, data = df_summary, FUN =
mean)

# Mean income by state (all education levels combined)
state_avg <- aggregate(INCOME_NUM ~ REGION + EST_ST, data = df_summary_agg, FUN = mean)
names(state_avg)[names(state_avg) == "INCOME_NUM"] <- "overall_mean"

# Rename education levels
df_summary_agg$EEDUC <- recode(df_summary_agg$EEDUC,
                               "less than hs" = "No HS",
                               "some hs" = "Some HS",
                               "HS diploma" = "HS Diploma",
                               "some coll" = "Some College",
                               "assoc deg" = "Associate's",
                               "bach deg" = "Bachelor's",
                               "adv deg" = "Advanced Degree")

# Assign divisions manually
assign_division <- function(states) {
  div <- rep(NA, length(states))
  div[states %in% c("Connecticut", "Maine", "Massachusetts", "New Hampshire",
                    "Rhode Island", "Vermont")] <- "New England"
  div[states %in% c("New Jersey", "New York", "Pennsylvania")] <- "Middle Atlantic"
  div[states %in% c("Ohio", "Indiana", "Illinois", "Michigan", "Wisconsin")] <- "East North Central"
  div[states %in% c("Minnesota", "Iowa", "Missouri", "North Dakota", "South Dakota",
                    "Nebraska", "Kansas")] <- "West North Central"
  div[states %in% c("Delaware", "District of Columbia", "Florida", "Georgia", "Maryland",
                    "North Carolina", "South Carolina", "Virginia", "West Virginia")] <- "South Atlantic"
  div[states %in% c("Alabama", "Kentucky", "Mississippi", "Tennessee")] <- "East South Central"
  div[states %in% c("Arkansas", "Louisiana", "Oklahoma", "Texas")] <- "West South Central"
  div[states %in% c("Montana", "Idaho", "Wyoming", "Colorado", "New Mexico", "Arizona",
                    "Utah", "Nevada")] <- "Mountain"
  div[states %in% c("Washington", "Oregon", "California", "Alaska", "Hawaii")] <- "Pacific"
  return(div)
}

df_summary_agg$DIVISION <- assign_division(df_summary_agg$EST_ST)
state_avg$DIVISION <- assign_division(state_avg$EST_ST)

# Plot
ggplot(df_summary_agg, aes(x = EST_ST, y = INCOME_NUM, color = EEDUC, group = EEDUC)) +
  geom_point(size = 3) +
  geom_line(data = state_avg, aes(x = EST_ST, y = overall_mean, group = 1),
            color = "black", size = 1) +
  facet_wrap(~ DIVISION, scales = "free_x") +
  labs(title = "Average Income by Education (Ages 45–65)",
       x = "State",
       y = "Average Income",
       color = "Highest Education Level") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Marwan: Marital Status by Income Across Regions

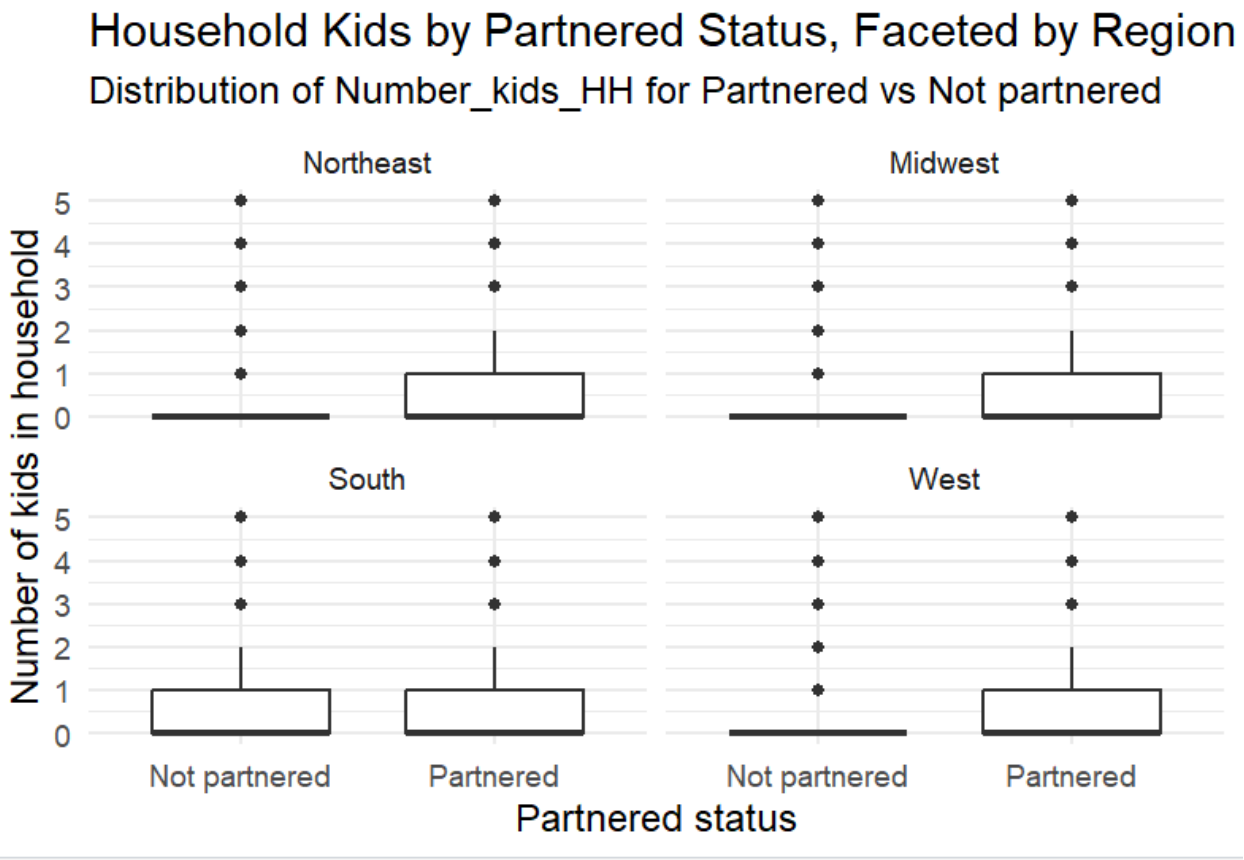



```
> ggplot(d_HHP2020_24, aes(x = income_midpoint_factor, fill = Mar_Stat)) +
+     geom_bar(position = "stack") +
+     facet_wrap(~ Region, ncol = 2) +
+     labs(
+         title = "Marital Status by Income Across Regions",
+         x = "Income Midpoint",
+         y = "Count",
+         fill = "Marital Status"
+     ) +
```

```
+     theme_minimal() +
+     theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
+     coord_cartesian(ylim = c(0, 40000))   # adjust this number to zoom in more/less
```

Married people are most common at higher income levels, while those who never married are more common at lower incomes. Divorced and widowed groups appear across all incomes but less at the top. The same pattern shows up in every region.

Aqeel: Distribution of Kids per Household across partnered status



In all regions, the partnered group has a higher median number of kids in the household compared to the non-partnered group except for the southern region where number of kids per household between partnered and non partnered variables are more similarly distributed.

```
library(tidyverse)
library(dplyr)
library(ggplot2)

d_HHP2020_24$partnered <- (d_HHP2020_24$Mar_Stat == "Married") |
  (d_HHP2020_24$Mar_Stat == "widowed") |
  (d_HHP2020_24$Mar_Stat == "divorced") |
  (d_HHP2020_24$Mar_Stat == "separated")


plot_df <- d_HHP2020_24 %>%
```

```
  mutate(
    partnered_lbl = if_else(partnered, "Partnered", "Not partnered"),
    Region = factor(Region, levels = c("Northeast", "Midwest", "South", "West"))
  ) %>%
  filter(!is.na(Number_kids_HH), !is.na(Region), !is.na(partnered))

kids_summary <- plot_df %>%
  group_by(Region, partnered_lbl) %>%
  summarise(
    n = n(),
    mean_kids = mean(Number_kids_HH, na.rm = TRUE),
    median_kids = median(Number_kids_HH, na.rm = TRUE),
    sd_kids = sd(Number_kids_HH, na.rm = TRUE),
    .groups = "drop"
  )

print(kids_summary)

p_box <- ggplot(plot_df, aes(x = partnered_lbl, y = Number_kids_HH)) +
  geom_boxplot(outlier.alpha = 0.15, width = 0.7) +
  facet_wrap(~ Region, nrow = 2) +
  labs(
    title = "Household Kids by Partnered Status, Faceted by Region",
    subtitle = "Distribution of Number_kids_HH for Partnered vs Not partnered",
    x = "Partnered status",
    y = "Number of kids in household"
  ) +
  theme_minimal(base_size = 12)

print(p_box)
```