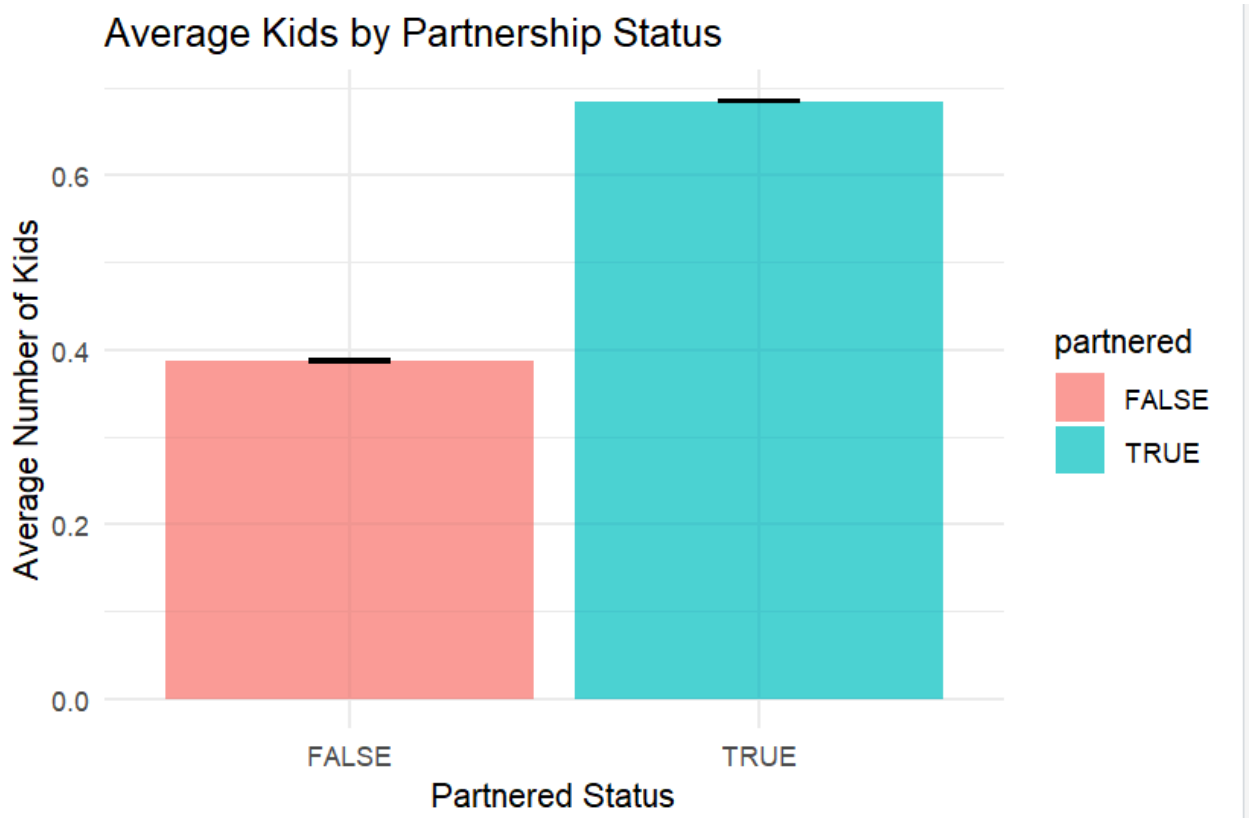
Aqeel Choudhury
9/15/25
B2000
Homework #3
Prompt: What are some possible variables in the household pulse data that have strong correlation

Hypothesis: Strong correlating variables may include Income, the number of kids per household, and education level

Analysis:

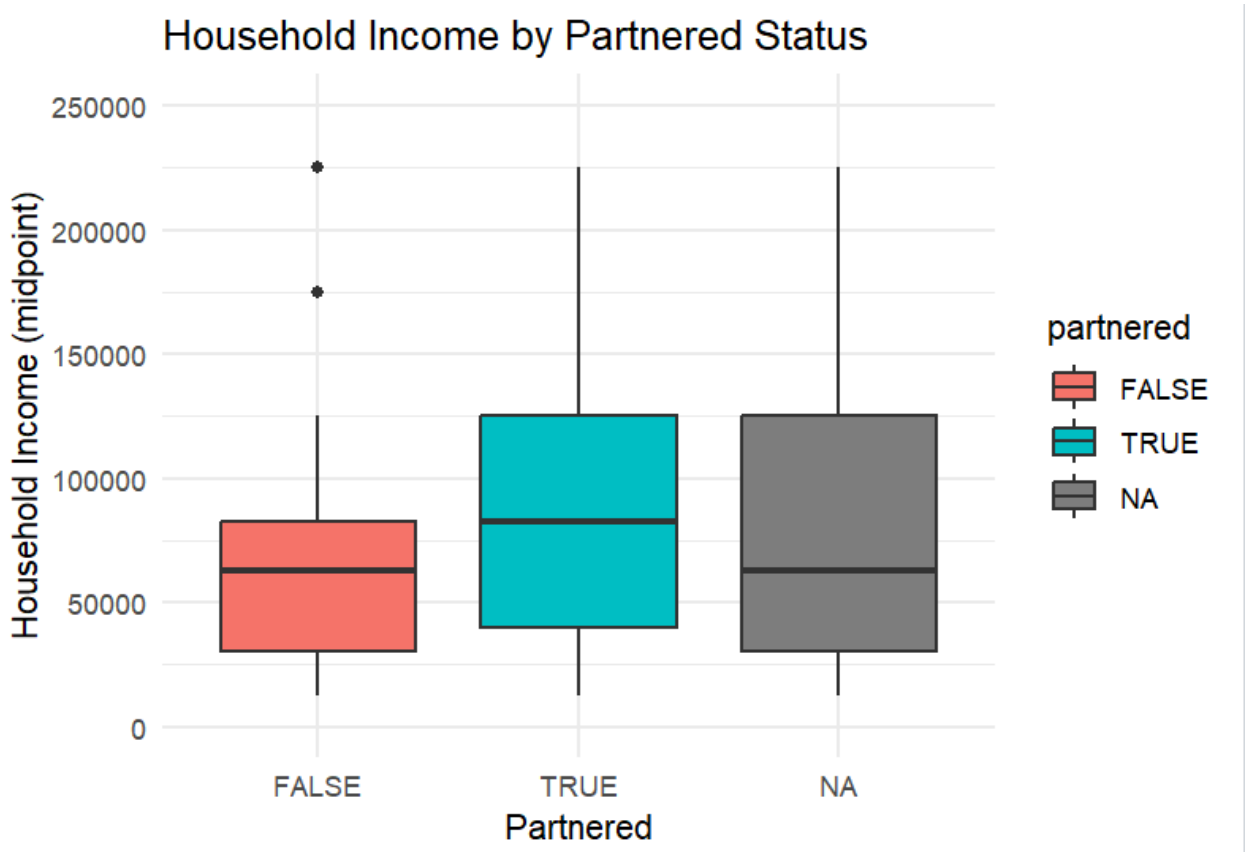
```
summary(d_HHP2020_24)
d_HHP2020_24$partnered <- (d_HHP2020_24$Mar_Stat == "Married") |
  (d_HHP2020_24$Mar_Stat == "widowed") |
  (d_HHP2020_24$Mar_Stat == "divorced") |
  (d_HHP2020_24$Mar_Stat == "separated")

library(tidyverse)
d_partnered <- d_HHP2020_24 %>%
  mutate(
    partnered = Mar_Stat %in% c("Married", "widowed", "divorced", "separated"),
    partnered_num = as.integer(partnered)  # 1 if TRUE, 0 if FALSE
  )
corr_out <- d_partnered %>%
  summarise(
    correlation = cor(partnered_num, Number_kids_HH, use = "complete.obs")
  )
d_partnered %>%
  group_by(partnered) %>%
  summarise(
    mean_kids = mean(Number_kids_HH, na.rm = TRUE),
    se = sd(Number_kids_HH, na.rm = TRUE) / sqrt(n()),
    .groups = "drop"
  ) %>%
  ggplot(aes(x = partnered, y = mean_kids, fill = partnered)) +
  geom_col(alpha = 0.7) +
  geom_errorbar(aes(ymin = mean_kids - se, ymax = mean_kids + se), width = 0.2) +
  labs(
    x = "Partnered Status",
    y = "Average Number of Kids",
    title = "Average Kids by Partnership Status"
  ) +
  theme_minimal()
```

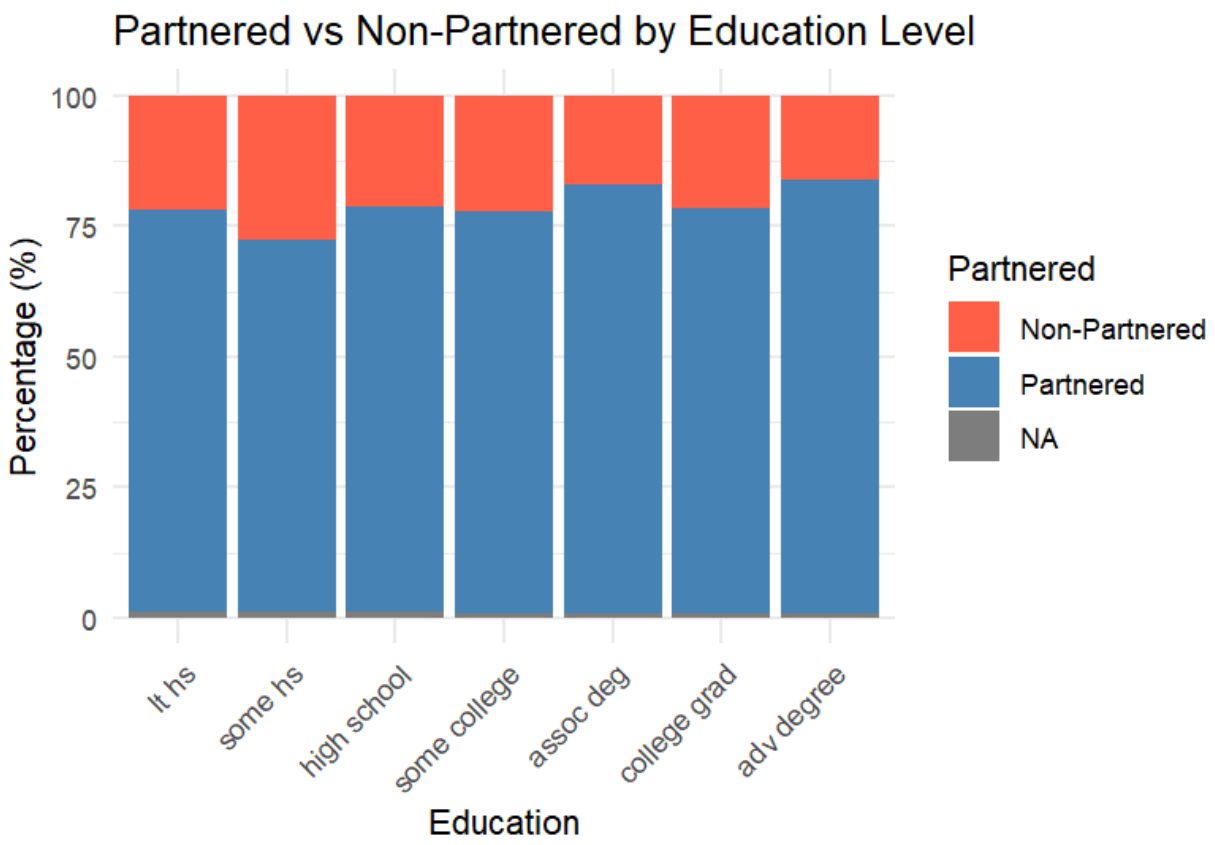



```
d_HHP2020_24 %>%
  group_by(partnered) %>%
  summarise(mean_income = mean(income_midpoint, na.rm = TRUE),
            median_income = median(income_midpoint, na.rm = TRUE),
            count = n())

ggplot(d_HHP2020_24, aes(x = partnered, y = income_midpoint, fill = partnered)) +
  geom_boxplot() +
  scale_y_continuous(limits = c(0, 250000)) +
  labs(title = "Household Income by Partnered Status",
       x = "Partnered",
       y = "Household Income (midpoint)") +
  theme_minimal()
```




```
edu_partnered <- d_HHP2020_24 %>%
  group_by(Education, partnered) %>%
  summarise(count = n(), .groups = "drop") %>%
  group_by(Education) %>%
  mutate(percent = count / sum(count) * 100)
ggplot(edu_partnered, aes(x = Education, y = percent, fill = partnered)) +
  geom_bar(stat = "identity", position = "stack") +
  labs(title = "Partnered vs Non-Partnered by Education Level",
       y = "Percentage (%)",
       fill = "Partnered") +
  scale_fill_manual(values = c("FALSE" = "tomato", "TRUE" = "steelblue"),
                    labels = c("Non-Partnered", "Partnered")) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Results: After further analysis we've discovered that partnered households tend to have more children per household than non partnered. Partnered households tend to have higher incomes than non partnered households on average. Partnered households also tend to have higher education levels than non partnered as a percentage with advanced degrees being higher as well (shocking)