Midterm 1

October 3rd, 2024

```r
library(ggplot2)
library(dplyr)
library(gtsummary)
library(openintro)
require(tidyverse)
library(readr)
labor <- read_csv("labor.csv")


summary(labor)
```

```
##  labor_force         kids_under6        kids6_18          age       
##  Length:753         Min.   :0.0000   Min.   :0.000   Min.   :30.00  
##  Class :character   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:36.00  
##  Mode  :character   Median :0.0000   Median :1.000   Median :43.00  
##                     Mean   :0.2337   Mean   :1.353   Mean   :42.54  
##                     3rd Qu.:0.0000   3rd Qu.:2.000   3rd Qu.:49.00  
##                     Max.   :2.0000   Max.   :8.000   Max.   :60.00  
##  wife_college       husband_college    family_income   
##  Length:753         Length:753         Min.   :-0.029  
##  Class :character   Class :character   1st Qu.:13.025  
##  Mode  :character   Mode  :character   Median :17.700  
##                                        Mean   :20.129  
##                                        3rd Qu.:24.466  
##                                        Max.   :96.000  
```

1.

Sampling Method Proposal: 1. Target Population: Women aged 18-65 in the U.S. 2. Sampling Frame: Use U.S. Census data to list eligible women, ensuring diversity in race, ethnicity, socio-economic status, education, and location (urban/rural). 3. Sampling Method: Stratified Random Sampling: Divide the population into groups (age, race, location). Use random sampling within each group. 4. Sample Size: Calculate a statistically significant sample size considering confidence level, margin of error, and population variability. 5. Data Collection: Use trained professionals to conduct surveys or interviews for consistent data. 6. Minimize Bias: Implement methods to reduce non-response and other biases, ensuring inclusivity.

2.

Population for Generalization: The study results aim to represent all U.S. women aged 18-65, including those with diverse backgrounds in race, ethnicity, socio-economic status, education, and both urban and rural settings.

Potential Biases and Concerns: Non-response Bias: Certain groups, like lower socio-economic women, may be less likely to respond, skewing results. Sampling Frame Bias: If Census data doesn't reflect the true population, some subgroups might be over- or underrepresented. Selection Bias: Flaws in participant selection could lead to a non-random sample. Language and Cultural Bias: Surveys only in English may exclude non-English speakers or culturally diverse groups. Undercoverage: Marginalized populations might be missing from Census data. Mitigating strategies include follow-ups and inclusive data collection methods.

3.

Proposed Variable: "Access to Childcare Services" Explanation: Access to affordable and reliable childcare can greatly impact a woman's ability to participate in the workforce. Women with childcare options are more likely to work since they can ensure their children are cared for during work hours. In contrast, limited or expensive childcare can discourage women from working or push them into lower-paying jobs that allow for more flexible parenting. This variable highlights a key factor influencing work-life balance and social support systems that affect women's labor force participation rates.

4. Explanation for Variable Treatment:

The variable "kids_under6" should be treated as a factor variable. my reasons: Categorical Nature: The values (0, 1, 2) represent distinct categories indicating childcare responsibilities rather than numbers for mathematical operations. Interpretability: Treating "kids_under6" as a factor improves clarity in statistical models by distinguishing between having no young children, one, or two. Predictive Modeling: Using it as a factor allows models (e.g., logistic regression) to compare each category against a baseline, enhancing the understanding of the impact of each level.

```
labor <- read.csv('labor.csv')
labor <- labor %>%
  mutate("Wife College" = as.factor(wife_college),
       "Husband College" = as.factor(husband_college))
summary(labor)
```

```
##  labor_force         kids_under6        kids6_18          age       
##  Length:753         Min.   :0.0000   Min.   :0.000   Min.   :30.00  
##  Class :character   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:36.00  
##  Mode  :character   Median :0.0000   Median :1.000   Median :43.00  
##                     Mean   :0.2337   Mean   :1.353   Mean   :42.54  
##                     3rd Qu.:0.0000   3rd Qu.:2.000   3rd Qu.:49.00  
##                     Max.   :2.0000   Max.   :8.000   Max.   :60.00  
##  wife_college       husband_college    family_income    Wife College
##  Length:753         Length:753         Min.   :-0.029   No :541     
##  Class :character   Class :character   1st Qu.:13.025   Yes:212     
##  Mode  :character   Mode  :character   Median :17.700               
##                                        Mean   :20.129               
##                                        3rd Qu.:24.466               
##                                        Max.   :96.000               
##  Husband College
##  No :295        
##  Yes:458        
##                 
##                 
##                 
##                 
```

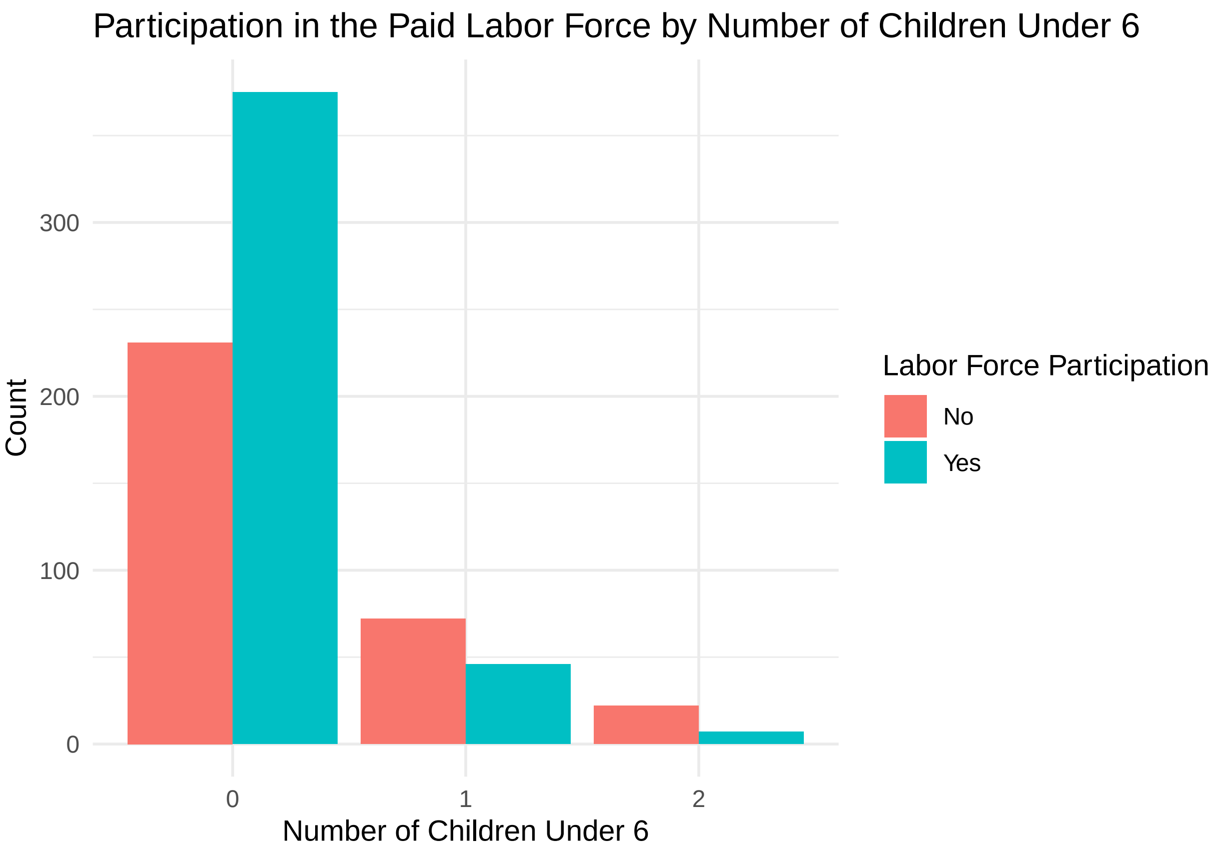
5.

```
# data("labor")
# summary_table <-
#   labor %>%
#   select(age,
#          kids_under6,
#          wife_college,
#          husband_college,
#          family_income) %>%
#   tbl_summary(
#     statistic = list(all_continuous() ~ "{mean} ({sd})",
#                      all_categorical() ~ "{n} ({p}%)"))
#
# summary_table
```

6.

```
labor <- labor %>%
  mutate(
    labor_force = factor(labor_force, levels = c("No", "Yes")),
    kids_under6 = as.factor(kids_under6)
  )

ggplot(labor, aes(x = kids_under6, fill = labor_force)) +
  geom_bar(position = "dodge") +
  labs(
    title = "Participation in the Paid Labor Force by Number of Children Under 6",
    x = "Number of Children Under 6",
    y = "Count",
    fill = "Labor Force Participation"
  ) +
  theme_minimal()
```

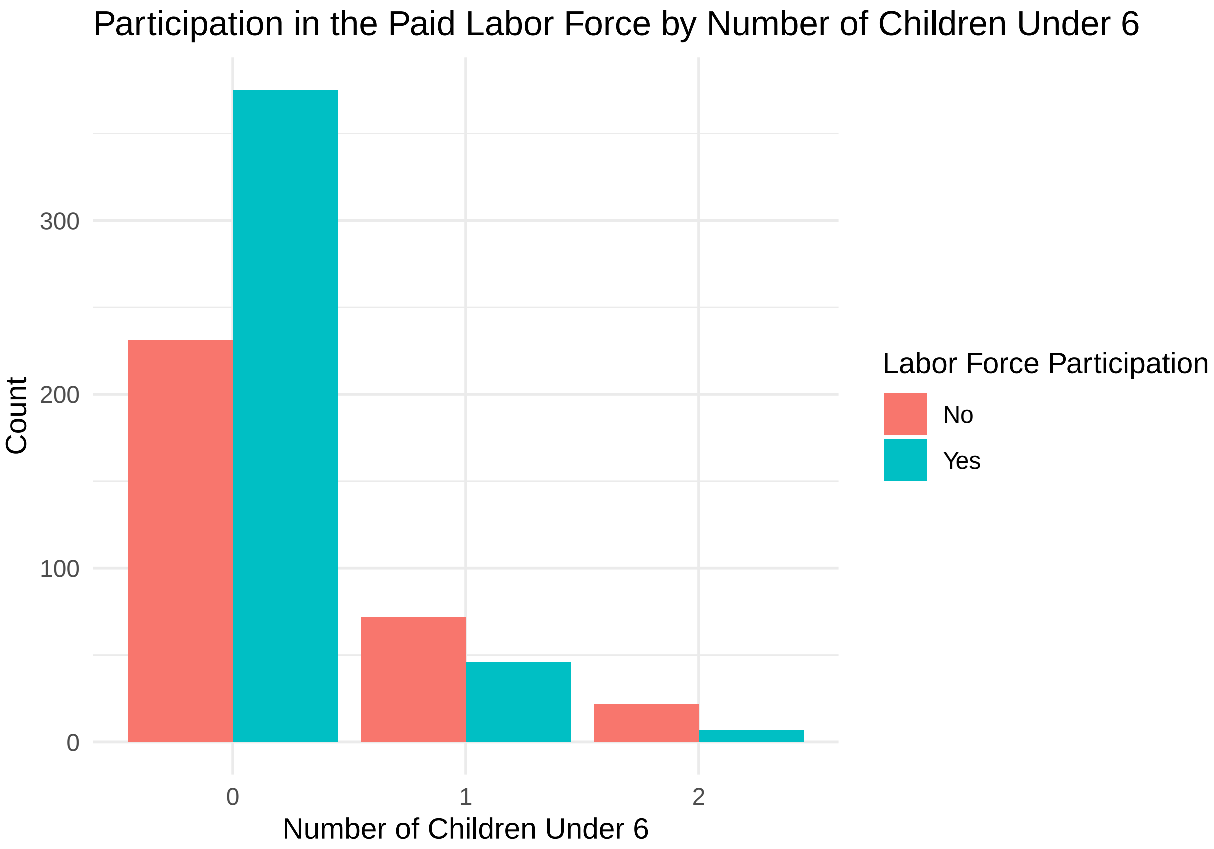


7.

```
kids_under6_summary <- labor %>%
  group_by(labor_force, kids_under6) %>%
  summarise(count = n()) %>%
  mutate(percent = round((count / sum(count)) * 100, 1))
print(kids_under6_summary)
```

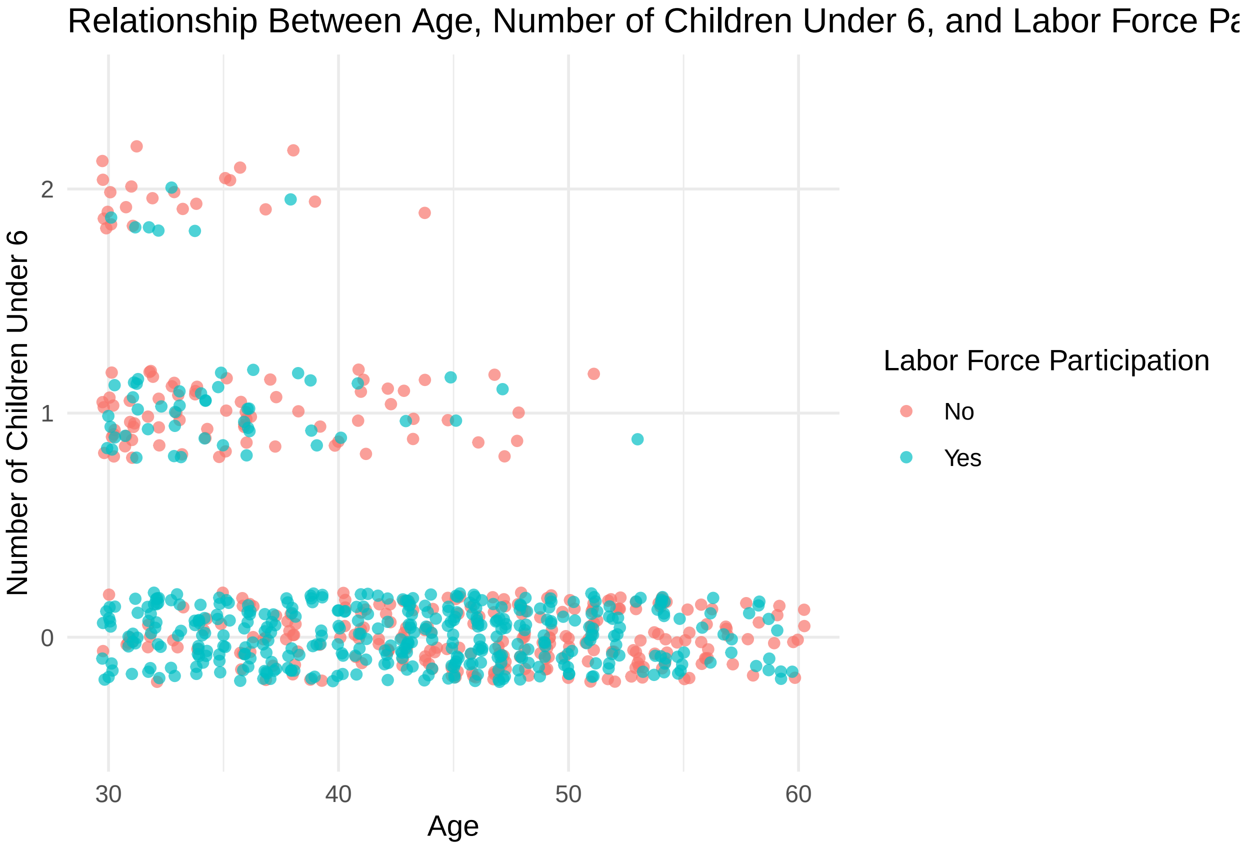
```
## # A tibble: 6 x 4
## # Groups:   labor_force [2]
##   labor_force kids_under6 count percent
##   <fct>       <fct>       <int>   <dbl>
## 1 No          0             231    71.1
## 2 No          1              72    22.2
## 3 No          2              22     6.8
## 4 Yes         0             375    87.6
## 5 Yes         1              46    10.7
## 6 Yes         2               7     1.6
```

```
ggplot(labor, aes(x = kids_under6, fill = labor_force)) +
  geom_bar(position = "dodge") +
  labs(
    title = "Participation in the Paid Labor Force by Number of Children Under 6",
    x = "Number of Children Under 6",
    y = "Count",
    fill = "Labor Force Participation"
  ) +
  theme_minimal()
```



8.

```
ggplot(labor, aes(x = age, y = kids_under6, color = labor_force)) +
  geom_jitter(width = 0.3, height = 0.2, alpha = 0.7) +
  labs(
    title = "Relationship Between Age, Number of Children Under 6, and Labor Force Participation",
    x = "Age",
    y = "Number of Children Under 6",
    color = "Labor Force Participation"
  ) +
  theme_minimal()
```



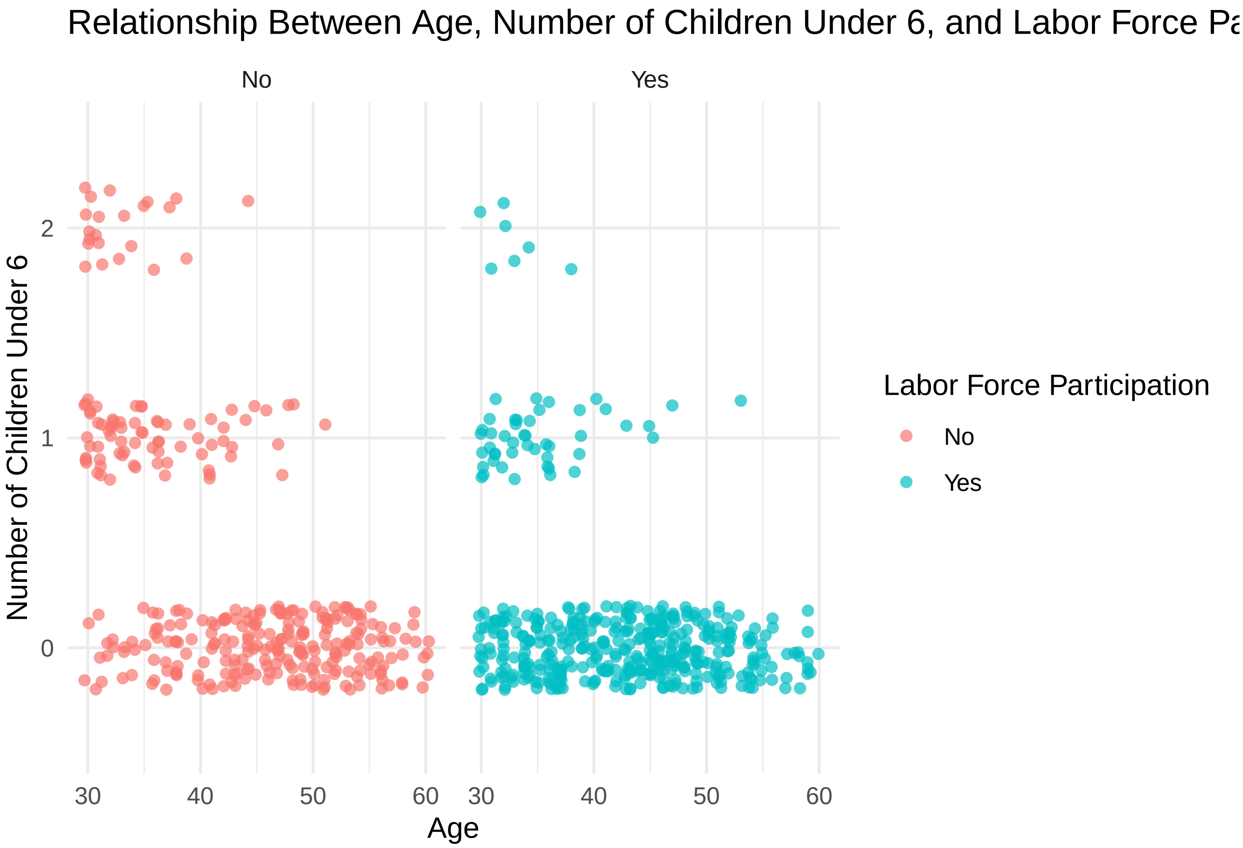
9.

```
age_kids_summary <- labor %>%
  group_by(labor_force, kids_under6) %>%
  summarise(
    avg_age = mean(age, na.rm = TRUE),
    count = n()
  )

print(age_kids_summary)
```

```
## # A tibble: 6 x 4
## # Groups:   labor_force [2]
##   labor_force kids_under6 avg_age count
##   <fct>       <fct>         <dbl> <int>
## 1 No          0              46.5   231
## 2 No          1              36.0    72
## 3 No          2              33.2    22
## 4 Yes         0              43.0   375
## 5 Yes         1              35.2    46
## 6 Yes            2                 32.9      7
```

```
ggplot(labor, aes(x = age, y = kids_under6, color = labor_force)) +
  geom_jitter(width = 0.3, height = 0.2, alpha = 0.7) +
  labs(
    title = "Relationship Between Age, Number of Children Under 6, and Labor Force Participation",
    x = "Age",
    y = "Number of Children Under 6",
    color = "Labor Force Participation"
  ) +
  theme_minimal() +
  facet_wrap(~labor_force)
```



my comment: This scatter plot visualizes the relationship between age, number of children under 6, and labor force participation. It shows two categories: those not in the labor force (pink) and those who are (blue). The plot suggests that younger women with more children are less likely to participate in the labor force.

10. (a)

```
n <- nrow(labor)
p_hat <- mean(labor$labor_force == "Yes")
q_hat <- 1 - p_hat
np <- n * p_hat
nq <- n * q_hat

print(paste("np =", np))
```

```
## [1] "np = 428"
```

```
print(paste("n(1-p) =", nq))
```

```
## [1] "n(1-p) = 325"
```

10. (b)

```
stderr <- sqrt(p_hat * q_hat / n)
z_value <- qnorm(0.975)
margin_of_error <- z_value * stderr

confidence_interval <- c(
  p_hat - margin_of_error,
  p_hat + margin_of_error
)

print(paste("95% Confidence Interval for the proportion of women in the workforce:",
            round(confidence_interval[1], 4), "to", round(confidence_interval[2], 4)))
```

```
## [1] "95% Confidence Interval for the proportion of women in the workforce: 0.533 to 0.6038"
```

11.

```
true_proportion_1970 <- 0.40

within_interval <- true_proportion_1970 >= confidence_interval[1] && true_proportion_1970 <= confidence_

print(paste("Is the 1970 proportion within the confidence interval?", within_interval))
```

```
## [1] "Is the 1970 proportion within the confidence interval? FALSE"
```

Reflection Questions 1. I feel generate a table including mean (sd) or n (%) is difficult for me. 2. E