Homework#3 - Possible variables that has a strong correlation with partnership status

2025-09-14

GROUP MEMBERS: Riyesh Nath, Bamba Cisse, Nasrin Khanam and Marwan Kenawy

---

Summary:

In this project, we will do a data exploration of household pulse data and try to find variables that can allows us to find a strong correlation with a person's partnership status.

The variables that we will test are:

- Effect of education on partnership status.
- Effect of Race on partnership status
- Effect of Age on partnership status

```
library(ggplot2)
library(tibble)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(gridExtra)
```

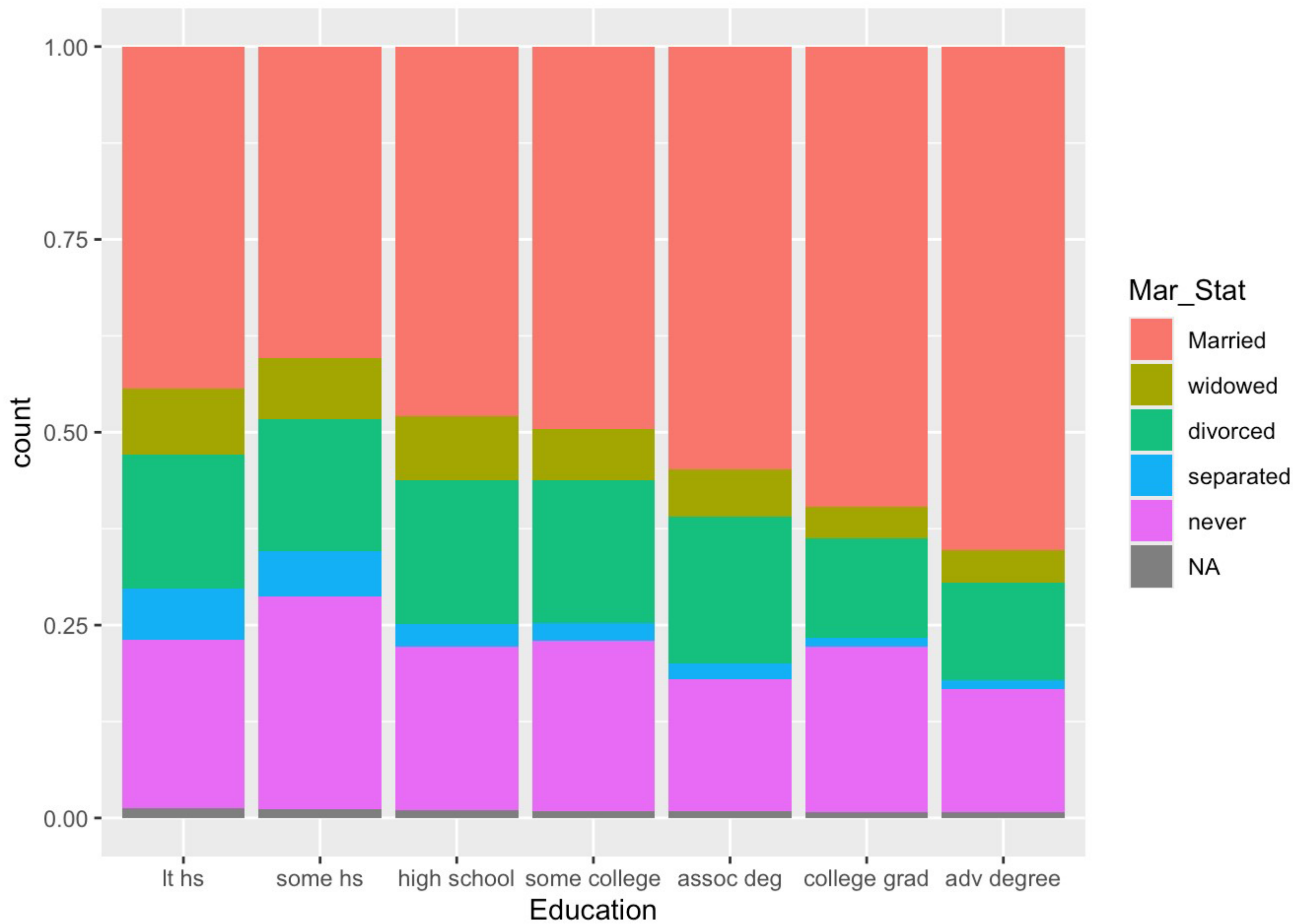
```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

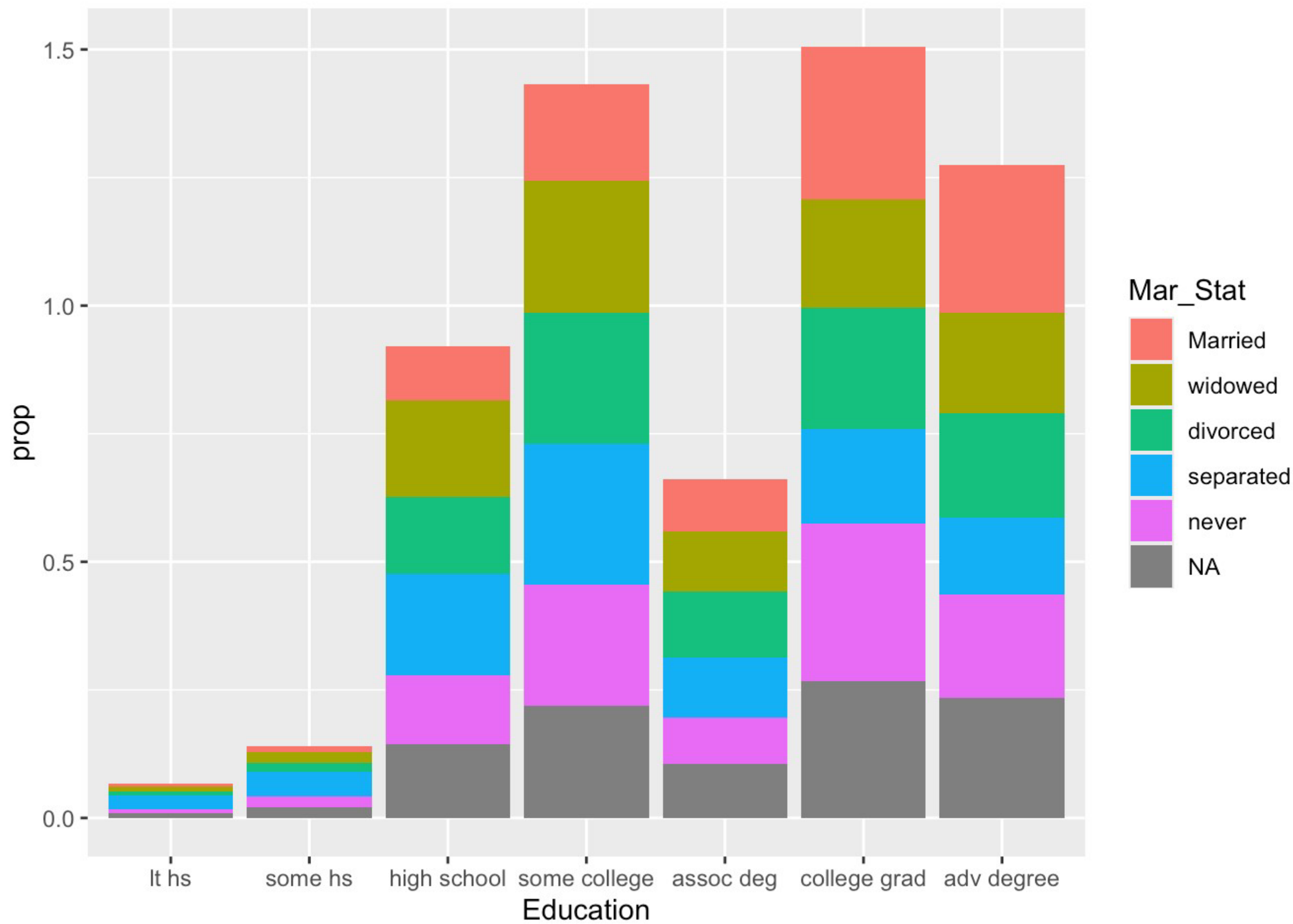
```
load("data/d_HHP2020_24.Rdata" )
attach(d_HHP2020_24)
```

Bamba Cisse – Education's effect on partnership status

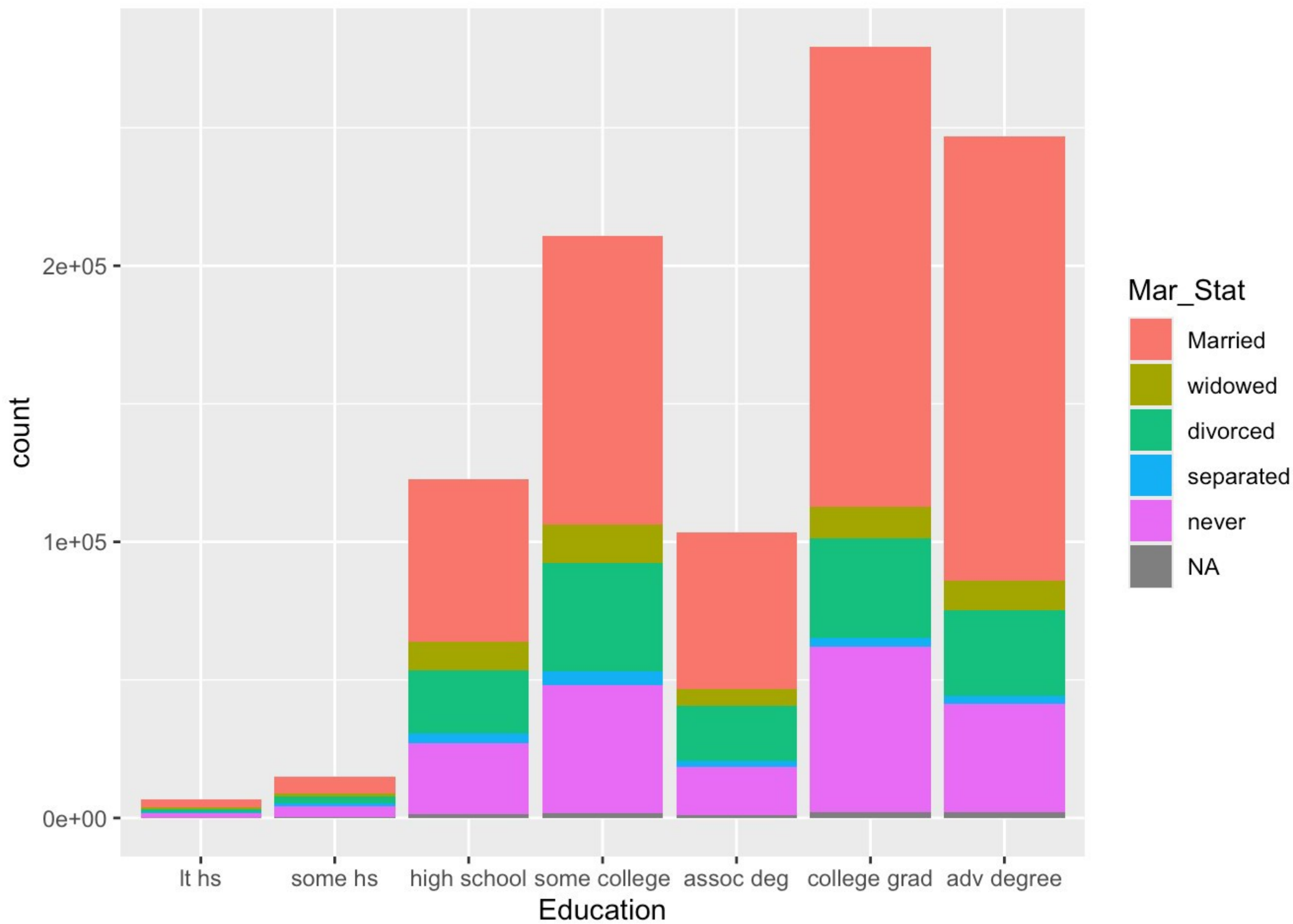
```
p <- ggplot(data = d_HHP2020_24,
            mapping = aes(x = Education, fill = Mar_Stat))
p + geom_bar(position = "fill")
```



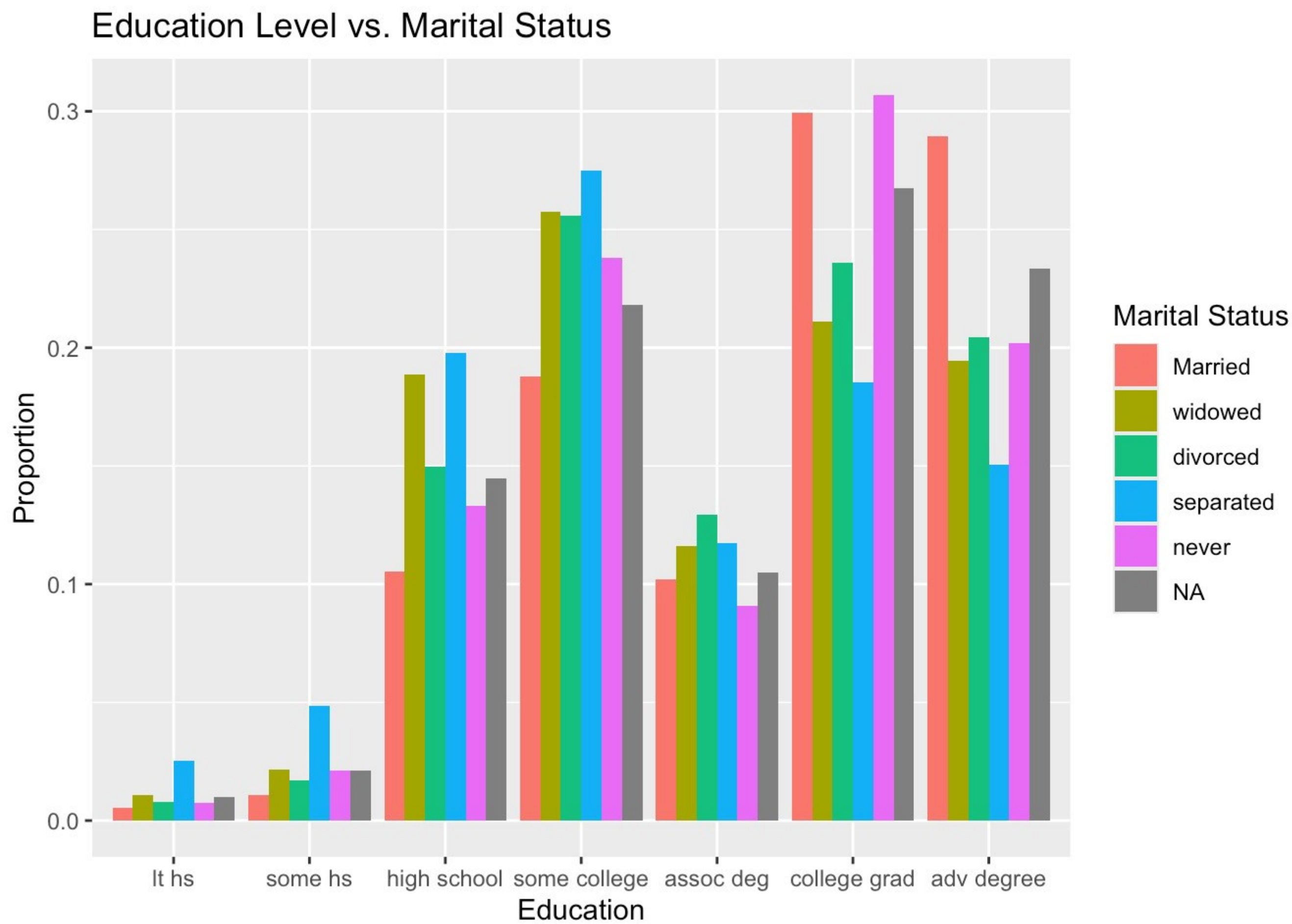

```
p + geom_bar(mapping = aes(
  y = after_stat(prop),
  group = Mar_Stat))
```




```
p <- ggplot(data = d_HHP2020_24,
            mapping = aes(x = Education, fill = Mar_Stat))
p + geom_bar()
```



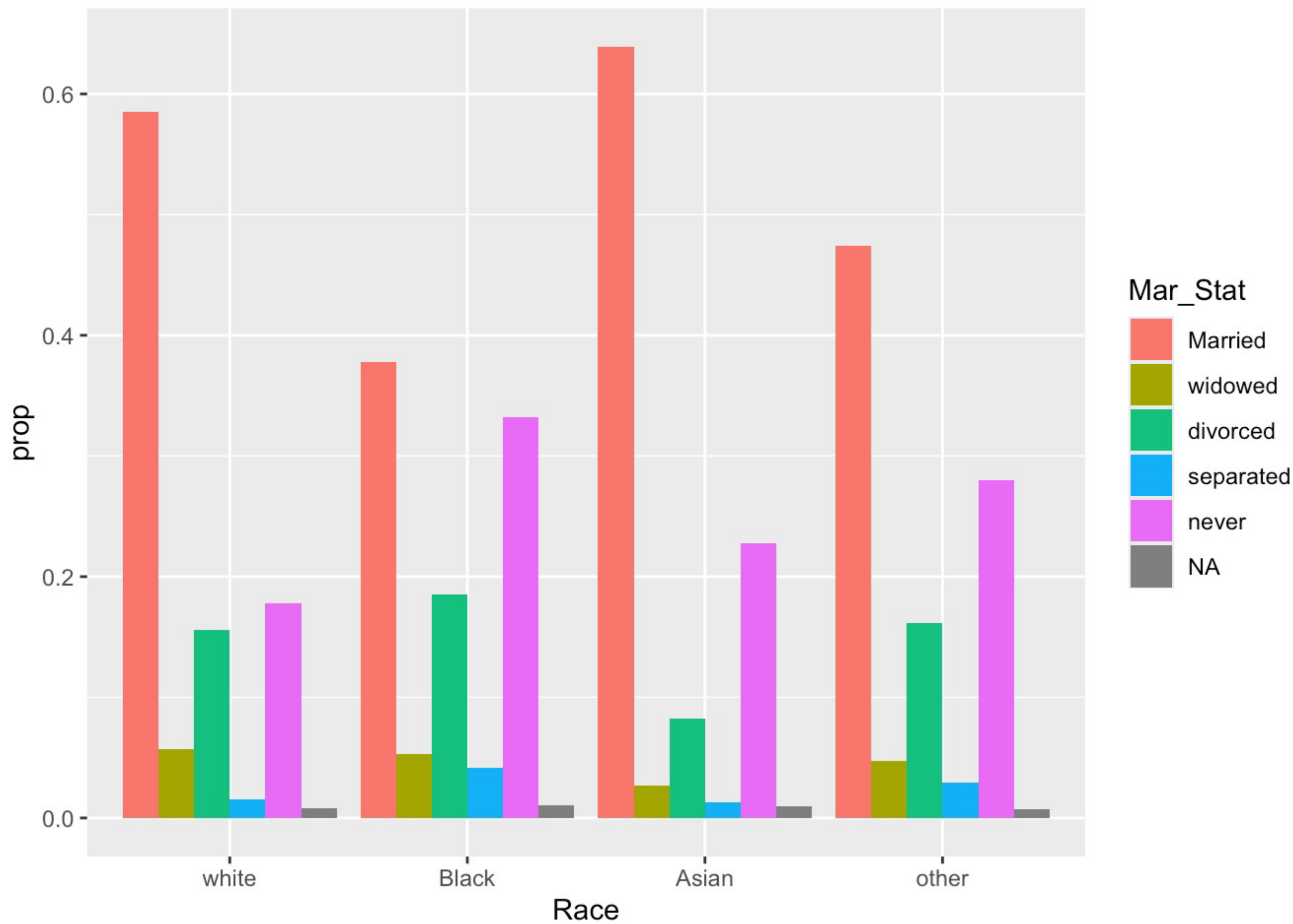

```
p + geom_bar( position = "dodge",
    mapping = aes(y = after_stat(prop), group = Mar_Stat, fill = Mar_Stat)
  ) +
  labs(title = "Education Level vs. Marital Status" ,
    x = "Education",y = "Proportion",fill = "Marital Status" )
```



Riyesh Nath – Race and Gender's effect on partnership status

First we will look at race's effect on partnership status:

```
data_groupby_race_part <- d_HHP2020_24 %>%
  count(Race, Mar_Stat) %>%
  group_by(Race) %>%
  mutate(prop = n / sum(n)) %>%
  ungroup()


ggplot(data = data_groupby_race_part, aes(x = Race, fill=Mar_Stat, y = prop)) +
  geom_col(position =  "dodge")
```



Here we see that we have highest proportion of married in Asian community, then white community, then other and finally black community. We also see that in black community there is close proportion of never married and married.

Maybe we can use chi sq test to see if race might have an affect on marriage rate.

```
d_only_married_ornot <- d_HHP2020_24 %>%
  mutate(Mar_Stat = if_else(Mar_Stat ==  "Married", "Married", "Not Married"))

print(table(d_only_married_ornot$Race,  d_only_married_ornot$Mar_Stat))
```

```
##
##          Married  No   Marrie
                      t       d
##   white  471546        32775
                            2
##  Black    30558         4942
                            1
##  Asian    31251         1715
                            0
##  other    23256         2543
                            1
```

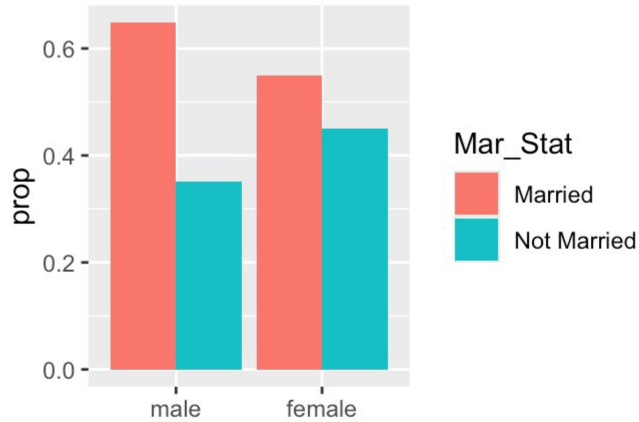
```
chisq.test(table(d_only_married_ornot$Race,  d_only_married_ornot$Mar_Stat))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(d_only_married_ornot$Race, d_only_married_ornot$Mar_Stat)
## X-squared = 15647, df = 3, p-value < 2.2e-16
```

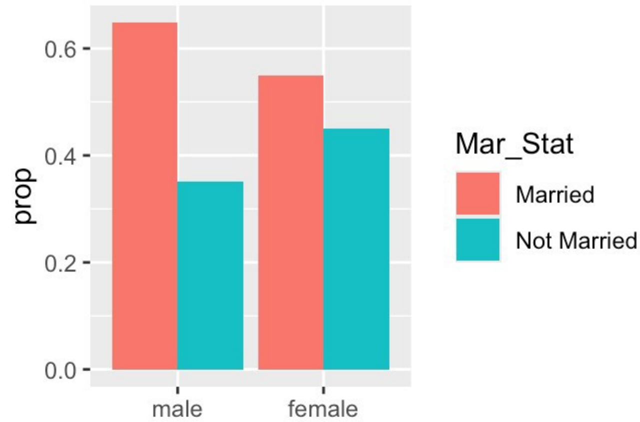
Using p-value less than .05, it seems that we can state that race does seem to have an affect on married status.

Now lets look at this further when we divide it by gender as well (we will filter trans due to the political and social complication which would make the analysis harder. Other is filter due to the ambiguity of other).

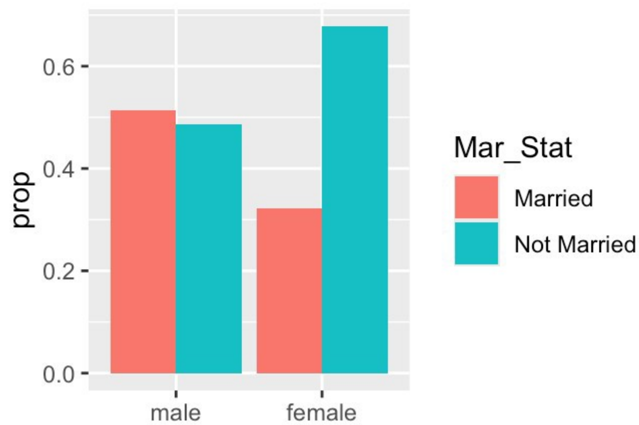
```r
d_HHP2020_24_female_male <- d_only_married_ornot %>%
  filter(Gender % in% c("male", "female"))


data_race_gender_partnership_black <- d_HHP2020_24_female_male %>%
  filter(Race == "Black", !is.na(Mar_Stat)) %>%
  count(Mar_Stat, Gender) %>%
  group_by(Gender) %>%
  mutate(prop = n / sum(n)) %>%
  ungroup()

plot_black_demo <- ggplot(data = data_race_gender_partnership_black,
       mapping = aes(x = Gender, fill=Mar_Stat, y=prop)) +
  geom_col(position =  "dodge") +
  labs(x = "Black demographic marriage status" )

data_race_gender_partnership_white <- d_HHP2020_24_female_male %>%
  filter(Race == "white", !is.na(Mar_Stat)) %>%
  count(Mar_Stat, Gender) %>%
  group_by(Gender) %>%
  mutate(prop = n / sum(n)) %>%
  ungroup()

plot_white_demo <-  ggplot(data = data_race_gender_partnership_white,
       mapping = aes(x = Gender, fill=Mar_Stat, y=prop)) +
  geom_col(position =  "dodge") +
  labs(x = "White demographic marriage status" )


data_race_gender_partnership_asian <- d_HHP2020_24_female_male %>%
  filter(Race == "Asian", !is.na(Mar_Stat)) %>%
  count(Mar_Stat, Gender) %>%
  group_by(Gender) %>%
  mutate(prop = n / sum(n)) %>%
  ungroup()

plot_asian_demo <- ggplot(data = data_race_gender_partnership_white,
       mapping = aes(x = Gender, fill=Mar_Stat, y=prop)) +
  geom_col(position =  "dodge") +
  labs(x = "Asian demographic marriage status" )

grid.arrange(plot_asian_demo, plot_white_demo, plot_black_demo, ncol =   2)
```



Asian demographic marriage status



White demographic marriage status



Black demographic marriage status

Using chi-square test for each subgroup for Race and then looking at Marriage or not Married, we see that gender has an affect.

```
d_HHP2020_24_female_male_black <- d_only_married_ornot %>%
  filter(Gender %in% c("male", "female"), Race == "Black") %>%
  droplevels()

chisq.test(table( d_HHP2020_24_female_male_black$Gende
  r, d_HHP2020_24_female_male_black$Mar_Stat
))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(d_HHP2020_24_female_male_black$Gender, d_HHP2020_24_female_male_black$Ma
r_Stat)
## X-squared = 2676.8, df = 1, p-value < 2.2e-16
```

```r
d_HHP2020_24_female_male_white <- d_HHP2020_24  %>%
  filter(Gender %in% c("male", "female"), Race == "white") %>%
  droplevels()

chisq.test(table( d_HHP2020_24_female_male_white$Gende
  r, d_HHP2020_24_female_male_white$Mar_Stat
))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(d_HHP2020_24_female_male_white$Gender, d_HHP2020_24_female_male_white$Ma
r_Stat)
## X-squared = 15462, df = 4, p-value < 2.2e-16
```

```r
d_HHP2020_24_female_male_asian <- d_HHP2020_24  %>%
  filter(Gender %in% c("male", "female"), Race == "Asian") %>%
  droplevels()

chisq.test(table( d_HHP2020_24_female_male_asian$Gende
  r, d_HHP2020_24_female_male_asian$Mar_Stat
))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(d_HHP2020_24_female_male_asian$Gender, d_HHP2020_24_female_male_asian$Ma
r_Stat)
## X-squared = 1327.3, df = 4, p-value < 2.2e-16
```

Looking at the ratio of female to male in Asian, White and Black demographic, we do see that there are a larger proportion of female vs male among the Black community than in other demographics. Could this be a factor for lack of marriage rate in Black community than other community? This needs to be tested as this hypothesis could claim that a person has a higher probability to marry someone from same Race. Unfortunately, our dataset does not give us information to test this claim.

```
data_black_community <- d_HHP2020_24_female_male %>%
  filter(Race == "Black") %>%
  count(Gender)

count_black_community <- ggplot(data = data_black_community,
       mapping = aes(x=Gender, y=n)) +
  geom_col() +
  labs(x = "Black Demographic Gender Ratio" )

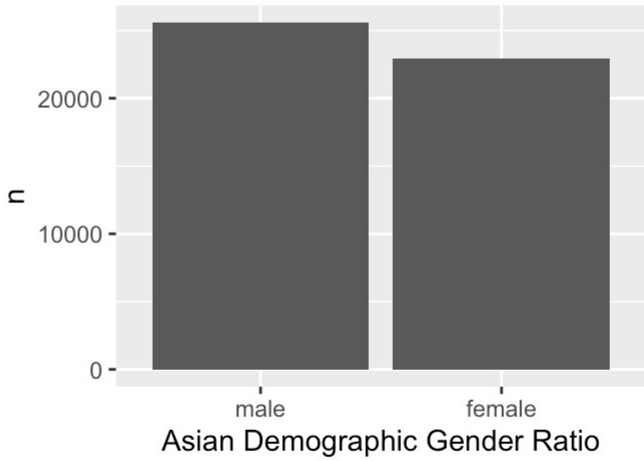

data_white_community <- d_HHP2020_24_female_male %>%
  filter(Race == "white") %>%
  count(Gender)

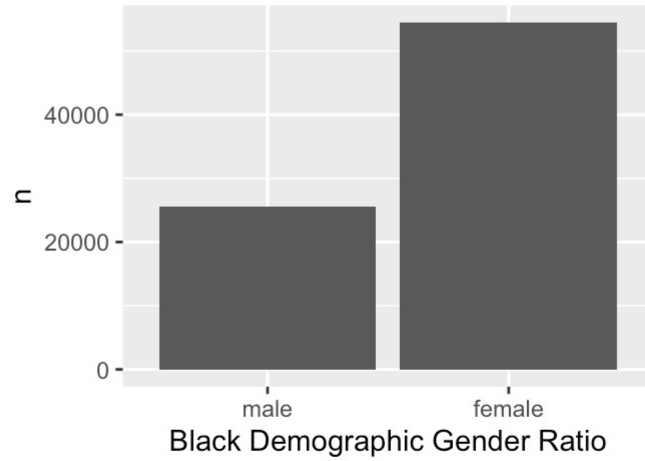
count_white_community <- ggplot(data = data_white_community,
       mapping = aes(x=Gender, y=n)) +
  geom_col() +
  labs(x = "White Demographic Gender Ratio" )

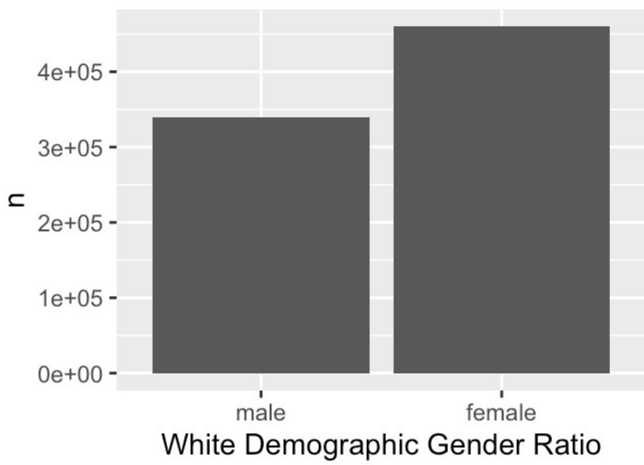
data_asian_community <- d_HHP2020_24_female_male %>%
  filter(Race == "Asian") %>%
  count(Gender)

count_asian_community <- ggplot(data = data_asian_community,
       mapping = aes(x=Gender, y=n)) +
  geom_col() +
  labs(x = "Asian Demographic Gender Ratio" )

grid.arrange(count_asian_community, count_black_community, count_white_community, ncol =
2)
```

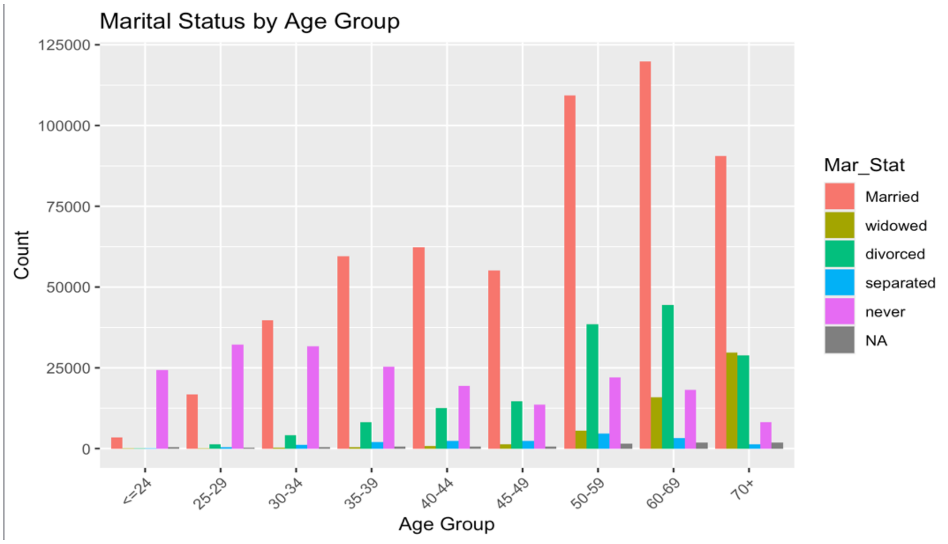






Effects of Age on Partnership Status

```
```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
library(ggplot2)
library(dplyr)
setwd("/Users/nasrinkhanam/Downloads")
unzip("d_HHP2020_24.zip")
load("d_HHP2020_24.Rdata")
```
Marital Status by Age Group
```{r}
d_HHP2020_24 <- d_HHP2020_24 %>%
  mutate(Age_group = cut(Age,
                         breaks = c(-Inf, 24, 29, 34, 39, 44, 49, 59, 69, Inf),
                         labels = c("<=24","25-29","30-34","35-39",
                                    "40-44","45-49","50-59","60-69","70+")))
ggplot(d_HHP2020_24, aes(x = Age_group, fill = Mar_Stat)) +
  geom_bar(position = "dodge") +
  labs(title = "Marital Status by Age Group",
       y = "Count", x = "Age Group") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```
```



This gives a very general view of the data, showing a natural progression of marital status for the most part increasing as one gets older. The “Married” category overtakes the “Never” substantially by the mid-30s. As the “Never” married declines, however, there is a rise in “Divorce”, which starts increasing more rapidly in the 40s onwards, causing the eventual dip in “Married”

```
# Marital Status by Age Groups (Selected Ranges)

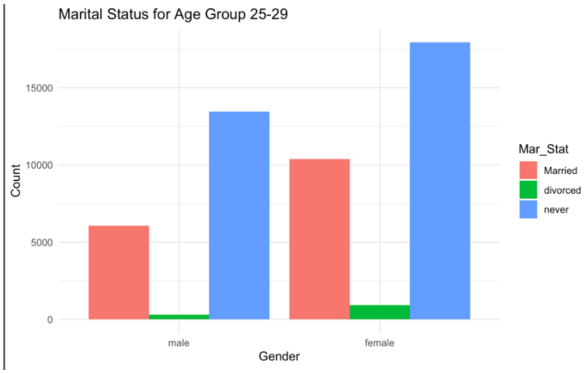
d_HHP2020_24 <- d_HHP2020_24 %>%

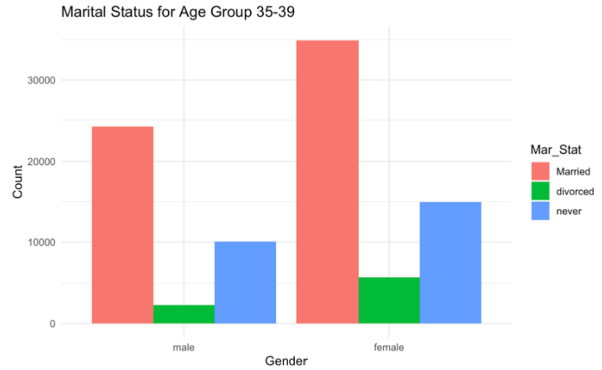
  mutate(Age_group = cut(Age,
                         breaks = c(-Inf, 24, 29, 34, 39, 44, 49, 59, 69, Inf),
                         labels = c("<=24","25-29","30-34","35-39",
                                    "40-44","45-49","50-59","60-69","70+")))

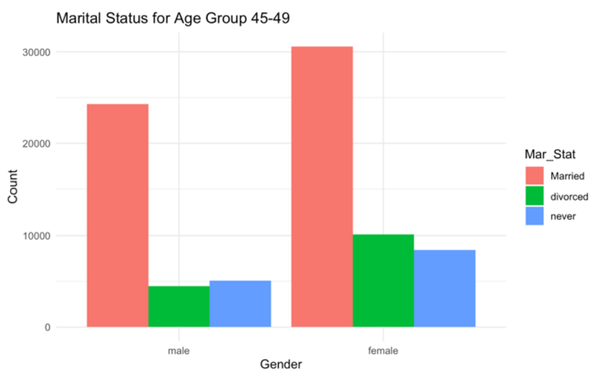
# Define function filters
plot_age_group <- function(data, group_label) {
  df <- data %>% filter(Age_group == group_label,
                        Gender %in% c("male","female"),
                        Mar_Stat %in% c("Married","never","divorced"))

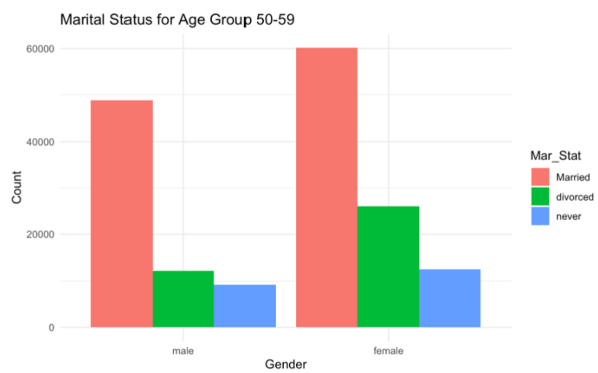
  ggplot(df, aes(x = Gender, fill = Mar_Stat)) +
    geom_bar(position = "dodge") +
    labs(title = paste("Marital Status for Age Group", group_label),
         x = "Gender", y = "Count") +
    theme_minimal()
}

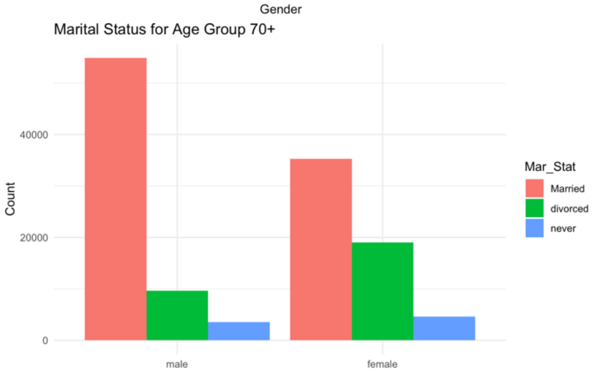
# Plots for each chosen age group
plot_age_group(d_HHP2020_24, "25-29")
plot_age_group(d_HHP2020_24, "35-39")
plot_age_group(d_HHP2020_24, "45-49")
plot_age_group(d_HHP2020_24, "50-59")
plot_age_group(d_HHP2020_24, "70+")
```











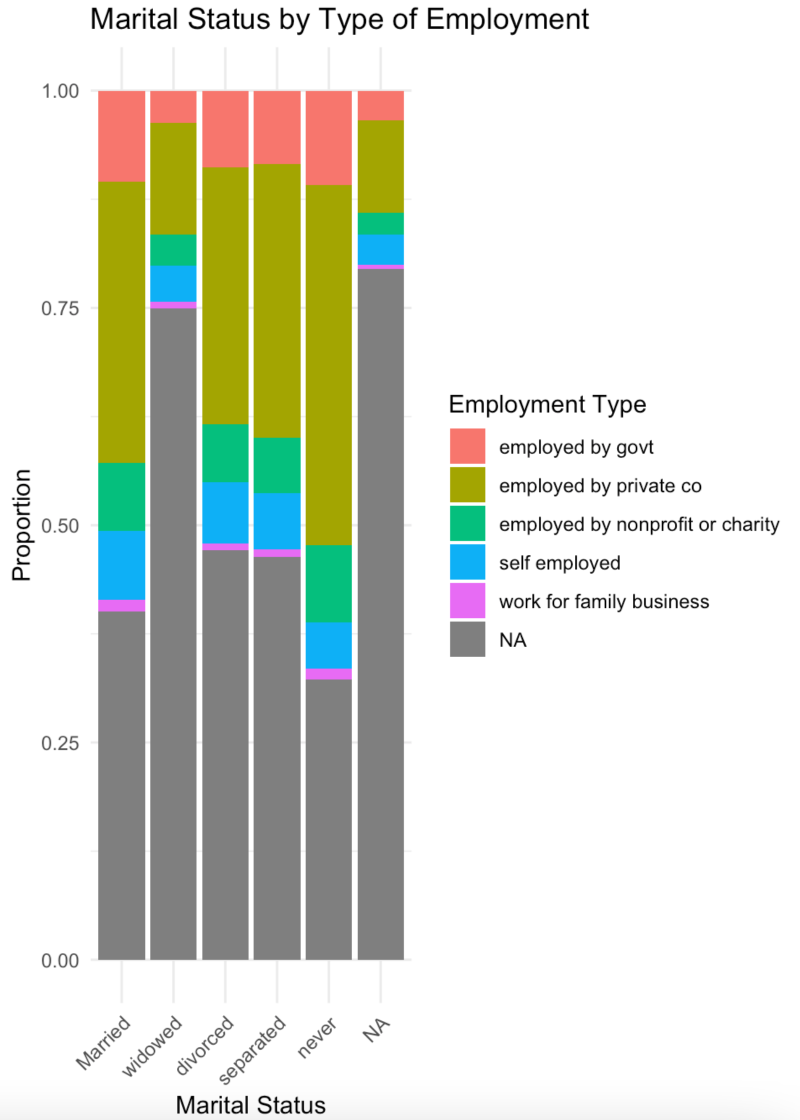
Men and Women in their 20s tend to fall mostly into "never" married, but women have a higher proportion of getting married. This suggests a timing difference between the two genders, as women tend to get married earlier than men. "Married" results in being the dominant partnering status for the rest of the age groups; however, it does fluctuate, and there's eventually a steady increase in "divorce," as the age group increases, with women dominating. People are more likely to get married the older they are, which is likely due to the desire to settle down however, oftentimes other important factors aren't taken into account in making that decision, which is seen in the decrease in the number of people who are married and an increase in divorce by the time people hit their 70s.

<u>How Employment Type Affects Marital Status</u>

```
> load("~/Downloads/d_HHP2020_24.Rdata")
> summary(d_HHP2020_24)
library(tidyverse)
library(ggplot2)


> ggplot(d_HHP2020_24, aes(x = Mar_Stat, fill = work_kind)) +
+     geom_bar(position = "fill") +
+     labs(
+         title = "Marital Status by Type of Employment",
+         x = "Marital Status",
+         y = "Proportion",
```

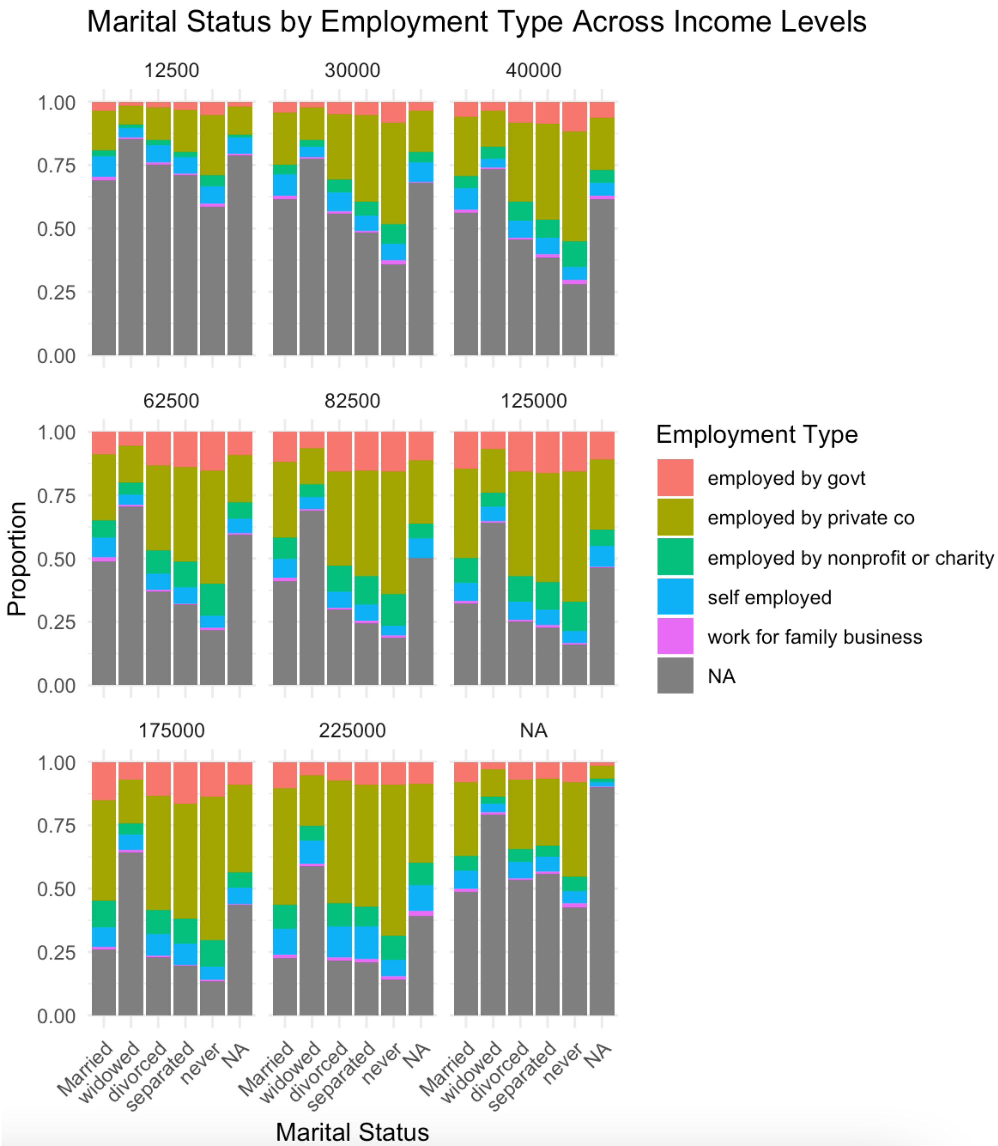
```
+       fill = "Employment Type"
+     ) +
+     theme_minimal() +
+     theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Married individuals are more likely to be employed by private companies or the government, while widowed and divorced individuals appear more frequently in the “not employed/NA” category. Those who have never married show up a lot in the “NA” group, which may suggest they’re less tied to formal jobs. Self-employment and nonprofit work are present in every group, but they make up a much smaller share compared to private company or government jobs. Overall, the chart shows that stable forms of employment are strongly associated with being married, while unemployment is more prevalent among never-married, divorced, and widowed individuals.

```
> ggplot(d_HHP2020_24, aes(x = Mar_Stat, fill = work_kind)) +
+     geom_bar(position = "fill") +
+     facet_wrap(~ income_midpoint_factor, ncol = 3) +
+     labs(
+       title = "Marital Status by Employment Type Across Income Levels",
+       x = "Marital Status",
```

```
+       y = "Proportion",
+       fill = "Employment Type"
+     ) +
+     theme_minimal() +
+     theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



When income is added as a second factor, we see more of a clearer picture. At relatively lower income levels (12,500–40,000), never-married individuals are strongly associated with unemployment (NA), while married individuals appear less connected to stable jobs. In the middle-income brackets (62,500–125,000), married individuals dominate, especially within private company and government employment, which shows the importance of stable, mid-level earnings in supporting marriage. At higher incomes (175,000–225,000), marriage becomes even more dominant, with most individuals employed in private companies and fewer in unemployment or alternative employment categories. Divorced, widowed, and separated people show up more often in the “NA” group, which could mean they’re less likely to be working or they may have job instability after changes in their marriage. All this shows that income strengthens the relationship between employment and marriage: stable, higher-paying jobs are linked with higher rates of marriage, while unemployment and low income are associated with never marrying or with separation.