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

Riyesh Nath

2025-09-14

GROUP MEMBERS: Bamba Cisse, Nasrin 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.
  -

```
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
```

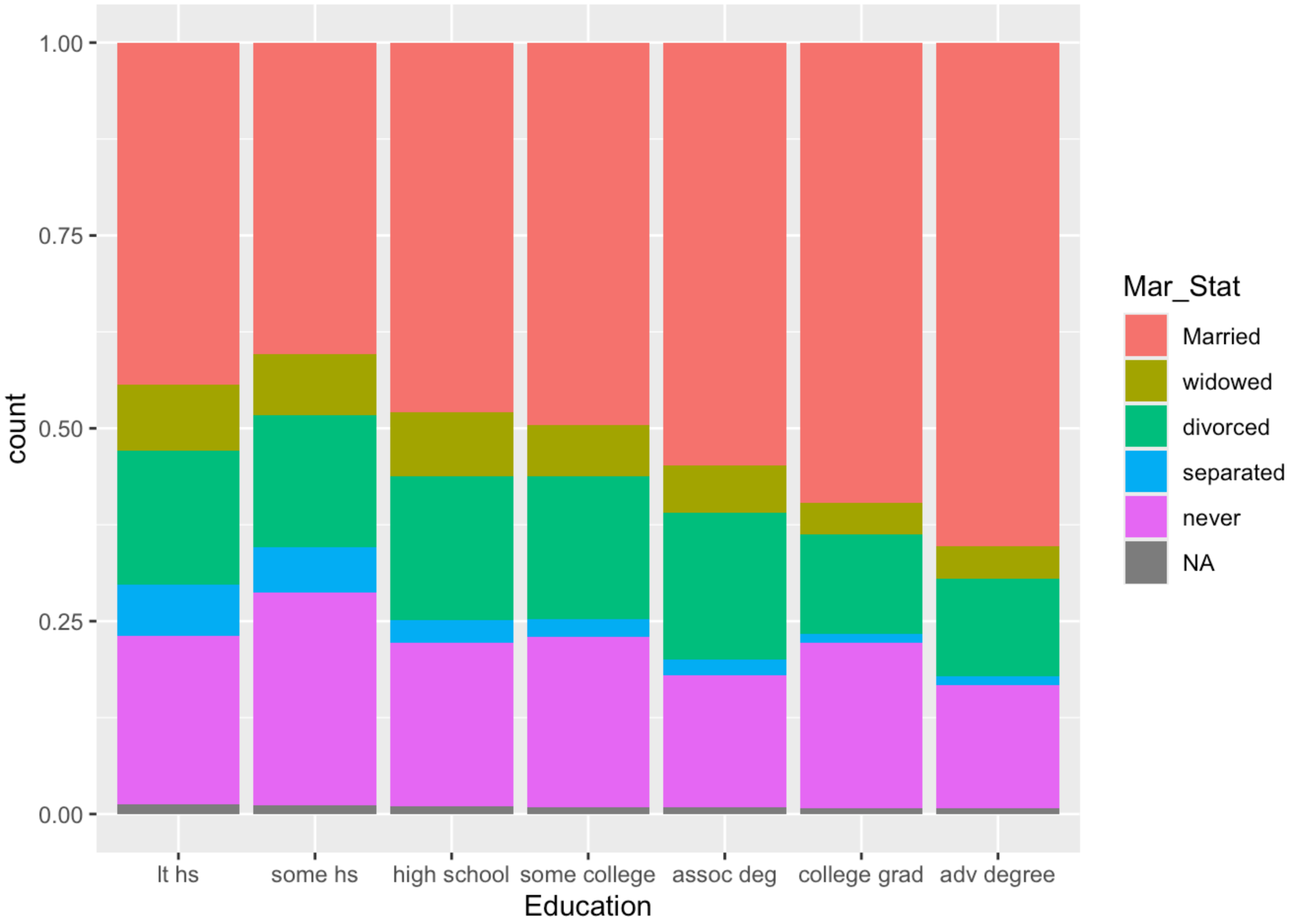
```
library(gridExtra)
```

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

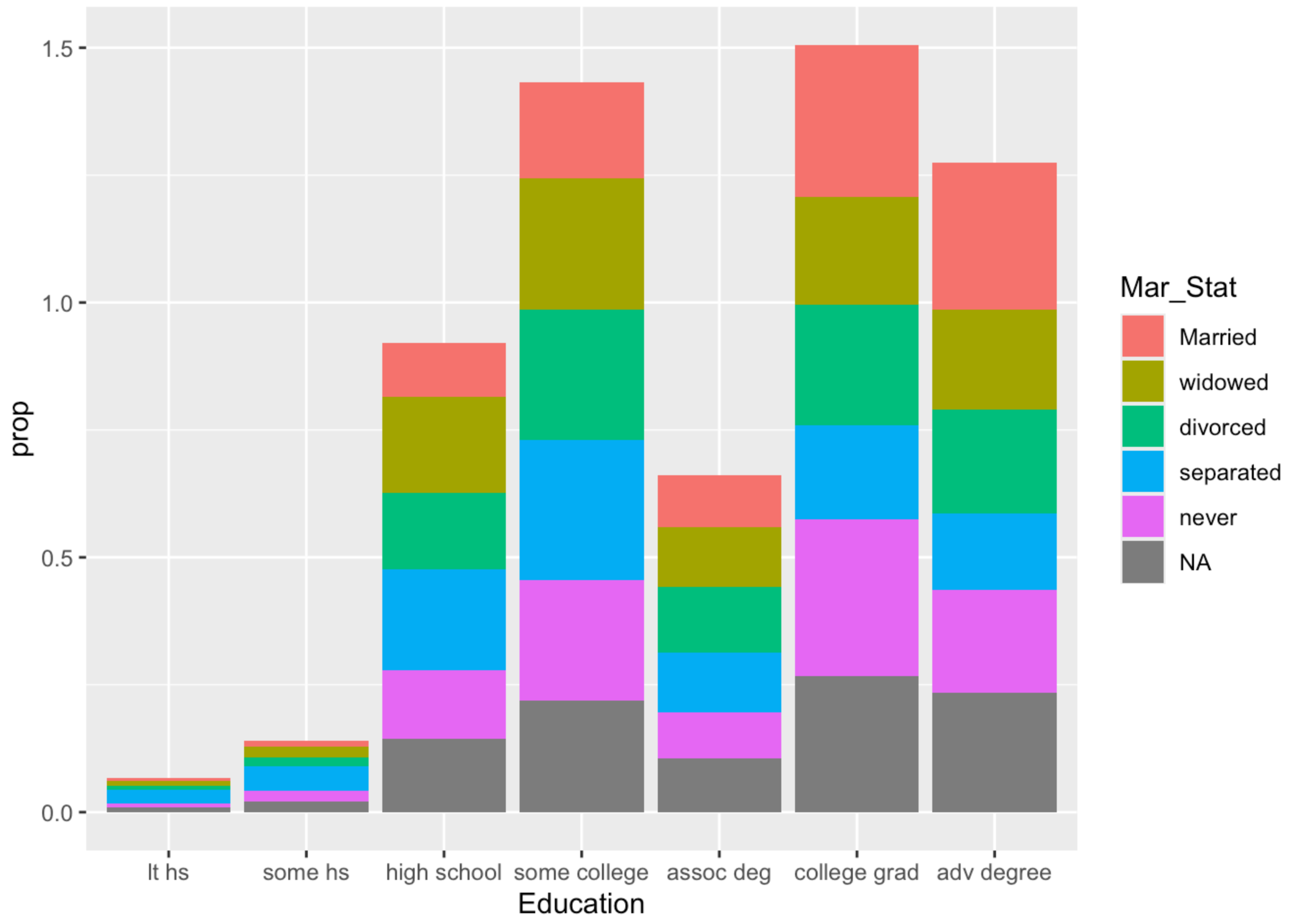
```
load("data/d_HHP2020_24.Rdata")
attach(d_HHP2020_24)
```

Bamba Cisse – Education’s effect on partnership status

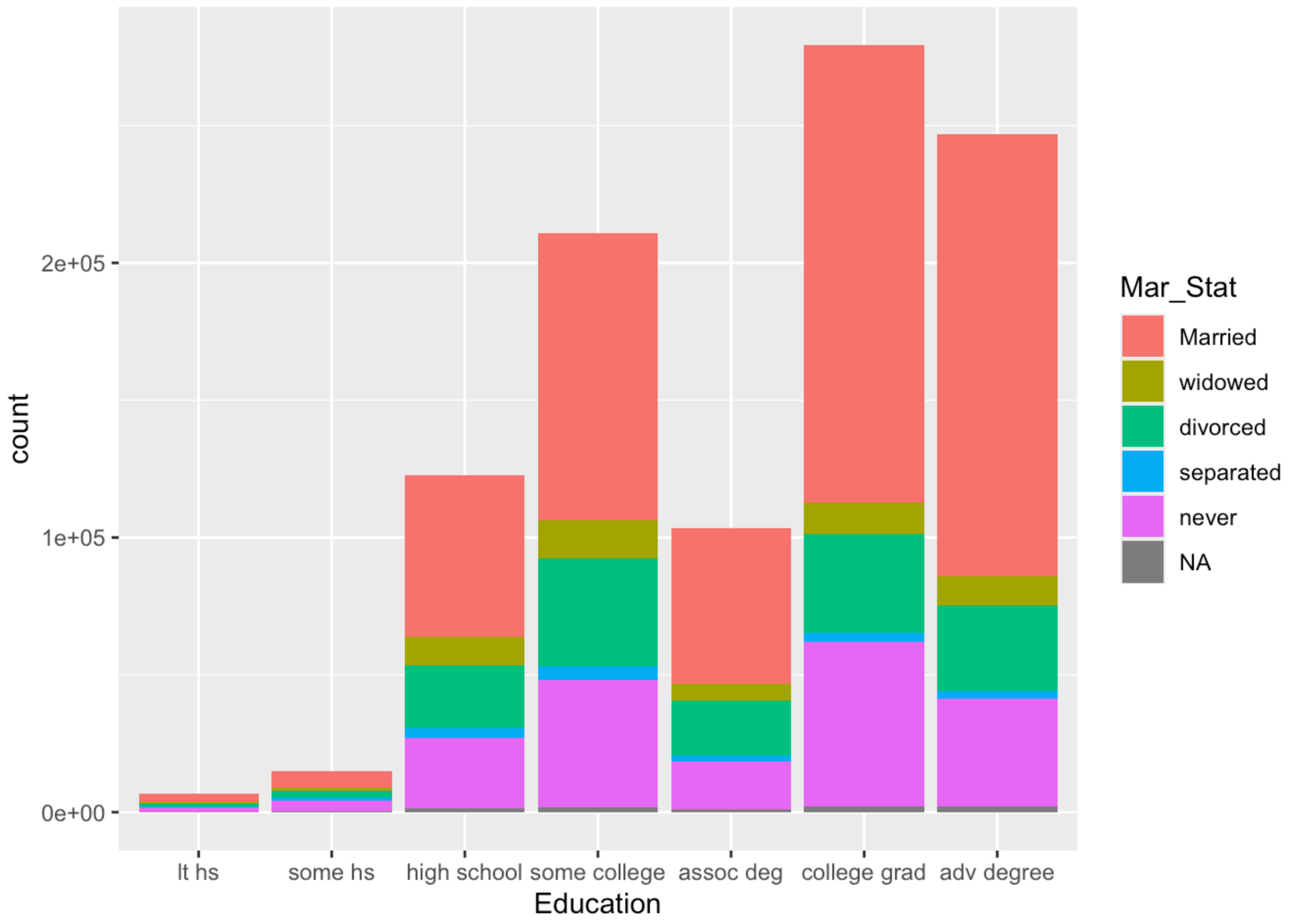
```
p <- ggplot(data = d_HHP2020_24,
            mapping = aes(x = Education, fill = Mar_Stat))
p + geom_bar(position = "fill")
```



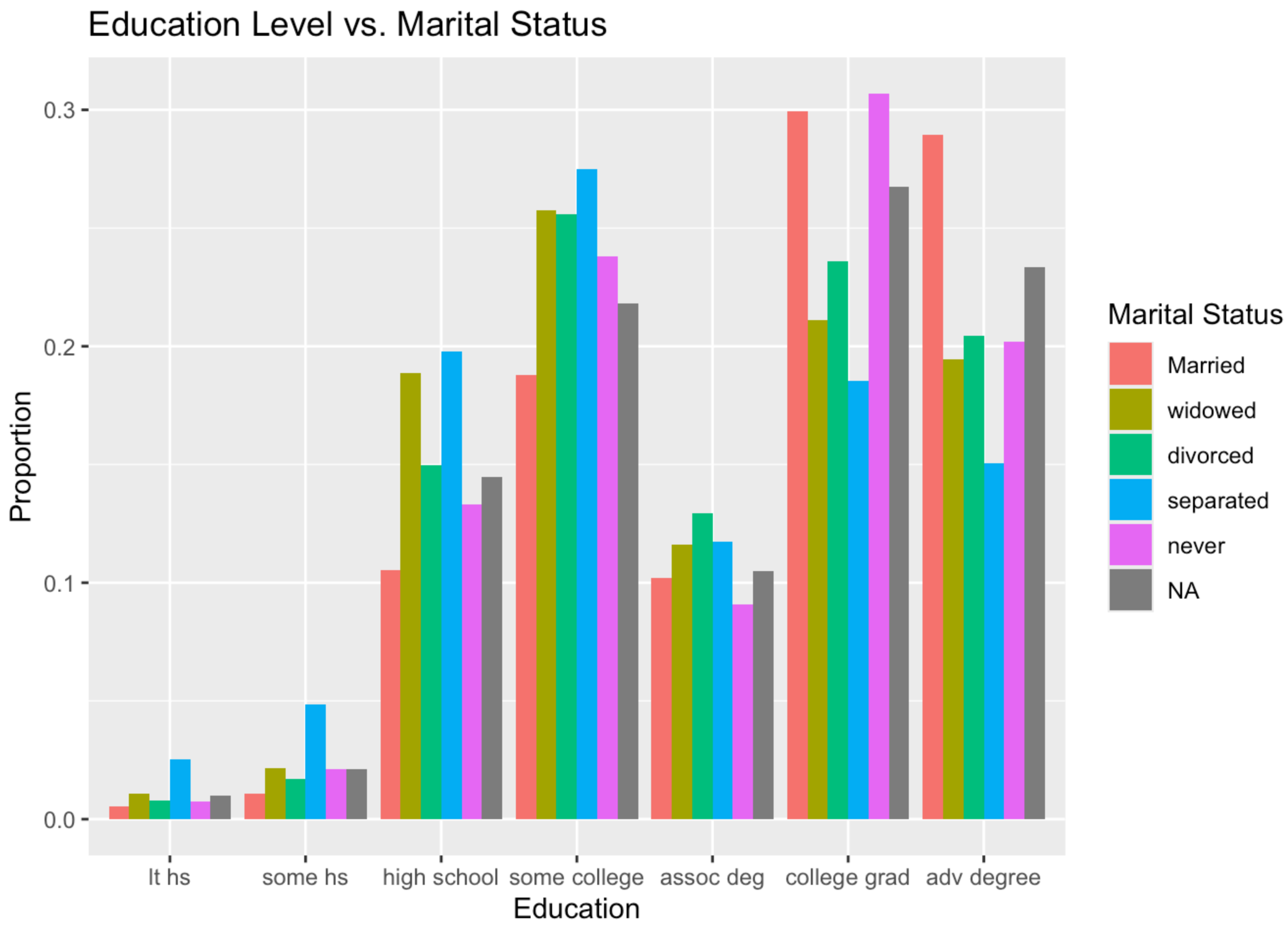
```
p + geom_bar(mapping = aes(
  y = after_stat(prop),
  group = Mar_Stat))
```



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



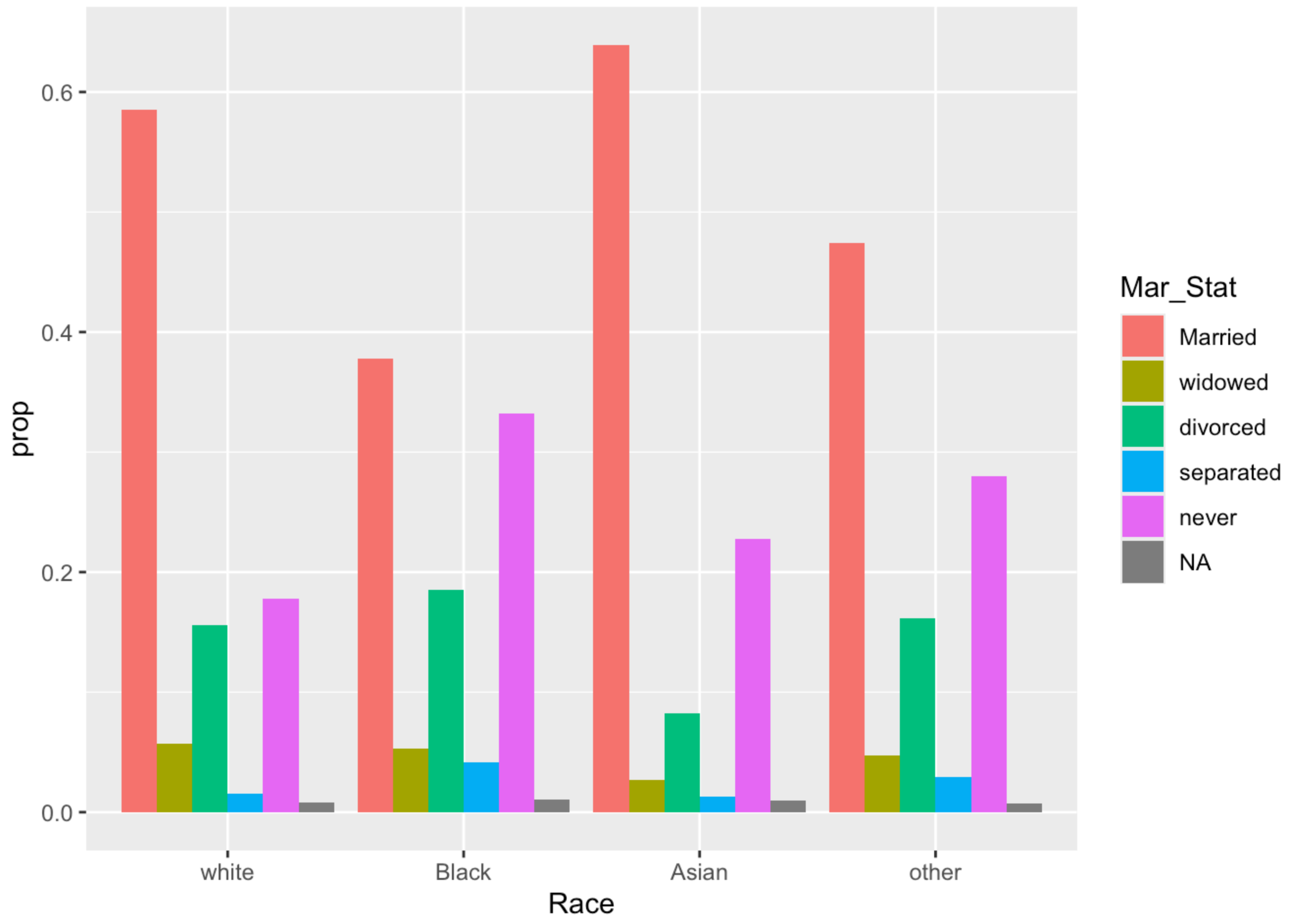
```
p + geom_bar( position = "dodge",
    mapping = aes(y = after_stat(prop), group = Mar_Stat, fill = Mar_Stat)
  ) +
  labs(title = "Education Level vs. Marital Status",
    x = "Education",y = "Proportion",fill = "Marital Status")
```



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 Not Married
##   white  471546      327752
##   Black   30558       49421
##   Asian   31251       17150
##   other   23256       25431
```

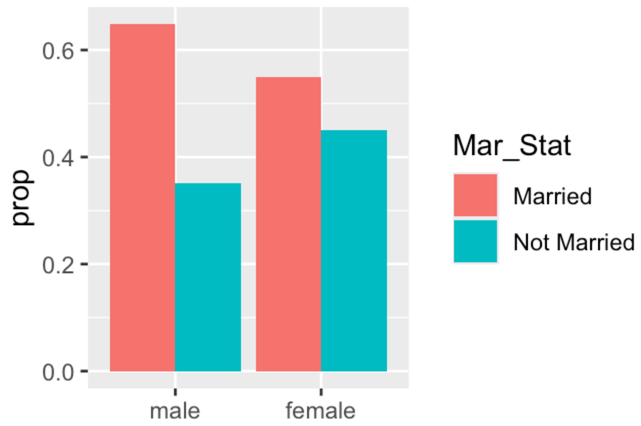
```
chisq.test(table(d_only_married_ornot$Race, d_only_married_ornot$Mar_Stat))
```

```
##
##  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
```

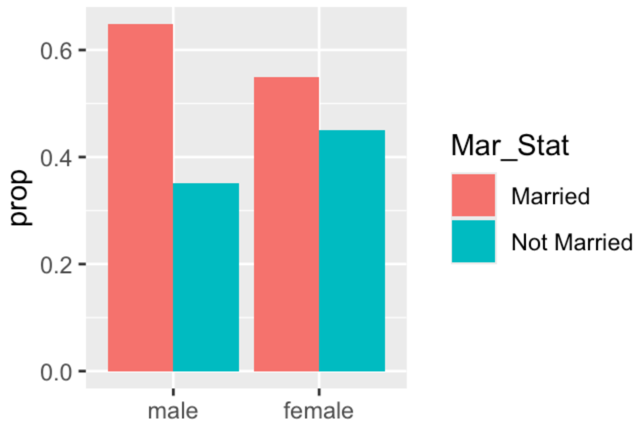
Using p-value less than .05, it seems that we can state that race does seem to have an affect on married status.

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).

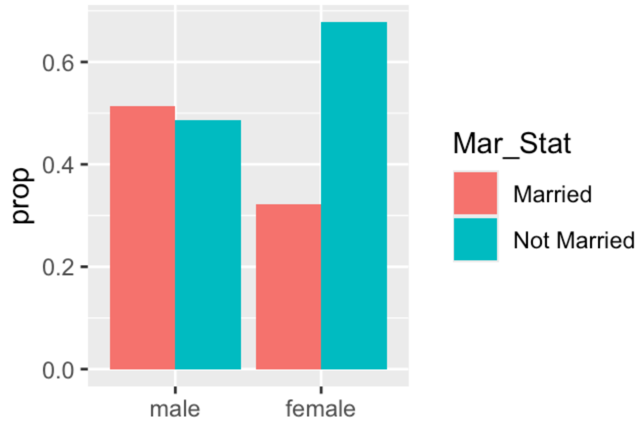
```r
d_HHP2020_24_female_male <- d_only_married_ornot %>%
  filter(Gender %in% c("male", "female"))


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$Gender,
  d_HHP2020_24_female_male_black$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
```

```
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$Gender,
  d_HHP2020_24_female_male_white$Mar_Stat
))
```

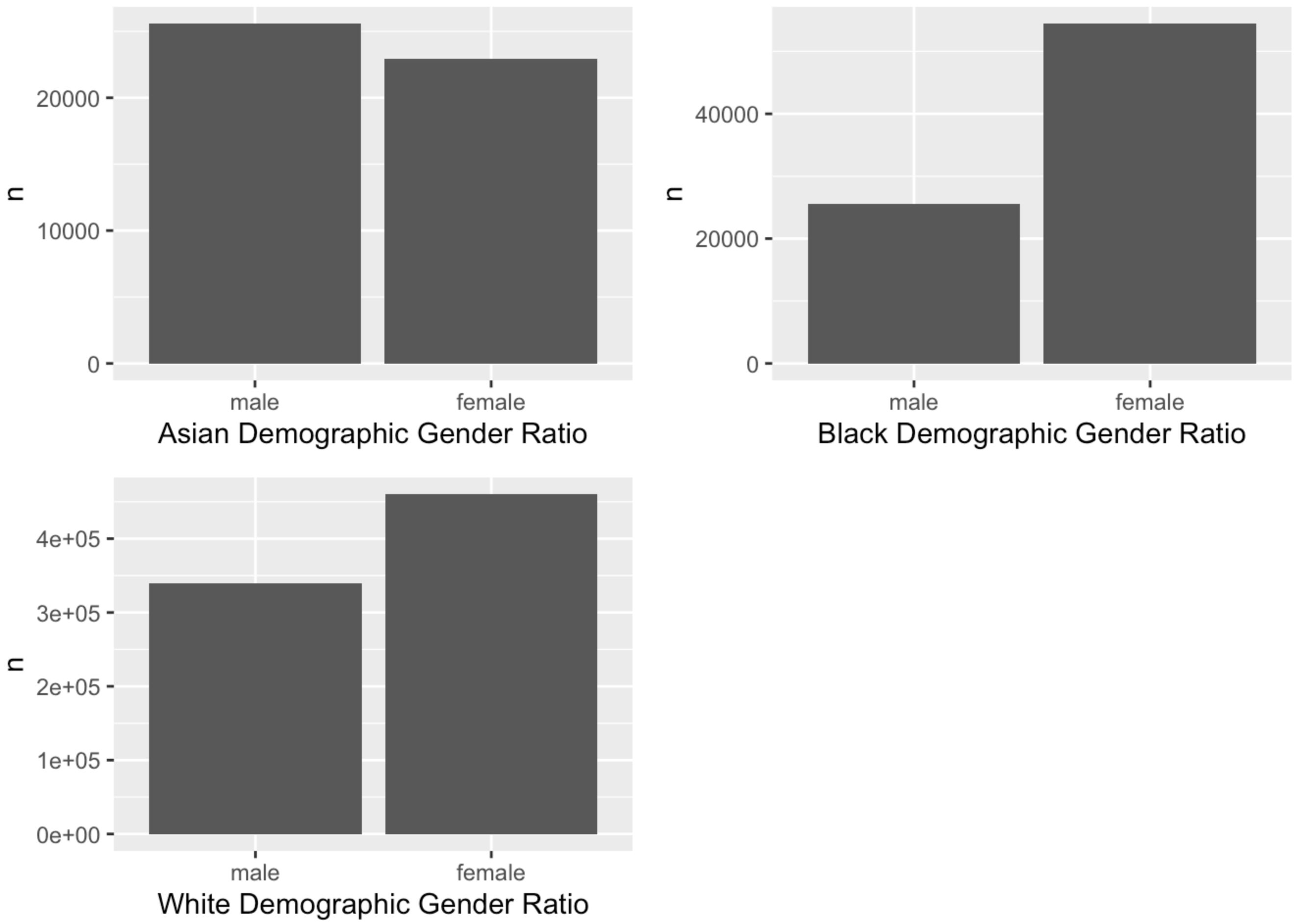
```
##
##  Pearson's Chi-squared test
##
## data:  table(d_HHP2020_24_female_male_white$Gender, d_HHP2020_24_female_male_white$Ma
r_Stat)
## X-squared = 15462, df = 4, p-value < 2.2e-16
```

```
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$Gender,
  d_HHP2020_24_female_male_asian$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)
```


n
20000
10000
0
male
female
Asian Demographic Gender Ratio
n
40000
20000
0
male
female
Black Demographic Gender Ratio
n
4e+05
3e+05
2e+05
1e+05
0e+00
male
female
White Demographic Gender Ratio