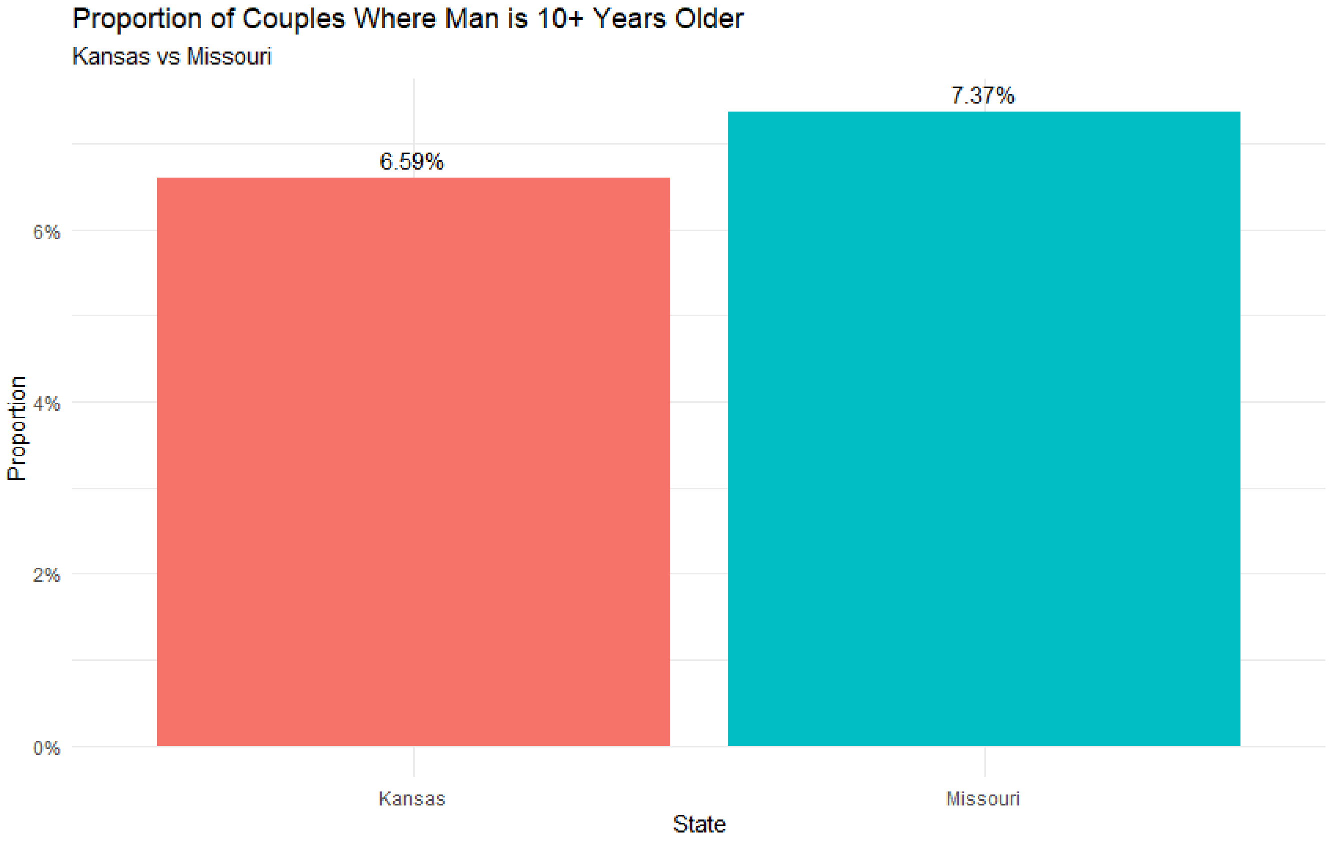
Michael Stewart, Riyesh Nath, Jason Seda and Maurice Agonsi

```
                                        ##
library(ggplot2)
library(tidyverse)
library(haven)

setwd("C:/Users/Michael/Desktop/ECON/Data")
load("ACS_2021_couples.RData")

# from Kevin: Let me fix up a couple of the variables with somewhat mysterious coding.

acs2021_couples$RACE <- fct_recode(as.factor(acs2021_couples$RACE),
                              "White" = "1",
                              "Black" = "2",
                              "American Indian or Alaska Native" = "3",
                              "Chinese" = "4",
                              "Japanese" = "5",
                              "Other Asian or Pacific Islander" = "6",
                              "Other race" = "7",
                              "two races" = "8",
                              "three races" = "9")

acs2021_couples$h_race <- fct_recode(as.factor(acs2021_couples$h_race),
                               "White" = "1",
                               "Black" = "2",
                               "American Indian or Alaska Native" = "3",
                               "Chinese" = "4",
                               "Japanese" = "5",
                               "Other Asian or Pacific Islander" = "6",
                               "Other race" = "7",
                               "two races" = "8",
                               "three races" = "9")

acs2021_couples$HISPAN <- fct_recode(as.factor(acs2021_couples$HISPAN),
                               "Not Hispanic" = "0",
                               "Mexican" = "1",
                               "Puerto Rican" = "2",
                               "Cuban" = "3",
                               "Other" = "4")
acs2021_couples$h_hispan <- fct_recode(as.factor(acs2021_couples$h_hispan),
                                "Not Hispanic" = "0",
                                "Mexican" = "1",
                                "Puerto Rican" = "2",
                                "Cuban" = "3",
                                "Other" = "4")
```

```
# dummy variable for if the man is more than *10* years older than the woman (modifying to include is
exactly 10 years older by changing < -10 to <= -10)

trad_data <- acs2021_couples %>% filter( (SEX == "Female") & (h_sex == "Male") )

trad_data$he_more_than_10yrs_than_her <- as.numeric(trad_data$age_diff <= -10)

# verifying that 1 corresponds to TRUE (man is 10+ years older than woman)

table(trad_data$he_more_than_10yrs_than_her,cut(trad_data$age_diff,c(-100,-10, -5, 0, 5, 10, 100)))

# first estimate from Kevin's code, changed to 10yrs

ols_out1 <- lm(he_more_than_10yrs_than_her ~ educ_hs + educ_somecoll + educ_college +
educ_advdeg + AGE, data = trad_data)
summary(ols_out1)

# Residuals:
#   Min       1Q   Median       3Q      Max
# -0.18687 -0.09641 -0.07386 -0.05087  0.99776
#
# Coefficients:
#   Estimate Std. Error t value Pr(>|t|)
# (Intercept)    2.129e-01  2.138e-03   99.57   <2e-16 ***
#   educ_hs       -3.348e-02  1.767e-03  -18.95   <2e-16 ***
#   educ_somecoll -4.461e-02  1.854e-03  -24.06   <2e-16 ***
#   educ_college  -6.600e-02  1.835e-03  -35.98   <2e-16 ***
#   educ_advdeg   -7.068e-02  1.934e-03  -36.55   <2e-16 ***
#   AGE          -1.628e-03  2.606e-05  -62.47   <2e-16 ***
#   ---
#   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# Residual standard error: 0.269 on 412269 degrees of freedom
# Multiple R-squared:  0.01302,        Adjusted R-squared:  0.013
# F-statistic:  1087 on 5 and 412269 DF,  p-value: < 2.2e-16
#
# As education level increases, couples less likely to have age gaps where man is 10+ years older.
Effect gets stronger with more education.
# As data shows woman's current age, vs age at marriage, could correlate to educated men marrying
younger, less educated women later in life
# R² = 0.013 model only explains about 1.3% of variation, quite low, suggesting there are other factors
determining age gaps beyond education and age
# All coefficients have ***, highly statistically significant (p < 0.001)

# Checking what values we have
names(trad_data)
```

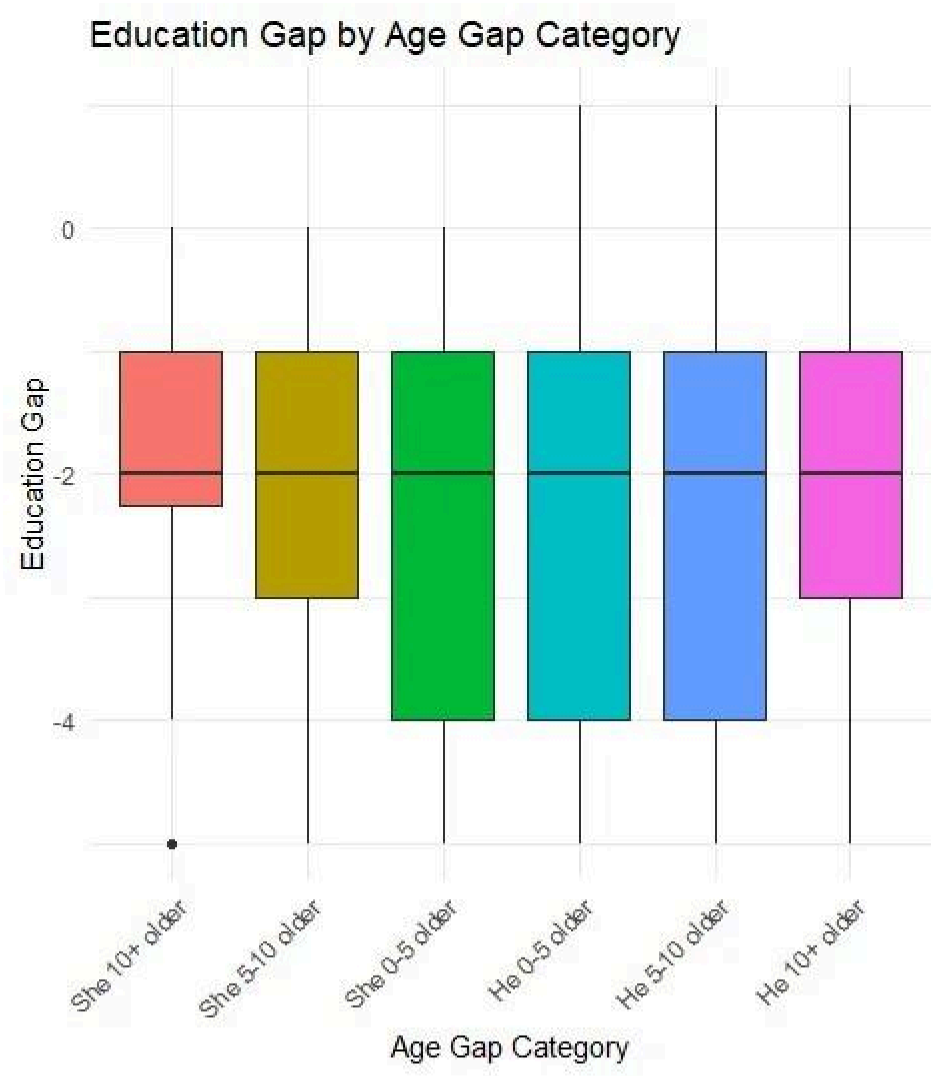
```
# Testing with numeric education variables
ols_out2 <- lm(he_more_than_10yrs_than_her ~ EDUC + h_educ + AGE, data = trad_data)
summary(ols_out2)

# Residuals:
#   Min       1Q   Median       3Q      Max
# -0.25997 -0.09607 -0.07329 -0.05008  1.00961
#
# Coefficients:
#   Estimate Std. Error t value Pr(>|t|)
# (Intercept)                    2.367e-01  4.130e-03  57.313  < 2e-16 ***
#   EDUCNursery school to grade 4    3.240e-02  7.432e-03   4.360 1.30e-05 ***
#   EDUCGrade 5, 6, 7, or 8         -1.870e-02  4.831e-03  -3.870 0.000109 ***
#   EDUCGrade 9                    -1.753e-02  5.957e-03  -2.943 0.003253 **
#   EDUCGrade 10                    -8.823e-03  5.550e-03  -1.590 0.111918
# EDUCGrade 11                   -7.640e-03  5.410e-03  -1.412 0.157874
# EDUCGrade 12                   -3.164e-02  3.628e-03  -8.722  < 2e-16 ***
#   EDUC1 year of college           -4.001e-02  3.789e-03 -10.559  < 2e-16 ***
#   EDUC2 years of college          -4.293e-02  3.830e-03 -11.209  < 2e-16 ***
#   EDUC4 years of college          -6.115e-02  3.717e-03 -16.450  < 2e-16 ***
#   EDUC5+ years of college         -6.854e-02  3.793e-03 -18.070  < 2e-16 ***
#   h_educNursery school to grade 4  2.384e-02  8.111e-03   2.940 0.003286 **
#   h_educGrade 5, 6, 7, or 8       -9.975e-03  5.025e-03  -1.985 0.047135 *
#   h_educGrade 9                  -5.856e-03  6.023e-03  -0.972 0.330890
# h_educGrade 10                 -2.545e-03  5.911e-03  -0.431 0.666815
# h_educGrade 11                 -3.710e-03  5.714e-03  -0.649 0.516084
# h_educGrade 12                 -2.494e-02  3.839e-03  -6.497 8.22e-11 ***
#   h_educ1 year of college        -2.136e-02  3.969e-03  -5.381 7.41e-08 ***
#   h_educ2 years of college       -2.922e-02  4.057e-03  -7.203 5.90e-13 ***
#   h_educ3 years of college       -3.606e-02  3.909e-03  -9.225  < 2e-16 ***
#   h_educ4 years of college       -1.856e-02  3.960e-03  -4.687 2.77e-06 ***
#   AGE                          -1.648e-03  2.615e-05 -63.000  < 2e-16 ***
#   ---
#   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# Residual standard error: 0.2688 on 412253 degrees of freedom
# Multiple R-squared:  0.01411,        Adjusted R-squared:  0.01406
# F-statistic:   281 on 21 and 412253 DF,  p-value: < 2.2e-16

# Interpretation:

# Woman's education matters more than man's education in predicting large age gaps
# Both partners having more education = less likely to have 10+ year age gap
# Still very low R² (0.014), only ~1.4% of variation
# Oddity: Some very low education levels (Nursery-Grade 4) show positive coefficients - these might be
small sample sizes or data quirks.
```

```
# Comparing states of Kansas and Missouri. The states border each other however Kansas allows for marriage as young as 15 in certain circumstances.
# Missouri does not allow marriage under 18 under any circumstances.

# Create comparison viz

# Check state codes
table(trad_data$STATEFIP)

# Step 1: Filter for Kansas and Missouri
ks_mo_data <- trad_data %>%
  filter(STATEFIP %in% c("Kansas", "Missouri"))

# Step 2: Check if it worked
nrow(ks_mo_data)  # Should show 11,794 (3778 + 8016)

# Step 3: Calculate proportions
age_gap_comparison <- ks_mo_data %>%
  group_by(STATEFIP) %>%
  summarize(
    prop_10plus_gap = mean(he_more_than_10yrs_than_her),
    n = n()
  )

# Step 4: View the results
print(age_gap_comparison)

# Step 5: Make the chart
ggplot(age_gap_comparison, aes(x = STATEFIP, y = prop_10plus_gap, fill = STATEFIP)) +
  geom_col() +
  geom_text(aes(label = paste0(round(prop_10plus_gap * 100, 2), "%")),
            vjust = -0.5) +
  labs(title = "Proportion of Couples Where Man is 10+ Years Older",
       subtitle = "Kansas vs Missouri",
       x = "State",
       y = "Proportion") +
  scale_y_continuous(labels = scales::percent) +
  theme_minimal() +
  theme(legend.position = "none")

# Histogram chart reveals slightly higher proportion in MO despite difference in laws.
```



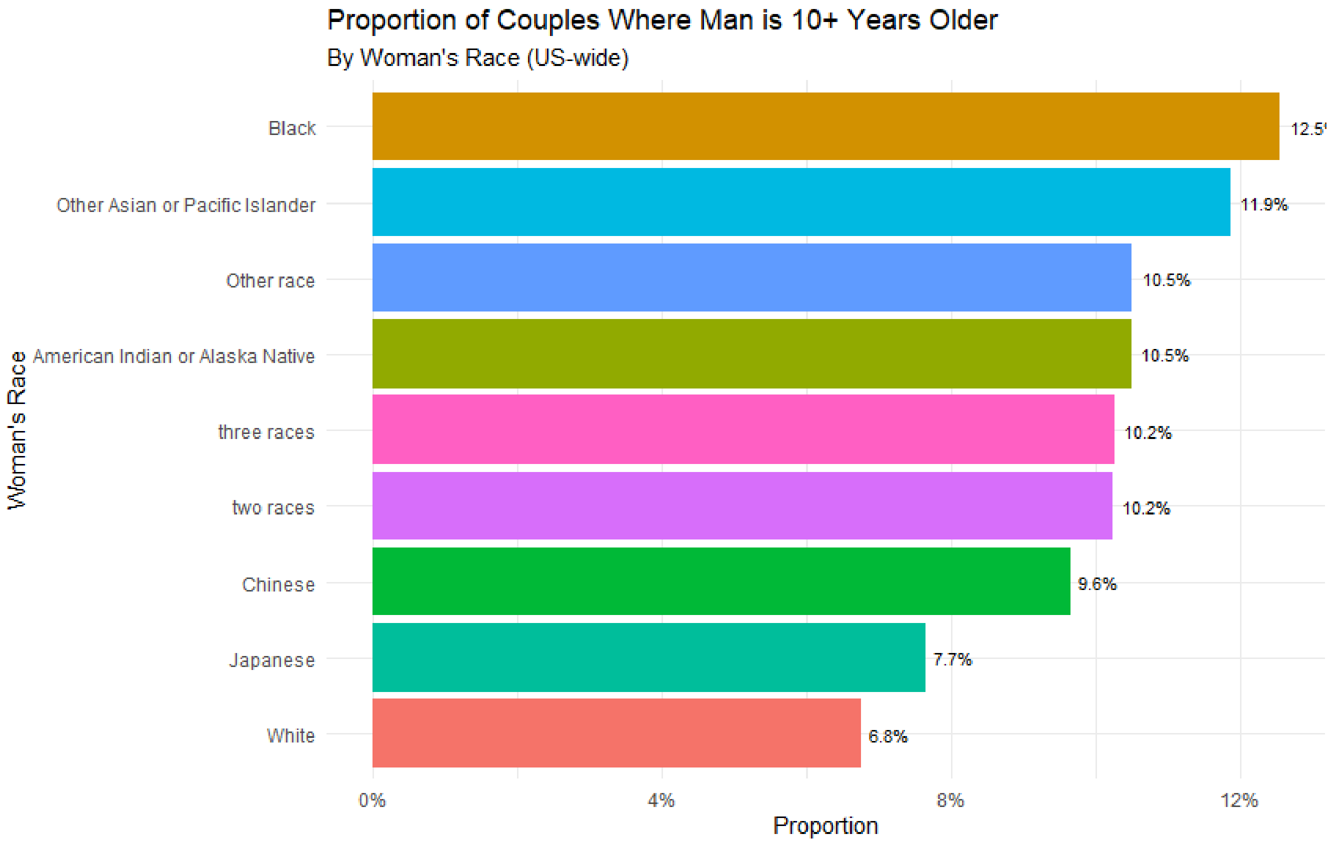
```
# Checking to see if education level changes anything

# Check the education variables
class(ks_mo_data$EDUC)
class(ks_mo_data$EDUCD)
class(ks_mo_data$h_educ)
class(ks_mo_data$h_educd)

# Convert education to simple levels
educ_to_level <- function(educ) {
  case_when(
    grepl("N/A|No schooling|Grade [1-9]|Grade 1[01]|12th grade, no diploma", educ) ~ 0,  # No HS
diploma
    grepl("Grade 12|high school|GED|Regular high school diploma", educ, ignore.case = TRUE) ~ 1,  #
HS diploma
    grepl("Some college|1 year|2 year|1 or more years", educ, ignore.case = TRUE) ~ 2,  # Some college
    grepl("Associate", educ, ignore.case = TRUE) ~ 3,  # Associates
    grepl("Bachelor|4 years of college", educ, ignore.case = TRUE) ~ 4,  # Bachelors
    grepl("Master|Professional|Doctoral|5\\+|6 years|7 years|8\\+", educ, ignore.case = TRUE) ~ 5,  #
Graduate degree
    TRUE ~ NA_real_
  )
}
```

```
# Re-apply to both partners
ks_mo_data$her_educ_level <- educ_to_level(as.character(ks_mo_data$EDUCD))
ks_mo_data$his_educ_level <- educ_to_level(as.character(ks_mo_data$h_educd))

# Recreate education gap
ks_mo_data$educ_gap <- ks_mo_data$his_educ_level - ks_mo_data$her_educ_level

# Check how many NAs
sum(is.na(ks_mo_data$educ_gap))
# Result: 14, acceptable

# Create age gap category for ks_mo_data
ks_mo_data$age_gap_category <- cut(-ks_mo_data$age_diff,
                                     breaks = c(-Inf, -10, -5, 0, 5, 10, Inf),
                                     labels = c("She 10+ older", "She 5-10 older",
                                               "She 0-5 older", "He 0-5 older",
                                               "He 5-10 older", "He 10+ older"))

# Create boxplot viz Kansas Missouri
ggplot(ks_mo_data, aes(x = age_gap_category, y = educ_gap, fill = age_gap_category)) +
  geom_boxplot() +
  labs(title = "Education Gap by Age Gap Category",
       x = "Age Gap Category",
       y = "Education Gap") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        legend.position = "none")
```



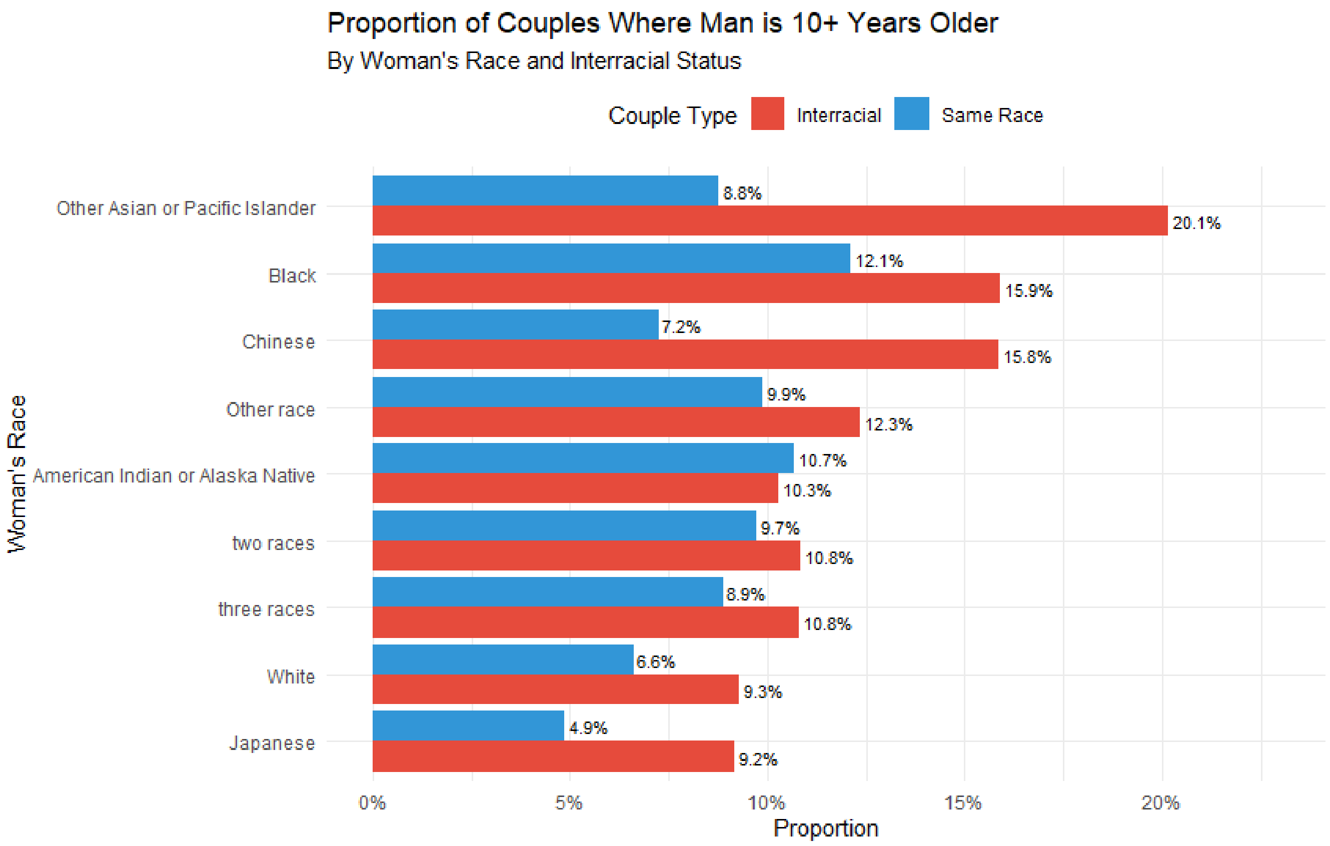
```
# The slightly higher proportion of 10+ year age gaps in Missouri probably isn't explained by education differences
# Marriage age laws prohibiting marriage under 18 might not be strongly related to actual age gaps

# Look into potential correlation between age gap and race (whole country, no longer looking at just Kansas and Missouri)

# See  distribution of races
table(trad_data$RACE)

# Calculate proportion of 10+ year age gaps by woman's race
age_gap_by_race <- trad_data %>%
  group_by(RACE) %>%
  summarize(
    prop_10plus_gap = mean(he_more_than_10yrs_than_her),
    n = n()
  ) %>%
  arrange(desc(prop_10plus_gap))
```

```
print(age_gap_by_race)

# Create Viz
ggplot(age_gap_by_race, aes(x = reorder(RACE, prop_10plus_gap), y = prop_10plus_gap, fill =
RACE)) +
  geom_col() +
  geom_text(aes(label = paste0(round(prop_10plus_gap * 100, 1), "%")),
            hjust = -0.2, size = 3) +
  coord_flip() +
  labs(title = "Proportion of Couples Where Man is 10+ Years Older",
       subtitle = "By Woman's Race (US-wide)",
       x = "Woman's Race",
       y = "Proportion") +
  scale_y_continuous(labels = scales::percent) +  # No limits now
  theme_minimal() +
  theme(legend.position = "none")
```



```
#Key finding: Black women and Asian/Pacific Islander women have notably higher rates of large age
gaps (12-13%) compared to White women (6.8%) and Japanese women (7.7%).

# Look at interracial couples
# Create variable whether couple is same race
```

```r
trad_data$same_race <- as.numeric(trad_data$RACE == trad_data$h_race)

# Compare age gaps in same-race vs different-race couples
interracial_comparison <- trad_data %>%
  group_by(same_race) %>%
  summarize(
    prop_10plus_gap = mean(he_more_than_10yrs_than_her),
    n = n()
  )

print(interracial_comparison)

# same_race prop_10plus_gap      n
# <dbl>           <dbl>  <int>
# 1        0          0.121   54628
# 2        1          0.0734 357647

# Are Black women's higher rates driven by being in interracial relationships more often? Or do Black women in same-race couples also have higher age gaps?

# Break down by race AND interracial status
race_interracial <- trad_data %>%
  mutate(couple_type = ifelse(same_race == 1, "Same Race", "Interracial")) %>%
  group_by(RACE, couple_type) %>%
  summarize(
    prop_10plus_gap = mean(he_more_than_10yrs_than_her),
    n = n()
  ) %>%
  filter(n >= 100)  # Only show groups with at least 100 observations

print(race_interracial)

# RACE                             couple_type prop_10plus_gap      n
# <fct>                           <chr>                <dbl>  <int>
# 1 White                          Interracial         0.0929  17335
# 2 White                          Same Race           0.0661 277275
# 3 Black                          Interracial         0.159    2285
# 4 Black                          Same Race           0.121   16967
# 5 American Indian or Alaska Native Interracial          0.103    1817
# 6 American Indian or Alaska Native Same Race           0.107    1955
# 7 Chinese                        Interracial         0.158    2374
# 8 Chinese                        Same Race           0.0724   6108
# 9 Japanese                       Interracial         0.0917   1320
# 10 Japanese                      Same Race           0.0488    717
# 11 Other Asian or Pacific Islander  Interracial         0.201    7207
# 12 Other Asian or Pacific Islander  Same Race           0.0877  19236
# 13 Other race                    Interracial         0.123    5325
# 14 Other race                    Same Race          0.0990  15973
# 15 two races                     Interracial        0.108   15639
# 16 two races                     Same Race          0.0974  18898
# 17 three races                   Interracial        0.108    1326
# 18 three races                   Same Race          0.0888    518

# Findings:

# Interracial effect largest for Asian women:
# Other Asian/Pacific Islander: 20.1% interracial vs 8.8% same-race (2.3x higher!)
# Chinese: 15.8% interracial vs 7.2% same-race (2.2x higher!)
# Japanese: 9.2% interracial vs 4.9% same-race (1.9x higher!)

# Black women have high rates regardless:
# Interracial: 15.9%
# Same-race: 12.1%
# Both are above average, but interracial is still higher

# White women show the pattern but less dramatically:
# Interracial: 9.3%
# Same-race: 6.6%

# American Indian/Alaska Native women are unique:
# Almost identical rates (10.3% vs 10.7%) - interracial status doesn't matter

# Visualization

# Create the visualization
ggplot(race_interracial, aes(x = reorder(RACE, prop_10plus_gap), y = prop_10plus_gap, fill =
couple_type)) +
  geom_col(position = "dodge") +
  geom_text(aes(label = paste0(round(prop_10plus_gap * 100, 1), "%")),
            position = position_dodge(width = 0.9),
            hjust = -0.1, size = 3) +
  coord_flip() +
  labs(title = "Proportion of Couples Where Man is 10+ Years Older",
       subtitle = "By Woman's Race and Interracial Status",
       x = "Woman's Race",
       y = "Proportion",
       fill = "Couple Type") +
  scale_y_continuous(labels = scales::percent, limits = c(0, 0.23)) +
  scale_fill_manual(values = c("Interracial" = "#E74C3C", "Same Race" = "#3498DB")) +
  theme_minimal() +
  theme(legend.position = "top")
```



Proportion of Couples Where Man is 10+ Years Older

By Woman's Race and Interracial Status