lab 8

Riyesh Nath and Jason Seda

2025-11-14

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.2
## ✓ lubridate 1.9.4     ✓ tibble    3.3.0
## ✓ purrr     1.1.0     ✓ tidyr     1.3.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
## Type 'citation("pROC")' for a citation.
##
##
## Attaching package: 'pROC'
##
##
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

In this homework, we attempt to see what parameters we can use to find if a given Male household head is 10 years of older than his Female partner.

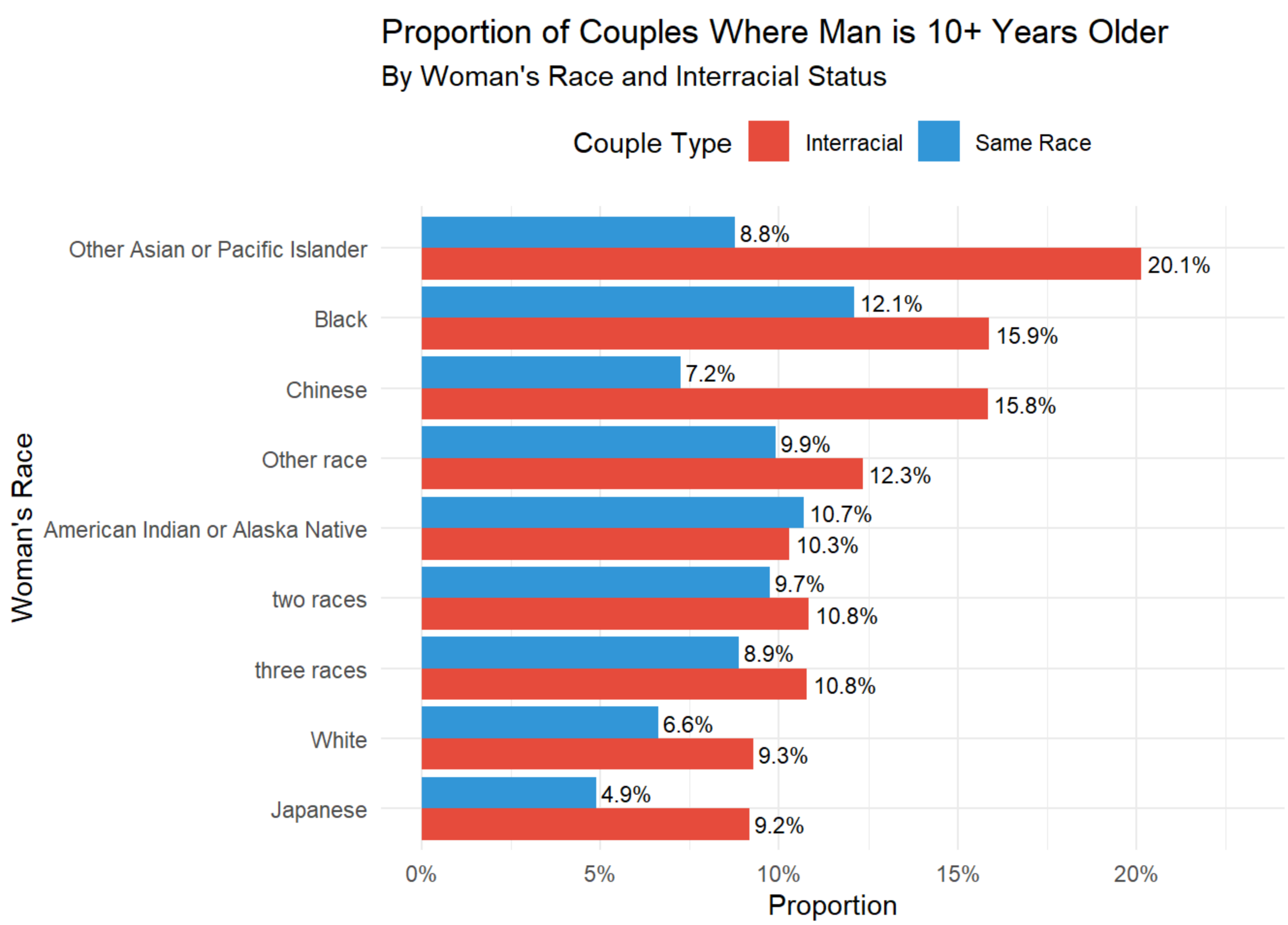
```
trad_data <- acs2021_couples %>% filter( (SEX == "Female") & (h_sex == "Male") )
trad_data$he_more_than_10yrs_than_her <- as.numeric(trad_data$age_diff <= -10)
```

We first check to see if interracial couple versus same race couple grouped by race has an affect on the proportion of women who are marrier to men older than 10 years. Looking at the plot below, we do see that race and couple type (interracial or same race) does have an affect.

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

```
## `summarise()` has grouped output by 'RACE'. You can override using the
## `.groups` argument.
```

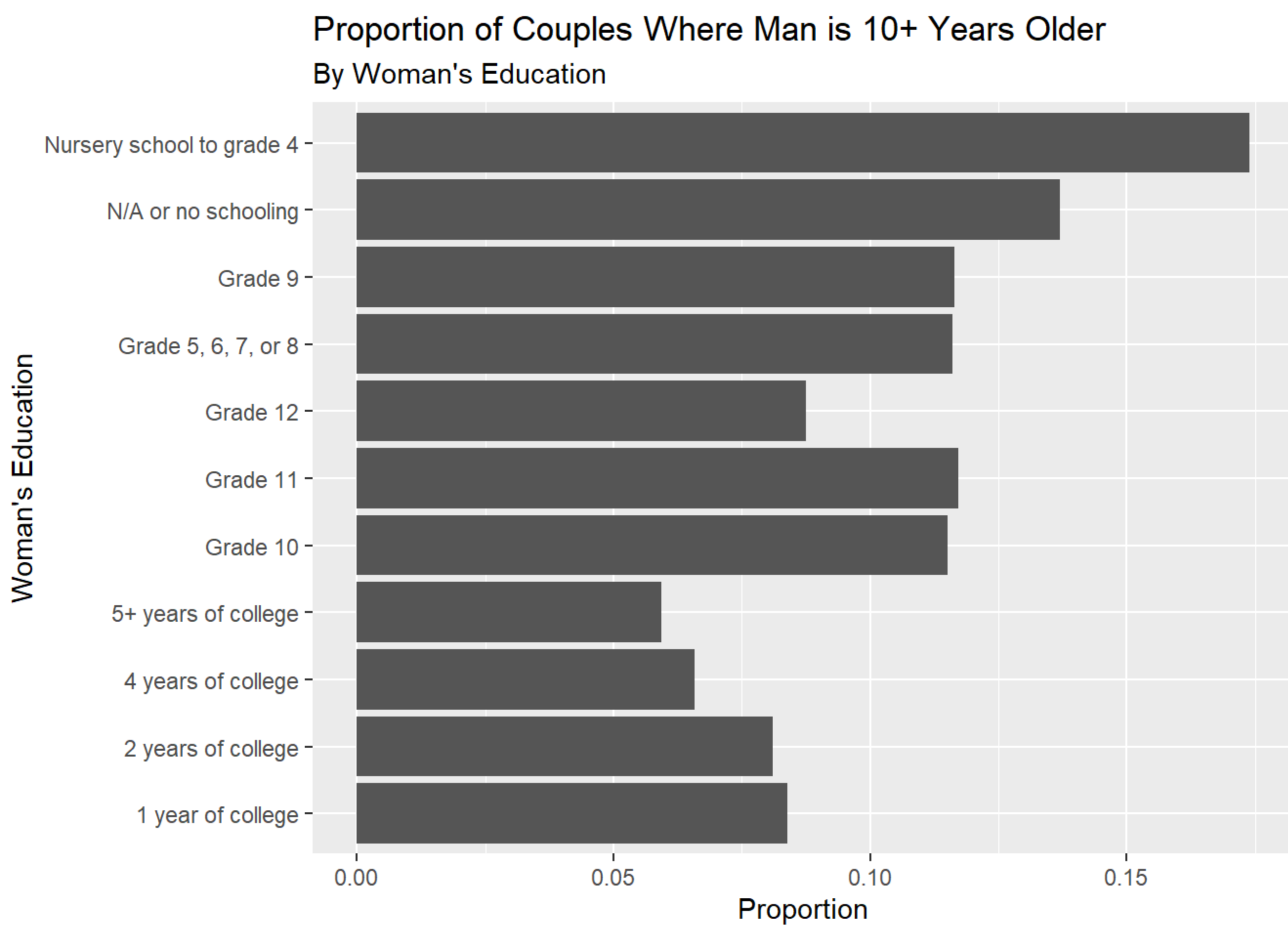
```
ggplot(race_interracial, aes(x = reorder(RACE, prop_10plus_gap), y = prop_10plus_gap, fill = cou
ple_type)) +
  geom_col(position = "dodge") +
  geom_text(aes(label = paste0(round(prop_10plus_gap * 100, 1), "%")),
            position = position_dodge(width = 0.9),
            hjust = -0.1, size = 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")
```



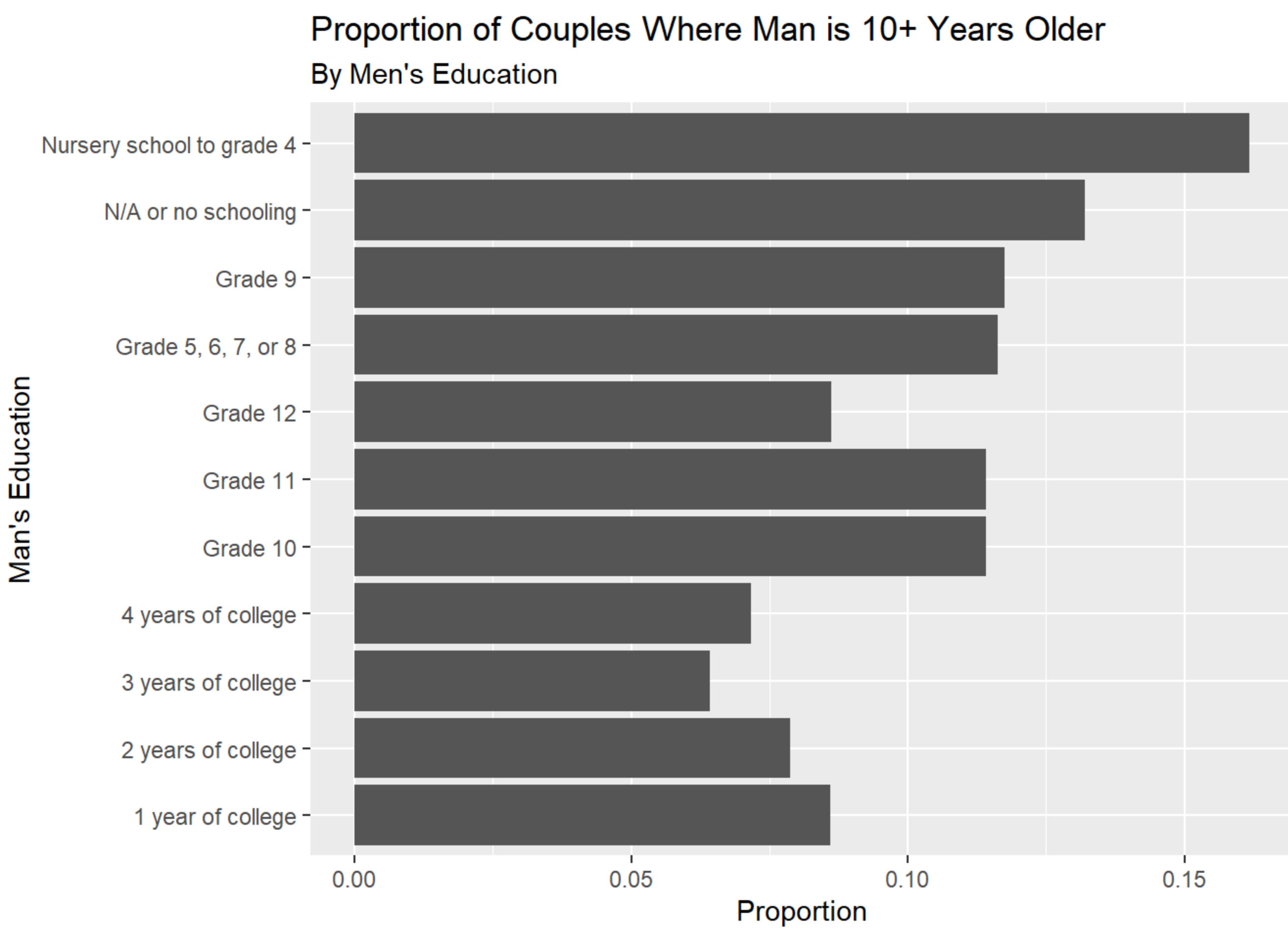
####

Looking at the proportion of partnership where men are older women based on education level of men and woman, we see similar pattern where we see a lower education level of both men and women affecting the age difference. Though the proportion isn't too significant.

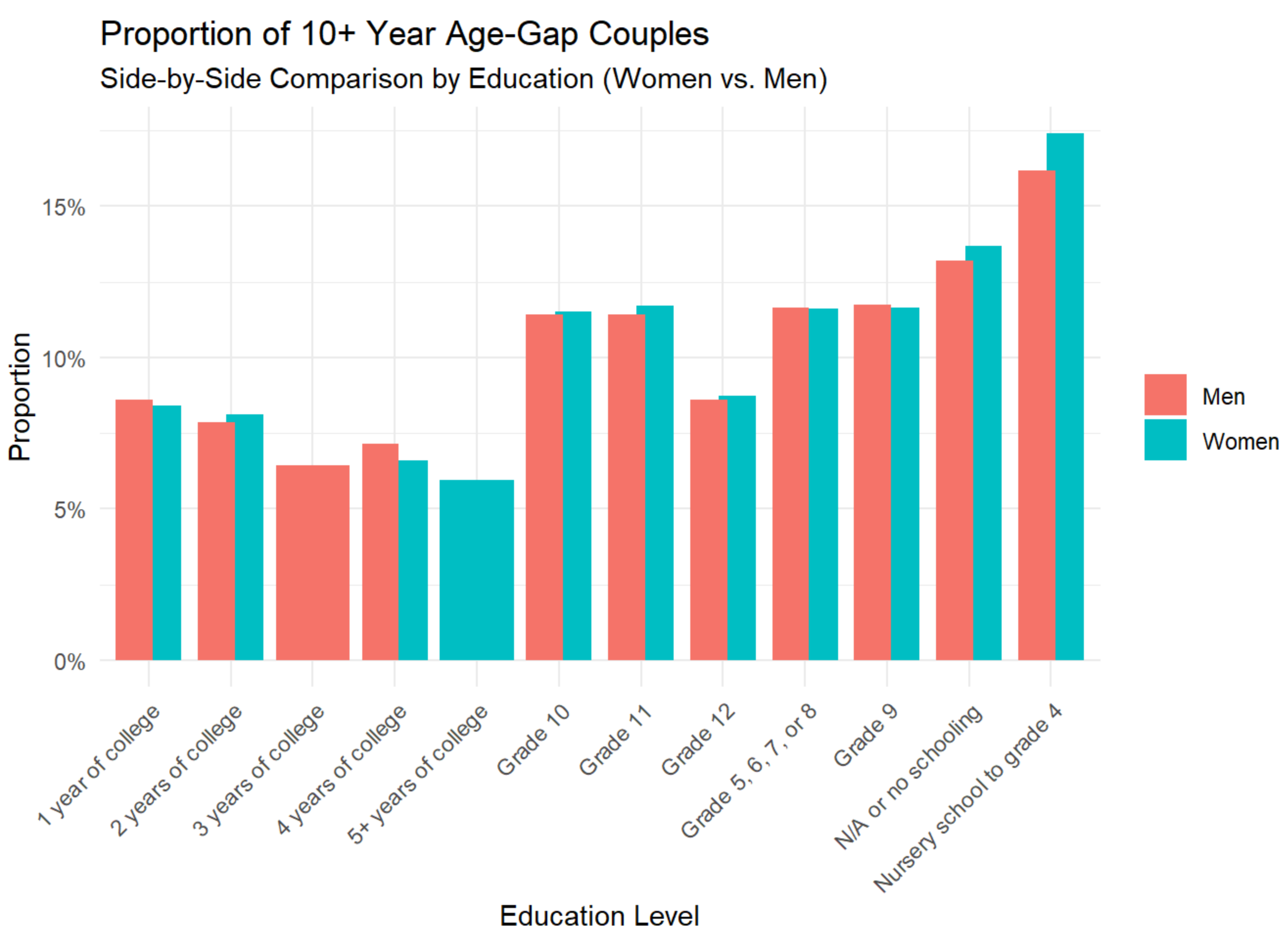
```
education_status_of_women_who_married_older_men_df <- as.data.frame.matrix(xtabs(data = trad_dat
a, formula = ~  EDUC + he_more_than_10yrs_than_her))
education_status_of_women_who_married_older_men_df <- rownames_to_column(education_status_of_wom
en_who_married_older_men_df, var="EDUC")

education_status_of_women_who_married_older_men_df$total <- education_status_of_women_who_marrie
d_older_men_df$`0` + education_status_of_women_who_married_older_men_df$`1`
education_status_of_women_who_married_older_men_df$prop <- education_status_of_women_who_married
_older_men_df$`1` / education_status_of_women_who_married_older_men_df$total
education_status_of_women_who_married_older_men_df <- education_status_of_women_who_married_olde
r_men_df %>%
  filter(!is.na(prop))

ggplot(data= education_status_of_women_who_married_older_men_df, aes(x = prop, y = EDUC)) +
  geom_col() +
  labs(title = "Proportion of Couples Where Man is 10+ Years Older",
       subtitle = "By Woman's Education",
       y = "Woman's Education",
       x = "Proportion")
```



```
education_status_of_men_who_married_younger_women_df <- as.data.frame.matrix(xtabs(data = trad_d
ata, formula = ~  h_educ + he_more_than_10yrs_than_her))
education_status_of_men_who_married_younger_women_df <- rownames_to_column(education_status_of_m
en_who_married_younger_women_df, var="h_educ")

education_status_of_men_who_married_younger_women_df$total <- education_status_of_men_who_marrie
d_younger_women_df$`0` + education_status_of_men_who_married_younger_women_df$`1`
education_status_of_men_who_married_younger_women_df$prop <- education_status_of_men_who_married
_younger_women_df$`1` / education_status_of_men_who_married_younger_women_df$total
education_status_of_men_who_married_younger_women_df <- education_status_of_men_who_married_youn
ger_women_df %>%
  filter(!is.na(prop))

ggplot(data= education_status_of_men_who_married_younger_women_df, aes(x = prop, y = h_educ)) +
  geom_col() +
  labs(title = "Proportion of Couples Where Man is 10+ Years Older",
       subtitle = "By Men's Education",
       y = "Man's Education",
       x = "Proportion")
```



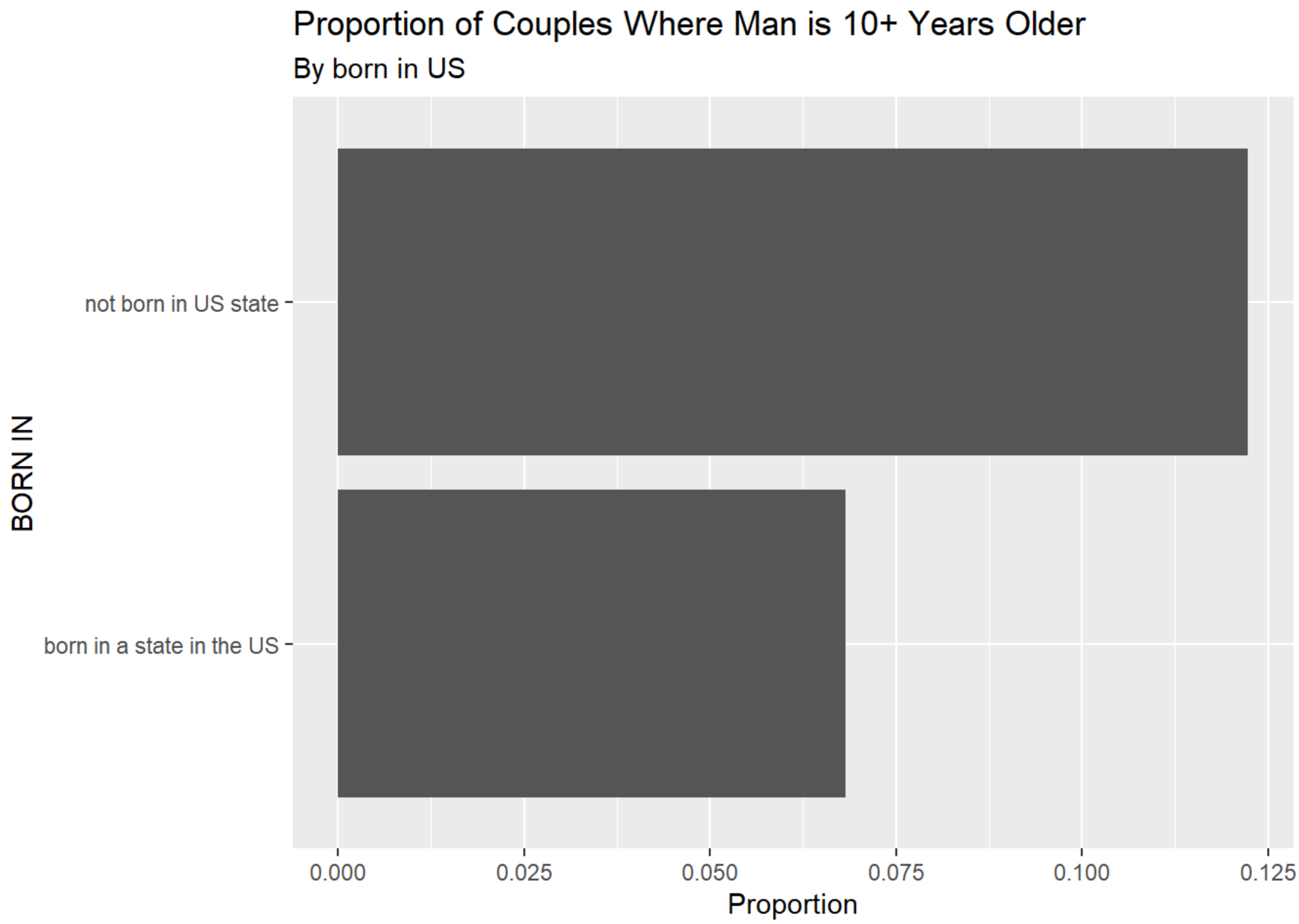
```
women_df <- education_status_of_women_who_married_older_men_df %>%
  mutate(group = "Women", education = EDUC) %>%
  select(group, education, prop)

men_df <- education_status_of_men_who_married_younger_women_df %>%
  mutate(group = "Men", education = h_educ) %>%
  select(group, education, prop)

combined_df <- bind_rows(women_df, men_df)
ggplot(combined_df, aes(x = education, y = prop, fill = group)) +
  geom_col(position = position_dodge(width = 0.7)) +
  scale_y_continuous(labels = scales::percent_format()) +
  labs(
    title = "Proportion of 10+ Year Age-Gap Couples",
    subtitle = "Side-by-Side Comparison by Education (Women vs. Men)",
    x = "Education Level",
    y = "Proportion",
    fill = ""
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1)
  )
```



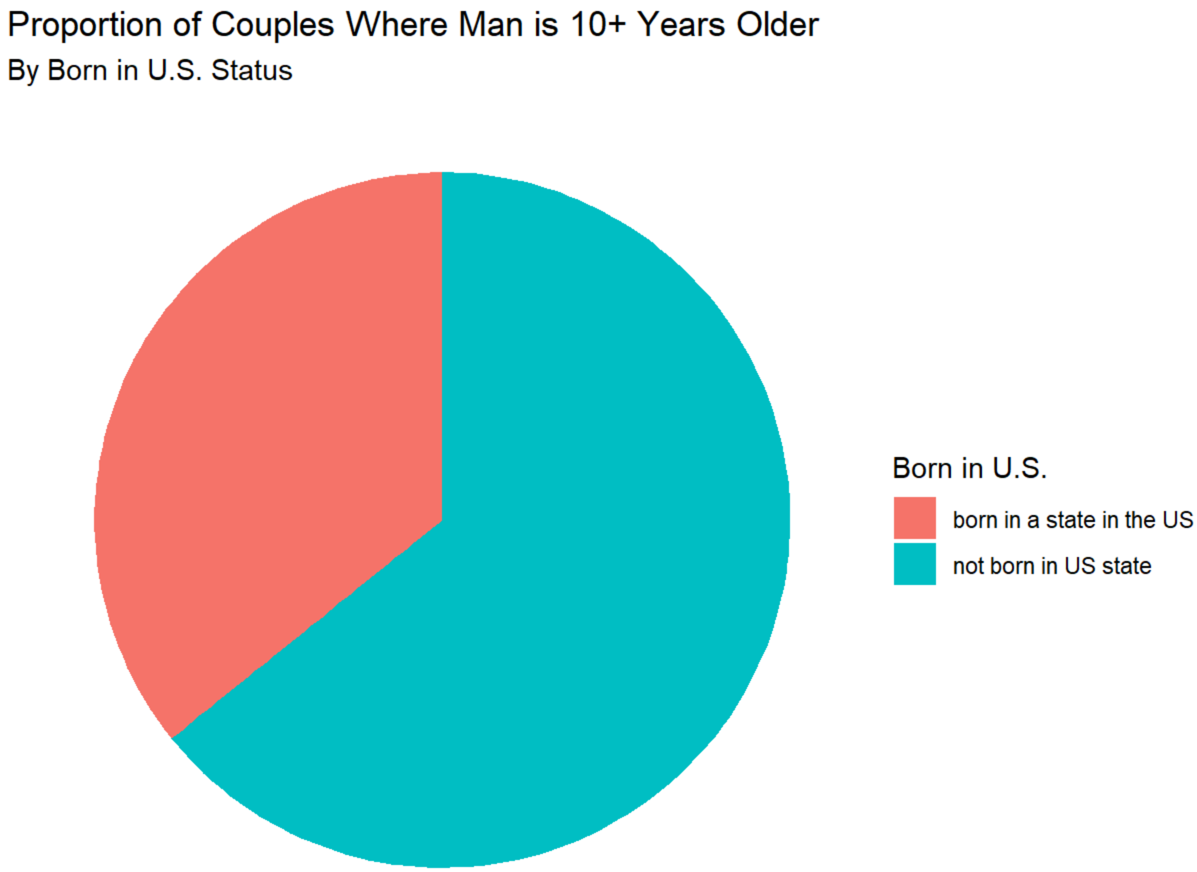
We also look at the proportion of those born in the US state against those not born in the US state. We see that a higher proportion of those not born in the US state have 10 years age gap between man and woman. However, here the proportion both both are still very low.

It is important to note, when isolating location (within US alone), there can be a clear favor for those not born in the US to have coupled with a significant older than 10+ years.

```
proportion_based_on_born_in_usa <- as.data.frame.matrix(xtabs(data= trad_data, formula = ~ born_in_USstate + he_more_than_10yrs_than_her))
proportion_based_on_born_in_usa <- rownames_to_column(proportion_based_on_born_in_usa, var="born_int_USstate")
proportion_based_on_born_in_usa$total <- proportion_based_on_born_in_usa$`0` + proportion_based_on_born_in_usa$`1`
proportion_based_on_born_in_usa$prop <- proportion_based_on_born_in_usa$`1` / proportion_based_on_born_in_usa$total

ggplot(data = proportion_based_on_born_in_usa, aes(y = born_int_USstate, x = prop)) +
  geom_col() +
  labs(
    title = "Proportion of Couples Where Man is 10+ Years Older",
       subtitle = "By born in US",
       y = "BORN IN",
       x = "Proportion"
  )
```



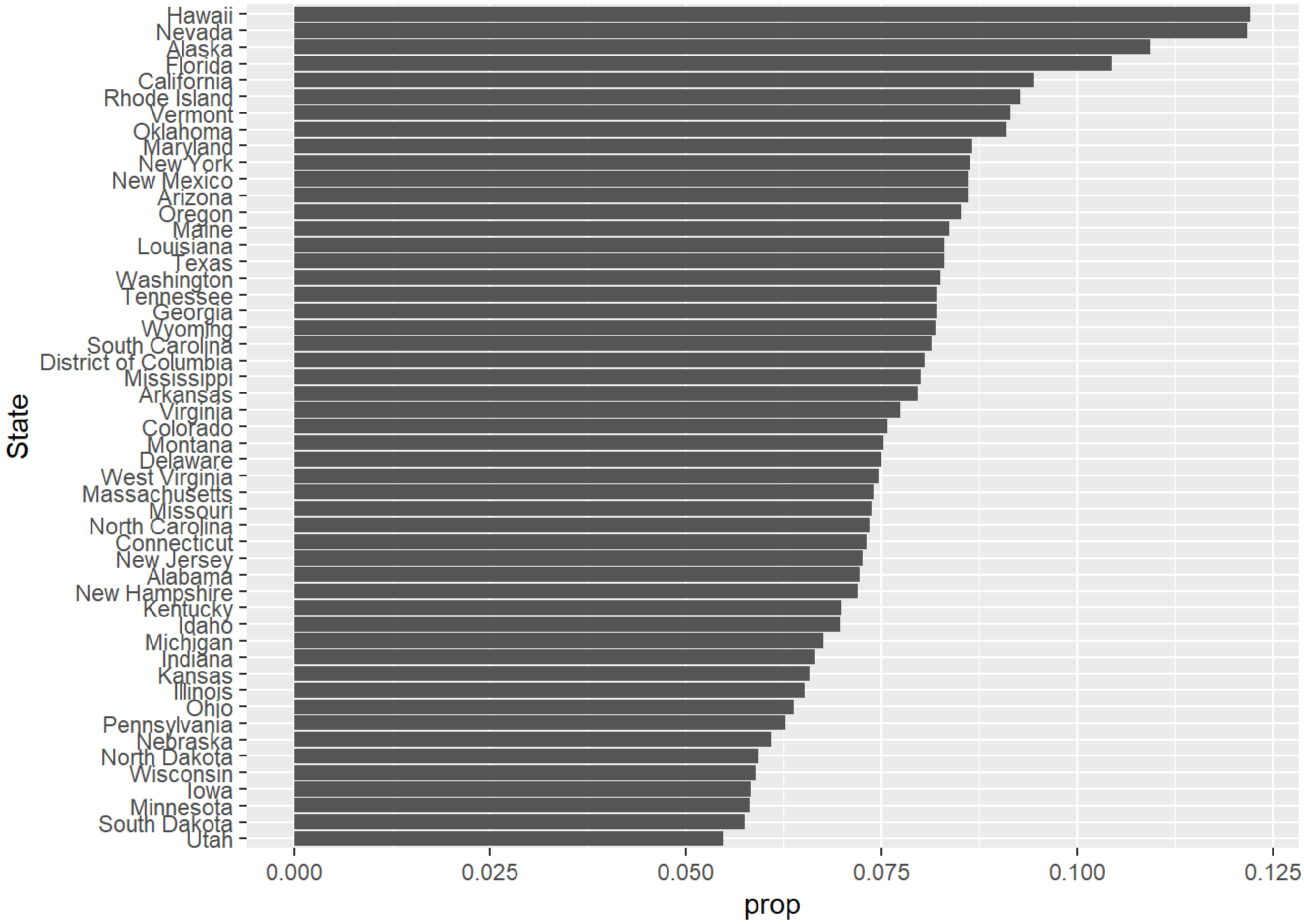

```
ggplot(proportion_based_on_born_in_usa, aes(x = "", y = prop, fill = born_int_USstate)) +
  geom_col(width = 1) +
  coord_polar(theta = "y") +
  labs(
    title = "Proportion of Couples Where Man is 10+ Years Older",
    subtitle = "By Born in U.S. Status",
    x = NULL,
    y = NULL,
    fill = "Born in U.S."
  ) +
  theme_void()
```



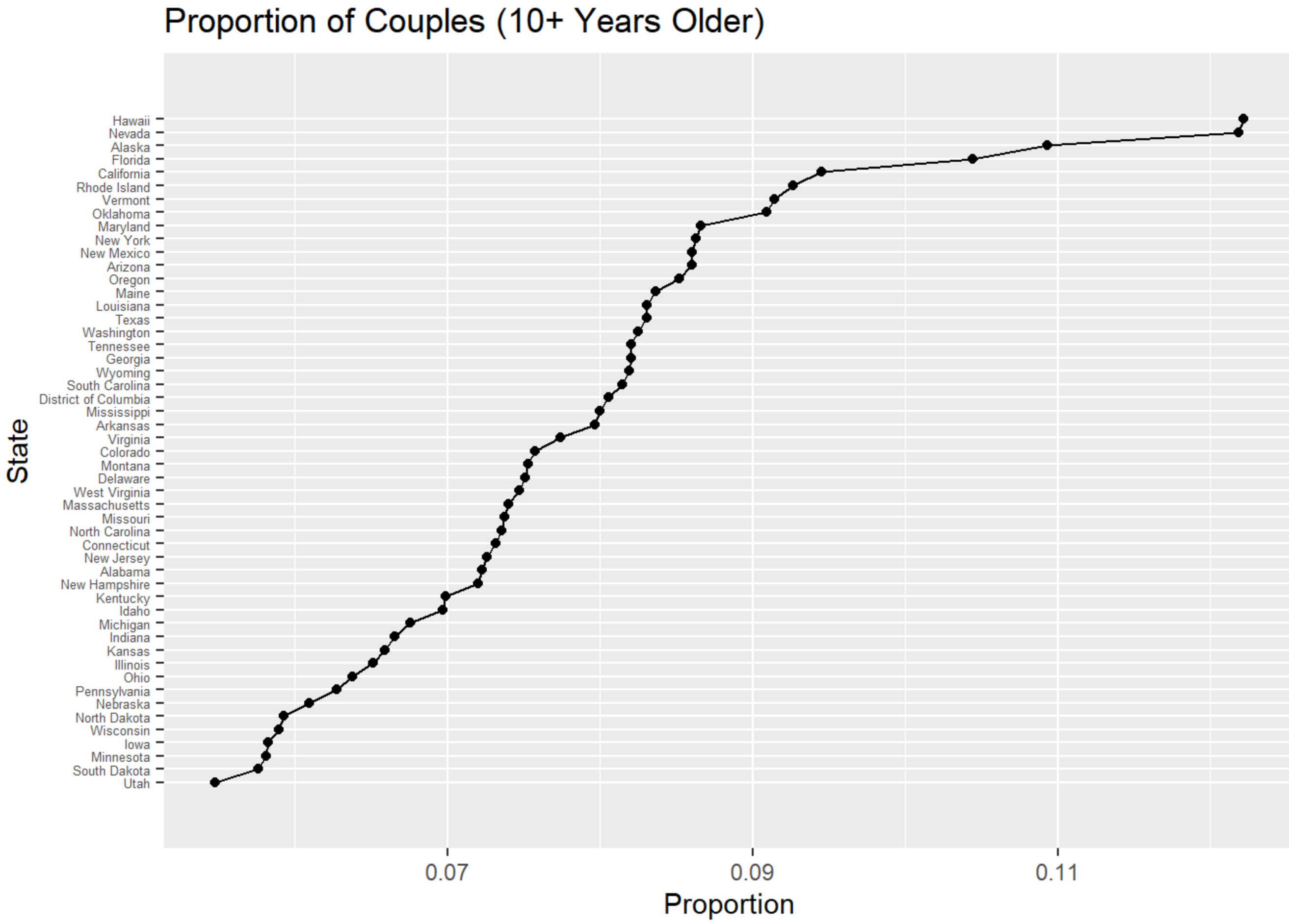
####

We also see that difference states may have an affect the proportions overall are very minor in the Majority of states. When reviewing state dependent, there are a few outlying states that could skew the graph in favor of larger age gap marriage, ie. seen as Hawaii, alaska, Nevada, and Florida with over 10% proportionally. Eliminating these states in particular from our statistical data may prove to unify the overall resulting consensus that the probability of a couple sub 10+ gap is siginificantly low. The opposite can hold true as well, by eliminating all other states asides from Hawaii, Alaska, Nevada and Florida could heavily influence the mean proportion towards a higher likelihood.

```
by_state <- as.data.frame.matrix(xtabs(data= trad_data, formula = ~ STATEFIP + he_more_than_10yr
s_than_her))

by_state <- rownames_to_column(by_state, var="State")
by_state$total <- by_state$`0` + by_state$`1`
by_state$prop <- by_state$`1` / by_state$total

by_state$State <- with(
  by_state,
  reorder(State, prop)
)

ggplot(data= by_state, aes(x= prop, y= State)) +
  geom_col()
```

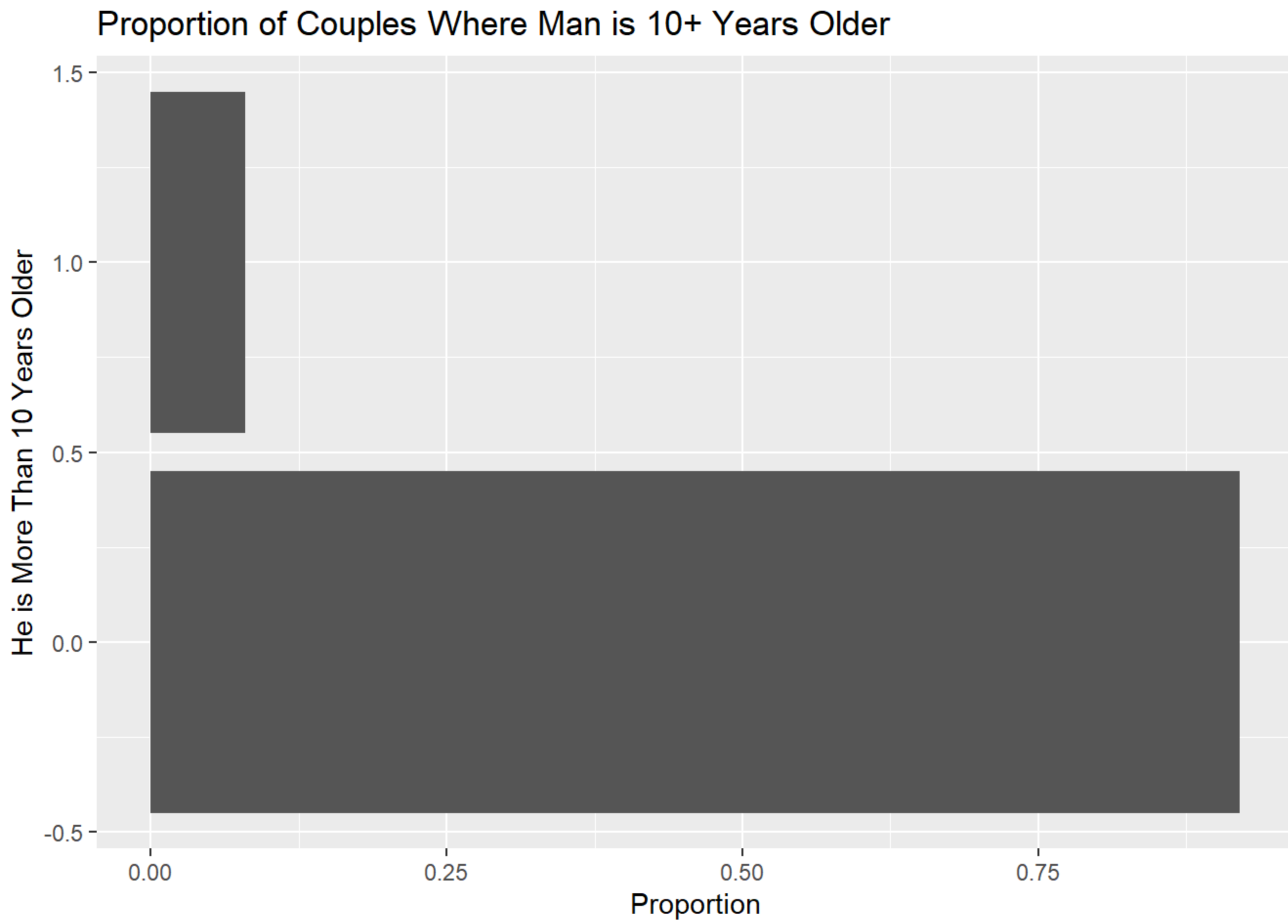



```
ggplot(data = by_state, aes(x = prop, y = State, group = 1)) +
  geom_line() +
  geom_point() +
  theme(axis.text.y = element_text(size = 5))+
  scale_y_discrete(expand = expansion(mult = 0.1)) +
  labs(
    title = "Proportion of Couples (10+ Years Older)",
    x = "Proportion",
    y = "State"
  )
```



The overall trend that we see with low proportion we see in all the graphs above may be attributed to overall low percent of Men who have married women 10 years or younger.

```
ggplot(trad_data, aes(y = he_more_than_10yrs_than_her)) +
  geom_bar(aes(x = after_stat(prop))) +
  labs(
    title = "Proportion of Couples Where Man is 10+ Years Older",
    x = "Proportion",
    y = "He is More Than 10 Years Older"
  )
```




```
logistic_modal <- glm(data = trad_data, formula = he_more_than_10yrs_than_her ~ same_race + RACE
+ h_race + EDUC + STATEFIP , family = binomial)

predict_val <- predict(logistic_modal, trad_data, type = "response")

roc_obj <- roc(trad_data$he_more_than_10yrs_than_her, predict_val)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
best_thr <- coords(roc_obj, "best", ret = "threshold")
print(best_thr)
```

```
##    threshold
## 1 0.08011561
```

```
summary(logistic_modal)
```

```
##
## Call:
## glm(formula = he_more_than_10yrs_than_her ~ same_race + RACE +
##     h_race + EDUC + STATEFIP, family = binomial, data = trad_data)
##
## Coefficients:
##                                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)                            -1.824771   0.064081 -28.476  < 2e-16
## same_race                              -0.380026   0.018079 -21.020  < 2e-16
## RACEBlack                               0.413198   0.041159  10.039  < 2e-16
## RACEAmerican Indian or Alaska Native    0.116098   0.062569   1.856 0.063521
## RACEChinese                             0.697289   0.055675  12.524  < 2e-16
## RACEJapanese                           -0.012198   0.090543  -0.135 0.892837
## RACEOther Asian or Pacific Islander     0.839300   0.032758  25.621  < 2e-16
## RACEOther race                          0.234575   0.037425   6.268 3.66e-10
## RACEtwo races                           0.210466   0.023572   8.929  < 2e-16
## RACEthree races                         0.205267   0.080824   2.540 0.011095
## h_raceBlack                             0.199137   0.039581   5.031 4.88e-07
## h_raceAmerican Indian or Alaska Native  0.047707   0.064169   0.743 0.457207
## h_raceChinese                          -0.624120   0.066097  -9.443  < 2e-16
## h_raceJapanese                         -0.337639   0.119450  -2.827 0.004704
## h_raceOther Asian or Pacific Islander  -0.512169   0.037967 -13.490  < 2e-16
## h_raceOther race                       -0.104633   0.039673  -2.637 0.008354
## h_racetwo races                        -0.090637   0.023551  -3.849 0.000119
## h_racethree races                      -0.349797   0.088308  -3.961 7.46e-05
## EDUCNursery school to grade 4           0.294381   0.072115   4.082 4.46e-05
## EDUCGrade 5, 6, 7, or 8                -0.124466   0.050472  -2.466 0.013663
## EDUCGrade 9                            -0.107385   0.065239  -1.646 0.099760
## EDUCGrade 10                           -0.124283   0.061196  -2.031 0.042264
## EDUCGrade 11                           -0.109382   0.059192  -1.848 0.064616
## EDUCGrade 12                           -0.394270   0.036882 -10.690  < 2e-16
## EDUC1 year of college                  -0.466669   0.039461 -11.826  < 2e-16
## EDUC2 years of college                 -0.481604   0.040284 -11.955  < 2e-16
## EDUC4 years of college                 -0.737469   0.038044 -19.385  < 2e-16
## EDUC5+ years of college                -0.858901   0.039525 -21.731  < 2e-16
## STATEFIPAlaska                          0.403022   0.129868   3.103 0.001914
## STATEFIPArizona                         0.176937   0.062147   2.847 0.004412
## STATEFIPArkansas                        0.095350   0.078416   1.216 0.224004
## STATEFIPCalifornia                      0.210989   0.052498   4.019 5.85e-05
## STATEFIPColorado                        0.125403   0.065926   1.902 0.057148
## STATEFIPConnecticut                     0.105209   0.075873   1.387 0.165551
## STATEFIPDelaware                        0.068778   0.118620   0.580 0.562036
## STATEFIPDistrict of Columbia            0.192082   0.158375   1.213 0.225194
## STATEFIPFlorida                         0.399524   0.053453   7.474 7.76e-14
## STATEFIPGeorgia                         0.111458   0.059799   1.864 0.062339
## STATEFIPHawaii                          0.444529   0.091576   4.854 1.21e-06
## STATEFIPIdaho                           0.018960   0.094568   0.200 0.841094
## STATEFIPIllinois                       -0.072646   0.058853  -1.234 0.217068
## STATEFIPIndiana                        -0.040918   0.065382  -0.626 0.531433
## STATEFIPIowa                           -0.147987   0.082102  -1.802 0.071470
## STATEFIPKansas                         -0.034245   0.082370  -0.416 0.677594
## STATEFIPKentucky                        0.034053   0.071514   0.476 0.633951
```

```
## STATEFIPLouisiana                     0.105645   0.071469   1.478 0.139357
## STATEFIPMaine                         0.272983   0.098202   2.780 0.005439
## STATEFIPMaryland                      0.193058   0.064163   3.009 0.002622
## STATEFIPMassachusetts                 0.131413   0.064399   2.041 0.041291
## STATEFIPMichigan                     -0.008786   0.060563  -0.145 0.884650
## STATEFIPMinnesota                    -0.128405   0.069337  -1.852 0.064040
## STATEFIPMississippi                   0.062487   0.082094   0.761 0.446565
## STATEFIPMissouri                      0.079033   0.065490   1.207 0.227512
## STATEFIPMontana                       0.133862   0.111208   1.204 0.228704
## STATEFIPNebraska                     -0.116582   0.097042  -1.201 0.229616
## STATEFIPNevada                        0.454726   0.070691   6.433 1.25e-10
## STATEFIPNew Hampshire                 0.125489   0.101903   1.231 0.218152
## STATEFIPNew Jersey                    0.046147   0.060636   0.761 0.446621
## STATEFIPNew Mexico                    0.149218   0.090649   1.646 0.099744
## STATEFIPNew York                      0.230496   0.054766   4.209 2.57e-05
## STATEFIPNorth Carolina                0.040144   0.059441   0.675 0.499446
## STATEFIPNorth Dakota                 -0.134495   0.139497  -0.964 0.334973
## STATEFIPOhio                         -0.071264   0.059644  -1.195 0.232158
## STATEFIPOklahoma                      0.183968   0.071929   2.558 0.010539
## STATEFIPOregon                        0.215086   0.069349   3.101 0.001925
## STATEFIPPennsylvania                 -0.070724   0.058634  -1.206 0.227744
## STATEFIPRhode Island                  0.341445   0.107332   3.181 0.001467
## STATEFIPSouth Carolina                0.141063   0.067073   2.103 0.035453
## STATEFIPSouth Dakota                 -0.150556   0.138281  -1.089 0.276255
## STATEFIPTennessee                     0.170795   0.063116   2.706 0.006809
## STATEFIPTexas                         0.113671   0.053592   2.121 0.033919
## STATEFIPUtah                         -0.226586   0.083609  -2.710 0.006727
## STATEFIPVermont                       0.406227   0.130623   3.110 0.001871
## STATEFIPVirginia                      0.089346   0.060651   1.473 0.140722
## STATEFIPWashington                    0.152169   0.061072   2.492 0.012716
## STATEFIPWest Virginia                 0.103957   0.094092   1.105 0.269227
## STATEFIPWisconsin                    -0.123390   0.067996  -1.815 0.069574
## STATEFIPWyoming                       0.185812   0.142130   1.307 0.191096
##
## (Intercept)                          ***
## same_race                            ***
## RACEBlack                            ***
## RACEAmerican Indian or Alaska Native  .
## RACEChinese                          ***
## RACEJapanese
## RACEOther Asian or Pacific Islander  ***
## RACEOther race                       ***
## RACEtwo races                        ***
## RACEthree races                      *
## h_raceBlack                          ***
## h_raceAmerican Indian or Alaska Native
## h_raceChinese                        ***
## h_raceJapanese                       **
## h_raceOther Asian or Pacific Islander ***
## h_raceOther race                     **
## h_racetwo races                      ***
## h_racethree races                    ***
```

```
## EDUCNursery school to grade 4          ***
## EDUCGrade 5, 6, 7, or 8                *
## EDUCGrade 9                            .
## EDUCGrade 10                           *
## EDUCGrade 11                           .
## EDUCGrade 12                           ***
## EDUC1 year of college                  ***
## EDUC2 years of college                 ***
## EDUC4 years of college                 ***
## EDUC5+ years of college                ***
## STATEFIPAlaska                         **
## STATEFIPArizona                        **
## STATEFIPArkansas
## STATEFIPCalifornia                     ***
## STATEFIPColorado                       .
## STATEFIPConnecticut
## STATEFIPDelaware
## STATEFIPDistrict of Columbia
## STATEFIPFlorida                        ***
## STATEFIPGeorgia                        .
## STATEFIPHawaii                         ***
## STATEFIPIdaho
## STATEFIPIllinois
## STATEFIPIndiana
## STATEFIPIowa                           .
## STATEFIPKansas
## STATEFIPKentucky
## STATEFIPLouisiana
## STATEFIPMaine                          **
## STATEFIPMaryland                       **
## STATEFIPMassachusetts                  *
## STATEFIPMichigan
## STATEFIPMinnesota                      .
## STATEFIPMississippi
## STATEFIPMissouri
## STATEFIPMontana
## STATEFIPNebraska
## STATEFIPNevada                         ***
## STATEFIPNew Hampshire
## STATEFIPNew Jersey
## STATEFIPNew Mexico                     .
## STATEFIPNew York                       ***
## STATEFIPNorth Carolina
## STATEFIPNorth Dakota
## STATEFIPOhio
## STATEFIPOklahoma                       *
## STATEFIPOregon                         **
## STATEFIPPennsylvania
## STATEFIPRhode Island                   **
## STATEFIPSouth Carolina                 *
## STATEFIPSouth Dakota
## STATEFIPTennessee                      **
```

```
## STATEFIPTexas                          *
## STATEFIPUtah                           **
## STATEFIPVermont                        **
## STATEFIPVirginia
## STATEFIPWashington                     *
## STATEFIPWest Virginia
## STATEFIPWisconsin                      .
## STATEFIPWyoming
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 229160  on 412274  degrees of freedom
## Residual deviance: 224194  on 412197  degrees of freedom
## AIC: 224350
##
## Number of Fisher Scoring iterations: 5
```

```
auc(roc_obj)
```

```
## Area under the curve: 0.6071
```

```
pred <- as.integer(predict_val >= 0.08011561)
table(pred, trad_data$he_more_than_10yrs_than_her)
```

```
##
## pred      0      1
##    0 250679  16594
##    1 128757  16245
```