Keyla Pereira

- Lab 7

Nancy Reyes Soto, Irania Matos, Kerra Sinanan, Heidy Collado.

We decided to see the factors associated with a man being at least 10 years older than his female partner. We focus on traditional different-sex couples (female respondent and male partner) and define the dependent variable y=1 if the man is 10+ years older, and 0 otherwise; a tabulation of age difference confirms that this coding is correct.

We estimate several Linear Probability Models to examine how education, age, and location relate to this outcome. The first model uses education dummies for the woman and includes her age, the second uses approximate years of schooling for both partners, and the third treats both partners' education as factors. Two additional models add geographic controls using STATEFIP and REGION to test whether location helps explain variation in age-gap relationships.

Coefficients in these models reflect changes in probability (0–1), so multiplying by 100 gives percentage-point effects. Because y=1 indicates the man is much older, positive coefficients increase the likelihood of a large age gap. I also test whether geographic variables matter jointly, and rejection of the null indicates meaningful regional or state-level differences.

To support the regression results, I include figures showing how the probability varies by women's education and across states, demonstrating both educational and geographic patterns in age-gap relationships.

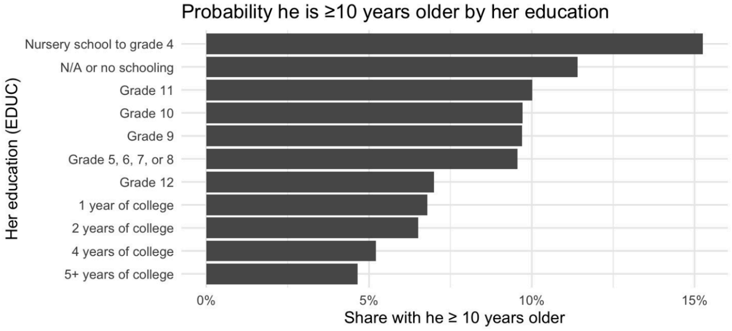

Figure 1: Probability he is ≥10 years older by her education

This figure shows that large age gaps are more common among women with lower levels of education, and the likelihood declines steadily as the woman's education increases.

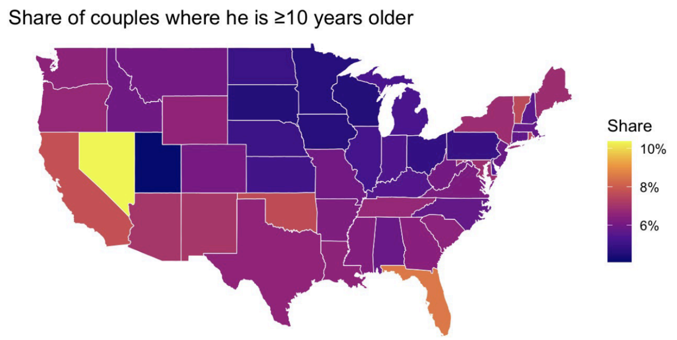

Figure 2: Share of couples where he is ≥10 years older (U.S. map)
This map illustrates geographic variation in age‑gap relationships, with several Western and Southern states showing higher rates and many Northeastern and Midwestern states showing lower rates.

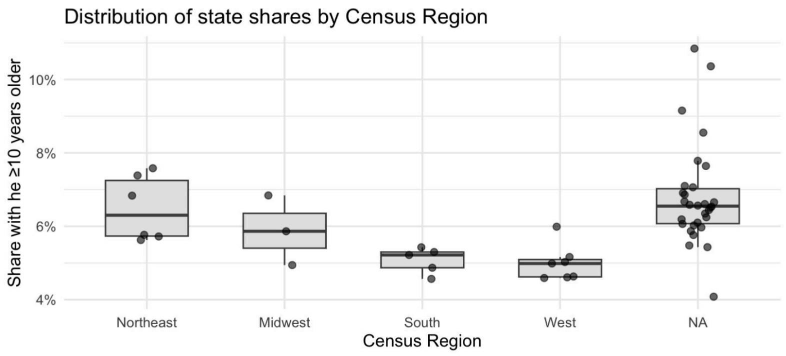

Figure 3: Distribution of state shares by Census Region
This boxplot summarizes differences across regions, indicating slightly higher rates of large age gaps in the South and West compared to the Midwest and Northeast.

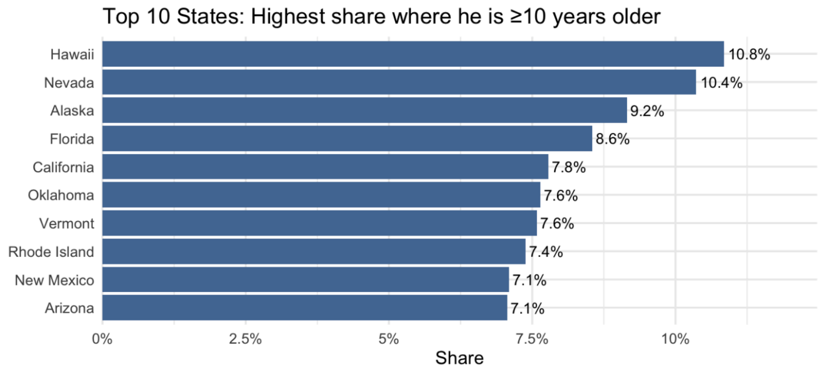

Figure 4: Top 10 states with the highest share where he is ≥10 years older
This ranked chart highlights the states with the largest age‑gap prevalence, with Hawaii, Nevada, Alaska, and Florida among the highest.

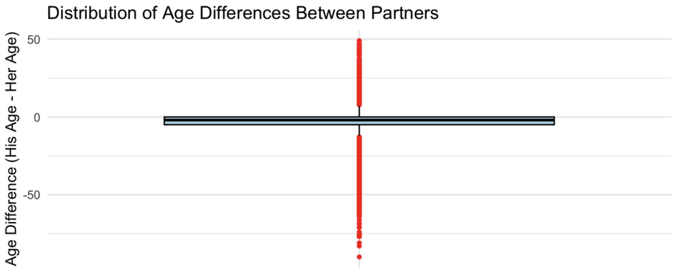

Figure 5: Distribution of Age Differences Between Partners
This boxplot summarizes the distribution of age differences between partners. The middle line inside the box represents the median age gap, while the box itself shows the interquartile range (IQR), which contains the middle 50% of couples. The whiskers extend to typical minimum and maximum values, and the red dots indicate outliers—couples with unusually large age differences in either direction. Overall, this plot shows that most couples have small age gaps, with a few extreme cases where one partner is much older.

```
load("/Users/keylapereira/Downloads/ACS_2021_couples.RData")

library(ggplot2)
library(tidyverse)
library(haven)
library(car)
library(sandwich)
library(lmtest)
library(maps)


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")


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)

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

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

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

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


trad_data <- trad_data %>%
  mutate(educ_numeric = as.numeric(EDUC),
         h_educ_numeric = as.numeric(h_educ))


ols_out2 <- lm(he_more_than_10yrs_than_her ~ educ_numeric + h_educ_numeric + AGE, data = trad_data)
summary(ols_out2)

ols_out3 <- lm(he_more_than_10yrs_than_her ~ factor(EDUC) + factor(h_educ) + AGE, data = trad_data)
summary(ols_out3)

ols_out4 <- lm(he_more_than_10yrs_than_her ~ educ_numeric + h_educ_numeric + AGE + factor(STATEFIP),
               data = trad_data)
summary(ols_out4)

ols_out5 <- lm(he_more_than_10yrs_than_her ~ educ_numeric + h_educ_numeric + AGE + factor(REGION),
               data = trad_data)
summary(ols_out5)


R_region <- grep("^factor\\(REGION\\)", names(coef(ols_out5)), value = TRUE)
car::linearHypothesis(ols_out5, R_region, vcov = vcovHC(ols_out5, type = "HC1"))

R_state <- grep("^factor\\(STATEFIP\\)", names(coef(ols_out4)), value = TRUE)
car::linearHypothesis(ols_out4, R_state, vcov = vcovHC(ols_out4, type = "HC1"))
```

```
plotA_df <- trad_data %>%
  mutate(educ_cat_f = factor(EDUC)) %>%
  group_by(educ_cat_f) %>%
  summarise(pct_he10 = mean(he_more_than_10yrs_than_her, na.rm = TRUE)) %>%
  ungroup()

ggplot(plotA_df, aes(x = reorder(educ_cat_f, pct_he10), y = pct_he10)) +
  geom_col() +
  coord_flip() +
  scale_y_continuous(labels = scales::percent) +
  labs(
    x = "Her education (EDUC)",
    y = "Share with he ≥ 10 years older",
    title = "Probability he is ≥10 years older by her education"
  ) +
  theme_minimal(base_size = 12)


state_rates <- trad_data %>%
  mutate(state_name = tolower(as.character(STATEFIP))) %>%
  group_by(state_name) %>%
  summarise(
    pct_he10 = mean(he_more_than_10yrs_than_her, na.rm = TRUE),
    n = n(),
    .groups = "drop"
  )

us_map <- ggplot2::map_data("state") %>% tibble::as_tibble()

map_df <- us_map %>%
  left_join(state_rates, by = c("region" = "state_name"))

ggplot(map_df, aes(long, lat, group = group, fill = pct_he10)) +
  geom_polygon(color = "white", linewidth = 0.2) +
  coord_fixed(1.3) +
  scale_fill_continuous(labels = function(x) paste0(round(x*100,1), "%"),
                        na.value = "grey90") +
  labs(
    title = "Share of couples where he is ≥10 years older",
    fill  = "Share"
  ) +
  theme_void(base_size = 12)


ggplot(map_df, aes(long, lat, group = group, fill = pct_he10)) +
  geom_polygon(color = "white", linewidth = 0.25) +
  coord_fixed(1.3) +
  scale_fill_viridis_c(
    option = "plasma",
    labels = function(x) paste0(round(x*100,1), "%"),
    na.value = "grey90"
  ) +
  labs(
    title = "Share of couples where he is ≥10 years older",
    fill  = "Share"
  ) +
  theme_void(base_size = 12)



state_rates_by_region <- trad_data %>%
  filter(!is.na(STATEFIP), !is.na(REGION)) %>%
```

```r
  group_by(STATEFIP, REGION) %>%
  summarise(
    pct_he10 = mean(he_more_than_10yrs_than_her, na.rm = TRUE),
    n = n(),
    .groups = "drop"
  )


region_labs <- c("1"="Northeast","2"="Midwest","3"="South","4"="West")
state_rates_by_region <- state_rates_by_region %>%
  mutate(REGION_L = if (is.numeric(REGION)) region_labs[as.character(REGION)]
         else if (is.factor(REGION)) region_labs[as.character(as.integer(REGION))]
         else region_labs[REGION]) %>%
  mutate(REGION_L = factor(REGION_L, levels = c("Northeast","Midwest","South","West")))


ggplot(state_rates_by_region, aes(x = REGION_L, y = pct_he10)) +
  geom_boxplot(width = 0.6, outlier.shape = NA, fill = "grey90", color = "grey30") +
  geom_jitter(width = 0.12, height = 0, alpha = 0.6, size = 2) +
  scale_y_continuous(labels = function(x) paste0(round(x*100), "%")) +
  labs(
    x = "Census Region",
    y = "Share with he ≥10 years older",
    title = "Distribution of state shares by Census Region"
  ) +
  theme_minimal(base_size = 12)

ggplot(top10_states, aes(x = State, y = pct_he10)) +
  geom_col(fill = "#4E79A7") +
  geom_text(aes(label = paste0(round(pct_he10*100,1), "%")),
            hjust = -0.1, size = 3.5) +
  coord_flip(clip = "off") +
  scale_y_continuous(labels = function(x) paste0(round(x*100,1), "%"),
                     expand = expansion(mult = c(0, .15))) +
  labs(
    title = "Top 10 States: Highest share where he is ≥10 years older",
    x = NULL,
    y = "Share"
  ) +
  theme_minimal(base_size = 12)


ggplot(trad_data, aes(x = "", y = age_diff)) +
  geom_boxplot(fill = "lightblue", color = "black", outlier.color = "red", outlier.shape = 16) +
  labs(
    title = "Distribution of Age Differences Between Partners",
    y = "Age Difference (His Age - Her Age)",
    x = ""
  ) +
  theme_minimal(base_size = 12)
```

The first article I chose is “CSR Actions, Brand Value, and Willingness to Pay a Premium Price for Luxury Brands: Does Long-Term Orientation Matter?” by Diallo et al. (2021). This study examines how corporate social responsibility (CSR) actions by luxury brands affect consumers’ willingness to pay higher prices. The authors used survey data from 1,049 participants in France and Tunisia and applied structural equation modeling to test how CSR influences brand value and purchase intentions. They found that CSR activities can sometimes lower willingness to pay, but this effect depends on cultural factors such as long-term orientation. In cultures with a strong future focus, CSR increases perceived brand value and willingness to pay more.

This second article looks at how the prices charged by online retailers influence how happy and loyal customers feel. The authors used data from BizRate.com customer reviews and compared prices from nine online stores. They applied a PLS model to see how price, price satisfaction, and delivery experiences are connected. The results show that higher prices make customers less satisfied with what they paid, and surprisingly, being satisfied with the price doesn’t always mean they’re happier with the delivery. Overall, the study suggests that simply offering low prices isn’t enough to keep customers loyal over time.