PREDOC Sample Data Task

George Rao

Table of contents

Load packages, import data, and clean variables

```
library(tidyverse)

raw_data <- read_csv("cps_women_lfp.csv")

# Reorder factors
data <- raw_data |>
  mutate_if(is_character, as_factor) |>
  mutate(
    education = fct_relevel(education, "< HS Diploma", "HS Diploma",
                            "Some college, no degree", "Associate's Degree",
                            "Bachelor's Degree", "Master's or Higher"),
    age = fct_relevel(age, "< 25", "25-34", "35-44", "45-54",
```

```
                       "55-64", "65-74", "75+"),
    wageinc_quantiles = fct_relevel(wageinc_quantiles, "0-19.99", "20-39.99",
                                    "40-59.99", "60-79.99", "80-100"),
    income_quantiles = fct_relevel(income_quantiles, "0-19.99", "20-39.99",
                                   "40-59.99", "60-79.99", "80-100")
  )
```

Part 1: Labor Force Participation

Create new variables and filter dataset

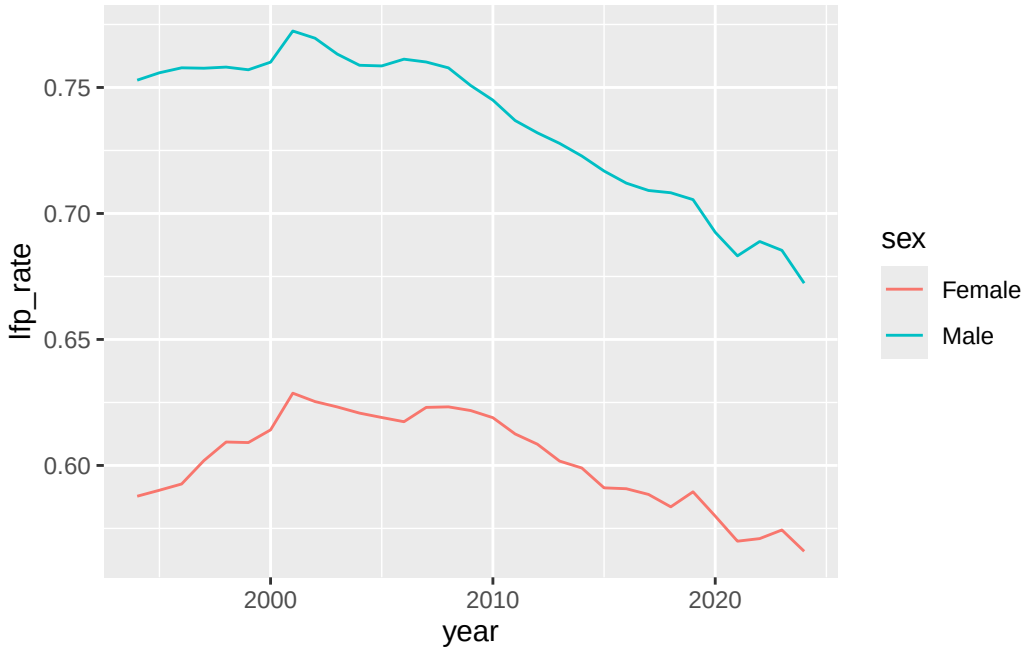
```
# Since LFP is the variable of interest, filter out NAs
all <- data |>
  filter(!is.na(lfp))

# Create new variables
all <- all |>
  mutate(
    lfp_lgl = lfp == "In labor force",
    college_lgl = college == "Has college degree",
    # This logical is true if we know the individual is self-employed,
    # but false otherwise, including if there is a missing value
    self_employed_lgl = if_else(self_employed == "Self-employed",
                                TRUE, FALSE, FALSE),
    lfp_lgl_excl_self = lfp_lgl & !self_employed_lgl
  )

# Create filtered data set of women only
women <- all |>
  filter(sex == "Female")

# Create filtered data set of women over 25 only
women_over_25 <- women |>
  filter(age != "< 25" & !is.na(age))
```

Question 1

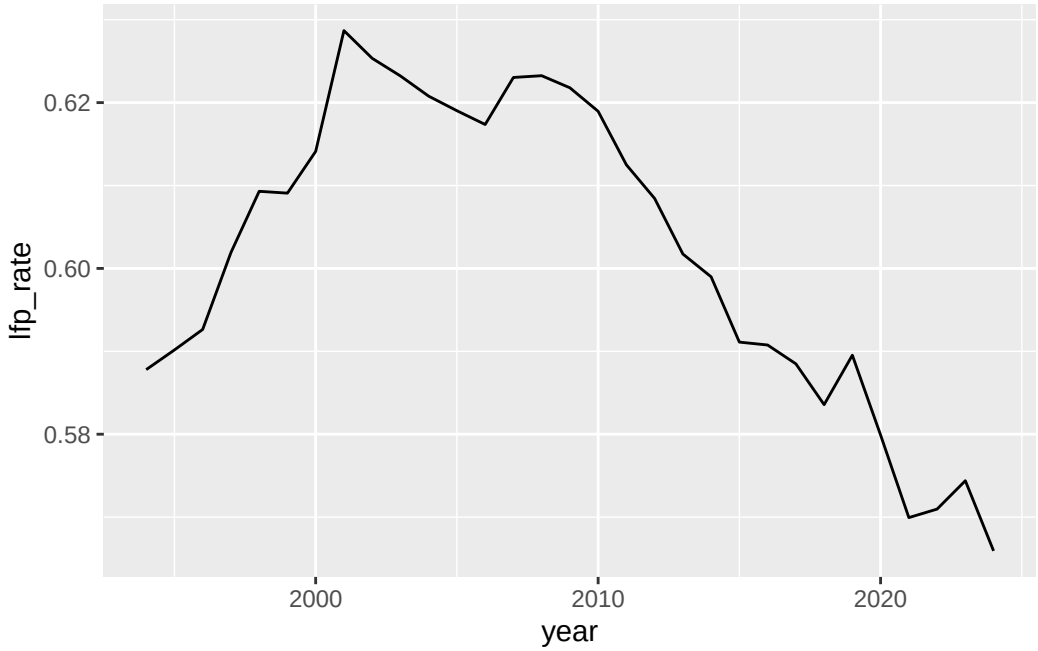
How has female labor force participation evolved since 1994? Please provide graphs and/or tables to support your answer.

Chart evolution of LFP over time

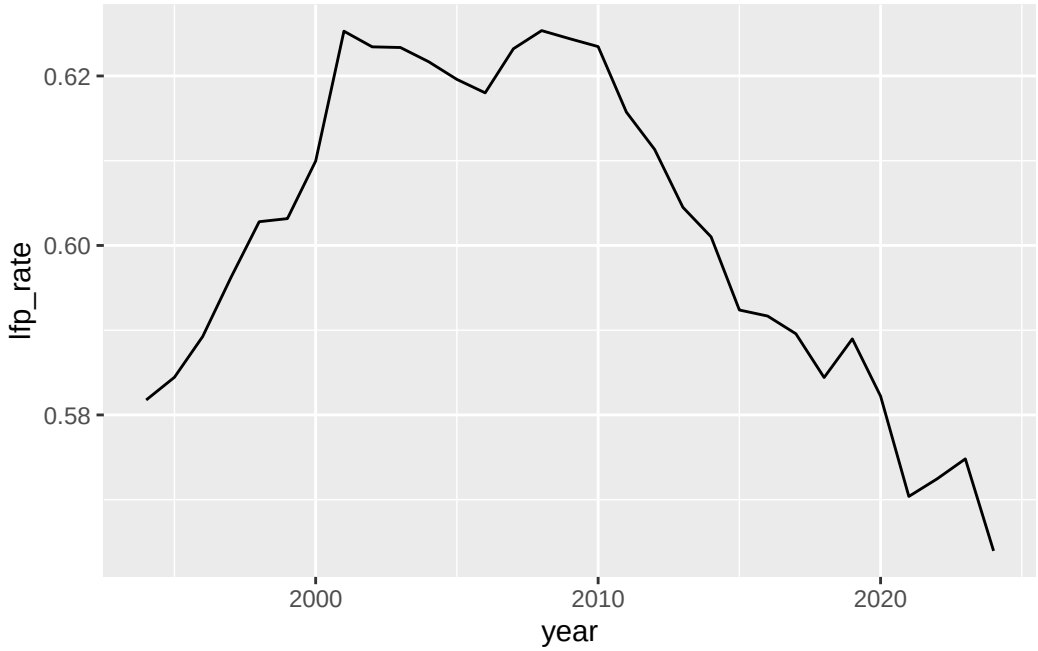
```
all |>
  summarize(
    .by = c(year, sex),
    lfp_rate = mean(lfp_lgl, na.rm = TRUE)
  ) |>
  ggplot(aes(x = year, y = lfp_rate, color = sex)) +
  geom_line()
```




```
women |>
  summarize(
    .by = year,
    lfp_rate = mean(lfp_lgl, na.rm = TRUE)
  ) |>
  ggplot(aes(x = year, y = lfp_rate)) +
  geom_line()
```




```
women_over_25 |>
  summarize(
    .by = year,
    lfp_rate = mean(lfp_lgl, na.rm = TRUE)
  ) |>
  ggplot(aes(x = year, y = lfp_rate)) +
  geom_line()
```



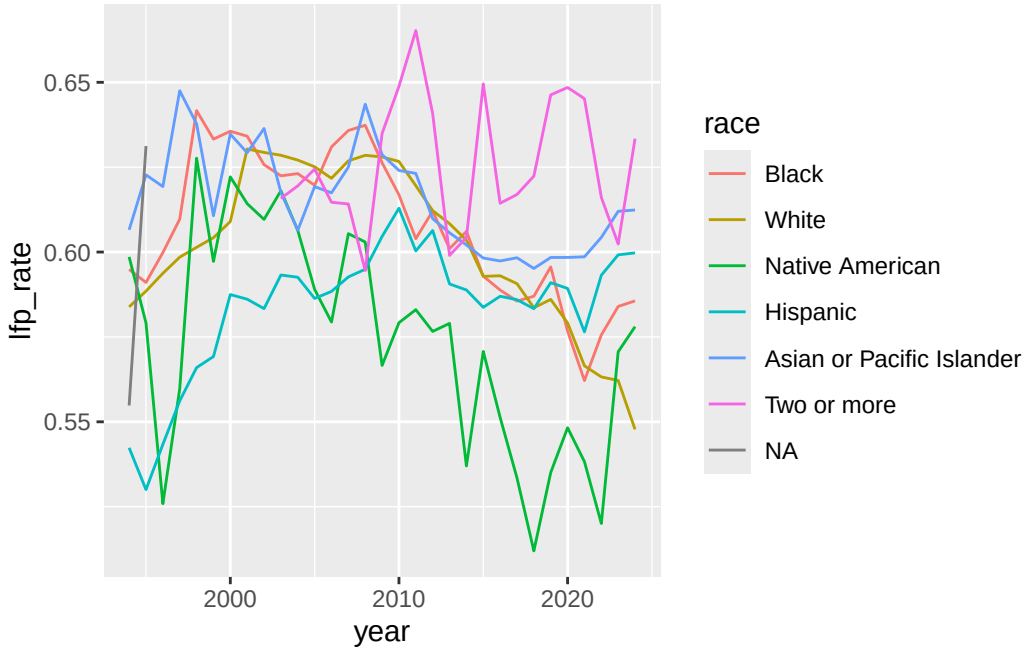
Female labor force participation increased from 1994 to the early 2000s, then has been falling ever since. At the same time, however, male labor force participation has fallen by even more since the early 2000s.

Question 2

> Among women older than 25, which groups (race, age, income percentile, etc.) of people had the biggest changes in labor force participation since 1994? Please provide at least three graphs and/or tables to support your answer.

Examine changes by race

```
women_over_25 |>
  summarize(
    .by = c(year, race),
    lfp_rate = mean(lfp_lgl, na.rm = TRUE)
  ) |>
  ggplot(aes(x = year, y = lfp_rate, color = race)) +
  geom_line()
```




```
women_over_25 |>
  summarize(
    .by = c(race),
    lfp_rate_1994 = mean(if_else(year == 1994, lfp_lgl, NA), na.rm = TRUE),
    lfp_rate_2024 = mean(if_else(year == 2024, lfp_lgl, NA), na.rm = TRUE),
    change = lfp_rate_2024 - lfp_rate_1994
  )
```

```
# A tibble: 7 x 4
  race                     lfp_rate_1994 lfp_rate_2024    change
  <fct>                            <dbl>         <dbl>     <dbl>
1 Black                            0.595         0.586  -0.00926
2 White                            0.584         0.548  -0.0361
3 Native American                  0.599         0.578  -0.0206
4 Hispanic                         0.542         0.600   0.0574
5 Asian or Pacific Islander        0.607         0.612   0.00584
6 <NA>                             0.555       NaN     NaN
7 Two or more                    NaN             0.633 NaN
```

The biggest changes by race were for the following groups:

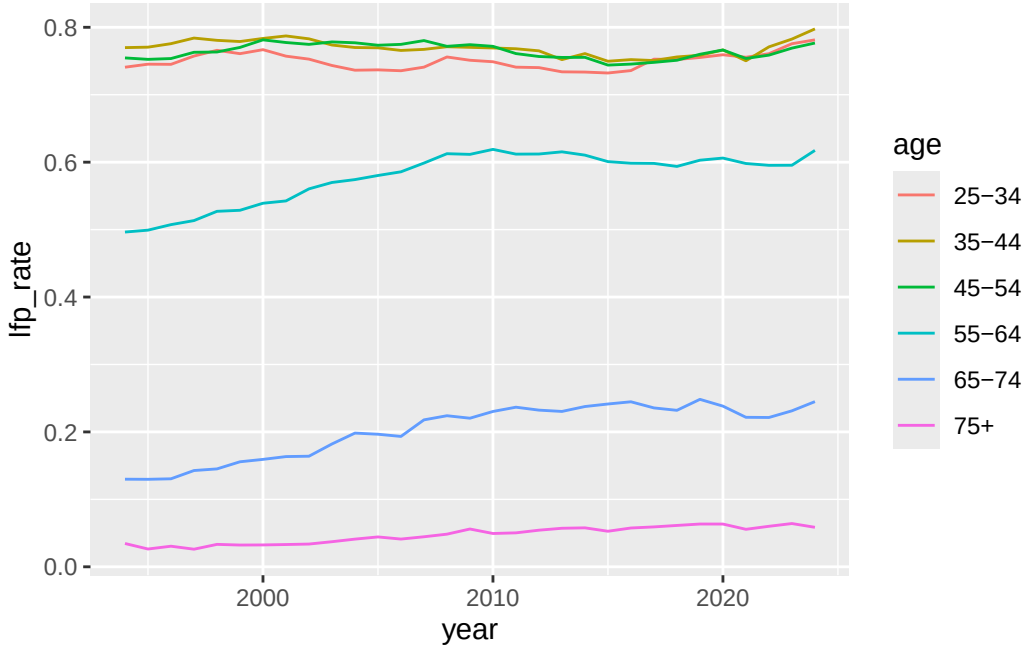
- White women experienced the biggest fall in LFP, by 3.6%.

- Hispanic women experienced the biggest rise in LFP, by 5.7%.

White women, being the largest group, have a similar rise-and-fall trend as the larger population.

Examine changes by age

```
women_over_25 |>
  summarize(
    .by = c(year, age),
    lfp_rate = mean(lfp_lgl, na.rm = TRUE)
  ) |>
  ggplot(aes(x = year, y = lfp_rate, color = age)) +
  geom_line()
```




```
women_over_25 |>
  summarize(
    .by = age,
    lfp_rate_1994 = mean(if_else(year == 1994, lfp_lgl, NA), na.rm = TRUE),
    lfp_rate_2024 = mean(if_else(year == 2024, lfp_lgl, NA), na.rm = TRUE),
    change = lfp_rate_2024 - lfp_rate_1994
```

```
  ) |>
  arrange(age)
```

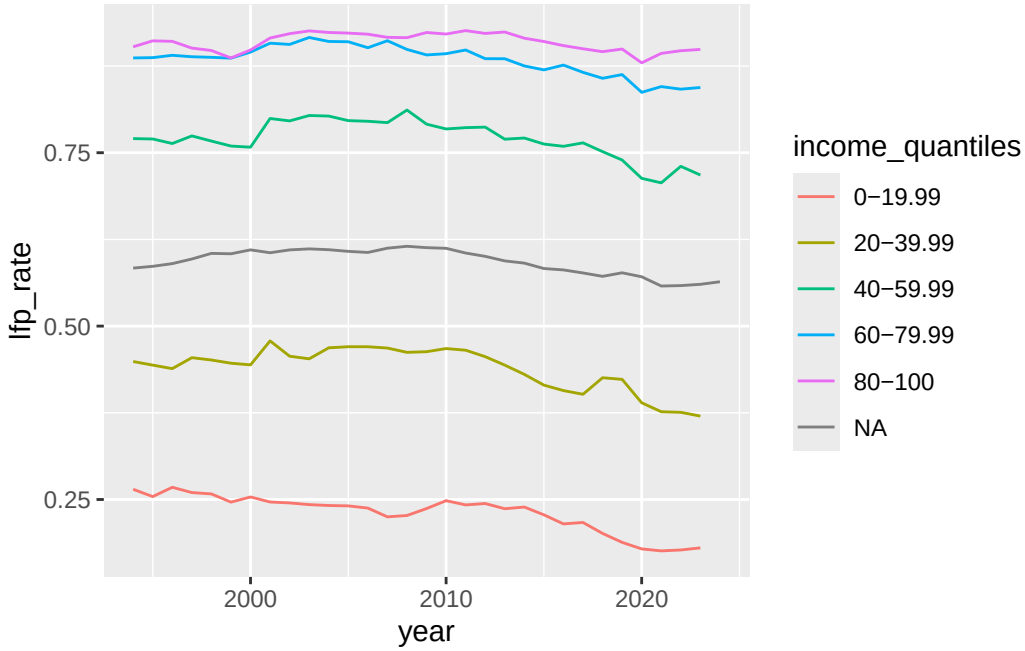
```
# A tibble: 6 x 4
  age   lfp_rate_1994 lfp_rate_2024 change
  <fct>         <dbl>         <dbl>  <dbl>
1 25-34        0.741         0.781  0.0406
2 35-44        0.770         0.798  0.0277
3 45-54        0.755         0.777  0.0221
4 55-64        0.496         0.617  0.121
5 65-74        0.130         0.245  0.115
6 75+          0.0345        0.0586 0.0240
```

The biggest changes by age were for the following groups:

- Women aged 55-64 experienced the biggest rise in LFP, by 12.1%.
- Women aged 65-74 also experienced a similarly big rise in LFP, by 11.5%.

Examine changes by income

```
women_over_25 |>
  summarize(
    .by = c(year, income_quantiles),
    lfp_rate = mean(lfp_lgl, na.rm = TRUE)
  ) |>
  ggplot(aes(x = year, y = lfp_rate, color = income_quantiles)) +
  geom_line()
```




```
women_over_25 |>
  summarize(
    .by = c(income_quantiles),
    lfp_rate_1994 = mean(if_else(year == 1994, lfp_lgl, NA), na.rm = TRUE),
    lfp_rate_2024 = mean(if_else(year == 2024, lfp_lgl, NA), na.rm = TRUE),
    # Since the 2024 data is not yet available, compare with the 2023 rate
    lfp_rate_2023 = mean(if_else(year == 2023, lfp_lgl, NA), na.rm = TRUE),
    change = lfp_rate_2023 - lfp_rate_1994
  )
```

```
# A tibble: 6 x 5
  income_quantiles lfp_rate_1994 lfp_rate_2024 lfp_rate_2023   change
  <fct>                    <dbl>         <dbl>         <dbl>    <dbl>
1 <NA>                     0.584         0.564         0.560 -0.0234
2 0-19.99                  0.265       NaN             0.180 -0.0845
3 20-39.99                 0.449       NaN             0.370 -0.0787
4 40-59.99                 0.770       NaN             0.718 -0.0525
5 60-79.99                 0.887       NaN             0.844 -0.0427
6 80-100                   0.903       NaN             0.899 -0.00380
```

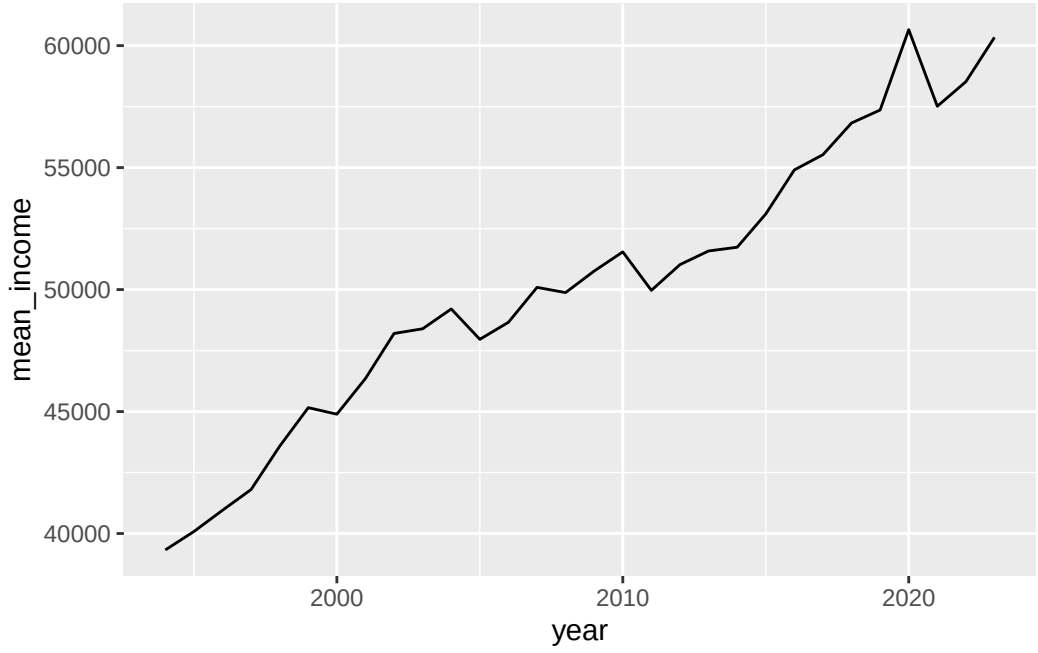
There is a clear trend: the lower a woman's income, the more her expected LFP drops from 1994 to 2023. The lowest quintile of female earners experienced an 8.5% drop in LFP, compared to the highest quintile, who experienced less than 1% of a drop.

Question 3

Use the data to examine trends among women older than 25 for each of the following factors from 1994 to 2024: (a) Wage and salary income (b) Social insurance income (c) Education attainment Based on these trends, what factors could be driving the patterns you found in Questions 1 and 2?

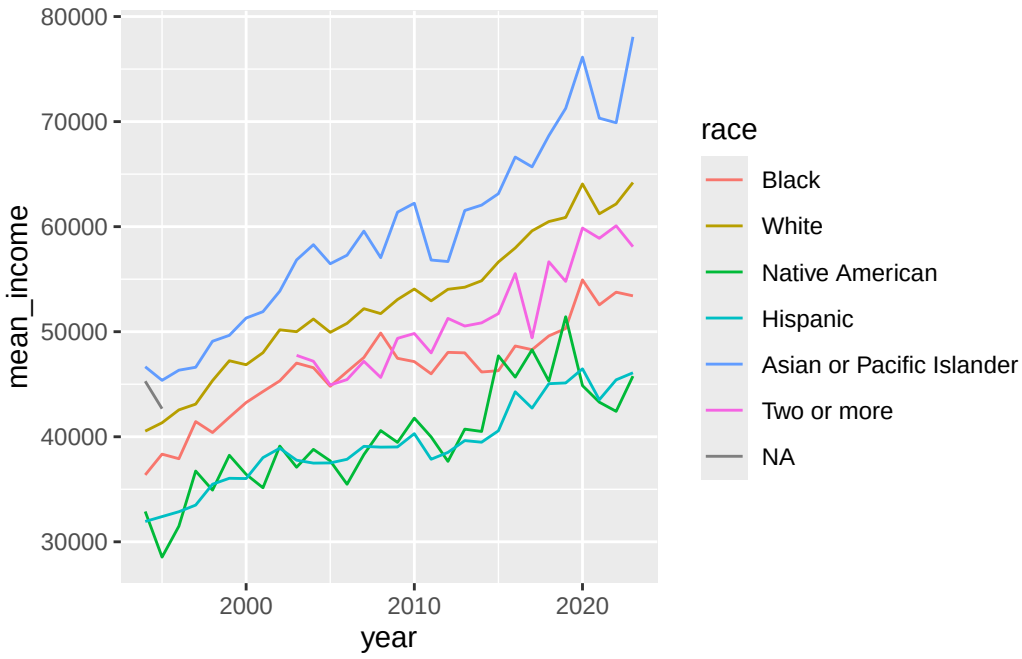
Examine wage and salary income per woman earning an income

```
women_over_25 |>
  filter(income != 0) |>
  summarize(
    .by = year,
    mean_income = mean(income, na.rm = TRUE)
  ) |>
  ggplot(aes(x = year, y = mean_income)) +
  geom_line()
```

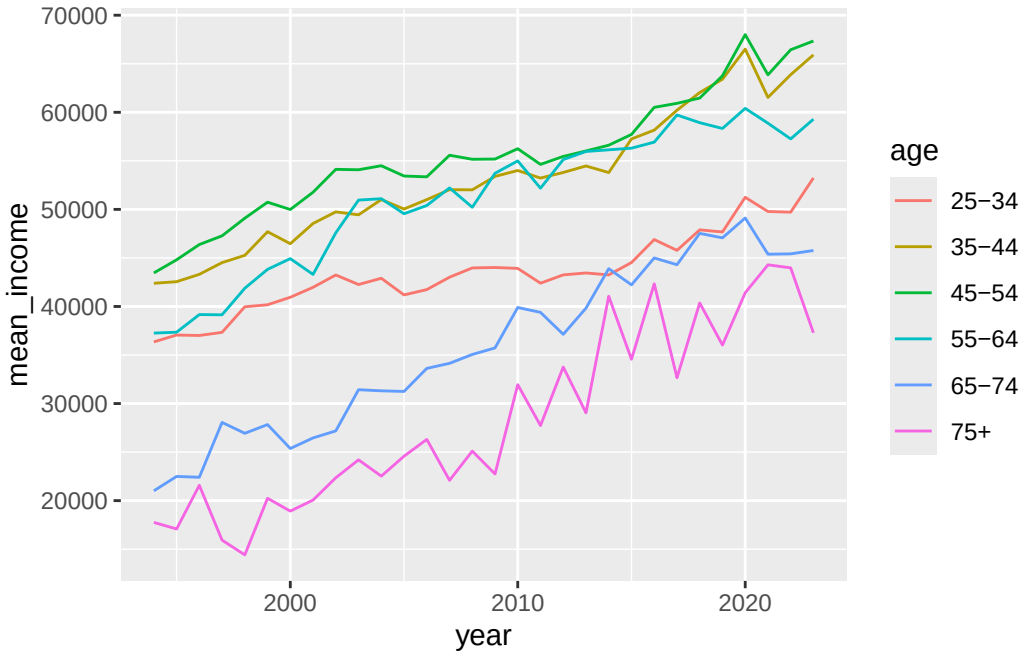



```
women_over_25 |>
  filter(income != 0) |>
  summarize(
```

```
    .by = c(year, race),
    mean_income = mean(income, na.rm = TRUE)
  ) |>
  ggplot(aes(x = year, y = mean_income, color = race)) +
  geom_line()
```



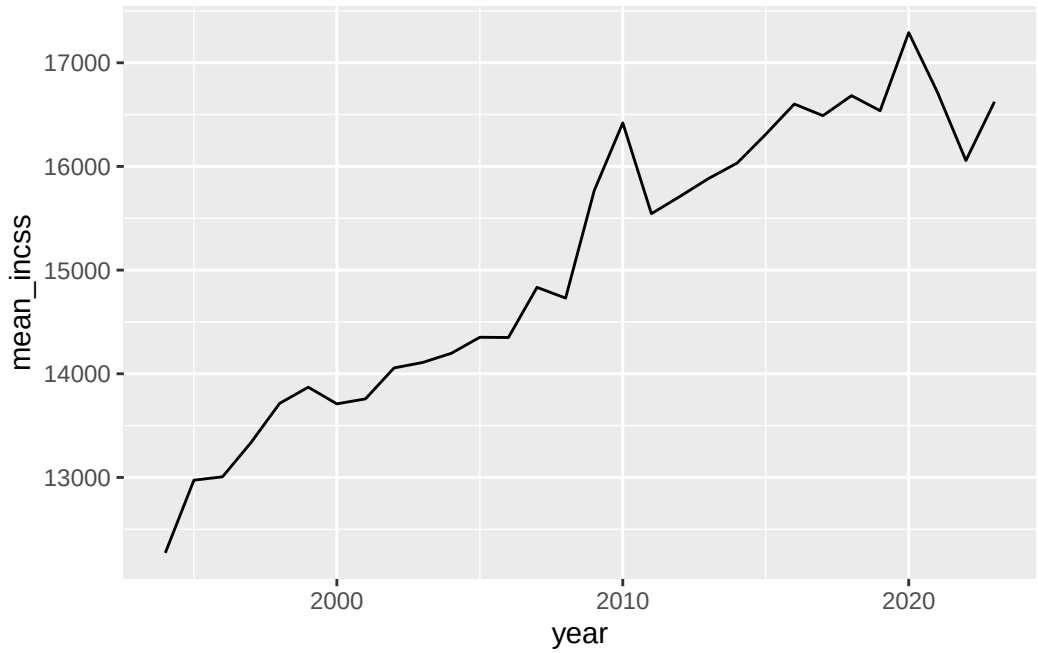

```
women_over_25 |>
  filter(income != 0) |>
  summarize(
    .by = c(year, age),
    mean_income = mean(income, na.rm = TRUE)
  ) |>
  ggplot(aes(x = year, y = mean_income, color = age)) +
  geom_line()
```



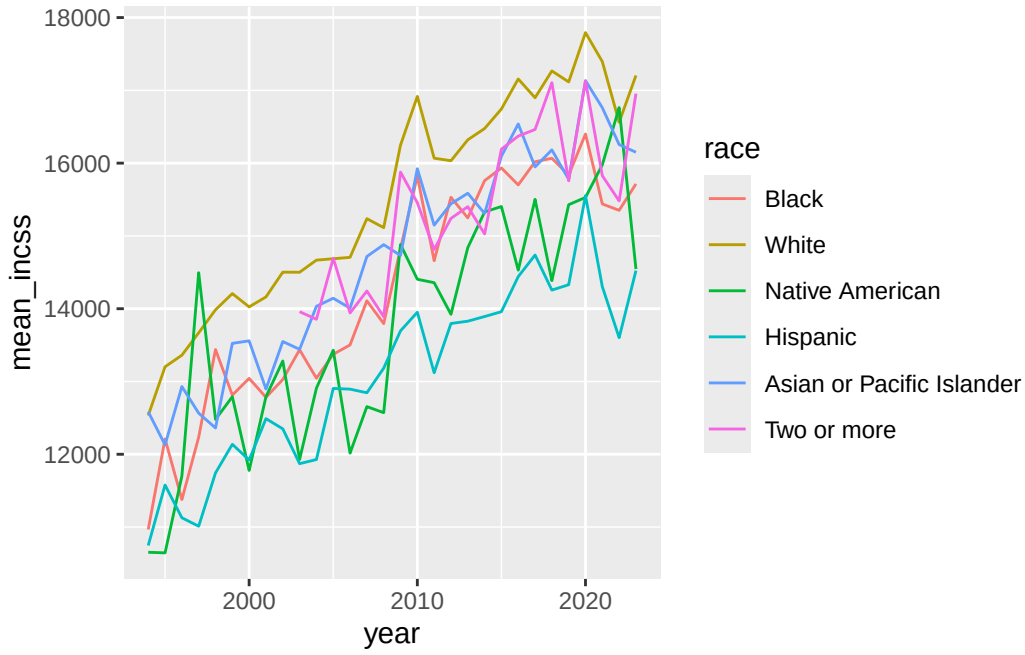
Average income has increased across the board. The most relevant observation is that income has increased the most for women aged 65 and older who are earning an income. This is not due to the effect of more elderly women working (those not working have been filtered out), but that the average working elderly woman is earning a higher income. This may be related to the increase in LFP for women over 65, with the logic that higher wages are associated with higher LFP, but it's hard to say that this relationship holds for all groups. For instance, Asian or Pacific Islander women experienced a dramatic increase in average income but had almost no change in LFP since 1994.

Examine average social insurance income per woman who was receiving it

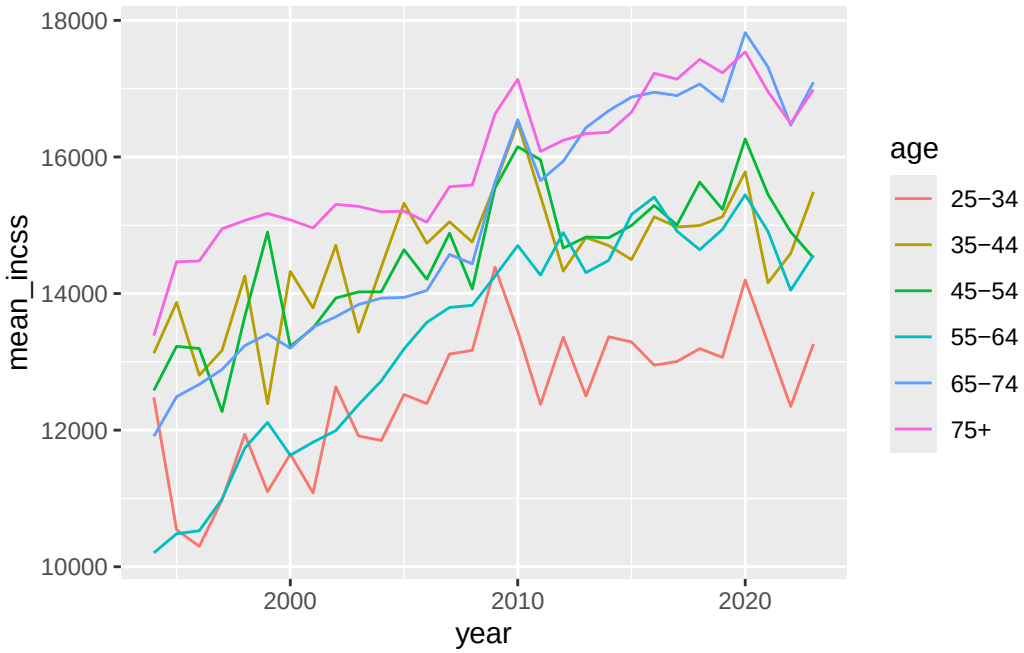
```
women_over_25 |>
  filter(incss != 0) |>
  summarize(
    .by = year,
    mean_incss = mean(incss)
  ) |>
  ggplot(aes(x = year, y = mean_incss)) +
  geom_line()
```




```
women_over_25 |>
  filter(incss != 0) |>
  summarize(
    .by = c(year, race),
    mean_incss = mean(incss)
  ) |>
  filter(!is.na(race)) |>
  ggplot(aes(x = year, y = mean_incss, color = race)) +
  geom_line()
```



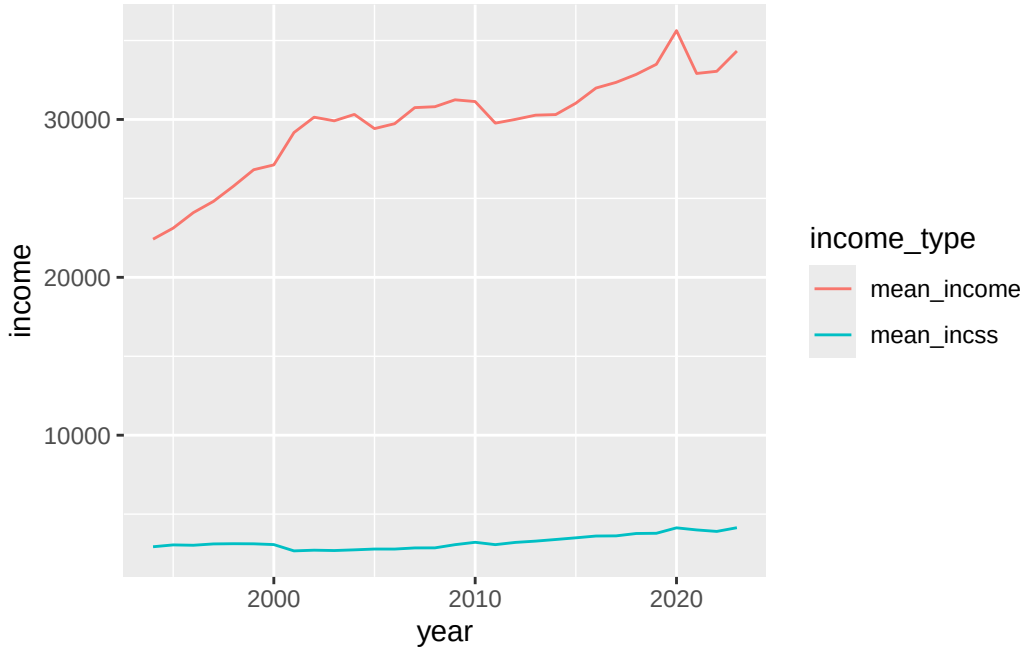

```
women_over_25 |>
  filter(incss != 0) |>
  summarize(
    .by = c(year, age),
    mean_incss = mean(incss)
  ) |>
  ggplot(aes(x = year, y = mean_incss, color = age)) +
  geom_line()
```



Average social insurance income has increased a large amount for women aged 55 to 64. However, similar to the previous section, it is unclear how this could be related to LFP.

Examine total average earned income versus social insurance income

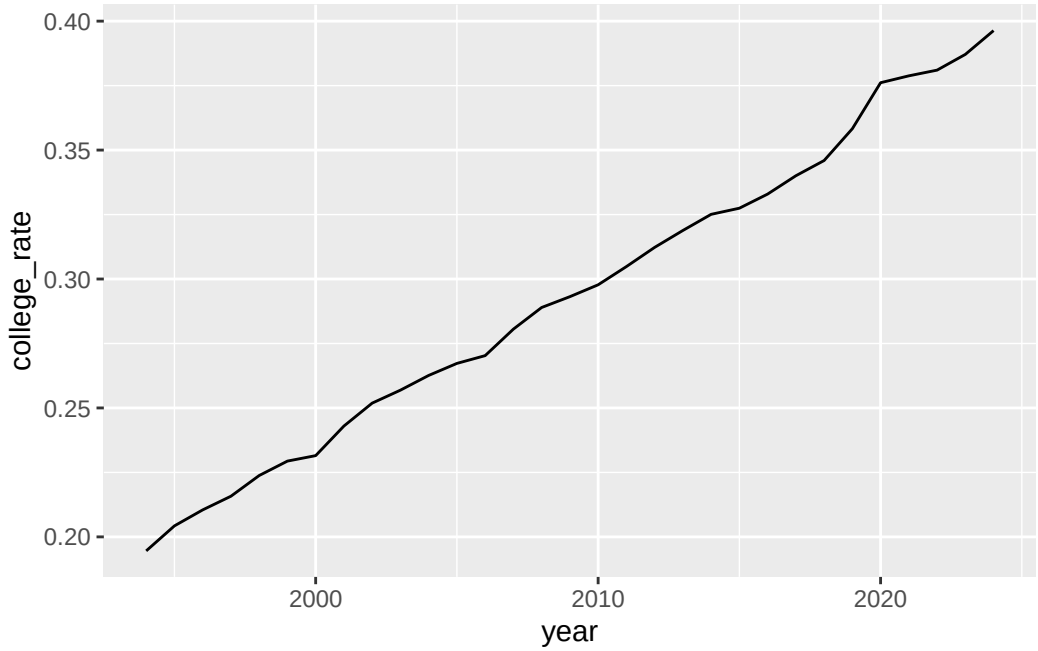
```
women_over_25 |>
  summarize(
    .by = year,
    mean_income = mean(income, na.rm = TRUE),
    mean_incss = mean(incss, na.rm = TRUE)
  ) |>
  pivot_longer(
    cols = c(mean_income, mean_incss),
    names_to = "income_type",
    values_to = "income"
  ) |>
  filter(year != 2024) |>
  ggplot(aes(x = year, y = income, color = income_type)) +
  geom_line()
```



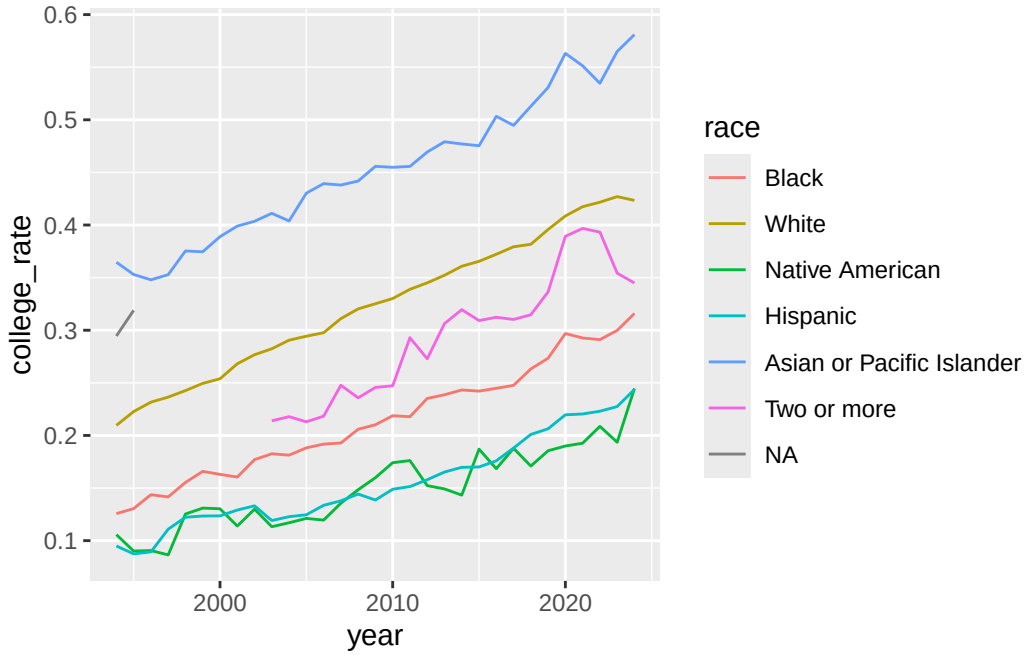
Both earned income and social insurance income are increasing over the period (averaged over all women, including those not earning the given kind of income).

Examine education attainment

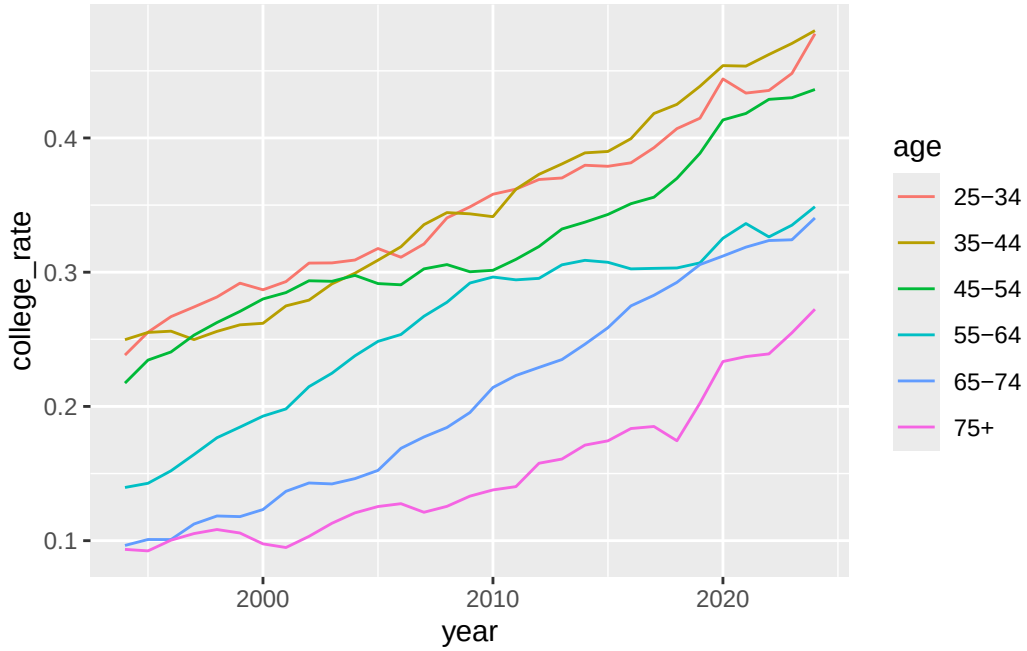
```
women_over_25 |>
  summarize(
    .by = year,
    college_rate = mean(college_lgl)
  ) |>
  ggplot(aes(x = year, y = college_rate)) +
  geom_line()
```




```
women_over_25 |>
  summarize(
    .by = c(year, race),
    college_rate = mean(college_lgl)
  ) |>
  ggplot(aes(x = year, y = college_rate, color = race)) +
  geom_line()
```




```
women_over_25 |>
  summarize(
    .by = c(year, age),
    college_rate = mean(college_lgl)
  ) |>
  ggplot(aes(x = year, y = college_rate, color = age)) +
  geom_line()
```



The rate of having a college degree is increasing over time for all races and ages. Women aged 75+ seem to have a comparably larger increase in the rate of having a college degree, but the effect seems small.

In conclusion for this question, wages, social insurance income, and education attainment are all increasing for most groups of women during the period since 1994. It is unclear, however, how this might be driving the overall trend in women's LFP. It does seem like there may be a connection between higher wages and education attainment for older women and their relatively large increase in LFP.

Question 4

> Between 1994 and 2024, which year had the steepest increase in female labor force participation relative to the previous year? What factors do you think are driving this pattern? Support your answers by using the data, referencing major events that happened around this time period, and/or citing previous studies.

Calculate year-over-year changes in female LFP

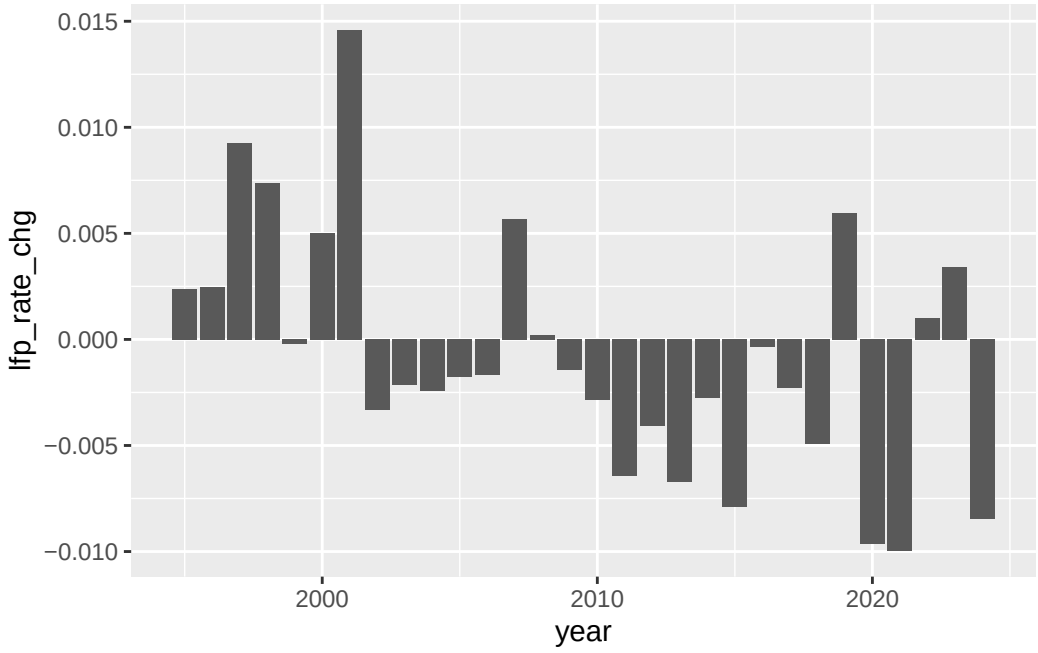
```
women_lfp_rate_chg <- women |>
  summarize(
```

```
    .by = year,
    lfp_rate = mean(lfp_lgl)
  ) |>
  mutate(
    lfp_rate_chg = lfp_rate - lag(lfp_rate)
  )

women_lfp_rate_chg |>
  filter(lfp_rate_chg == max(lfp_rate_chg, na.rm = TRUE))
```

```
# A tibble: 1 x 3
   year lfp_rate lfp_rate_chg
  <dbl>    <dbl>        <dbl>
1  2001    0.629       0.0146
```

```
women_lfp_rate_chg |>
  ggplot(aes(x = year, y = lfp_rate_chg)) +
  geom_col()
```

```
Warning: Removed 1 row containing missing values or values outside the scale range
(`geom_col()`).
```



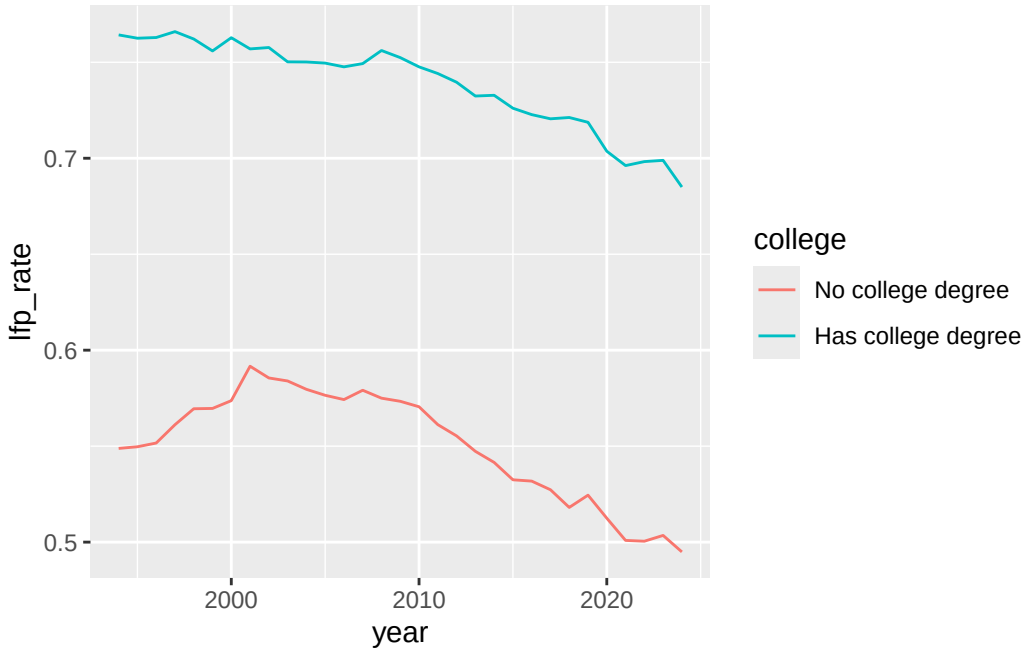
The year with the steepest increase in female labor force participation relative to the previous year was 2001, with an increase of 1.5% (absolute change in the proportion). This lines up with the September 11 attacks and the subsequent War in Afghanistan, during which women may have stepped in to fill any labor market gaps left by men serving in the war.

Question 5

> How has labor force participation for college-educated and not college-educated women evolved since 1994? Please provide graphs and/or tables to support your answer.

Examine LFP by college education status

```
women |>
  summarize(
    .by = c(year, college),
    lfp_rate = mean(lfp_lgl)
  ) |>
  ggplot(aes(x = year, y = lfp_rate, color = college)) +
  geom_line()
```



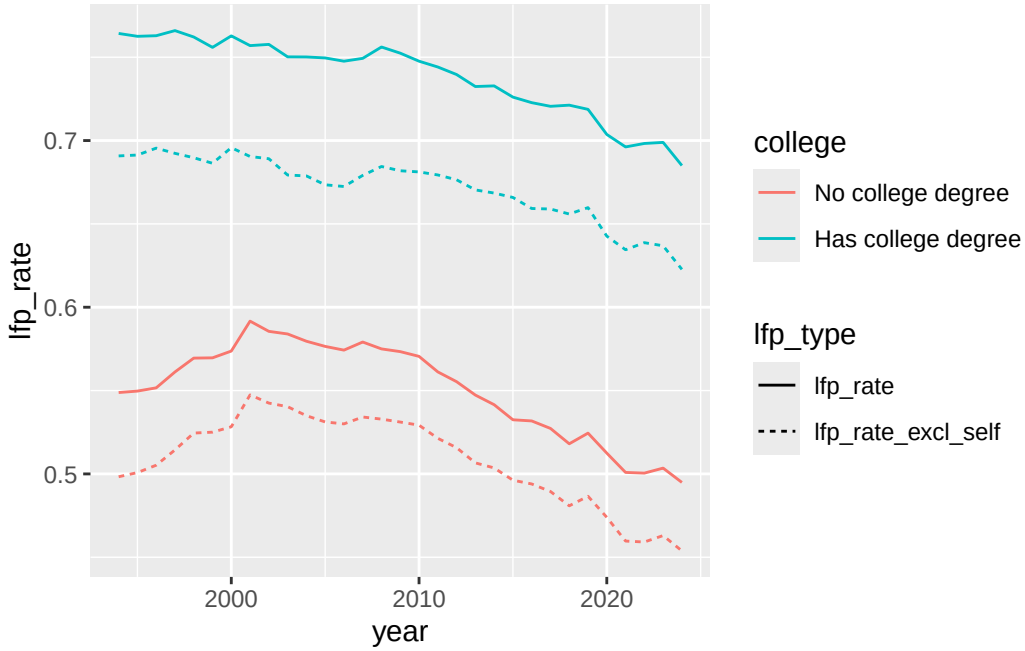
College-educated women experienced a slightly larger drop in LFP since 1994 than women without a college education. Furthermore, the drop for college-educated women has been more or less monotonic, with a steady decrease almost every year. The LFP rate for women without a college education, however, initially rose until the early 2000s, after which it experienced a sharper decline.

Question 6

> Create an alternative measure of labor force participation that excludes individuals from the labor force if they are self-employed in their main job (lfp = 0 if self-employed in main job). Using the new measure, describe how labor force participation for college- educated and not college-educated women has evolved since 1994. Please provide graphs and/or tables to support your answer.

Examine alternative measure of LFP by college education status

```
women |>
  summarize(
    .by = c(year, college),
    lfp_rate = mean(lfp_lgl),
    lfp_rate_excl_self = mean(lfp_lgl_excl_self)
  ) |>
  pivot_longer(
    cols = c(lfp_rate, lfp_rate_excl_self),
    names_to = "lfp_type",
    values_to = "lfp_rate"
  ) |>
  ggplot(aes(x = year, y = lfp_rate, color = college, linetype = lfp_type)) +
  geom_line()
```



The alternative measure of LFP is everywhere lower than the regular measure. The effect of changing to the alternative measure is larger for women with a college degree than for women without a college degree, indicating that a relatively larger proportion of women with a college degree are self-employed. This seems to make sense given that self-employed workers are usually in a skilled or white-collar profession. However, any effects of the recent rise of the gig economy do not seem to be captured in the data, perhaps because they are small relative to the magnitude of the labor force as a whole.

Using the alternative measure, the pattern of women's LFP evolution by college education status is not materially different than when using the original measure.

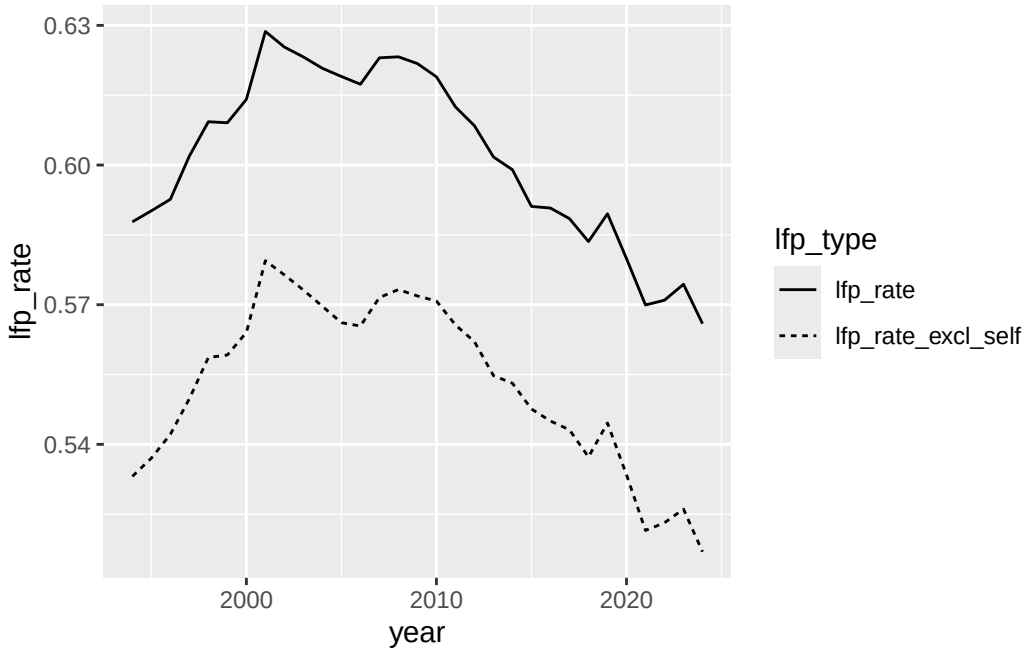
Question 7

> How does our labor market analysis change when we use the new measure? Which measure do you prefer? Explain.

Examine alternative measure of LFP for different cross sections of the data

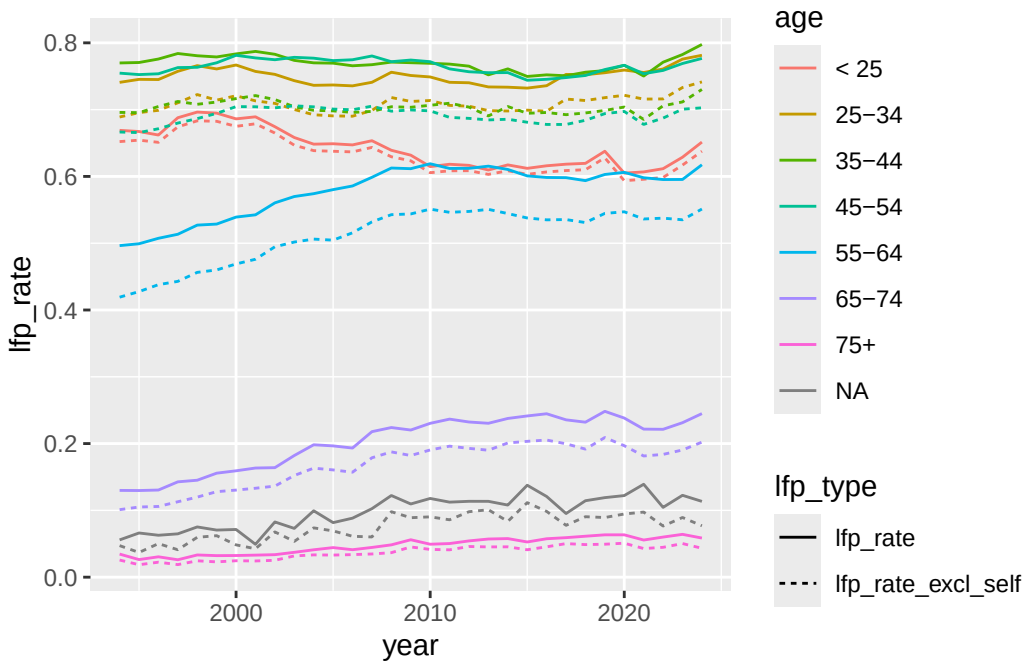
```
women |>
  summarize(
    .by = year,
```

```
    lfp_rate = mean(lfp_lgl),
    lfp_rate_excl_self = mean(lfp_lgl_excl_self)
  ) |>
  pivot_longer(
    cols = c(lfp_rate, lfp_rate_excl_self),
    names_to = "lfp_type",
    values_to = "lfp_rate"
  ) |>
  ggplot(aes(x = year, y = lfp_rate, linetype = lfp_type)) +
  geom_line()
```

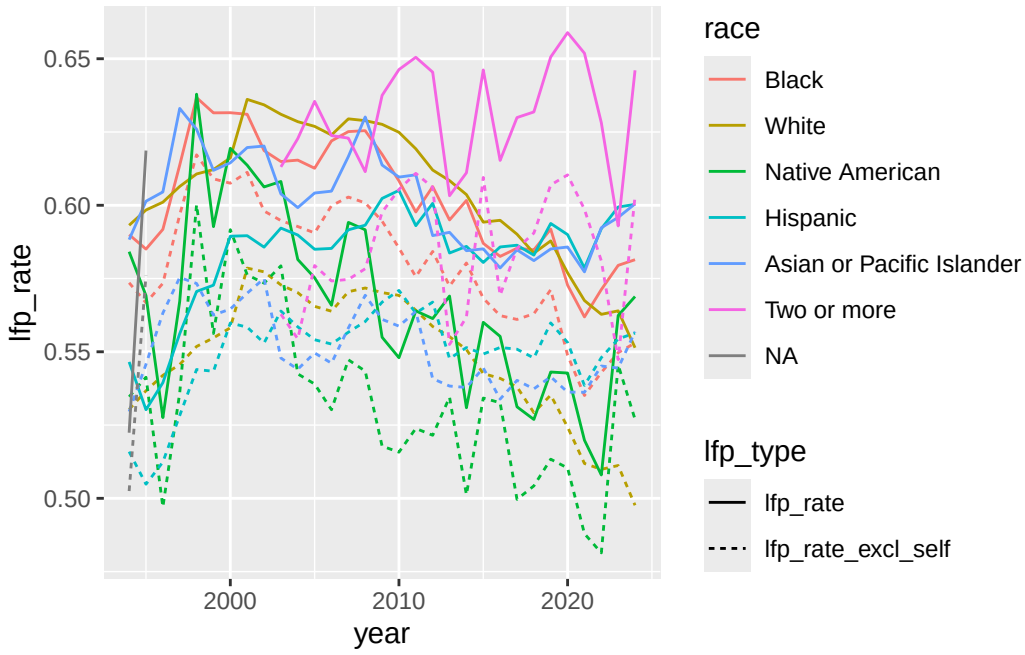



```
women |>
  summarize(
    .by = c(year, age),
    lfp_rate = mean(lfp_lgl),
    lfp_rate_excl_self = mean(lfp_lgl_excl_self)
  ) |>
  pivot_longer(
    cols = c(lfp_rate, lfp_rate_excl_self),
    names_to = "lfp_type",
    values_to = "lfp_rate"
  ) |>
```

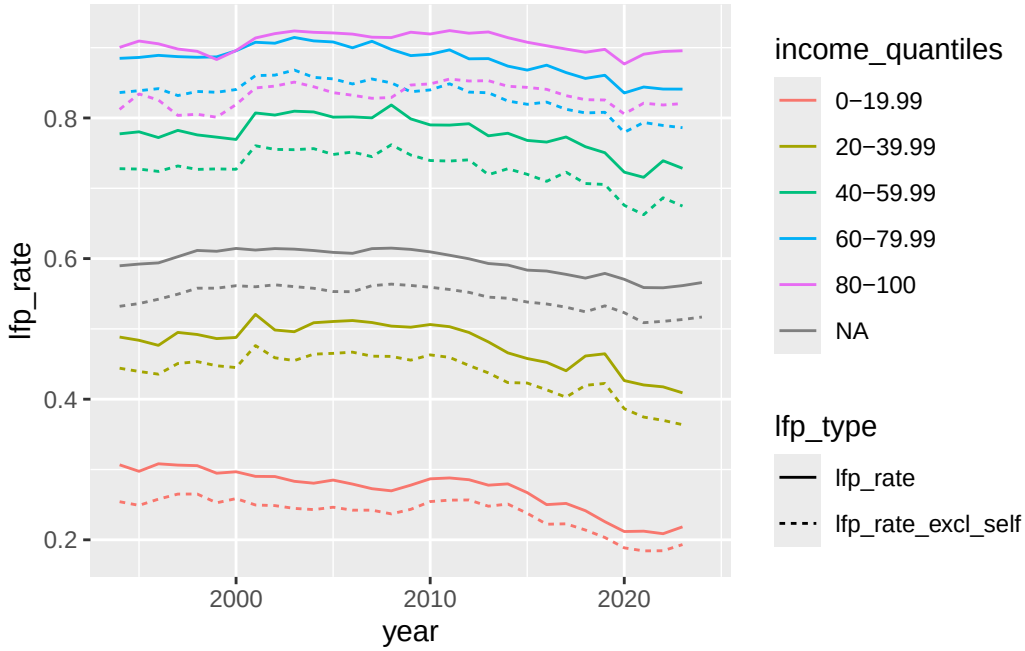
```
  ggplot(aes(x = year, y = lfp_rate, color = age, linetype = lfp_type)) +
  geom_line()
```




```
women |>
  summarize(
    .by = c(year, race),
    lfp_rate = mean(lfp_lgl),
    lfp_rate_excl_self = mean(lfp_lgl_excl_self)
  ) |>
  pivot_longer(
    cols = c(lfp_rate, lfp_rate_excl_self),
    names_to = "lfp_type",
    values_to = "lfp_rate"
  ) |>
  ggplot(aes(x = year, y = lfp_rate, color = race, linetype = lfp_type)) +
  geom_line()
```




```
women |>
  summarize(
    .by = c(year, income_quantiles),
    lfp_rate = mean(lfp_lgl),
    lfp_rate_excl_self = mean(lfp_lgl_excl_self)
  ) |>
  pivot_longer(
    cols = c(lfp_rate, lfp_rate_excl_self),
    names_to = "lfp_type",
    values_to = "lfp_rate"
  ) |>
  ggplot(aes(x = year, y = lfp_rate,
             color = income_quantiles, linetype = lfp_type)) +
  geom_line()
```



Without analyzing the effect of changing to the new measure on different cross sections of the data, it is hard to say for certain, but it seems that the effect is simply to lower LFP across the board, albeit in differing magnitudes for different demographics.

At the end of the day, the way labor force participation should be measured depends on the goals of the economic analyst. If LFP is simply intended to show how much of the population is working, then self-employment should clearly count as employment because self-employed people are indeed working. In this case, due to the differing effects on different demographics, filtering out self-employed people would distort the numbers.

However, there could be applications of LFP for which it makes sense to filter out self-employed people. Perhaps the self-employed are less likely to try to find a new job if they lose work, or are otherwise unwilling to work if not for themselves; then, it might make sense to exclude them if the goal is to use LFP as a proxy for the size of the active labor market.

Part 2: Telework

Create new variables and filter dataset

The most important concept here is that `had_telework` is created while the data is grouped by `cpsidp`. This means that any individual who had telework during COVID is categorized as having had telework for all years. The point is to be able to follow the same individuals

(who had telework during COVID) and observe their outcomes post-COVID; the years before COVID can be disregarded into missing values later.

```
all <- data |>
  mutate(
    employed_lgl = employed == "Employed",
    lfp_lgl = lfp == "In labor force",
    covid_telework_lgl =
      covid_telework == "Telework from 2021-2022 due to COVID"
  ) |>
  mutate(
    .by = cpsidp,
    had_telework = any(covid_telework_lgl)
    )

women <- all |>
  filter(sex == "Female")

women_over_25 <- women |>
  filter(age != "< 25")
```

Question 1

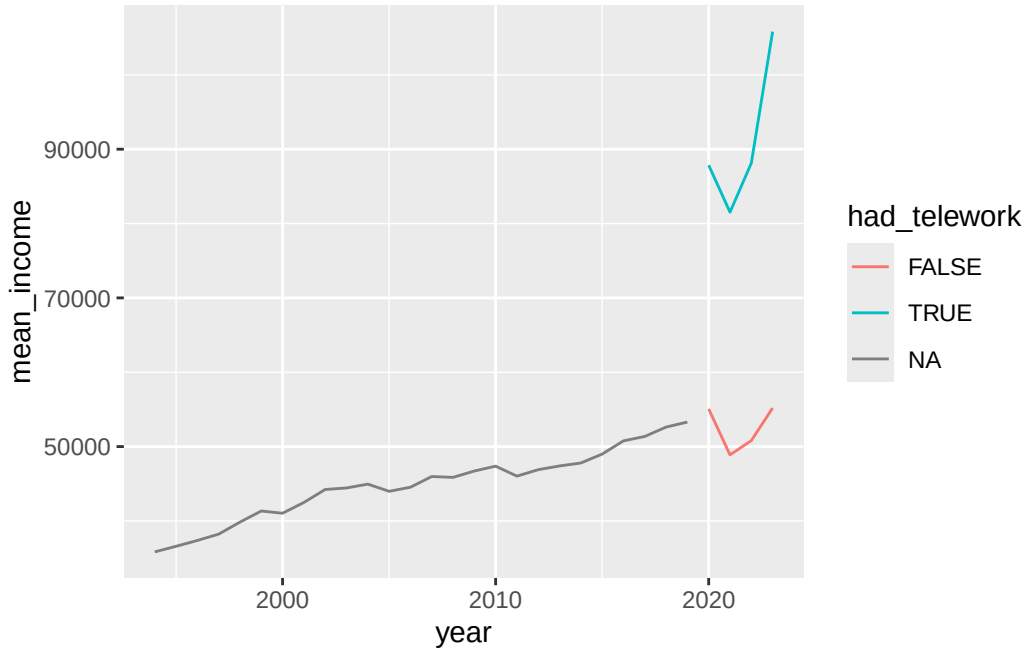
> Since the rise of telework in 2020, how have wages, employment, and labor force participation changed for women who had telework from 2020-2024 and women who did not? Please provide at least three graphs and/or tables to support your answer.

I will attempt to answer this question by comparing women who had telework during COVID and women who did not have telework during COVID.

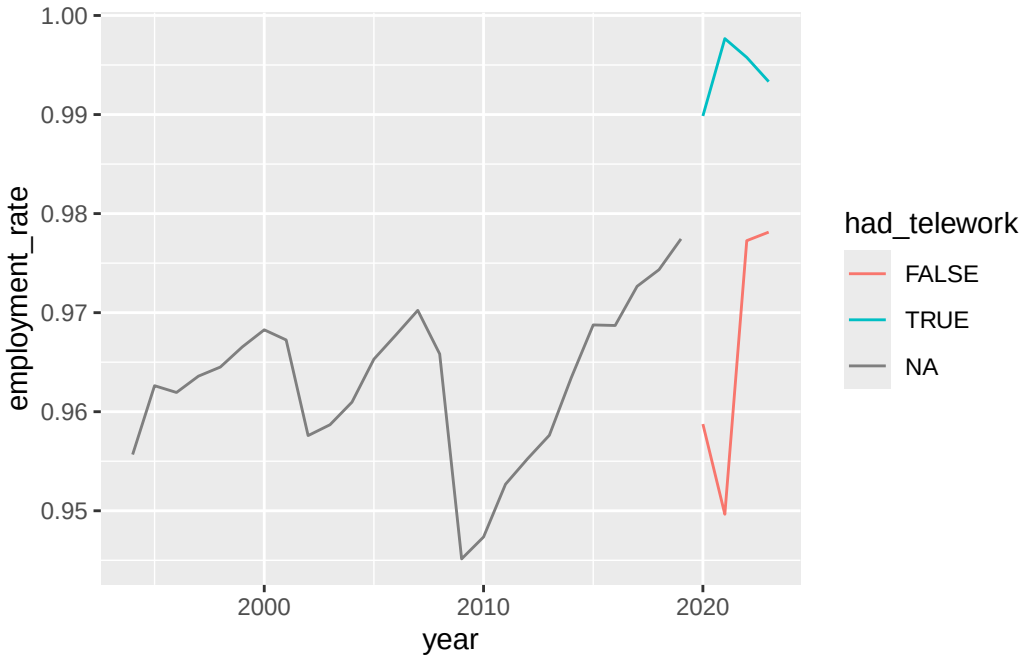
Examine labor market outcomes by teleworking status during COVID

```
summary_women_by_telework <- women |>
  filter(income != 0) |>
  summarize(
    .by = c(year, had_telework),
    mean_income = mean(income, na.rm = TRUE),
    employment_rate = mean(employed_lgl, na.rm = TRUE),
    lfp_rate = mean(lfp_lgl, na.rm = TRUE)
```

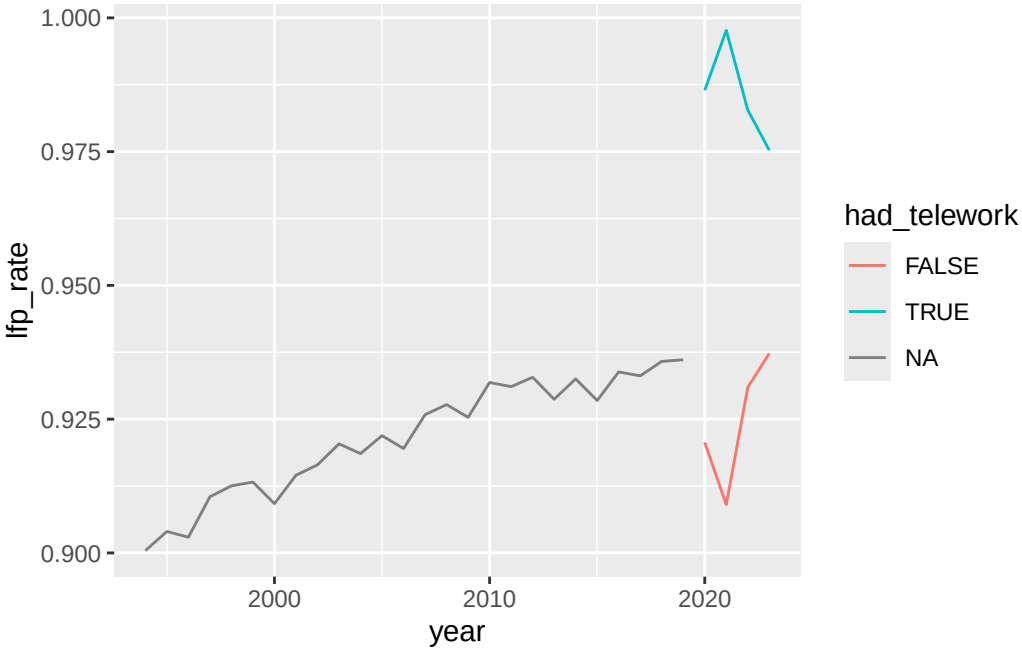
```
  ) |>
  mutate(
    had_telework = case_when(
      year <= 2019 ~ NA,
      year >= 2020 ~ !is.na(had_telework)
    )
  )

summary_women_by_telework |>
  ggplot(aes(x = year, y = mean_income, color = had_telework)) +
  geom_line()
```




```
summary_women_by_telework |>
  ggplot(aes(x = year, y = employment_rate, color = had_telework)) +
  geom_line()
```




```
summary_women_by_telework |>
  ggplot(aes(x = year, y = lfp_rate, color = had_telework)) +
  geom_line()
```



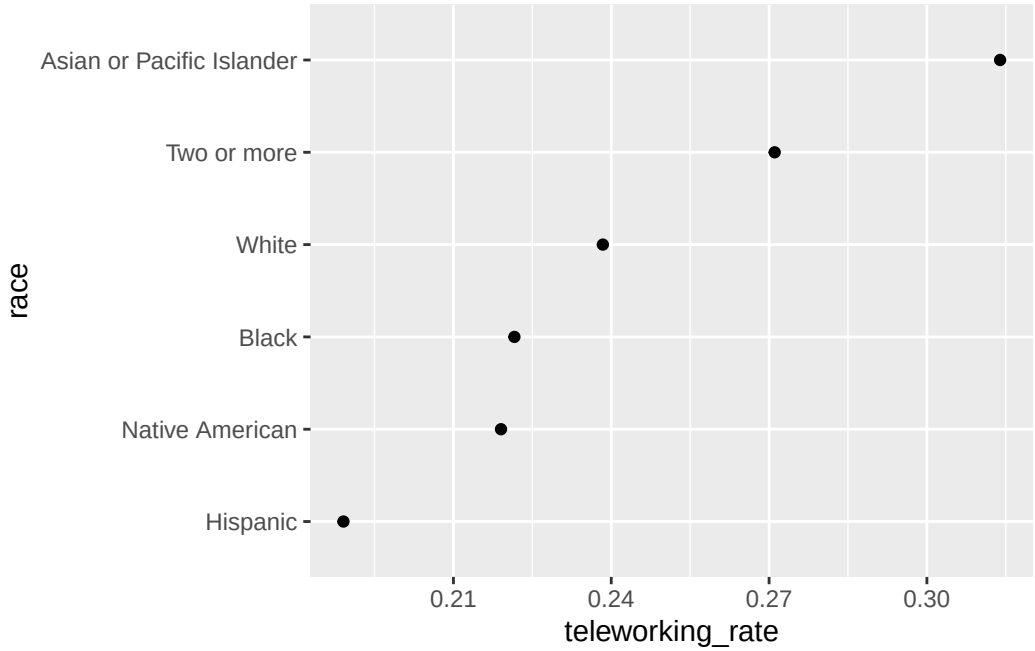
Women who were able to telework during COVID had markedly better labor outcomes even after 2022, with much higher incomes, employment rates, and LFP rates. There seems to be some regression to the mean in employment rates and LFP rates in 2022 and beyond. For incomes, the effect seems to persist or even expand past 2022.

Question 2

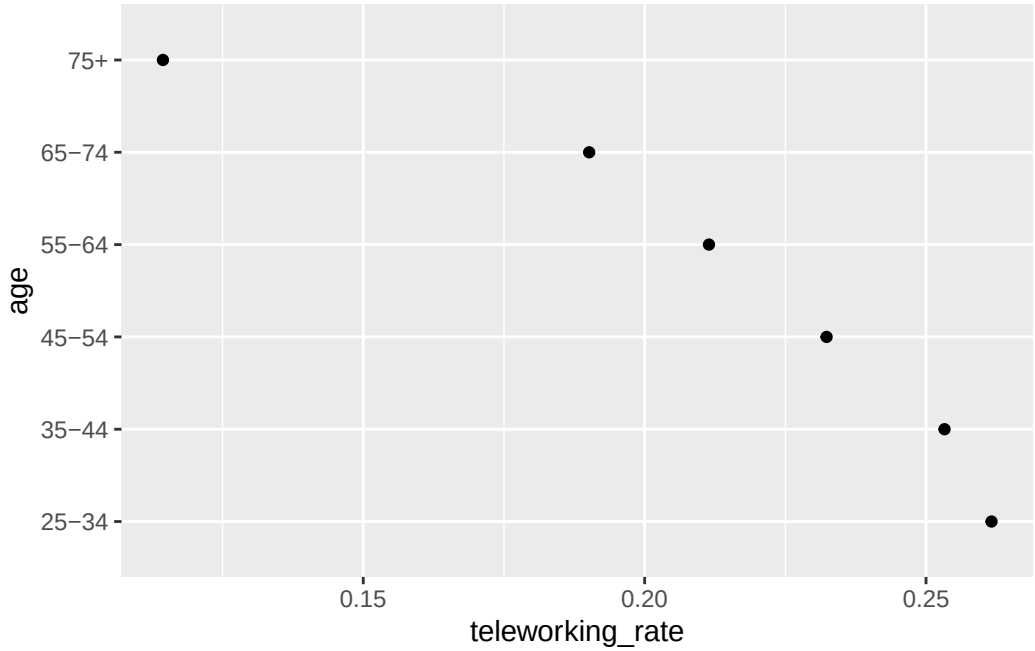
> For which groups of women older than 25 was telework due to the pandemic most common in 2021? Based on these patterns, what can you infer about the relationship between economic well-being and the ability to telework between 2021? Please provide at least three graphs and/or tables to support your answer.

Examine 2021 teleworking rates for women older than 25 by demographic factors

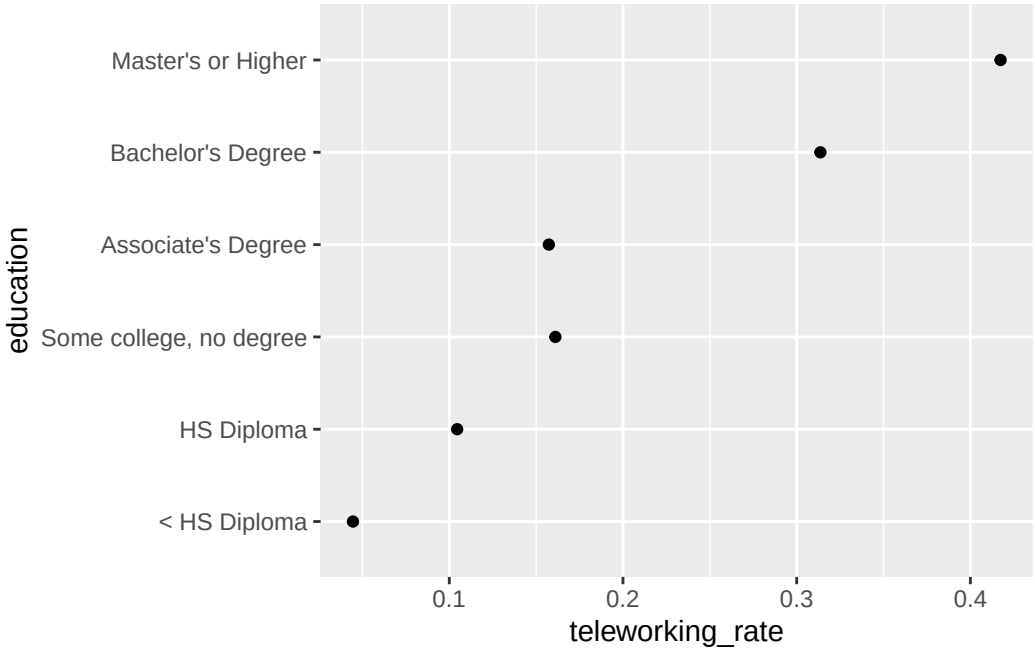
```
women_over_25 |>
  filter(year == 2021) |>
  summarize(
    .by = race,
    teleworking_rate = mean(covid_telework_lgl, na.rm = TRUE)
  ) |>
  mutate(
    race = fct_reorder(race, teleworking_rate)
  ) |>
  ggplot(aes(x = teleworking_rate, y = race)) +
  geom_point()
```




```
women_over_25 |>
  filter(year == 2021) |>
  summarize(
    .by = age,
    teleworking_rate = mean(covid_telework_lgl, na.rm = TRUE)
  ) |>
  ggplot(aes(x = teleworking_rate, y = age)) +
  geom_point()
```




```
women_over_25 |>
  filter(year == 2021) |>
  summarize(
    .by = education,
    teleworking_rate = mean(covid_telework_lgl, na.rm = TRUE)
  ) |>
  ggplot(aes(x = teleworking_rate, y = education)) +
  geom_point()
```



White or Asian/Pacific Islander women had higher rates of telework in 2021 than Black, Native American, and Hispanic women. The more advanced a woman's level of education, the more likely she was able to telework. Although a direct observation of the relationship between income and teleworking rate is unavailable in the data, the aforementioned demographic factors are highly correlated with income. It can therefore be inferred that women with higher incomes were much more likely to be able to telework in 2021.

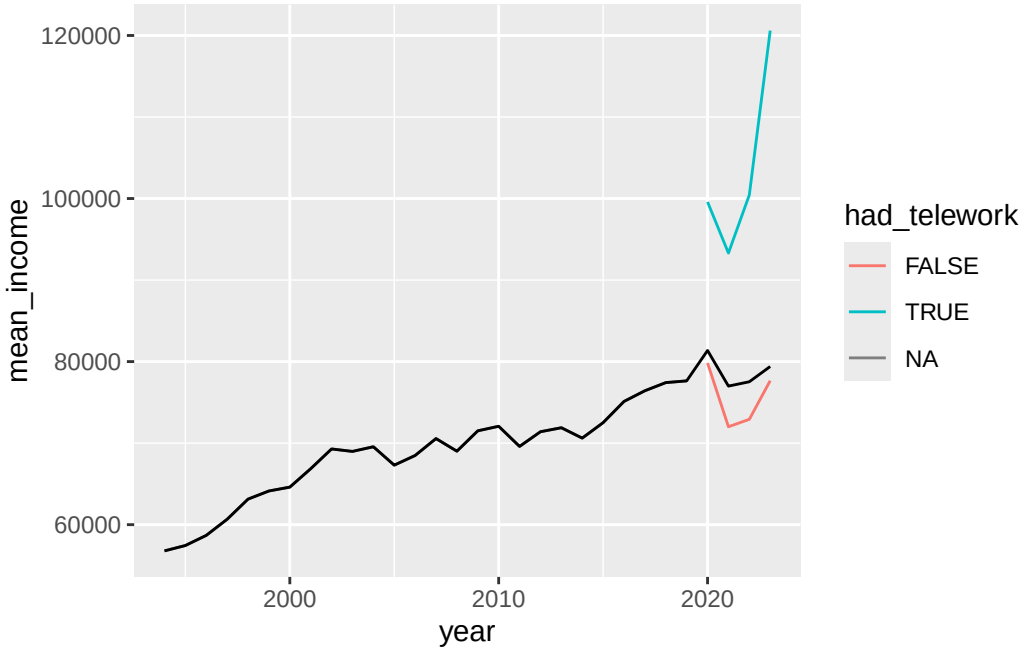
Question 3

> Predict what trends in wages, employment, and labor force participation for college-educated women from 2020 to 2024 would have looked like if telework was not an option. What does this tell you about the economic impacts of telework during the COVID-19 pandemic? Please support your answer with graphs and/or tables.
>
> Hint: Look at trends from previous years that had similar economic contexts. Also, feel free to explore the variables you haven't used yet.

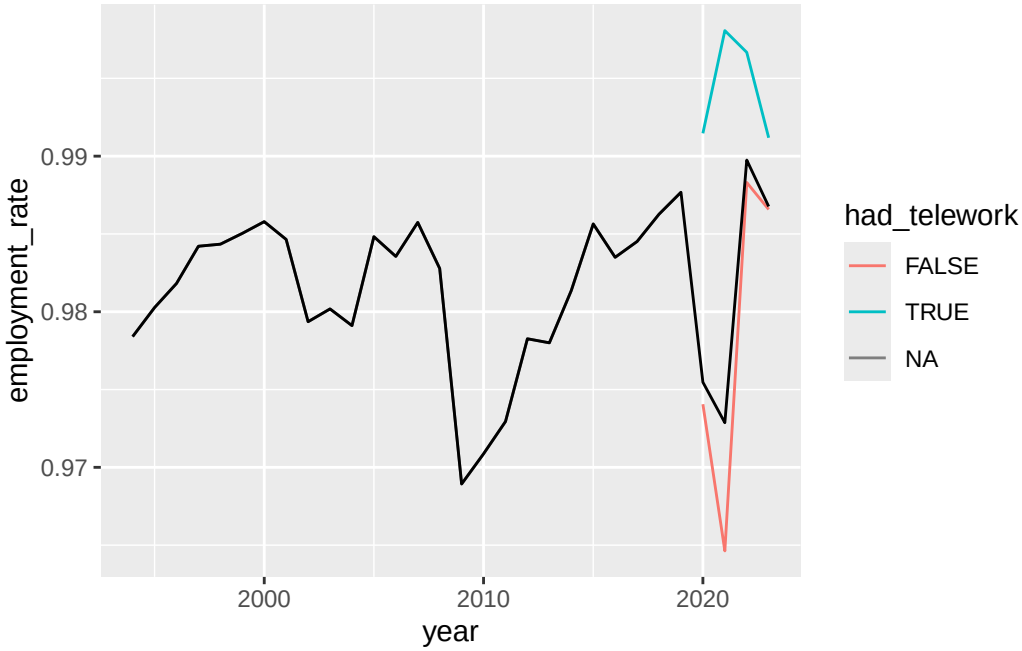
For timing reasons, the answer to this question is kept to a simplistic analysis.

Examine labor market outcomes for college-educated women by teleworking status during COVID

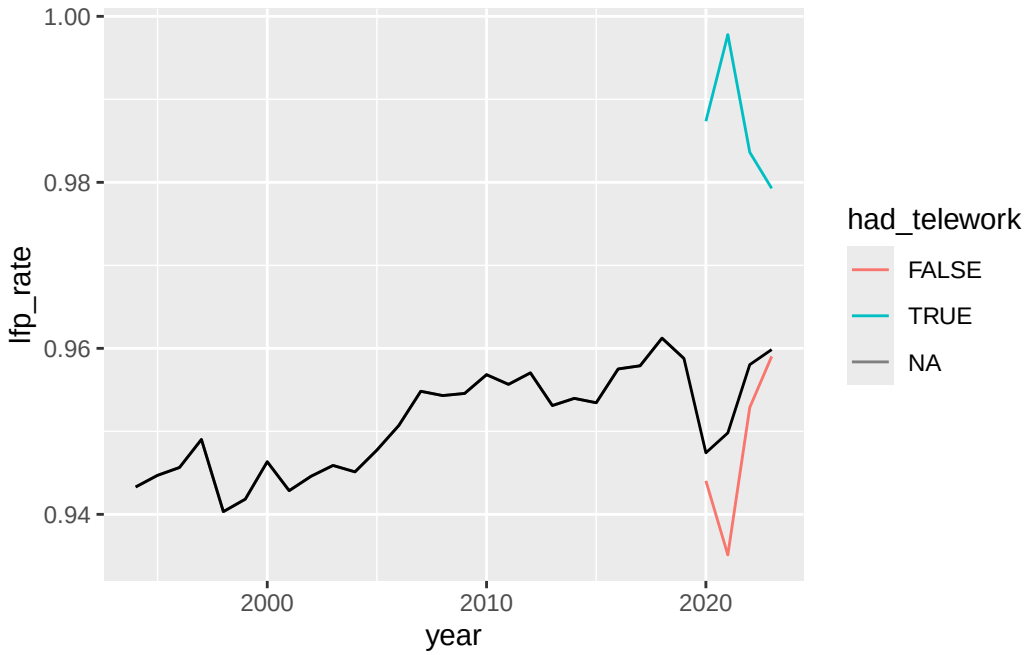
```
summary_college_women_by_telework <- women |>
  filter(income != 0, college == "Has college degree") |>
  summarize(
    .by = c(year, had_telework),
    mean_income = mean(income, na.rm = TRUE),
    employment_rate = mean(employed_lgl, na.rm = TRUE),
    lfp_rate = mean(lfp_lgl, na.rm = TRUE)
  ) |>
  mutate(
    had_telework = case_when(
      year <= 2019 ~ NA,
      year >= 2020 ~ !is.na(had_telework)
    )
  )

summary_college_women <- women |>
  filter(income != 0, college == "Has college degree") |>
  summarize(
    .by = year,
    mean_income = mean(income, na.rm = TRUE),
    employment_rate = mean(employed_lgl, na.rm = TRUE),
    lfp_rate = mean(lfp_lgl, na.rm = TRUE)
  )

summary_college_women_by_telework |>
  ggplot(aes(x = year, y = mean_income)) +
  geom_line(aes(color = had_telework)) +
  geom_line(data = summary_college_women)
```




```
summary_college_women_by_telework |>
  ggplot(aes(x = year, y = employment_rate)) +
  geom_line(aes(color = had_telework)) +
  geom_line(data = summary_college_women)
```




```
summary_college_women_by_telework |>
  ggplot(aes(x = year, y = lfp_rate)) +
  geom_line(aes(color = had_telework)) +
  geom_line(data = summary_college_women)
```



Under this simplistic analysis (not attempting to consider a counterfactual), telework caused positive economic impacts for college-educated women during and after the pandemic. Under the assumption that the complete absence of telework during COVID would have caused all college-educated women to experience the same effects as those college-educated women without telework actually experienced during COVID, we can say that the difference between the black and red lines in the above charts represents the impacts of telework. Incomes were higher, employment rates were slightly higher, and LFP rates were higher as a result of telework.