INFO 3370 Final Project

Code

Import and Setup

```
library(readr)
```

```
Warning: package 'readr' was built under R version 4.2.3
```

```
library(haven)
```

```
Warning: package 'haven' was built under R version 4.2.3
```

```
library(tidyverse)
```

```
Warning: package 'tidyverse' was built under R version 4.2.3
```

```
Warning: package 'ggplot2' was built under R version 4.2.3
```

```
Warning: package 'tibble' was built under R version 4.2.3
```

```
Warning: package 'dplyr' was built under R version 4.2.3
```

```
Warning: package 'lubridate' was built under R version 4.2.3
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.2     v purrr     1.0.1
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.5.0     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(scales)
```

```
Warning: package 'scales' was built under R version 4.2.3
```

```
Attaching package: 'scales'

The following object is masked from 'package:purrr':

    discard

The following object is masked from 'package:readr':

    col_factor
```

```
library(dplyr)
library(plotly)
```

```
Warning: package 'plotly' was built under R version 4.2.3
```

```
Attaching package: 'plotly'

The following object is masked from 'package:ggplot2':

    last_plot

The following object is masked from 'package:stats':
```

```
    filter

The following object is masked from 'package:graphics':

    layout
```

```
df_main <- read_sav("data/usa_00001.sav.gz")
df_main$RACE <- as.numeric(df_main$RACE)
```

Data Cleaning

```
# Define income classes
df_main_income_class <- df_main |>
  mutate(income_class = case_when(
    INCWAGE <= 30000 ~ "Lower Class",
    INCWAGE > 30000 & INCWAGE <= 94000 ~ "Middle Class",
    INCWAGE > 94000 ~ "Upper Class",
    TRUE ~ "Unknown"
  ))

# Filter and clean up data
health <- df_main_income_class |>
  mutate(RACE = case_when(
    RACE == 1 ~ "White",
    RACE == 2 ~ "Black/African American",
    RACE == 3 ~ "American Indian or Alaska Native",
    RACE == 4 ~ "Chinese",
    RACE == 5 ~ "Japanese",
    RACE == 6 ~ "Other Asian or Pacific Islander",
  )) |>
  mutate(
    HCOVPUB = ifelse(HCOVPUB == 1, 0, 1),
    HCOVPRIV = ifelse(HCOVPRIV == 1, 0, 1)
  ) |>
  mutate(
    insurance_type =
      if_else(HCOVPUB == 0 & HCOVPRIV == 1, "Private",
      if_else(HCOVPUB == 1 & HCOVPRIV == 0, "Public",
      if_else(HCOVPUB == 1 & HCOVPRIV == 1, "Both",
      if_else(HCOVPUB == 0 & HCOVPRIV == 0, "None", NA))))
```

```
  )

# Count cases dropped
dropped_cases <- nrow(df_main) - nrow(df_main_income_class)
print(paste("Cases dropped at defining income classes:", dropped_cases))
```

```
[1] "Cases dropped at defining income classes: 0"
```

```
dropped_cases <- nrow(df_main_income_class) - nrow(health)
print(paste("Cases dropped at filtering step:", dropped_cases))
```

```
[1] "Cases dropped at filtering step: 0"
```

```
health_prop <- health |>
  filter(insurance_type != "None") |>
  drop_na(income_class, insurance_type, RACE, YEAR) |>
  group_by(income_class, RACE, YEAR) |>
  summarize(weighted_count = sum(PERWT))
```

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

```
# Count cases dropped
dropped_cases <- nrow(health) - nrow(health_prop)
print(paste("Cases dropped at summarizing step:", dropped_cases))
```

```
[1] "Cases dropped at summarizing step: 1577702"
```

```
#  group_by(income_class, insurance_type, YEAR) |>
#  mutate(percent = round(weighted_count / sum(weighted_count), 4))

health_22 <- health |>
  filter(INCWAGE > 0, YEAR == 2022) |>
  drop_na(insurance_type) |>
  drop_na(RACE)
```

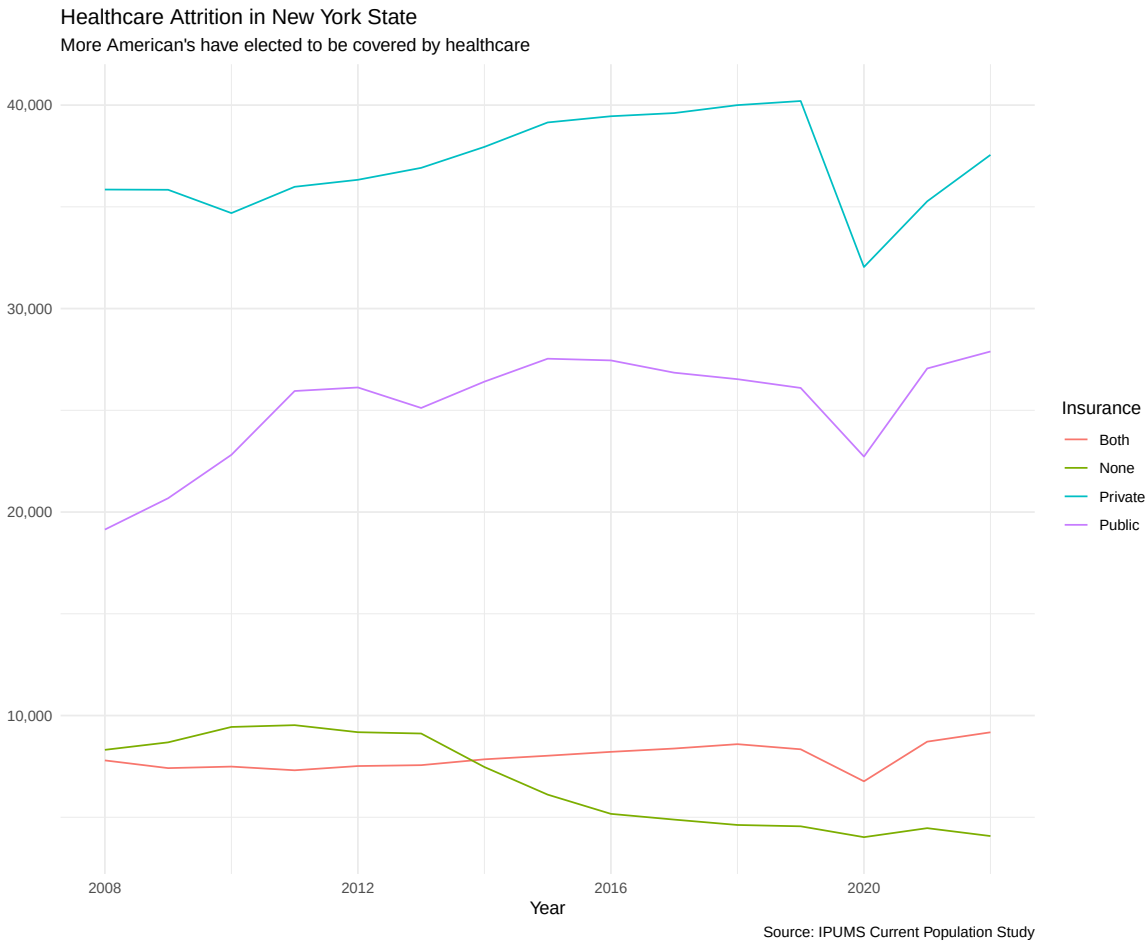
```r
# Calculate proportions
health_22_prop <- health_22 |>
  group_by(income_class, insurance_type, RACE) |>
  summarize(weighted_count = sum(PERWT)) |>
  group_by(income_class, insurance_type) |>
  mutate(percent = round(weighted_count / sum(weighted_count), 4))
```

```
`summarise()` has grouped output by 'income_class', 'insurance_type'. You can
override using the `.groups` argument.
```

Visualization on Trends

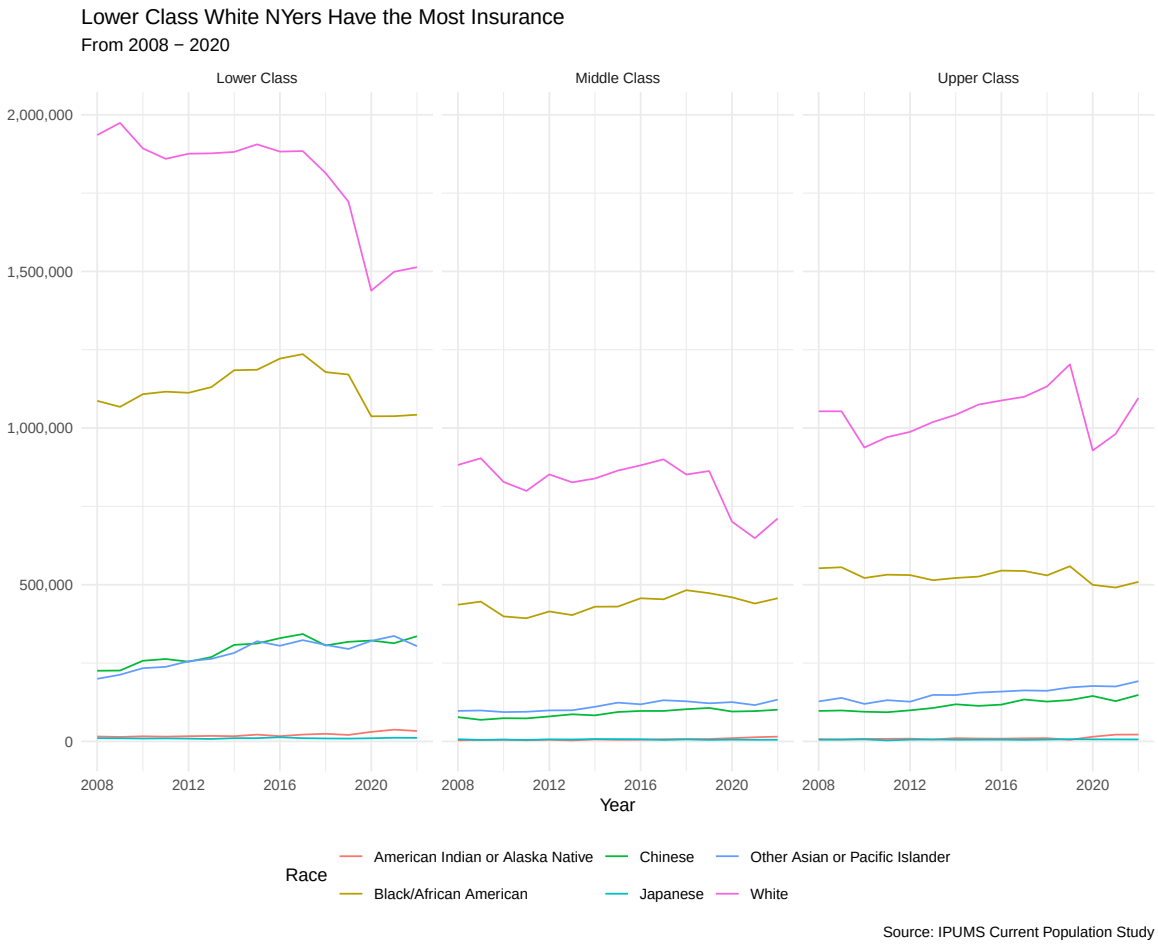
```r
health |>
  drop_na(insurance_type) |>
  group_by(YEAR, insurance_type) |>
  summarize (count = n()) |>
  ggplot(aes(YEAR, count, color = insurance_type, group = insurance_type)) +
  geom_line() +
  scale_y_continuous(label = comma) +
  theme(legend.position = "bottom") +
  theme_minimal() +
  labs(
    title = "Healthcare Attrition in New York State",
    subtitle = "More American's have elected to be covered by healthcare",
    x = "Year",
    y = "",
    color = "Insurance",
    caption = "Source: IPUMS Current Population Study"
  )
```

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




```
health_prop |>
  ggplot(aes(YEAR, weighted_count, color = RACE, group = RACE)) +
  geom_line() +
  facet_wrap(~income_class) +
  scale_y_continuous(label = comma) +
  theme_minimal() +
    theme(legend.position = "bottom") +
  labs(
    title = "Lower Class White NYers Have the Most Insurance",
    subtitle = "From 2008 - 2020",
    x = "Year",
    y = "",
    color = "Race",
    caption = "Source: IPUMS Current Population Study"
```

```
)
```

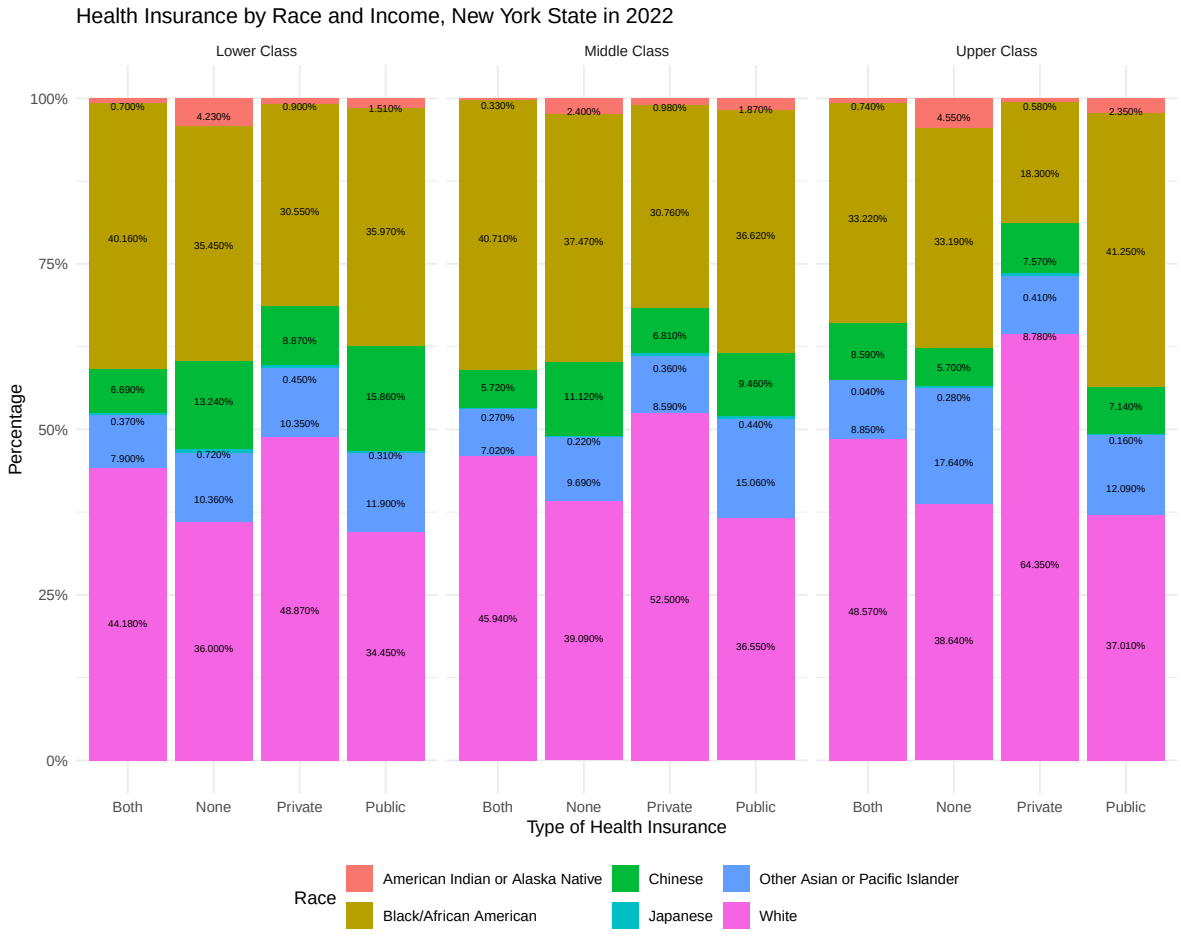


Final Visualization

```
# Create bar chart visualization
health_22_prop_bar <- ggplot(data = health_22_prop, aes(x = insurance_type,
                                                        y = percent,
                                    fill = RACE)) +
  geom_bar(position = "fill", stat = "identity") +
  scale_y_continuous(labels = scales::percent) +
  facet_wrap(~income_class) +
  theme_minimal() +
  theme(legend.position = "bottom") +
```

```
  geom_text(aes(label = scales::percent(percent), y = percent + 0.02),
            position = position_fill(vjust = 0.5), size = 2) +
  labs(
    title = "Health Insurance by Race and Income, New York State in 2022",
    x = "Type of Health Insurance",
    y = "Percentage",
    fill = "Race"
  )

health_22_prop_bar
```



Write-Up

Introduction

How do income and racial inequality impact access to healthcare services? Understanding this question is essential for advancing social justice and improving public health outcomes. Addressing this query can help identify and confront barriers to access, which can lead to better health outcomes for entire communities and populations. Additionally, findings can inform evidence-based policies aimed at reducing healthcare disparities, ultimately creating a more equitable healthcare system. Furthermore, improving access to healthcare can have economic benefits, including increased productivity and reduced healthcare costs.

The Data

To analyze our research question, our unit of analysis is individual people in New York state and our target population is all individuals residing within the geographic boundaries of New York State. This includes people of all ages, genders, races, ethnicities, socioeconomic statuses, and health conditions who are currently living in New York state. We obtained data from IPUMS USA, which gathers microdata from U.S. censuses and encompasses decennial census records spanning from 1790 to 2010, as well as American Community Surveys (ACS) conducted from 2000 onwards. To analyze our research question in a relevant time frame and our desired location, we only examined data from the year 2022 and within New York state. This data intends to represents the population of New York state in 2022, providing a snapshot of the health and socioeconomic characteristics of individuals residing in the state during that year.

Predictors and Outcome

In our visualization to investigate our research question, we use a few predictors. Our first predictor is "Income Class", which we categorize into three classes: "Lower Class", "Middle Class", and "Upper Class." To determine these income class categories, we used the `INCWAGE` variable from our data set, which reports each census respondent's total pre-tax wage and salary income, and divided the incomes into "Lower Class" if the individual had an income of $30,000 or less, "Middle Class" if the individual had an income greater than $30,000 but less than or equal to $94,000, and "Upper Class" if the individual had an income greater than $94,000. We obtained these income boundaries from the Census Bureau's "Income in the United States: 2022" report. Our second predictor is "Race," which we obtained from our data set's `RACE` variable to be categorized as "White", "Black/African American", "American Indian or Alaska Native", "Chinese", "Japanese", and "Other Asian or Pacific Islander." The outcome variable in our visualization is "Type of Health Insurance" which we classified as "Private", "Public", "Both" or "None." These were determined from our data set's variables `HCOVPUB` (public health coverage) and `HCOVPRIV` (private health coverage). We use this outcome variable to

represent health care access because health insurance coverage serves as a practical proxy for access, reflecting financial barriers and influencing individuals' utilization of health services. Disparities in insurance coverage by income and race can highlight underlying mechanisms contributing to unequal access. By examining these intersections, we gain insight into how socioeconomic factors shape individuals' ability to obtain and utilize health care services.

Data Cleaning and Sample Restrictions

Initially, individuals are categorized into income classes based on their wage income (`INCWAGE`), with "Lower Class" for incomes less than or equal to $30,000, "Middle Class" for incomes between $30,000 and $94,000, and "Upper Class" for incomes exceeding $94,000. Cases with missing or undefined income values are labeled as "Unknown". No cases were dropped after this step. Subsequently, the data is filtered to include observations with positive income values and belonging to the year 2022. Additionally, the racial variable (`RACE`) is recoded into more interpretable labels, and binary indicators (`HCOVPUB` and `HCOVPRIV`) for public and private health insurance coverage are created. The `insurance_type` variable is then derived based on combinations of these indicators, indicating whether individuals have private, public, both, or no health insurance. Observations with missing values in `insurance_type` or `RACE` are subsequently removed from the dataset. This resulted in 1539577 cases being dropped. Additionally, to include percentages in the visualization, percentages were calculated as the weighted count of individuals (weighted by the `PERWT` variable in our data set) divided by the total weighted count of individuals in each income class, insurance type, and race group. The separate summary data set from this resulted in 38323 cases being dropped.

Analysis

In our visualization, the summary statistic we use is the percentage/proportion of individuals within each combination of income class and insurance type group, broken down by race. So, for each combination of income class and insurance type, the percentage represents the proportion of a given race within that group relative to the total population of individuals in that income class and insurance type. This allows for a comparison of the distribution of health insurance types across different income classes and racial groups, while accounting for the survey weights.

Our visualization shows that across all income classes, people of the White race represent the highest proportion of private health insurance holders. We can see that within "Upper Class" individuals that have private health insurance, over half (64.350%) of those individuals are "White." Additionally, within "Lower Class" individuals that have public health insurance, "Black/African Americans" represent the greatest proportion (35.970%). It is also interesting to note that "Japanese" individuals represent the smallest proportion across all groups of health insurance and income classes. Furthermore, the groups that represent the largest proportions across all income and health insurance category combinations are "White" and "Black/African

American" individuals. This data suggests that income and racial inequalities significantly influence access to health care services, with disparities evident in health insurance coverage among different demographic groups.