DATA 606 Data Project Proposal

AJ Strauman-Scott

Research Question

Which demographic and socio-economic differentials have the most prominent effect on mortality rates within the United States?

Data Source

To answer my research question, I will examine the National Longitudinal Mortality Study. The National Longitudinal Mortality Study (NLMS) is a national, longitudinal, mortality study sponsored by the National Heart, Lung, and Blood Institute the National Cancer Institute, the National Institute on Aging (all part of the National Institutes of Health), the National Center for Health Statistics (part of the Center for Disease Control and Prevention) and the U.S. Census Bureau. The NLMS consists of a database developed for the purpose of studying the effects of demographic and socio-economic characteristics on differentials in U.S. mortality rates. It consists of Annual Social and Economic Supplements which cover the period from March 1973 to March 2011, Current Population Surveys for February 1978, April 1980, August 1980, December 1980, and September 1985, and one 1980 Census cohort, 39 cohorts in all. These are combined with death certificate information to identify mortality status and cause of death.

Data collection

The NLMS is a collection of Annual Social and Economic Supplements which cover the period from March 1973 to March 2011, Current Population Surveys for February 1978, April 1980, August 1980, December 1980, and September 1985, and one 1980 Census cohort, 39 cohorts in all. These are combined with death certificate information to identify mortality status and cause of death.

This data has come from the US Census, and the National Institute of Health.

Type of study

This study will be observational.

Dependent Variable

The dependent variable is life expectancy

Independent Variable(s)

The independent variable are the different identities and situations of Americans. How do these influence average life expectancy within the population?

Relevant summary statistics

```
library(tidyverse)

nlms_11 <- read_csv('/Users/opportunity/Documents/MSDS/Spring2024/DATA606/project/data/NLMS_PublicUse_5
```

The 11-year follow up file consists of a subset of the 39 NLMS cohorts included in the full NLMS that can be followed prospectively for 11 years. We'll examine this data for our EDA and summary statistics.

The following is a comparison of the NLMS survey's demographics proportions to the current census proportions.

Keep in mind the data is merged from interviews that happened anytime from 1980 to 2000. The dates in the publicly available dataset have been standardized for anonymity, so while we know our sample' is disproportionately white's numbers, calculating exactly how those numbers compare to the population proportion involves averaging values across several decades or arbitrarily chosing a point for comparison.

```
nlms_11 |>
  group_by(sex) |>
  summarise(count = n(),
            proportion = round(((n() / nrow(nlms_11)) * 100), 1))
```

```
## # A tibble: 2 x 3
##     sex  count proportion
##   <dbl>  <int>      <dbl>
## 1     1 880617         48
## 2     2 954455         52
```

There are 880,617 men (48%) and 954,455 women (52%) observed for 11 years in this data. Currently women make up 50.4% of Americans.

```
nlms_11 |>
  group_by(race) |>
  summarise(count = n(),
            proportion = round(((n() / nrow(nlms_11)) * 100), 1))
```

```
## # A tibble: 6 x 3
##    race   count proportion
##   <dbl>   <int>      <dbl>
## 1     1 1584581       86.3
## 2     2  180475        9.8
## 3     3   18859        1
## 4     4   43128        2.4
## 5     5    5139        0.3
## 6    NA    2890        0.2
```

This data contains:

- 1,584,581 white participants, or 86.3%
- 180,475 Black participants, or 9.8%

- 18,859 American Indian or Alaskan Native participants, or 1%
- 43,128 Asian or Pacific Islander participants, or 2.4%
- 5,139 participants who identify as "Other, nonwhite", or 0.3% and
- 2,890 participants who do not have race identifying data, or 0.2%

These numbers differ with some significance from today's census proportions of race among Americans, with white representing 75.5%, Black representing 13.6%, Asian at 6.3% and mixed race at 3%. Our data is overly white compared to today's population.

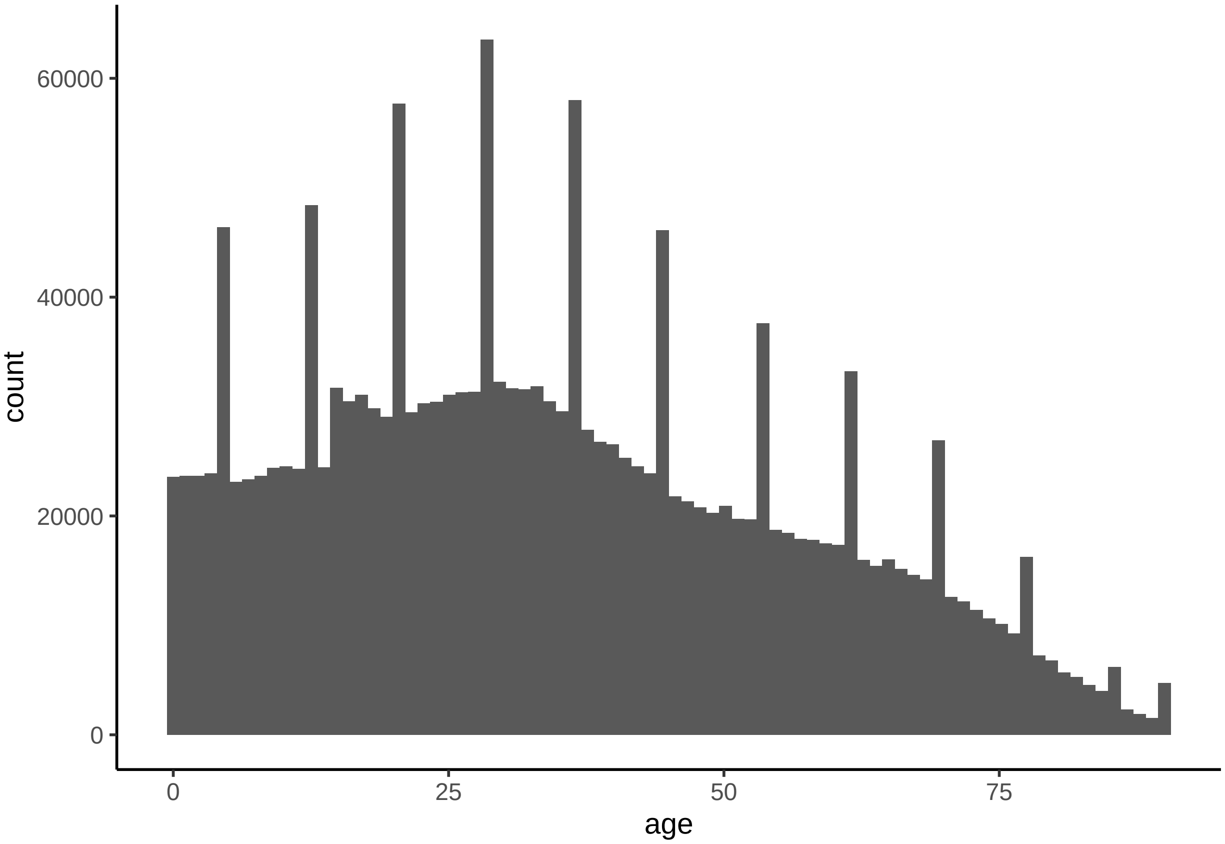
```
nlms_11 |>
  group_by(ms) |>
  summarise(count = n(),
            proportion = round(((n() / nrow(nlms_11)) * 100), 1))
```

```
## # A tibble: 6 x 3
##      ms  count proportion
##   <dbl>  <int>      <dbl>
## 1     1 862222       47
## 2     2 102406        5.6
## 3     3 106600        5.8
## 4     4  32053        1.7
## 5     5 370565       20.2
## 6    NA 361226       19.7
```

AMong participants, 47% are married, 5.6% are widowed, 5.8% are divorced and 1.7% are separated.

What ages are the participants?

```
ggplot(nlms_11, aes(x=age)) +
  geom_histogram(bins=80) +
  theme_classic()
```



I struggled to find an appropriate binwidth to accurately demonstrate the shape of the distribution. There are several ages with outlier numbers scatters among the distribution. This will require more investigation.

How many participants have health insurance?

```
nlms_11 |>
  group_by(histatus) |>
  summarise(count = n(),
            proportion = round(((n() / nrow(nlms_11)) * 100), 1))
```

```
## # A tibble: 3 x 3
##   histatus   count proportion
##      <dbl>   <int>      <dbl>
## 1        0  202642       11
## 2        1 1054665       57.5
## 3       NA  577765       31.5
```

More than half of participants have health insurance - 57.5%.

How many of the participants will die during the 11 years of observation?

```
nlms_11 |>
  group_by(inddea) |>
  summarise(count = n(),
            proportion = round(((n() / nrow(nlms_11)) * 100), 1))
```

```
## # A tibble: 2 x 3
##   inddea   count proportion
##    <dbl>   <int>      <dbl>
## 1      0 1674322       91.2
## 2      1  160750        8.8
```

Over nine out of every ten participants will die during the 11 years of observation during this study.