DATA 490 Project

Ed Dungo

2023-10-12

Library

```
library(ipumsr)
library(writexl)
library(ggplot2)
library(readr)
library(dplyr)
library(haven)
library(openxlsx)
library(gplots)
library(scatterplot3d)
library(tidymodels)
library(ggcorrplot)
```

Import Datasets

```
ddi <- read_ipums_ddi("meps_00002.xml")
ipums_df <- read_ipums_micro(ddi)
```

```
## Use of data from IPUMS MEPS is subject to conditions including that users should cite the data appro
```

```
nchs_df <- read.csv("nchs.csv")
```

Briefly Clean IPUMS Dataset

```
# remove unnecessary columns
ipums_df <- ipums_df %>% select(-(2:16)) # in range
ipums_df <- ipums_df %>%
  select(-one_of(c('HIOTHGOVA', 'HIOTHGOVB'))) # remove specific columns

# 996 means missing values, so we remove from IPUMS dataset
ipums_df <- subset(ipums_df, AGE != 996)

# only show survey year greater than or equal to 1996 in the IPUMS dataset
nchs_df <- subset(nchs_df, YEAR >= 1996)
```

```r
# remove values "Both Sexes" in the SEX column in NCHS dataset
nchs_df <- nchs_df[!grepl("Both Sexes", nchs_df$SEX), ]

# remove values "All Races" in the RACE column in NCHS dataset
nchs_df <- nchs_df[!grepl("All Races", nchs_df$RACE), ]
```

```r
# rename RACEA to RACE in IPUMS
colnames(ipums_df)[colnames(ipums_df) == "RACEA"] <- "RACE"

# remove other races besides "White" and "Black" in the IPUMS dataset
# note that "White" and "Black makes up the majority of IPUMS,
# so it's ok to take it out
ipums_df <- ipums_df %>%
  filter(RACE == 100 | RACE == 200)

# change values of the NCHS's SEX and RACE variables
nchs_df <- nchs_df %>%
  mutate(RACE = ifelse(RACE == "White", 100,
                       ifelse(RACE == "Black", 200, RACE)))
nchs_df <- nchs_df %>%
  mutate(SEX = ifelse(SEX == "Male", 1, ifelse(SEX == "Female", 2, SEX)))

# convert ipums_df's tbl_df format to data.frame
ipums_df <- as.data.frame(ipums_df)

a <- ipums_df$HIPRIVATE
b <- ipums_df$HICHAMPANY
c <- ipums_df$HIMACHIP
d <- ipums_df$HIMCARE

# make a new column called insurance types
ipums_df$INSURANCETYPE <-
  ifelse(a == 2 & b == 2 | a == 2 & c == 2 | a == 2 & d == 2 |
           b == 2 & c == 2 | b == 2 & d == 2 | c == 2 & d == 2,
         "Two Insurance",
         ifelse(a == 2, "Private",
                ifelse(b == 2, "CHAMPUS, TRICARE, or CHAMP-VA",
                       ifelse(c == 2, "Medicaid and/or SCHIP",
                              ifelse(d == 2, "Medicare",
                                     "No Insurance")))))

# Remove health coverage columns
ipums_df <- ipums_df %>% select(-(5:9)) # in range

# convert data types
nchs_df$RACE <- as.integer(nchs_df$RACE)
nchs_df$SEX <- as.integer(nchs_df$SEX)

# merge datasets on YEAR, RACE, and SEX
merged_df <- ipums_df %>%
  left_join(nchs_df, by = c("YEAR", "RACE", "SEX"))

# remove empty values from the merged_df
```

```r
merged_df <- merged_df[complete.cases(merged_df), ]

# change NCHS variables' values
nchs_df <- nchs_df %>%
  mutate(RACE = ifelse(RACE == 100, "White",
                       ifelse(RACE == 200, "Black", RACE)))
nchs_df <- nchs_df %>%
  mutate(SEX = ifelse(SEX == 1, "Male",
                      ifelse(SEX == 2, "Female", SEX)))

# change IPUMS variables' values
ipums_df <- ipums_df %>%
  mutate(RACE = ifelse(RACE == 100, "White",
                       ifelse(RACE == 200, "Black", RACE)))
ipums_df <- ipums_df %>%
  mutate(SEX = ifelse(SEX == 1, "Male",
                      ifelse(SEX == 2, "Female", SEX)))

# change merged_df variables' values
merged_df <- merged_df %>%
  mutate(RACE = ifelse(RACE == 100, "White",
                       ifelse(RACE == 200, "Black", RACE)))
merged_df <- merged_df %>%
  mutate(SEX = ifelse(SEX == 1, "Male", ifelse(SEX == 2, "Female", SEX)))

# rename columns in merged_df
colnames(merged_df) <- c("Year", "Age", "Sex", "Race", "Insurance Type",
                         "Average Life Expectancy", "Age-Adjusted Death Rate")
```

```r
save(merged_df, file = "merged_df.RData")
write.csv(merged_df, file = "merged_df.csv")
save(ipums_df, file = "ipums_df.RData")
save(nchs_df, file = "nchs_df.RData")
```