in_class_week2_sols.R

moffer

2022-10-07

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
# download the data from this website: https://data.cdc.gov/NCHS/Provisional-COVID-19-Deaths-by-Race-an
# and read it in locally, using `read_csv()`
dat <- read_csv('../data/Provisional_COVID-19_Deaths_by_Race_and_Hispanic_Origin__and_Age.csv')
```

```
## Rows: 6489 Columns: 13
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (7): Data as of, Start Date, End Date, State, Age group, Race and Hispan...
## dbl (6): COVID-19 Deaths, Total Deaths, Pneumonia Deaths, Pneumonia and COVI...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# check out the data using `dat`, `summary(dat)`, and `View(dat)`.
dat
```

```
## # A tibble: 6,489 x 13
##    `Data as of` `Start Date` `End Date` State       `Age group` `Race and Hisp~`
##    <chr>        <chr>        <chr>      <chr>       <chr>       <chr>
##  1 10/05/2022   01/01/2020   10/01/2022 United Sta~ All Ages    Total Deaths
##  2 10/05/2022   01/01/2020   10/01/2022 United Sta~ All Ages    Non-Hispanic Wh~
##  3 10/05/2022   01/01/2020   10/01/2022 United Sta~ Under 1 ye~ Non-Hispanic Wh~
##  4 10/05/2022   01/01/2020   10/01/2022 United Sta~ 0-17 years  Non-Hispanic Wh~
##  5 10/05/2022   01/01/2020   10/01/2022 United Sta~ 1-4 years   Non-Hispanic Wh~
##  6 10/05/2022   01/01/2020   10/01/2022 United Sta~ 5-14 years  Non-Hispanic Wh~
##  7 10/05/2022   01/01/2020   10/01/2022 United Sta~ 15-24 years Non-Hispanic Wh~
##  8 10/05/2022   01/01/2020   10/01/2022 United Sta~ 18-29 years Non-Hispanic Wh~
##  9 10/05/2022   01/01/2020   10/01/2022 United Sta~ 25-34 years Non-Hispanic Wh~
## 10 10/05/2022   01/01/2020   10/01/2022 United Sta~ 30-49 years Non-Hispanic Wh~
## # ... with 6,479 more rows, and 7 more variables: `COVID-19 Deaths` <dbl>,
## #   `Total Deaths` <dbl>, `Pneumonia Deaths` <dbl>,
## #   `Pneumonia and COVID-19 Deaths` <dbl>, `Influenza Deaths` <dbl>,
```

```
## #   `Pneumonia, Influenza, or COVID-19 Deaths` <dbl>, Footnote <chr>
```

```
summary(dat)
```

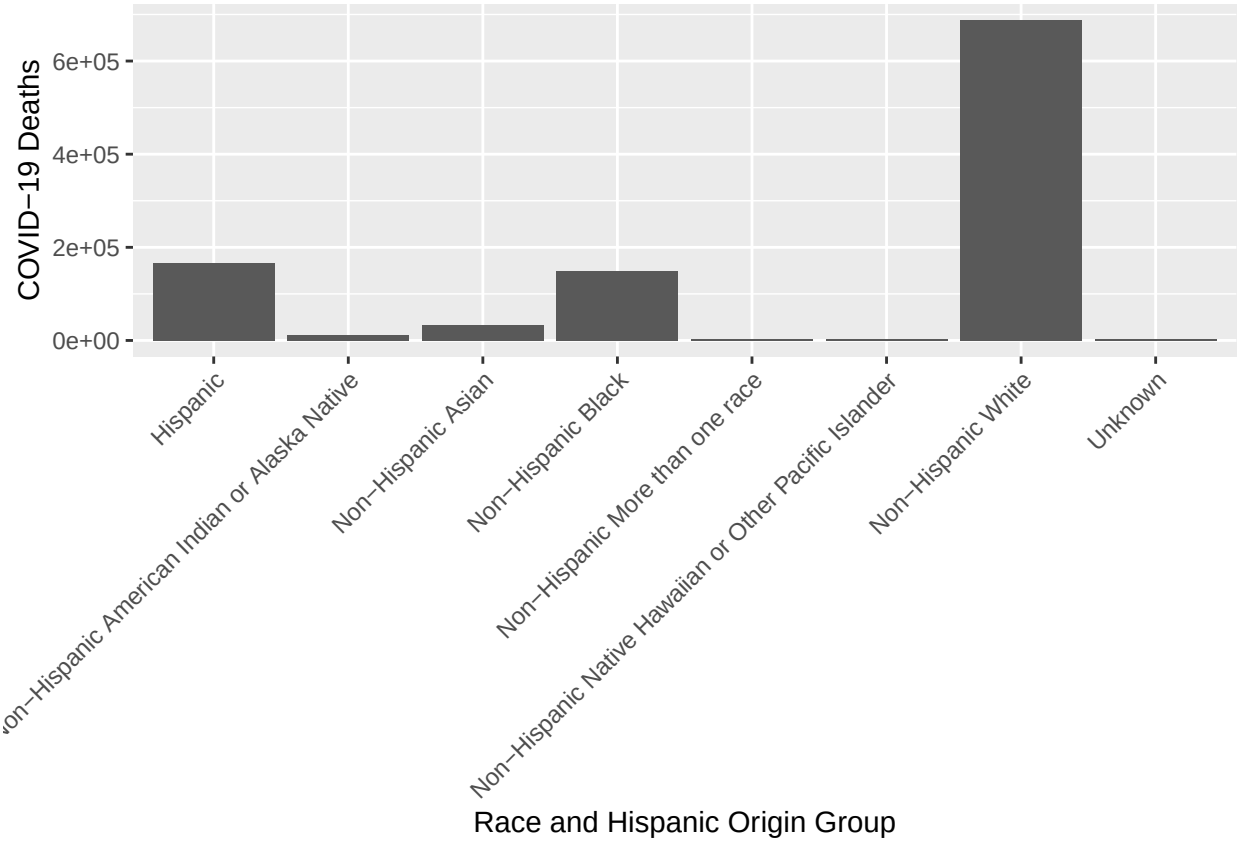
```
##   Data as of         Start Date          End Date            State
##  Length:6489        Length:6489        Length:6489        Length:6489
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   Age group         Race and Hispanic Origin Group COVID-19 Deaths
##  Length:6489        Length:6489                    Min.   :      0
##  Class :character   Class :character               1st Qu.:      0
##  Mode  :character   Mode  :character               Median :     12
##                                                    Mean   :   1058
##                                                    3rd Qu.:     97
##                                                    Max.   :1055967
##                                                    NA's   :1995
##   Total Deaths     Pneumonia Deaths Pneumonia and COVID-19 Deaths
##  Min.   :      0   Min.   :     0   Min.   :     0
##  1st Qu.:     24   1st Qu.:     0   1st Qu.:     0
##  Median :    106   Median :    11   Median :     0
##  Mean   :   8034   Mean   :   985   Mean   :   517
##  3rd Qu.:    770   3rd Qu.:    88   3rd Qu.:    42
##  Max.   :9225883   Max.   :969256   Max.   :536751
##  NA's   :1269      NA's   :2068     NA's   :1783
##  Influenza Deaths   Pneumonia, Influenza, or COVID-19 Deaths   Footnote
##  Min.   :    0.00   Min.   :      0.0                        Length:6489
##  1st Qu.:    0.00   1st Qu.:      0.0                        Class :character
##  Median :    0.00   Median :     18.0                        Mode  :character
##  Mean   :   10.79   Mean   :   1518.4
##  3rd Qu.:    0.00   3rd Qu.:    137.8
##  Max.   :12381.00   Max.   :1499035.0
##  NA's   :1419       NA's   :2059
```

```
# View(dat)

# filter the data so that you're just using data for the United states as a whole,
# and drop observations where the column `Race and Hispanic Origin Group` is 'Total Deaths'
dat <- dat %>%
  filter(State == 'United States',
         `Race and Hispanic Origin Group` != 'Total Deaths')

# Consider the age groups in this data. Do you want to keep all of them?
# Filter out any age groups you don't want to keep.
table(dat$`Age group`)
```

```
##
##        0-17 years         1-4 years       15-24 years       18-29 years
##                 8                 8                 8                 8
##       25-34 years       30-49 years       35-44 years       45-54 years
##                 8                 8                 8                 8
##        5-14 years       50-64 years       55-64 years       65-74 years
```

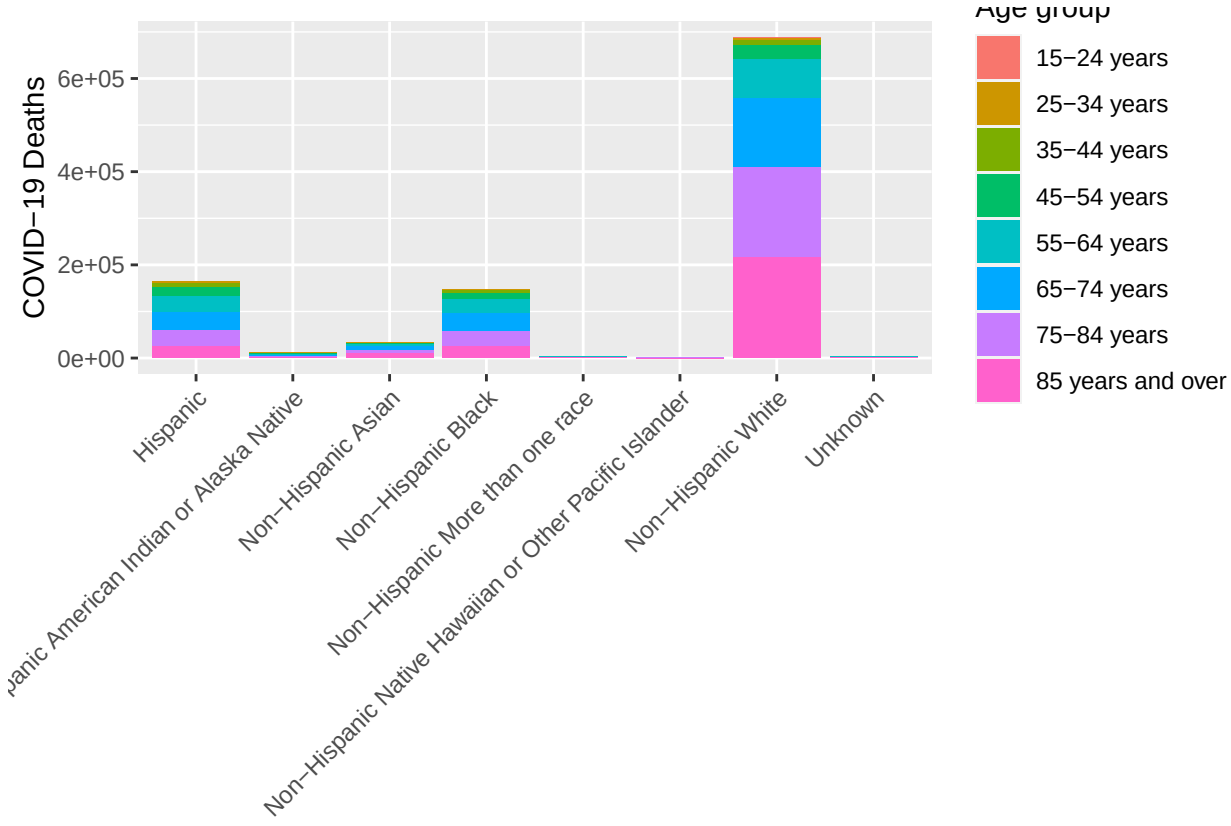
```
##                 8                 8                 8                 8
##       75-84 years 85 years and over         All Ages      Under 1 year
##                 8                 8                 8                 8
```

```
dat <- dat %>%
  filter(`Age group` %in% c('15-24 years', '25-34 years', '35-44 years',
                            '45-54 years', '55-64 years', '65-74 years',
                            '75-84 years', '85 years and over', 'All Ages'))
# create a data set, `dat0`, that is just the data where the column `Age group`
# is 'All Ages'
dat0 <- dat %>%
  filter(`Age group` == 'All Ages')

# create another data set, `dat1`, that is the date where the column `Age group`
# is anything but 'All Ages'

dat1 <- dat %>%
  filter(`Age group` != 'All Ages')

# in dat0, plot COVID-19 deaths by `Race and Hispanic Origin Group` using `geom_col`.
# put the x axis labels at 45 degrees (you can google how to do this)
ggplot(dat0, aes(y = `COVID-19 Deaths`,
                 x = `Race and Hispanic Origin Group`)) +
  geom_col() +
  theme(axis.text.x = element_text(angle = 45, hjust=1))
```

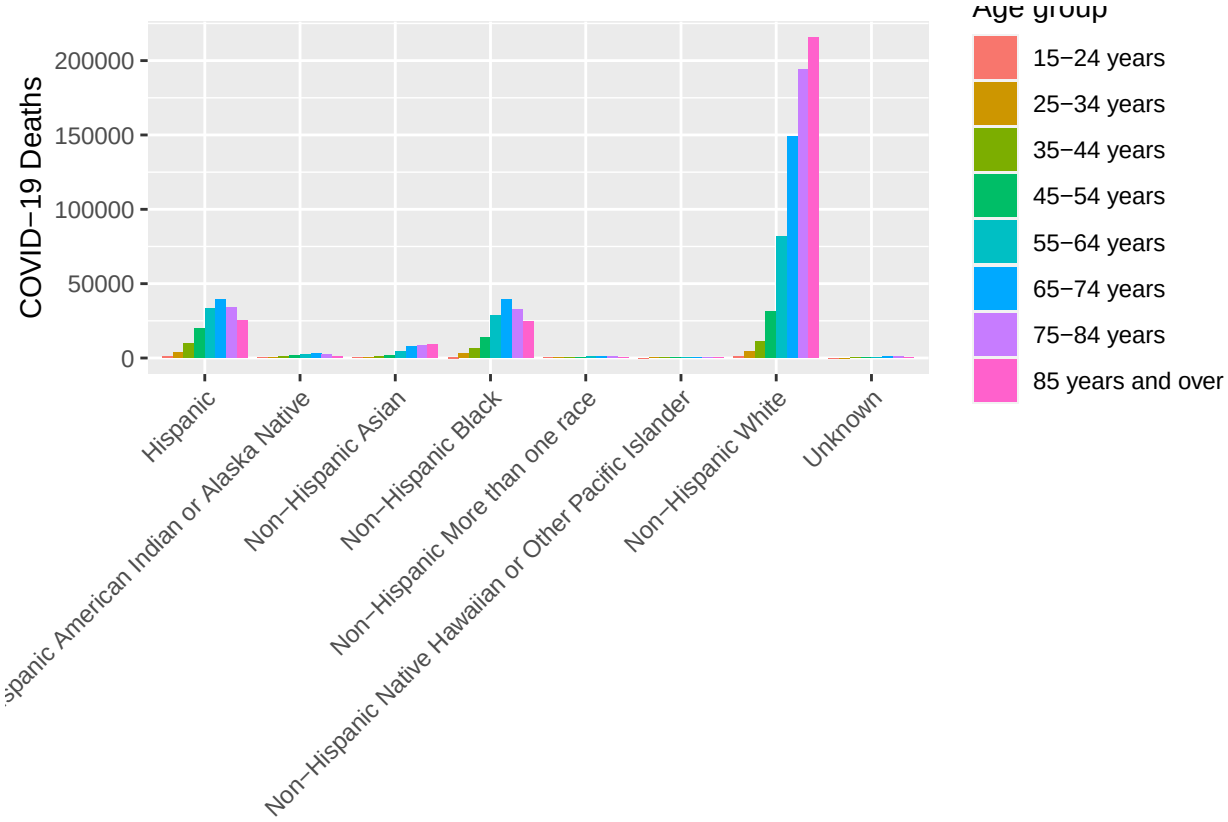



```
# in dat1, plot COVID-19 deaths by `Race and Hispanic Origin Group` using `geom_col`,
# AND set the fill by `Age group`
```

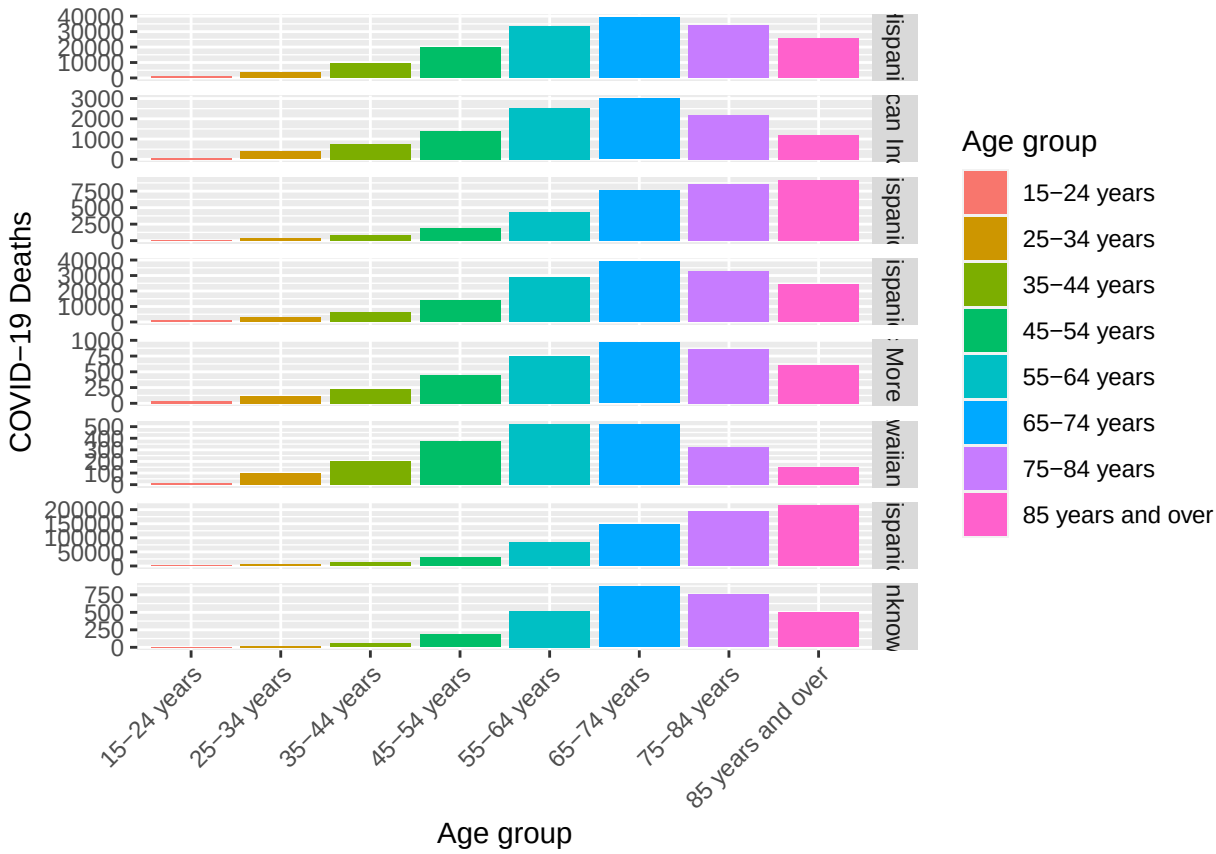
```
ggplot(dat1, aes(y = `COVID-19 Deaths`,
                 x = `Race and Hispanic Origin Group`,
                 fill = `Age group`)) +
  geom_col() +
  theme(axis.text.x = element_text(angle = 45, hjust=1))
```




```
# Do the last plot again, but try using the argument `position = 'dodge'` in
# `geom_col()`. Which do you like better? Why?
ggplot(dat1, aes(y = `COVID-19 Deaths`,
                 x = `Race and Hispanic Origin Group`,
                 fill = `Age group`)) +
  geom_col(position = 'dodge') +
  theme(axis.text.x = element_text(angle = 45, hjust=1))
```




```
# Try using `facet_grid()` where you have separate plots for each race and
# hispanic origin group, with COVID deaths on the Y axis, and age on the X axis
ggplot(dat1, aes(y = `COVID-19 Deaths`,
                 x = `Age group`,
                 fill = `Age group`)) +
  facet_grid(vars(`Race and Hispanic Origin Group`), scales = 'free') +
  geom_col(position = 'identity') +
  theme(axis.text.x = element_text(angle = 45, hjust=1))
```




```
# Which plot do you think is the best way to summarize this data? Why?

# What other data would you need to say something about the relative risk for
# by age and race/hispanic origin?

# Use ggsave() to save your best plot, and email it to mollyow@uchicago.edu
# with this script, and the names of students in your group.
```