Fixing Visualization Pitfalls

INFO 526 Data Analysis and Visualization

Adriana Picoral

Data on homicides by firearms in the US

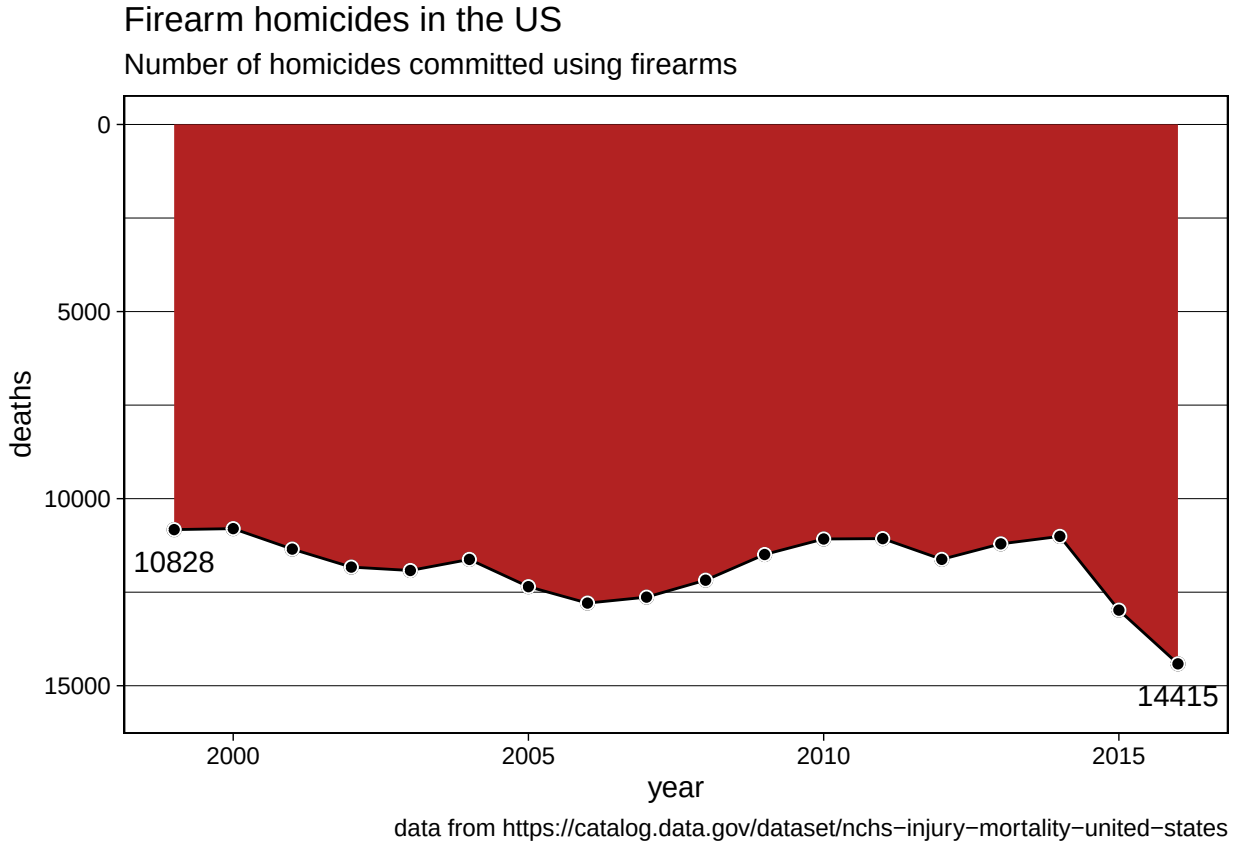
In this tutorial we will be using a subset of NCHS - Injury Mortality: United States data, focusing on mortality related to firearms with homicide as the intent, with data aggregated for all sexes, ages, and races.

```
library(tidyverse)
firearm_homicides <- read_csv("data/firearm-homicide-injury-deaths.csv")
```

Bad plot

You will find below a replication of the 2014 Gun Deaths in Florida plot, with inverted y axis.

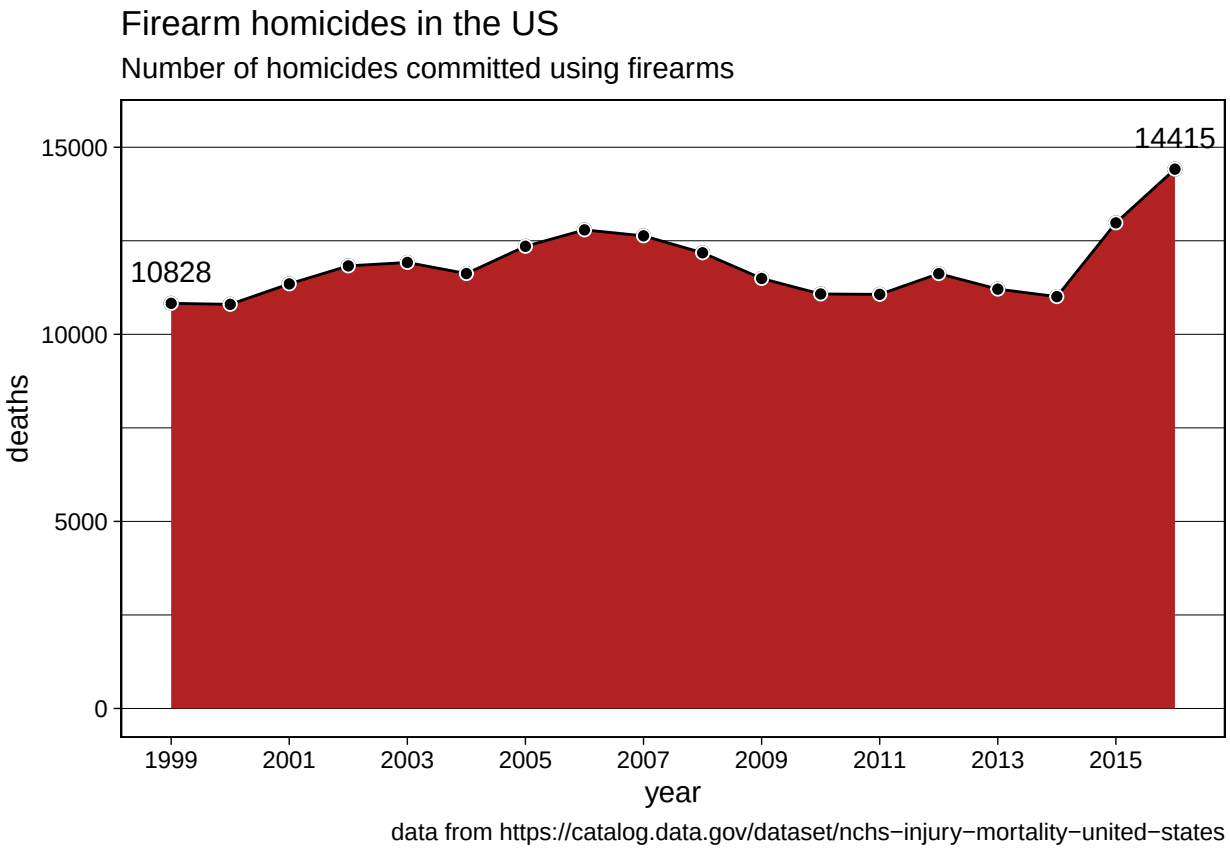
Although the original plot had labels for the first and last data points, the plot is still misleading because the y axis is rarely inverted, and at a quick glance, the number of homicides by firearm appears to be going down, when it's actually going up.



Fixed plot

Several elements in the original plot are good – the use of points in addition to a line, the red shading makes it striking, and the labels are mostly helpful. Here's how the plot would look with the not inverted y axis and more labels for years:

```
firearm_homicides %>%
  mutate(annotation = ifelse(year == 1999 |
                               year == 2016,
                             deaths, NA)) %>%
  ggplot(aes(x = year,
             y = deaths)) +
  geom_area(fill = "firebrick") +
  geom_line() +
  geom_point(size = 2) +
  geom_point(size = 2,
             color = "white",
             shape = 21) +
  geom_text(aes(label = annotation,
                vjust = -1)) +
  scale_y_continuous(limits = c(0, 15500)) +
  scale_x_continuous(breaks = seq(1999, 2016, by = 2)) +
  theme_linedraw() +
  theme(panel.grid.major.x = element_blank(),
        panel.grid.minor.x = element_blank()) +
  labs(title = "Firearm homicides in the US",
       subtitle = "Number of homicides committed using firearms",
       caption = "data from https://catalog.data.gov/dataset/nchs-injury-mortality-united-states")
```



Google Mobility data

In this part of the tutorial we will be using Google Mobility Data for the US. The raw data downloaded from Google's website was merged, pivoted, and cleaned.

```
us_mobility_data <- read_csv("data/us-mobility-data.csv")
glimpse(us_mobility_data)
```

```
## Rows: 7,051,256
## Columns: 5
## $ date                 <date> 2020-02-15, 2020-02-15, 2020-02-15, 2020-02-15, ~
## $ year                 <dbl> 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2~
## $ month                <chr> "February", "February", "February", "February", "~
## $ place                <chr> "retail and recreation percent", "grocery and pha~
## $ change_from_baseline <dbl> 6, 2, 15, 3, 2, -1, 7, 1, 16, 2, 0, -1, 6, 0, 28,~
```

Bad plot

In the stacked bar below, the mean change from baseline (percentage of change based on pre-pandemic baseline) is calculated by month, year, and place and then plotted in a stacked bar plot. Note that month is an ordered categorical variable, but it's displayed in alphabetical order. Also, the scale for each year is specific to the range of data for that year, creating different scales across years, making comparisons very difficult.

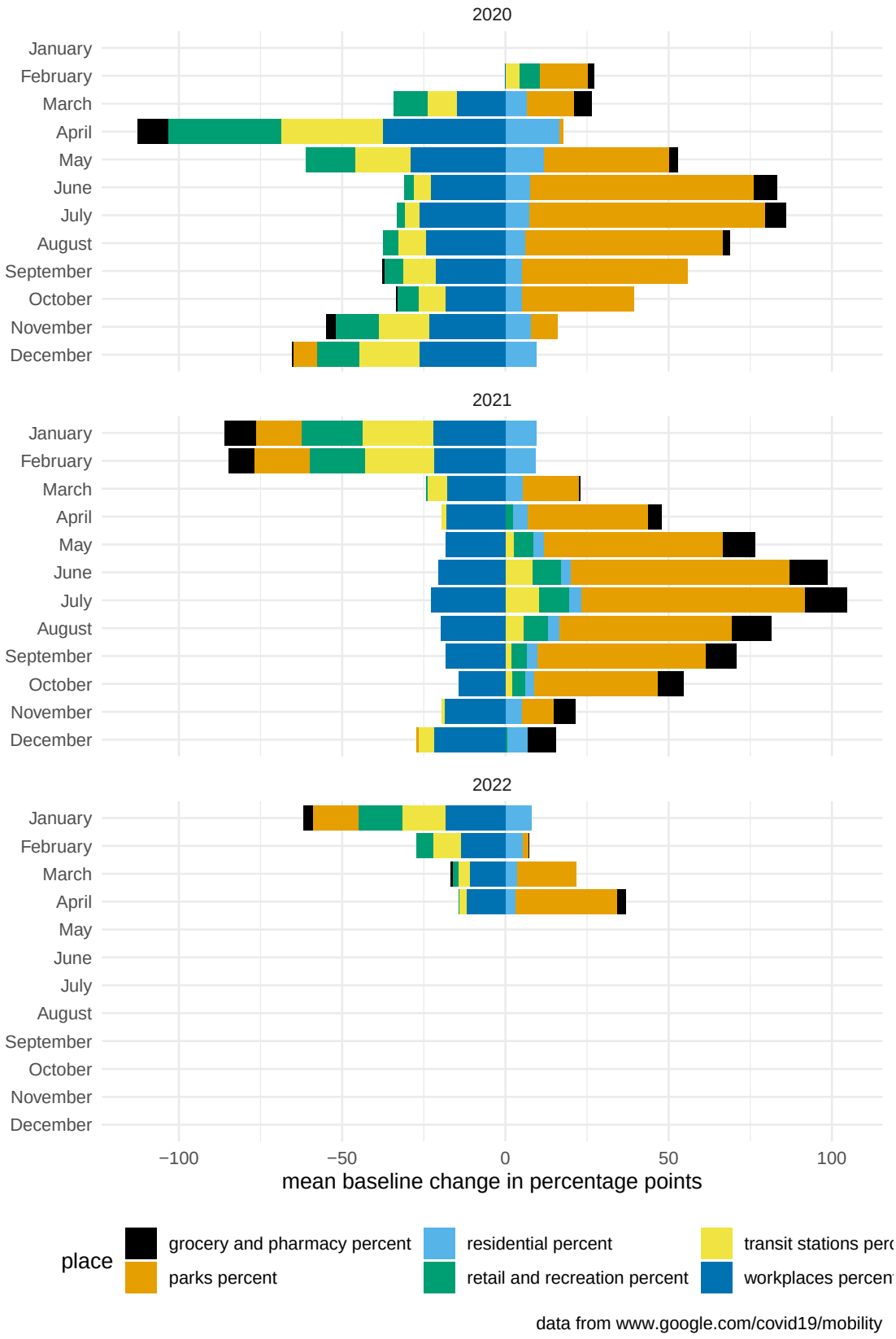
Fixed plot

We need to ensure the order of the months is correct by making the variable into a factor, with the correct order of levels. We then swap the x and y axes to have the months be read horizontally from left to right. We ensure the scales are all the same across different years, and we use a colorblind color scheme for our fill mapping. Finally we add labels that explain what we are displaying.

```
library(ggthemes)

us_mobility_data %>%
  group_by(month, year, place) %>%
  summarize(mean_baseline_change = mean(change_from_baseline)) %>%
  mutate(month = factor(month,
                        levels = c("January",
                                   "February",
                                   "March",
                                   "April",
                                   "May",
                                   "June",
                                   "July",
                                   "August",
                                   "September",
                                   "October",
                                   "November",
                                   "December"))) %>%
  ggplot(aes(y = fct_rev(month),
```

```
            x = mean_baseline_change,
            fill = place)) +
geom_col() +
scale_fill_colorblind() +
theme_minimal() +
theme(legend.position = "bottom") +
facet_wrap(~year, ncol = 1) +
labs(title = "Mobility change in percentage",
     subtitle = "in the US, across places",
     y = "",
     x = "mean baseline change in percentage points",
     caption = "data from www.google.com/covid19/mobility")
```


Mobility change in percentage
in the US, across places
2020
January
February
March
April
May
June
July
August
September
October
November
December
2021
January
February
March
April
May
June
July
August
September
October
November
December
2022
January
February
March
April
May
June
July
August
September
October
November
December
−100
−50
0
50
100
mean baseline change in percentage points
place
grocery and pharmacy percent
parks percent
residential percent
retail and recreation percent
transit stations perc
workplaces percen
data from www.google.com/covid19/mobility