Police Line of Duty Deaths

William Lovejoy

6/13/2022

Data from https://www.kaggle.com/datasets/mayureshkoli/police-deaths-in-usa-from-1791-to-2022

```
library(tidyverse)
library(ggridges)
library(treemap)
library(lubridate)
```

To start with, we load in our libraries. We'll be using lubridate to handle the dates in the dataset and tidyverse for general data work and visualizations. We will also be using ggridges and treemap for some specialized graphs.

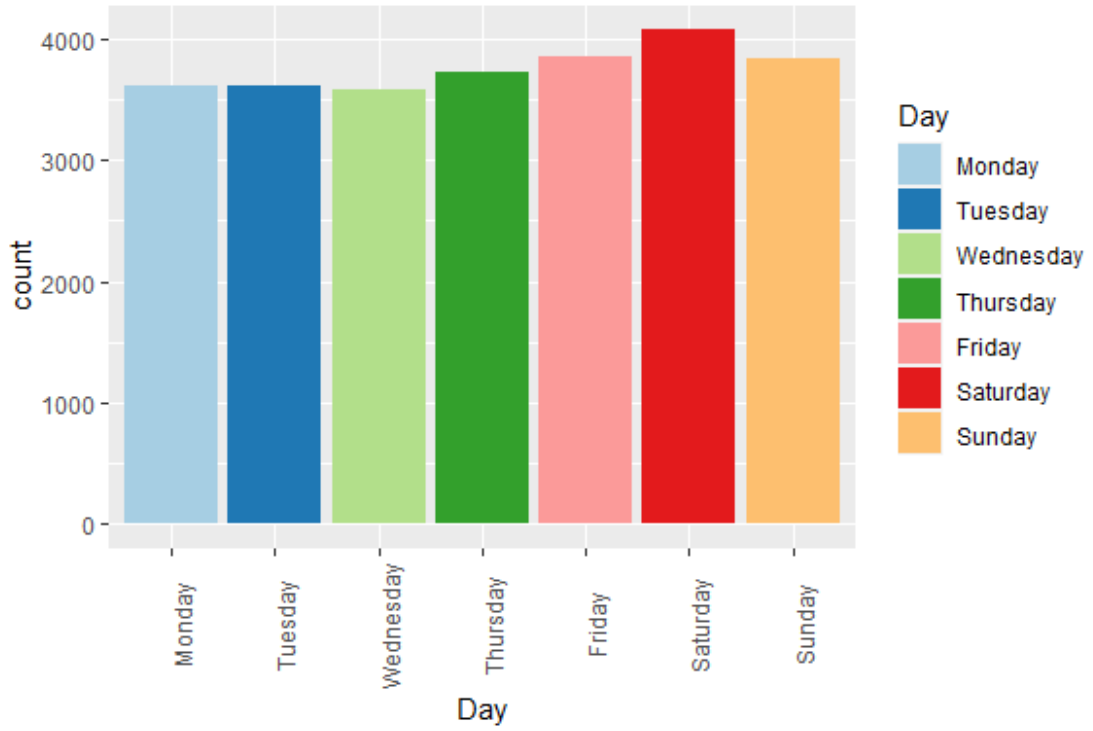
```
df <- read.csv("C:\\Users\\William Lovejoy\\Documents\\Codes\\R\\DataScience\
\police_deaths\\police_deaths_in_america.csv")
df$Year <- year(as.Date(as.character(df$Year), format = "%Y"))
df$Day <- factor(df$Day, levels = c("Monday", "Tuesday", "Wednesday",
                                    "Thursday","Friday", "Saturday",
                                    "Sunday"))
df$Month <- factor(df$Month, levels = c("January", "February", "March", "Apri
l",
                                        "May", "June", "July", "August",
                                        "September", "October", "November",
                                        "December"))
glimpse(df)
```

```
## Rows: 26,269
## Columns: 10
## $ Rank           <chr> "Constable", "Sheriff", "Deputy Sheriff", "Marshal"
, "D~
## $ Name           <chr> "Darius Quimby", "Cornelius Hogeboom", "Isaac Smith
", "~
## $ Cause_of_Death <chr> "Stabbed", "Gunfire", "Gunfire", "Gunfire", "Gunfir
e", ~
## $ Date           <chr> "Monday, January 3, 1791", "Saturday, October 22, 1
791"~
## $ Year           <dbl> 1791, 1791, 1792, 1794, 1797, 1797, 1798, 1804, 180
6, 1~
## $ Month          <fct> January, October, May, January, June, November, Sep
temb~
## $ Day            <fct> Monday, Saturday, Thursday, Saturday, Thursday, Sun
day,~
## $ Department     <chr> "Albany County Constable's Office, NY", "Columbia C
```

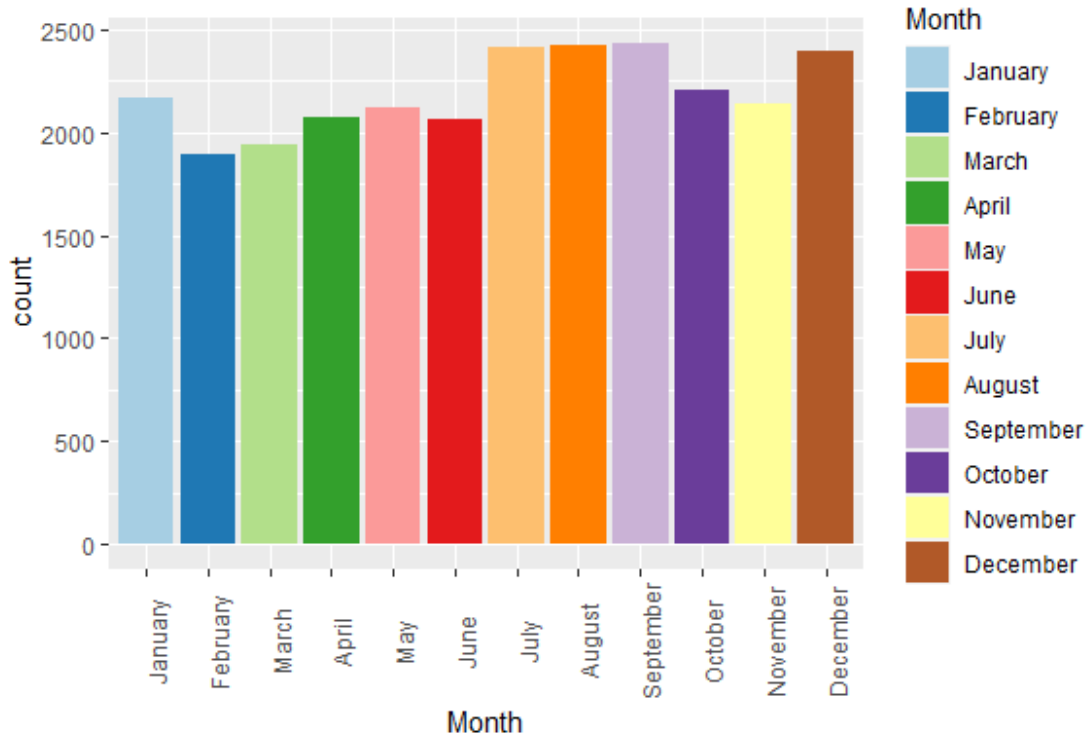
```
ount~
## $ State          <chr> "New York", "New York", "New York", "United States"
, "N~
## $ K9_Unit        <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0~
```

After loading in the data, we'll clean up the dates and create specified columns for the year, month, and weekday. Then we can start exploring the data.

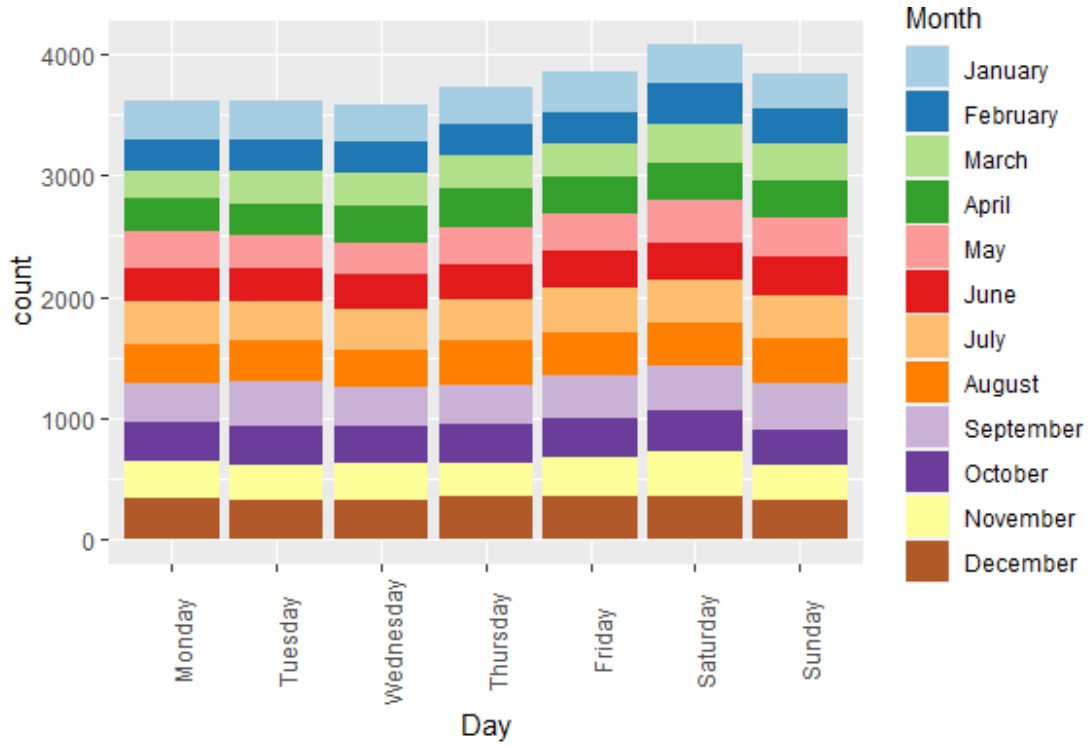
```
ggplot(df, aes(x = Day, fill = Day)) + geom_bar(position = "dodge") +
  scale_fill_brewer(palette = "Paired") +
  theme(axis.text.x = element_text(angle = 90))
```




```
ggplot(df, aes(x = Month, fill = Month)) + geom_bar(position = "dodge") +
  scale_fill_brewer(palette = "Paired") +
  theme(axis.text.x = element_text(angle = 90))
```



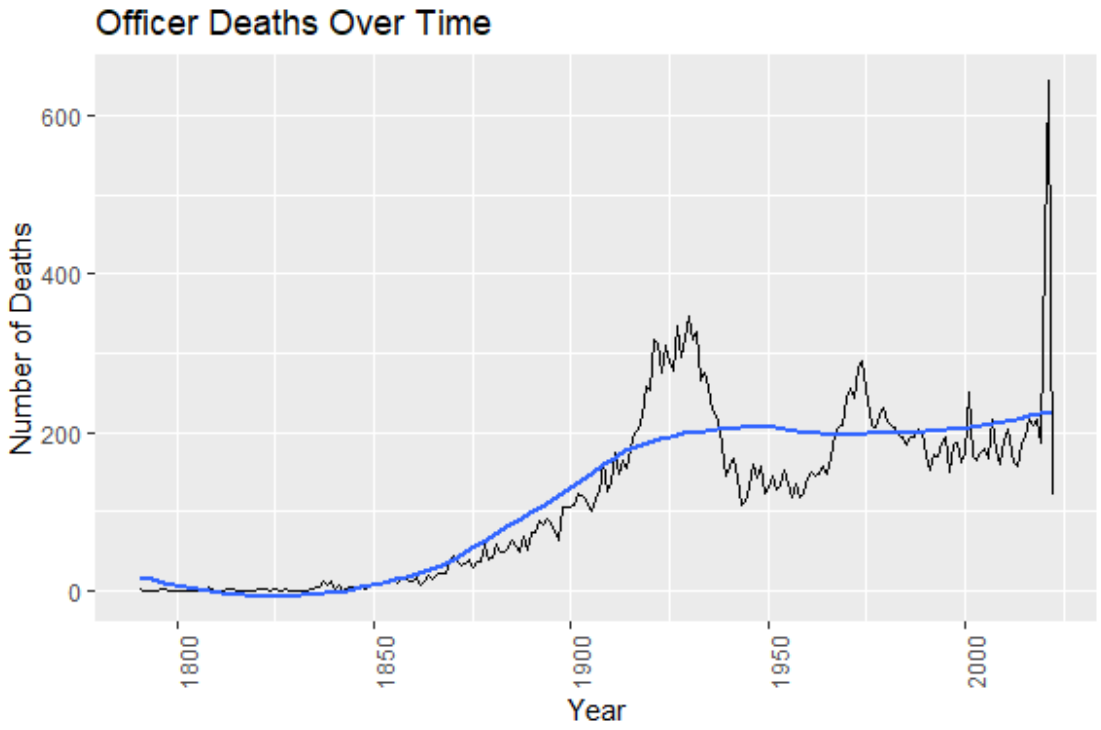

```
ggplot(df, aes(x = Day)) + geom_bar(aes(fill = Month)) +
  scale_fill_brewer(palette = "Paired") +
  theme(axis.text.x = element_text(angle = 90))
```



Our first set of graphs shows that the deadliest weekday for a cop is Saturday, and the deadliest month is September. Although neither of these is by a large amount. Friday and Sunday are almost as deadly as Saturday while July, August, and December are almost equal to September.

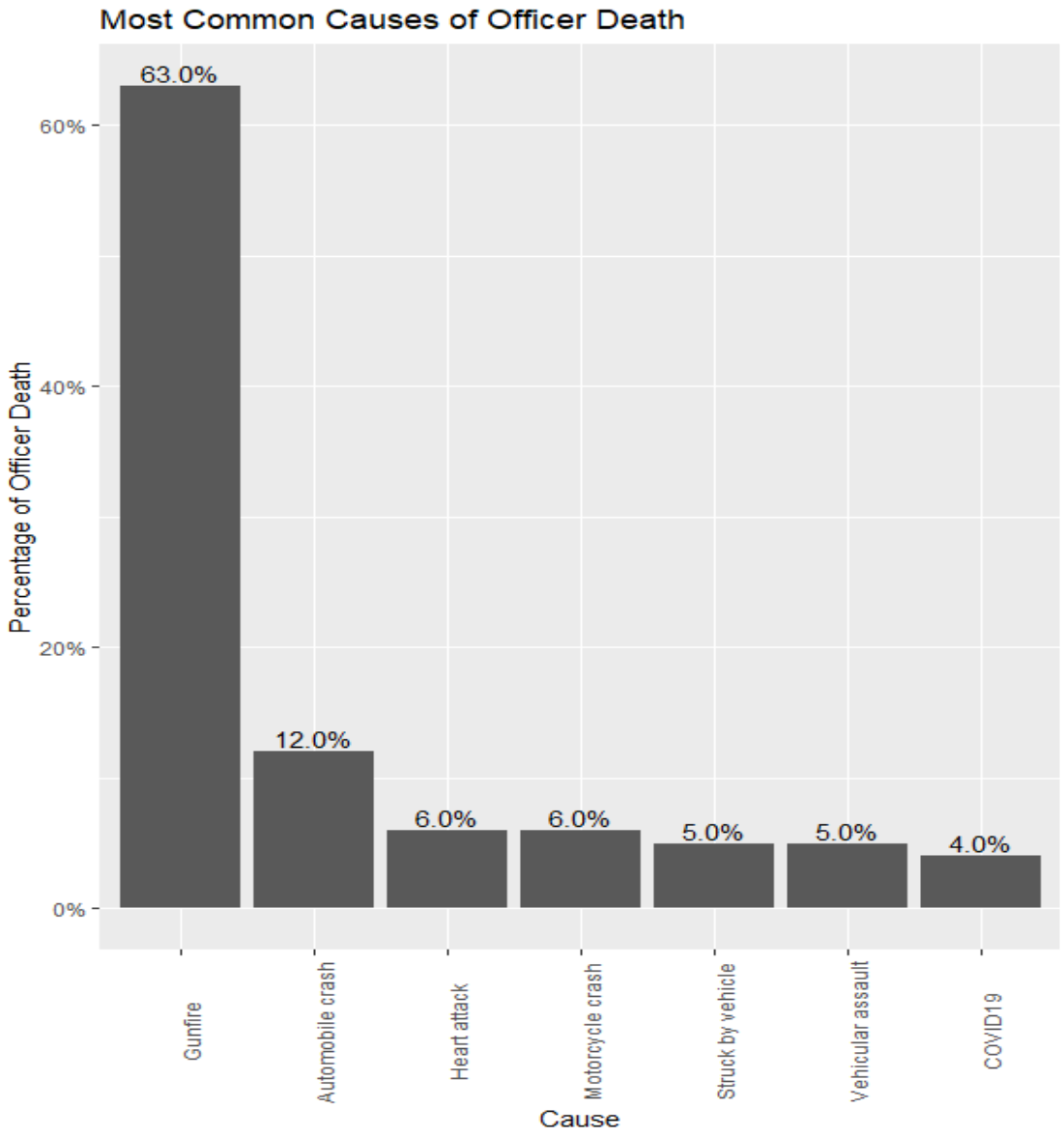
```
df %>%
  group_by(Year) %>%
  summarize(n = n()) %>%
  ungroup() %>%
  ggplot(aes(x = Year, y = n)) + geom_line() + geom_smooth(se = FALSE) +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_x_continuous() +
  labs(title = "Officer Deaths Over Time") + xlab("Year") +
  ylab("Number of Deaths")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



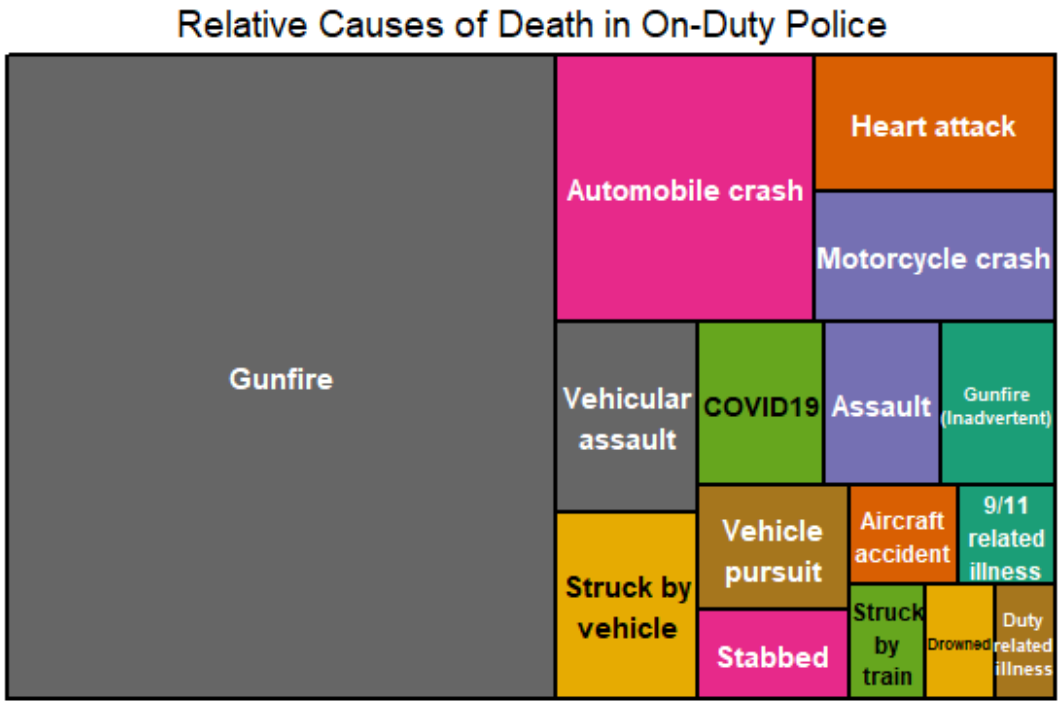
When graphing the total number of officer deaths by year, we can see a stable increase in annual deaths over time. This doesn't necessarily mean that being a US police officer is becoming more dangerous. It can simply mean that there are more police in America, so the number of deaths in the line of duty is growing while the overall proportion stays relatively constant. There are 3 notable spikes in annual deaths though. The first occurs in the 1920s, immediately after WWI. This was also during the Great Depression and the Prohibition Era. The second occurs in the early 1970s lining up with the end of the Vietnam war. The final spike is currently happening in the midst of a global pandemic. All three of these time points are marked by large scale social problems, with 2 of them situated immediately next to large war efforts.

```r
df %>%
  drop_na(Cause_of_Death) %>%
  group_by(Cause_of_Death) %>%
  count(sort = TRUE) %>%
  ungroup() %>%
  slice(1:7) %>%
  mutate(pct = round(n / sum(n), digits = 2),
         Cause_of_Death = fct_reorder(Cause_of_Death, desc(pct))) %>%
  ggplot(aes(x = Cause_of_Death, y = pct, label = pct)) + geom_col() +
  geom_text(aes(label = scales::percent(pct)), vjust = -.25) +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_y_continuous(labels = scales::percent_format(accuracy = 1)) +
  labs(title = "Most Common Causes of Officer Death") + xlab("Cause") +
  ylab("Percentage of Officer Death")
```



The most common cause of death overall is Gunfire, and by a large margin. Line of duty deaths by gunfire make up over 60% of total officer deaths between 1791 and 2022. Despite the fact that the disease has only been around for 2 and a half years (Jan. 2020 - May 2022), accounting for 1.08% of the available time, COVID19 is responsible for almost 4% of total officer deaths in the line of duty.

```
df %>%
  group_by(Cause_of_Death) %>%
  summarize(n = n()) %>%
  filter(n > 250) %>%
  ungroup() %>%
  treemap(index = "Cause_of_Death", vSize = "n", type = "index",
          palette = "Dark2",
          title = "Relative Causes of Death in On-Duty Police")
```

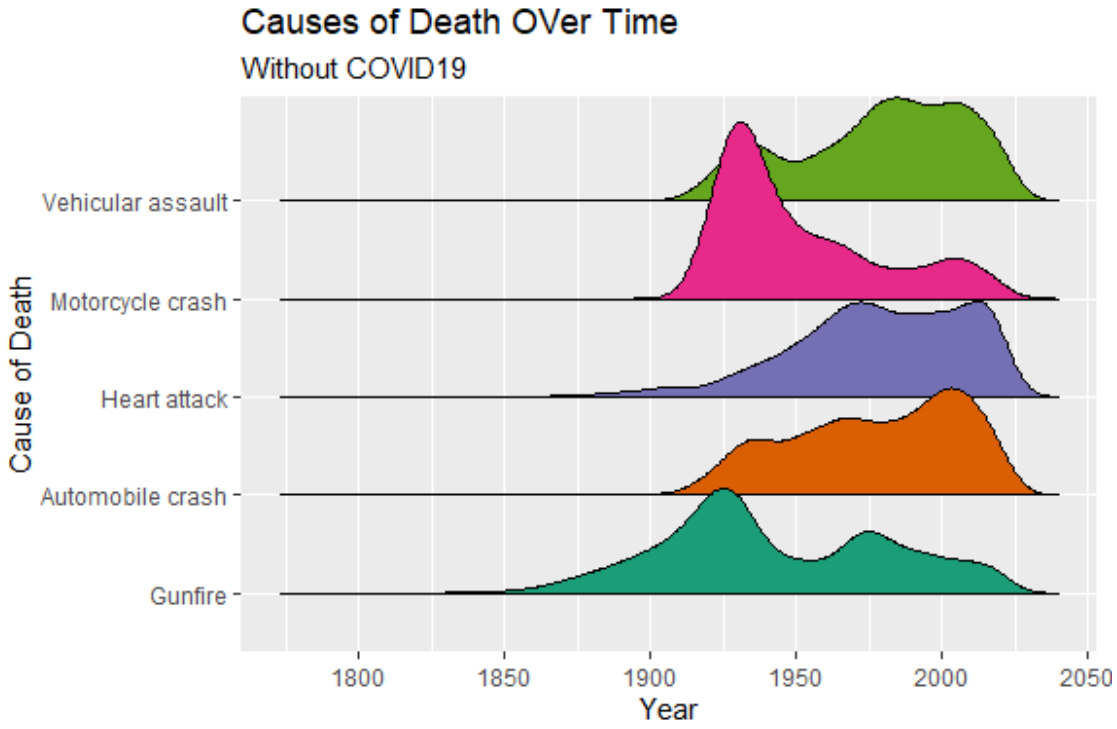


Our treemap lets us see how large each of these categories is in comparison to the whole. Gunfire takes up more than half the space, with automobile crashes coming in second.

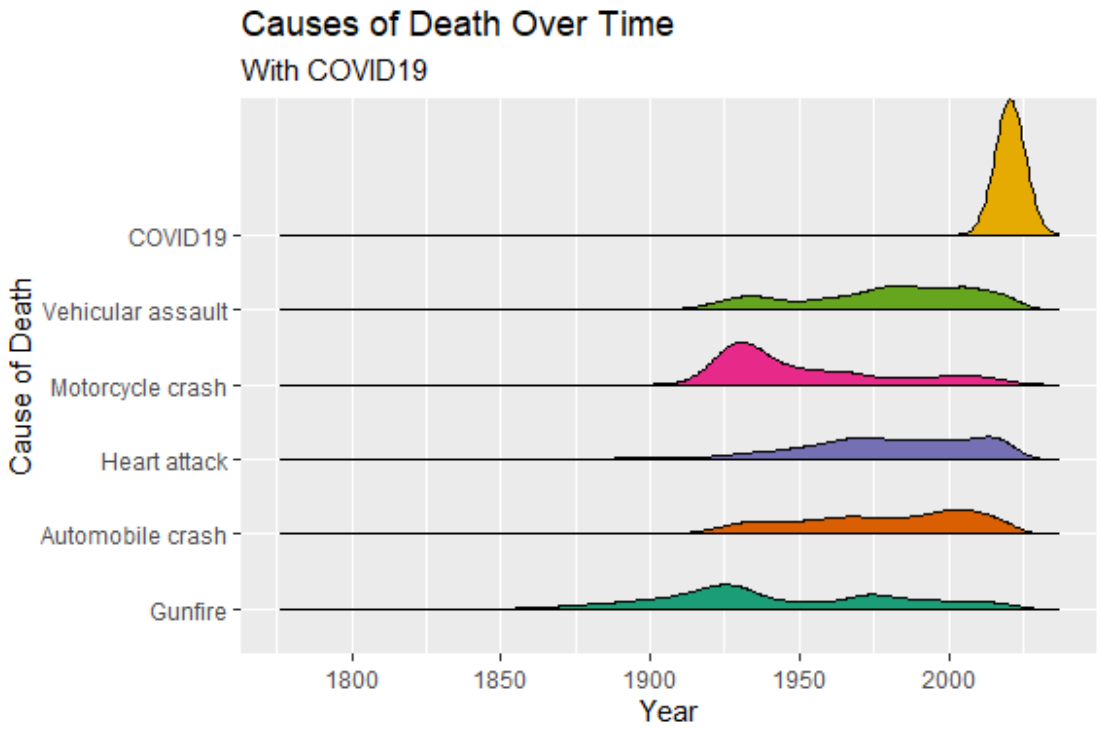
```
df$Cause_of_Death <- factor(df$Cause_of_Death,
                            levels = c("Gunfire", "Automobile crash",
                                       "Heart attack", "Motorcycle crash",
                                       "Vehicular assault", "COVID19"))


df %>%
  subset(Cause_of_Death %in% c("Gunfire", "Automobile crash", "Heart attack",
```

```
                              "Motorcycle crash", "Vehicular assault")) %>%
  ggplot(aes(x = Year, y = Cause_of_Death, fill = Cause_of_Death)) +
  geom_density_ridges2() + theme(legend.position = "none") +
  labs(title = "Causes of Death OVer Time", subtitle = "Without COVID19",
       y = "Cause of Death") + scale_fill_brewer(palette = "Dark2")
```

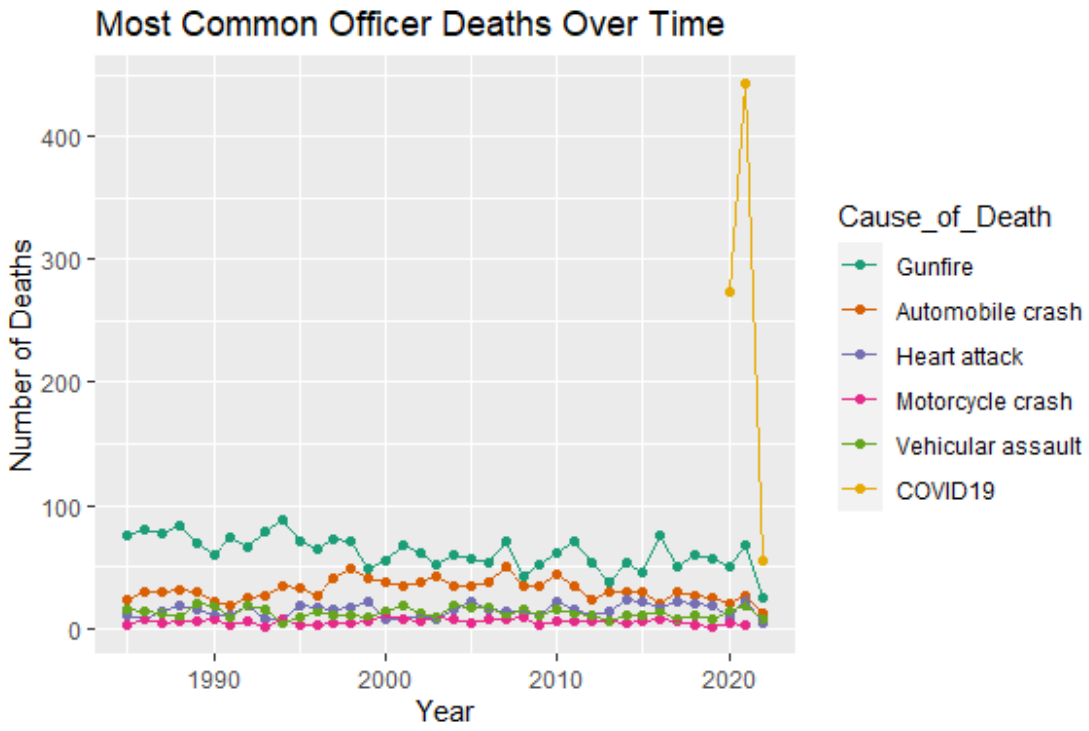



```
df %>%
  subset(Cause_of_Death %in% c("Gunfire", "Automobile crash", "Heart attack",
                               "Motorcycle crash", "Vehicular assault",
                               "COVID19")) %>%
  ggplot(aes(x = Year, y = Cause_of_Death, fill = Cause_of_Death)) +
  geom_density_ridges2() + theme(legend.position = "none") +
  labs(title = "Causes of Death Over Time", subtitle = "With COVID19",
       y = "Cause of Death") + scale_fill_brewer(palette = "Dark2")
```



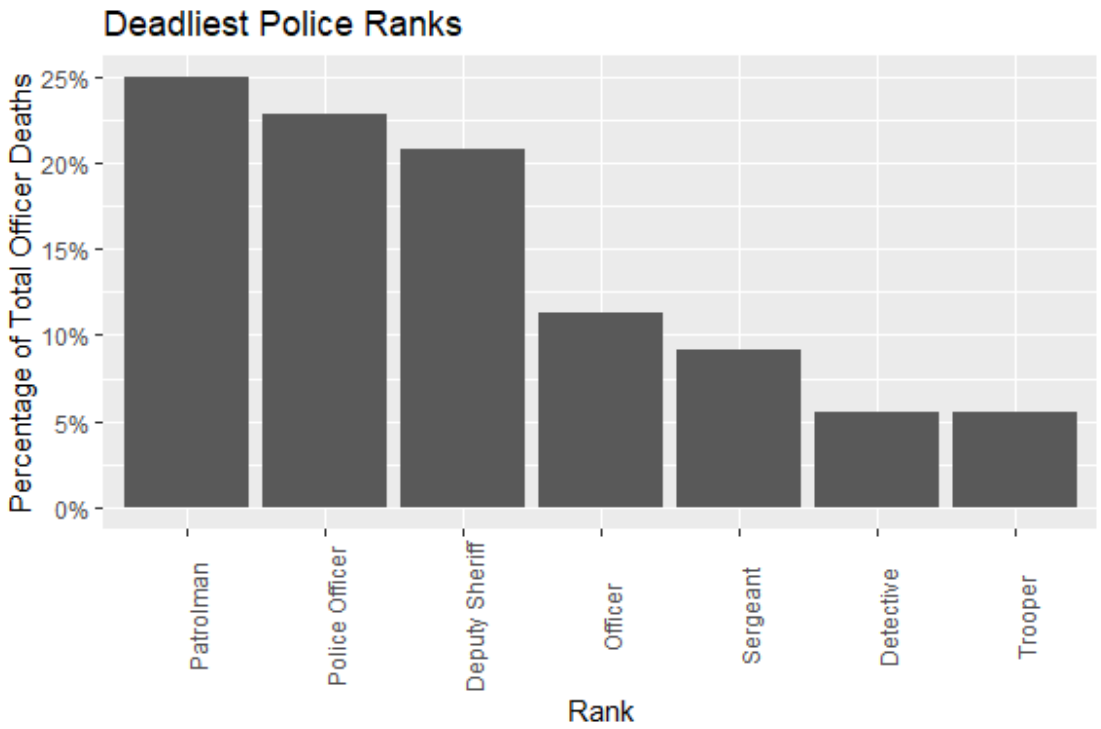
The density of officer deaths due to COVID19 is so large, that in a ridgeline graph, it changes the scale and washes out the other causes of death. We have to remove those deaths from analysis to get a better overall picture.

```
df %>%
  subset(Year >= 1985 &
           Cause_of_Death %in% c("Gunfire", "Automobile crash", "Heart attack
",
                                 "Motorcycle crash", "Vehicular assault",
                                 "COVID19")) %>%
  group_by(Year, Cause_of_Death) %>%
  summarize(n = n()) %>%
  ungroup() %>%
  ggplot(aes(x = Year, y = n, group = Cause_of_Death)) + geom_point(aes(color
= Cause_of_Death)) +
  geom_line(aes(group = Cause_of_Death, color = Cause_of_Death)) +
  scale_x_continuous() + scale_color_brewer(palette = "Dark2") +
  labs(title = "Most Common Officer Deaths Over Time", x = "Year",
       y = "Number of Deaths")
```



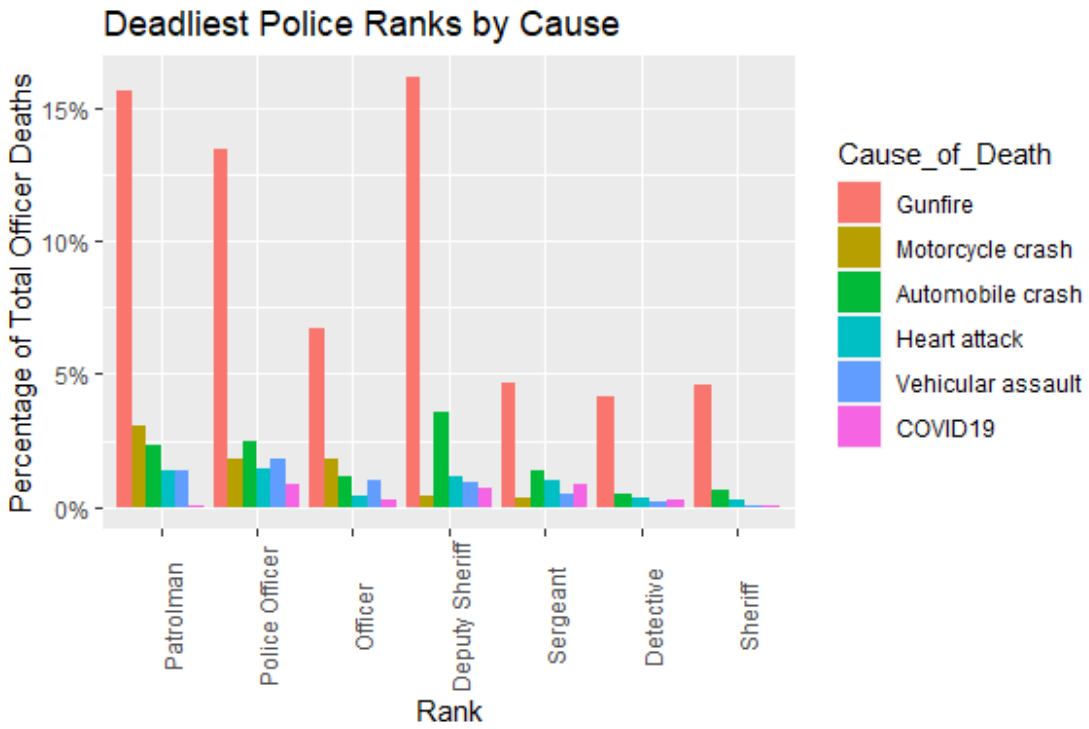
For a more focused view, these are all line of duty deaths post 1985. We can see that gunfire holds steady for number of deaths over the other main causes until 2020 where it is massively misplaced by COVID19. The first yellow dot representing COVID19 deaths is far above the dot representing gunfire deaths, meaning that COVID19 was the first and only cause of death to top gunfire since 1791.

```
df %>%
  group_by(Rank) %>%
  count(sort = TRUE) %>%
  ungroup() %>%
  slice(1:7) %>%
  mutate(pct = n / sum(n),
         Rank = fct_reorder(Rank, desc(pct))) %>%
  ggplot(aes(x = Rank, y = pct)) + geom_col() +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_y_continuous(labels = scales::percent_format(accuracy = 1)) +
  labs(title = "Deadliest Police Ranks") + xlab("Rank") +
  ylab("Percentage of Total Officer Deaths")
```



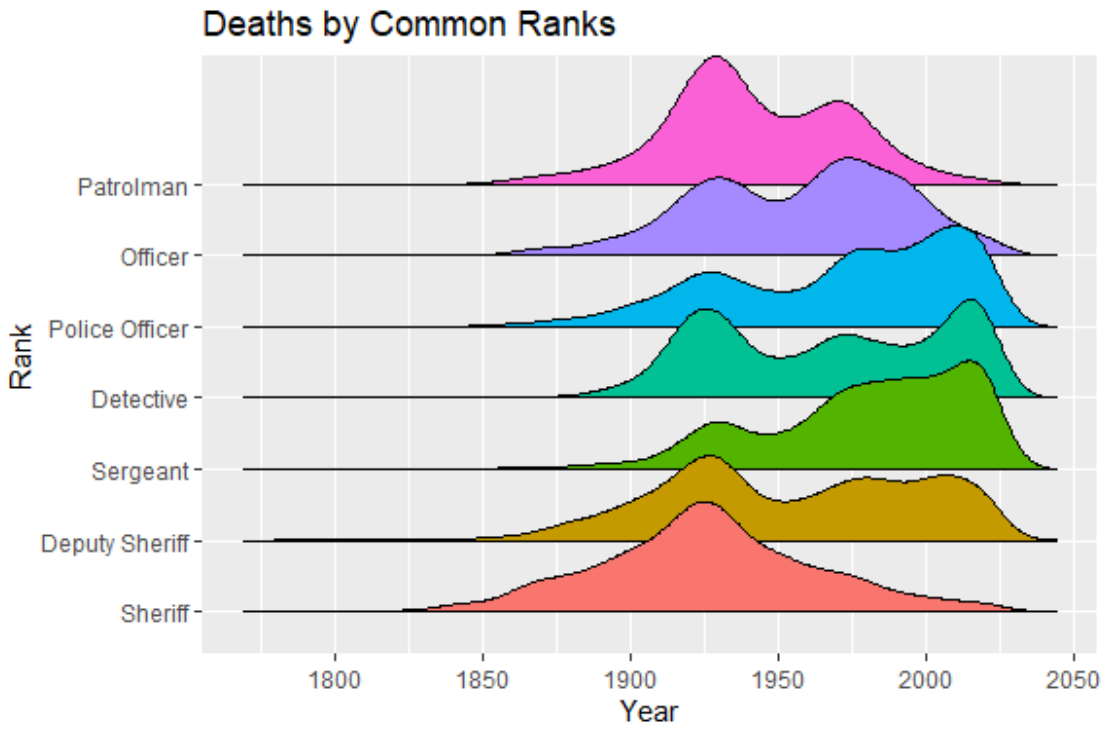
Now we can take a look at which police rank is the most dangerous. The highest percentage of officer deaths belonged to those with ranks of Patrolman. While not a well known contemporary rank, it is still present in the Midwest and Southern states. Next up is Police Officer, which is used to describe any officer in uniform below the rank of Detective. These 2 ranks tend to spend the most time of any rank responding to calls and interacting with civilians.

```
df %>%
  group_by(Rank, Cause_of_Death) %>%
  count(sort = TRUE) %>%
  ungroup() %>%
  subset(Rank %in% c("Deputy Sheriff", "Patrolman", "Police Officer", "Office
r",
                     "Sergeant", "Sheriff", "Detective") &
         Cause_of_Death %in% c("Gunfire", "Automobile crash", "Heart attack",
                               "Motorcycle crash", "Vehicular assault",
                               "COVID19")) %>%
  mutate(pct = n / sum(n),
         Rank = fct_reorder(Rank, desc(pct)),
         Cause_of_Death = fct_reorder(Cause_of_Death, desc(pct))) %>%
  ggplot(aes(x = Rank, y = pct)) + geom_col(aes(fill = Cause_of_Death),
                                            position = "dodge") +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_y_continuous(labels = scales::percent_format(accuracy = 1)) +
  labs(title = "Deadliest Police Ranks by Cause") + xlab("Rank") +
  ylab("Percentage of Total Officer Deaths")
```



When splitting up the rank deaths by cause, we can see that gunfire is still the total most common. While not as present as before, we can see that in the 2.5 years since the pandemic started, COVID19 is not the bottom ranked cause of death in 3 of these 7 ranks (Deputy Sheriff, Sergeant, and Detective).

```
df %>%
  subset(Rank %in% c("Deputy Sheriff", "Patrolman", "Police Officer", "Office
r",
                     "Sergeant", "Sheriff", "Detective")) %>%
  mutate(Rank = factor(Rank, levels = c("Sheriff", "Deputy Sheriff", "Sergean
t",
                                        "Detective", "Police Officer",
                                        "Officer", "Patrolman"))) %>%
  ggplot(aes(x = Year, y = Rank, fill = Rank)) + geom_density_ridges2() +
  theme(legend.position = "none") + labs(title = "Deaths by Common Ranks")
```



We can see how different ranks have become safer or more dangerous over time. The worst time to be a sheriff would have been in the 1920s during the Prohibition Era. This was the time in which bootleggers would make runs through rural areas and were more likely to run into Sheriffs than other forms of law enforcement. This is also the origins of organized crime in America, typically beginning with smuggling of alcohol and later weapons. Other sources of danger for police ranks such as patrolmen and detectives.