Data 607 Final Project

Coco Donovan

2023-05-13

Glaring Limitation:

There is no readily available data for before 2003, so I cannot make quantitative statements from my own analysis of crime trends from before 2003 (I would love to based on reporting I have seen; however that would not be my own work).

Necessary Packages

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(readr)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.0
## v lubridate 1.9.2     v tibble    3.2.1
## v purrr     1.0.1     v tidyr     1.3.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Police Incident Data: 2003 - 2017

```
old_police_files  <- list.files(pattern = '^Police_Data_....\\.csv')

old_police_tables <- lapply(old_police_files, read.csv, header = TRUE)

old_police_data <- do.call(rbind , old_police_tables)
```

Police Incident Data: 2018 - Current

```
present_police_files  <- list.files(pattern = 'Present_Police_Data_....\\.csv')

present_police_tables <- lapply(present_police_files, read.csv, header = TRUE)

present_police_data <- do.call(rbind , present_police_tables)
```

SF City Budget

```
budget <- read.csv('Budget.csv')
```

SF Yearly Population

```
sf_pop <- read.csv('SF_POP.csv')

sf_pop$Population <- 1000 * sf_pop$Population

sf_pop <- sf_pop %>%
  separate(Year, c('year', 'month', 'date'), '-') %>%
  select(year, Population)

sf_pop$year <- as.integer(sf_pop$year)

sf_pop <- sf_pop %>%
  filter(year >= 2003)
```

Joining New and Old Incident Counts, Reports, and the ratio of the two:

```
old_police_data <- old_police_data %>%
  separate(Date, c("month", "day","year"), "/")

old_police_data$year <- as.integer(old_police_data$year)

old_counts_and_incidents <- old_police_data %>%
  group_by(year) %>%
  summarize(count_reports = as.double(n()), count_incidents = as.double(length(unique(IncidntNum))))

new_counts_and_incidents <- present_police_data %>%
```

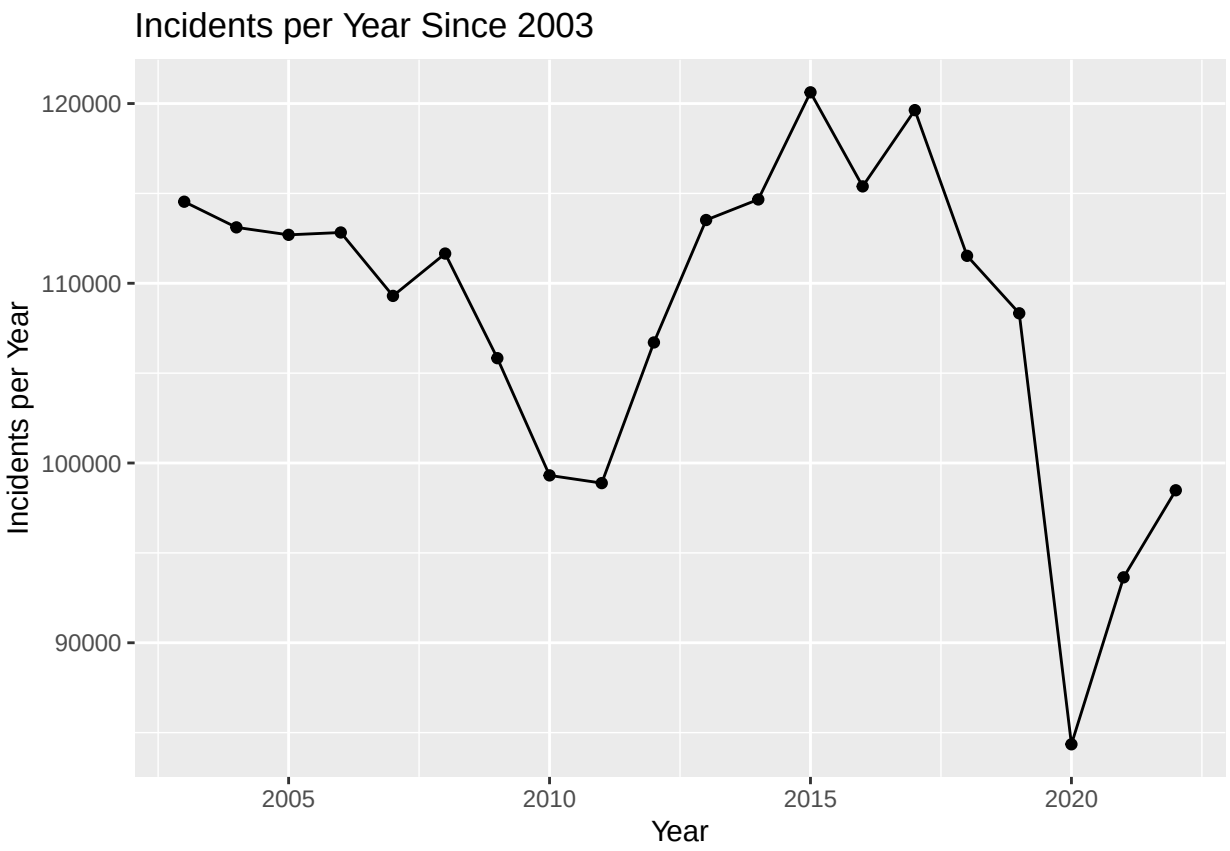
```
  rename('year' = 'Incident.Year') %>%
  group_by(year) %>%
  summarize(count_reports = as.double(n()), count_incidents = as.double(length(unique(Incident.Number))))

counts_and_incidents <- rbind(new_counts_and_incidents, old_counts_and_incidents)

counts_and_incidents <- counts_and_incidents %>%
  group_by(year) %>%
  mutate(ratio = count_reports/count_incidents) %>%
  arrange(desc(year))

knitr::kable(counts_and_incidents)
```

year	count_reports	count_incidents	ratio
2023	44748	32720	1.367604
2022	134843	98483	1.369201
2021	128775	93640	1.375214
2020	118313	84351	1.402627
2019	148272	108330	1.368707
2018	152707	111528	1.369226
2017	149487	119633	1.249546
2016	145994	115390	1.265222
2015	151459	120623	1.255640
2014	144844	114664	1.263204
2013	147664	113518	1.300798
2012	135464	106708	1.269483
2011	126713	98883	1.281444
2010	127758	99311	1.286444
2009	134309	105833	1.269065
2008	135242	111653	1.211271
2007	131771	109297	1.205623
2006	131856	112825	1.168677
2005	137048	112695	1.216096
2004	142054	113110	1.255893
2003	142803	114538	1.246774

```
# cool point: look at 2020...it had the lowest number of incidents in this entire time frame (cannot wa
```

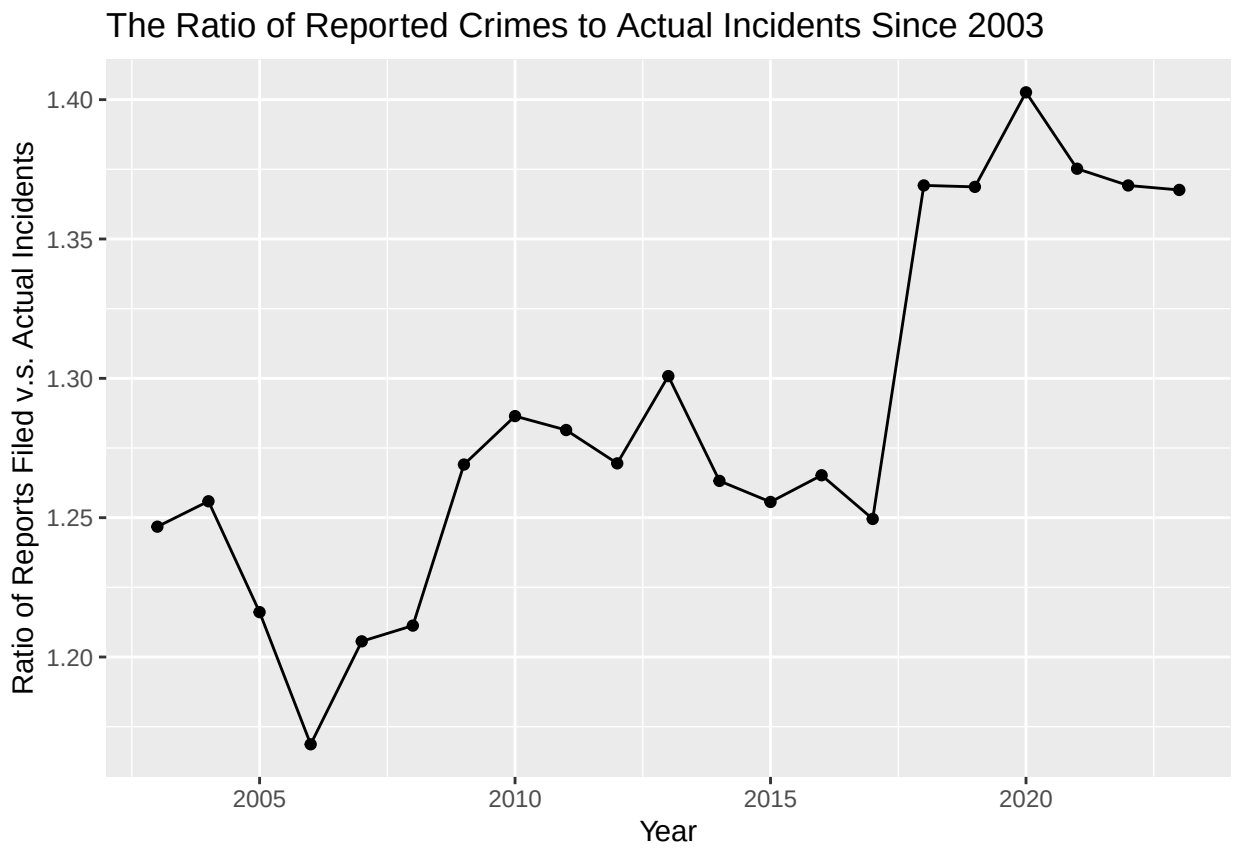
Visualizations: Incidents Over Time

```
# I did not include 2023 just yet, because the year is only a about 1/3 of the
# way over and would not contribute to reliable comparisions for raw counts

counts_and_incidents_no_2023 <- subset(counts_and_incidents, year != 2023)

ggplot(counts_and_incidents_no_2023, aes(x= year, y= count_incidents)) +
  geom_point() +
  geom_line() +
  labs(x = 'Year', y = 'Incidents per Year', title = 'Incidents per Year Since 2003')
```



Visualization: Ratio of reports vs acutal incidents

```
ggplot(counts_and_incidents, aes(x= year, y= ratio)) +
  geom_point() +
  geom_line() +
  labs(x = 'Year', y = 'Ratio of Reports Filed v.s. Actual Incidents', title = 'The Ratio of Reported C
```



Adding population to counts and incidents

```
years_and_counts <- full_join(sf_pop, counts_and_incidents, by = 'year')
```

Crime per 100K

```
years_and_counts <- years_and_counts %>%
  mutate(incidents_per_100k = 100000 * count_incidents/Population) %>%
  arrange(desc(year))

knitr::kable(years_and_counts)
```

year	Population	count_reports	count_incidents	ratio	incidents_per_100k
2023	NA	44748	32720	1.367604	NA
2022	808437	134843	98483	1.369201	12181.90
2021	811253	128775	93640	1.375214	11542.64
2020	870393	118313	84351	1.402627	9691.14
2019	878826	148272	108330	1.368707	12326.67
2018	879676	152707	111528	1.369226	12678.30
2017	877471	149487	119633	1.249546	13633.84

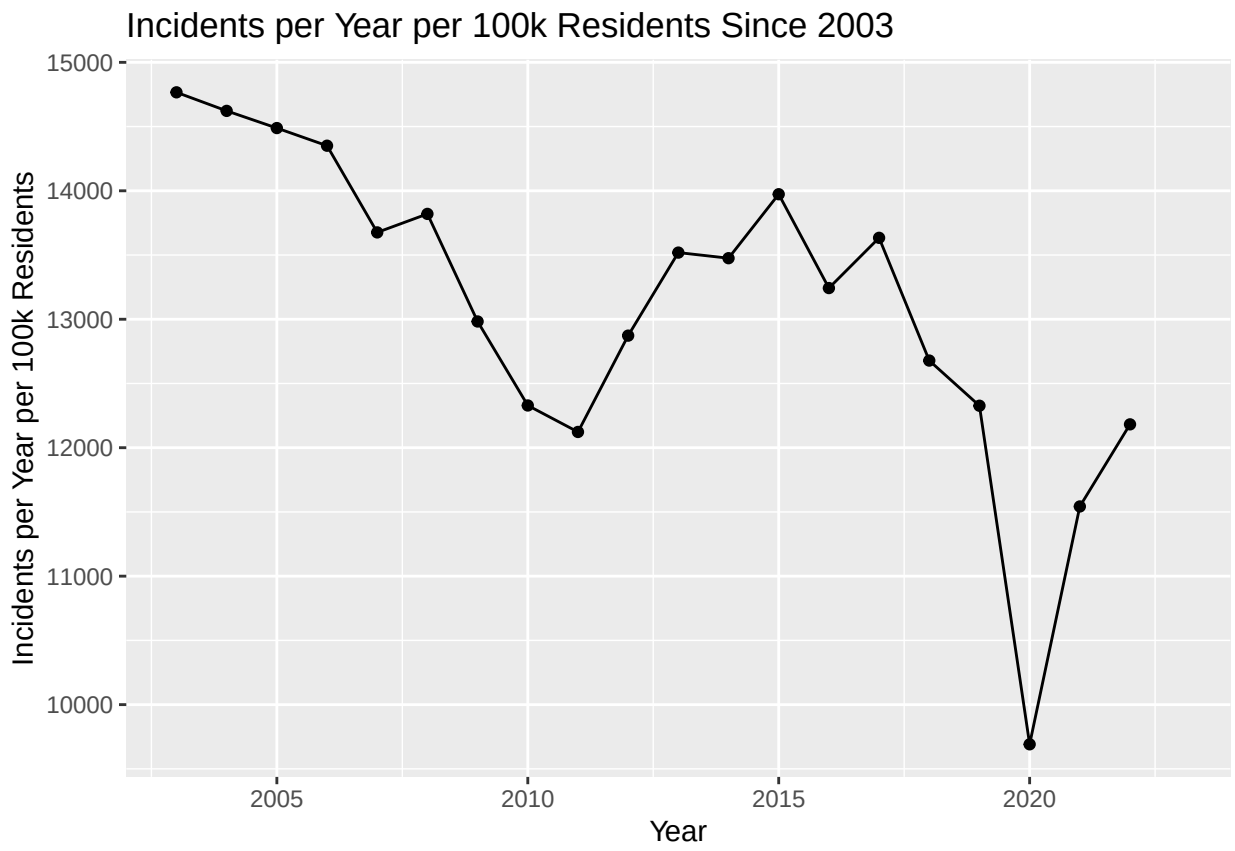
year	Population	count_reports	count_incidents	ratio	incidents_per_100k
2016	871343	145994	115390	1.265222	13242.78
2015	863237	151459	120623	1.255640	13973.34
2014	850918	144844	114664	1.263204	13475.33
2013	839695	147664	113518	1.300798	13518.96
2012	828963	135464	106708	1.269483	12872.47
2011	815694	126713	98883	1.281444	12122.56
2010	805519	127758	99311	1.286444	12328.82
2009	815184	134309	105833	1.269065	12982.71
2008	807904	135242	111653	1.211271	13820.08
2007	799185	131771	109297	1.205623	13676.06
2006	786187	131856	112825	1.168677	14350.91
2005	777835	137048	112695	1.216096	14488.29
2004	773556	142054	113110	1.255893	14622.08
2003	775663	142803	114538	1.246774	14766.46

Visualization: Crime Incidents per 100k

```
ggplot(years_and_counts, aes(x= year, y= incidents_per_100k)) +
  geom_point() +
  geom_line() +
  labs(x = 'Year', y = 'Incidents per Year per 100k Residents', title = 'Incidents per Year per 100k Re
```

```
## Warning: Removed 1 rows containing missing values (‘geom_point()‘).
```

```
## Warning: Removed 1 row containing missing values (‘geom_line()‘).
```

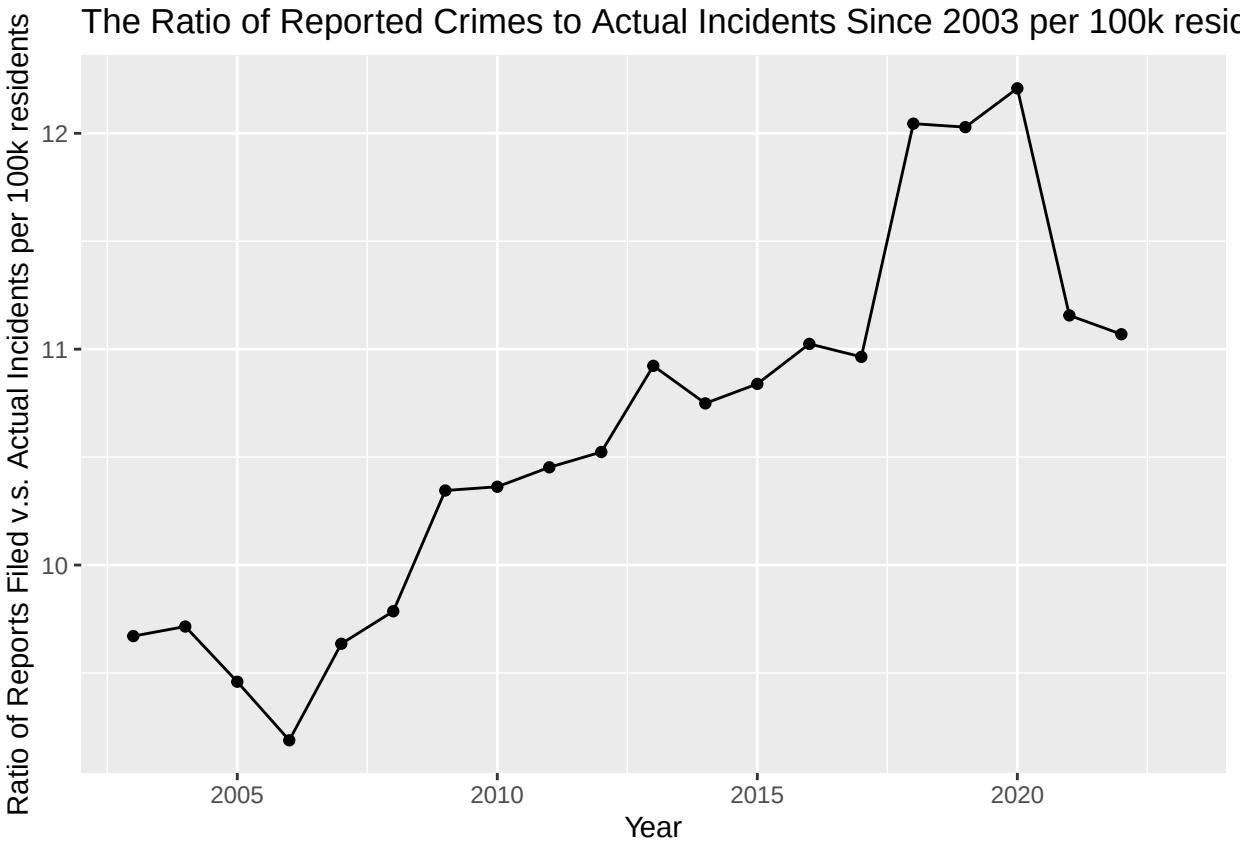


Visualization: Ratio of Reports vs Incidents per 100K over time

```
years_and_count_ratio_100k <-years_and_counts %>%
  mutate(ratio_per_100k = count_reports / incidents_per_100k) %>%
  arrange(desc(year))

ggplot(years_and_count_ratio_100k, aes(x= year, y= ratio_per_100k)) +
  geom_point() +
  geom_line() +
  labs(x = 'Year', y = 'Ratio of Reports Filed v.s. Actual Incidents per 100k residents', title = 'The
```

```
## Warning: Removed 1 rows containing missing values (‘geom_point()‘).
```

```
## Warning: Removed 1 row containing missing values (‘geom_line()‘).
```



Ideas: - Compare row count vs incident number count (maybe test for statistical significance) - Check resolution rate over time (per year) and does that differ across crime types? -Crime per 100,000 -Are there more crime heavy months over the years? - what do specific types of crime counts look like over time? (compare non-violent and violent crime)