Analyzing Major Crime Reports in Toronto over the years*

Allen Uy

January 15, 2024

Prototype of Term Paper 1. Reports of major crime in Toronto reached a peak in 2023. Based on analysis of data from OpenDataToronto there has been a steady increase of major crime reports since data collection in 2014.

1 Introduction

Basically a single Quarto doc of my scripts I'm likely using with modification for Term Paper 1 for Mini-essay #2 for easy code review. Major crime indicators are listed as: assault, auto theft, break and enter, homicide, robbery, sexual violation, and theft over.

The remainder of this paper is structured as follows. Section 2....

2 Data

The data was gathered from OpenDataToronto (see Gelfand (2022)) and analyzed using R (R Core Team (2023)). Data was cleaned using janitor (Firke (2023)).

```
#### Simulate data ####
set.seed(42)

sim_data <- tibble(
  report_id = c(1:100),
  report_year = sample(2013:2023, 100, replace=TRUE)
)
```

*Code and data are available at: https://github.com/varygx/TorontoMajorCrime

```r
#### Clean data ####
raw_data <- read_csv("../../inputs/data/raw_data.csv")
```

```
Rows: 372899 Columns: 27
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (14): EVENT_UNIQUE_ID, REPORT_MONTH, REPORT_DOW, OCC_MONTH, OCC_DOW, DI...
dbl  (11): X_id, REPORT_YEAR, REPORT_DAY, REPORT_DOY, REPORT_HOUR, OCC_YEAR,...
date  (2): REPORT_DATE, OCC_DATE

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.
```

```r
cleaned_data <- clean_names(raw_data)

cleaned_data <- cleaned_data |> select("report_year")

#### Save data ####
write_csv(cleaned_data, "../../outputs/data/analysis_data.csv")
```

```r
#### Test data ####
cleaned_data <- read_csv("../../outputs/data/analysis_data.csv")
```

```
Rows: 372899 Columns: 1
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (1): report_year

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.
```

```r
cleaned_data$report_year |> min() == 2014
```

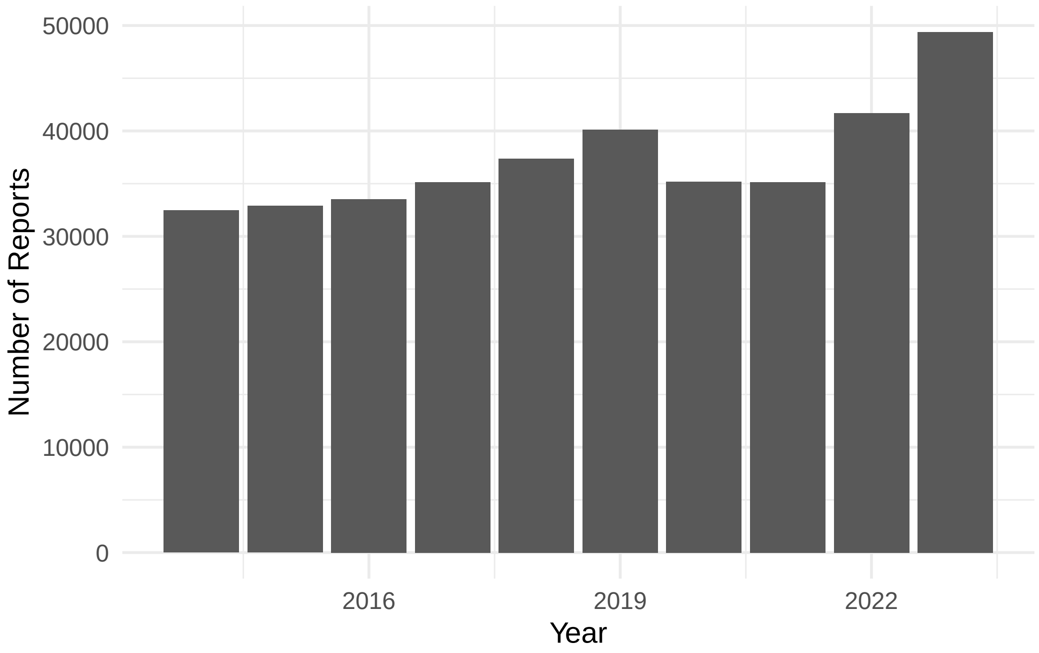
```
[1] TRUE
```

```r
cleaned_data$report_year |> max() == 2023
```

```
[1] TRUE
```

3 Results

```r
#### Visualize data ####
cleaned_data |> ggplot(aes(x=report_year)) +
  geom_bar() +
  theme_minimal() +
  labs(x = "Year", y = "Number of Reports")
```



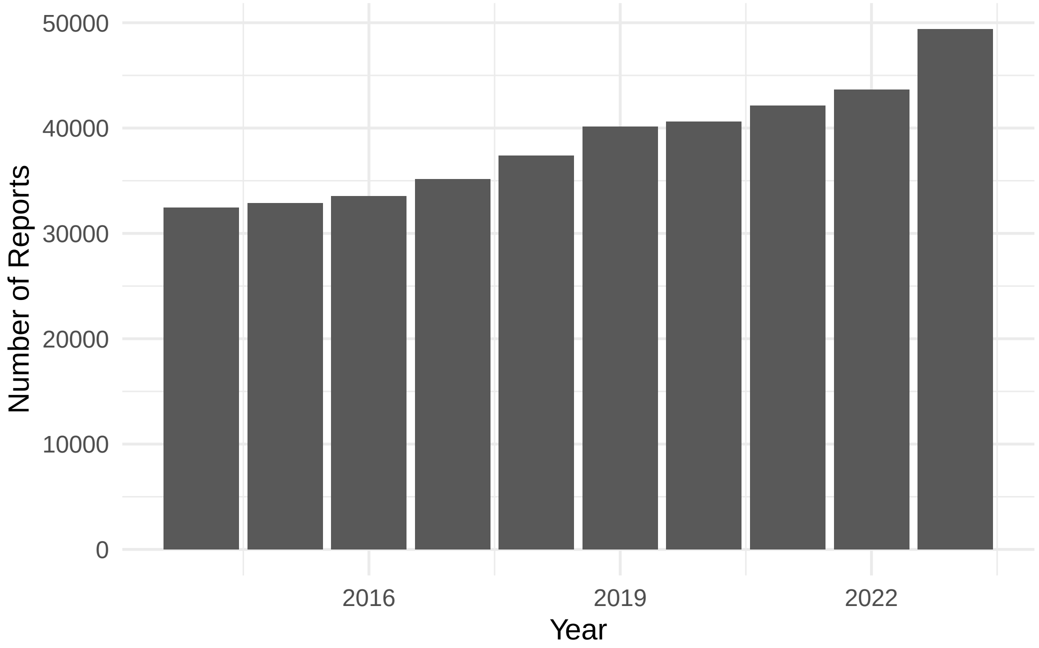
We can see a steady rise in reports from 2014 to 2019 with a sudden decrease in 2020. This lines up with the COVID pandemic and gives a plausible explanation as less people were likely to be outside. Below is a graph that extrapolates data for 2020 to 2022 based on a linear model fitted to 2014 to 2019 data.

```r
year_count <- cleaned_data |> count(report_year)
pre_covid <- year_count |> filter(report_year <= 2019)
fit <- lm(n ~ report_year, data = pre_covid)

predicted_years <- tibble(report_year = c(2020:2022))
predicted_counts <- predict(fit, newdata=predicted_years, type="response")
predicted_data <- tibble(report_year = predicted_years$report_year, n=predicted_counts)
```

```
extrapolated_data <- year_count |>
  mutate(n = ifelse(report_year %in% predicted_data$report_year, predicted_data$n, n))

extrapolated_data |> ggplot(aes(x=report_year, y=n)) +
  geom_bar(stat="identity") +
  theme_minimal() +
  labs(x = "Year", y = "Number of Reports")
```



References

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://github.com/sfirke/janitor.

Gelfand, Sharla. 2022. *Opendatatoronto: Access the City of Toronto Open Data Portal*. https://sharlagelfand.github.io/opendatatoronto/.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.