NYPD Shooting Incident Data Report

Meg McGrath

2025-11-15

This project will use data that comes directly from the NYPD, reviewed by the Office of Management and Planning. It is a collection of information on all shooting incidents in NYC from 2006 to 2024.
We will take a look at the trends of incidents based on the time of year, as well as the overall trends from year to year.

```
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- 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
```

```
# read in data

nypd_shootings <- read_csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLO
```

```
## Rows: 29744 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (12): OCCUR_DATE, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION...
## dbl   (5): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, Latitude, Longitude
## num   (2): X_COORD_CD, Y_COORD_CD
## lgl   (1): STATISTICAL_MURDER_FLAG
## time  (1): OCCUR_TIME
##
## 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.
```

```
# summarize data
# look at number of rows, columns, variables
# look at summary of each variable

glimpse(nypd_shootings)
```

```
## Rows: 29,744
## Columns: 21
## $ INCIDENT_KEY            <dbl> 231974218, 177934247, 255028563, 25384540, 726~
## $ OCCUR_DATE              <chr> "08/09/2021", "04/07/2018", "12/02/2022", "11/~
## $ OCCUR_TIME              <time> 01:06:00, 19:48:00, 22:57:00, 01:50:00, 01:58~
## $ BORO                    <chr> "BRONX", "BROOKLYN", "BRONX", "BROOKLYN", "BRO~
## $ LOC_OF_OCCUR_DESC       <chr> NA, NA, "OUTSIDE", NA, NA, NA, NA, NA, NA, NA,~
## $ PRECINCT                <dbl> 40, 79, 47, 66, 46, 42, 71, 69, 75, 69, 40, 42~
## $ JURISDICTION_CODE       <dbl> 0, 0, 0, 0, 0, 2, 0, 2, 0, 0, 0, 2, 0, 0, 2, 0~
## $ LOC_CLASSFCTN_DESC      <chr> NA, NA, "STREET", NA, NA, NA, NA, NA, NA, NA, ~
## $ LOCATION_DESC           <chr> NA, NA, "GROCERY/BODEGA", "PVT HOUSE", "MULTI ~
## $ STATISTICAL_MURDER_FLAG <lgl> FALSE, TRUE, FALSE, TRUE, TRUE, FALSE, TRUE, F~
## $ PERP_AGE_GROUP          <chr> NA, "25-44", "(null)", "UNKNOWN", "25-44", "18~
## $ PERP_SEX                <chr> NA, "M", "(null)", "U", "M", "M", NA, NA, "M",~
## $ PERP_RACE               <chr> NA, "WHITE HISPANIC", "(null)", "UNKNOWN", "BL~
## $ VIC_AGE_GROUP           <chr> "18-24", "25-44", "25-44", "18-24", "<18", "18~
## $ VIC_SEX                 <chr> "M", "M", "M", "M", "F", "M", "M", "M", "M", "~
## $ VIC_RACE                <chr> "BLACK", "BLACK", "BLACK", "BLACK", "BLACK", "~
## $ X_COORD_CD              <dbl> 1006343.0, 1000082.9, 1020691.0, 985107.3, 100~
## $ Y_COORD_CD              <dbl> 234270.0, 189064.7, 257125.0, 173349.8, 247502~
## $ Latitude                <dbl> 40.80967, 40.68561, 40.87235, 40.64249, 40.845~
## $ Longitude               <dbl> -73.92019, -73.94291, -73.86823, -73.99691, -7~
## $ Lon_Lat                 <chr> "POINT (-73.92019278899994 40.80967347200004)"~
```

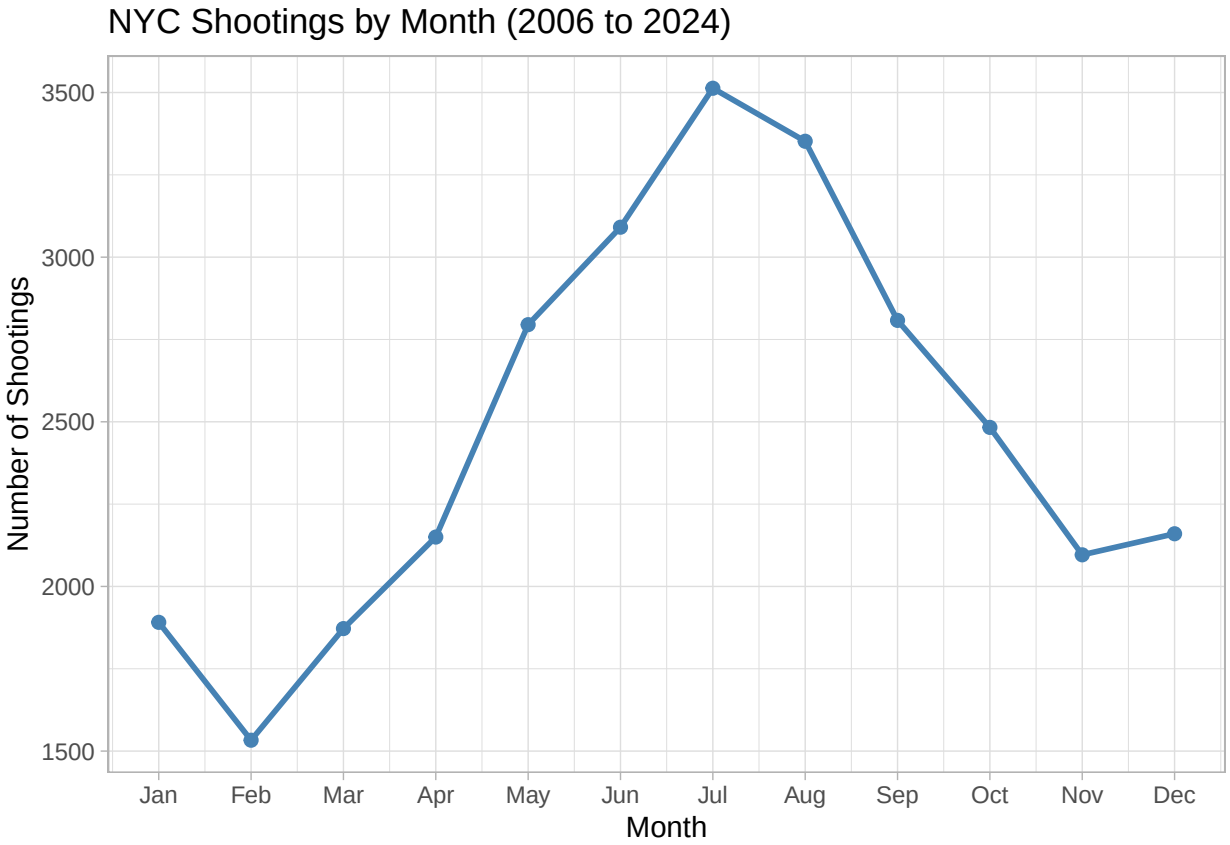
There are 29,744 rows representing 29,744 separate incidents. There are 21 columns representing information about each incident including information about the location, the perpetrator and the victim.
For the sake of this report, we only care about the month and year of the indcident occurence.

```
# clean data
# change OCCUR_DATE from chr to date
# create month and hour variable for future analysis
# select desired variables

nypd_clean <- nypd_shootings %>%
  mutate(
    occur_date = mdy(OCCUR_DATE),
    occur_month = month(occur_date),
  ) %>%
  select(
    occur_date, occur_month
  )
```

It is worth noting that there were some missing data points in most categories, sometimes marked as UNKNOWN, NA or (null). However there are no missing data points for dates, so we do not need to account for how to manage missing data.

```
# look at shootings by month
# group and total by month
# create visual

shootings_by_month <- nypd_clean %>%
  group_by(occur_month) %>%
  summarise(total_shootings = n())
```

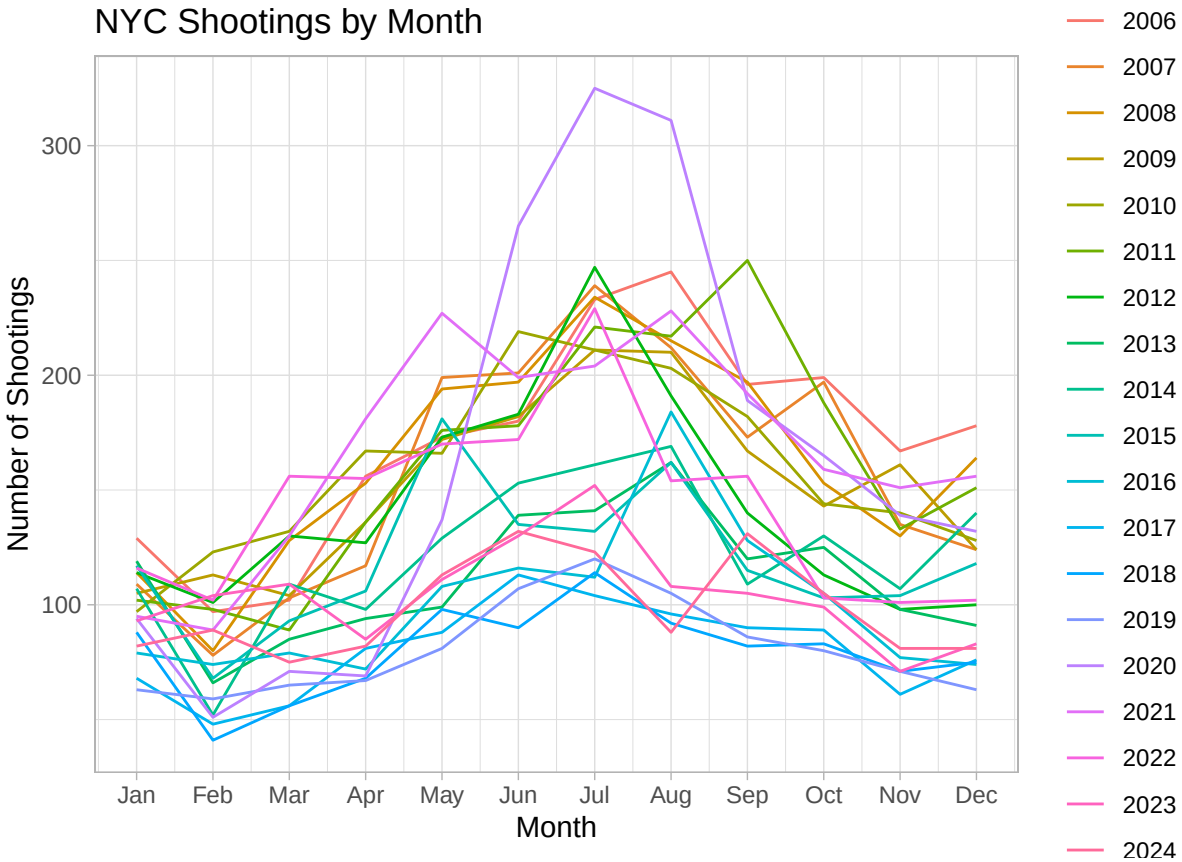
```
ggplot(shootings_by_month, aes(x = occur_month, y = total_shootings)) +
  geom_line(color = "steelblue", linewidth = 1) +
  geom_point(color = "steelblue", size = 2) +
  scale_x_continuous(breaks = 1:12, labels = month.abb) +
  labs(
    title = "NYC Shootings by Month (2006 to 2024)",
    x = "Month",
    y = "Number of Shootings"
  ) +
  theme_light()
```



Looking at shootings per month, there is a clear trend that increases during summer months, peaking in July, and decreases during winter months, hitting the lowest point in February.
This is likely due to the number of people outside and interacting during warmer weather. With more people, there are likely more interactions between people, more crimes of opportunity, more personal conflicts, and unfortunately more shooting incidents.
There is little room for bias in this analysis, due to the fact that there is no missing data and dates of incidents are not open to interpretation.
However, we could check for possible outliers that could be skewing the data. We will address this by separating the data by year.

```
# summarise data by month and year
# create visual

month_year <- nypd_clean %>%
  group_by(year = year(occur_date), month = month(occur_month)) %>%
```

```
  summarise(total_shootings = n())

ggplot(month_year, aes(x = month, y = total_shootings, group = year, color = factor(year))) +
  geom_line() +
  scale_x_continuous(breaks = 1:12, labels = month.abb) +
  labs(
    title = "NYC Shootings by Month",
    x = "Month",
    y = "Number of Shootings"
  ) +
  theme_light()
```

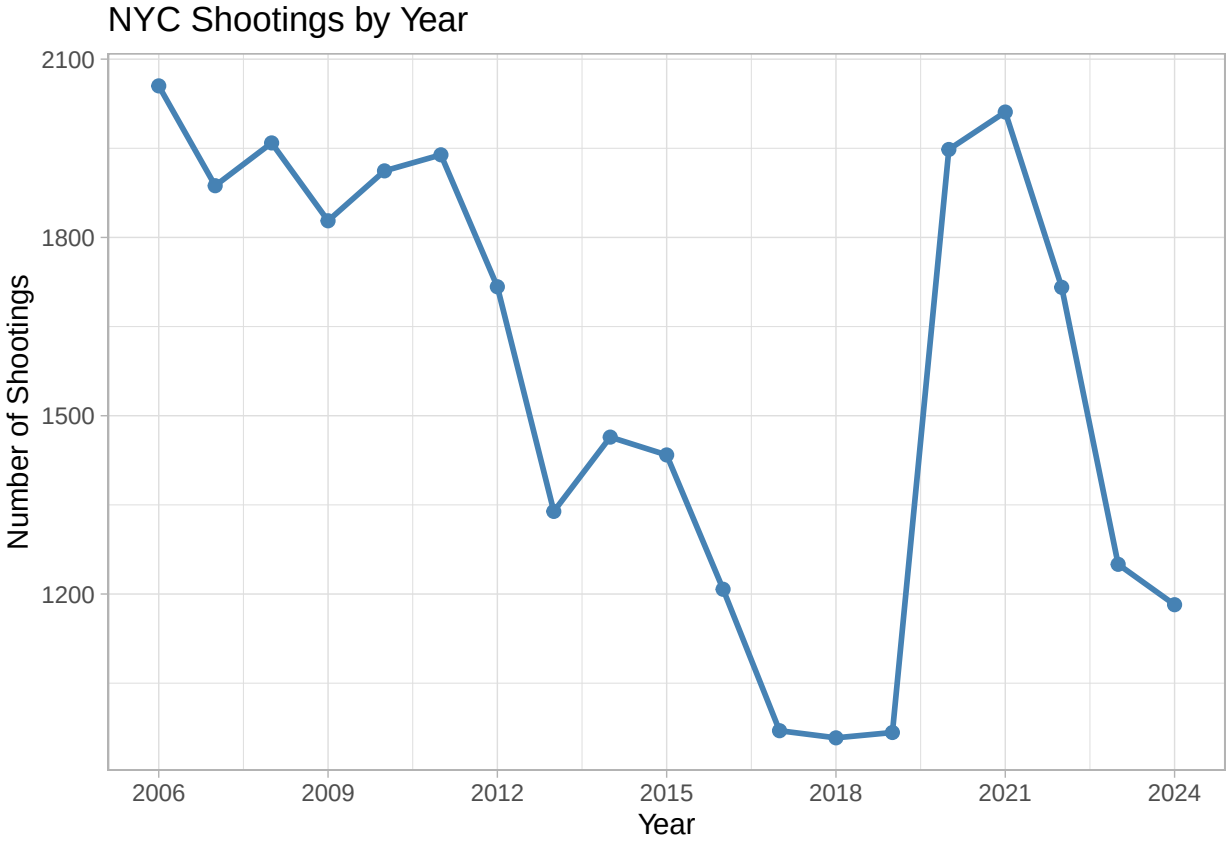


Separating the data by month and year, the trends are less obvious, but we still see similar trends of more NYC shooting incidents during summer months than winter months, confirming our previous analysis.
This information could be used by the NYPD in training, to make officers aware of higher risks during certain months.
This information could also be used to inform communities of the higher risk of gun violence during certain times of year.

Next, we will look at the shooting incident trends from year to year.

```
# summarise data by year

shootings_by_year <- nypd_clean %>%
  group_by(year = year(occur_date)) %>%
  summarise(total_shootings = n())
```

```
ggplot(shootings_by_year, aes(x = year, y = total_shootings)) +
  geom_line(color = "steelblue", linewidth = 1) +
  geom_point(color = "steelblue", size = 2) +
  scale_x_continuous(breaks = seq(2006,2024, by = 3)) +
  labs(
    title = "NYC Shootings by Year",
    x = "Year",
    y = "Number of Shootings"
  ) +
  theme_light()
```



There was a significant downward trend from 2011 to 2019, then a huge spike in 2020.
This seems to contradict our conclusions from our analysis of monthly trends, that more people remaining in their homes during colder weather led to less shootings. Following this same logic, the COVID shutdown in 2020 should have led to less shootings.
But we can account for this spike without discarding our previous conclusions. The Black Lives Matter protests following the death of George FLoyd also occured in 2020, and this led to increased tensions, especially between communities and their police departments. This is the most likely cause of the major spike in shootings in 2020 and 2021.
We can see from our analysis that from 2022 to 2024, the number of shootings have continued to drop, similar to the previous drop beginning in 2011.

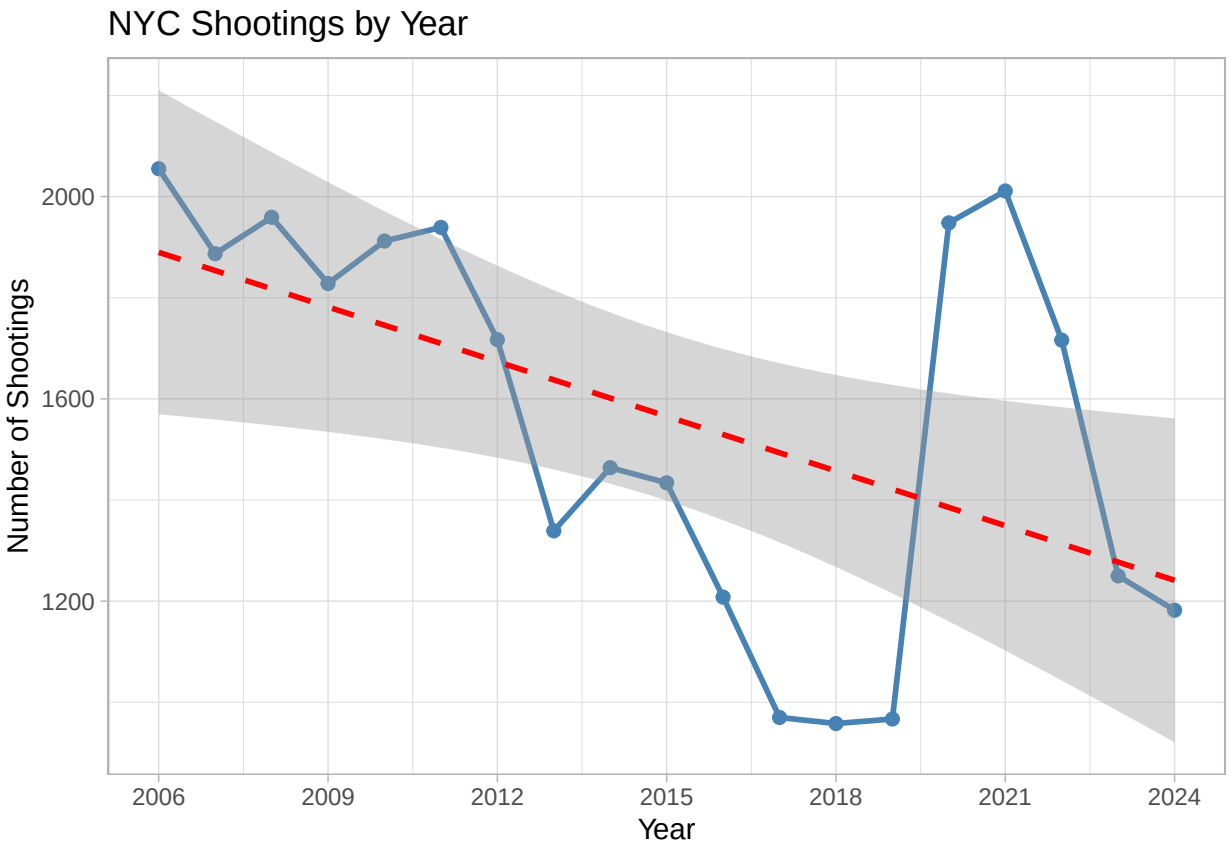
Next, we will add a linear regression model to our year by year visual.

```
# create linear regression model and plot
```

```r
year_model <- lm(total_shootings ~ year, data = shootings_by_year)
summary(year_model)
```

```
##
## Call:
## lm(formula = total_shootings ~ year, data = shootings_by_year)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -523.42 -218.01   33.32  165.84  661.68
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 74158.50   29035.88   2.554   0.0205 *
## year          -36.03      14.41  -2.500   0.0229 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 344 on 17 degrees of freedom
## Multiple R-squared:  0.2688, Adjusted R-squared:  0.2258
## F-statistic: 6.251 on 1 and 17 DF,  p-value: 0.02294
```

```r
ggplot(shootings_by_year, aes(x = year, y = total_shootings)) +
  geom_line(color = "steelblue", linewidth = 1) +
  geom_point(color = "steelblue", size = 2) +
  geom_smooth(method = "lm", se = TRUE, color = "red", linetype = "dashed") +
  scale_x_continuous(breaks = seq(2006,2024, by = 3)) +
  labs(
    title = "NYC Shootings by Year",
    x = "Year",
    y = "Number of Shootings"
  ) +
  theme_light()
```



Adding a linear regression model to the data shows an overall downward trend in NYC shooting incidents. The model has limitations. The major spike in 2020 does not fit the model, nor does the more rapid decline from 2011 to 2019.

Yet looking at how the data fits the model over the entire time period, it appears that shooting incidents have once again returned to a decreasing rate similar to the beginning of the analysis, meaning that the city has recovered from the huge disruption of 2020 and 2021.

We could conclude that the police departments play a large role in the level of overall gun violence, since gun violence sky-rocketed in 2020 and 2021 when the NYPD was struggling with community support.
But it is also a possibility that the drop in gun violence had nothing to do with NYPD, but rather something else, like more community outreach programs.
Either conclusion, without further evidence, could be considered an example of bias in data analysis.

This report simply shows that gun violence is trending downwards in NYC. We can see the effect, but not necessarily the cause. However, this report does demonstrate that it would be valuable to do continued research into what factors are aiding these downward trends. It is likely a combination of many things, including better police department support and more community outreach programs. It would be worthwhile to further examine the factors that are supporting this downward trend in gun violence in NYC, and how we could replicate these practices in other cities.

```
sessionInfo()
```

```
## R version 4.4.2 (2024-10-31 ucrt)
## Platform: x86_64-w64-mingw32/x64
## Running under: Windows 11 x64 (build 26100)
##
## Matrix products: default
```

```
##
##
## locale:
## [1] LC_COLLATE=English_United States.utf8
## [2] LC_CTYPE=English_United States.utf8
## [3] LC_MONETARY=English_United States.utf8
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.utf8
##
## time zone: America/Denver
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] lubridate_1.9.3 forcats_1.0.0   stringr_1.5.1   dplyr_1.1.4
##  [5] purrr_1.0.2     readr_2.1.5     tidyr_1.3.1     tibble_3.2.1
##  [9] ggplot2_3.5.1   tidyverse_2.0.0
##
## loaded via a namespace (and not attached):
##  [1] generics_0.1.3    stringi_1.8.4     lattice_0.22-6    hms_1.1.3
##  [5] digest_0.6.37     magrittr_2.0.3    evaluate_1.0.1    grid_4.4.2
##  [9] timechange_0.3.0  fastmap_1.2.0     Matrix_1.7-1      mgcv_1.9-1
## [13] scales_1.3.0      cli_3.6.3         rlang_1.1.4       crayon_1.5.3
## [17] bit64_4.5.2       munsell_0.5.1     splines_4.4.2     withr_3.0.2
## [21] yaml_2.3.10       tools_4.4.2       parallel_4.4.2    tzdb_0.4.0
## [25] colorspace_2.1-1  curl_6.0.1        vctrs_0.6.5       R6_2.5.1
## [29] lifecycle_1.0.4   bit_4.5.0.1       vroom_1.6.5       pkgconfig_2.0.3
## [33] pillar_1.10.0     gtable_0.3.6      glue_1.8.0        xfun_0.49
## [37] tidyselect_1.2.1  rstudioapi_0.17.1 knitr_1.49        farver_2.1.2
## [41] htmltools_0.5.8.1 nlme_3.1-166      rmarkdown_2.29    labeling_0.4.3
## [45] compiler_4.4.2
```