NYPD Shooting Incident Data (Historic)

G. E.

Intro

NYPD Shooting Incident Data (Historic) Metadata Updated: April 19, 2025

This document analyzes shooting incident numbers in NYC from 2006-2024. We'll explore how shooting incidents are affected by location and time.

> **Dataset Description**: This is a breakdown of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year. This data is manually extracted every quarter and reviewed by the Office of Management Analysis and Planning before being posted on the NYPD website. Each record represents a shooting incident in NYC and includes information about the event, the location and time of occurrence. In addition, information related to suspect and victim demographics is also included. This data can be used by the public to explore the nature of shooting/criminal activity. Please refer to the attached data footnotes for additional information about this dataset.

Setup the R Environment

To setup the R environment, libraries `lubridate` and `tidyverse` were imported.

Import the Data

The dataset was imported via URL, made publicly available by the City of New York.

Dataset URL: "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"

```
shootings_url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD" # nol
shootings <- read_csv(shootings_url)
```

Data Cleaning

Data was cleaned by ensuring proper date conversions and removing variables deemed unnecessary for the given research instance.

```
shootings_clean <- shootings %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE),
         OCCUR_TIME = hms(OCCUR_TIME),
         across(c(BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION_DESC,
                  PERP_SEX, PERP_RACE, VIC_AGE_GROUP, VIC_SEX, VIC_RACE),
                as.factor)) %>%
  select(-c(X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat))
```

Data Summary

```
summary(shootings_clean)
```

```
##   INCIDENT_KEY         OCCUR_DATE           OCCUR_TIME
##  Min.   :  9953245   Min.   :2006-01-01   Min.   :0S
##  1st Qu.: 67321140   1st Qu.:2009-10-29   1st Qu.:3H 30M 45S
##  Median :109291972   Median :2014-03-25   Median :15H 15M 0S
##  Mean   :133850951   Mean   :2014-10-31   Mean   :12H 46M 10.8747982786153S
##  3rd Qu.:214741917   3rd Qu.:2020-06-29   3rd Qu.:20H 44M 0S
##  Max.   :299462478   Max.   :2024-12-31   Max.   :23H 59M 0S
##
##             BORO       LOC_OF_OCCUR_DESC    PRECINCT      JURISDICTION_CODE
##  BRONX        : 8834   INSIDE :  682     Min.   :  1.00   Min.   :0.0000
##  BROOKLYN     :11685   OUTSIDE: 3466     1st Qu.: 44.00   1st Qu.:0.0000
##  MANHATTAN    : 3977   NA's   :25596     Median : 67.00   Median :0.0000
##  QUEENS       : 4426                     Mean   : 65.23   Mean   :0.3181
##  STATEN ISLAND:  822                     3rd Qu.: 81.00   3rd Qu.:0.0000
##                                          Max.   :123.00   Max.   :2.0000
##                                                           NA's   :2
##   LOC_CLASSFCTN_DESC                   LOCATION_DESC   STATISTICAL_MURDER_FLAG
##  STREET    : 2639    MULTI DWELL - PUBLIC HOUS: 5188   Mode :logical
##  HOUSING   :  643    MULTI DWELL - APT BUILD  : 3042   FALSE:23979
##  DWELLING  :  341    (null)                   : 2526   TRUE :5765
##  COMMERCIAL:  276    PVT HOUSE                : 1010
##  OTHER     :   74    GROCERY/BODEGA           :  775
##  (Other)   :  175    (Other)                  : 2226
##  NA's      :25596    NA's                     :14977
##  PERP_AGE_GROUP       PERP_SEX             PERP_RACE     VIC_AGE_GROUP
##  Length:29744       (null): 1628   BLACK         :12323   <18    : 3081
##  Class :character   F     :  461   WHITE HISPANIC: 2667   1022   :    1
##  Mode  :character   M     :16845   UNKNOWN       : 1838   18-24  :10677
##                     U     : 1500   (null)        : 1628   25-44  :13563
##                     NA's  : 9310   BLACK HISPANIC: 1487   45-64  : 2118
##                                    (Other)       :  491   65+    :  236
##                                    NA's          : 9310   UNKNOWN:   68
##  VIC_SEX                              VIC_RACE
##  F: 2891   AMERICAN INDIAN/ALASKAN NATIVE:   13
##  M:26841   ASIAN / PACIFIC ISLANDER      :  478
##  U:   12   BLACK                         :20999
##            BLACK HISPANIC                : 2930
##            UNKNOWN                       :   72
##            WHITE                         :  741
##            WHITE HISPANIC                : 4511
```

Missing Data

```
missing_report <- shootings_clean %>%
  summarize(across(everything(), ~ round(mean(is.na(.x)) * 100, 2))) %>%
  pivot_longer(everything(), names_to = "column", values_to = "pct_missing") %>%
```

```
  filter(pct_missing > 0)
print("Missing Data by Percetage > 0")
```

```
## [1] "Missing Data by Percetage > 0"
```

```
missing_report
```

```
## # A tibble: 7 x 2
##   column              pct_missing
##   <chr>                     <dbl>
## 1 LOC_OF_OCCUR_DESC          86.0
## 2 JURISDICTION_CODE           0.01
## 3 LOC_CLASSFCTN_DESC         86.0
## 4 LOCATION_DESC              50.4
## 5 PERP_AGE_GROUP             31.4
## 6 PERP_SEX                   31.3
## 7 PERP_RACE                  31.3
```

```
shootings_clean <- shootings_clean %>%
  select(-c(LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC))
```

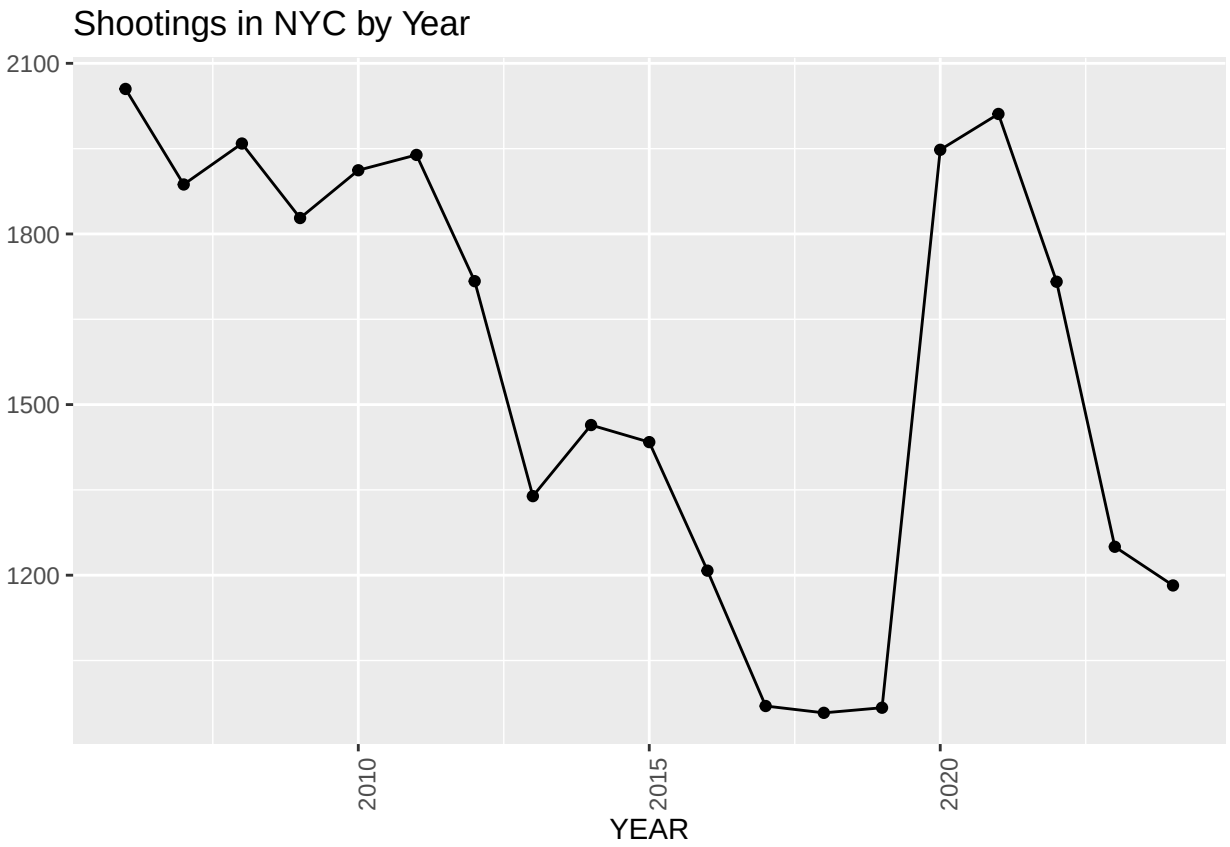
Missing data fields for perpetrator descriptions were retained due to their implication that the traits were "unconfirmed", rather than not being captured due to poor data integrity. Its possible that the missing descriptions are due to a case being unresolved. A large number of entries missing such fields may also indicate greater levels of sophistication and planning by the perpetrator. Another implication may be that people struggling with vision or memory are being targeted.

However, variables `LOC_OF_OCCUR_DESC` and `LOC_CLASSFCTN_DESC` both had a significant percentage of values missing (over 80%). For this reason, it was decided that excluding them from further analysis would be most efficient, as their potential for extracting insight was deemed insignificant.

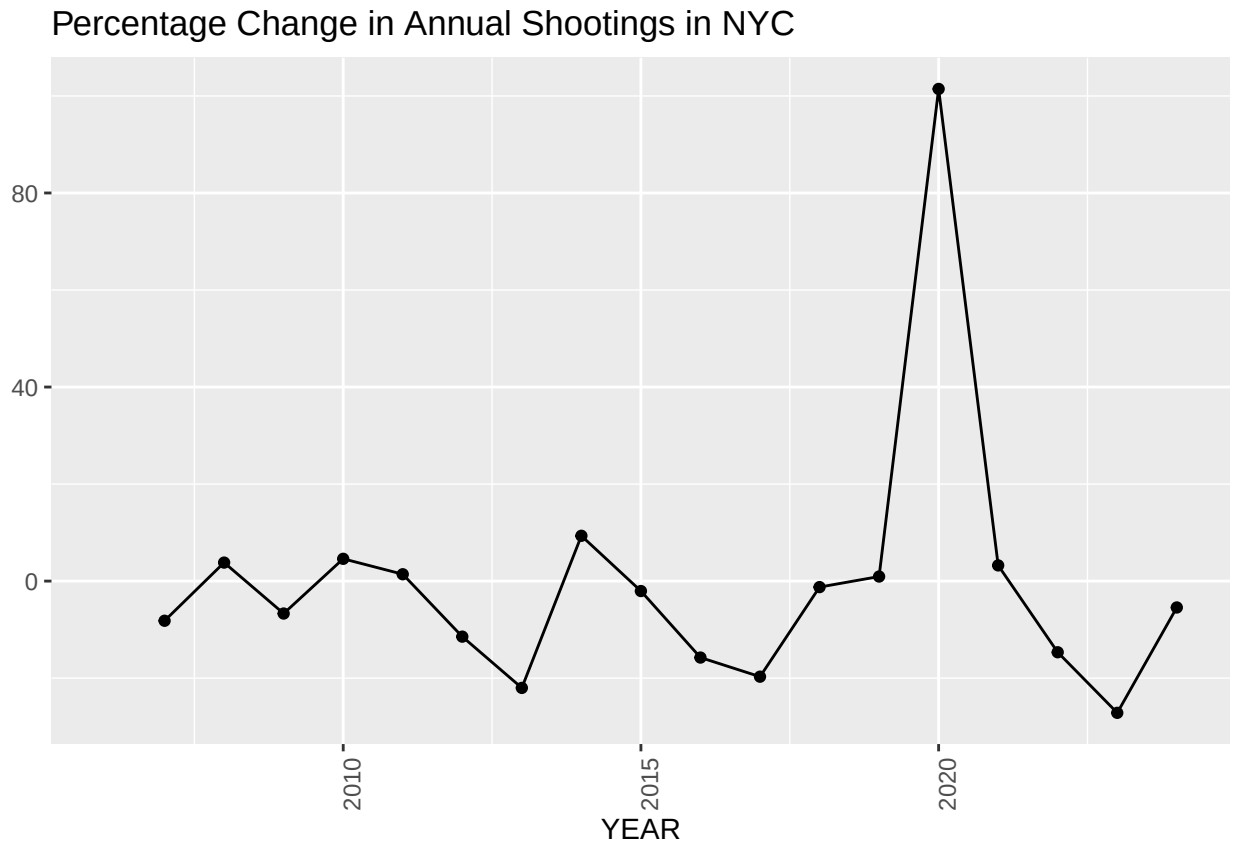
Visualize and Analyze

NYC Shootings Trends by Year

```
shootings_clean <- shootings_clean %>% # add year and percent change
  mutate(YEAR = year(OCCUR_DATE))
shootings_clean %>% # graph annual
  count(YEAR) %>%
  ggplot(aes(x = YEAR, y = n)) +
  geom_line() +
  geom_point() +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title = "Shootings in NYC by Year", y = NULL)
```

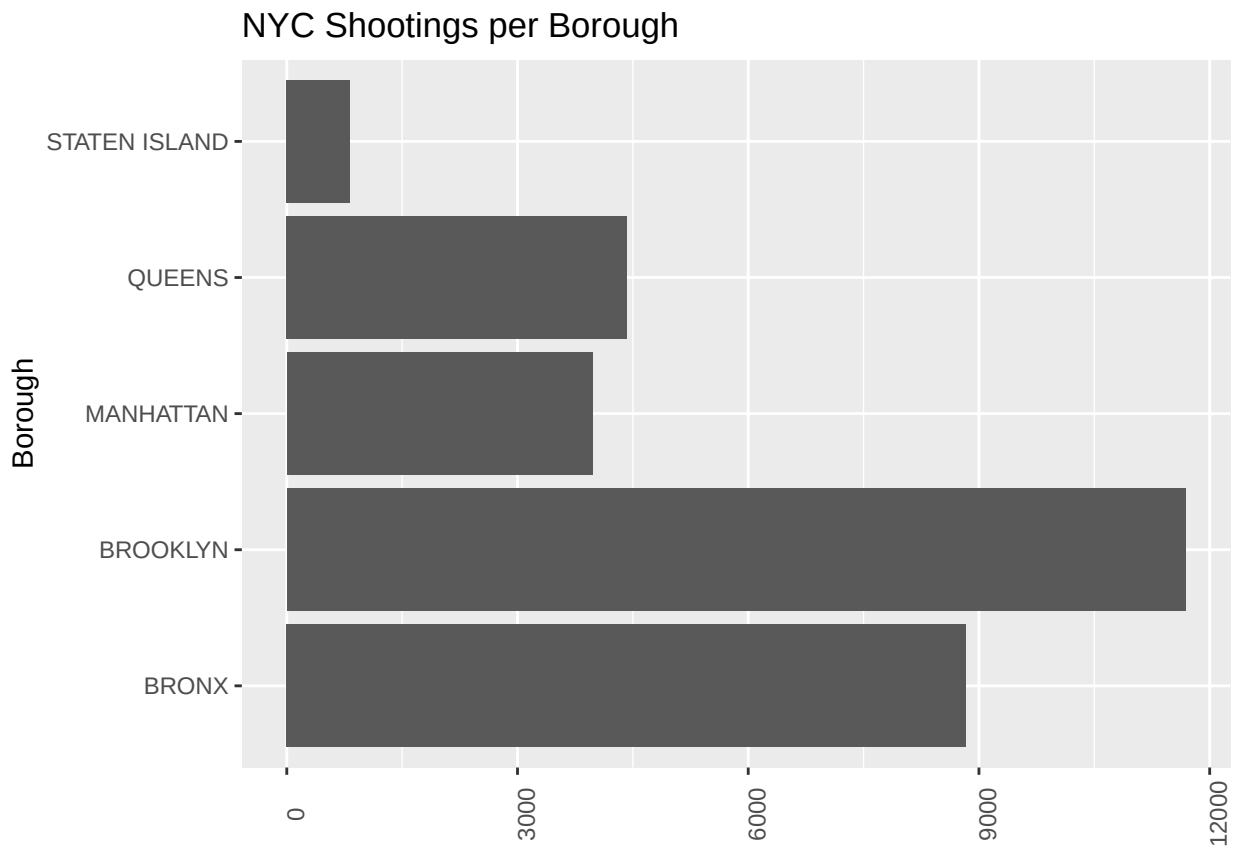



```
shootings_clean %>% # graph percent change
  count(YEAR) %>%
  mutate(pct_change = (n - lag(n)) / lag(n) * 100) %>%
  ggplot(aes(x = YEAR, y = pct_change)) +
  geom_line() +
  geom_point() +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle=90)) +
  labs(title = "Percentage Change in Annual Shootings in NYC", y = NULL)
```



NYC Shootings by Borough

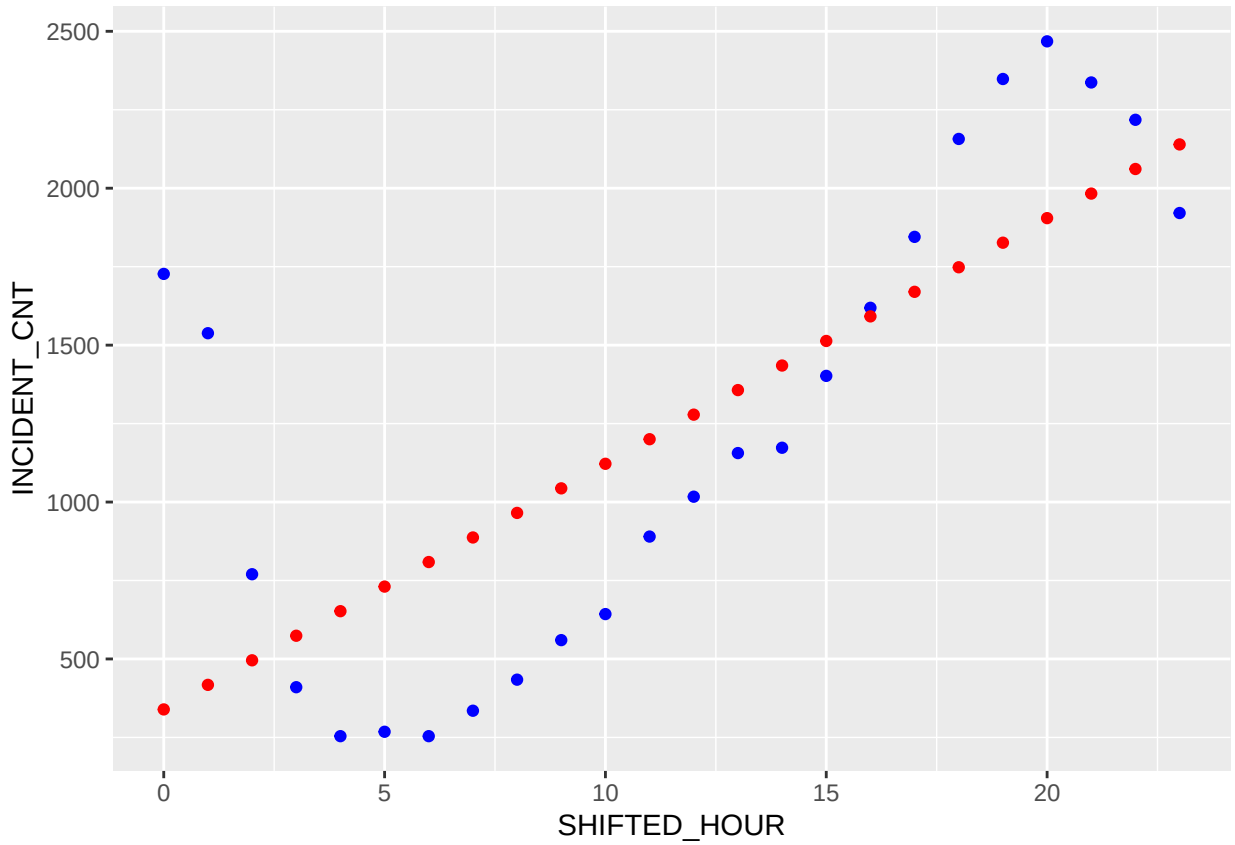
```
shootings_clean %>%
  count(BORO) %>%
  ggplot(aes(x = BORO, y = n)) +
  geom_col() +
  coord_flip() +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(title = "NYC Shootings per Borough",
       x = "Borough",
       y = NULL)
```



Modeling

Shootings by Time of Day

```
shootings_time <- shootings_clean %>%
  mutate(HOUR = hour(OCCUR_TIME),
         SHIFTED_HOUR = (HOUR - 3) %% 24) %>%
  count(SHIFTED_HOUR, name = "INCIDENT_CNT")
mod <- lm(INCIDENT_CNT ~ SHIFTED_HOUR, data = shootings_time)
shootings_time_pred <- shootings_time %>% mutate(pred = predict(mod))
shootings_time_pred %>%
  ggplot() +
  geom_point(aes(x = SHIFTED_HOUR, y = INCIDENT_CNT), color = "blue") +
  geom_point(aes(x = SHIFTED_HOUR, y = pred), color = "red")
```



When examining the number of shooting incidents by hour, a trend was revealed indicating that the likelihood of an incident was at its lowest around 6am, and increases to its peak at 11pm in a surprisingly linear manner. The times were shifted by 3 hours to better highlight the linearity. `0` on the graph represents "3am" or "0300 hours".

Additional Questions Raised

These illuminating visualizations sparked greater questions for further analysis. The inclusion of population and population density, as well as average net worth and income for particular boroughs could be a worthwhile pursuit. Additionally, seemingly unusual data points like the number of skyscrapers and tourist attractions in a given area could also yield interesting insights.

Shootings were on the decline until spiking in 2020. Was this due to the impacts of COVID and lockdowns? Did data collection standards suddenly jump, boosting the percent of shootings recorded?

Shooting incidents and time have a shockingly linear relationship; how might this be affected by the presence of "late night" businesses and activities like nightclubs and 24/7 restaurants?

Bias

Potential Bias Sources

Bias has many entry points into this research process. Below are a list of potential biases. Notice how they coincide with the **Additional Questions Raised** section above.

- Views on race.

- Police perception.
- Opinions on class and wealth.
- Prior experience with crime.
- Gun handling experience.
- Opinions of the locations themselves (specific boroughs, New York City, New York State, the United States, and North America).

Bias Mitigation Techniques

Several techniques were used to reduce introduced biases.

A fundamental starting point is to engage in self-reflection to develop an understanding of personal biases. Like the exercise proposed in the **Bias sources** lecture, take some time to write down all the biases you may have, and do your best to measure its quality and magnitude. Much like those surveys where answers scale from "*Strongly Agree*" to "*Strongly Disagree*".

Next, approach the problem from a differing point of view. Engage in a self-dialogue with these opposing viewpoints, or consult others who may be able to engage in this conversation with you, without it becoming emotionally charged. Seek to achieve a greater level of understanding.

Another useful bias techniques could be metadata obfuscation, such that a human can't impose their own bias but are still able to conduct effective research. For instance, anonymizing the characteristics of perpetrators and victims. In terms of data modeling, bias inherent in the model could be reduced by comparing the outputs of multiple, different models.

Conclusion

Results show that shooting indicents are disproportionately reported and captured in the Brooklyn and Bronx boroughs. The rate of incidents surged in 2020 to 100% and immediately returned to oscillating within a range of -20% and 10%.

Session Info

```
sessionInfo()
```

```
## R version 4.4.2 (2024-10-31)
## Platform: aarch64-apple-darwin20
## Running under: macOS Sequoia 15.5
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRlapack.dylib;  LAPACK v
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## time zone: America/Denver
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] forcats_1.0.0   stringr_1.5.1   dplyr_1.1.4     purrr_1.0.2
```

```
##  [5] readr_2.1.5     tidyr_1.3.1     tibble_3.3.0    ggplot2_3.5.2
##  [9] tidyverse_2.0.0 lubridate_1.9.4
##
## loaded via a namespace (and not attached):
##  [1] bit_4.6.0          gtable_0.3.6       jsonlite_1.8.9     crayon_1.5.3
##  [5] compiler_4.4.2     tidyselect_1.2.1   parallel_4.4.2     jquerylib_0.1.4
##  [9] scales_1.4.0       yaml_2.3.10        fastmap_1.2.0      R6_2.5.1
## [13] labeling_0.4.3     generics_0.1.4     curl_6.4.0         knitr_1.50
## [17] bslib_0.9.0        pillar_1.10.1      RColorBrewer_1.1-3 tzdb_0.5.0
## [21] rlang_1.1.6        utf8_1.2.4         stringi_1.8.4      cachem_1.1.0
## [25] xfun_0.52          sass_0.4.10        bit64_4.6.0-1      timechange_0.3.0
## [29] cli_3.6.5          withr_3.0.2        magrittr_2.0.3     digest_0.6.37
## [33] grid_4.4.2         vroom_1.6.5        hms_1.1.3          lifecycle_1.0.4
## [37] vctrs_0.6.5        evaluate_1.0.3     glue_1.8.0         farver_2.1.2
## [41] rmarkdown_2.29     tools_4.4.2        pkgconfig_2.0.3    htmltools_0.5.8.1
```