First RMD

Kevin

2025-06-30

The setup imports used librarys.

R Shootings data set

This project looks at the NYPD shooting reports. I will clean the data, create visualizations, and use this to make a model. The question I will focus on is the seasonal trend of shootings in NYC.

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

```
## 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.
```

```
shootings <- shootings %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE)) %>%
  select(-LOC_OF_OCCUR_DESC, -JURISDICTION_CODE, -LOC_CLASSFCTN_DESC, -LOCATION_DESC)
summary(shootings)
```

```
##   INCIDENT_KEY         OCCUR_DATE          OCCUR_TIME           BORO
##  Min.   :  9953245   Min.   :2006-01-01   Length:29744      Length:29744
##  1st Qu.: 67321140   1st Qu.:2009-10-29   Class1:hms        Class :character
##  Median :109291972   Median :2014-03-25   Class2:difftime   Mode  :character
##  Mean   :133850951   Mean   :2014-10-31   Mode  :numeric
##  3rd Qu.:214741917   3rd Qu.:2020-06-29
##  Max.   :299462478   Max.   :2024-12-31
##
##     PRECINCT      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP       PERP_SEX
##  Min.   :  1.00   Mode :logical           Length:29744       Length:29744
##  1st Qu.: 44.00   FALSE:23979             Class :character   Class :character
##  Median : 67.00   TRUE :5765              Mode  :character   Mode  :character
##  Mean   : 65.23
```

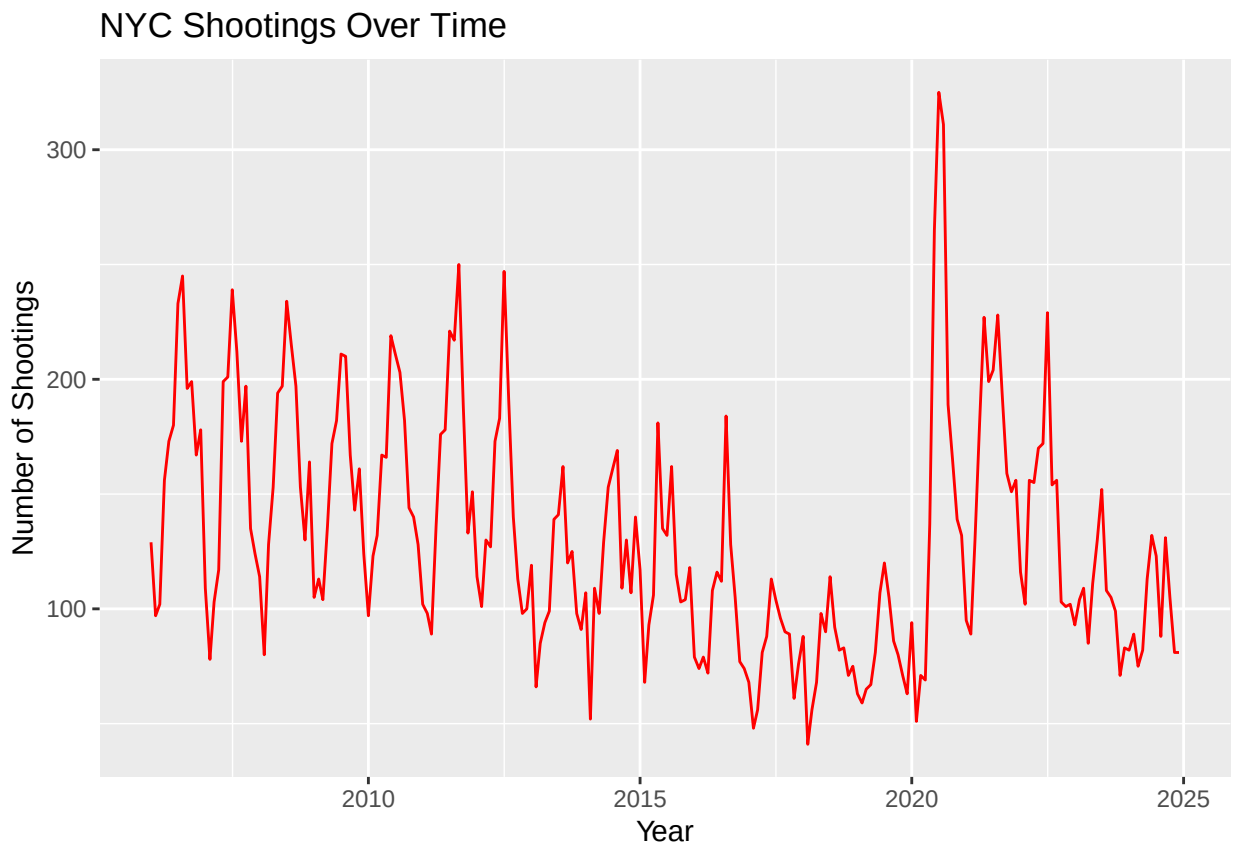
```
##  3rd Qu.: 81.00
##  Max.   :123.00
##
##   PERP_RACE         VIC_AGE_GROUP        VIC_SEX            VIC_RACE
##  Length:29744       Length:29744       Length:29744       Length:29744
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    X_COORD_CD        Y_COORD_CD        Latitude       Longitude
##  Min.   : 914928   Min.   :125757   Min.   :40.51   Min.   :-74.25
##  1st Qu.:1000094   1st Qu.:183042   1st Qu.:40.67   1st Qu.:-73.94
##  Median :1007826   Median :195506   Median :40.70   Median :-73.91
##  Mean   :1009442   Mean   :208722   Mean   :40.74   Mean   :-73.91
##  3rd Qu.:1016739   3rd Qu.:239980   3rd Qu.:40.83   3rd Qu.:-73.88
##  Max.   :1066815   Max.   :271128   Max.   :40.91   Max.   :-73.70
##                                     NA's   :97      NA's   :97
##    Lon_Lat
##  Length:29744
##  Class :character
##  Mode  :character
##
##
##
##
```

This data set describes shootings in New York City from 2006 to 2024. Variables include date and time, borough, precinct, perpetrator age, and many more. I removed some codes and columns such as locations that were not useful to me. Some columns such as info on the PERP have missing data. If we continue to work with this, we would either need to filter out empty values, or not use these columns.

Visualizations and Analysis

Plot 1: Shootings by Month

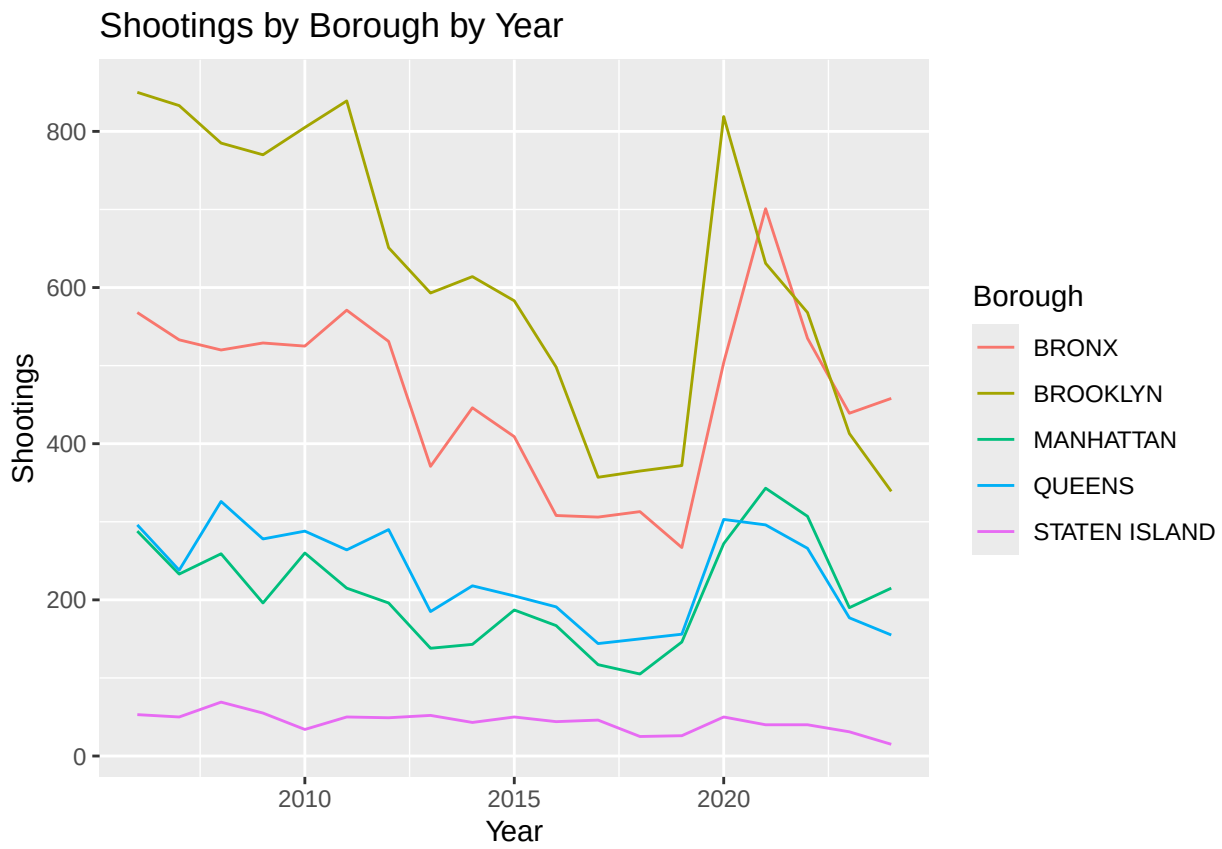
```
monthly_shootings <- shootings %>%
  mutate(month = lubridate::floor_date(OCCUR_DATE, unit = "month")) %>%
  count(month, name = "shootings")
ggplot(monthly_shootings, aes(x = month, y = shootings)) +
  geom_line(color = "red") +
  labs(
    title = "NYC Shootings Over Time",
    x = "Year",
    y = "Number of Shootings"
  )
```



This visualization shows that there may be a seasonal trend. It also shows that there was a peak in a month in 2020. Potential future questions could ask these seasonal trend, and if covid plays a role in the large peak.

```
yearly_shootings <- shootings %>%
  mutate(year = lubridate::year(OCCUR_DATE)) %>%
  group_by(year, BORO) %>%
  summarize(shootings = n(), .groups = "drop")

ggplot(yearly_shootings, aes(x = year, y = shootings, color = BORO)) +
  geom_line() +
  labs(
    title = "Shootings by Borough by Year",
    x = "Year",
    y = "Shootings",
    color = "Borough"
  )
```



This visualization breaks down the shootings by year and by Borough. Again this shows the peak in 2020. This also shows that the Bronx and Brooklyn have the highest number of shootings. One additional question would be how does this compare to the per capita shootings. Another question would be what caused the trend downwards, and can this be repeated.

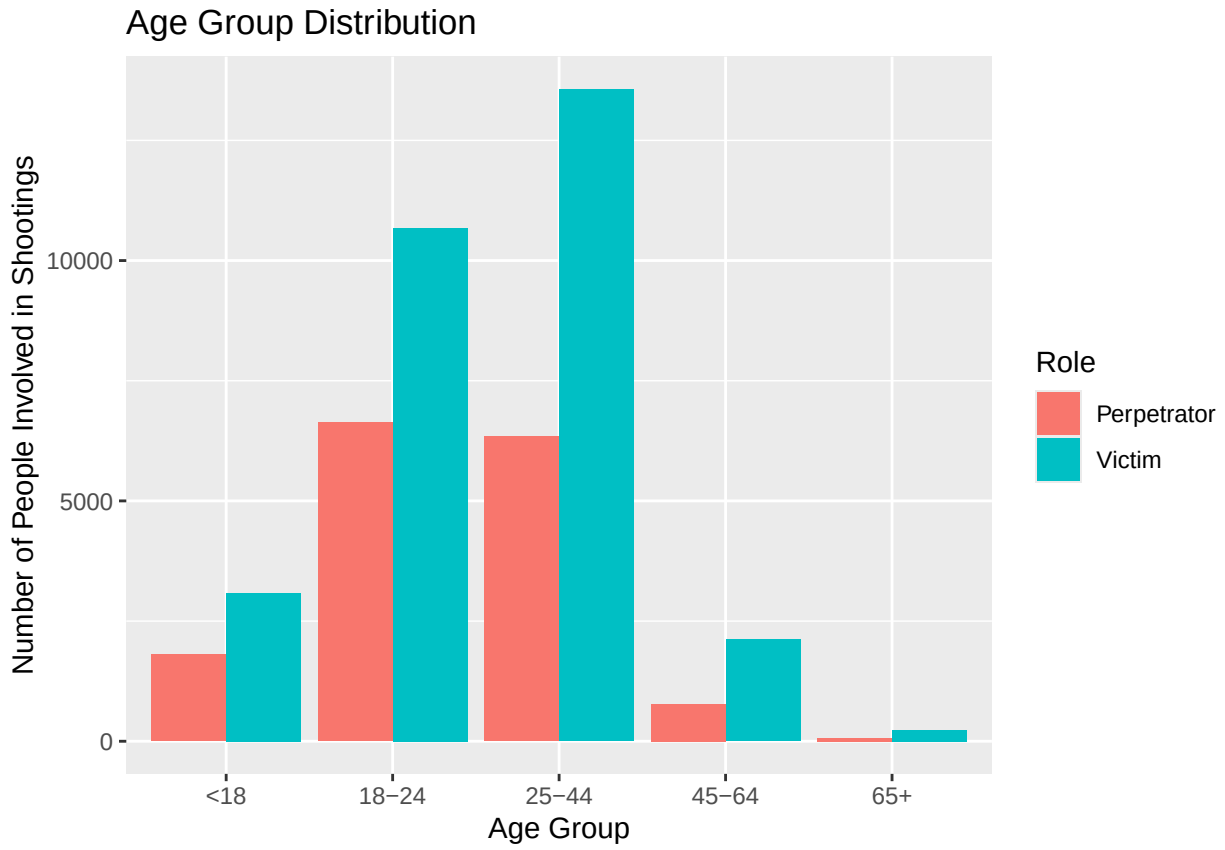
```
age_data <- shootings %>%
  pivot_longer(cols = c(PERP_AGE_GROUP, VIC_AGE_GROUP), names_to = "Role", values_to = "AgeGroup") %>%
  filter(!is.na(AgeGroup))

clean_age_data <- shootings %>%
  pivot_longer(cols = c(PERP_AGE_GROUP, VIC_AGE_GROUP),
               names_to = "Role", values_to = "AgeGroup") %>%
  filter(AgeGroup %in% c("<18", "18-24", "25-44", "45-64", "65+"))  # Keep only valid groups

clean_age_data$AgeGroup <- factor(clean_age_data$AgeGroup,
  levels = c("<18", "18-24", "25-44", "45-64", "65+"))

ggplot(clean_age_data, aes(x = AgeGroup, fill = Role)) +
  geom_bar(position = "dodge") +
  labs(
    title = "Age Group Distribution",
    x = "Age Group",
    y = "Number of People Involved in Shootings",
    fill = "Role"
  ) +
  scale_fill_discrete(
```

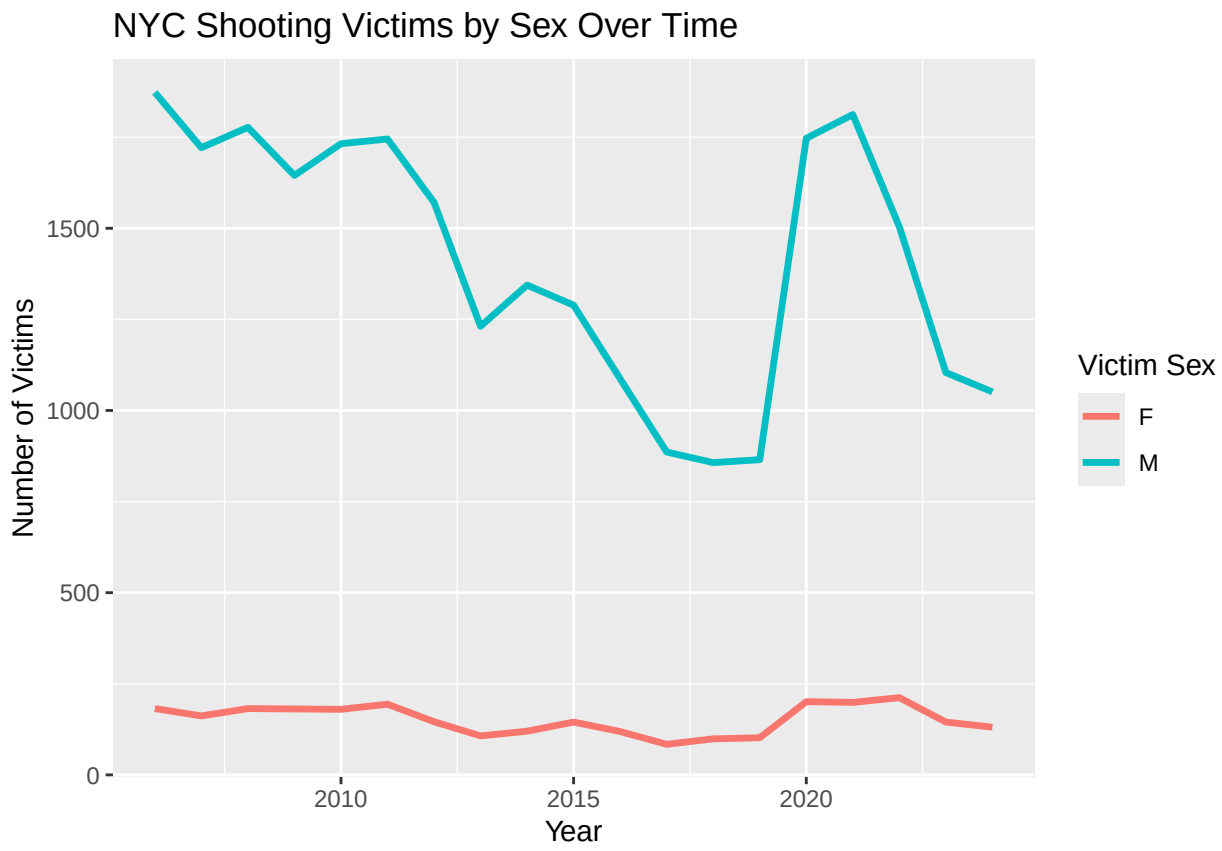
```
    labels = c("PERP_AGE_GROUP" = "Perpetrator", "VIC_AGE_GROUP" = "Victim")
  )
```




```
sex_time <- shootings %>%
  mutate(year = lubridate::year(OCCUR_DATE)) %>%
  filter(!is.na(VIC_SEX), VIC_SEX != "UNKNOWN", VIC_SEX != "U") %>%
  count(year, VIC_SEX)

# Plot
ggplot(sex_time, aes(x = year, y = n, color = VIC_SEX)) +
  geom_line(size = 1.2) +
  labs(
    title = "NYC Shooting Victims by Sex Over Time",
    x = "Year",
    y = "Number of Victims",
    color = "Victim Sex")
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



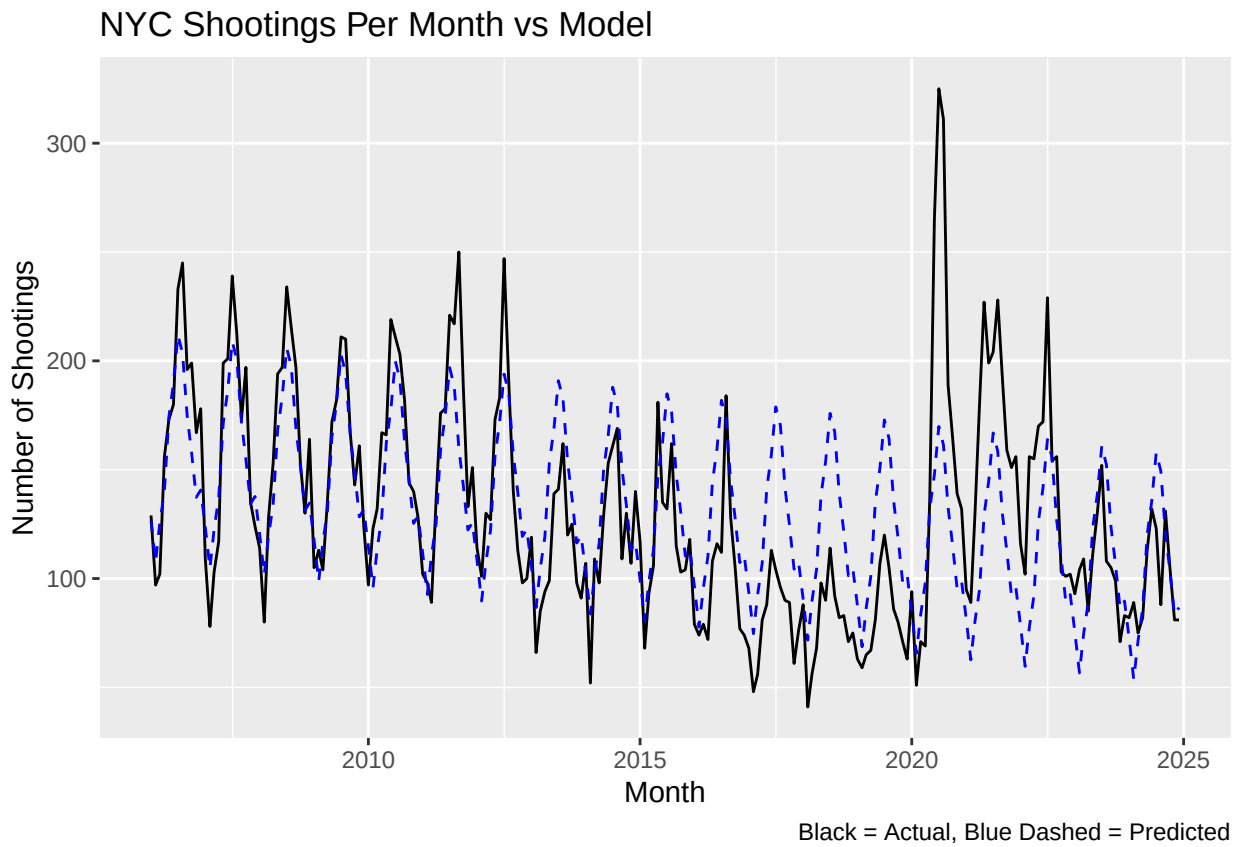

```
monthly_shootings <- monthly_shootings %>%
  mutate(
    time_index = row_number(),
    month_factor = factor(lubridate::month(month, label = TRUE))
  )
model <- lm(shootings ~ time_index + month_factor, data = monthly_shootings)
summary(model)
```

```
##
## Call:
## lm(formula = shootings ~ time_index + month_factor, data = monthly_shootings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -75.414 -24.679  -3.673  18.554 155.116
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     159.10207    4.96654  32.035  < 2e-16 ***
## time_index       -0.25018    0.03762  -6.651 2.38e-10 ***
## month_factor.L   41.91221    8.57694   4.887 2.00e-06 ***
## month_factor.Q  -84.42430    8.56513  -9.857  < 2e-16 ***
## month_factor.C  -26.13623    8.56513  -3.051  0.00256 **
## month_factor^4   50.70221    8.56513   5.920 1.26e-08 ***
## month_factor^5    6.66955    8.56513   0.779  0.43702
```

```
## month_factor^6   -2.80864    8.56513  -0.328  0.74329
## month_factor^7   -6.63338    8.56513  -0.774  0.43951
## month_factor^8    5.47610    8.56513   0.639  0.52328
## month_factor^9    3.56102    8.56513   0.416  0.67800
## month_factor^10   7.99088    8.56513   0.933  0.35189
## month_factor^11  -5.76779    8.56513  -0.673  0.50141
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.33 on 215 degrees of freedom
## Multiple R-squared:  0.4936, Adjusted R-squared:  0.4654
## F-statistic: 17.47 on 12 and 215 DF,  p-value: < 2.2e-16
```

```r
monthly_shootings <- monthly_shootings %>%
  mutate(predicted = predict(model))

ggplot(monthly_shootings, aes(x = month)) +
  geom_line(aes(y = shootings), color = "black") +
  geom_line(aes(y = predicted), color = "blue", linetype = "dashed") +
  labs(
    title = "NYC Shootings Per Month vs Model",
    x = "Month",
    y = "Number of Shootings",
    caption = "Black = Actual, Blue Dashed = Predicted"
  )
```



I ran a model to check the linear regression of time and seasonal effects. I had to add variables to track the

passage of time and a value to represent the month of the year. The model shows both a strong downward trend over time, and a strong seasonal effect. There is a large outlier in 2020, most likely due to covid. The R-squared value of 0.49 shows that this model explains about half the variation, meaning time and season are important predictors of this violence.

Conclusion

This markdown file shows an initial look into the NYC Shooting data set. The first code chunk imports the data via a CSV and does some organization. Then I plotted the data to show two visualizations. The first looks at shootings per month and we can see seasonal trends. The second breaks the shootings down by Borough and year, to see the trends across Boroughs. I used a model to fit a linear regression to the first plot. It shows that time and season are strong predictors of the shootings, but not the only factors. One source of bias is that this was an assigned project. I completed this as I was watching lectures from the class, so I followed similar techniques. One way to counter this bias is to not draw conclusions past what is shown. This analysis is would only be a start to more in depth questions which could be explored in a more in depth study. Other sources of bias include the method of data collection, the fact that it involves the NYPD and how the shootings were reported. My session info

```
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/Los_Angeles
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] lubridate_1.9.4 forcats_1.0.0   stringr_1.5.1   dplyr_1.1.4
##  [5] purrr_1.0.4     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] bit_4.5.0.1       gtable_0.3.6      crayon_1.5.3      compiler_4.4.2
##  [5] tidyselect_1.2.1  parallel_4.4.2    scales_1.3.0      yaml_2.3.10
##  [9] fastmap_1.2.0     R6_2.5.1          labeling_0.4.3    generics_0.1.3
## [13] curl_6.2.0        knitr_1.49        munsell_0.5.1     pillar_1.10.1
## [17] tzdb_0.4.0        rlang_1.1.5       stringi_1.8.4     xfun_0.50
## [21] bit64_4.6.0-1     timechange_0.3.0  cli_3.6.3         withr_3.0.2
## [25] magrittr_2.0.3    digest_0.6.37     grid_4.4.2        vroom_1.6.5
```

```
## [29] rstudioapi_0.17.1 hms_1.1.3         lifecycle_1.0.4   vctrs_0.6.5
## [33] evaluate_1.0.3    glue_1.8.0        farver_2.1.2      colorspace_2.1-1
## [37] rmarkdown_2.29    tools_4.4.2       pkgconfig_2.0.3   htmltools_0.5.8.1
```