A Look at NYPD Reported Murders Over Time: 2006 - 2024

2026-01-16

For this project, I am interested in looking at how the number of murders in New York City (NYC) changes from 2006 to 2024. To investigate this question, we will import and tidy data on historic shootings provided by the New York City Police Department, visualize and analyze the changes to the data over time, and conclude with a discussion of next steps and bias mitigation.

Import data

First we will read in a CSV file containing the historic data on NYPD shootings.

```
library("tidyverse")
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.6
## v forcats   1.0.1     v stringr   1.6.0
## v ggplot2   4.0.1     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.2
## v purrr     1.2.0
## -- 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
```

```
library("lubridate")
```

```
# Define URL and read in CSV file
url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv"
nypd_data <- read_csv(url)
```

```
## 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.
```

We can view a summary of the data upon first import, before we do any tidying.

```r
# Print a summary of the data
summary(nypd_data)
```

```
##   INCIDENT_KEY        OCCUR_DATE          OCCUR_TIME
##  Min.   :  9953245   Length:29744       Min.   :00:00:00.000000
##  1st Qu.: 67321140   Class :character   1st Qu.:03:30:45.000000
##  Median :109291972   Mode  :character   Median :15:15:00.000000
##  Mean   :133850951                      Mean   :12:46:10.874798
##  3rd Qu.:214741917                      3rd Qu.:20:44:00.000000
##  Max.   :299462478                      Max.   :23:59:00.000000
##
##      BORO           LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE
##  Length:29744       Length:29744       Min.   :  1.00   Min.   :0.0000
##  Class :character   Class :character   1st Qu.: 44.00   1st Qu.:0.0000
##  Mode  :character   Mode  :character   Median : 67.00   Median :0.0000
##                                        Mean   : 65.23   Mean   :0.3181
##                                        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
##  Length:29744       Length:29744       Mode :logical
##  Class :character   Class :character   FALSE:23979
##  Mode  :character   Mode  :character   TRUE :5765
##
##
##
##
##  PERP_AGE_GROUP       PERP_SEX           PERP_RACE          VIC_AGE_GROUP
##  Length:29744       Length:29744       Length:29744       Length:29744
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_SEX             VIC_RACE           X_COORD_CD        Y_COORD_CD
##  Length:29744       Length:29744       Min.   : 914928   Min.   :125757
##  Class :character   Class :character   1st Qu.:1000094   1st Qu.:183042
##  Mode  :character   Mode  :character   Median :1007826   Median :195506
##                                        Mean   :1009442   Mean   :208722
##                                        3rd Qu.:1016739   3rd Qu.:239980
##                                        Max.   :1066815   Max.   :271128
##
##     Latitude       Longitude        Lon_Lat
##  Min.   :40.51   Min.   :-74.25   Length:29744
##  1st Qu.:40.67   1st Qu.:-73.94   Class :character
##  Median :40.70   Median :-73.91   Mode  :character
##  Mean   :40.74   Mean   :-73.91
##  3rd Qu.:40.83   3rd Qu.:-73.88
##  Max.   :40.91   Max.   :-73.70
##  NA's   :97      NA's   :97
```

Tidy and Transform

Next, we will tidy and transform the data and then re-check a summary of the data and complete sanity checks before proceeding further.

```
# Change appropriate variables to factor and date types

# Change the data column to a date object (the time column is already a time object)
nypd_data <- nypd_data %>%
              mutate(OCCUR_DATE = mdy(OCCUR_DATE))

# Define the factors
col_names <- sapply(nypd_data, function(col) length(unique(col)) < 15 & !is.logical(col))
nypd_data <- nypd_data %>%
  mutate(across(names(col_names)[col_names], as.factor))

# Get rid of any columns not needed
nypd_data <- nypd_data %>%
              select(-c(Latitude, Longitude, Lon_Lat, X_COORD_CD, Y_COORD_CD))

# Check the data summary again
summary(nypd_data)
```

```
##   INCIDENT_KEY         OCCUR_DATE           OCCUR_TIME
##  Min.   :  9953245   Min.   :2006-01-01   Min.   :00:00:00.000000
##  1st Qu.: 67321140   1st Qu.:2009-10-29   1st Qu.:03:30:45.000000
##  Median :109291972   Median :2014-03-25   Median :15:15:00.000000
##  Mean   :133850951   Mean   :2014-10-31   Mean   :12:46:10.874798
##  3rd Qu.:214741917   3rd Qu.:2020-06-29   3rd Qu.:20:44:00.000000
##  Max.   :299462478   Max.   :2024-12-31   Max.   :23:59:00.000000
##
##             BORO       LOC_OF_OCCUR_DESC    PRECINCT      JURISDICTION_CODE
##  BRONX        : 8834   INSIDE :  682     Min.   :  1.00   0   :24957
##  BROOKLYN     :11685   OUTSIDE: 3466     1st Qu.: 44.00   1   :  109
##  MANHATTAN    : 3977   NA's   :25596     Median : 67.00   2   : 4676
##  QUEENS       : 4426                     Mean   : 65.23   NA's:    2
##  STATEN ISLAND:  822                     3rd Qu.: 81.00
##                                          Max.   :123.00
##
##   LOC_CLASSFCTN_DESC LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  STREET    : 2639    Length:29744       Mode :logical           18-24  :6630
##  HOUSING   :  643    Class :character   FALSE:23979             25-44  :6342
##  DWELLING  :  341    Mode  :character   TRUE :5765              UNKNOWN:3148
##  COMMERCIAL:  276                                               <18    :1805
##  OTHER     :   74                                               (null) :1628
##  (Other)   :  175                                               (Other): 847
##  NA's      :25596                                               NA's   :9344
##    PERP_SEX              PERP_RACE     VIC_AGE_GROUP   VIC_SEX
##  (null): 1628   BLACK         :12323   <18    : 3081   F: 2891
##  F     :  461   WHITE HISPANIC: 2667   1022   :    1   M:26841
##  M     :16845   UNKNOWN       : 1838   18-24  :10677   U:   12
##  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_RACE
##  AMERICAN INDIAN/ALASKAN NATIVE:   13
##  ASIAN / PACIFIC ISLANDER      :  478
##  BLACK                         :20999
##  BLACK HISPANIC                : 2930
##  UNKNOWN                       :   72
##  WHITE                         :  741
##  WHITE HISPANIC                : 4511
```

As we take a look at the data after it's been tidied, we can do some initial assessments.

- The data ranges from 2006 to 2024 and time from 0:00 to 24:00, which is all logical.
- All 5 boroughs are included, though we do not know if some may be over or underrepresented.
- Confirmed that even though there are only 78 police precincts in New York City, their numbers do in fact go up to 123 (some numbers are skipped), so this column does make sense.
- The 2 columns that track age groups include some illogical numbers (e.g. 1028, 1020, 940, 1022, etc). This would need to be investigated more, however it is not data that we are including in this report's analysis.

Missing Data

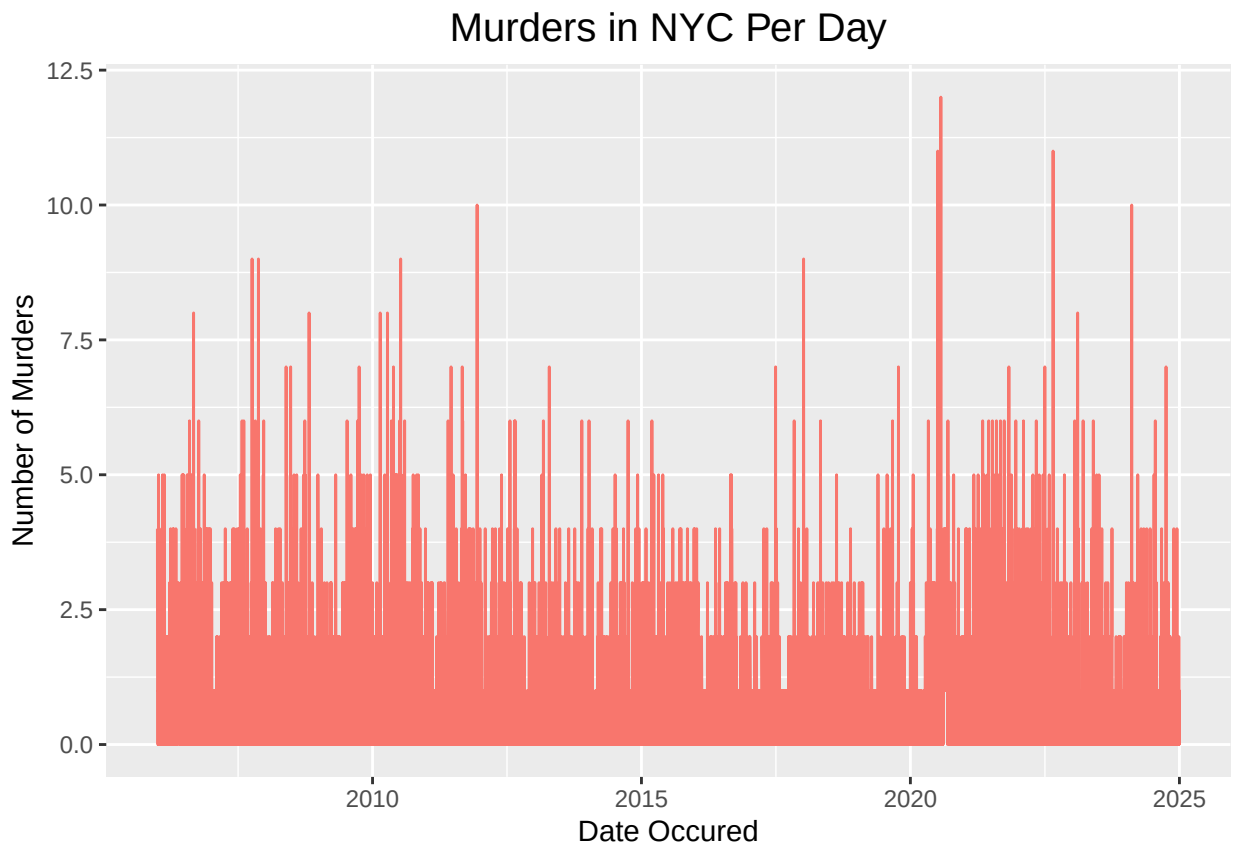
There are several columns that are missing data for some incidents. These include the classification and description of where the shootings occurred, and the age group, race, and sex of the people involved. A few of these columns contain missing data formatted with multiple different undefined terms, including "null", "NA", and "Unknown" all in the same column. We could deal with these values by converting them all to the same undefined term, but before doing this I would want to first read through any documentation on the data set to see if the different terms refer to different information that we may want to use.

However, with all that being said, these instances of missing data do not apply to any of the columns used for the analysis in this report, so I feel confident moving forward with my analysis.

Visualizations and Analysis

We can start the analysis of how the murders in NYC changes over time by first producing a basic plot of the number of murders per day.

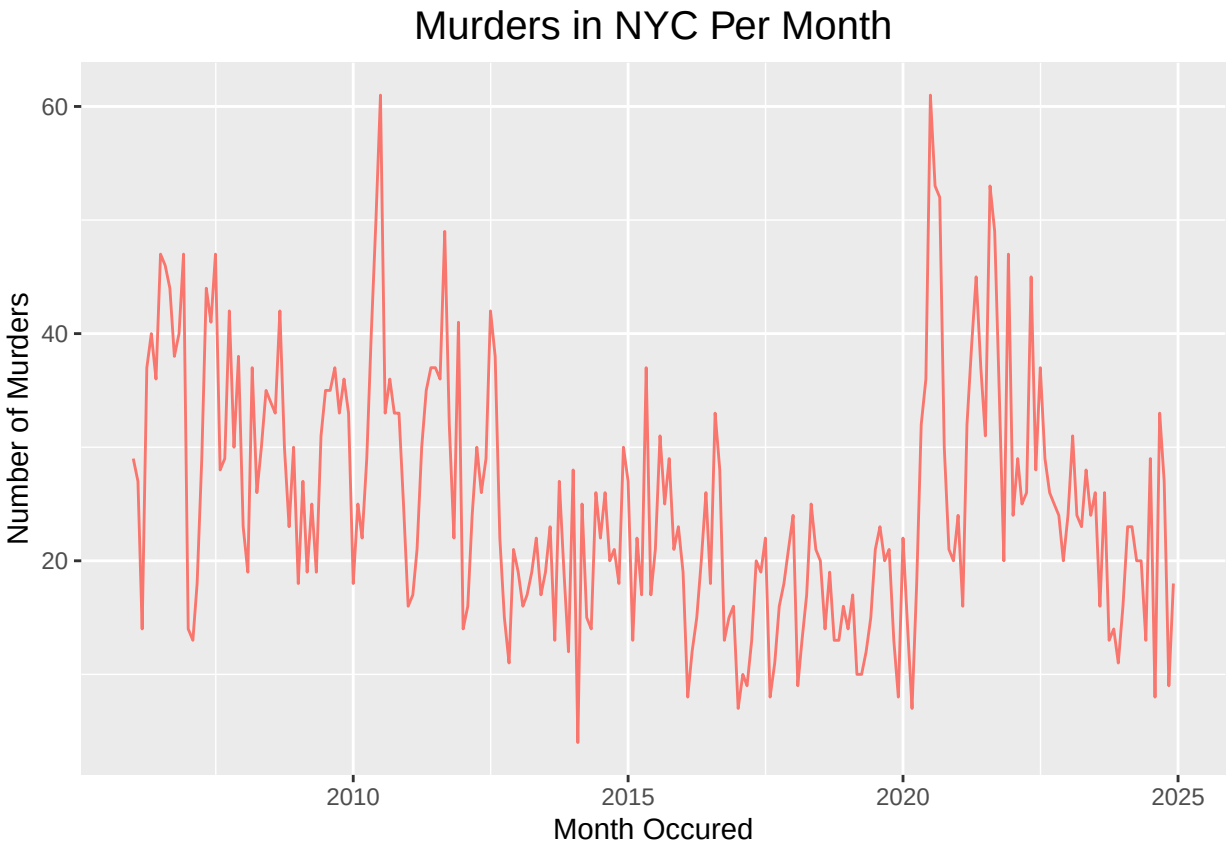
```r
# Group rows by date and summarize by summing the number of murder flags
nypd_data %>%
  group_by(OCCUR_DATE) %>%
  summarize(MURDERS = sum(STATISTICAL_MURDER_FLAG)) %>%
  ungroup() %>%
  ggplot(aes(x = OCCUR_DATE, y = MURDERS)) +
  geom_line(aes(color = "Murders"), show.legend = FALSE) +
  labs(title="Murders in NYC Per Day", x="Date Occured", y="Number of Murders") +
  theme(plot.title = element_text(size=15, hjust = 0.5))
```



It might be easier to see trends over the time period by grouping our crimes into months instead of days, so we'll plot this as well.

```
# Groups all dates in the same month by the first day of that month
# and summarize by summing the number of murder flags
nypd_data_months <- nypd_data %>%
    mutate(OCCUR_MONTH = floor_date(OCCUR_DATE, "month")) %>%
    group_by(OCCUR_MONTH) %>%
    summarize(MURDERS = sum(STATISTICAL_MURDER_FLAG)) %>%
    ungroup() %>%
    select(OCCUR_MONTH, MURDERS, everything())

# Plot
nypd_data_months %>%
  ggplot(aes(x = OCCUR_MONTH, y = MURDERS)) +
  geom_line(aes(color = "Murders"), show.legend = FALSE) +
  labs(title="Murders in NYC Per Month", x="Month Occured", y="Number of Murders") +
  theme(plot.title = element_text(size=15, hjust = 0.5))
```



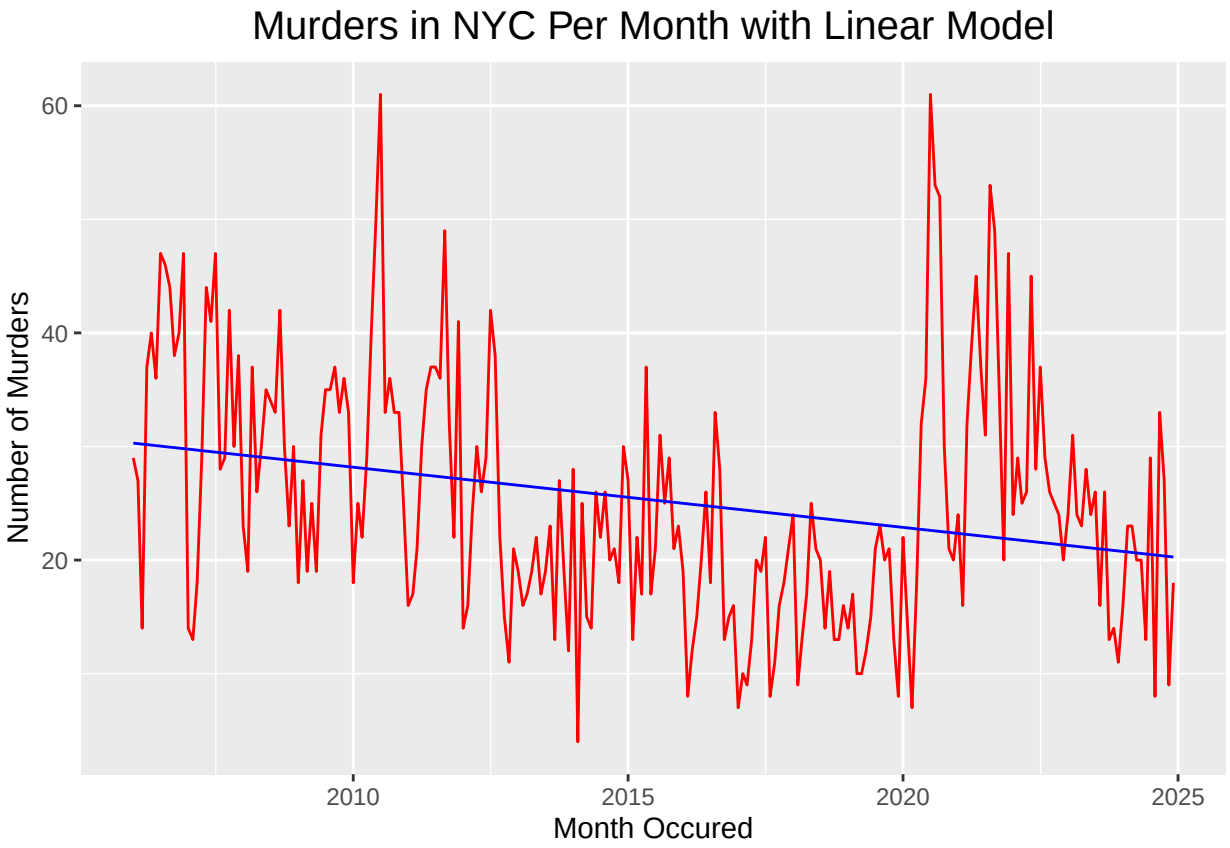
From this plot we can see more trends emerge. One that jumps out is an overall decrease in murders from around 2013 through 2020. Next, we can check to see if there is any overall trend to the number of murders per month by applying a linear model to the data.

```
# Create a linear model based on the data
model <- lm(MURDERS ~ OCCUR_MONTH, data = nypd_data_months)
summary(model)
```

```
##
## Call:
## lm(formula = MURDERS ~ OCCUR_MONTH, data = nypd_data_months)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.012  -7.685  -1.314   6.441  38.386
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 49.3741176  5.8403395   8.454 3.49e-15 ***
## OCCUR_MONTH -0.0014509  0.0003492  -4.155 4.62e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.56 on 226 degrees of freedom
## Multiple R-squared:  0.07095,    Adjusted R-squared:  0.06684
## F-statistic: 17.26 on 1 and 226 DF,  p-value: 4.625e-05
```

```r
# Apply our model and plot actual vs modeled data
nypd_data_modeled <- nypd_data_months %>% mutate(MODELED = predict(model))

nypd_data_modeled %>% ggplot() +
    geom_line(aes(x = OCCUR_MONTH, y = MURDERS), color = "red") +
    geom_line(aes(x = OCCUR_MONTH, y = MODELED), color = "blue") +
    labs(title="Murders in NYC Per Month with Linear Model",
         x="Month Occured", y="Number of Murders") +
    theme(plot.title = element_text(size=15, hjust = 0.5))
```



It's obvious that this data is not perfectly fit by a linear model. However, it can provide some initial insight that the overall number of reported murders in New York City has decreased over the time period this data covers. According to our model, at the start of our data in January 2006, the average number of murders as approximately 30 per month, and by the end of our data in December 2024 that number dropped to approximately 20 per month.

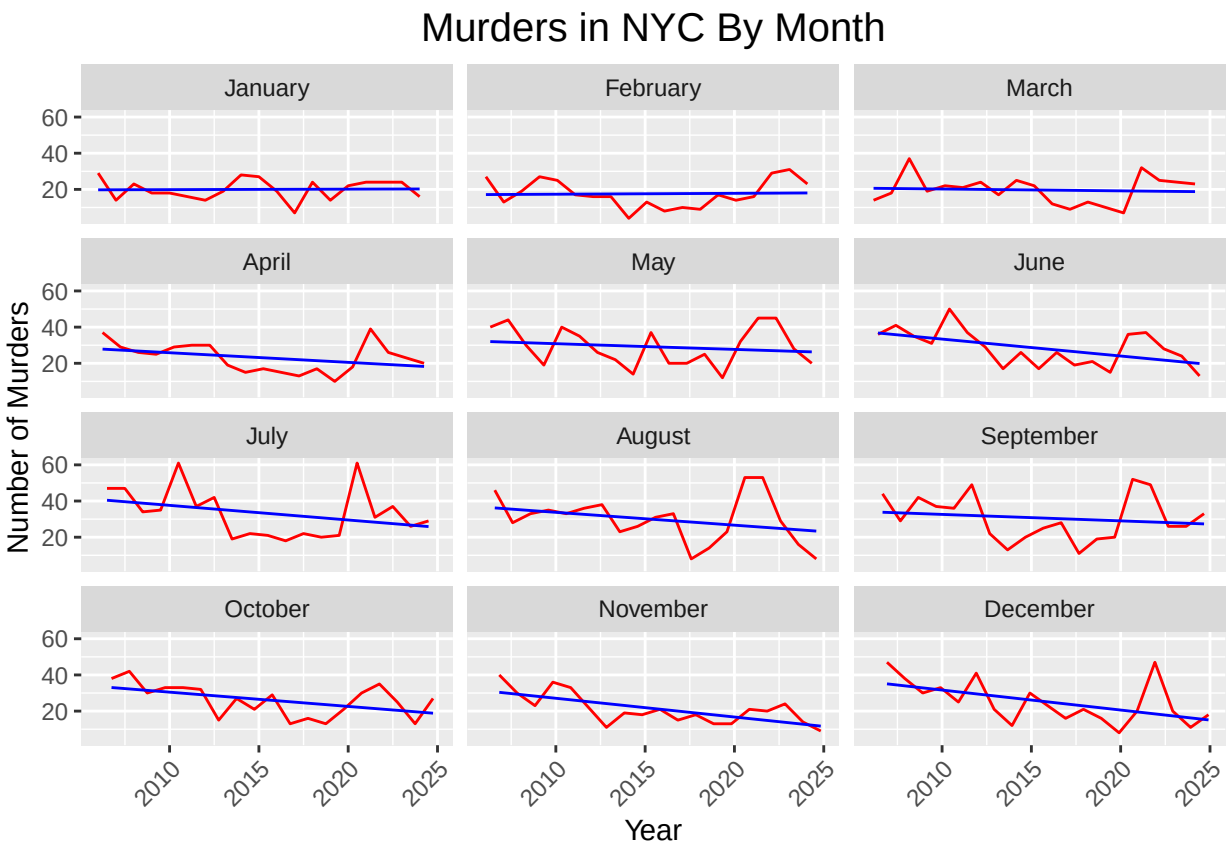
To dive slightly deeper, I decided to next look at how the number of murders varies for each month. To do this, I plotted the number of murders for each month across the years, and fit each of these with their own models.

```r
# Add a column for the month name only
nypd_data_by_month <- nypd_data_months %>% mutate(MONTH = month(OCCUR_MONTH))

# Create number to name vector for labeller
month_names <- c(month.name)
names(month_names) <- c(1:12)
```

```r
# Create models for each month
model_data <- nypd_data_by_month %>%
  group_by(MONTH) %>%
  do({
    model <- lm(MURDERS ~ OCCUR_MONTH, data = .)
    data.frame(
      OCCUR_MONTH = seq(min(.$OCCUR_MONTH), max(.$OCCUR_MONTH), by="year"),
      MURDERS = predict(model, newdata = data.frame(OCCUR_MONTH =
                                                      seq(min(.$OCCUR_MONTH),
                                                          max(.$OCCUR_MONTH),
                                                          by="year")))
    )
  }) %>%
  ungroup()

# Plot data and models for each month
ggplot(nypd_data_by_month, aes(x = OCCUR_MONTH, y = MURDERS)) +
    geom_line(color = "red") +
    labs(x = "Year", y = "Number of Murders") +
    theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1)) +
    facet_wrap(~ MONTH, nrow = 4, labeller = labeller(MONTH = month_names)) +
    geom_line(data=model_data, color = "blue") +
    labs(title="Murders in NYC By Month") +
    theme(plot.title = element_text(size=15, hjust = 0.5))
```



From these plots, you can see the differences in the total number of murders for certain months. According to the model, on average January has a lower total number of murders than September. You can also see that there are differences in how the data is trending. For example, October, November, and December have had their number of murders decrease over the time period, while January, February, and March have mostly stayed the same over the years. You can also see that some months like May and September have had larger oscillations in their numbers over the years.

Looking back at the graph that included all the months together, there was a large dip in murders from 2013-2020. However, comparing that trend to these month by month plots shows that this dip isn't apparent in all the months individually. It seems like that overall dip might have been more influenced by the numbers in July and September than some other months like January. This could help us narrow down what happened in New York to cause this multi-year dip. For example, maybe there were programs in place during these months that helped decrease the overall murder rate.

Conclusion

In this project, we imported our data in a reproducible manner, tidied it and performed sanity checks to ensure its validity, and then did analysis in order to answer the question of how the number of murders in New York City are changing over time. Through plotting and modeling the data, we found that the number of murders in New York City do appear to have decreased from 2006 to 2024. We also found the trends varied greatly when examining each month individually. However, both these conclusions were found through a simple procedure and much more work can be done to better understand the data.

Additional Questions

When thinking about next steps to investigate the data, there are plenty of other ideas that could be explored. For example, it would be interesting to look at if a non-linear model could fit more of the trends of this data. Also, it is reported that the data set counts shootings that killed multiple people as different shootings, so it would be interesting to see if any features in the data change when you combine these multi-victim shooting instances.

Bias Identification

When identifying bias in this report, its good to look at both the biases present in the data, and my own personal biases.

Looking at bias in the data that could affect our analysis, an important thing to note is that the data only includes crimes that police are aware of and investigate. This data set could potentially be missing crimes where victims, families, or witnesses would not want to talk with the police, such as those involving immigrants or other marginalized groups with histories of distrust in the police. Therefore, when using this data set, we cannot make assumptions about crime in New York City in general - only reported crime.

My own personal bias could also impact the conclusions we draw from this data. For example, I have never lived in New York City and this could impact my understanding of the best ways to group this data. Perhaps things like weather and local events would be more useful to group the data by than months, but I don't know that. I am biased by my experiences in my own city and am potentially applying those experiences to New York without taking into account how they may be different.

There is also bias in what I have chosen to research in the first place. I intentionally chose not to analyze anything involving the race of the perpetrators and victims because as a white person, I know that I will always be under-informed about issues of race and policing in America and wanted to avoid any misunderstandings from my own lack of knowledge. However, this is still bias and clearly not a long term solution as these topics need to be studied.

Session Info

```
sessionInfo()
```

```
## R version 4.5.2 (2025-10-31)
## Platform: x86_64-apple-darwin20
## Running under: macOS Sequoia 15.7.3
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.5-x86_64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.5-x86_64/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/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_1.0.1   stringr_1.6.0   dplyr_1.1.4
##  [5] purrr_1.2.0     readr_2.1.6     tidyr_1.3.2     tibble_3.3.0
##  [9] ggplot2_4.0.1   tidyverse_2.0.0
##
## loaded via a namespace (and not attached):
##  [1] bit_4.6.0          gtable_0.3.6       crayon_1.5.3       compiler_4.5.2
##  [5] tidyselect_1.2.1   parallel_4.5.2     scales_1.4.0       yaml_2.3.12
##  [9] fastmap_1.2.0      R6_2.6.1           labeling_0.4.3     generics_0.1.4
## [13] curl_7.0.0         knitr_1.51         pillar_1.11.1      RColorBrewer_1.1-3
## [17] tzdb_0.5.0         rlang_1.1.6        stringi_1.8.7      xfun_0.55
## [21] S7_0.2.1           bit64_4.6.0-1      timechange_0.3.0   cli_3.6.5
## [25] withr_3.0.2        magrittr_2.0.4     digest_0.6.39      grid_4.5.2
## [29] vroom_1.6.7        rstudioapi_0.17.1  hms_1.1.4          lifecycle_1.0.4
## [33] vctrs_0.6.5        evaluate_1.0.5     glue_1.8.0         farver_2.1.2
## [37] rmarkdown_2.30     tools_4.5.2        pkgconfig_2.0.3    htmltools_0.5.9
```