NYPD

Ben Lewis

2025-01-10

Step 1: Start an Rmd Document

Start an Rmd document that describes and imports the shooting project dataset in a reproducible manner.

```
library(tidyverse)
```

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

Step 2: Tidy and Transform Your Data

Add to your Rmd document a summary of the data and clean up your dataset by changing appropriate variables to factor and date types and getting rid of any columns not needed. Show the summary of your data to be sure there is no missing data. If there is missing data, describe how you plan to handle it.

```
glimpse(df)
```

```
## Rows: 28,562
## Columns: 21
## $ INCIDENT_KEY            <dbl> 244608249, 247542571, 84967535, 202853370, 270~
## $ OCCUR_DATE              <chr> "05/05/2022", "07/04/2022", "05/27/2012", "09/~
## $ OCCUR_TIME              <time> 00:10:00, 22:20:00, 19:35:00, 21:00:00, 21:00~
## $ BORO                    <chr> "MANHATTAN", "BRONX", "QUEENS", "BRONX", "BROO~
## $ LOC_OF_OCCUR_DESC       <chr> "INSIDE", "OUTSIDE", NA, NA, NA, NA, NA, NA, N~
## $ PRECINCT                <dbl> 14, 48, 103, 42, 83, 23, 113, 77, 48, 49, 73, ~
## $ JURISDICTION_CODE       <dbl> 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ LOC_CLASSFCTN_DESC      <chr> "COMMERCIAL", "STREET", NA, NA, NA, NA, NA, NA~
## $ LOCATION_DESC           <chr> "VIDEO STORE", "(null)", NA, NA, NA, "MULTI DW~
## $ STATISTICAL_MURDER_FLAG <lgl> TRUE, TRUE, FALSE, FALSE, FALSE, FALSE, TRUE, ~
## $ PERP_AGE_GROUP          <chr> "25-44", "(null)", NA, "25-44", "25-44", NA, N~
## $ PERP_SEX                <chr> "M", "(null)", NA, "M", "M", NA, NA, NA, NA, "~
## $ PERP_RACE               <chr> "BLACK", "(null)", NA, "UNKNOWN", "BLACK", NA,~
## $ VIC_AGE_GROUP           <chr> "25-44", "18-24", "18-24", "25-44", "25-44", "~
## $ VIC_SEX                 <chr> "M", "M", "M", "M", "M", "M", "M", "M", "M", "~
## $ VIC_RACE                <chr> "BLACK", "BLACK", "BLACK", "BLACK", "BLACK", "~
## $ X_COORD_CD              <dbl> 986050, 1016802, 1048632, 1014493, 1009149, 99~
## $ Y_COORD_CD              <dbl> 214231.0, 250581.0, 198262.0, 242565.0, 190104~
## $ Latitude                <dbl> 40.75469, 40.85440, 40.71063, 40.83242, 40.688~
## $ Longitude               <dbl> -73.99350, -73.88233, -73.76777, -73.89071, -7~
## $ Lon_Lat                 <chr> "POINT (-73.9935 40.754692)", "POINT (-73.8823~
```

```r
df <- df[, c("INCIDENT_KEY", "OCCUR_DATE")]
df$OCCUR_DATE = mdy(df$OCCUR_DATE)
```

```r
df$OCCUR_YEAR = year(df$OCCUR_DATE)
df$OCCUR_MONTH = month(df$OCCUR_DATE)
sum(is.na(df))
```

```
## [1] 0
```

```r
glimpse(df)
```

```
## Rows: 28,562
## Columns: 4
## $ INCIDENT_KEY <dbl> 244608249, 247542571, 84967535, 202853370, 27078636, 2303~
## $ OCCUR_DATE   <date> 2022-05-05, 2022-07-04, 2012-05-27, 2019-09-24, 2007-02-~
## $ OCCUR_YEAR   <dbl> 2022, 2022, 2012, 2019, 2007, 2021, 2021, 2021, 2021, 202~
## $ OCCUR_MONTH  <dbl> 5, 7, 5, 9, 2, 7, 6, 7, 5, 12, 9, 12, 4, 11, 7, 5, 8, 6, ~
```

```r
summary(select(df, -INCIDENT_KEY))
```

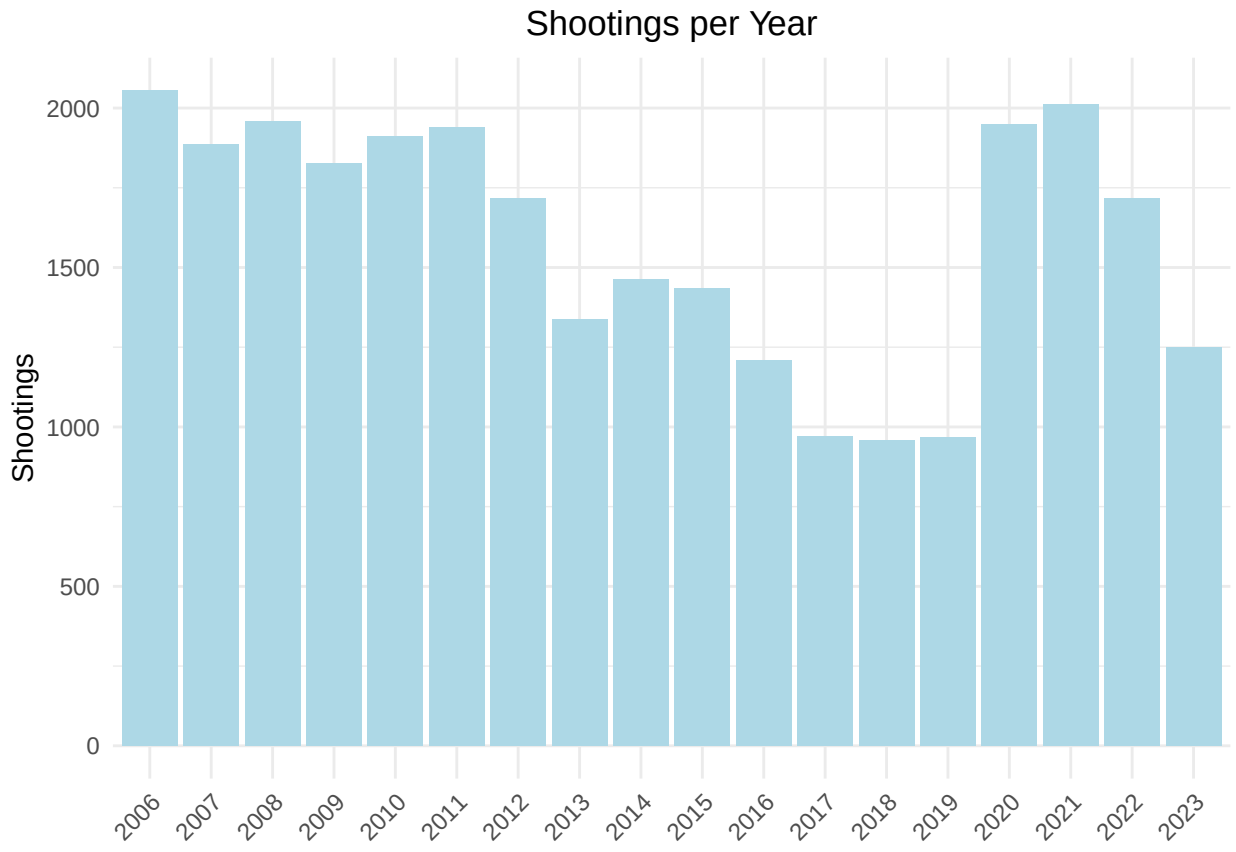
```
##    OCCUR_DATE           OCCUR_YEAR    OCCUR_MONTH
##  Min.   :2006-01-01   Min.   :2006   Min.   : 1.000
##  1st Qu.:2009-09-04   1st Qu.:2009   1st Qu.: 5.000
##  Median :2013-09-20   Median :2013   Median : 7.000
##  Mean   :2014-06-07   Mean   :2014   Mean   : 6.805
##  3rd Qu.:2019-09-29   3rd Qu.:2019   3rd Qu.: 9.000
##  Max.   :2023-12-29   Max.   :2023   Max.   :12.000
```

Step 3: Add Visualizations and Analysis

Add at least two different visualizations & some analysis to your Rmd. Does this raise additional questions that you should investigate?

```r
year_counts = as.data.frame(table(df$OCCUR_YEAR))
colnames(year_counts) <- c("Year", "Shootings")


ggplot(year_counts, aes(x = Year, y = Shootings)) +
  geom_bar(stat = "identity", fill = "lightblue") +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    plot.title = element_text(hjust = 0.5)
  ) +
  labs(
    title = "Shootings per Year",
    x = NULL
  )
```



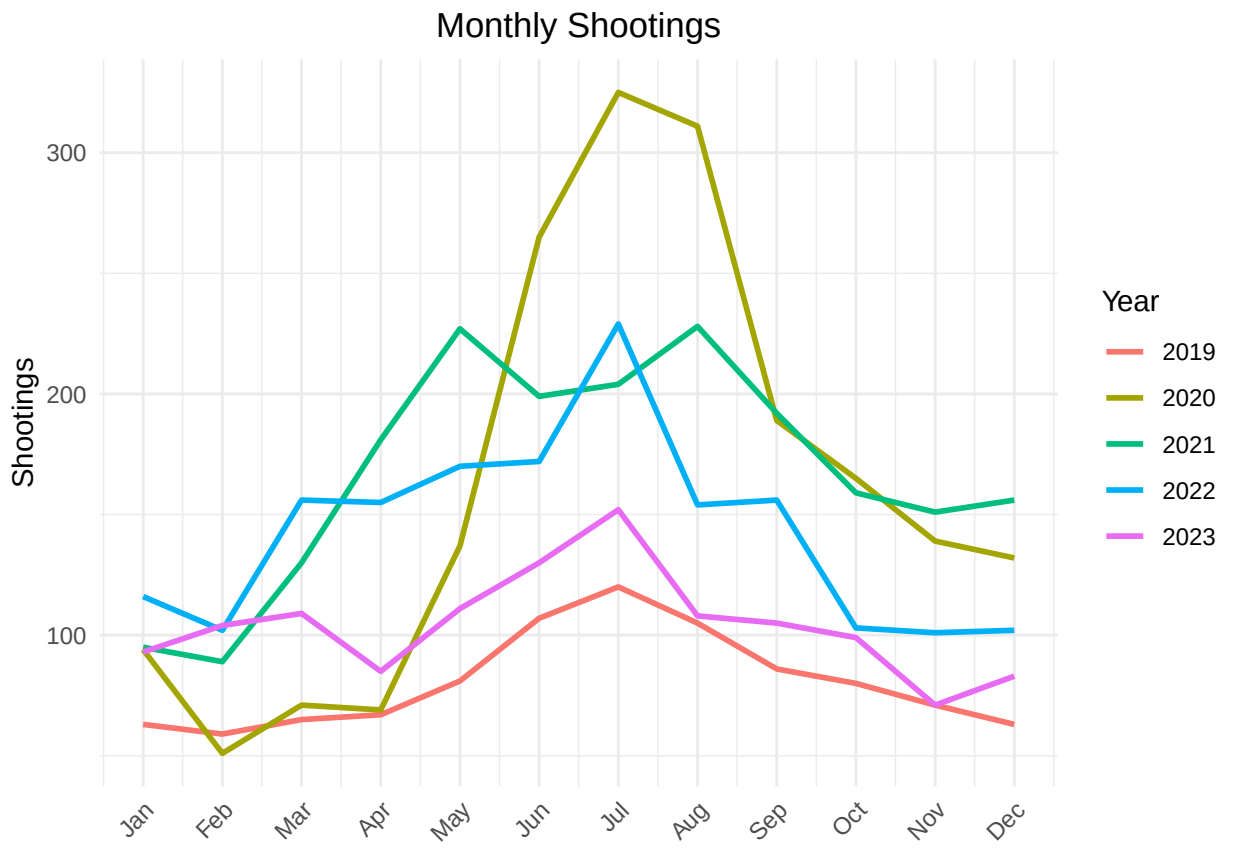

```r
df_2019_to_2023 <- df %>%
  filter(OCCUR_YEAR >= 2019 & OCCUR_YEAR <= 2023)

monthly_shootings <- df_2019_to_2023 %>%
  group_by(OCCUR_YEAR, OCCUR_MONTH) %>%
  summarise(incident_count = n(), .groups = "drop")

ggplot(monthly_shootings, aes(x = OCCUR_MONTH, y = incident_count, color = as.factor(OCCUR_YEAR), group
  geom_line(size = 1) +
  scale_x_continuous(breaks = 1:12, labels = month.abb) +  # Show months as abbreviations
  theme_minimal() +
  labs(
    title = "Monthly Shootings",
    x = NULL,
    y = "Shootings",
    color = "Year"
  ) +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    plot.title = element_text(hjust = 0.5)
  )
```

```
## 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.
```



Additional questions include crime per capita. I assume a lot of people in NYC left in 2020 to areas with more space.

Step 4: Add Bias Identification

Write the conclusion to your project report and include any possible sources of bias. Be sure to identify what your personal bias might be and how you have mitigated that.