NYPD Shooting Data

2024-03-20

Data and Background

This document analyzes historic shooting data from the NYPD between January 2006 and December 2022. The data is obtained from https://catalog.data.gov/dataset and comes directly from the City of New York data repository. It includes data like time and date, location, and victim demographics.

This document will attempt to identify shooting incident trends over time and create a seasonal forecast for predicting future incidents.

In reviewing this data two main visualizations were produced:

- A **time of day** analysis for shooting incidents
- A **time of year** analysis of shooting incidents

This document also highlights a seasonal predictive model on the time of year analysis.

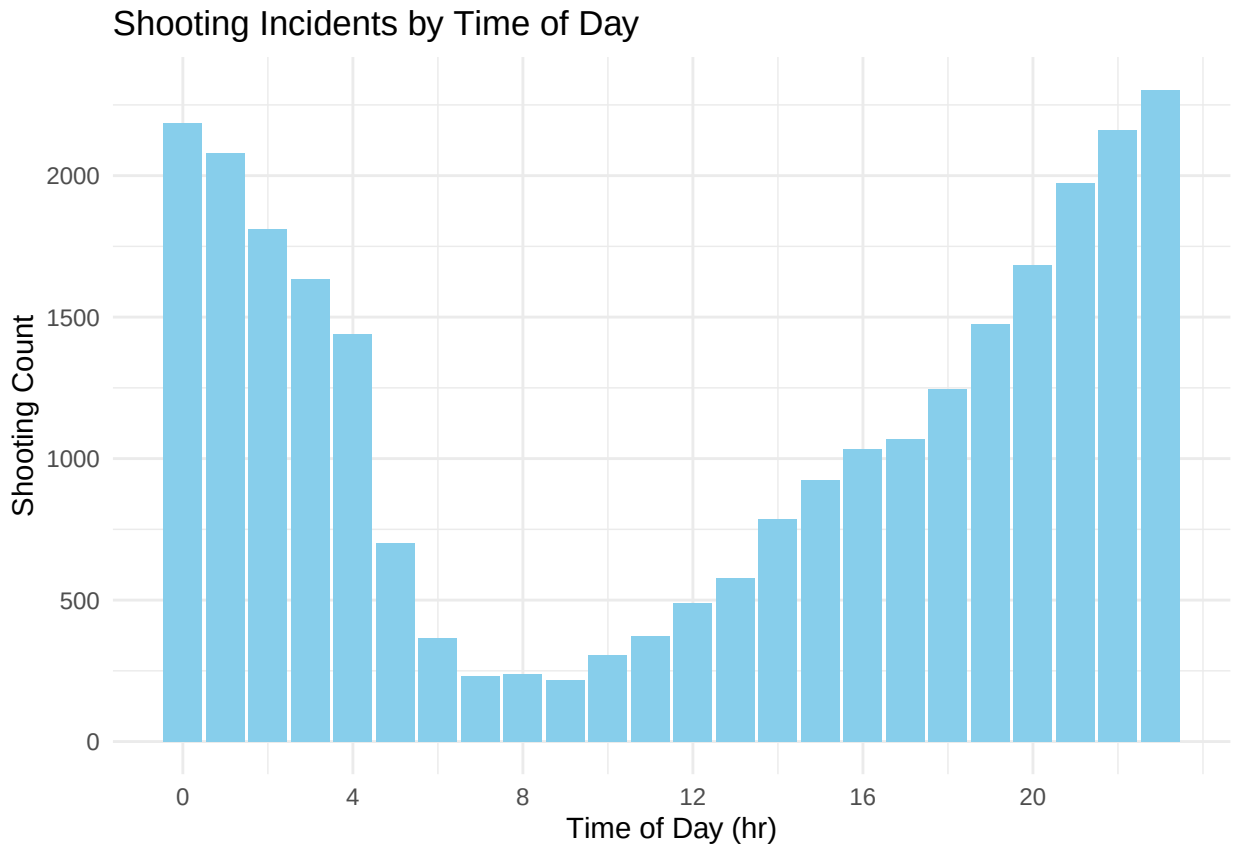
```
#Read data
nypd_shoot_incidence <- read_csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=
```

```
#Clean data to only have Date and Time
nypd_shoot_incidence <- nypd_shoot_incidence %>%
  select(OCCUR_DATE,OCCUR_TIME) %>% #Only select Date and Time
  mutate(OCCUR_DATE = mdy(OCCUR_DATE)) %>% #Modify date column as date
  arrange(OCCUR_DATE) %>% #Order column by date ascending
  rename(Date = 'OCCUR_DATE', Time = 'OCCUR_TIME') #Rename columns
```

Shooting Incidents by Time of Day

```
#Plot Time of Day vs Shooting Incidents
nypd_shoot_incidence_df <- as.data.frame(nypd_shoot_incidence) #save as dataframe

nypd_shoot_incidence_df <- nypd_shoot_incidence_df %>% #add hour column and extract from time
  mutate(Hour = hour(Time))

shooting_counts <- nypd_shoot_incidence_df %>% #count and summarize the shootings per hour
  group_by(Hour) %>%
  summarize(shooting_count = n())

#plot as bar chart
ggplot(shooting_counts, aes(x=Hour, y=shooting_count)) +
  geom_bar(stat = "identity", fill = "skyblue") +
```

```
  labs(title= "Shooting Incidents by Time of Day", x = "Time of Day (hr)", y= "Shooting Count")+
  scale_x_continuous(breaks = seq(0, 23, by = 4)) +
  theme_minimal()
```

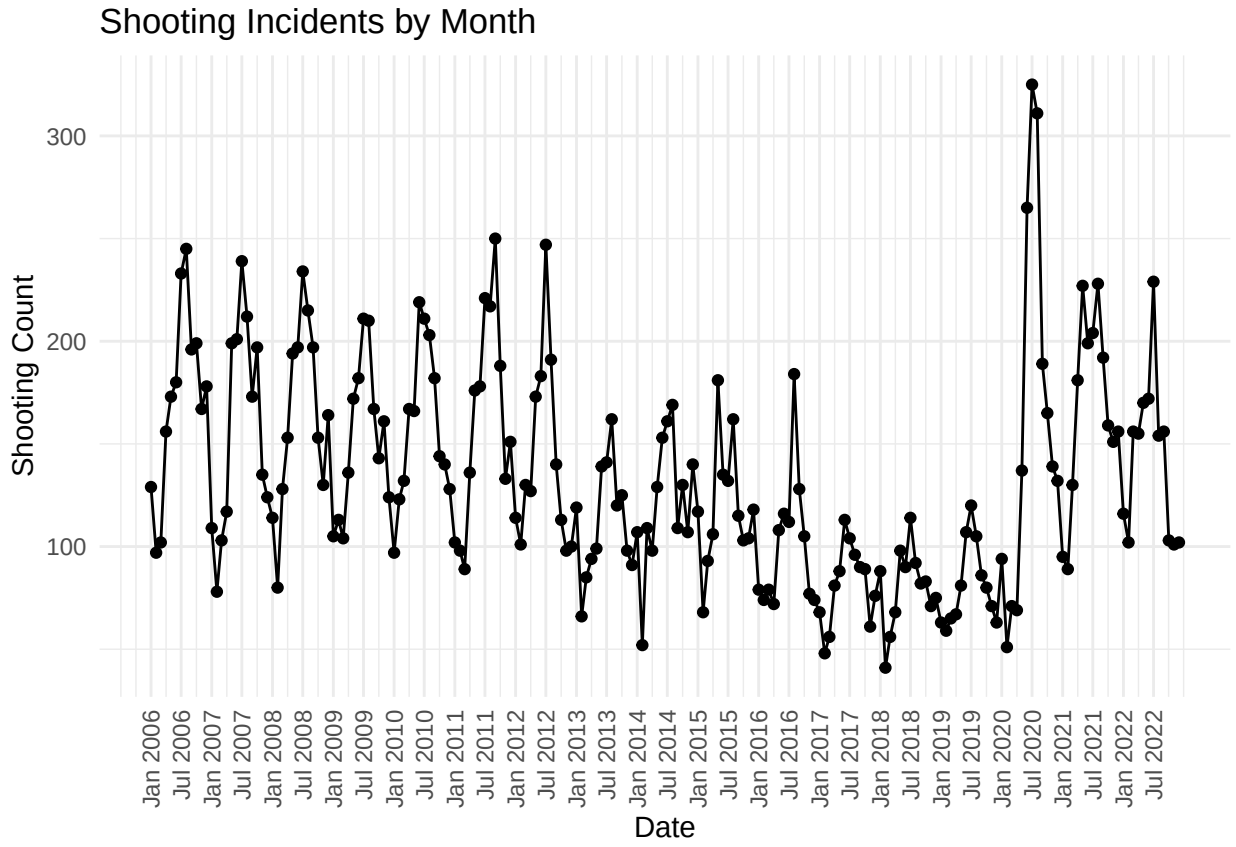


Analysis: This chart shows that the vast majority of shootings occur during the nigh time hours with the peak time being between 11pm and 12am across all days and locations. The time with the least amount of shooting incidents is between 9am and 10am.

Shooting Incidents by Time of Year

```
#Plot Time of Year vs Shooting Incidents
nypd_shoot_incidence_df <- as.data.frame(nypd_shoot_incidence) #save as dataframe

nypd_shoot_incidence_df <- nypd_shoot_incidence_df %>% #Add month and year columns
  mutate(Month = format(Date,"%m"),
         Year = format(Date,"%Y"))

shooting_counts <- nypd_shoot_incidence_df %>% #Count shooting incidents by month
  group_by(Year,Month) %>%
  summarize(shooting_count = n())

shooting_counts$Date <- as.Date(paste(shooting_counts$Year,shooting_counts$Month, "01", sep = "-")) #cr

min_date <- min(shooting_counts$Date)
```

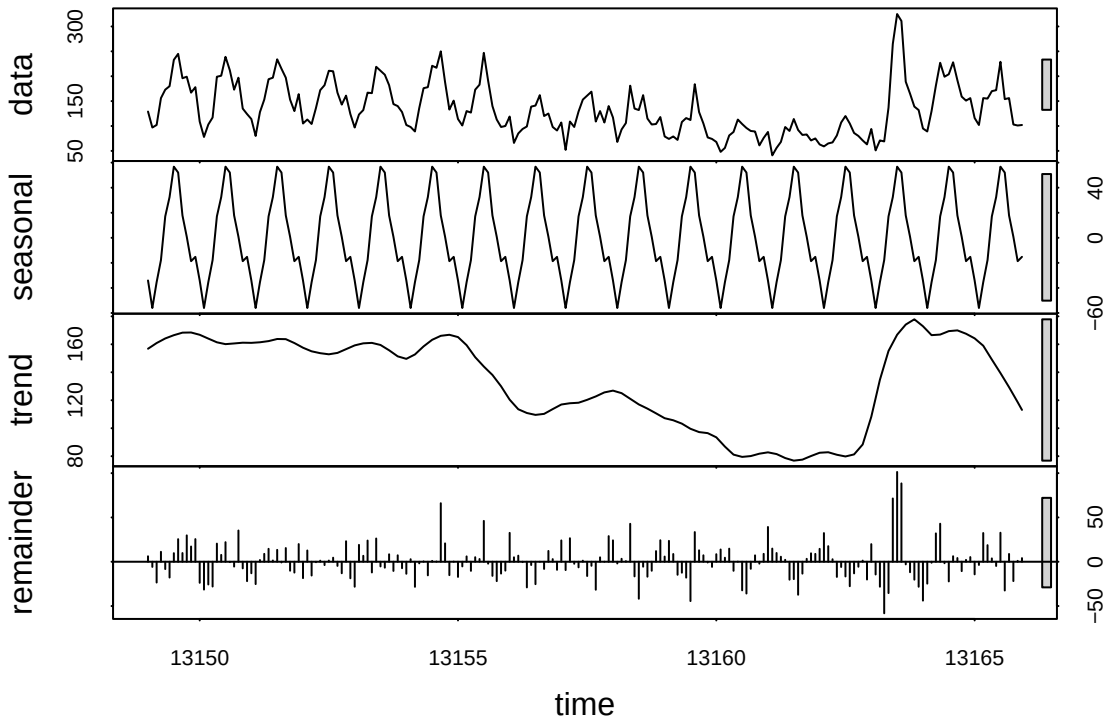
```r
max_date <- max(shooting_counts$Date)

#plot
ggplot(shooting_counts, aes(x=Date, y=shooting_count)) +
  geom_line()+
  geom_point()+
  scale_x_date(labels = scales::date_format("%b %Y"), breaks = seq(min_date, max_date, by ="6 months"))+
  labs(title= "Shooting Incidents by Month", x = "Date", y= "Shooting Count")+
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Analysis: This chart shows that shootings steadily decreased between January 2006 and January 2020. After January 2020, shootings rose dramatically, coinciding with the COVID-19 Pandemic. Seasonality is also very clearly shown in this plot and will be explored in the subsequent modeling.

Shooting Incidents Seasonality and Forecasting

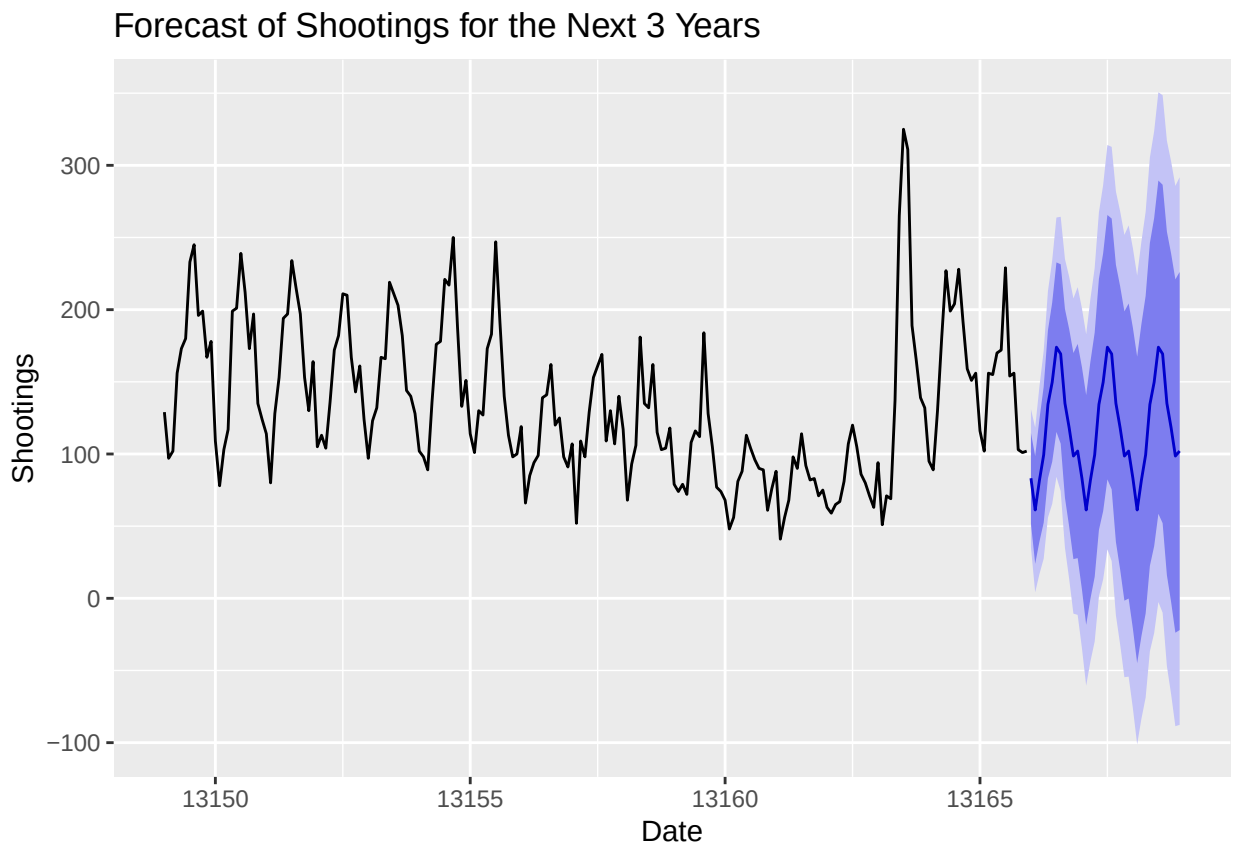
```r
#Analyze seasonality
ts_data <- ts(shooting_counts$shooting_count, start = min(shooting_counts$Date), frequency=12)
decomp <- stl(ts_data, s.window = "periodic")
plot(decomp)
```



Analysis: This output shows a very strong seasonal trend where shooting incidents peak during the summer months each year and are at a minimum during the winter.

Utilizing the seaonality data, we can forecast the predicted shooting incident counts for the next 3 years shown below:

```
#Plot 3 year forecast
forecast_values <- forecast(decomp, h = 36)
autoplot(forecast_values, xaxt = "n") +
  labs(title = "Forecast of Shootings for the Next 3 Years", x = "Date", y = "Shootings")
```



Conclusion

The time of day analysis shows that shooting incidents occur most frequently at night. The time of year analysis shows a clear seasonal trend that has shooting incidents peak during the summer months. Utilizing the seasonal trends, a forecast model can be produced which predicts the next 3 years of time including uncertainty bands.

Bias

There are several sources of bias in this data. The biggest source of bias in this NYPD shooting incident data is underreporting bias. There could be incidents that go unreported or are inaccurately reported. This could be due to the fact that there might be a reluctance to report due to fear of retaliation or even a lack of trust in law enforcement. Another source of bias is selection bias since only shooting incidents that had witnesses or a police presence were reported. Many incidents likely go unreported and are missing from the data set. Lastly, another source of bias is temporal bias. This bias may show changes in shooting incidents over time if reporting avenues, policies, or even community dynamics has shifted over the years. This could help explain artifacts that can be misattributed to the data as trends.

Session Info

```
## R version 4.3.1 (2023-06-16 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 11 x64 (build 22631)
```

```
##
## 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/New_York
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] forecast_8.22.0 lubridate_1.9.3 forcats_1.0.0   stringr_1.5.0
##  [5] dplyr_1.1.3     purrr_1.0.2     readr_2.1.4     tidyr_1.3.0
##  [9] tibble_3.2.1    ggplot2_3.4.4   tidyverse_2.0.0
##
## loaded via a namespace (and not attached):
##  [1] utf8_1.2.3        generics_0.1.3    stringi_1.7.12    lattice_0.21-8
##  [5] hms_1.1.3         digest_0.6.33     magrittr_2.0.3    evaluate_0.21
##  [9] grid_4.3.1        timechange_0.3.0  fastmap_1.1.1     nnet_7.3-19
## [13] fansi_1.0.4       scales_1.2.1      cli_3.6.1         crayon_1.5.2
## [17] rlang_1.1.1       bit64_4.0.5       munsell_0.5.0     withr_2.5.0
## [21] yaml_2.3.7        tools_4.3.1       parallel_4.3.1    tzdb_0.4.0
## [25] colorspace_2.1-0  curl_5.0.2        vctrs_0.6.3       R6_2.5.1
## [29] zoo_1.8-12        lifecycle_1.0.3   tseries_0.10-55   bit_4.0.5
## [33] vroom_1.6.3       urca_1.3-3        pkgconfig_2.0.3   pillar_1.9.0
## [37] gtable_0.3.4      quantmod_0.4.26   glue_1.6.2        Rcpp_1.0.11
## [41] xfun_0.40         lmtest_0.9-40     tidyselect_1.2.0  rstudioapi_0.15.0
## [45] knitr_1.44        farver_2.1.1      nlme_3.1-162      htmltools_0.5.6
## [49] labeling_0.4.3    xts_0.13.2        rmarkdown_2.24    timeDate_4032.109
## [53] fracdiff_1.5-3    compiler_4.3.1    quadprog_1.5-8    TTR_0.24.4
```