NY Shootings Report

Sid Sethi

2025-04-24

Analysis Background

This report analyzes historic NYPD Shooting Incident data (Historic). Each record in the dataset represents a shooting event that took place in NY, along with location, time, suspect age, victim age, suspect race and victim race. The report also indicates if the shooting end in a murder. The source of this data is NYC Shooting Dataset

Step 1: Import the data

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

Step 2: Read the data

```
shootings <- read_csv(url_in)
```

Step 3: Convert to date format

```
shootings <- shootings %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE),
         YEAR = year(OCCUR_DATE))
```

Step 4: Factoring and Removing Long Lat

Making all variables "unknown" if no value is there

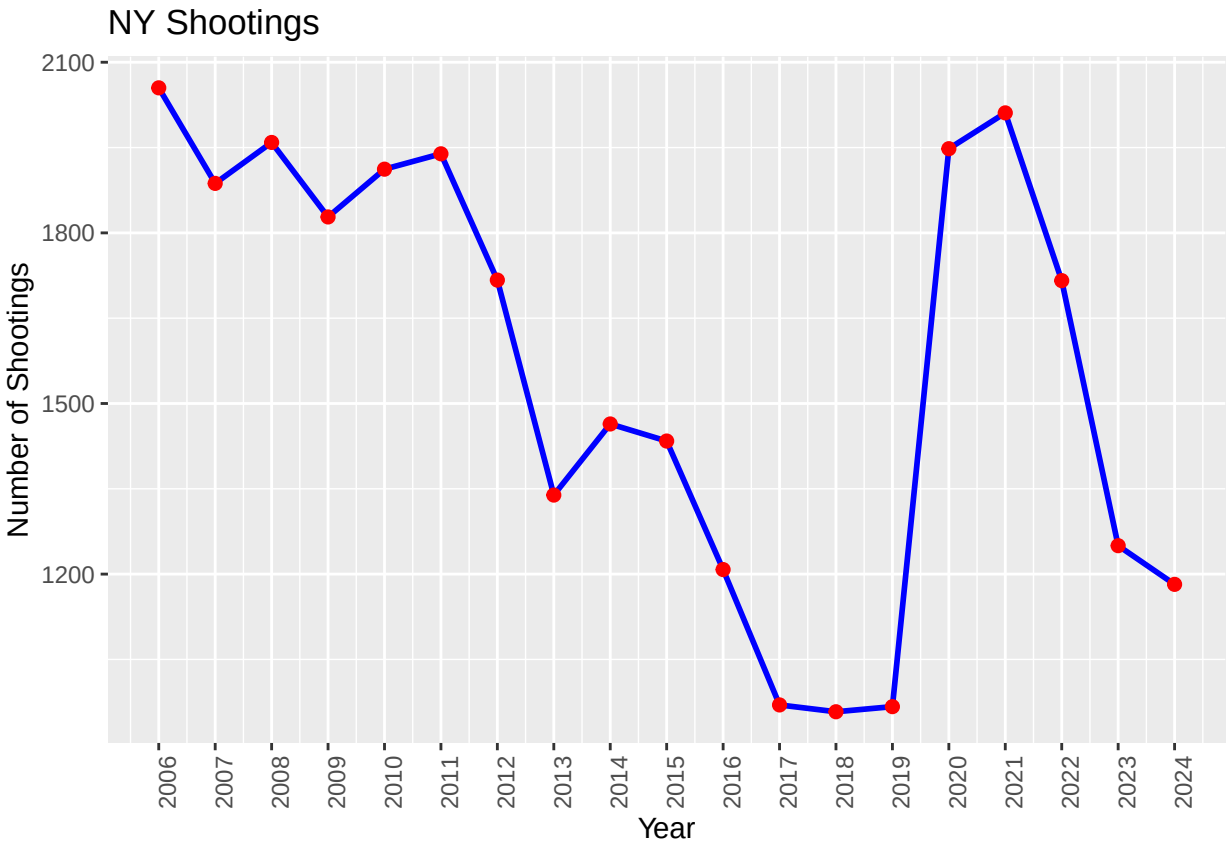
```
valid_ages <- c("<18", "18-24", "25-44", "45-64", "65+")

shootings <- shootings %>% mutate(BORO = fct_explicit_na(factor(BORO), na_level = "Unknown"),
LOC_OF_OCCUR_DESC = fct_explicit_na(factor(LOC_OF_OCCUR_DESC), na_level = "Unknown"),
PRECINCT = fct_explicit_na(factor(PRECINCT), na_level = "Unknown"),
JURISDICTION_CODE = fct_explicit_na(factor(JURISDICTION_CODE), na_level = "Unknown"),
LOC_CLASSFCTN_DESC = fct_explicit_na(factor(LOC_CLASSFCTN_DESC), na_level = "Unknown"),
LOCATION_DESC = fct_explicit_na(factor(LOCATION_DESC), na_level = "Unknown"),
STATISTICAL_MURDER_FLAG = fct_explicit_na(factor(STATISTICAL_MURDER_FLAG), na_level = "Unknown"),
```

```
PERP_SEX = fct_explicit_na(factor(PERP_SEX), na_level = "Unknown"),
PERP_RACE = fct_explicit_na(factor(PERP_RACE), na_level = "Unknown"),
VIC_AGE_GROUP = fct_explicit_na(factor(VIC_AGE_GROUP), na_level = "Unknown"),
VIC_SEX = fct_explicit_na(factor(VIC_SEX), na_level = "Unknown"),
VIC_RACE = fct_explicit_na(factor(VIC_RACE), na_level = "Unknown"),
PERP_AGE_GROUP = case_when(
  PERP_AGE_GROUP %in% valid_ages ~PERP_AGE_GROUP,
  TRUE ~ "Unknown"
), VIC_AGE_GROUP = case_when(
  VIC_AGE_GROUP %in% valid_ages ~PERP_AGE_GROUP,
  TRUE ~ "Unknown"
))


#removing long lat
shootings <- shootings %>% select(-c(Latitude, Longitude, Lon_Lat))
```

Analysis

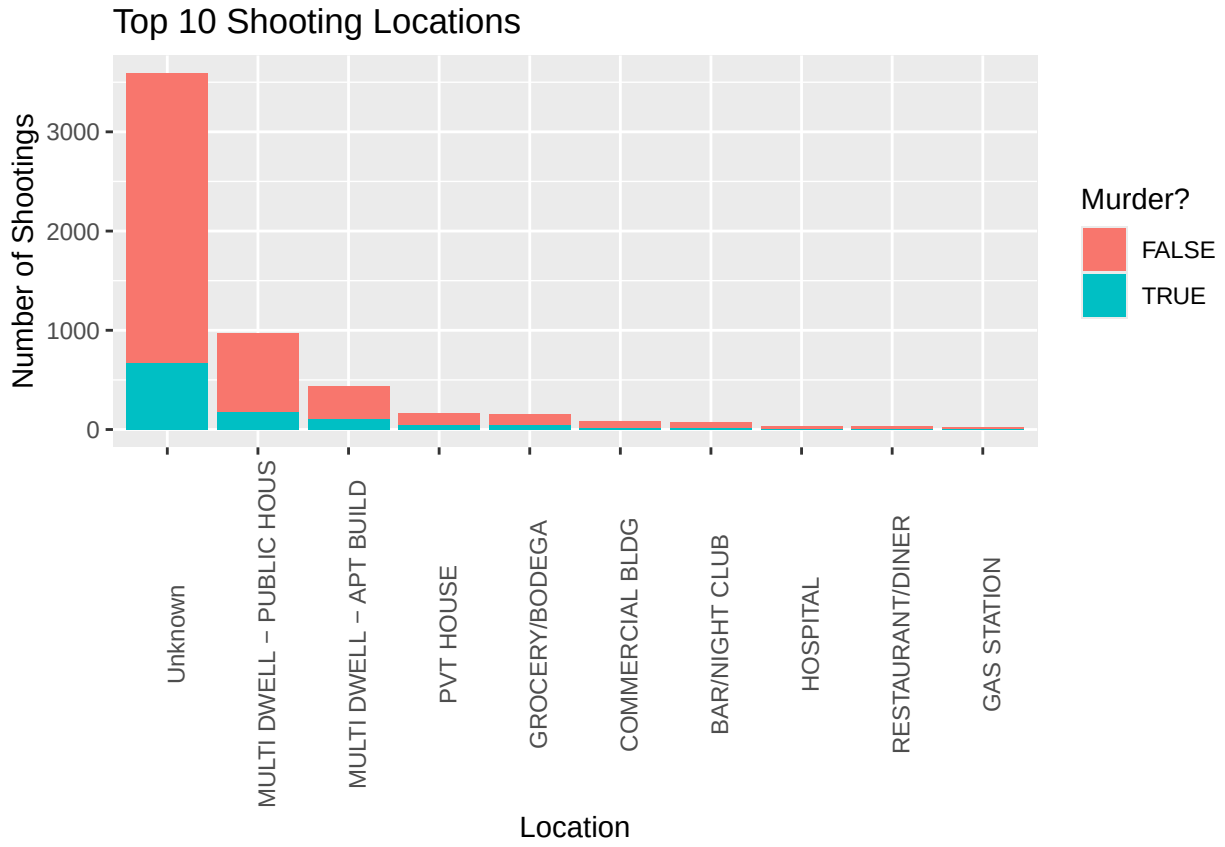
The analysis I want to perform using this dataset is that I want to look at when most shootings happened. I also want to see if there are particular locations where shootings are more prevalent. In those areas, I want to look at during what times shootings occur most and by who. Also, I want to look at when shootings are the most fatal.

Step 1: When did shootings take place most frequently?



Conclusion: Using this graph, I am able to conclude that there was a significant spike in shootings between 2020 and 2022 (likely tied to the pandemic).

Step 2: Where did shootings take place most frequently in 2020-2022 given there was a global pandemic.



It seems like there was a spike in crime in between 2020-2022. Most of it happened in housing only (likely due to covid). Multi Dwell - Public Housing, Apt build, PVT House.

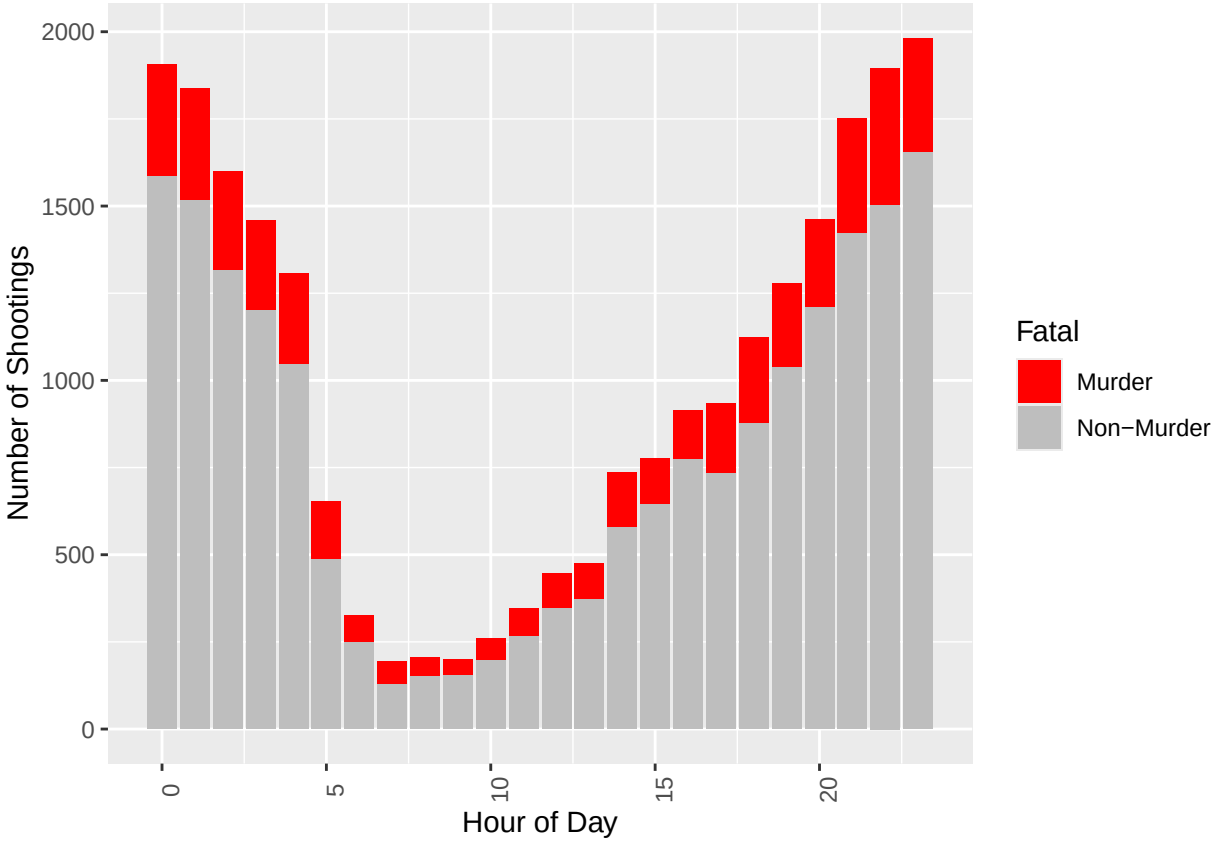
Step 3: At What times did shootings take place?

```
shootings_non_cov <- shootings %>%
  mutate(hr = hour(OCCUR_TIME),
         yr = year(OCCUR_DATE),
         murder = ifelse(STATISTICAL_MURDER_FLAG == TRUE, "Murder", "Non-Murder")) %>%
  filter(!yr %in% c(2020, 2021, 2022)) %>%

  group_by(hr, murder) %>% summarise(Count = n())
```

```
## `summarise()` has grouped output by 'hr'. You can override using the `.groups`
## argument.
```

```
ggplot(shootings_non_cov, aes(x= hr, y= Count, fill = murder)) +
geom_col() +
  scale_fill_manual(values = c("Murder" = "red", "Non-Murder" = "grey")) +
```

```
  labs(
    x = "Hour of Day",
    y = "Number of Shootings",
    fill = "Fatal",
    Title = "Shooting Hrs (Non Covid Years)"
  ) +
  theme(axis.text.x = element_text(angle = 90))
```



We are able to see that during the Non Covid years (not including 2022-2022), shootings increase after 2 PM. Also, it is important to note that from 7PM onwards shootings generally increase in fatality. Now, I want to see if COVID had an impact on this? Did shootings start earlier since a large majority was unemployed?

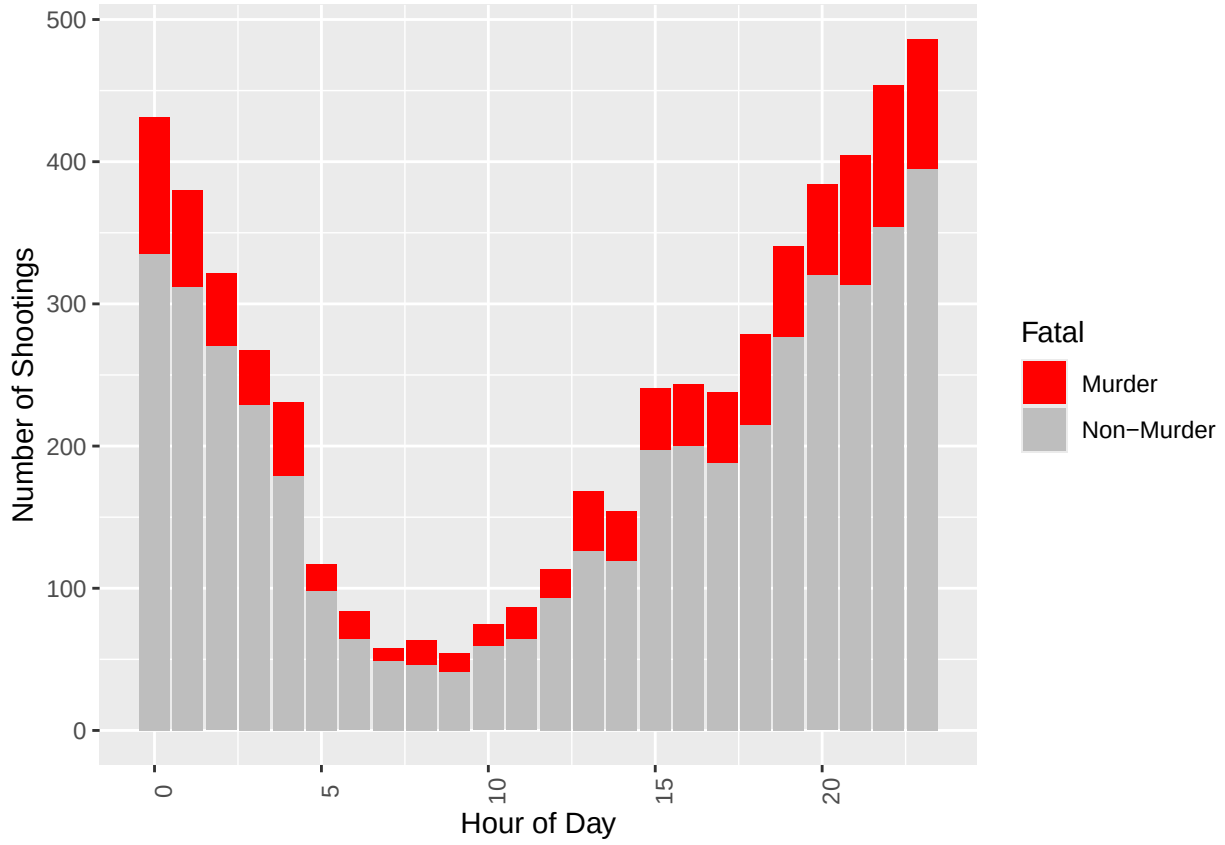
Step 4: Did time of shooting change during COVID?

```
shootings_covid_1 <- shootings %>%
  mutate(hr = hour(OCCUR_TIME),
         yr = year(OCCUR_DATE),
         murder = ifelse(STATISTICAL_MURDER_FLAG == TRUE, "Murder", "Non-Murder")) %>%
  filter(yr %in% c(2020, 2021, 2022)) %>%

  group_by(hr, murder) %>% summarise(Count = n())
```

```
## `summarise()` has grouped output by 'hr'. You can override using the `.groups`
```

```
## argument.
```

```
ggplot(shootings_covid_1, aes(x= hr, y= Count, fill = murder)) +
geom_col() +
  scale_fill_manual(values = c("Murder" = "red", "Non-Murder" = "grey")) +
  labs(
    x = "Hour of Day",
    y = "Number of Shootings",
    fill = "Fatal"
  ) +
  theme(axis.text.x = element_text(angle = 90))
```




```
shootings_covid_2 = shootings %>% group_by() %>%
  summarise(
    total_shootings = n(),
    fatal = sum(STATISTICAL_MURDER_FLAG == TRUE)) %>%
  mutate(murder_rate = fatal/total_shootings

  )
```

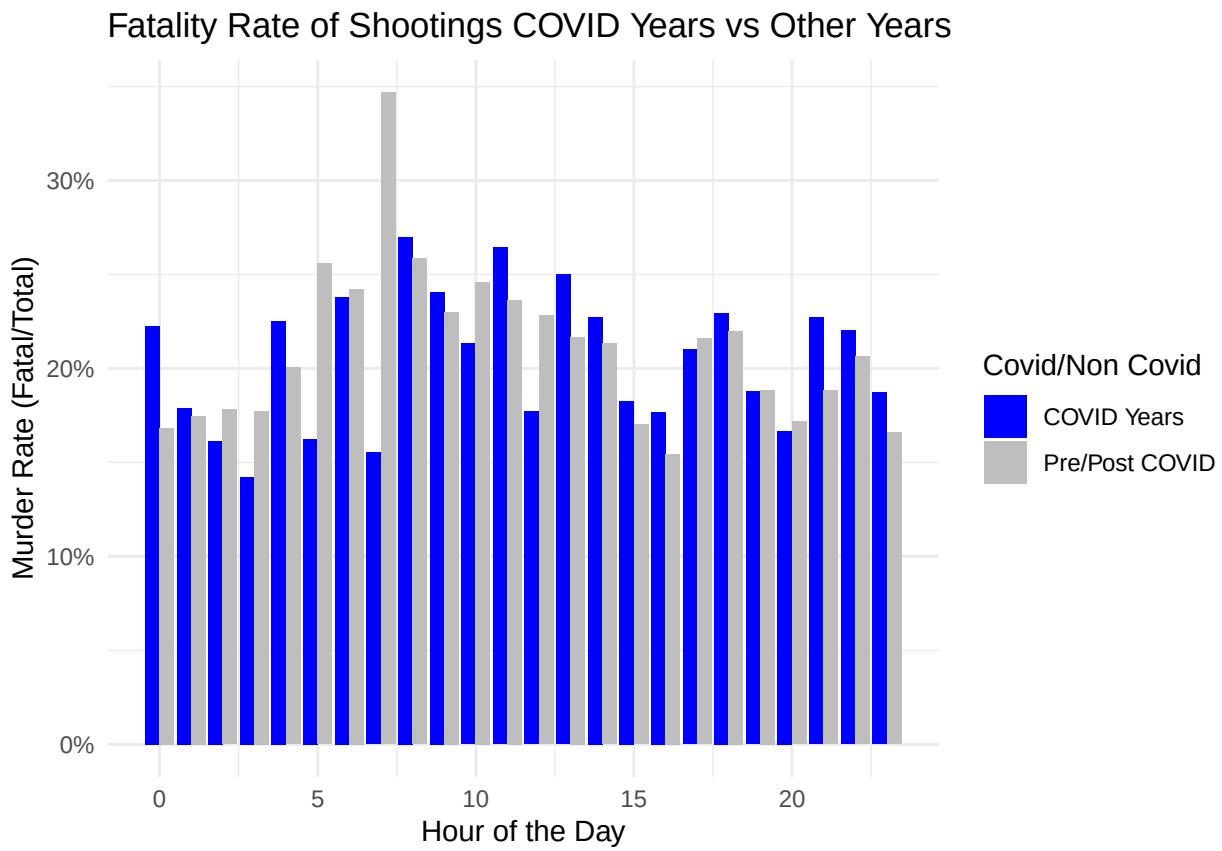
Looking at the COVID years, we see that in those 3 years, shootings followed the same trajectory; rapidly increasing after 2PM.

Step 5: Checking if In Covid Shootings Were More Fatal?

```
shootings_period <- shootings %>%
  mutate(Year = year(OCCUR_DATE), Hour = hour(hms(OCCUR_TIME)),
    MurderFlag = ifelse(STATISTICAL_MURDER_FLAG == TRUE, "Murder", "Non-Murder"),
    Period = ifelse(Year %in% c(2020, 2021, 2022), "COVID Years", "Pre/Post COVID")
  )

shootings_summary_period <- shootings_period %>%
  group_by(Hour, Period) %>%
  summarise( Total_Shootings = n(), Fatal_Shootings = sum(STATISTICAL_MURDER_FLAG == TRUE, na.rm = TRUE)

  ) %>%
  mutate(
    Murder_Rate = Fatal_Shootings / Total_Shootings
  )
```

```
## `summarise()` has grouped output by 'Hour'. You can override using the
## `.groups` argument.
```

```
ggplot(shootings_summary_period, aes(x = Hour, y = Murder_Rate, fill = Period)) +
  geom_col(position = "dodge") +
  scale_fill_manual(values = c("COVID Years" = "blue", "Pre/Post COVID" = "grey")) +
  labs(
    title = "Fatality Rate of Shootings COVID Years vs Other Years",
    x = "Hour of the Day",
    y = "Murder Rate (Fatal/Total)",
    fill = "Covid/Non Covid"
  ) +
  scale_y_continuous(labels = scales::percent_format(accuracy = 1)) +
  theme_minimal()
```



With this analysis, we can conclude that for the majoity of the hours, COVID years are more fatal due to lack of resources available and the fact that there were more shootings.

Model

```
shootings_model_noncov <- shootings %>%
  mutate(murder = ifelse(STATISTICAL_MURDER_FLAG == TRUE, 1, 0), Hour = hour(hms(OCCUR_TIME))) %>% filte
  filter(PERP_RACE != "Unknown") %>% filter(!YEAR %in% c(2020,2021,2022))

model_age <- glm(murder ~ PERP_AGE_GROUP, data = shootings_model_noncov, family = "binomial")

summary(model_age)
```

```
##
## Call:
## glm(formula = murder ~ PERP_AGE_GROUP, family = "binomial", data = shootings_model_noncov)
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -1.49356    0.06661 -22.422  < 2e-16 ***
## PERP_AGE_GROUP18-24  0.15918    0.07427   2.143 0.032083 *
## PERP_AGE_GROUP25-44  0.48444    0.07396   6.550 5.77e-11 ***
## PERP_AGE_GROUP45-64  0.84144    0.10991   7.656 1.92e-14 ***
## PERP_AGE_GROUP65+    1.00108    0.27869   3.592 0.000328 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 13859  on 12702  degrees of freedom
## Residual deviance: 13745  on 12698  degrees of freedom
## AIC: 13755
##
## Number of Fisher Scoring iterations: 4
```

This shows that as age increases of the perp. chance of fatality also increases. Now lets see if this changed during covid?

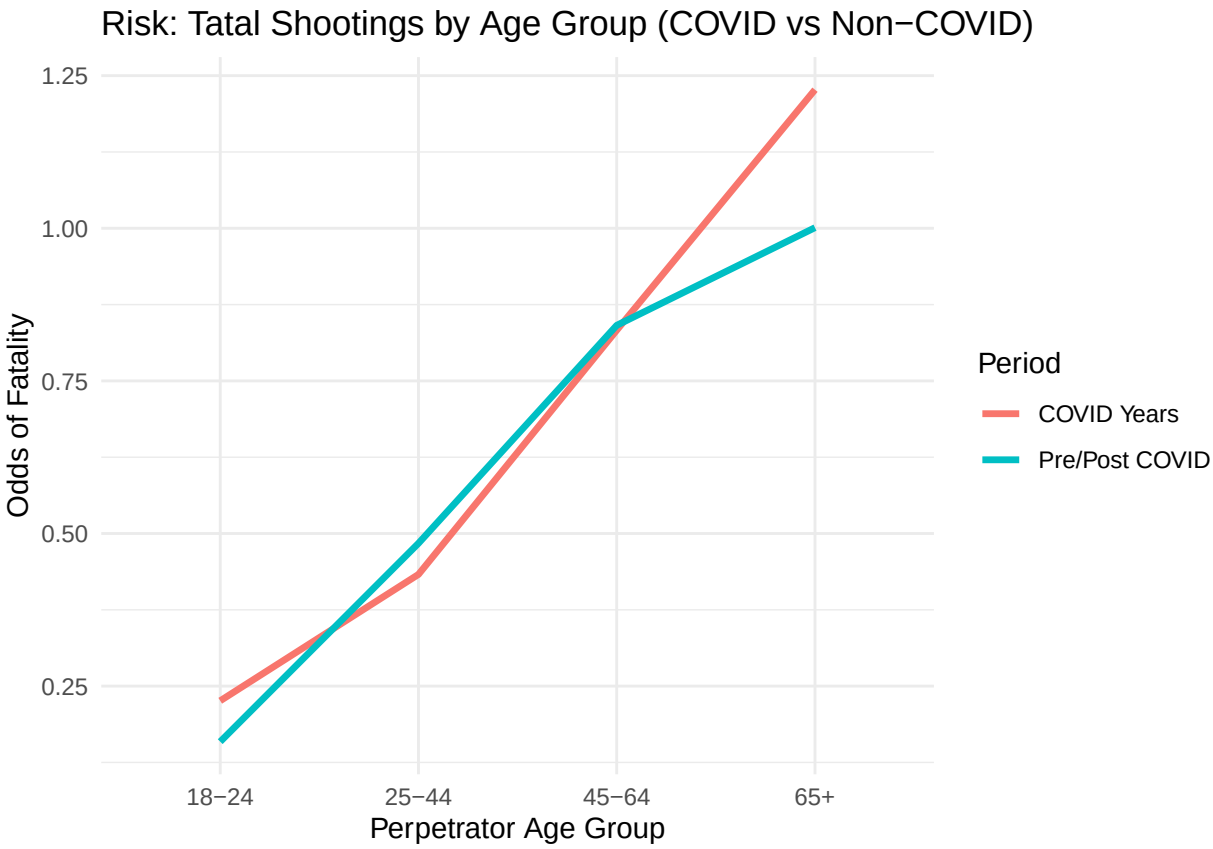
```
shootings_model_cov <- shootings %>%
  mutate(murder = ifelse(STATISTICAL_MURDER_FLAG == TRUE, 1, 0), Hour = hour(hms(OCCUR_TIME))) %>%
  filter(PERP_AGE_GROUP != "Unknown") %>% filter(PERP_RACE != "Unknown") %>%  filter(YEAR %in% c(2020,2

model_age <- glm(murder ~ PERP_AGE_GROUP, data = shootings_model_cov, family = "binomial")

summary(model_age)
```

```
##
## Call:
## glm(formula = murder ~ PERP_AGE_GROUP, family = "binomial", data = shootings_model_cov)
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -1.4500     0.1472  -9.853  < 2e-16 ***
## PERP_AGE_GROUP18-24   0.2262     0.1652   1.370  0.17081
## PERP_AGE_GROUP25-44   0.4326     0.1591   2.718  0.00656 **
## PERP_AGE_GROUP45-64   0.8332     0.2105   3.959 7.53e-05 ***
## PERP_AGE_GROUP65+     1.2269     0.6868   1.786  0.07403 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3283.9  on 2915  degrees of freedom
## Residual deviance: 3261.9  on 2911  degrees of freedom
## AIC: 3271.9
##
## Number of Fisher Scoring iterations: 4
```

```
risk <- tibble(
  AgeGroup = c("18-24", "25-44", "45-64", "65+"),
  COVID = c(0.226, 0.433, 0.833, 1.227),
  Non_COVID = c(0.159, 0.484, 0.841, 1.001)
)
```

```
ggplot(risk, aes(x = AgeGroup)) +
  geom_line(aes(y = COVID, color = "COVID Years", group = 1), size = 1.2) +
  geom_line(aes(y = Non_COVID, color = "Pre/Post COVID", group = 1), size = 1.2) +
```

```
  labs(
    title = "Risk: Tatal Shootings by Age Group (COVID vs Non-COVID)",
    x = "Perpetrator Age Group",
    y = "Odds of Fatality",
    color = "Period"
  ) +
  theme_minimal()
```



Whether COVID or not, Fatality odds were higher for older age groups. I had expected older groups to be weaker during COVID, but they had a higher success rate in COVID.

Conclusion

Overall, COVID was one of the deadliest times in NY, with shootings inclining significantly. We also noticed shootings increase in housing areas. Most of the shootings increases significantly after 3PM. The older generation was the most deadlier, even outside of COVID. In this whole analysis, Bias could be present to missing data points, only neighbourhoods with higher policing may have recorded data, COVID playing a factor and weaking immune systems, hence causing more fatalities.

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.