Analyzing NYPD Shooting Data - Seasonal Variation, Demographics and Locations - Plus Modeling

Packages used: tidyverse, lubridate (Please install the libraries tidyverse and lubridate if you are reproducing the code)

```
#install.packages("tidyverse")
#install.packages("lubridate")
```

```
library(tidyverse)
library(lubridate)
```

NYPD Shooting Incident Data

In this report, I'm going to share my analysis on the NYPD shooting incidents data obtained from data.gov. The dataset is covering more than 23,000 incidents happening in NYC from 2006 to 2020.

You may wonder, why do we want to look at data of the shooting incidents? The answer is simple. By analyzing the data in different dimensions, such as the date and time of the incidents, the details of the perpetrators and victims, or the location of the incidents, we can think of ways to prevent shooting incidents strategically.

Focus of This Analysis

We will look at a few dimensions in particular of the shooting incidents data.

- The **seasonal variation** in the number of shooting incidents across the year. At what time of the year are shooting incidents more common? Is there a pattern we can observe, like a high season and a low season?
- **Demographics** of the perpetrators and victims. Who are they?

We will look at these two dimensions in conjunction to see if we can get some useful findings.

Besides, we will also look at:

- **Time** of the shootings
- **Location** of the shootings. In what kind of premises did they take place?

Is there any correlation between the location and the time?

Lastly, we will **model** the data to see if there is a correlation between the numbers of total cases and cases committed by perpetrators of age 18-24.

- **Modeling**: Relationship between total cases and cases committed by individuals between 18-24yo

Importing and Cleaning the Data

We will first **import the data** from NYC's data repository. This data set covers the date and time of occurrence, details of the involved perpetrators and victims, as well as the location of the incidents (including boroughs, precincts, and coordinates).

```
nypd_data <- read_csv('https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv')
```

Here is a summary of the data.

```
summary(nypd_data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:23585       Length:23585       Length:23585
##  1st Qu.: 55322804   Class :character   Class1:hms         Class :character
##  Median : 83435362   Mode  :character   Class2:difftime    Mode  :character
##  Mean   :102280741                      Mode  :numeric
##  3rd Qu.:150911774
##  Max.   :230611229
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.000     Length:23585       Mode :logical
##  1st Qu.: 44.00   1st Qu.:0.000     Class :character   FALSE:19085
##  Median : 69.00   Median :0.000     Mode  :character   TRUE :4500
##  Mean   : 66.21   Mean   :0.333
##  3rd Qu.: 81.00   3rd Qu.:0.000
##  Max.   :123.00   Max.   :2.000
##                   NA's   :2
##  PERP_AGE_GROUP       PERP_SEX          PERP_RACE         VIC_AGE_GROUP
##  Length:23585       Length:23585       Length:23585       Length:23585
##  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:23585       Length:23585       Min.   : 914928   Min.   :125757
##  Class :character   Class :character   1st Qu.: 999925   1st Qu.:182539
##  Mode  :character   Mode  :character   Median :1007654   Median :193470
##                                        Mean   :1009379   Mean   :207300
##                                        3rd Qu.:1016782   3rd Qu.:239163
##                                        Max.   :1066815   Max.   :271128
##
##     Latitude       Longitude        Lon_Lat
##  Min.   :40.51   Min.   :-74.25   Length:23585
##  1st Qu.:40.67   1st Qu.:-73.94   Class :character
##  Median :40.70   Median :-73.92   Mode  :character
##  Mean   :40.74   Mean   :-73.91
##  3rd Qu.:40.82   3rd Qu.:-73.88
##  Max.   :40.91   Max.   :-73.70
##
```

We will now **clean up** the dataset.

```
nypd_data_cleaned <- nypd_data %>%
  select(-c(X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat, STATISTICAL_MURDER_FLAG)) %>%
  mutate(OCCUR_DATE=as.Date(OCCUR_DATE, format = "%m/%d/%Y")) %>%
  mutate(hour=hour(OCCUR_TIME)) %>%
  mutate(minute=minute(OCCUR_TIME))
```

We have **removed some of the unused columns** such as latitude and longtitude in our analysis. We have also transformed the date column into date format.

We are also **transforming columns** such as the boroughs, and the age group, sex and race of perpetrators and victims into **factors**.

```
nypd_data_cleaned$BORO = as.factor(nypd_data_cleaned$BORO)
nypd_data_cleaned$PERP_AGE_GROUP = as.factor(nypd_data_cleaned$PERP_AGE_GROUP)
nypd_data_cleaned$PERP_SEX = as.factor(nypd_data_cleaned$PERP_SEX)
nypd_data_cleaned$PERP_RACE = as.factor(nypd_data_cleaned$PERP_RACE)
nypd_data_cleaned$VIC_AGE_GROUP = as.factor(nypd_data_cleaned$VIC_AGE_GROUP)
nypd_data_cleaned$VIC_SEX = as.factor(nypd_data_cleaned$VIC_SEX)
nypd_data_cleaned$VIC_RACE = as.factor(nypd_data_cleaned$VIC_RACE)
```

The summary of the cleaned data set shows that the number of rows are still the same (23585), which means **all rows are preserved**.

```
summary(nypd_data_cleaned)
```

```
##   INCIDENT_KEY         OCCUR_DATE          OCCUR_TIME
##  Min.   :  9953245   Min.   :2006-01-01   Length:23585
##  1st Qu.: 55322804   1st Qu.:2008-12-31   Class1:hms
##  Median : 83435362   Median :2012-02-27   Class2:difftime
##  Mean   :102280741   Mean   :2012-10-05   Mode  :numeric
##  3rd Qu.:150911774   3rd Qu.:2016-03-02
##  Max.   :230611229   Max.   :2020-12-31
##
##             BORO         PRECINCT      JURISDICTION_CODE LOCATION_DESC
##  BRONX        :6701   Min.   :  1.00   Min.   :0.000     Length:23585
##  BROOKLYN     :9734   1st Qu.: 44.00   1st Qu.:0.000     Class :character
##  MANHATTAN    :2922   Median : 69.00   Median :0.000     Mode  :character
##  QUEENS       :3532   Mean   : 66.21   Mean   :0.333
##  STATEN ISLAND: 696   3rd Qu.: 81.00   3rd Qu.:0.000
##                       Max.   :123.00   Max.   :2.000
##                                        NA's   :2
##  PERP_AGE_GROUP PERP_SEX              PERP_RACE     VIC_AGE_GROUP   VIC_SEX
##  18-24  :5508   F   :  335   BLACK         :10025   <18    : 2525   F: 2204
##  25-44  :4714   M   :13490   WHITE HISPANIC: 1988   18-24  : 9003   M:21370
##  UNKNOWN:3148   U   : 1499   UNKNOWN       : 1836   25-44  :10303   U:   11
##  <18    :1368   NA's: 8261   BLACK HISPANIC: 1096   45-64  : 1541
##  45-64  : 495                WHITE         :  255   65+    :  154
##  (Other):  57                (Other)       :  124   UNKNOWN:   59
##  NA's   :8295                NA's          : 8261
##                            VIC_RACE          hour           minute
##  AMERICAN INDIAN/ALASKAN NATIVE:    9   Min.   : 0.00   Min.   : 0.00
```

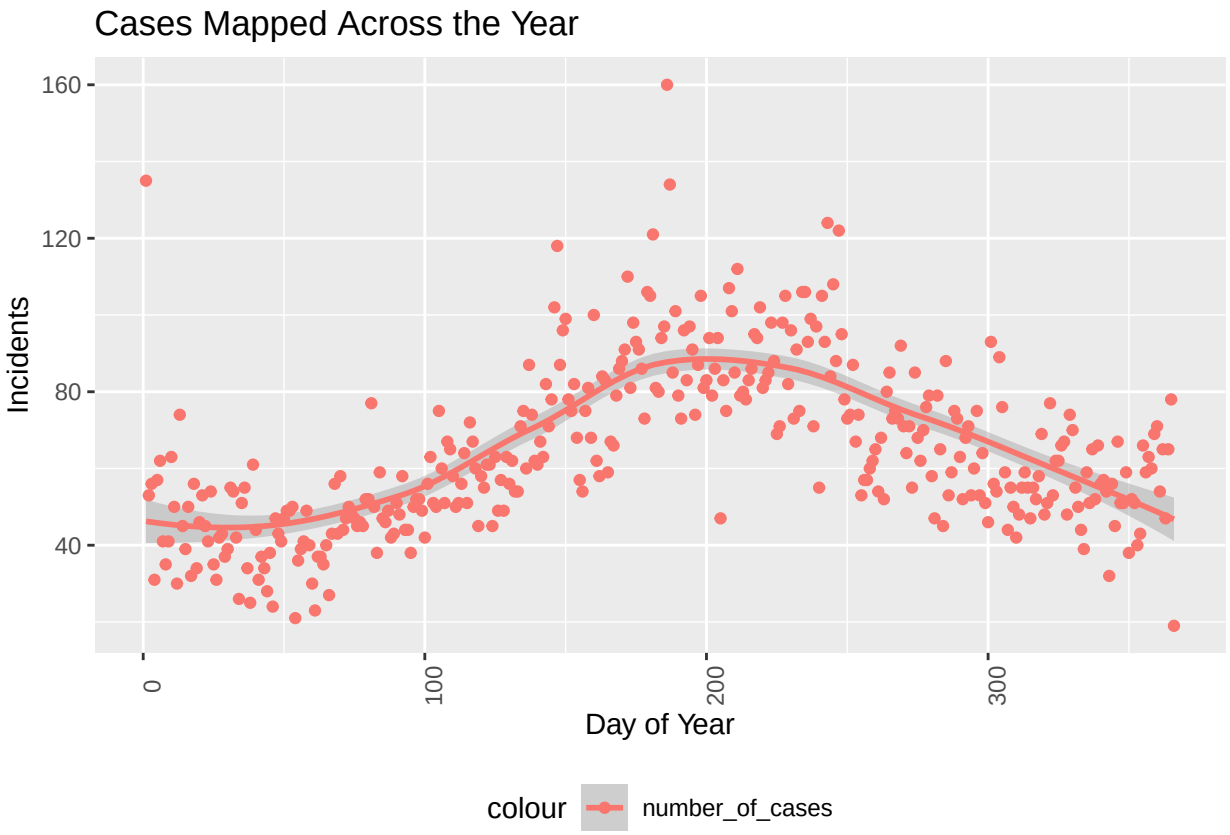
```
##  ASIAN / PACIFIC ISLANDER      :  327   1st Qu.: 3.00   1st Qu.:14.00
##  BLACK                         :16869   Median :15.00   Median :30.00
##  BLACK HISPANIC                : 2245   Mean   :12.08   Mean   :28.21
##  UNKNOWN                       :   65   3rd Qu.:20.00   3rd Qu.:44.00
##  WHITE                         :  620   Max.   :23.00   Max.   :59.00
##  WHITE HISPANIC                : 3450
```

In case any row is missing, we should always revisit the steps involved in cleaning the data, find out where the data was mistakenly dropped out, and try to fix that.

PART ONE: SEASONAL VARIATION OF SHOOTING INCIDENTS

We will look at the date of occurence to see if any pattern can be observed. We will map the cases across the year. We first count the number of cases per day of year, from day 1 to day 366 (which means Jan 1 to Dec 31). We will then plot out the graph, using day of year as x-axis, and incident numbers as y-axis.

```
daily_count <- nypd_data_cleaned %>%
  mutate(day_of_year = yday(OCCUR_DATE)) %>%
  group_by(day_of_year) %>%
  summarize(number_of_cases = n())

daily_count_plotted <- daily_count %>%
  ggplot(aes(x = day_of_year, y = number_of_cases)) +
  geom_smooth(aes(y = number_of_cases, color = "number_of_cases")) +
  geom_point(aes(y = number_of_cases, color = "number_of_cases")) +
  theme(legend.position="bottom", axis.text.x = element_text(angle = 90)) +
  labs(title = "Cases Mapped Across the Year", y = "Incidents", x = "Day of Year")
daily_count_plotted
```



Observations and Further Questions Regarding the Date of Cases

We can see that there is a **yearly low season in around Jan to Mar, and a high season in Jul and Aug (days 180 to 240)**.

This may lead us to **another question for further analysis**. Are there more cases happening in summer related to young people in general? What is the correlation between the occurrence date and the age/race of the perpetrators/victims?

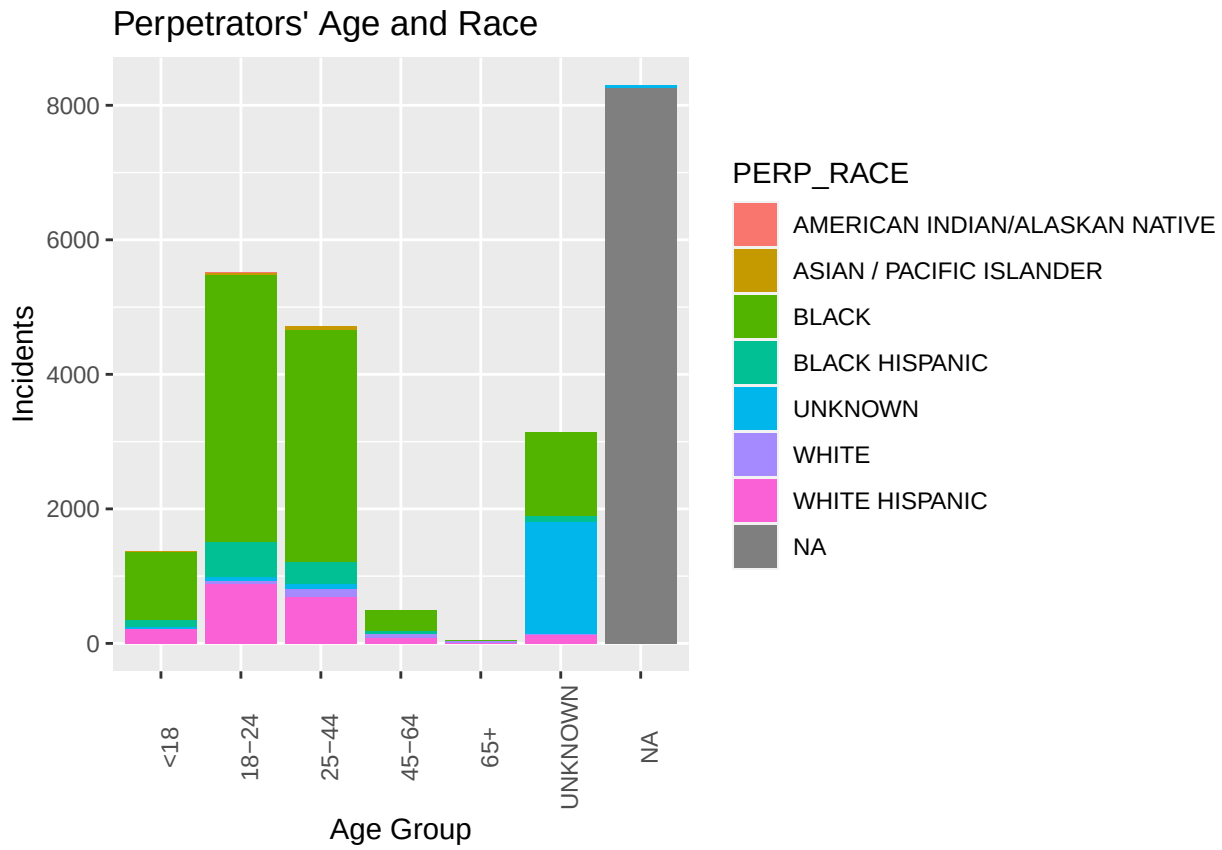
Analyzing the Perpetrators and Victims

To shed light on this observation, we will now take a look at the demographics of the perpetrators and victims.

We will first remove a few problematic outlier rows that do not have a reasonable age group. After that, we group the cases by their age groups. We can see that the top age group is 25-44, followed by 18-24. There is also a high NA rate, perhaps due to unresolved cases.

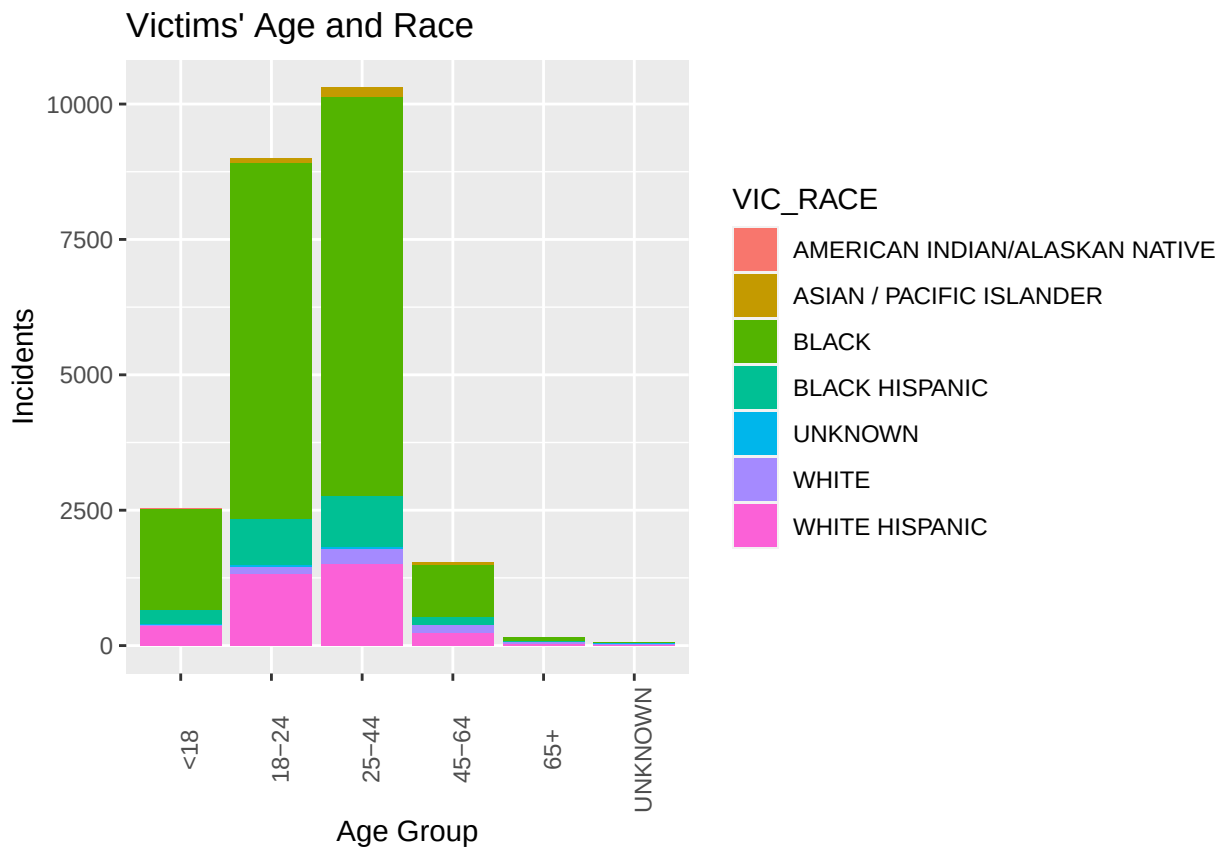
```
nypd_data_cleaned_perp <- nypd_data_cleaned %>%
  filter(PERP_AGE_GROUP != "1020" | is.na(PERP_AGE_GROUP)) %>%
  filter(PERP_AGE_GROUP != "224" | is.na(PERP_AGE_GROUP)) %>%
  filter(PERP_AGE_GROUP != "940" | is.na(PERP_AGE_GROUP))

grouped_by_perp_age <- nypd_data_cleaned_perp %>%
```

```
  group_by(PERP_AGE_GROUP, PERP_RACE, PERP_SEX) %>%
  tally()

grouped_by_perp_age_chart_with_race <- grouped_by_perp_age %>%
  ggplot(aes(x = PERP_AGE_GROUP, y = n, fill=PERP_RACE)) +
  geom_bar(position="stack", stat="identity") +
  theme(legend.position="right", axis.text.x = element_text(angle = 90)) +
  labs(title = "Perpetrators' Age and Race", y = "Incidents", x = "Age Group")
grouped_by_perp_age_chart_with_race
```



We will then look at the **victims' age groups and race**. We can observe similar patterns as in the perpetrators data.

```
nypd_data_cleaned_vic <- nypd_data_cleaned

grouped_by_vic_age <- nypd_data_cleaned_vic %>%
  group_by(VIC_AGE_GROUP, VIC_RACE, VIC_SEX) %>%
  tally()

grouped_by_vic_age_chart_with_race <- grouped_by_vic_age %>%
  ggplot(aes(x = VIC_AGE_GROUP, y = n, fill=VIC_RACE)) +
  geom_bar(position="stack", stat="identity") +
  theme(legend.position="right", axis.text.x = element_text(angle = 90)) +
  labs(title = "Victims' Age and Race", y = "Incidents", x = "Age Group")
grouped_by_vic_age_chart_with_race
```



At the top, you can see the perpetrators, and on the bottom are the victims. These charts also show the race but we may just ignore them for now. We can see that for both perpetrators and victims, the age groups 18-24 and 25-44 are the top two groups. And if we consider the relative population sizes of these two age groups, we can deduce that the age group 18-24 has the highest rate of involvement among all age groups.

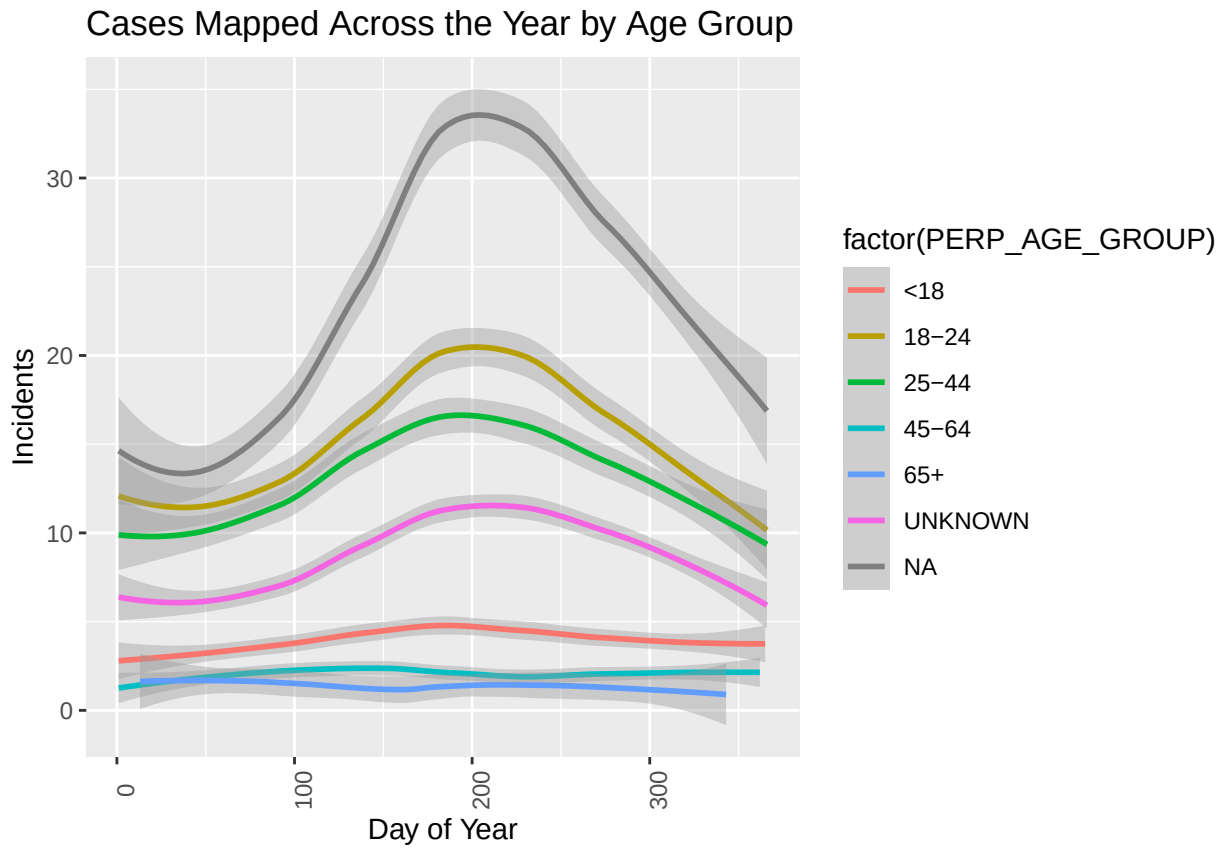
Observations and Further Questions Regarding Age Groups and Seasonal Variation

This is an important finding that can explain the variation in the incident number across the year. During summer, in July and August, young people are having summer vacation. They may have more free time on their own, which might result in more conflicts outside of schools. This hypothesis is verified in the following chart.

If we modify the first chart and break down the incident number by the perpetrators' age, we can get a closer look at the numbers involving each age group. You can see that for older groups such as 45-64 and over 65, the number remains rather consistent or even decreases slightly during summer. However, the numbers for 18-24 and 25-44 increases significantly during summer. Especially for 18-24, the rate of increase is much more dramatic. This verifies our hypothesis that the increase in the total incident number in summer may be related to the fact that young people are having summer vacation.

```
daily_count_with_age <- nypd_data_cleaned_perp %>%
  mutate(day_of_year = yday(OCCUR_DATE)) %>%
  group_by(day_of_year, PERP_AGE_GROUP) %>%
  summarize(number_of_cases_by_age_group = n())

daily_count_with_age_plotted <- daily_count_with_age %>%
  ggplot(aes(x = day_of_year, y = number_of_cases_by_age_group)) +
```

```
  geom_smooth(aes(y = number_of_cases_by_age_group, color = factor(PERP_AGE_GROUP))) +
  theme(legend.position="right", axis.text.x = element_text(angle = 90)) +
  labs(title = "Cases Mapped Across the Year by Age Group", y = "Incidents", x = "Day of Year")
daily_count_with_age_plotted
```



Other than age, another factor for the seasonal variation is the weather. The weather is more favorable in summer, and the freezing cold in winter might keep more people home. This may reduce the chance of conflicts.

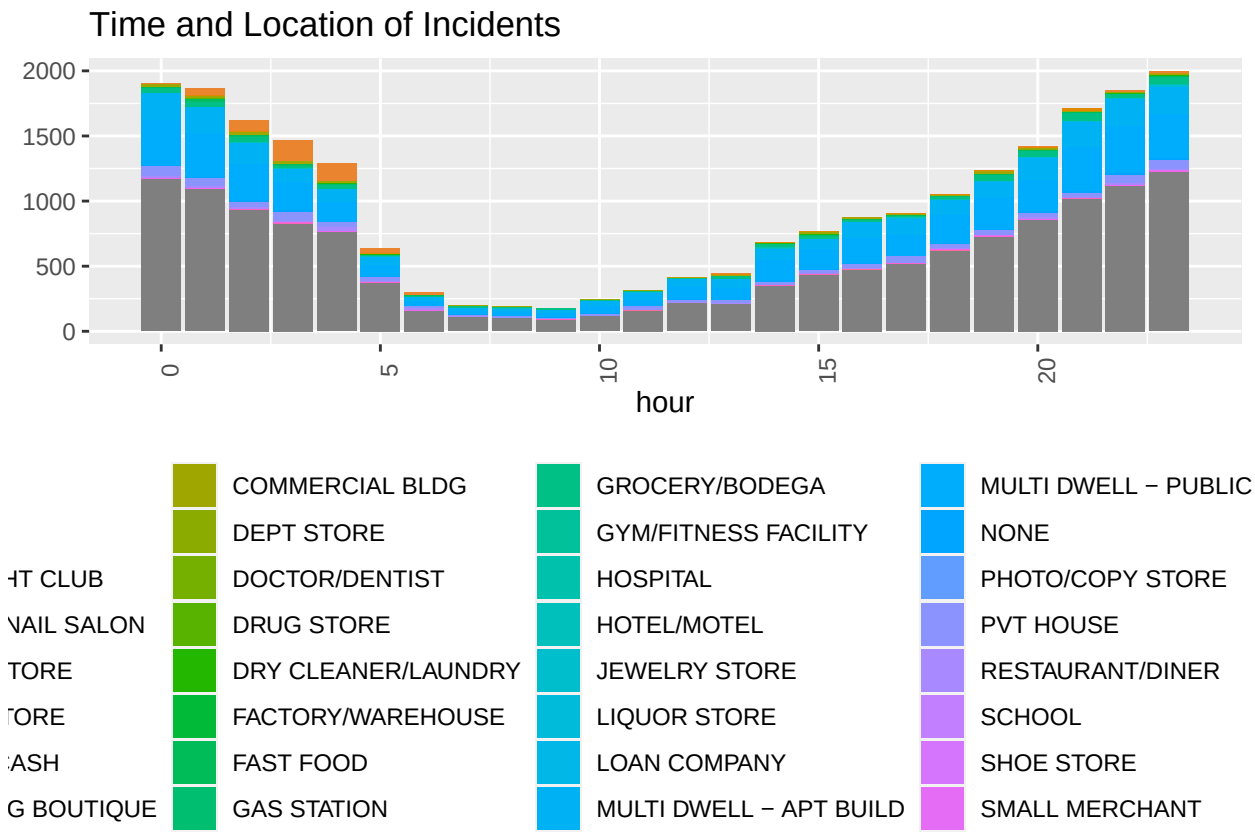
PART TWO - HAPPENING TIME/HOUR AND LOCATION OF SHOOTING INCIDENTS

After looking at the variation of the incident number across the year, we will now look at the time and location of the shootings. We are grouping the NYPD shooting incidents data by the hour first.

The following chart shows the number of incidents in different types of location in each of the 24 hours of a day.

```
hourly_plot <- nypd_data_cleaned %>%
  ggplot(aes(x = hour, fill = LOCATION_DESC)) +
  geom_bar() +
  theme(legend.position="bottom", axis.text.x = element_text(angle = 90)) +
```

```
  labs(title = "Time and Location of Incidents", y = NULL)
hourly_plot
```



The lowest numbers are found in the morning, between 6am - 11am. The number continues to rise throughout the afternoon and evening, until it reaches the peak at midnight, when the number becomes almost 10 times of the number in the morning. The number then goes down after midnight gradually. This result is perhaps not too surprising, especially if we consider the location of the shootings

The location with the largest number is the blue segment here, which are apartments. It takes up a relatively constant percentage in each hour. Other than apartments, we can also see that incidents in restaurants and diners (colored in violet) start to rise in the evening, reach a peak at midnight, and retreat slowly until the wee hours. Another notable location is the orange color, which are bars and night clubs. The number swells significantly between midnight and 4am. And the highest number at bars and night clubs is recorded between 3-4am.

So, with apartments being the top location, followed by restaurants or diners and bars or night clubs, it makes perfect sense that shootings happen mostly at night. And if we consider the location in relation to the seasonal variation discussed earlier, we may come to another hypothesis that during winter, less people may go out for dinner, go to bars, or stay out late. Therefore, the incident number is also lower.

MODELING: Predicting the number of cases involving perpertrators in the age group 18-24 with the total case number

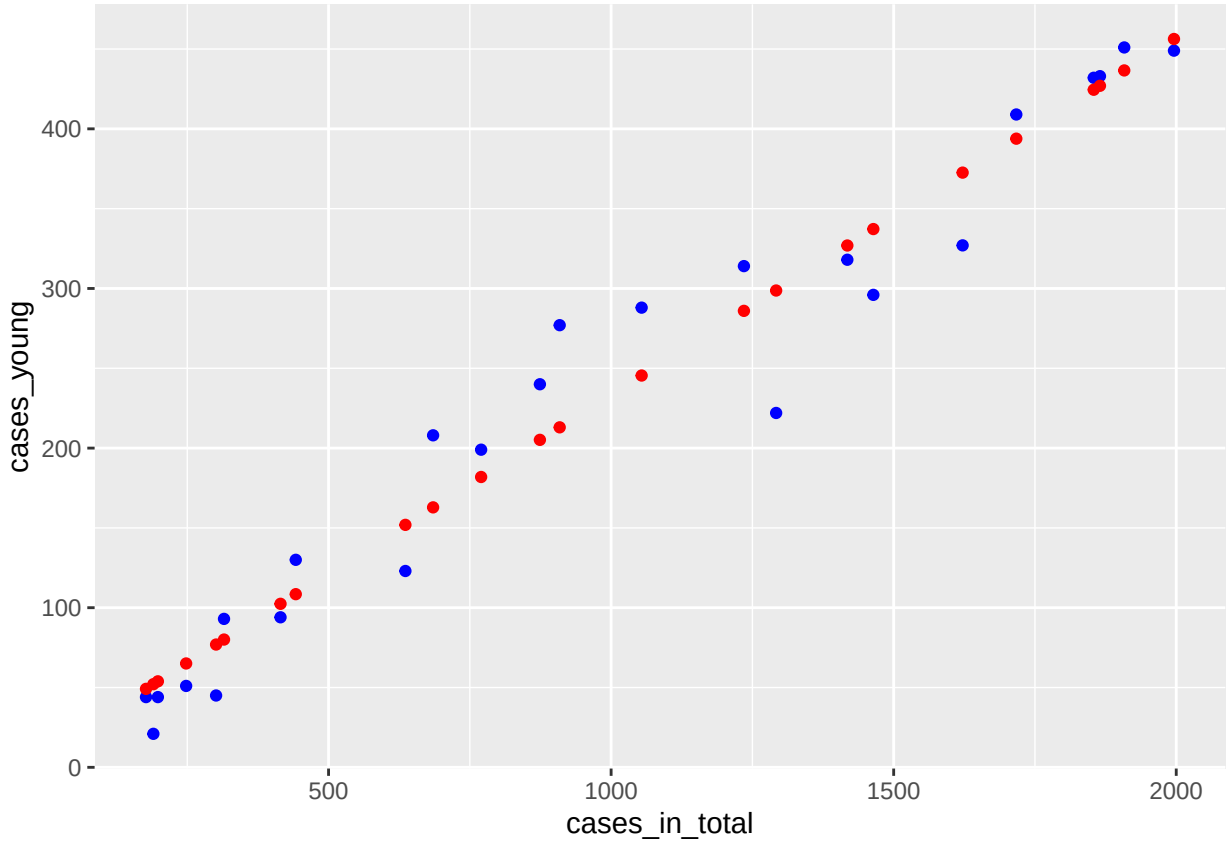
Lastly, we will try to model our data to predict the number of cases committed by perpetrators aged 18-24, by using the total number of cases. We will first segment the data by the 24 hours of a day. We will then run our linear model for prediction. After that, we plot the predicted cases involving perpetrators between 18 and 24 together with the actual recorded cases. We can see that there is a direct linear relationship between the total cases and involving the age group 18-24.

```
nypd_data_for_model <- nypd_data_cleaned %>%
  group_by(hour)

nypd_data_young <- nypd_data_for_model %>%
  filter(PERP_AGE_GROUP == "18-24") %>%
  group_by(hour) %>%
  summarise(cases_young=n())

nypd_data_total_cases <- nypd_data_for_model %>%
  group_by(hour) %>%
  summarise(cases_in_total=n())

combined <- left_join(nypd_data_young, nypd_data_total_cases, by = c("hour"="hour"))


mod <- lm(cases_young ~ cases_in_total, data = combined)
summary(mod)
```

```
##
## Call:
## lm(formula = cases_young ~ cases_in_total, data = combined)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -76.72 -17.77   0.43  18.21  64.00
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     9.55234   12.76199   0.748    0.462
## cases_in_total  0.22382    0.01099  20.365 9.13e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.31 on 22 degrees of freedom
## Multiple R-squared:  0.9496, Adjusted R-squared:  0.9473
## F-statistic: 414.7 on 1 and 22 DF,  p-value: 9.129e-16
```

```
combined_w_pred <- combined %>%
  mutate(pred = predict(mod))

combined_w_pred_plotted <- combined_w_pred %>%
  ggplot() +
  geom_point(aes(x=cases_in_total, y=cases_young), color="blue") +
```

```
  geom_point(aes(x=cases_in_total, y=pred), color="red")
combined_w_pred_plotted
```



Bias and Mitigation

In this analysis, we may have the bias that young people are more likely to be involved (being either the victim or the perpetrator) in these shooting incidents. This is a bias because we haven't analyzed in detail the relative population sizes of these demographic groups in NYC.

To carry out a fair comparison, we should get further information on the population sizes and compute the cases per thousand/million people for each of the demographic groups. We should never have any prejudice based on a person's age, race, sex. More data should be pulled in and comparisons should be done to get an objective view and analysis on these data sets later on. And it's also good to consider the borough/precinct population sizes if we were to analyze the cases in terms of their location or community demographics.

Conclusion

To wrap up our analysis, we have observed that there is a seasonal variation in the number of shooting incidents across the year. The low season is January to March, while the peak season is July to August. The higher number in summer may be attributed to the fact that young people are having summer vacation and the weather is milder.

The second observation is that most shooting incidents take place in the night, with the number peaking at midnight. This has close relationships with the location.

Lastly, we have modeled our data to predict the number of cases involving the 18-24 years old age group, by using the total number of cases in each hour.

With these observations, it may be easier for the related departments to take preemptive measures in combating gun violence. We hope that gun violence can be eliminated completely in NYC soon.