NYC Shooting

Ye Lwin Oo

2024-11-30

Required Library Packages

```
library(tidyverse)
library(grid)
library(knitr)
installed.packages()[names(sessionInfo()$otherPkgs), "Version"]
```

Download NYC Shooting Data

```
#download NYC shooting data
url_in <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD/"
filename <- "NYPD_Shooting_Incident_Data__Historic_.csv"
data <- read_csv(filename)
```

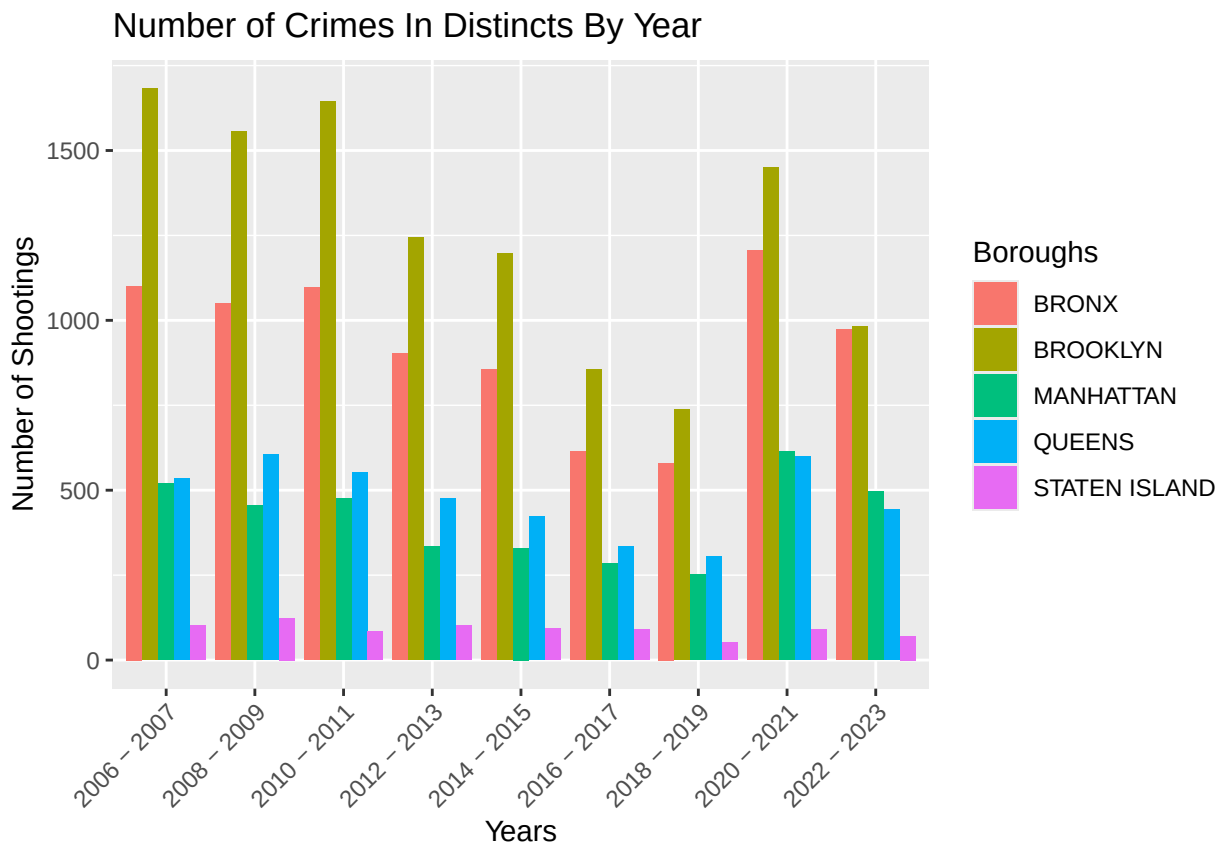
```
filtered_data <- data %>% select(-c(INCIDENT_KEY, JURISDICTION_CODE, X_COORD_CD:last_col()))
#converting data type for occur date
filtered_data <- filtered_data %>% mutate(DATE = mdy(OCCUR_DATE), .after = OCCUR_DATE) %>% select(-c(OC
```

Analysis of the Relation between Number of Shootings in Each Borough and Time

Number of Shootings In Boroughs By Year

```
#get data for each boroughs when the shootings usually occurs within 2 years range
year_data <- filtered_data %>% mutate(YEAR = as.numeric(format(DATE, "%Y")))
range_size = 2
year_data <- year_data %>% mutate(YEAR_RANGE = paste(
  floor((YEAR - min(YEAR)) / range_size) * range_size + min(YEAR),
  floor((YEAR - min(YEAR)) / range_size) * range_size + min(YEAR) + (range_size - 1),
  sep = " - "))

yearly_shooting <- year_data %>% count(YEAR_RANGE, BORO)
ggplot(yearly_shooting, aes(x = YEAR_RANGE, y = n , fill = BORO)) + geom_bar(stat = "identity", positio
  labs(title = "Number of Crimes In Distincts By Year", x = "Years", y = "Number of Shootings", fill =
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



The plot above shows that Brooklyn has the highest shootings among all the countries for all years across followed by Bronx. Staten Island has the lowest number of shootings. The downloaded data does not contain the population in each county but from the data here, Brooklyn has the highest population and Staten Island has the lowest since 1990. In 2022, Brooklyn has 2,561,225, Bronx has 1,356,476, Manhattan has 1,597,451, Queens has 2,252,196and Staten Island has 490,687. From the plot, Bronx has the second highest shootings although Queens has much higher population compared to Bronx. It would be interesting to look further at the relation of number of shootings and the population density of each county to understand which county has the highest shootings per population.

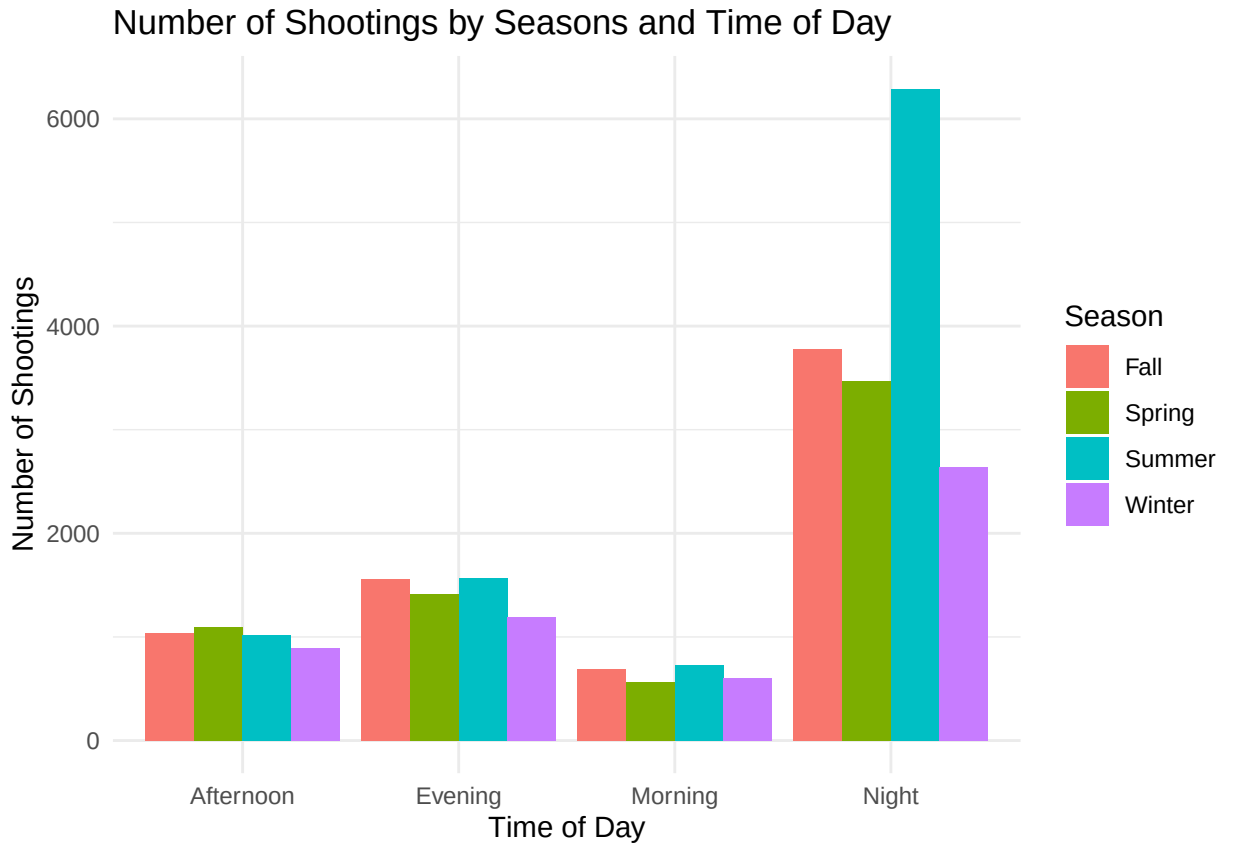
Number of Shootings Occurred Based On Time Of the Year and the Day

```
#split the occur dates into seasons
season_data <- filtered_data %>%
  mutate(SEASON = case_when(
    format(DATE, "%m") %in% c("12", "01", "02") ~ "Winter",
    format(DATE, "%m") %in% c("03", "04", "05") ~ "Spring",
    format(DATE, "%m") %in% c("06", "07", "08") ~ "Summer",
    format(DATE, "%m") %in% c("09", "10", "11") ~ "Fall"
  ))
#split the occur time
season_data <- season_data %>%
  mutate(
    HOUR = hour(season_data$OCCUR_TIME),
    TIME_OF_DAY = case_when(
```

```
      HOUR >= 5 & HOUR < 12 ~ "Morning",
      HOUR >= 12 & HOUR < 17 ~ "Afternoon",
      HOUR >= 17 & HOUR < 21 ~ "Evening",
      TRUE ~ "Night"))

# Count occurrences by season and time of day
counts <- season_data %>% count(SEASON, TIME_OF_DAY)

# Plot data
ggplot(counts, aes(x = TIME_OF_DAY, y = n, fill = SEASON)) + geom_bar(stat = "identity", position = "do
  labs(title = "Number of Shootings by Seasons and Time of Day", x = "Time of Day", y = "Number of Shoo
  theme_minimal()
```



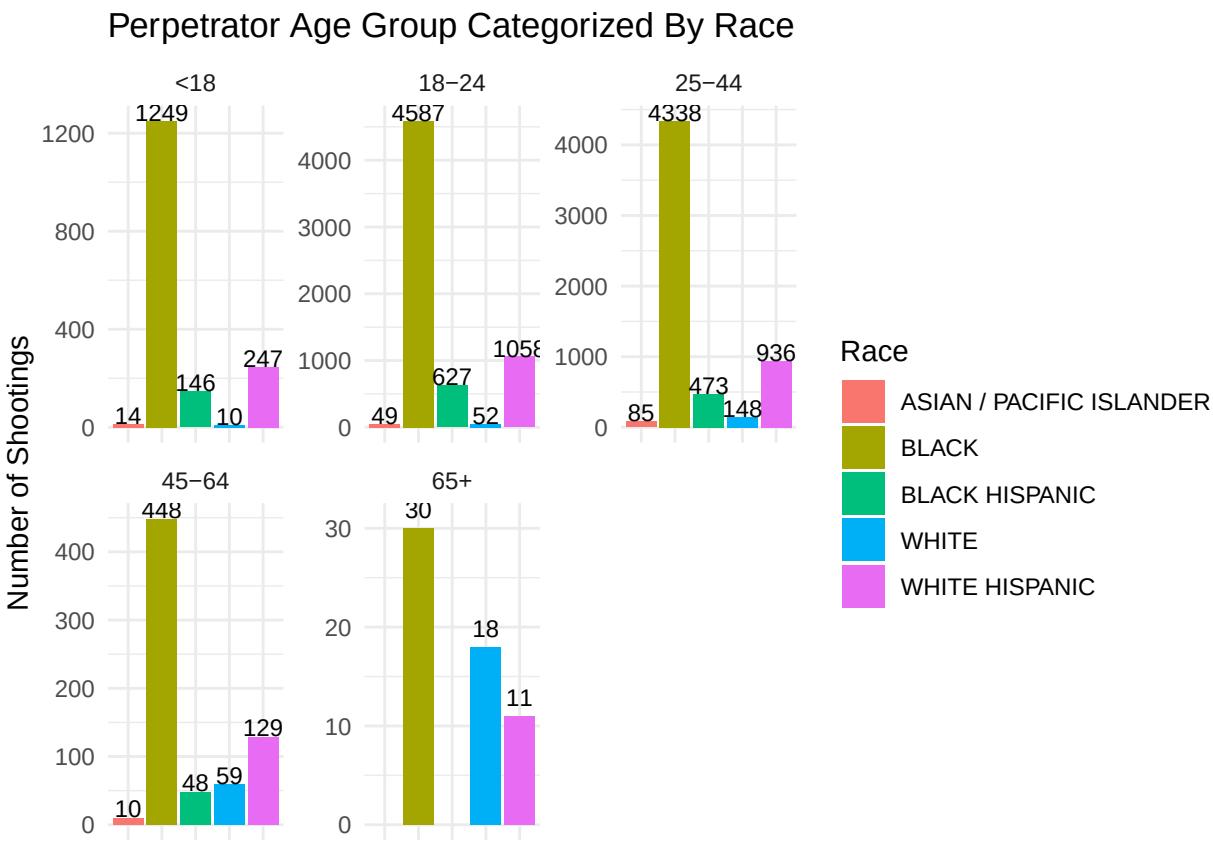
The plot above shows that most of the shootings occurred at night and during summer. This is intuitive since crimes mostly occur at night and people do not go outside much during winter because of the cold temperature. It is also interesting to note that the earlier the time of the day is, the lower the number of shootings. Moreover, there is not much different in number of shootings in afternoon, evening and morning regardless of the season. These results could also be because it is difficult to carry weapons when the daylight is still present.

Analysis of the Relation between Number Of Shootings and Race/Age Group

Relation of Perpetrator Age Group and Race and Number of Shootings

```
#data for perp age and race
perp_vic_data <- filtered_data %>% mutate(PERP_AGE_GROUP = case_when(
  PERP_AGE_GROUP == "(null)" | PERP_AGE_GROUP == "UNKNOWN"| PERP_AGE_GROUP == "1020"| PERP_AGE_GROUP ==
                                                          TRUE ~ as.character(PERP_AGE_GROUP)),
                                    PERP_RACE = case_when(PERP_RACE == "(null)" | PERP_RACE == "UNKNOWN"
                                                          TRUE ~ as.character(PERP_RACE)))
perp_vic_data <- perp_vic_data %>% mutate(VIC_AGE_GROUP = case_when(VIC_AGE_GROUP == "(null)" | VIC_AGE
                                                            TRUE ~ as.character(VIC_AGE_GROUP)),
                                      VIC_RACE = case_when(VIC_RACE == "(null)" | VIC_RACE == "UNKNOWN"
                                                           TRUE ~ as.character(VIC_RACE)))
#take out one category because of lack of data
perp_count_data <- perp_vic_data %>% count(PERP_AGE_GROUP,PERP_RACE)  %>%
  replace_na(list(PERP_AGE_GROUP = "Missing", PERP_RACE = "Missing")) %>%
  filter(PERP_AGE_GROUP != "Missing" & PERP_RACE != "Missing" & n >5)

ggplot(perp_count_data, aes(x = PERP_RACE, y = n, fill = PERP_RACE)) + geom_bar(stat = "identity", posi
  labs(title = "Perpetrator Age Group Categorized By Race",x = NULL,y = "Number of Shootings",fill = "Ra
  facet_wrap(~ PERP_AGE_GROUP, scales = "free_y") +
  theme(
    axis.text.x = element_blank(),  # Remove X-axis labels
    axis.title.x = element_text(size = 12),  # Optional: If you want to keep the X-axis title
    axis.text.x.bottom = element_blank()
  ) +
  # Add labels for bars exceeding the threshold
  geom_text(
    data = perp_count_data,  # Label bars
    aes(label = n, y = n + 1),  # Position labels slightly above the bars
    size = 3, color = "black", vjust = 0
  )
```



The plot above shows that age group between 18-24 and 25-44 Black person commit shooting the most. Some of the data were excluded because either race or age group was only recorded. The table below showed the number of missing information on perpetrator. It would be interesting to investigate further on the population percentage and median income levels of each race in NYC to analyse if there is any relation with the number of shootings. It is also important to note that age below 18 has more than a thousand shooting cases even though the legal age to carry a gun in NYC is 21 years old.

Number of Missing Perpetrator Information

The table below shows the number of missing information on perpetrator. This is interesting to note that mostly age of the perpetrator was not reported. In these incomplete information, we can still observe that age group between 18-44 and black were reported most.

```
# Filter out rows where race or age_group is missing
perp_missing_data <- perp_vic_data %>%
  filter(is.na(PERP_RACE) | is.na(PERP_AGE_GROUP)) %>%
  group_by(PERP_RACE, PERP_AGE_GROUP) %>%
  summarise(missing_count = n(), .groups = 'drop') %>%
  slice(-n())

colnames(perp_missing_data) <- c("Race", "Age Group", "Number of Incidents")
kable(perp_missing_data, caption = "Perpetrator Age/Race Missing In Incidents")
```

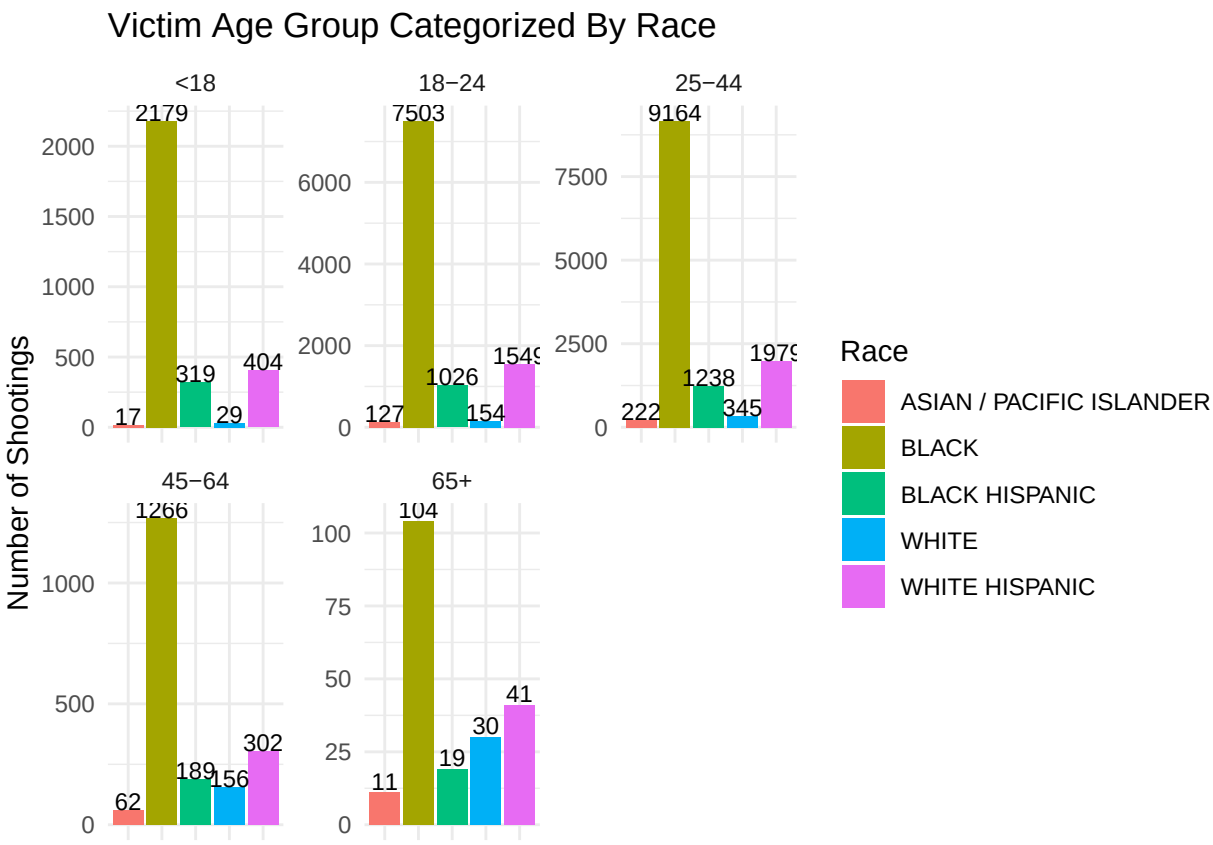
Table 1: Perpetrator Age/Race Missing In Incidents

Race	Age Group	Number of Incidents
ASIAN / PACIFIC ISLANDER	NA	11
BLACK	NA	1251
BLACK HISPANIC	NA	93
WHITE	NA	11
WHITE HISPANIC	NA	129
NA	18-24	64
NA	25-44	60
NA	45-64	5
NA	65+	1
NA	<18	16

Relation of Victim Age Group and Race and Number of Shootings

```
vic_count_data <- perp_vic_data %>% count(VIC_AGE_GROUP,VIC_RACE)  %>%
  replace_na(list(VIC_AGE_GROUP = "Missing", VIC_RACE = "Missing")) %>%
  filter(VIC_AGE_GROUP != "Missing" & VIC_RACE != "Missing" & n >5)

ggplot(vic_count_data, aes(x = VIC_RACE, y = n, fill = VIC_RACE)) + geom_bar(stat = "identity", positio
  labs(title = "Victim Age Group Categorized By Race",x = NULL,y = "Number of Shootings" ,fill = "Race"
  facet_wrap(~ VIC_AGE_GROUP, scales = "free_y") +
  theme(
    axis.text.x = element_blank(),  # Remove X-axis labels
    axis.title.x = element_text(size = 12),  # Optional: If you want to keep the X-axis title
    axis.text.x.bottom = element_blank()
  ) +
  # Add labels for bars
  geom_text(
    data = vic_count_data,
    aes(label = n, y = n + 1),  # Position labels slightly above the bars
    size = 3, color = "black", vjust = 0
  )
```



The plot above shows that age group between 18-24 and 25-44 Black person was victimized the most. Similar to the above analysis, some of the data were excluded because either race or age group was only recorded. It is also interesting to note that the number of shootings trend for victims age group follows the same pattern as the perpetrator age group, where the age group 25-44 was mostly related to the shootings, followed by age 18-24, $<$ 18, 45-64 and 65+ in order. The victim race pattern is also very similar to the perpetrator race that White Hispanic being the second and black Hispanic being the third.

Number of Missing Victim Information

The table below shows the number of missing information on victim, The information is more complete than that of perpetrator. This makes sense since the reports will most likely to have complete victim information than perpetrator information.

```
# Filter out rows where race or age_group is missing
vic_missing_data <- perp_vic_data %>%
  filter(is.na(VIC_RACE) | is.na(VIC_AGE_GROUP)) %>%
  group_by(VIC_RACE, VIC_AGE_GROUP) %>%
  summarise(missing_count = n(), .groups = 'drop') %>%
  slice(-n())

colnames(vic_missing_data) <- c("Race", "Age Group", "Number of Incidents")
kable(vic_missing_data, caption = "Victim Age/Race Missing In Incidents")
```

Table 2: Victim Age/Race Missing In Incidents

Race	Age Group	Number of Incidents
ASIAN / PACIFIC ISLANDER	NA	1
BLACK	NA	18
BLACK HISPANIC	NA	4
WHITE	NA	14
WHITE HISPANIC	NA	8
NA	18-24	20
NA	25-44	21
NA	45-64	6
NA	<18	4

Modeling the Severity of Shooting Incident Based on Victim Age Group

The relation between the crime victim age group and whether the shooting resulted in the victim's death is analyzed. Victim age groups are categorized and filtered out the incomplete data. The generalized linear model is used to model the probability of the response variable STATISTICAL_MURDER_FLAG, which is a logical type as a function of the predictor, age group, which is the categorical type .The predicted probability will fall within the range of [0,1]. As can be seen from the table, the probability that the victim's death occur gets higher with the age group. We have observed that age group between 18-24 and 25-44 were victimized the most but the most serious consequence could happen for the older age group.

```
perp_vic_data_clean <- na.omit(perp_vic_data)
perp_vic_data_clean $VIC_RACE <- as.factor(perp_vic_data_clean $VIC_RACE)
perp_vic_data_clean $PERP_RACE <- as.factor(perp_vic_data_clean $PERP_RACE)

perp_vic_data_clean$VIC_AGE_GROUP <- as.factor(perp_vic_data_clean$VIC_AGE_GROUP)
perp_vic_data_clean$PERP_AGE_GROUP <- as.factor(perp_vic_data_clean$PERP_AGE_GROUP)

model_age <- glm(STATISTICAL_MURDER_FLAG ~ VIC_AGE_GROUP, data = perp_vic_data_clean, family = binomial)

# Create a dataset with unique age groups
age_data <- data.frame(VIC_AGE_GROUP = levels(perp_vic_data_clean$VIC_AGE_GROUP))

# Predict probabilities for each age group
age_data$predicted_prob <- predict(model_age, newdata = age_data, type = "response")

colnames(age_data) <- c("Age Group", "Predicted Probability")

# View the table
kable(age_data)
```

Age Group	Predicted Probability
<18	0.1616766
1022	0.0000035
18-24	0.2302772
25-44	0.2430196

Age Group	Predicted Probability
45-64	0.2673797
65+	0.3793103

Conclusion

From the above analysis, Brooklyn has the most number of shootings. It can also be seen that most of the shootings occur at night during summer and the shootings are mostly related to a black person age between 18 and 44.

The number of shootings in boroughs do not necessarily reflect how dangerous the borough is. When I looked up the population of NYC counties, Queens has the population almost as the same as Brooklyn. Queens has relatively much lower number of shootings compared to Brooklyn and Bronx, which has almost half population compared to Queens. If we utilize the population data, Bronx seems to be have as much if not more number of shootings per population as Brooklyn.

Moreover, this data did not mention how the records were obtained. This could be combination of the shootings where the perpetrator was arrested and the case was closed or some bystander reported to the police and the crime was not solved. If so, there could be a racial bias in the reporting and the age group might not be as accurate. To mitigate some of these, I excluded the data that has missing either the age group or the race to filter some data that might not be accurate.