NYPD Shooting Analysis

CVo2

2024-10-11

This report analyzes the NYPD shooting incident dataset to explore patterns in time of day, day of the week, and location of incidents. The goal is to provide insights into when and where shooting incidents are most likely to occur, and whether location influences the likelihood of incidents happening at night

Set up and load library

```
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate)
```

Start an rmd and inport dataset

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

```
## Rows: 28562 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (12): OCCUR_DATE, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION...
## dbl   (7): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD...
## lgl   (1): STATISTICAL_MURDER_FLAG
## time  (1): OCCUR_TIME
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
summary(nypd_data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO          
##  Min.   :  9953245   Length:28562       Length:28562      Length:28562      
##  1st Qu.: 65439914   Class :character   Class1:hms        Class :character  
##  Median : 92711254   Mode  :character   Class2:difftime   Mode  :character  
##  Mean   :127405824                      Mode  :numeric                      
##  3rd Qu.:203131993                                                          
##  Max.   :279758069                                                          
##                                                                             
##  LOC_OF_OCCUR_DESC     PRECINCT     JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:28562       Min.   :  1.0   Min.   :0.0000    Length:28562      
##  Class :character   1st Qu.: 44.0   1st Qu.:0.0000    Class :character  
##  Mode  :character   Median : 67.0   Median :0.0000    Mode  :character  
##                     Mean   : 65.5   Mean   :0.3219                      
##                     3rd Qu.: 81.0   3rd Qu.:0.0000                      
##                     Max.   :123.0   Max.   :2.0000                      
##                                     NA's   :2                           
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP    
##  Length:28562       Mode :logical           Length:28562      
##  Class :character   FALSE:23036             Class :character  
##  Mode  :character   TRUE :5526              Mode  :character  
##                                                               
##                                                               
##                                                               
##                                                               
##    PERP_SEX          PERP_RACE         VIC_AGE_GROUP        VIC_SEX         
##  Length:28562       Length:28562       Length:28562       Length:28562      
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##                                                                             
##    VIC_RACE           X_COORD_CD        Y_COORD_CD        Latitude    
##  Length:28562       Min.   : 914928   Min.   :125757   Min.   :40.51  
##  Class :character   1st Qu.:1000068   1st Qu.:182912   1st Qu.:40.67  
##  Mode  :character   Median :1007772   Median :194901   Median :40.70  
##                     Mean   :1009424   Mean   :208380   Mean   :40.74  
##                     3rd Qu.:1016807   3rd Qu.:239814   3rd Qu.:40.82  
##                     Max.   :1066815   Max.   :271128   Max.   :40.91  
##                                                        NA's   :59     
##    Longitude        Lon_Lat         
##  Min.   :-74.25   Length:28562      
##  1st Qu.:-73.94   Class :character  
##  Median :-73.92   Mode  :character  
##  Mean   :-73.91                     
##  3rd Qu.:-73.88                     
##  Max.   :-73.70                     
##  NA's   :59                         
```

Data Preparation and Cleaning

```
# Remove column
nypd_data_clean = nypd_data %>%
  select(-c(X_COORD_CD:Lon_Lat, PRECINCT, JURISDICTION_CODE))
# Change format
nypd_data_clean = nypd_data_clean %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE))
nypd_data_clean <- nypd_data_clean %>%
  mutate(OCCUR_TIME = hms(OCCUR_TIME))
# After remove column and change format, show the summary
summary(nypd_data_clean)
```

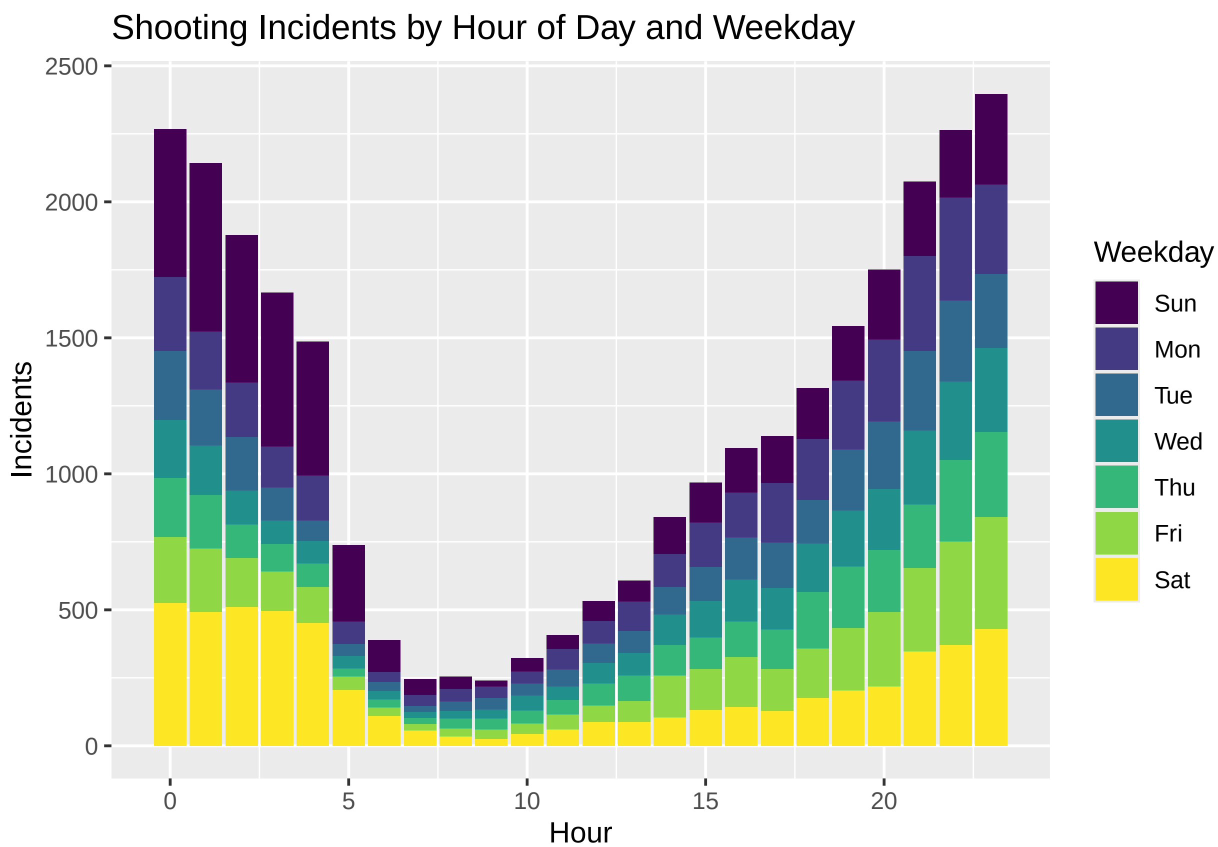
```
##   INCIDENT_KEY         OCCUR_DATE           OCCUR_TIME
##  Min.   :  9953245   Min.   :2006-01-01   Min.   :0S
##  1st Qu.: 65439914   1st Qu.:2009-09-04   1st Qu.:3H 30M 0S
##  Median : 92711254   Median :2013-09-20   Median :15H 15M 0S
##  Mean   :127405824   Mean   :2014-06-07   Mean   :12H 44M 16.713115328057S
##  3rd Qu.:203131993   3rd Qu.:2019-09-29   3rd Qu.:20H 45M 0S
##  Max.   :279758069   Max.   :2023-12-29   Max.   :23H 59M 0S
##      BORO            LOC_OF_OCCUR_DESC  LOC_CLASSFCTN_DESC LOCATION_DESC
##  Length:28562       Length:28562       Length:28562       Length:28562
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP       PERP_SEX
##  Mode :logical           Length:28562       Length:28562
##  FALSE:23036             Class :character   Class :character
##  TRUE :5526              Mode  :character   Mode  :character
##
##
##
##   PERP_RACE         VIC_AGE_GROUP        VIC_SEX            VIC_RACE
##  Length:28562       Length:28562       Length:28562       Length:28562
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
```

Visualization & Analysis 1: Shooting day of the week / Time of Day

```
nypd_data_clean %>%
  mutate(Weekday = wday(OCCUR_DATE, label = TRUE), Hour = hour(OCCUR_TIME)) %>%
  group_by(Weekday, Hour) %>%
  summarize(Incidents = n()) %>%
```

```r
  ggplot(aes(x = Hour, y = Incidents, fill = Weekday)) +
  geom_bar(stat = "identity") +
  labs(title = "Shooting Incidents by Hour of Day and Weekday")
```

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



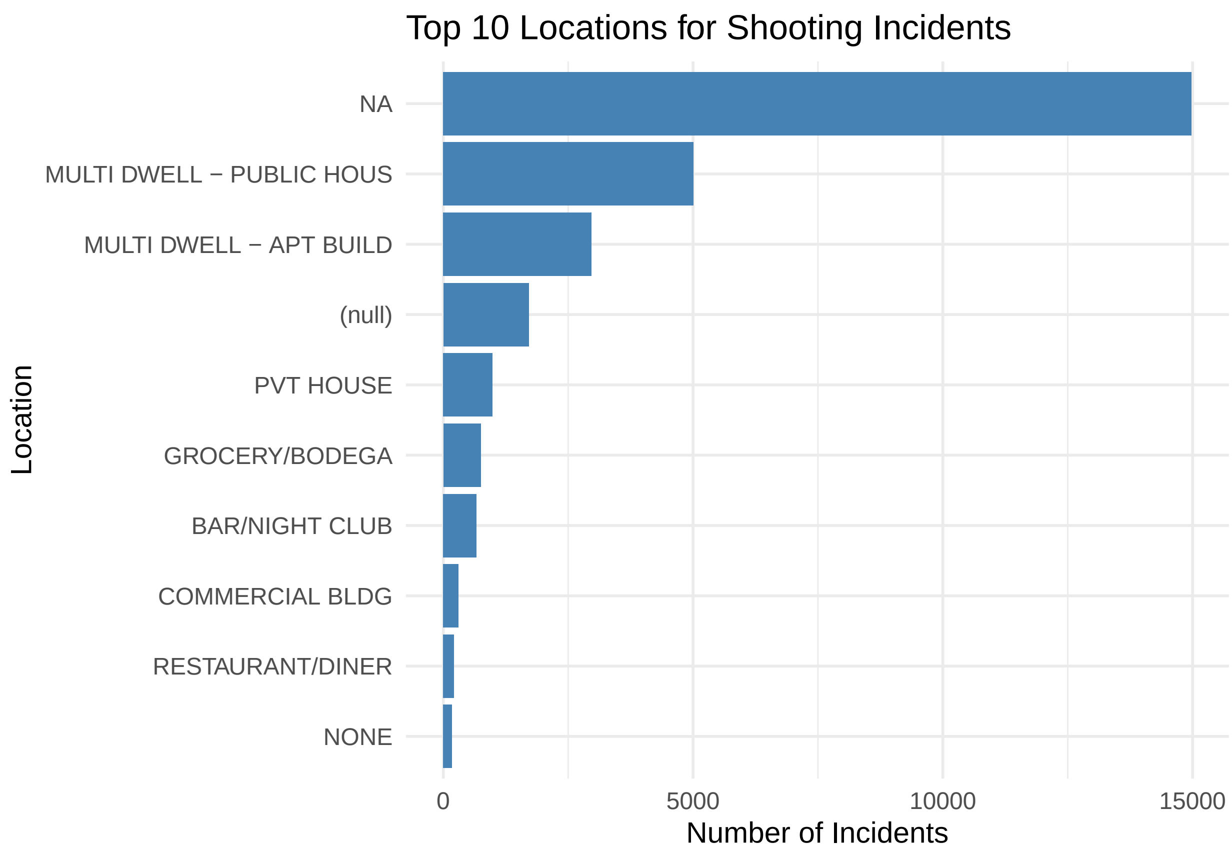
Analysis: Late-night hours on weekends are the most active times for shooting incidents, with a particular concentration between midnight and early morning. This may suggest that social activities or certain environmental factors contribute to the increased number of incidents during these periods.

Visualization and Analysis 2 : Location-Specific Analysis

Based on Analysis 1, we see that shootings peak during late-night hours, especially on weekends. However, knowing when shootings occur isn't enough—we also need to know where they happen. Analysis 2 focuses on identifying the most common locations for shootings, helping us understand the environmental factors at play and where targeted interventions might be needed.

```r
nypd_data_clean %>%
  group_by(LOCATION_DESC) %>%
  summarize(Incidents = n()) %>%
  arrange(desc(Incidents)) %>%
  top_n(10, Incidents) %>% # Display only the top 10 locations
```

```r
ggplot(aes(x = reorder(LOCATION_DESC, Incidents), y = Incidents)) +
geom_bar(stat = "identity", fill = "steelblue") +
coord_flip() + # Flip coordinates for better readability
labs(
  title = "Top 10 Locations for Shooting Incidents",
  x = "Location",
  y = "Number of Incidents"
 ) +
theme_minimal()
```



Analysis:

The analysis shows that a significant number of incidents fall under the category "NA," indicating missing or incomplete location data, which affects the clarity of location-based insights and suggests a need for data collection improvements. Multi-dwelling residences, particularly public housing and apartment buildings, are the most frequent known locations for shooting incidents, highlighting the need for targeted safety interventions in these densely populated areas. Other significant locations, such as private houses, grocery/bodegas, and bars/nightclubs, also see notable incidents, suggesting that both residential and public spaces require varied strategies to address the issue effectively.

Model

After identifying the key locations where shooting incidents occur, the next logical step is to examine how these locations influence the likelihood of incidents happening during late-night hours. By using a logistic regression model, we can determine which locations are most strongly associated with late-night shootings.

This analysis will help us understand not only where shootings are occurring but also when they are most likely to happen, providing deeper insights into potential patterns and allowing for more targeted interventions based on both time and location.

```r
# Add a binary variable for late-night (Yes/No)
nypd_data_clean <- nypd_data_clean %>%
  mutate(Late_Night = ifelse(hour(OCCUR_TIME) >= 22 | hour(OCCUR_TIME) < 5, 1, 0))

# Fit a logistic regression model to see if the location influences late-night shootings
model <- glm(Late_Night ~ LOCATION_DESC, data = nypd_data_clean, family = binomial)
summary(model)
```

```
##
## Call:
## glm(formula = Late_Night ~ LOCATION_DESC, family = binomial,
##     data = nypd_data_clean)
##
## Coefficients:
##                                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)                              -0.374873   0.049203  -7.619 2.56e-14
## LOCATION_DESCATM                        -14.191195 882.743376  -0.016 0.987174
## LOCATION_DESCBANK                       -14.191195 509.652128  -0.028 0.977786
## LOCATION_DESCBAR/NIGHT CLUB               2.031765   0.116401  17.455  < 2e-16
## LOCATION_DESCBEAUTY/NAIL SALON           -1.224515   0.250044  -4.897 9.72e-07
## LOCATION_DESCCANDY STORE                  0.662555   0.765346   0.866 0.386658
## LOCATION_DESCCHAIN STORE                 -1.416887   1.081244  -1.310 0.190053
## LOCATION_DESCCHECK CASH                 -14.191195 882.743377  -0.016 0.987174
## LOCATION_DESCCLOTHING BOUTIQUE           -1.416887   0.765346  -1.851 0.064126
## LOCATION_DESCCOMMERCIAL BLDG              0.243103   0.125044   1.944 0.051878
## LOCATION_DESCDEPT STORE                 -14.191195 294.247796  -0.048 0.961534
## LOCATION_DESCDOCTOR/DENTIST             -14.191195 882.743377  -0.016 0.987174
## LOCATION_DESCDRUG STORE                  -0.924410   0.653195  -1.415 0.157006
## LOCATION_DESCDRY CLEANER/LAUNDRY          0.249710   0.357647   0.698 0.485051
## LOCATION_DESCFACTORY/WAREHOUSE            0.374873   0.708817   0.529 0.596895
## LOCATION_DESCFAST FOOD                    0.344101   0.182202   1.889 0.058949
## LOCATION_DESCGAS STATION                  0.047660   0.240696   0.198 0.843039
## LOCATION_DESCGROCERY/BODEGA               0.095741   0.088650   1.080 0.280146
## LOCATION_DESCGYM/FITNESS FACILITY         0.374873   1.001210   0.374 0.708092
## LOCATION_DESCHOSPITAL                    -0.241313   0.243840  -0.990 0.322351
## LOCATION_DESCHOTEL/MOTEL                  0.317714   0.341760   0.930 0.352557
## LOCATION_DESCJEWELRY STORE              -14.191195 235.923096  -0.060 0.952035
## LOCATION_DESCLIQUOR STORE                 0.087191   0.315663   0.276 0.782383
## LOCATION_DESCLOAN COMPANY               -14.191195 882.743377  -0.016 0.987174
## LOCATION_DESCMULTI DWELL - APT BUILD      0.184012   0.061504   2.992 0.002773
## LOCATION_DESCMULTI DWELL - PUBLIC HOUS    0.259306   0.056767   4.568 4.93e-06
## LOCATION_DESCNONE                         0.179972   0.159674   1.127 0.259690
## LOCATION_DESCPHOTO/COPY STORE            14.940941 882.743377   0.017 0.986496
## LOCATION_DESCPVT HOUSE                    0.226074   0.080701   2.801 0.005089
## LOCATION_DESCRESTAURANT/DINER             0.796086   0.148789   5.350 8.77e-08
## LOCATION_DESCSCHOOL                      14.940941 882.743377   0.017 0.986496
## LOCATION_DESCSHOE STORE                  -1.822352   1.055240  -1.727 0.084176
## LOCATION_DESCSMALL MERCHANT              -0.983251   0.376966  -2.608 0.009099
## LOCATION_DESCSOCIAL CLUB/POLICY LOCATI    0.966924   0.249317   3.878 0.000105
## LOCATION_DESCSTORAGE FACILITY           -14.191195 882.743377  -0.016 0.987174
```

```
## LOCATION_DESCSTORE UNCLASSIFIED         -0.008119   0.338441  -0.024 0.980860
## LOCATION_DESCSUPERMARKET                -0.788278   0.514705  -1.532 0.125642
## LOCATION_DESCTELECOMM. STORE           -14.191195 266.157147  -0.053 0.957478
## LOCATION_DESCVARIETY STORE              -1.927712   1.049962  -1.836 0.066360
## LOCATION_DESCVIDEO STORE                 2.320783   1.070177   2.169 0.030113
##
## (Intercept)                            ***
## LOCATION_DESCATM
## LOCATION_DESCBANK
## LOCATION_DESCBAR/NIGHT CLUB            ***
## LOCATION_DESCBEAUTY/NAIL SALON         ***
## LOCATION_DESCCANDY STORE
## LOCATION_DESCCHAIN STORE
## LOCATION_DESCCHECK CASH
## LOCATION_DESCCLOTHING BOUTIQUE         .
## LOCATION_DESCCOMMERCIAL BLDG           .
## LOCATION_DESCDEPT STORE
## LOCATION_DESCDOCTOR/DENTIST
## LOCATION_DESCDRUG STORE
## LOCATION_DESCDRY CLEANER/LAUNDRY
## LOCATION_DESCFACTORY/WAREHOUSE
## LOCATION_DESCFAST FOOD                 .
## LOCATION_DESCGAS STATION
## LOCATION_DESCGROCERY/BODEGA
## LOCATION_DESCGYM/FITNESS FACILITY
## LOCATION_DESCHOSPITAL
## LOCATION_DESCHOTEL/MOTEL
## LOCATION_DESCJEWELRY STORE
## LOCATION_DESCLIQUOR STORE
## LOCATION_DESCLOAN COMPANY
## LOCATION_DESCMULTI DWELL - APT BUILD   **
## LOCATION_DESCMULTI DWELL - PUBLIC HOUS ***
## LOCATION_DESCNONE
## LOCATION_DESCPHOTO/COPY STORE
## LOCATION_DESCPVT HOUSE                 **
## LOCATION_DESCRESTAURANT/DINER          ***
## LOCATION_DESCSCHOOL
## LOCATION_DESCSHOE STORE                .
## LOCATION_DESCSMALL MERCHANT            **
## LOCATION_DESCSOCIAL CLUB/POLICY LOCATI ***
## LOCATION_DESCSTORAGE FACILITY
## LOCATION_DESCSTORE UNCLASSIFIED
## LOCATION_DESCSUPERMARKET
## LOCATION_DESCTELECOMM. STORE
## LOCATION_DESCVARIETY STORE             .
## LOCATION_DESCVIDEO STORE               *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 18783  on 13584  degrees of freedom
## Residual deviance: 18164  on 13545  degrees of freedom
##   (14977 observations deleted due to missingness)
```

```
## AIC: 18244
##
## Number of Fisher Scoring iterations: 13
```

What we can tell from the model:

Locations like bars, nightclubs, multi-dwelling residences (public housing and apartment buildings), and restaurants/diners are strong predictors of late-night shootings. These findings can help target safety interventions in these areas, especially during high-risk hours.

Locations such as ATMs, banks, and gas stations do not appear to significantly contribute to late-night shootings, likely because they are not typical late-night gathering spots.

The model reveals some limitations due to missing data, which should be addressed for future analysis.

Predicted Probabilities of Late-Night Shootings by Location

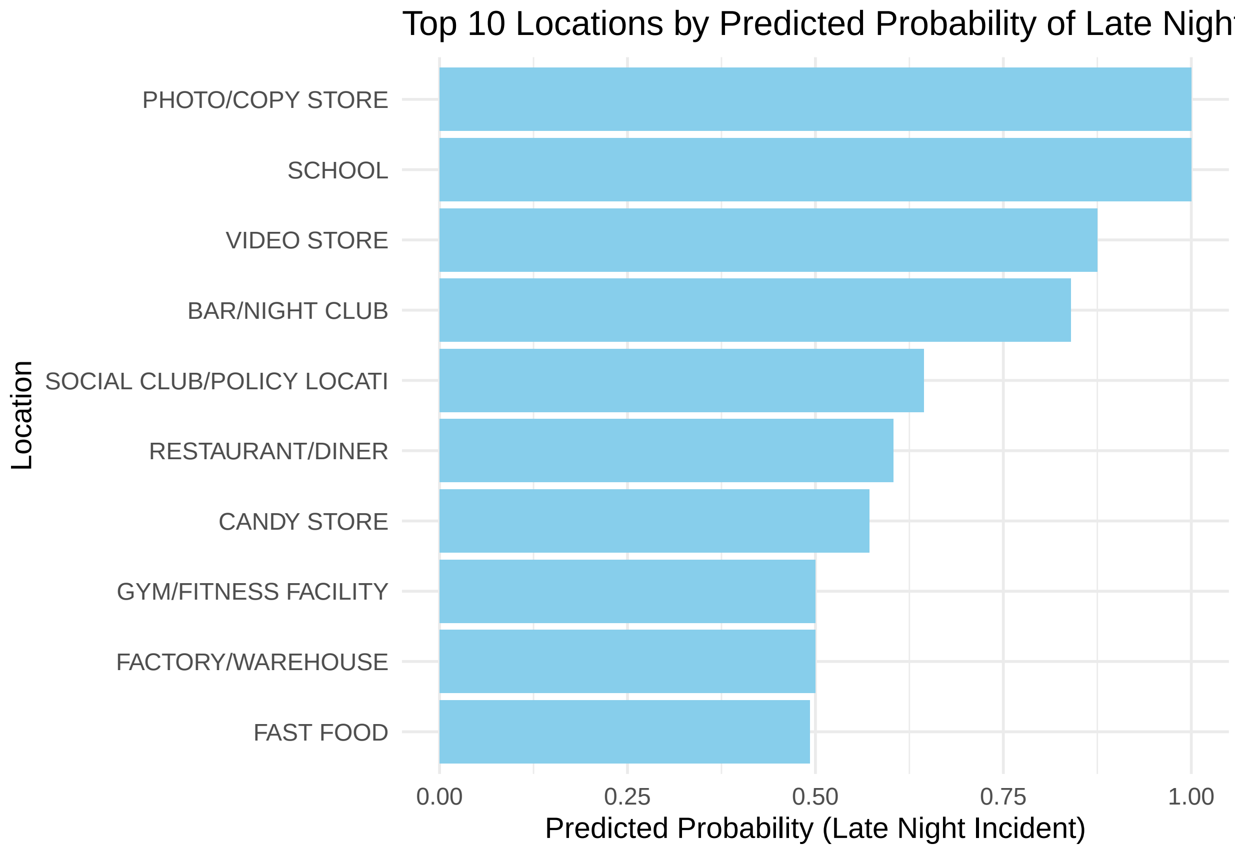
After fitting the logistic regression model, it's important to move beyond the coefficients to understand the practical impact of each location on the likelihood of late-night shootings. By predicting the probabilities of a late-night incident occurring at different locations, we can quantify the relative risk across various environments. This step helps identify the riskiest locations in practical terms, allowing us to pinpoint where late-night shootings are most likely to occur. Visualizing these probabilities gives us a clearer picture of which locations require targeted interventions based on their predicted likelihood of incidents.

```r
# Create a new dataset with the unique locations for prediction
locations <- unique(nypd_data_clean$LOCATION_DESC)
prediction_data <- data.frame(LOCATION_DESC = locations)

# Get predicted probabilities from the logistic regression model
prediction_data$predicted_prob <- predict(model, newdata = prediction_data, type = "response")

# Select the top 10 locations with the highest predicted probabilities
top_10_predictions <- prediction_data %>%
  arrange(desc(predicted_prob)) %>%
  top_n(10, predicted_prob)

# Plot predicted probabilities for the top 10 locations
ggplot(top_10_predictions, aes(x = reorder(LOCATION_DESC, predicted_prob), y = predicted_prob)) +
  geom_col(fill = "skyblue") +
  coord_flip() +
  labs(title = "Top 10 Locations by Predicted Probability of Late Night Shootings",
       x = "Location",
       y = "Predicted Probability (Late Night Incident)") +
  theme_minimal()
```



The chart highlights the Top 10 locations with the highest predicted probabilities of late-night shootings. PHOTO/COPY STORE, SCHOOL, and BAR/NIGHT CLUB show the highest risks, with probabilities close to 1.0. Locations like VIDEO STORE, SOCIAL CLUB, and RESTAURANT/DINER also have elevated risks, likely due to late-night activity. Some unexpected locations, like CANDY STORE and GYM, appear, which may suggest data irregularities or specific local contexts. These insights can help target safety interventions during high-risk times at key locations.

Conclusion:

This analysis of NYPD shooting incidents reveals key patterns in when and where shootings are most likely to occur. Late-night hours, especially on weekends, are high-risk times, and locations like multi-dwelling residences (public housing, apartments) and bars/nightclubs are the most common hotspots. Logistic regression provided further insights by predicting the probability of late-night shootings at different locations, helping identify areas for targeted interventions.

However, the analysis faces limitations due to missing data and limited predictor variables. Addressing these gaps in future studies will improve the accuracy and reliability of the findings.

Potential Biases and Limitations:

Missing Data: A significant number of incidents have missing location information (NA), which could introduce bias if certain locations are systematically underreported. Improving location data collection would enhance analysis.

Time Coverage: The dataset may lack broader time coverage or exclude seasonal trends, limiting the generalization of results. Expanding the timeframe could reveal deeper patterns.

Limited Predictors: The analysis focuses on time and location but excludes other factors like socioeconomic conditions. Adding more variables in future models would provide a fuller picture.

Model Assumptions: The logistic regression assumes a linear relationship, which may not fully capture the complexity of shooting incidents. Exploring non-linear models or interactions could improve the analysis.