NYPD Shooting Incident Data Report

SungHak Hong

2021 11 27

Project Step 1: Start an Rmd Document

```
#Import packages
library(tidyverse) #supports loading, filtering, and saving database
library(lubridate) #manages time and date data on database
library(ggplot2) #supports creating graphs
```

```
#Load Data
#The function "read_csv" imports CSV file to DataFrame format
df = read_csv("C:\\Users\\vmfl7\\Downloads\\NYPD_Shooting_Incident_Data__Historic_.csv")
```

```
## Rows: 23568 Columns: 19

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (10): OCCUR_DATE, BORO, LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX, PERP_R...
## dbl   (5): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, Latitude, Longitude
## 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.
```

```
#check if the load is completed
head(df)
```

```
## # A tibble: 6 x 19
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO          PRECINCT JURISDICTION_CODE
##          <dbl> <chr>      <time>     <chr>            <dbl>             <dbl>
## 1    201575314 08/23/2019 22:10      QUEENS             103                 0
## 2    205748546 11/27/2019 15:54      BRONX               40                 0
## 3    193118596 02/02/2019 19:40      MANHATTAN           23                 0
## 4    204192600 10/24/2019 00:52      STATEN ISLAND      121                 0
## 5    201483468 08/22/2019 18:03      BRONX               46                 0
## 6    198255460 06/07/2019 17:50      BROOKLYN            73                 0
## # ... with 13 more variables: LOCATION_DESC <chr>,
## #   STATISTICAL_MURDER_FLAG <lgl>, PERP_AGE_GROUP <chr>, PERP_SEX <chr>,
## #   PERP_RACE <chr>, VIC_AGE_GROUP <chr>, VIC_SEX <chr>, VIC_RACE <chr>,
## #   X_COORD_CD <dbl>, Y_COORD_CD <dbl>, Latitude <dbl>, Longitude <dbl>,
## #   Lon_Lat <chr>
```

Project Step 2: Tidy and Transform Your Data

```
#select columns what I will use in this project
df2 = df %>% select(INCIDENT_KEY, OCCUR_DATE, OCCUR_TIME, BORO, STATISTICAL_MURDER_FLAG,
                    PERP_AGE_GROUP, PERP_SEX, PERP_RACE, VIC_AGE_GROUP, VIC_SEX, VIC_RACE)
#convert blank values and "U" as "UNKNOWN"
df2 = df2%>%
  replace_na(list(PERP_AGE_GROUP = "UNKNOWN", PERP_SEX = "UNKNOWN", PERP_RACE = "UNKNOWN"))
df2$PERP_SEX = recode(df2$PERP_SEX, U = "UNKNOWN")
df2$VIC_SEX   = recode(df2$VIC_SEX, U = "UNKNOWN")

#Remove unrealistic values on perpetrator age group
df2 = subset(df2, PERP_AGE_GROUP == "<18" | PERP_AGE_GROUP == "18-24" |
               PERP_AGE_GROUP == "25-44" | PERP_AGE_GROUP == "45-64" |
               PERP_AGE_GROUP == "65+" | PERP_AGE_GROUP == "UNKNOWN")

# Change data type of INCIDENT_KEY to character, and others to factor
df2$INCIDENT_KEY = as.character(df2$INCIDENT_KEY)
df2$BORO = as.factor(df2$BORO)
df2$PERP_AGE_GROUP = as.factor(df2$PERP_AGE_GROUP)
df2$PERP_SEX = as.factor(df2$PERP_SEX)
df2$PERP_RACE = as.factor(df2$PERP_RACE)
df2$VIC_AGE_GROUP = as.factor(df2$VIC_AGE_GROUP)
df2$VIC_SEX = as.factor(df2$VIC_SEX)
df2$VIC_RACE = as.factor(df2$VIC_RACE)


#Show summary of the cleaned data
summary(df2)
```

```
##  INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME                    BORO
##  Length:23565       Length:23565       Length:23565      BRONX        :6698
##  Class :character   Class :character   Class1:hms        BROOKLYN     :9721
##  Mode  :character   Mode  :character   Class2:difftime   MANHATTAN    :2921
##                                        Mode  :numeric    QUEENS       :3527
##                                                          STATEN ISLAND: 698
##
##
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP     PERP_SEX
##  Mode :logical           <18    : 1354   F      :  334
##  FALSE:19077             18-24  : 5448   M      :13302
##  TRUE :4488              25-44  : 4613   UNKNOWN: 9929
##                          45-64  :  481
##                          65+    :   54
##                          UNKNOWN:11615
##
##                              PERP_RACE     VIC_AGE_GROUP     VIC_SEX
##  AMERICAN INDIAN/ALASKAN NATIVE:    2   <18    : 2525   F      : 2195
##  ASIAN / PACIFIC ISLANDER      :  120   18-24  : 8999   M      :21350
##  BLACK                         : 9854   25-44  :10285   UNKNOWN:   20
##  BLACK HISPANIC                : 1081   45-64  : 1536
##  UNKNOWN                       :10294   65+    :  155
##  WHITE                         :  255   UNKNOWN:   65
##  WHITE HISPANIC                : 1959  
##                           VIC_RACE    
##  AMERICAN INDIAN/ALASKAN NATIVE:    9  
##  ASIAN / PACIFIC ISLANDER      :  320  
##  BLACK                         :16845  
##  BLACK HISPANIC                : 2244  
##  UNKNOWN                       :  102  
##  WHITE                         :  615  
##  WHITE HISPANIC                : 3430  
```

Project Step 3: Add Vidualizations and Analysis

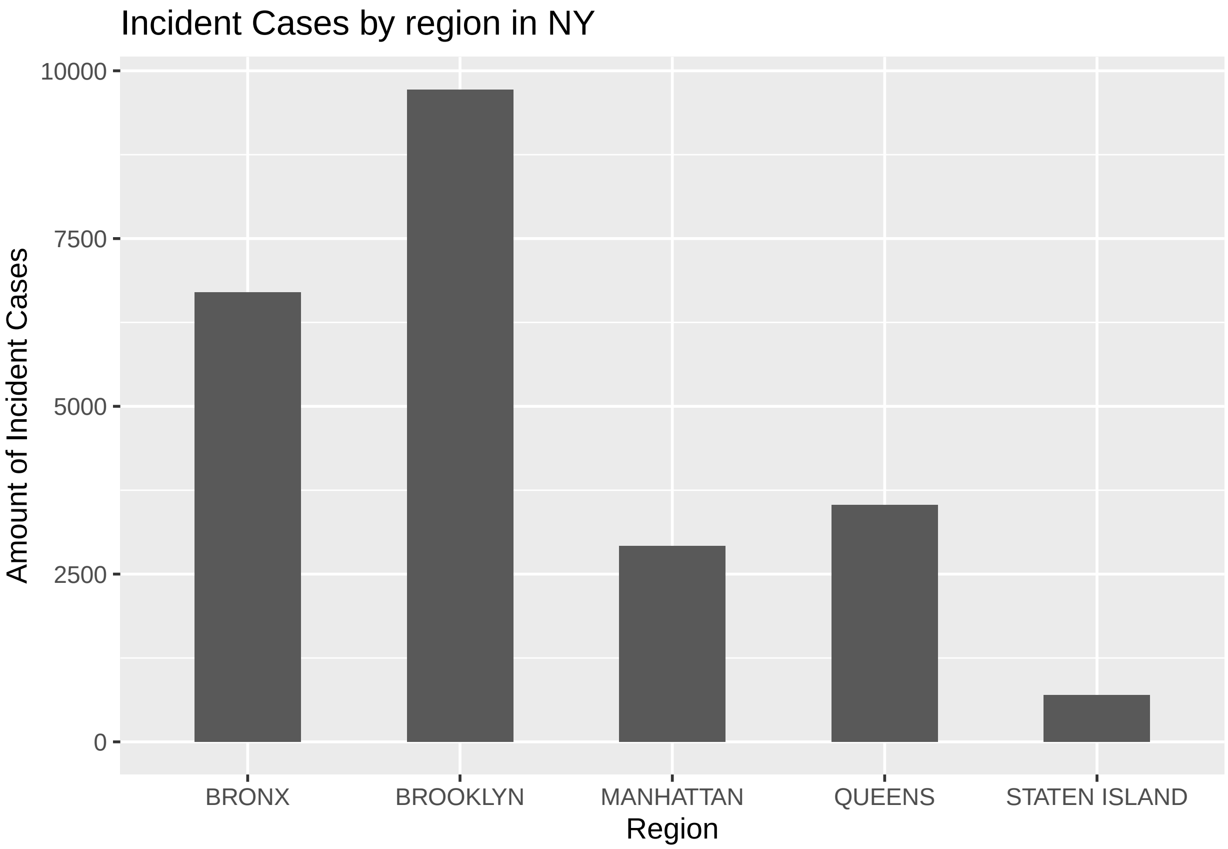
1. Place: which region in New York has the most incident cases?

The table and histogram below show the number of cases in 5 regions including: Bronx, Brooklyn, Manhattan, Queens, and Staten Island.

```
summary(df2$BORO)
```

```
##          BRONX      BROOKLYN     MANHATTAN        QUEENS STATEN ISLAND
##           6698          9721          2921          3527           698
```

```
g1 <- ggplot(df2, aes(x = BORO)) +
  geom_bar(width=0.5) +
  labs(title = "Incident Cases by region in NY", x = "Region",
       y = "Amount of Incident Cases")

g1
```

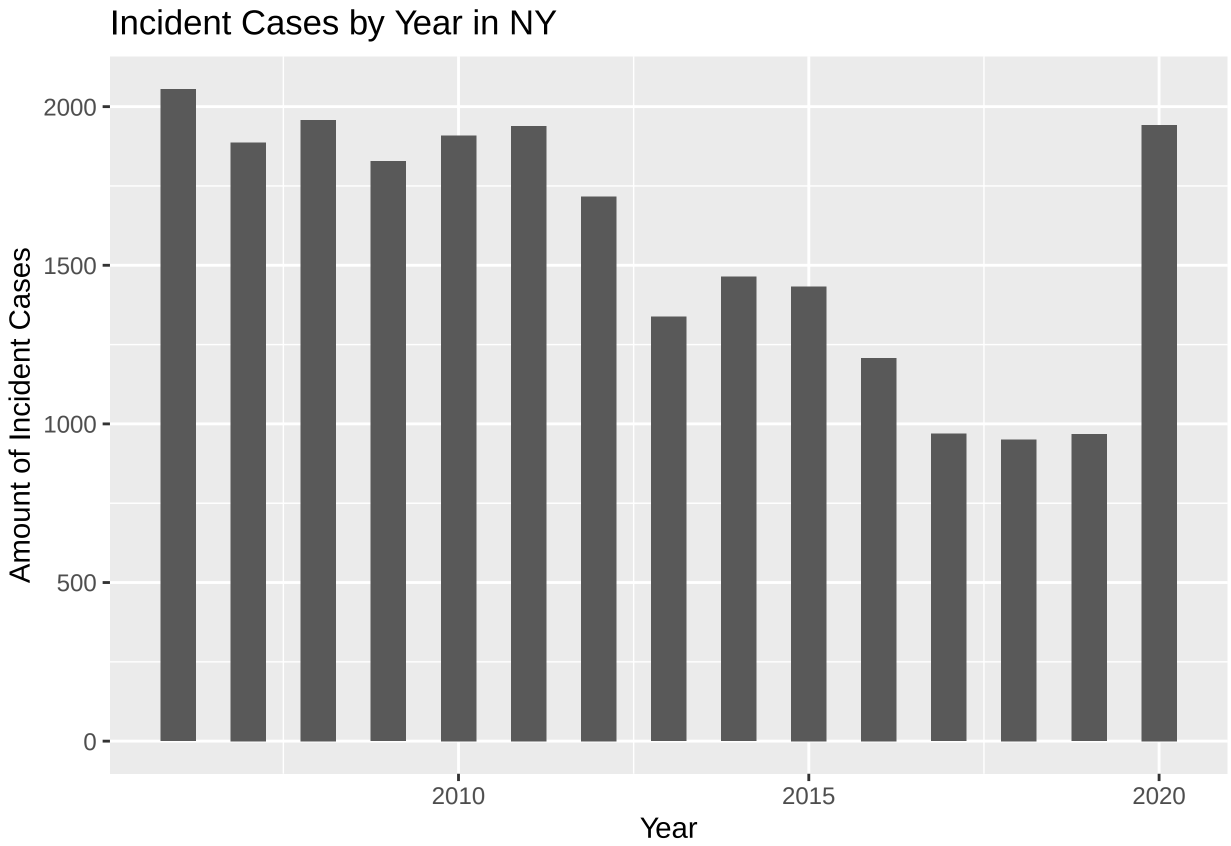


- Shooting incidents happened at Brooklyn the most followed by Bronx, Queens, Manhattan, and Staten Island, respectively.

2. Time: What time do incident cases occur the most?

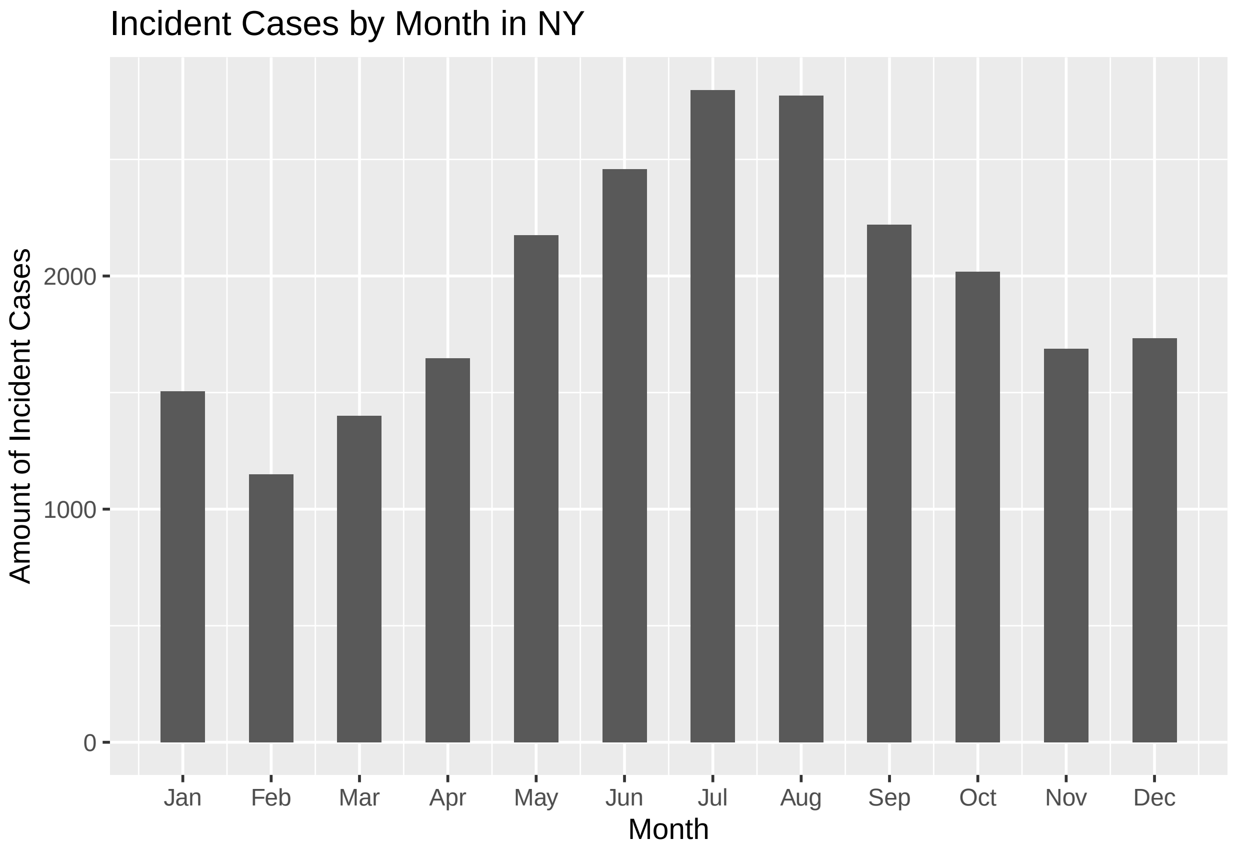
The below histograms and line graph show the number of cases by specific time period: years, months, days, and hours.

```r
df2$OCCUR_mdy = mdy(df2$OCCUR_DATE)
df2$OCCUR_YEAR = year(df2$OCCUR_mdy)
df2$OCCUR_MONTH = month(df2$OCCUR_mdy)
df2$OCCUR_DAY = wday(df2$OCCUR_mdy)
df2$OCCUR_HOUR = hour(df2$OCCUR_TIME)

#Year
g2 <- ggplot(df2, aes(x = OCCUR_YEAR)) +
  geom_bar(width=0.5) +
  labs(title = "Incident Cases by Year in NY", x = "Year", y = "Amount of Incident Cases")

g2
```



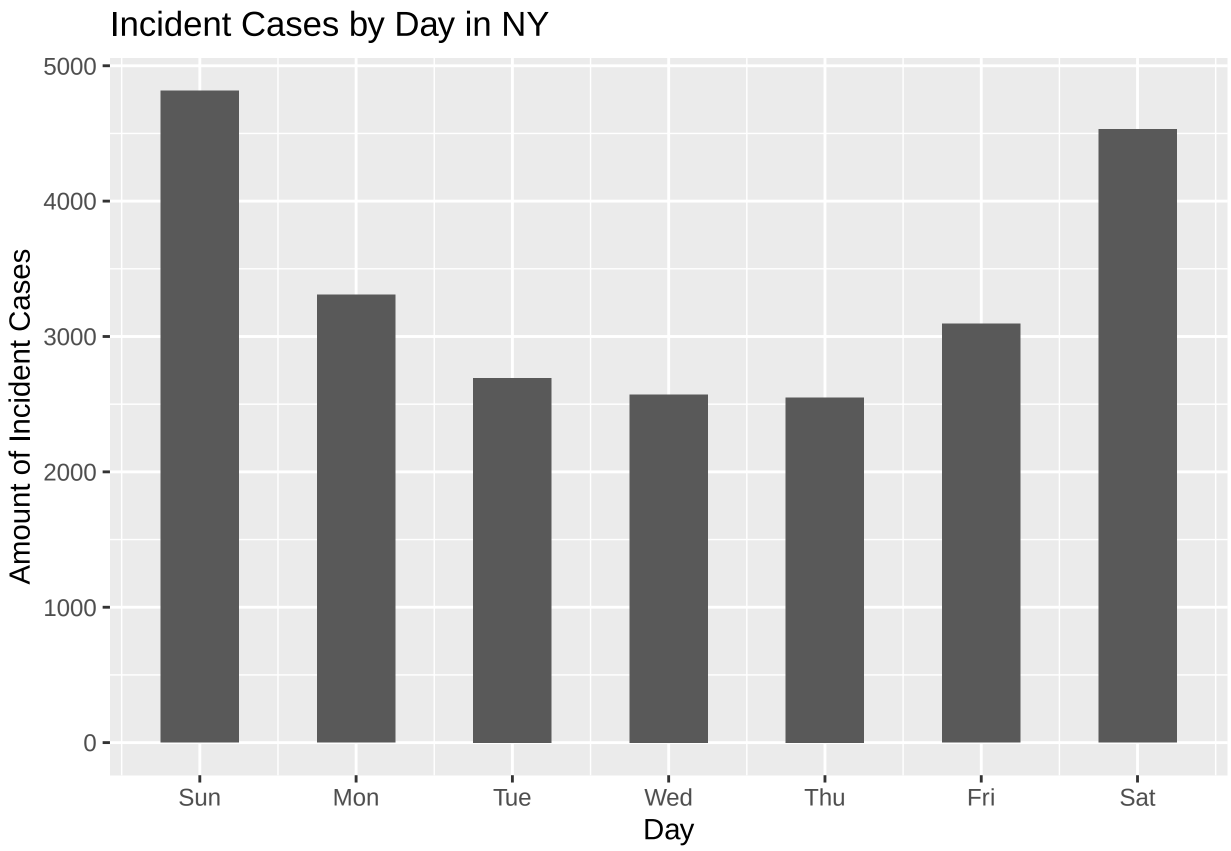
- From 2006 through 2019, the incident decreased continuously.
- In 2020, the cases increased significantly (almost doubled from 2019).

```r
g3 <- ggplot(df2, aes(x = OCCUR_MONTH)) +
  geom_bar(width=0.5) +
  labs(title = "Incident Cases by Month in NY", x = "Month", y = "Amount of Incident Cases")
g3 <- g3 + scale_x_continuous(breaks=1:12,
  labels=c("Jan","Feb","Mar","Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov", "Dec"))

g3
```



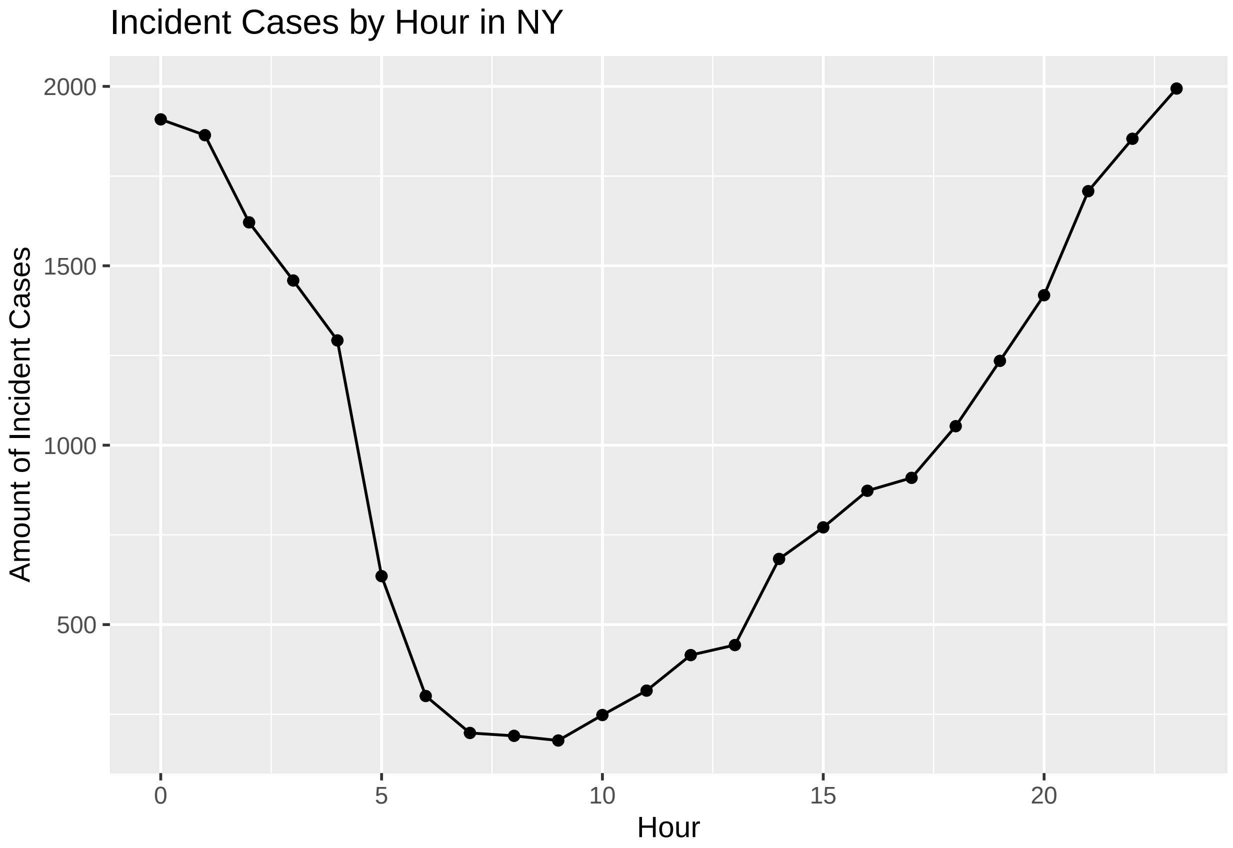
- The incident cases occurred the most in July and the least in February.
- There were more shooting incidents during summer (June, July, August) than any other seasons.

```r
g4 <- ggplot(df2, aes(x = OCCUR_DAY)) +
  geom_bar(width=0.5) +
  labs(title = "Incident Cases by Day in NY", x = "Day", y = "Amount of Incident Cases")

g4 <- g4 + scale_x_continuous(breaks=1:7,
  labels=c("Sun", "Mon","Tue","Wed","Thu","Fri","Sat"))


g4
```



- The incident cases occurred the most on weekends(Sunday, Saturday).

```
df3 = df2 %>%
  group_by(OCCUR_HOUR) %>%
  count()

g5 <- ggplot(df3, aes(x = OCCUR_HOUR, y = n)) +
  geom_line() +
  geom_point() +
  labs(title = "Incident Cases by Hour in NY", x = "Hour", y = "Amount of Incident Cases")

g5
```



- Shooting occurred the most at 23pm, and the least at 9 am.
- Shooting occurred more at night than at daytime.

Project Step 4: Add Bias Identification

I have never been to New York City, and therefore, all I've heard about New York City's public safety is from news and friends who live there. Before doing this project, I thought that Manhattan is the most dangerous place in New York City. However, the data show that the shooting incidents occurred in Brooklyn the most. Another interesting point I realized from the data is that the number of shooting incidents had gradually decreased from 2006 through 2019 and suddenly increased significantly in 2020. I think the increase might have been caused by the COVID-19 pandemic. However, this assumption cannot be supported only by the data used for this project and requires further research and analyses.

Reference

https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic
https://www.r-graph-gallery.com/index.html