Project for Week 4

Ryan Ordonez

2023-10-08

This project is to identify certain boroughs of New York to impliment after school community centers to help reduce crime in the most incident prone areas for our youth.

Load Packages

```
library(tidyverse)
library(reshape2)
library(pROC)
```

Import Data

```
# breaking down the url so it doesn't run off the page of the pdf doc
url_part1 <- "https://data.cityofnewyork.us/api/views/833y-fsy8/"
url_part2 <- "rows.csv?accessType=DOWNLOAD"
NYPD_shootings <- read.csv(paste0(url_part1, url_part2))
```

This data is pulled from the Data.gov website. It lists every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year. A link to the .csv file is here: https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD

Inspect the Data

```
head(NYPD_shootings)
```

```
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME     BORO LOC_OF_OCCUR_DESC PRECINCT
## 1    228798151 05/27/2021   21:30:00   QUEENS                       105
## 2    137471050 06/27/2014   17:40:00    BRONX                        40
## 3    147998800 11/21/2015   03:56:00   QUEENS                       108
## 4    146837977 10/09/2015   18:30:00    BRONX                        44
## 5     58921844 02/19/2009   22:58:00    BRONX                        47
## 6    219559682 10/21/2020   21:36:00 BROOKLYN                        81
##   JURISDICTION_CODE LOC_CLASSFCTN_DESC LOCATION_DESC STATISTICAL_MURDER_FLAG
## 1                 0                                                   false
## 2                 0                                                   false
## 3                 0                                                    true
```

```
## 4                 0                                                  false
## 5                 0                                                   true
## 6                 0                                                   true
##   PERP_AGE_GROUP PERP_SEX PERP_RACE VIC_AGE_GROUP VIC_SEX       VIC_RACE
## 1                                          18-24       M          BLACK
## 2                                          18-24       M          BLACK
## 3                                          25-44       M          WHITE
## 4                                            <18       M WHITE HISPANIC
## 5          25-44        M     BLACK         45-64       M          BLACK
## 6                                          25-44       M          BLACK
##   X_COORD_CD Y_COORD_CD Latitude Longitude
## 1    1058925   180924.0 40.66296 -73.73084
## 2    1005028   234516.0 40.81035 -73.92494
## 3    1007668   209836.5 40.74261 -73.91549
## 4    1006537   244511.1 40.83778 -73.91946
## 5    1024922   262189.4 40.88624 -73.85291
## 6    1004234   186461.7 40.67846 -73.92795
##                                     Lon_Lat
## 1 POINT (-73.73083868899994 40.662964620000025)
## 2  POINT (-73.92494232599995 40.81035186300006)
## 3  POINT (-73.91549174199997 40.74260663300004)
## 4  POINT (-73.91945661499994 40.83778200300003)
## 5  POINT (-73.85290950899997 40.88623791800006)
## 6 POINT (-73.92795224099996 40.678456718000064)
```

Tidy and Transform the data

```r
# Convert appropriate variables to factor types
NYPD_shootings$BORO = as.factor(NYPD_shootings$BORO)
NYPD_shootings$LOC_OF_OCCUR_DESC = as.factor(NYPD_shootings$LOC_OF_OCCUR_DESC)
NYPD_shootings$PRECINCT = as.factor(NYPD_shootings$PRECINCT)
NYPD_shootings$JURISDICTION_CODE = as.factor(NYPD_shootings$JURISDICTION_CODE)
NYPD_shootings$LOC_CLASSFCTN_DESC = as.factor(NYPD_shootings$LOC_CLASSFCTN_DESC)
NYPD_shootings$STATISTICAL_MURDER_FLAG = as.factor(NYPD_shootings$STATISTICAL_MURDER_FLAG)
NYPD_shootings$PERP_AGE_GROUP = as.factor(NYPD_shootings$PERP_AGE_GROUP)
NYPD_shootings$PERP_SEX = as.factor(NYPD_shootings$PERP_SEX)
NYPD_shootings$PERP_RACE = as.factor(NYPD_shootings$PERP_RACE)
NYPD_shootings$VIC_AGE_GROUP = as.factor(NYPD_shootings$VIC_AGE_GROUP)
NYPD_shootings$VIC_SEX = as.factor(NYPD_shootings$VIC_SEX)
NYPD_shootings$VIC_RACE = as.factor(NYPD_shootings$VIC_RACE)
# Remove unnecessary columns
NYPD_shootings$Lon_Lat = NULL
NYPD_shootings$X_COORD_CD = NULL
NYPD_shootings$Y_COORD_CD = NULL
NYPD_shootings$JURISDICTION_CODE = NULL
NYPD_shootings$LOC_CLASSFCTN_DESC = NULL
NYPD_shootings$LOCATION_DESC = NULL
NYPD_shootings$STATISTICAL_MURDER_FLAG = NULL
NYPD_shootings$Latitude = NULL
NYPD_shootings$Longitude = NULL
NYPD_shootings$LOC_OF_OCCUR_DESC = NULL
```

```r
# Mark null, unknown and empty entries as NA for easy processing
NYPD_shootings$PERP_AGE_GROUP[NYPD_shootings$PERP_AGE_GROUP %in%
                                c("", "UNKNOWN", "(null)")] <- NA
```

```r
# Keep only specific age groups
valid_age_groups <- c("<18", "18-24", "25-44", "45-64", "65+", NA)
NYPD_shootings <- NYPD_shootings[NYPD_shootings$PERP_AGE_GROUP %in% valid_age_groups, ]
```

```r
summary(NYPD_shootings)
```

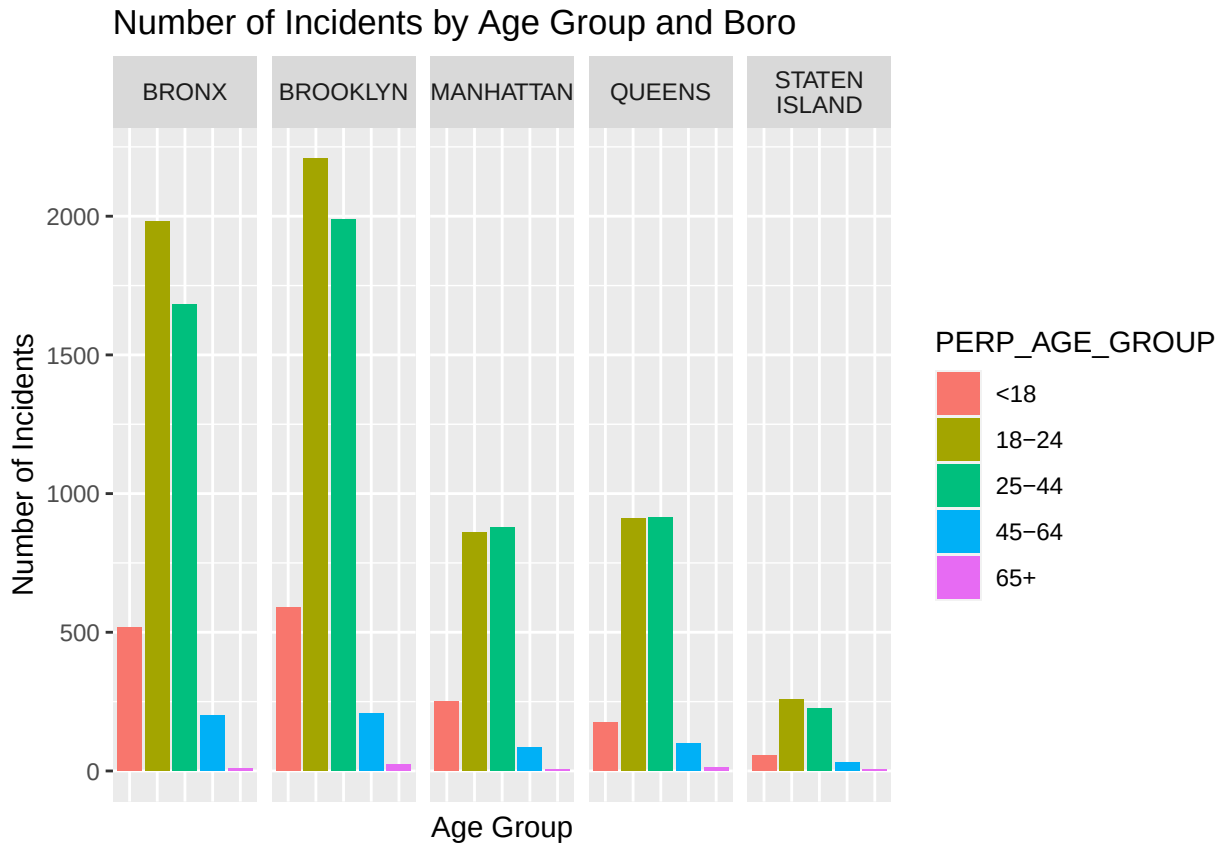
```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME
##  Min.   :  9953245   Length:27309       Length:27309
##  1st Qu.: 63859989   Class :character   Class :character
##  Median : 90374290   Mode  :character   Mode  :character
##  Mean   :120862700
##  3rd Qu.:188810231
##  Max.   :261190187
##
##             BORO          PRECINCT     PERP_AGE_GROUP    PERP_SEX
##  BRONX        : 7935   75     : 1557   18-24  : 6222         : 9310
##  BROOKLYN     :10932   73     : 1452   25-44  : 5687   (null):  640
##  MANHATTAN    : 3572   67     : 1216   <18    : 1591   F     :  424
##  QUEENS       : 4094   44     : 1020   45-64  :  617   M     :15436
##  STATEN ISLAND:  776   79     : 1012   65+    :   60   U     : 1499
##                        47     :  952   (Other):    0
##                        (Other):20100   NA's   :13132
##           PERP_RACE     VIC_AGE_GROUP   VIC_SEX
##  BLACK         :11431   <18    : 2839   F: 2615
##                : 9310   1022   :    1   M:24683
##  WHITE HISPANIC: 2339   18-24  :10085   U:   11
##  UNKNOWN       : 1836   25-44  :12279
##  BLACK HISPANIC: 1314   45-64  : 1863
##  (null)        :  640   65+    :  181
##  (Other)       :  439   UNKNOWN:   61
##                             VIC_RACE
##  AMERICAN INDIAN/ALASKAN NATIVE:   10
##  ASIAN / PACIFIC ISLANDER      :  404
##  BLACK                         :19438
##  BLACK HISPANIC                : 2646
##  UNKNOWN                       :   66
##  WHITE                         :  698
##  WHITE HISPANIC                : 4047
```

Visual Analysis

```r
# Filter out NA values from the dataset
NYPD_shootings_filtered <- NYPD_shootings %>% filter(!is.na(PERP_AGE_GROUP))

# Wrap the BORO column text for better display
NYPD_shootings_filtered$BORO <- str_wrap(NYPD_shootings_filtered$BORO, width = 10)
```

```r
# Create a bar chart
ggplot(NYPD_shootings_filtered, aes(x = PERP_AGE_GROUP)) +
  geom_bar(aes(fill = PERP_AGE_GROUP), position = "dodge") +
  ggtitle("Number of Incidents by Age Group and Boro") +
  xlab("Age Group") +
  ylab("Number of Incidents") +
  facet_wrap(~ BORO, ncol = length(unique(NYPD_shootings_filtered$BORO))) +
  theme(
    axis.text.x = element_blank(),  # Remove x-axis text
    axis.ticks.x = element_blank()  # Remove x-axis ticks
  )
```



The above chart identifies that most incidents occur in the Bronx and Brooklyn boroughs by perpetrators in the two age groups of 18-24 and 25-44. I will now dive into the hours in which most of these incidents occur within these boroughs.

Time-Based Analysis for Bronx and Brooklyn

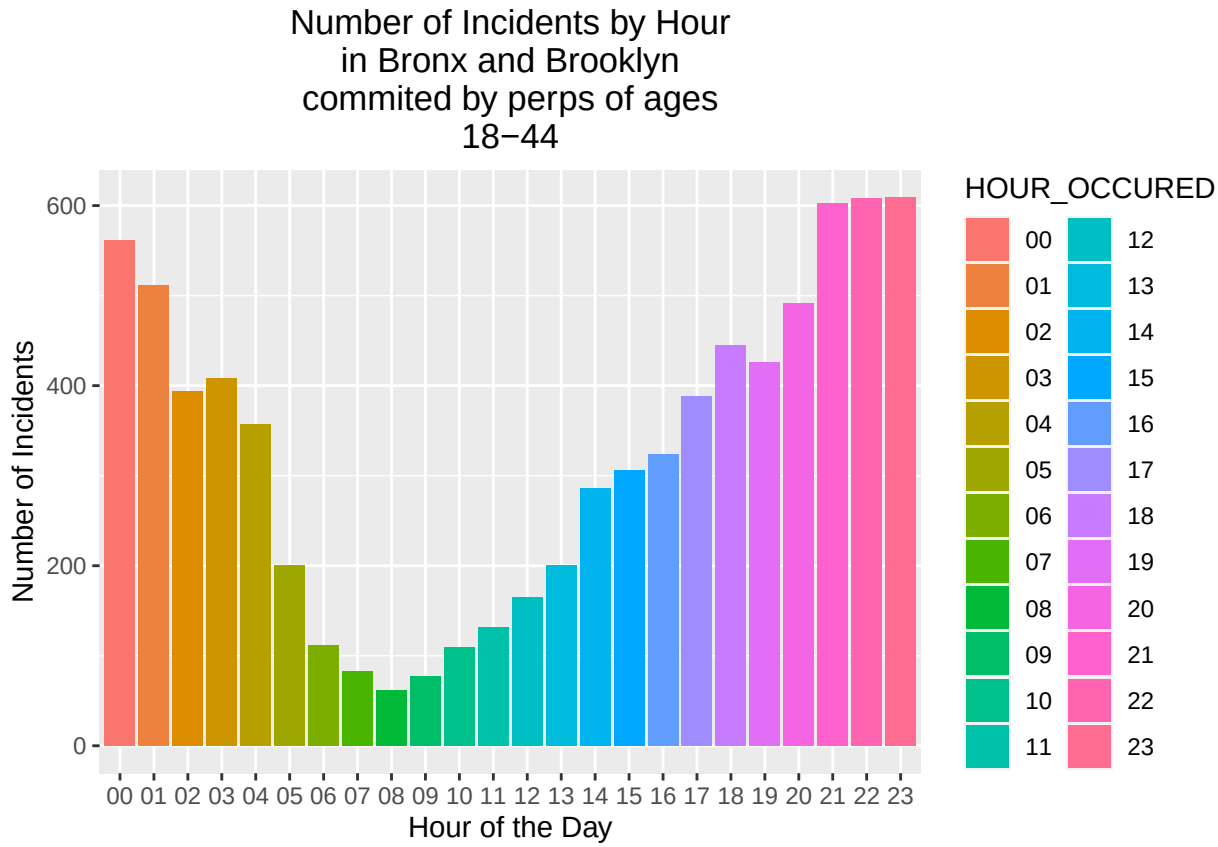
```r
# Filter data for Bronx and Brooklyn
NYPD_Bronx_Brooklyn <- NYPD_shootings %>% filter(BORO %in% c("BRONX", "BROOKLYN"))
```

```r
# Extract hour information (this assumes 'TIME_OCCURED' is a POSIX time object)
NYPD_Bronx_Brooklyn$HOUR_OCCURED <- format(as.POSIXct(NYPD_Bronx_Brooklyn$OCCUR_TIME, format="%H:%M:%S"
```

```r
# Filter the data for the age groups 18-24 and 25-44
filtered_data <- NYPD_Bronx_Brooklyn %>%
  filter(PERP_AGE_GROUP %in% c('18-24', '25-44'))

# Prepare the title with wrapped text
wrapped_title <- strwrap("Number of Incidents by Hour in Bronx and Brooklyn commited by perps of ages 18
                         width = 30) %>%
  paste(collapse = "\n")

# Create a bar chart for incidents by time of day in Bronx and Brooklyn
ggplot(filtered_data, aes(x = HOUR_OCCURED)) +
  geom_bar(aes(fill = HOUR_OCCURED), position = "dodge") +
  ggtitle(wrapped_title) +
  theme(plot.title = element_text(hjust = 0.5)) +  # Center the title
  xlab("Hour of the Day") +
  ylab("Number of Incidents")
```



This chart shows that incidents by these two age groups in these two boroughs occur mostly from 1400 (2pm) to 0400 (4am). An after school facility in these areas may help reduce the amount of crimes committed in these areas between these times. Some questions to take into consideration is what hours should the facility be open, what type of activities would it provide to reduce crime, how can we promote the facility, and how can we provide transportation to and from the facility for our youth?.

Modeling

```
selected_data <- filtered_data %>%
  select(HOUR_OCCURED, BORO, PERP_AGE_GROUP)

selected_data$HOUR_OCCURED <- as.numeric(selected_data$HOUR_OCCURED)
selected_data$BORO <- as.factor(selected_data$BORO)
selected_data$PERP_AGE_GROUP <- as.factor(selected_data$PERP_AGE_GROUP)

# Linear regression
model <- glm(HOUR_OCCURED ~ BORO + PERP_AGE_GROUP, data = selected_data, family = "gaussian")

summary(model)
```

```
##
## Call:
## glm(formula = HOUR_OCCURED ~ BORO + PERP_AGE_GROUP, family = "gaussian",
##     data = selected_data)
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           13.1079     0.1587  82.589  < 2e-16 ***
## BOROBROOKLYN           0.2352     0.1839   1.279    0.201
## PERP_AGE_GROUP25-44   -0.8565     0.1839  -4.657 3.26e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 66.17712)
##
##     Null deviance: 521618  on 7861  degrees of freedom
## Residual deviance: 520086  on 7859  degrees of freedom
## AIC: 55277
##
## Number of Fisher Scoring iterations: 2
```

Based on the GLM results, one key finding is the significant negative coefficient for the perpetrator age group 25-44, which suggests that crimes committed by individuals in this older age group occur, on average, 0.86 hours earlier than those in the reference age group. This could be interpreted as older individuals being involved in crime earlier in the day, leaving room to infer that younger individuals (such as ages 18-24, if they are the reference group) are more likely to be involved in crimes that occur later in the day.

Moreover, while the coefficient for the Brooklyn borough is not statistically significant, it does indicate a positive effect on the hour crimes occur, suggesting crimes happen slightly later in Brooklyn compared to the reference borough. When combining these insights, it's reasonable to propose that after-school activities targeting youth ages 18-24 in Brooklyn could potentially fill the time gap where they're more likely to engage in criminal activities, thus reducing crime rates in this particular demographic and area.

Conclusion

The data clearly shows trends in the age groups of the perpetrators. Specifically, the 18-24 year-old group leads in the number of incidents, and this is most noticeable in Brooklyn borough.

Some questions that arise from this data include the differences in policing policies in each borough. There's also a concern about the large number of unreported or missing age groups that could potentially skew our understanding of incidents by age group.

In terms of bias, how data is reported can differ from precinct to precinct. Additionally, the unique policies of each precinct could introduce underlying biases, especially when incidents by age group have historically been treated differently.

On a personal note, I place a high degree of trust in data. While I have tried to address missing or erroneous entries, there's always the risk of the data being off or misleading. To mitigate this, I've focused on maintaining the integrity of the majority of correct data while considering the impact of erroneous entries. I also plan on mitigating my own bias towards quantitative data by corroborating these findings with other data sources and incorporating them into a single, comprehensive report.