NYPD INCIDENT REPORT

H A

-|-|-|-|-|-|-|-|-|-|-|Step 1|-|-|-|-|-|-|-|-|-|-|-

Introduction

This project details every shooting incident in NYC from 2006 to 2021, reviewed by the NYPD's Office of Management Analysis and Planning. The data source is https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic (https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic)

Research Question

Which Boro has the highest number of shooting incidents to allocate more resource?

Installing libraries

```
library(tidyverse)
```

```
## Warning: package 'stringr' was built under R version 4.3.3
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.4.4     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to be
come errors
```

```
library(readr)
library(lubridate)
```

###import data from url https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD (https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD)

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

```
## 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.
```

-|-|-|-|-|-|-|-|-|-|-|Step 2|-|-|-|-|-|-|-|-|-|-|-

Tidy and Transformation

```
# read top 5 rows
head(df,5)
```

```
## # A tibble: 5 × 21
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO      LOC_OF_OCCUR_DESC PRECINCT
##          <dbl> <chr>      <time>     <chr>     <chr>                <dbl>
## 1    244608249 05/05/2022 00:10      MANHATTAN INSIDE                  14
## 2    247542571 07/04/2022 22:20      BRONX     OUTSIDE                 48
## 3     84967535 05/27/2012 19:35      QUEENS    <NA>                   103
## 4    202853370 09/24/2019 21:00      BRONX     <NA>                    42
## 5     27078636 02/25/2007 21:00      BROOKLYN  <NA>                    83
## # ℹ 15 more variables: JURISDICTION_CODE <dbl>, LOC_CLASSFCTN_DESC <chr>,
## #   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>
```

```
# read last 5 rows
tail(df,5)
```

```
## # A tibble: 5 × 21
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO      LOC_OF_OCCUR_DESC PRECINCT
##          <dbl> <chr>      <time>     <chr>     <chr>                <dbl>
## 1    265354835 03/19/2023 23:48      BRONX     INSIDE                  47
## 2    272968931 08/16/2023 02:46      BRONX     OUTSIDE                 41
## 3    270489846 06/27/2023 12:27      BRONX     INSIDE                  41
## 4    271021661 07/08/2023 11:27      QUEENS    OUTSIDE                102
## 5    271818283 07/24/2023 23:38      MANHATTAN OUTSIDE                 28
## # ℹ 15 more variables: JURISDICTION_CODE <dbl>, LOC_CLASSFCTN_DESC <chr>,
## #   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>
```

Add to your Rmd document a summary of the data and clean up your dataset by changing appropriate variables to factor and date types and getting rid of any columns not needed. Show the summary of your data to be sure there is no missing data. If there is missing data, describe how you plan to handle it.

Data Exploration

```
# list column names
names(df)
```

```
##  [1] "INCIDENT_KEY"           "OCCUR_DATE"
##  [3] "OCCUR_TIME"             "BORO"
##  [5] "LOC_OF_OCCUR_DESC"      "PRECINCT"
##  [7] "JURISDICTION_CODE"      "LOC_CLASSFCTN_DESC"
##  [9] "LOCATION_DESC"          "STATISTICAL_MURDER_FLAG"
## [11] "PERP_AGE_GROUP"         "PERP_SEX"
## [13] "PERP_RACE"              "VIC_AGE_GROUP"
## [15] "VIC_SEX"                "VIC_RACE"
## [17] "X_COORD_CD"             "Y_COORD_CD"
## [19] "Latitude"               "Longitude"
## [21] "Lon_Lat"
```

```
# number of columns in dataset
ncol(df)
```

```
## [1] 21
```

Converting data types

```
library(dplyr)
df <- df %>%
  mutate(INCIDENT_KEY = as.factor(INCIDENT_KEY),
         OCCUR_DATE = as.Date(OCCUR_DATE, format = "%Y-%m-%d"),  # Specify format
         OCCUR_TIME = as.POSIXct(OCCUR_TIME, format = ""),
         BORO = as.factor(BORO),
         PRECINCT = as.factor(PRECINCT),
         STATISTICAL_MURDER_FLAG = as.logical(STATISTICAL_MURDER_FLAG),
         VIC_AGE_GROUP = as.factor(VIC_AGE_GROUP),
         VIC_SEX = as.factor(VIC_SEX),
         VIC_RACE = as.factor(VIC_RACE))
```

Summary of data

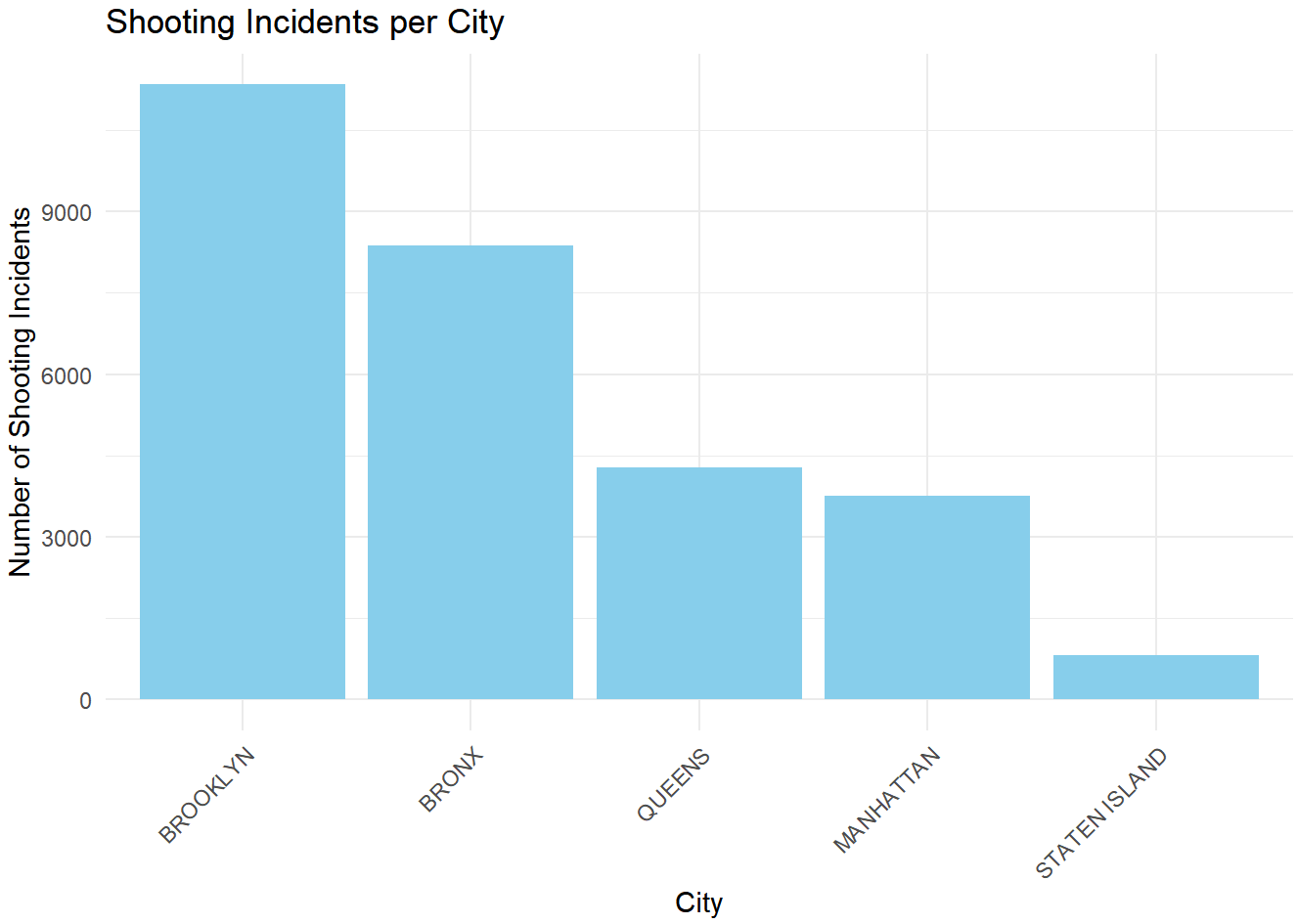
```
summary(df)
```

```
##     INCIDENT_KEY     OCCUR_DATE      OCCUR_TIME
##  173354054:   18   Min.   :NA      Min.   :1970-01-01 00:00:00.0000
##  263503175:   16   1st Qu.:NA      1st Qu.:1970-01-01 03:30:00.0000
##  23749375 :   12   Median :NA      Median :1970-01-01 15:15:00.0000
##  24717013 :   12   Mean   :NaN     Mean   :1970-01-01 12:44:16.7131
##  33478089 :   12   3rd Qu.:NA      3rd Qu.:1970-01-01 20:45:00.0000
##  33706902 :   12   Max.   :NA      Max.   :1970-01-01 23:59:00.0000
##  (Other)  :28480   NA's   :28562
##             BORO          PRECINCT     STATISTICAL_MURDER_FLAG VIC_AGE_GROUP
##  BRONX        : 8376   75     : 1628   Mode :logical           <18    : 2954
##  BROOKLYN     :11346   73     : 1500   FALSE:23036             1022   :    1
##  MANHATTAN    : 3762   67     : 1259   TRUE :5526              18-24  :10384
##  QUEENS       : 4271   44     : 1076                           25-44  :12973
##  STATEN ISLAND:  807   79     : 1045                           45-64  : 1981
##                        47     : 1006                           65+    :  205
##                        (Other):21048                           UNKNOWN:   64
##  VIC_SEX                             VIC_RACE
##  F: 2760   AMERICAN INDIAN/ALASKAN NATIVE:   11
##  M:25790   ASIAN / PACIFIC ISLANDER      :  440
##  U:   12   BLACK                         :20235
##            BLACK HISPANIC                : 2795
##            UNKNOWN                       :   70
##            WHITE                         :  728
##            WHITE HISPANIC                : 4283
```

-|-|-|-|-|-|-|-|-|-|-|- Step 3 -|-|-|-|-|-|-|-|-|-|-|-

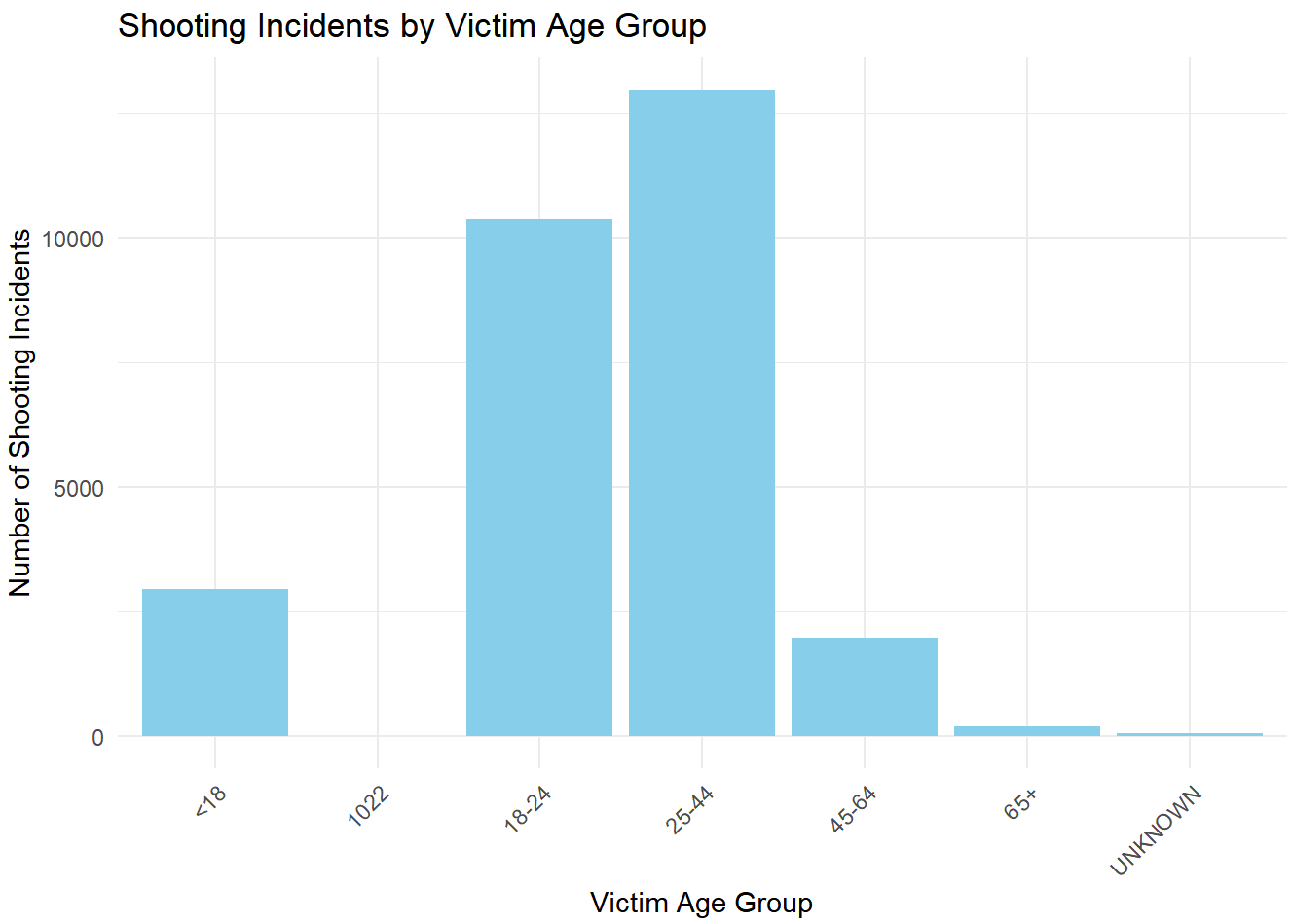
```
library(dplyr)
library(ggplot2)

# Convert OCCUR_DATE to Date type
df$OCCUR_DATE <- as.Date(df$OCCUR_DATE, format = "%Y-%m-%d")

# Extract year from OCCUR_DATE
df$YEAR <- lubridate::year(df$OCCUR_DATE)

# Count shooting incidents per city
city_shootings <- df %>%
  group_by(BORO) %>%
  summarise(shootings = n())

# Count shooting incidents per year
yearly_shootings <- df %>%
  group_by(YEAR) %>%
  summarise(shootings = n())
```

```
# Visualization 1: Shooting incidents per city
ggplot(city_shootings, aes(x = reorder(BORO, -shootings), y = shootings)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Shooting Incidents per City",
       x = "City",
       y = "Number of Shooting Incidents") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



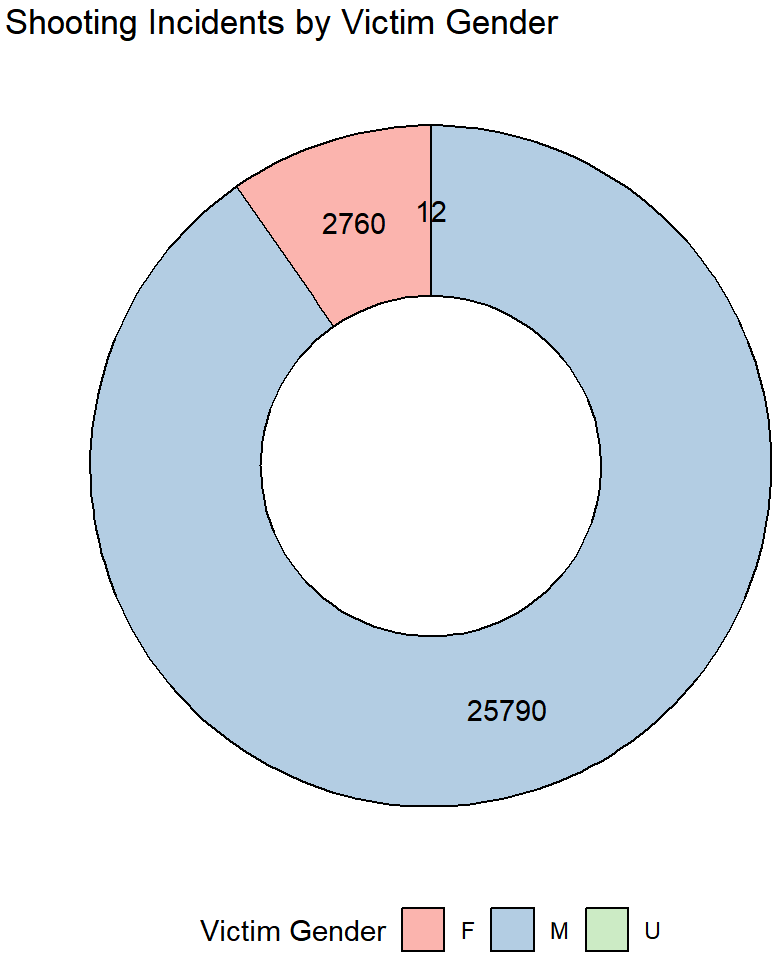
Shooting Incidents by Age_Group

```
library(ggplot2)

# shooting incidents per victim age group
age_group_shootings <- df %>%
  group_by(VIC_AGE_GROUP) %>%
  summarise(shootings = n())

# bar plot for shooting incidents by victim age group
ggplot(age_group_shootings, aes(x = VIC_AGE_GROUP, y = shootings)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Shooting Incidents by Victim Age Group",
       x = "Victim Age Group",
       y = "Number of Shooting Incidents") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Shooting Incidents by Victim Gender

```
age_group_shootings <- df %>%
  group_by(VIC_SEX) %>%
  summarise(shootings = n())

ggplot(age_group_shootings, aes(x = 2, y = shootings, fill = VIC_SEX)) +
  geom_bar(stat = "identity", width = 1, color = "black") +
  coord_polar(theta = "y") +
  xlim(0.5, 2.5) +
  theme_void() +
  geom_text(aes(label = shootings), position = position_stack(vjust = 0.5)) +
  labs(title = "Shooting Incidents by Victim Gender", fill = "Victim Gender") +
  theme(legend.position = "bottom") +
  scale_fill_brewer(palette = "Pastel1")
```



-|-|-|-|-|-|-|-|-|-|-|Step 4|-|-|-|-|-|-|-|-|-|-|-

Conclusion

We need to allocate more resources and put more police officers in Brooklyn. Perhaps move some officers from Staten Island to Brooklyn since they the fewest incidents. Work we the local social services to work with the youth to prevent crimes from happening in the first place.

Bias

1 - There is a potential bias since we don’t know how data was collected. 2 - lack of demographic information make any decision incomplete. 3 - Too many missing values