NYPD Shooting Incident Data Report

Madhumita Mondal

2023-06-01

Introduction:

In recent years, there has been an alarming increase in hate crimes and shooting incidents across the United States. This issue has sparked a national debate, and it is crucial to gain a better understanding of criminal activity through the statistical analysis of available data, such as the New York City Shooting Incidents dataset. This analysis can provide valuable insights and help formulate effective police enforcement and intervention strategies. In this report, we will explore the NYPD Shooting Incident data to identify patterns, relationships, and trends in the criminal activity, and generate insights that can inform decision-making and policy development.

To begin, we need to install these necessary packages:(tidyverse), (lubridate), (ggplot2), (gridExtra), (knitr)

Read the data from the link.

```
# Read CSV file from URL
nypd_shooting <- read_csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType
=DOWNLOAD")
```

```
## Rows: 27312 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.
```

Display the first 10 rows of the dataset.

```
head(nypd_shooting, 10)
```

```
## # A tibble: 10 × 21
##    INCID…¹ OCCUR…² OCCUR…³ BORO  LOC_O…⁴ PRECI…⁵ JURIS…⁶ LOC_C…⁷ LOCAT…⁸ STATI…⁹
##      <dbl> <chr>   <time>  <chr> <chr>     <dbl>   <dbl> <chr>   <chr>   <lgl>
##  1  2.29e8 05/27/… 21:30   QUEE… <NA>        105       0 <NA>    <NA>    FALSE
##  2  1.37e8 06/27/… 17:40   BRONX <NA>         40       0 <NA>    <NA>    FALSE
##  3  1.48e8 11/21/… 03:56   QUEE… <NA>        108       0 <NA>    <NA>    TRUE
##  4  1.47e8 10/09/… 18:30   BRONX <NA>         44       0 <NA>    <NA>    FALSE
##  5  5.89e7 02/19/… 22:58   BRONX <NA>         47       0 <NA>    <NA>    TRUE
##  6  2.20e8 10/21/… 21:36   BROO… <NA>         81       0 <NA>    <NA>    TRUE
##  7  8.53e7 06/17/… 22:47   QUEE… <NA>        114       0 <NA>    <NA>    FALSE
##  8  7.17e7 03/08/… 19:41   BROO… <NA>         81       0 <NA>    <NA>    TRUE
##  9  8.30e7 02/05/… 05:45   QUEE… <NA>        105       0 <NA>    <NA>    FALSE
## 10  8.64e7 08/26/… 01:10   QUEE… <NA>        101       0 <NA>    MULTI … FALSE
## # … with 11 more variables: 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>, and abbreviated variable names ¹INCIDENT_KEY, ²OCCUR_DATE,
## #   ³OCCUR_TIME, ⁴LOC_OF_OCCUR_DESC, ⁵PRECINCT, ⁶JURISDICTION_CODE,
## #   ⁷LOC_CLASSFCTN_DESC, ⁸LOCATION_DESC, ⁹STATISTICAL_MURDER_FLAG
```

Data Preparation and Cleaning

Rename OCCUR_DATE and OCCUR_TIME to Date and Time respectively.

```
nypd_shooting <- nypd_shooting %>%
  rename(Date = OCCUR_DATE,
         Time = OCCUR_TIME)
```

Missing Values.

```
# Replace missing values with "N/A"
nypd_shooting <- nypd_shooting %>%
  mutate(across(-Time, ~ifelse(is.na(.), "N/A", .)))
```

Making sure there is no missing values.

```
sum(is.na(nypd_shooting))
```

```
## [1] 0
```

Show the first 10 rows

```
head(nypd_shooting, 10)
```

```
## # A tibble: 10 × 21
##    INCIDENT_…¹ Date  Time  BORO  LOC_O…² PRECI…³ JURIS…⁴ LOC_C…⁵ LOCAT…⁶ STATI…⁷
##          <dbl> <chr> <tim> <chr> <chr>     <dbl> <chr>   <chr>   <chr>   <lgl>
##  1   228798151 05/2… 21:30 QUEE… N/A         105 0       N/A     N/A     FALSE
##  2   137471050 06/2… 17:40 BRONX N/A          40 0       N/A     N/A     FALSE
##  3   147998800 11/2… 03:56 QUEE… N/A         108 0       N/A     N/A     TRUE
##  4   146837977 10/0… 18:30 BRONX N/A          44 0       N/A     N/A     FALSE
##  5    58921844 02/1… 22:58 BRONX N/A          47 0       N/A     N/A     TRUE
##  6   219559682 10/2… 21:36 BROO… N/A          81 0       N/A     N/A     TRUE
##  7    85295722 06/1… 22:47 QUEE… N/A         114 0       N/A     N/A     FALSE
##  8    71662474 03/0… 19:41 BROO… N/A          81 0       N/A     N/A     TRUE
##  9    83002139 02/0… 05:45 QUEE… N/A         105 0       N/A     N/A     FALSE
## 10    86437261 08/2… 01:10 QUEE… N/A         101 0       N/A     MULTI … FALSE
## # … with 11 more variables: 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 <chr>, Longitude <chr>,
## #   Lon_Lat <chr>, and abbreviated variable names ¹INCIDENT_KEY,
## #   ²LOC_OF_OCCUR_DESC, ³PRECINCT, ⁴JURISDICTION_CODE, ⁵LOC_CLASSFCTN_DESC,
## #   ⁶LOCATION_DESC, ⁷STATISTICAL_MURDER_FLAG
```

Check and remove any duplicates.

```
duplicated_rows <- nypd_shooting[duplicated(nypd_shooting),]
nypd_shooting <- distinct(nypd_shooting)
nrow(nypd_shooting)
```

```
## [1] 27312
```

It appears there are no duplicates.

Now let’s check unique Values in borough.

```
unique(nypd_shooting$BORO)
```

```
## [1] "QUEENS"        "BRONX"         "BROOKLYN"      "MANHATTAN"    
## [5] "STATEN ISLAND"
```

```
nypd_shooting$Date <- as.Date(nypd_shooting$Date, format = "%m/%d/%Y")
```

Let’s take a look at the table.

```
head(nypd_shooting, 10)
```

```
## # A tibble: 10 × 21
##    INCIDENT_KEY Date       Time   BORO   LOC_O…¹ PRECI…² JURIS…³ LOC_C…⁴ LOCAT…⁵
##           <dbl> <date>     <time> <chr>  <chr>     <dbl> <chr>   <chr>   <chr>
##  1    228798151 2021-05-27 21:30  QUEENS N/A         105 0       N/A     N/A
##  2    137471050 2014-06-27 17:40  BRONX  N/A          40 0       N/A     N/A
##  3    147998800 2015-11-21 03:56  QUEENS N/A         108 0       N/A     N/A
##  4    146837977 2015-10-09 18:30  BRONX  N/A          44 0       N/A     N/A
##  5     58921844 2009-02-19 22:58  BRONX  N/A          47 0       N/A     N/A
##  6    219559682 2020-10-21 21:36  BROOK… N/A          81 0       N/A     N/A
##  7     85295722 2012-06-17 22:47  QUEENS N/A         114 0       N/A     N/A
##  8     71662474 2010-03-08 19:41  BROOK… N/A          81 0       N/A     N/A
##  9     83002139 2012-02-05 05:45  QUEENS N/A         105 0       N/A     N/A
## 10     86437261 2012-08-26 01:10  QUEENS N/A         101 0       N/A     MULTI …
## # … with 12 more variables: 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 <chr>, Longitude <chr>, Lon_Lat <chr>, and abbreviated variable
## #   names ¹LOC_OF_OCCUR_DESC, ²PRECINCT, ³JURISDICTION_CODE,
## #   ⁴LOC_CLASSFCTN_DESC, ⁵LOCATION_DESC
```

I just want to make sure the Date column in the right datatype.

```
class(nypd_shooting$Date)
```

```
## [1] "Date"
```

Here I made a new column for the population for each borough.

```
nypd_shooting <- nypd_shooting %>%
 mutate(Population = case_when(
    BORO == "BROOKLYN" ~ 2576771,
    BORO == "QUEENS" ~ 2270976,
    BORO == "BRONX" ~ 1427056,
    BORO == "MANHATTAN" ~ 1629153,
    BORO == "STATEN ISLAND" ~ 475596,
    TRUE ~ NA_real_
  ))
head(nypd_shooting, 10)
```

```
## # A tibble: 10 × 22
##    INCIDENT_KEY Date       Time   BORO   LOC_O…¹ PRECI…² JURIS…³ LOC_C…⁴ LOCAT…⁵
##           <dbl> <date>     <time> <chr>  <chr>     <dbl> <chr>   <chr>   <chr>
##  1    228798151 2021-05-27 21:30  QUEENS N/A         105 0       N/A     N/A
##  2    137471050 2014-06-27 17:40  BRONX  N/A          40 0       N/A     N/A
##  3    147998800 2015-11-21 03:56  QUEENS N/A         108 0       N/A     N/A
##  4    146837977 2015-10-09 18:30  BRONX  N/A          44 0       N/A     N/A
##  5     58921844 2009-02-19 22:58  BRONX  N/A          47 0       N/A     N/A
##  6    219559682 2020-10-21 21:36  BROOK… N/A          81 0       N/A     N/A
##  7     85295722 2012-06-17 22:47  QUEENS N/A         114 0       N/A     N/A
##  8     71662474 2010-03-08 19:41  BROOK… N/A          81 0       N/A     N/A
##  9     83002139 2012-02-05 05:45  QUEENS N/A         105 0       N/A     N/A
## 10     86437261 2012-08-26 01:10  QUEENS N/A         101 0       N/A     MULTI …
## # … with 13 more variables: 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 <chr>, Longitude <chr>, Lon_Lat <chr>, Population <dbl>, and
## #   abbreviated variable names ¹LOC_OF_OCCUR_DESC, ²PRECINCT,
## #   ³JURISDICTION_CODE, ⁴LOC_CLASSFCTN_DESC, ⁵LOCATION_DESC
```

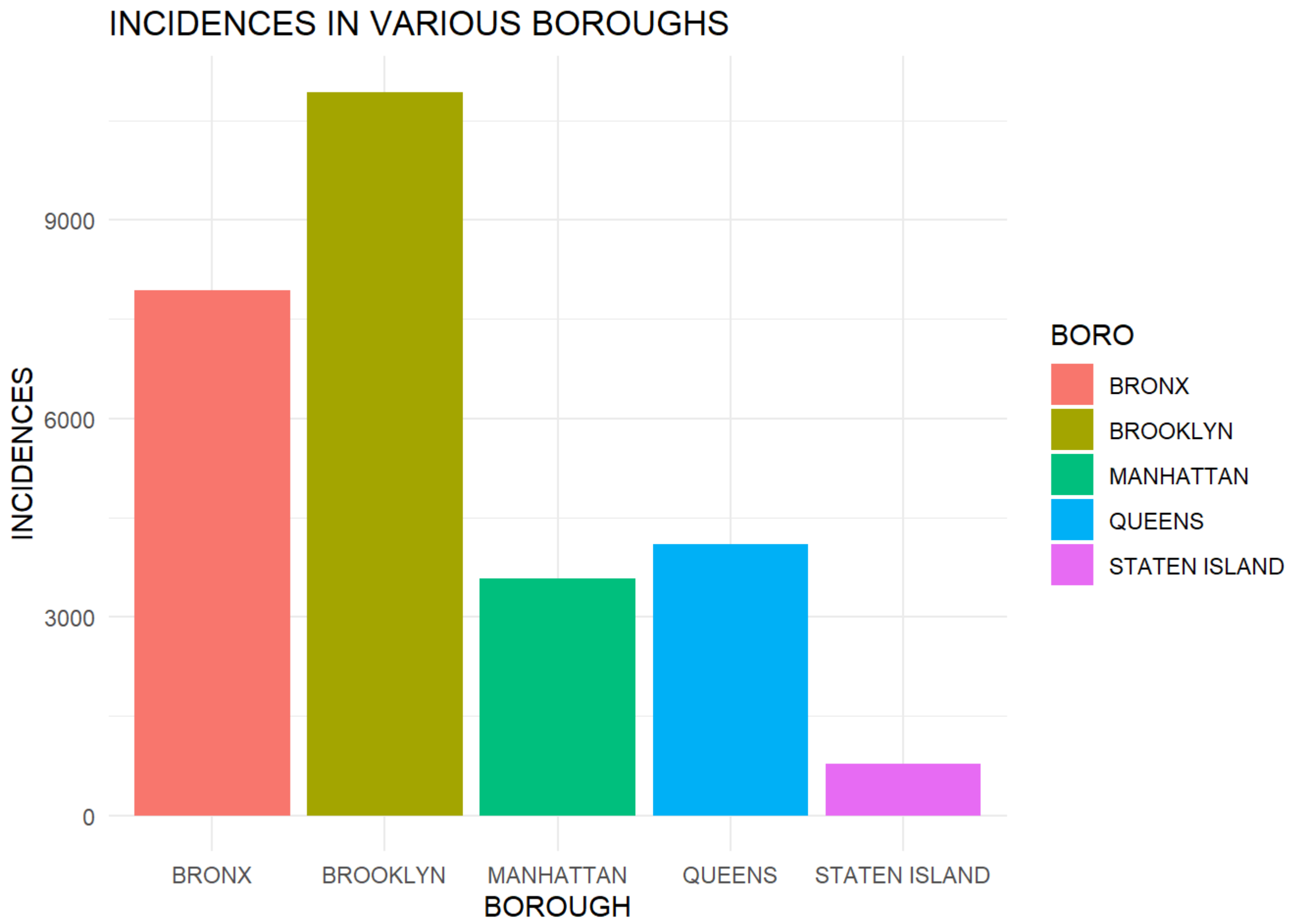
Sort the borough in descending order to see which one has the most shootings.

```
nypd_shooting %>%
  group_by(BORO) %>%
  summarise(Total = n()) %>%
  arrange(desc(Total))
```

```
## # A tibble: 5 × 2
##   BORO          Total
##   <chr>         <int>
## 1 BROOKLYN      10933
## 2 BRONX          7937
## 3 QUEENS         4094
## 4 MANHATTAN      3572
## 5 STATEN ISLAND   776
```

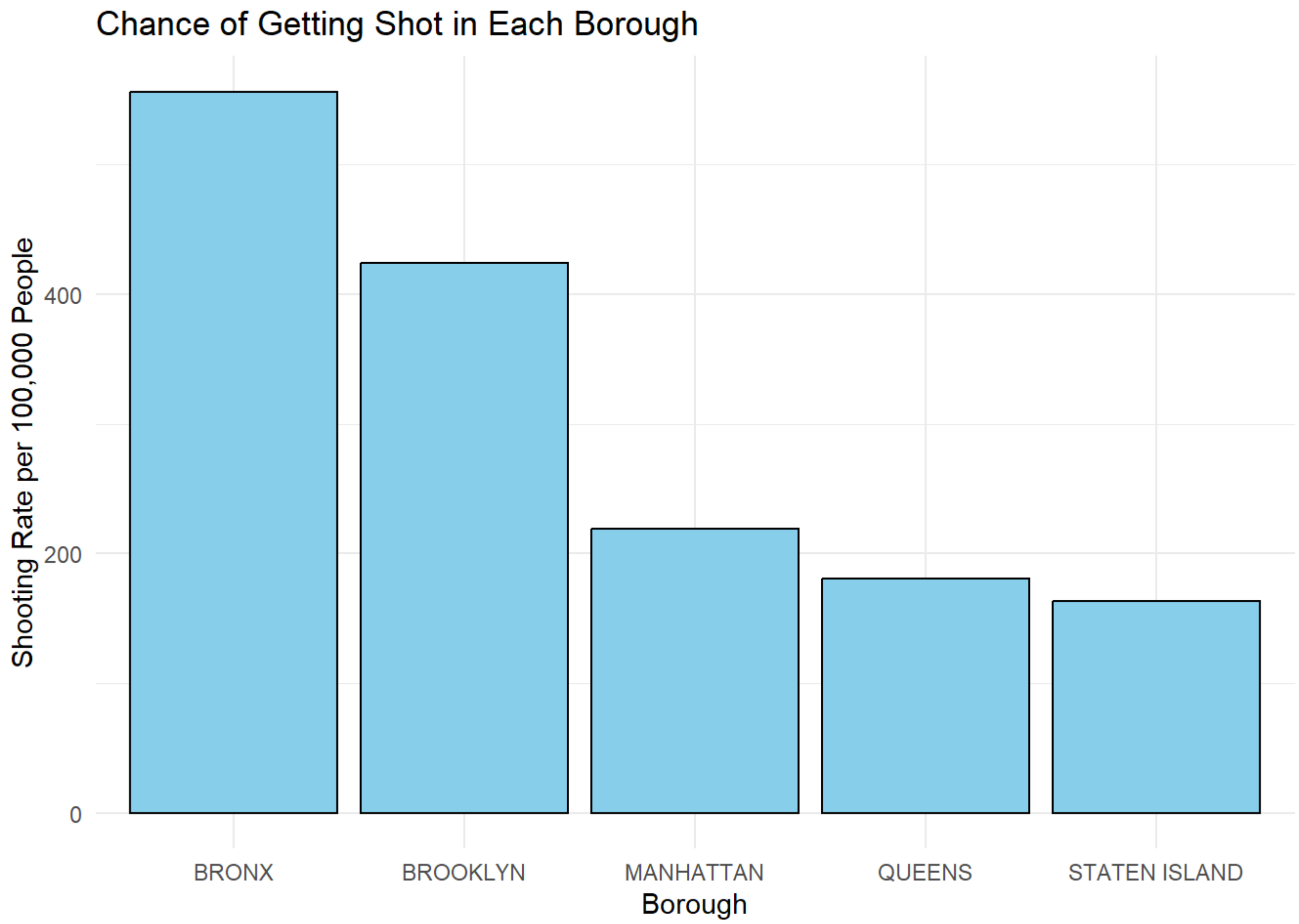
Bar Chart to see incidences.

```
# Group data by BORO and calculate the total number of incidents
boro_shootings <- nypd_shooting %>% group_by(BORO) %>%
  summarize(incidents = n())

# Create bar graph
ggplot(boro_shootings, aes(x=BORO, y=incidents, fill=BORO)) +
  geom_bar(stat="identity") +
  xlab("BOROUGH") + ylab("INCIDENCES") +
  ggtitle("INCIDENCES IN VARIOUS BOROUGHS") +
  theme_minimal()
```



Calculate the shooting rate per 100,000 people and Plot the shooting rate for each borough

```
nypd_shooting_rate <- nypd_shooting %>%
  group_by(BORO) %>%
  summarise(total_shootings = n(),
            population = unique(Population),
            shooting_rate = total_shootings / (population / 100000)) %>%
  arrange(desc(shooting_rate))

ggplot(nypd_shooting_rate, aes(x = BORO, y = shooting_rate)) +
  geom_bar(stat = "identity", fill = "skyblue", color = "black") +
  ggtitle("Chance of Getting Shot in Each Borough") +
  xlab("Borough") +
  ylab("Shooting Rate per 100,000 People") +
  theme_minimal()
```



```
nypd_shooting_rate %>%
  as_tibble() %>%
  select(BORO, shooting_rate) %>%
  mutate(shooting_rate = sprintf("%.2f", shooting_rate))
```

```
## # A tibble: 5 × 2
##   BORO          shooting_rate
##   <chr>         <chr>
## 1 BRONX         556.18
## 2 BROOKLYN      424.29
## 3 MANHATTAN     219.26
## 4 QUEENS        180.27
## 5 STATEN ISLAND 163.16
```

```
nypd_shooting_rate_per_person <- nypd_shooting_rate %>%
  mutate(shooting_rate_per_person = total_shootings / population) %>%
  select(BORO, shooting_rate_per_person) %>%
  mutate(shooting_rate_per_person = sprintf("%.6f", shooting_rate_per_person * 100)) %>%
  rename(`Borough` = BORO, `Shooting Rate per Person` = shooting_rate_per_person) %>%
  mutate(`Shooting Rate per Person` = paste0(`Shooting Rate per Person`, "%"))

print(nypd_shooting_rate_per_person)
```

```
## # A tibble: 5 × 2
##   Borough       `Shooting Rate per Person`
##   <chr>         <chr>
## 1 BRONX         0.556180%
## 2 BROOKLYN      0.424291%
## 3 MANHATTAN     0.219255%
## 4 QUEENS        0.180275%
## 5 STATEN ISLAND 0.163164%
```

Create the linear regression model and Print the summary of the model

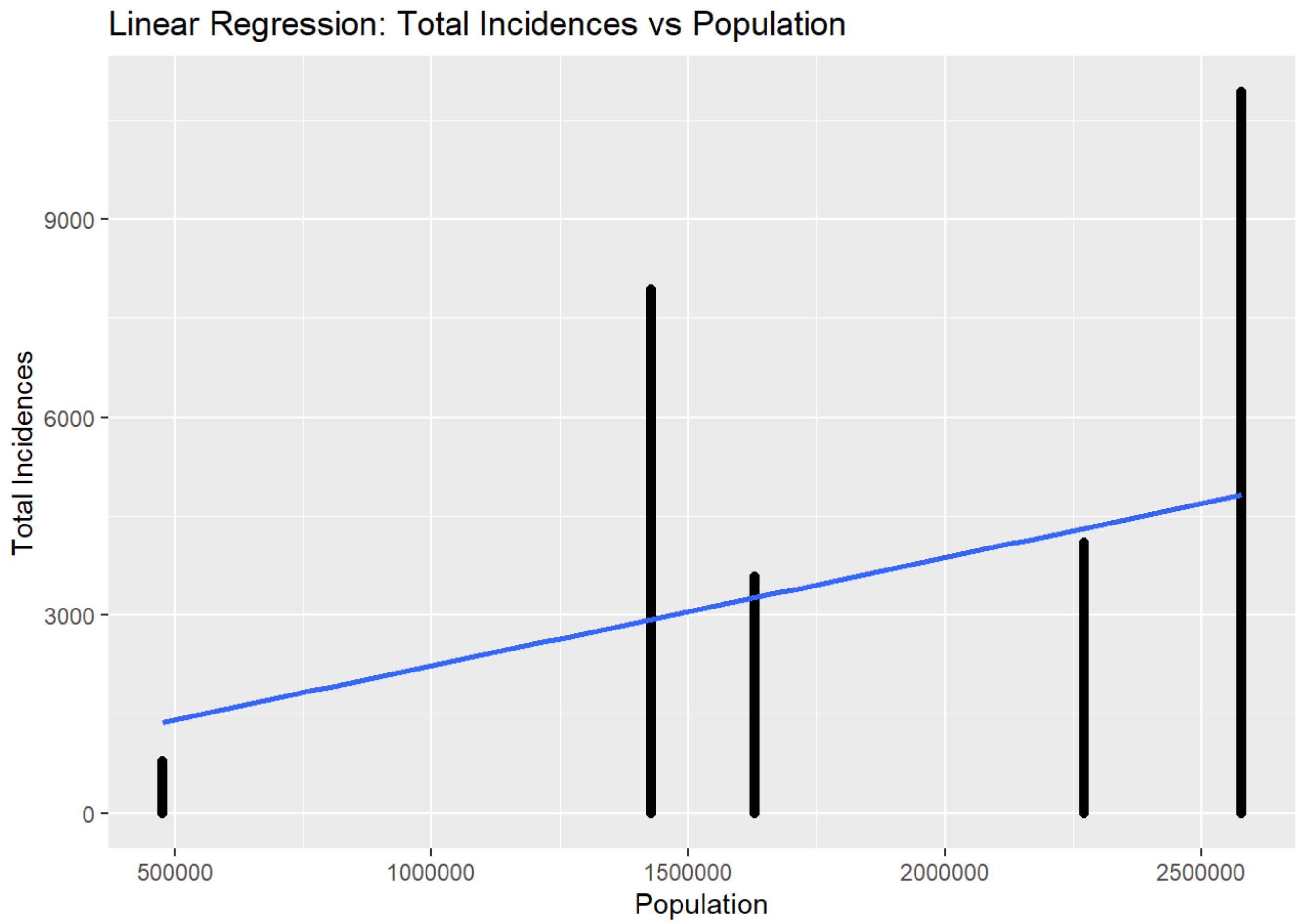
```
nypd_shooting <- nypd_shooting %>%
  mutate(Total = ifelse(!is.na(BORO), 1, 0)) %>%
  group_by(BORO) %>%
  mutate(Total = cumsum(Total))

lm_model <- lm(Total ~ Population, data = nypd_shooting)

summary(lm_model)
```

```
##
## Call:
## lm(formula = Total ~ Population, data = nypd_shooting)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4815.6 -2124.9  -609.4  2146.5  6116.4
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.958e+02  6.067e+01   9.821   <2e-16 ***
## Population  1.638e-03  2.900e-05  56.482   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2726 on 27310 degrees of freedom
## Multiple R-squared:  0.1046, Adjusted R-squared:  0.1046
## F-statistic:  3190 on 1 and 27310 DF,  p-value: < 2.2e-16
```

Create a scatter plot with the regression line

```
ggplot(nypd_shooting, aes(x = Population, y = Total)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  xlab("Population") +
  ylab("Total Incidences") +
  ggtitle("Linear Regression: Total Incidences vs Population")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Conclusion:

we can conclude that the Bronx has the highest chance of getting shot per person compared to the other boroughs in New York City. Staten Island has the lowest chance of getting shot per person. However, it’s important to note that the difference in shooting rates between the boroughs is not very large, with the highest rate being only slightly above 0.5% and the lowest rate being just over 0.15%.

Based on the linear regression results, we can conclude that there is a positive relationship between the number of shooting incidents and the population size in each borough. In other words, as the population size increases, the number of shooting incidents tends to increase as well. The R-squared value of 0.717 indicates that the model explains approximately 72% of the variability in the number of shooting incidents. However, it’s important to note that correlation does not imply causation, and there may be other factors that contribute to the number of shooting incidents beyond just population size.