NYD Shooting Project

Zeeshan Amjad

2024-08-10

R Markdown

This is an R Markdown document for NYPD shooting project for Data Science as Field course.

Loading packages

- tidyverse
- dplyr
- ggplot2

```
library('tidyverse')
library('dplyr')
install.packages('ggplot2')
```

Loading New York Shooting Data

Read the data from the url https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv

```
cases_url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv"
cases <- read_csv(cases_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.
```

See the specification of the data

Use spec() to see the full column specification to understand the data

```
spec(cases)
```

```
## cols(
##   INCIDENT_KEY = col_double(),
##   OCCUR_DATE = col_character(),
##   OCCUR_TIME = col_time(format = ""),
##   BORO = col_character(),
##   LOC_OF_OCCUR_DESC = col_character(),
##   PRECINCT = col_double(),
##   JURISDICTION_CODE = col_double(),
##   LOC_CLASSFCTN_DESC = col_character(),
##   LOCATION_DESC = col_character(),
##   STATISTICAL_MURDER_FLAG = col_logical(),
##   PERP_AGE_GROUP = col_character(),
##   PERP_SEX = col_character(),
##   PERP_RACE = col_character(),
##   VIC_AGE_GROUP = col_character(),
##   VIC_SEX = col_character(),
##   VIC_RACE = col_character(),
##   X_COORD_CD = col_double(),
##   Y_COORD_CD = col_double(),
##   Latitude = col_double(),
##   Longitude = col_double(),
##   Lon_Lat = col_character()
## )
```

Preparing the data

Change the data type

For preparing the data, change the date into date format

```
cases <- cases %>%
  mutate(
    OCCUR_DATE = as.Date(OCCUR_DATE, "%m/%d/%Y"))
```

Remove the unnecessary columns

Remove the following from the data - X_COORD_CD - Y_COORD_CD - Latitude - Longitude - Lon_Lat

```
cases <- cases %>%
  select(-c(X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat))
```

Display summary

```
summary(cases)
```

```
##   INCIDENT_KEY         OCCUR_DATE          OCCUR_TIME           BORO
##  Min.   :  9953245   Min.   :2006-01-01   Length:28562      Length:28562
##  1st Qu.: 65439914   1st Qu.:2009-09-04   Class1:hms        Class :character
##  Median : 92711254   Median :2013-09-20   Class2:difftime   Mode  :character
##  Mean   :127405824   Mean   :2014-06-07   Mode  :numeric
##  3rd Qu.:203131993   3rd Qu.:2019-09-29
##  Max.   :279758069   Max.   :2023-12-29
##
##  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
##  Length:28562
##  Class :character
##  Mode  :character
##
##
##
##
```

Group by

Group by one level

Now doing the group by of one level, here we are doing group by - boro - year - location - gender - time

```
group_by_boro <- cases %>%
  group_by(BORO) %>%
  summarise(INCIDENTS = n())

group_by_year <- cases %>%
  group_by(YEAR = format(OCCUR_DATE, "%Y")) %>%
```

```
  summarise(INCIDENTS = n())

group_by_location <- cases %>%
  group_by(LOCATION_DESC) %>%
  summarise(INCIDENTS = n())

group_by_gender <- cases %>%
  filter(!is.na(cases$PERP_SEX)) %>%
  filter(if_any(everything(), is.na)) %>%
  group_by(PERP_SEX) %>%
  summarise(INCIDENTS = n())

group_by_time <- cases %>%
  group_by(OCCUR_TIME) %>%
  summarise(INCIDENTS = n())
```

Group by two levels

Now doing the group by of two levels, here we are doing group by - boro, year - year, boro - location, year - year, location

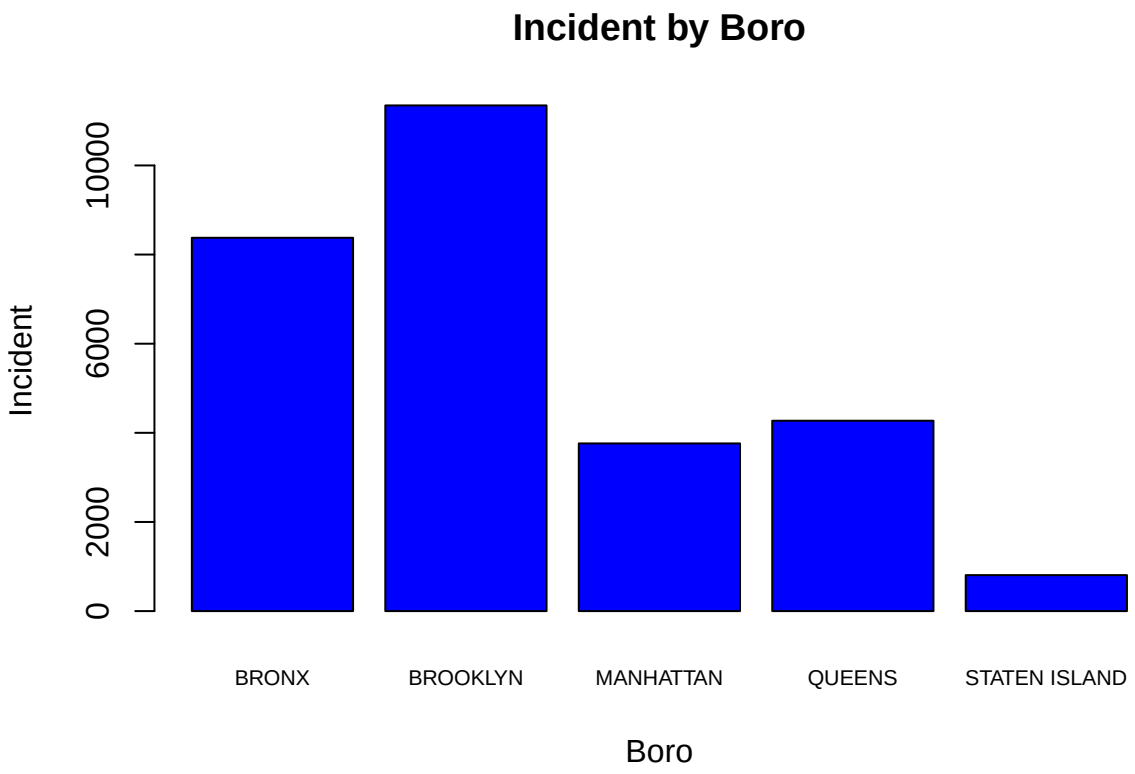
```
group_by_boro_year <- cases %>%
  group_by(BORO, YEAR = format(OCCUR_DATE, "%Y")) %>%
  summarise(INCIDENTS = n(), .groups="keep")

group_by_year_boro <- cases %>%
  group_by(YEAR = format(OCCUR_DATE, "%Y"), BORO) %>%
  summarise(INCIDENTS = n(), .groups="keep")

group_by_location_year <- cases %>%
  group_by(LOCATION_DESC, , YEAR = format(OCCUR_DATE, "%Y")) %>%
  summarise(INCIDENTS = n(), .groups="keep")

group_by_year_location <- cases %>%
  group_by(YEAR = format(OCCUR_DATE, "%Y"), LOCATION_DESC) %>%
  summarise(INCIDENTS = n(), .groups="keep")
```

Visualize Data

Now visualization the data of NYPD shooting incidents

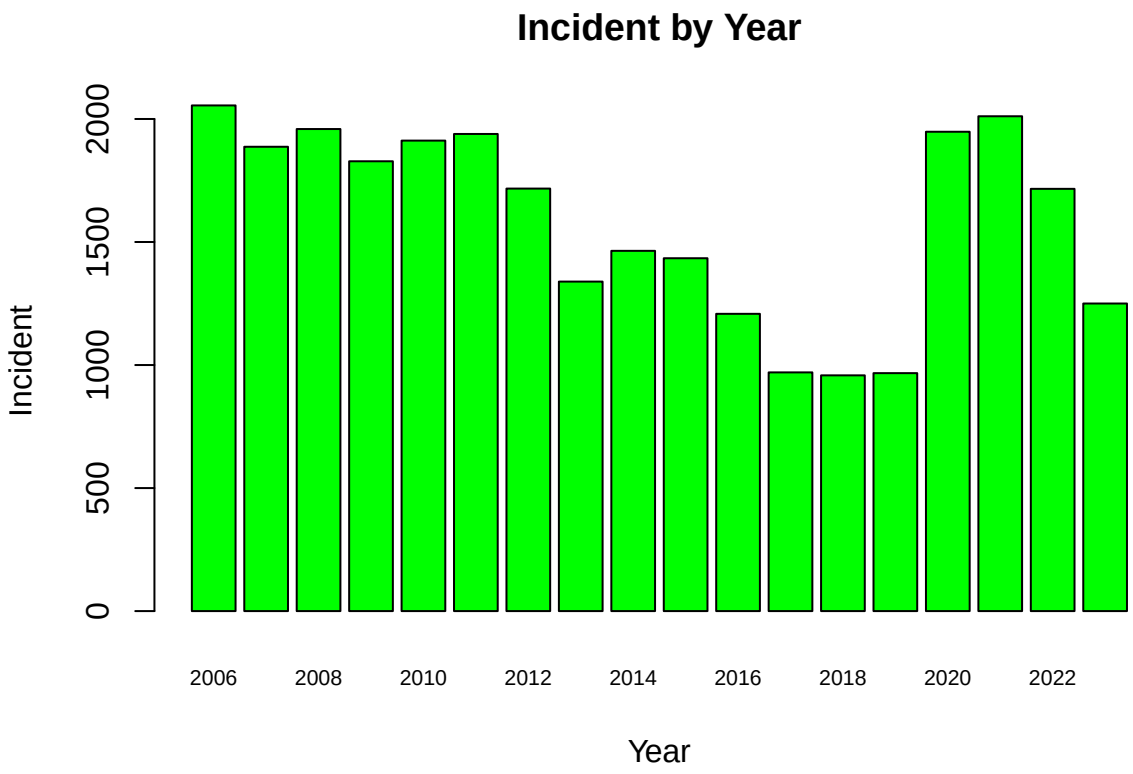
incident by boro

```
barplot(group_by_boro$INCIDENTS, names.arg = group_by_boro$BORO,
        col = "blue", cex.names = 0.7, main = "Incident by Boro",
        xlab = "Boro", ylab = "Incident")
```



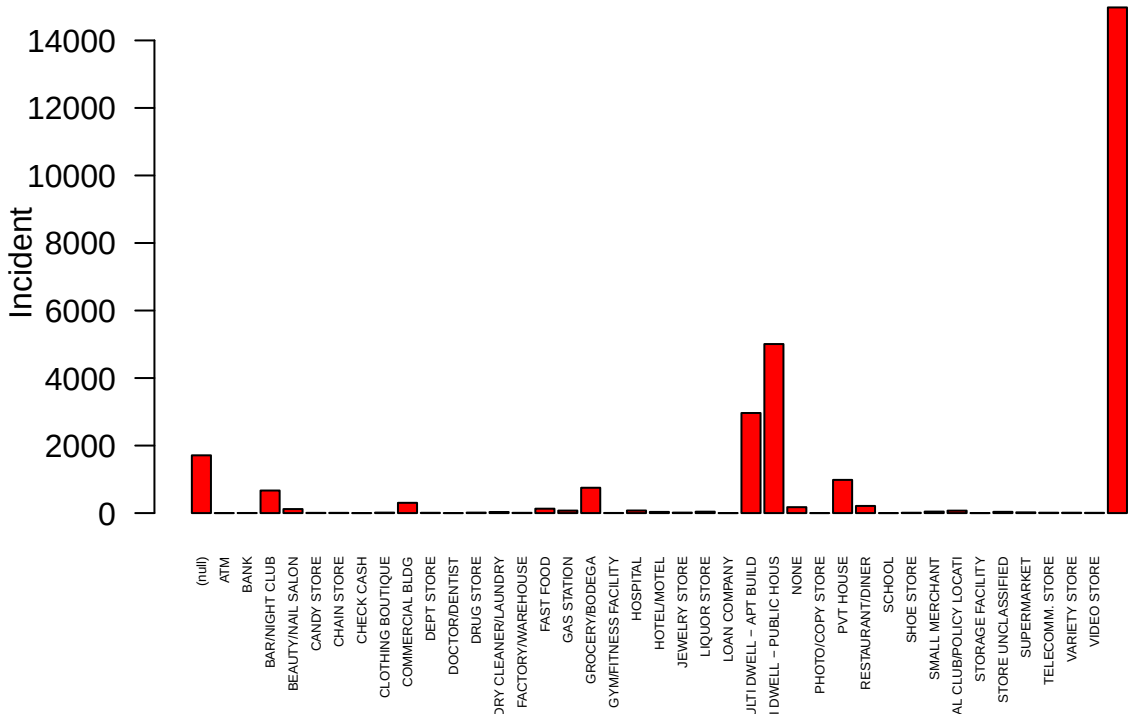
incident by year

```
barplot(group_by_year$INCIDENTS, names.arg = group_by_year$YEAR,
        col = "green", cex.names = 0.7, main = "Incident by Year",
        xlab = "Year", ylab = "Incident")
```



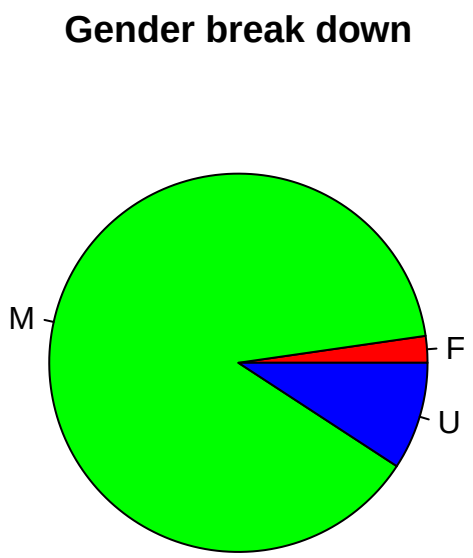
incident by location

```
barplot(group_by_location$INCIDENTS, names.arg = group_by_location$LOCATION_DESC,
        col = "red", las=2, cex.names = 0.4, main = "Incident by Location", ylab = "Incident")
```



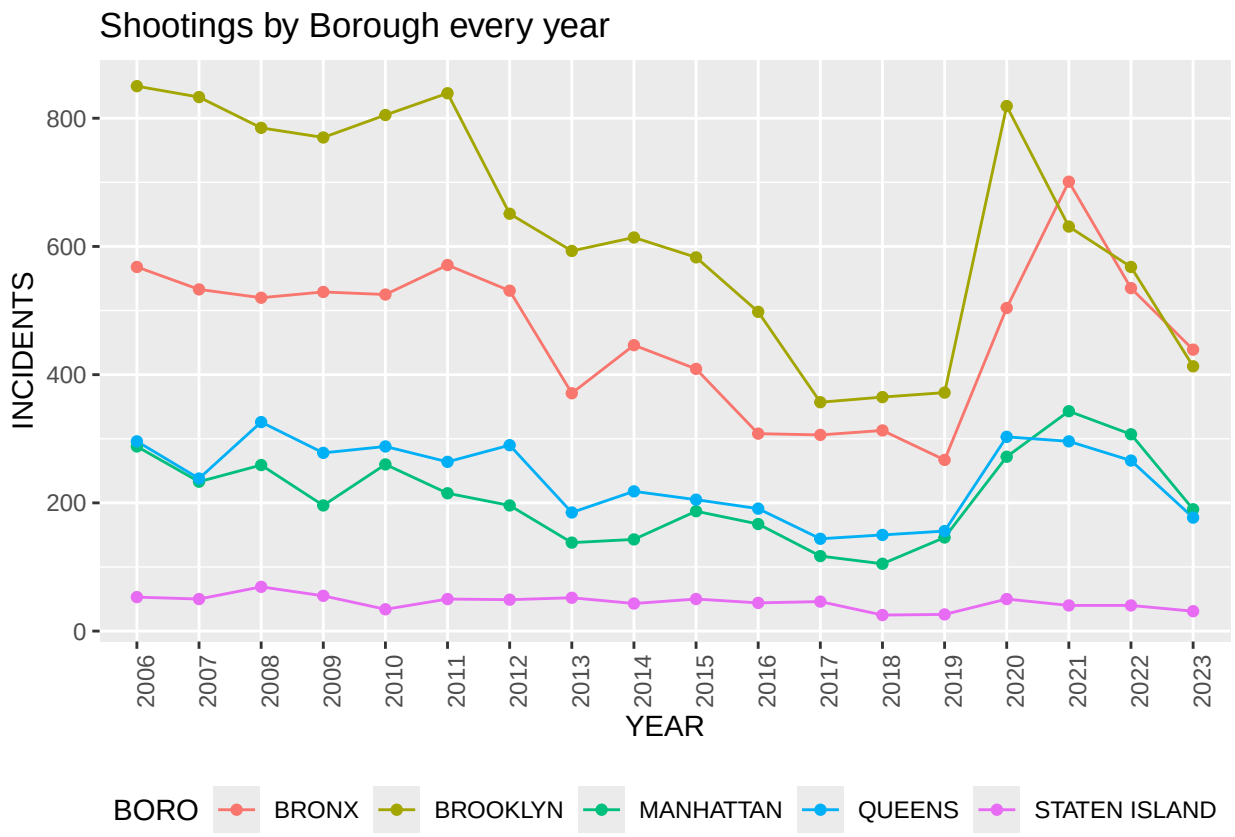
gender breakdown

```
pie(group_by_gender$INCIDENTS, labels = group_by_gender$PERP_SEX,
    main = "Gender break down", col = rainbow(length(group_by_gender$INCIDENTS)))
```



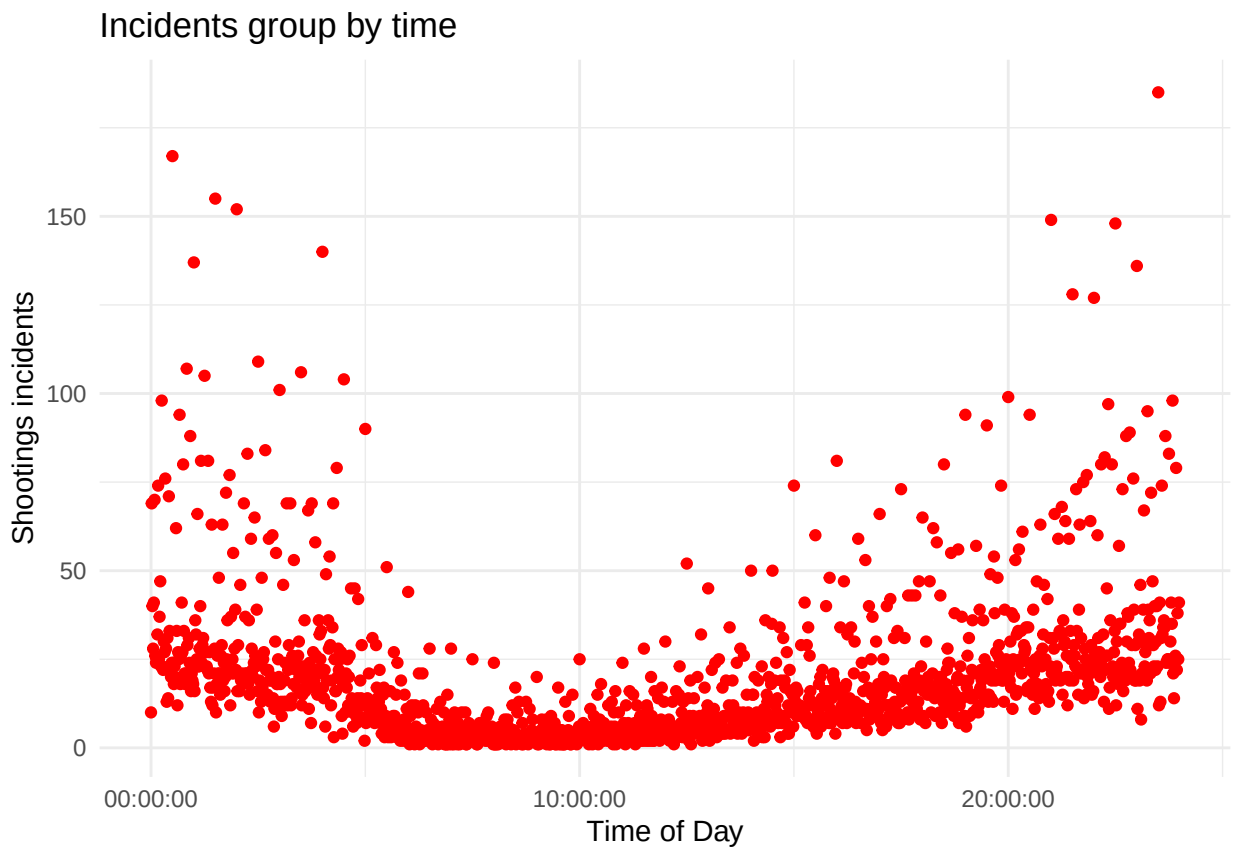
incident by boro ever year

```
group_by_boro_year  %>%
     ggplot(aes(x=YEAR, y=INCIDENTS, group = BORO)) +
     geom_line(aes(color = BORO)) +
     geom_point(aes(color = BORO)) +
     theme(legend.position = "bottom",
           axis.text.x = element_text(angle = 90)) +
     labs(title = "Shootings by Borough every year")
```



incident by time

```
ggplot(group_by_time, aes(x = OCCUR_TIME, y = INCIDENTS)) +
  geom_point(col="red") +
  labs(title = "Incidents group by time",
       x = "Time of Day", y = "Shootings incidents") +
  theme_minimal()
```



Model

This visualization is interesting and we can create a model based on time and number of shooting incidents and trying to see if it fits linearly. This visualization shows that more number of shooting incidents are happening at the night. Now let's create a linear model for it.

```
model <- lm(INCIDENTS ~ OCCUR_TIME, data = group_by_time)

summary(model)
```

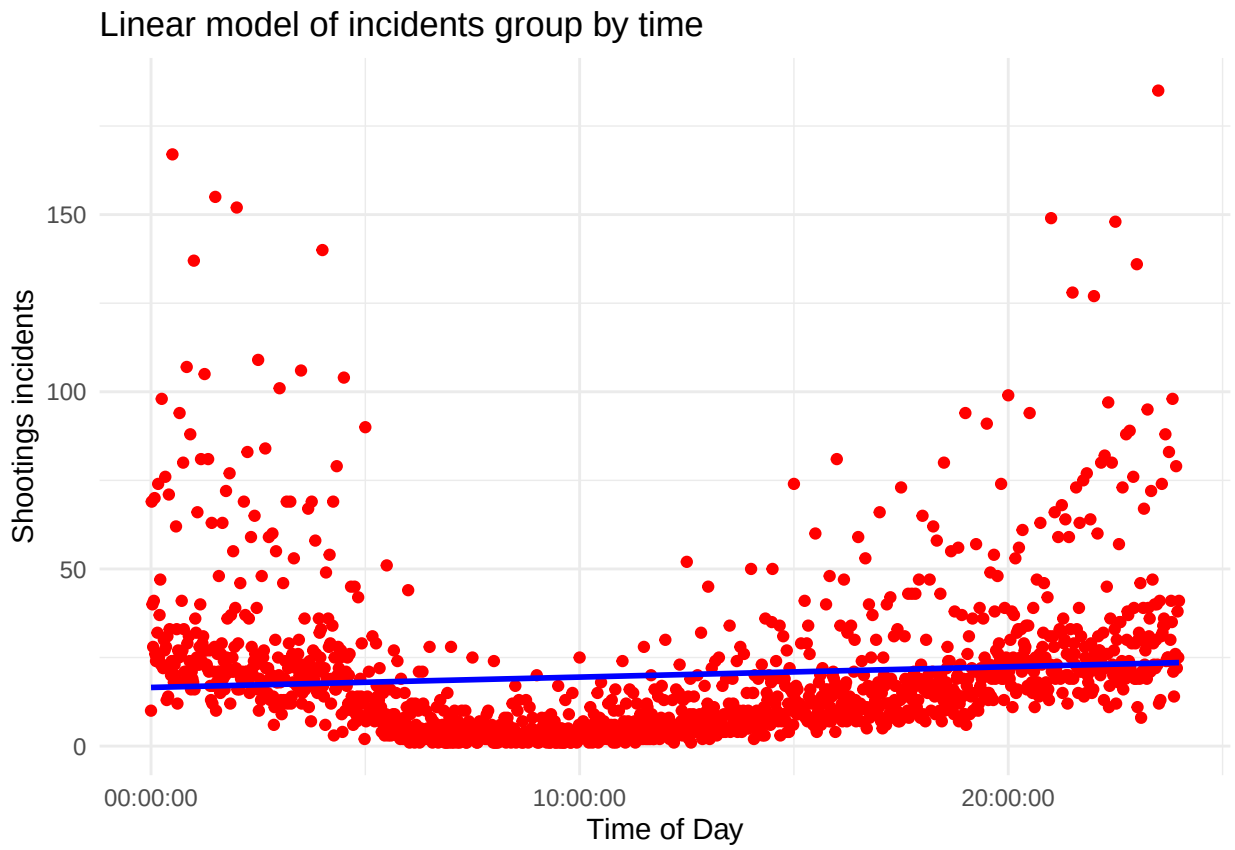
```
##
## Call:
## lm(formula = INCIDENTS ~ OCCUR_TIME, data = group_by_time)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.237 -13.686  -5.625   4.651 161.577
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.655e+01  1.140e+00  14.520  < 2e-16 ***
## OCCUR_TIME  8.119e-05  2.278e-05   3.564 0.000378 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 21.53 on 1421 degrees of freedom
## Multiple R-squared:  0.008858,   Adjusted R-squared:  0.00816
## F-statistic:  12.7 on 1 and 1421 DF,  p-value: 0.0003779
```

Now trying to see the linear model on this visualization.

```
ggplot(group_by_time, aes(x = OCCUR_TIME, y = INCIDENTS)) +
  geom_point(col="red") +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Linear model of incidents group by time",
       x = "Time of Day", y = "Shootings incidents") +
  theme_minimal()
```

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



These visualization shows that there are less number of shooting incident on Staten Island boro making it the safest and most number of shooting incident in Brooklyn. It also shows that Male are more likely to be a perpetrator than female.

When we visualize the data over time then for year 2017 to 2019 the shooting incident decreases and it is also evident if we look at the for each boro over the time.

One possible hypothesis could be since Staten Island is not directly connected with land except two bridges, therefore it would be difficult for shooter to escape from there. However we need to further investigate to validate or invalidate this hypothesis.

Another important observation is the time frame of 2017 to 2019, the number of shooting instances reduces in all boro and further investigation needed to understand the reason behind this. There may be another factor that may cause this.

We can further explore the relationship between location and number of shooting incident, but we have lots of missing data for location, so we can't make a concrete decision based on the existing data.

Bais

It is bais to assume that number of shooting is less on Island as compare to main land. Another important bais maight be who is doing this shooting? It is from the law enforcement or crimnal? Is there a really good law and order sitaution in staten island or the geography plays an important role here.

Conclusion

Although this dataset provide useful information and we can get new insight by grouping and visualization this, but this data is not sufficient to create any meaningful conclusion. There are two important reason for that - Some important data points are missing like location for lots of rows and who did the shooting, law enforcement, criminal or someone else - We may need to include some other datasets in our study to come up with some meaningful conclusion or even can find the missing link, relationship, causation in the current dataset.