NYPD Shooting Incident Data Report

2023-11-27

R code chunk "load_library" to load libraries

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate)
```

1. Introduction

This report is to analyze NYPD Shooting data on public dataset made available on NYPD website. R Markdown is used for this analysis.

About dataset

This is a breakdown of every shooting incident that occurred in NYC going back to 2013 through the end of the previous calendar year. This data is manually extracted every quarter and reviewed by the Office of Management Analysis and Planning before being posted on the NYPD website. Each record represents a shooting incident in NYC and includes information about the event.

Questions which will be addressed through this analysis

1. Is shooting rate increasing or decreasing year over year ? How much effective is law and order?
2. Are there any specific Boroughs in New York, which are more impacted by these shooting events ?
3. Are there any specific age groups which are more victimized due to shooting incidents ?

2. Report - Analysis and Visualizations

We will be analysing and visualizing these trends:

- Trend of shooting incident per year
- Trend of shooting incident per year for each Borough.
- Trend of shooting incident per year based on Victim's age category.

Step 1 - Identify and Import the Data

Import dataset titled NYPD Shooting Incident Data (Historic) from city of New York site.

R code chunk "get_nypd_shooting_data" to tidy raw imported dataset

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

Step 2 - Tidy up datasets

Let's tidy up the raw imported dataset. For our analysis, we will only keep these fields and exclude rest of the fields.

Column Name	Column Description
OCCUR_DATE	Exact date of the shooting incident
BORO	Borough where the shooting incident occurred
VIC_AGE_GROUP	Victim's age within a category

R code chunk "tidy_nypd_raw_dataset" to tidy raw imported dataset

```
nypd_cases <- nypd_raw_dataset %>%
  #Change OCCUR_DATE from Char to Date Type
  mutate(OCCUR_DATE = mdy(OCCUR_DATE)) %>%
  #Remove feilds not needed for further analysis
  select(-c(INCIDENT_KEY,
            LOC_OF_OCCUR_DESC:PERP_RACE,
            VIC_SEX:Lon_Lat )) %>%
  #Change Feild names to lower case.
  rename( "occur_date" = "OCCUR_DATE",
          "boro" = "BORO" ,
          "vic_age_group" = "VIC_AGE_GROUP")

# Sample records from dataframe nypd_cases
head(nypd_cases)
```

```
## # A tibble: 6 x 4
##   occur_date OCCUR_TIME boro     vic_age_group
##   <date>     <time>     <chr>    <chr>
## 1 2021-05-27 21:30      QUEENS   18-24
## 2 2014-06-27 17:40      BRONX    18-24
## 3 2015-11-21 03:56      QUEENS   25-44
## 4 2015-10-09 18:30      BRONX    <18
## 5 2009-02-19 22:58      BRONX    45-64
## 6 2020-10-21 21:36      BROOKLYN 25-44
```

Step 3 - Transform and Analyze datasets

Create a dataframe having summary count of shooting incident by year.

R code chunk "analyze_nypd_shooting_sum_by_year" to summarise shooting incident based by year

```
nypd_shooting_sum_by_year <- nypd_cases %>%
# Create occur_year filed from occur_date
  mutate(occur_year = year(occur_date)) %>%
# Group by Year
  group_by(occur_year) %>%
# summarise by year into shooting_incidents field
  summarise(shooting_incidents= n() ) %>%
  select(occur_year, shooting_incidents)

# Sample records from dataframe nypd_shooting_sum_by_year
head(nypd_shooting_sum_by_year)
```

```
## # A tibble: 6 x 2
##   occur_year shooting_incidents
##        <dbl>              <int>
## 1       2006               2055
## 2       2007               1887
## 3       2008               1959
## 4       2009               1828
## 5       2010               1912
## 6       2011               1939
```

Create a dataframe having summary count of shooting incident per borough by year.

R code chunk "analyze_nypd_shooting_sum_by_boro" to tidy raw imported dataset

```
nypd_shooting_sum_by_boro <- nypd_cases %>%
# Create occur_year filed from occur_date
  mutate(occur_year = year(occur_date)) %>%
# Group by Year and borough
  group_by(occur_year, boro) %>%
# summarise by year into shooting_incidents field
  summarise(shooting_incidents = n() ) %>%
  select(occur_year, boro, shooting_incidents)
```

```
## `summarise()` has grouped output by 'occur_year'. You can override using the
## `.groups` argument.
```

```
# Sample records from dataframe nypd_shooting_sum_by_year
head(nypd_shooting_sum_by_boro)
```

```
## # A tibble: 6 x 3
## # Groups:   occur_year [2]
##   occur_year boro          shooting_incidents
##        <dbl> <chr>                      <int>
## 1       2006 BRONX                        568
## 2       2006 BROOKLYN                     850
## 3       2006 MANHATTAN                    288
## 4       2006 QUEENS                       296
## 5       2006 STATEN ISLAND                 53
## 6       2007 BRONX                        533
```

Create a dataframe having summary count of shooting by victim age_group per year.

R code chunk "analyze_nypd_shooting_sum_by_vic_age_group" to tidy raw imported dataset

```
nypd_shooting_sum_by_vic_age_group <- nypd_cases %>%
# Removing one record having invalid age group of 1022
  filter(!vic_age_group == "1022") %>%
  mutate(occur_year =year(occur_date)) %>%
  group_by(occur_year,vic_age_group) %>%
  summarise(shooting_incidents = n() ) %>%
  select(occur_year, vic_age_group, shooting_incidents)
```

```
## `summarise()` has grouped output by 'occur_year'. You can override using the
## `.groups` argument.
```

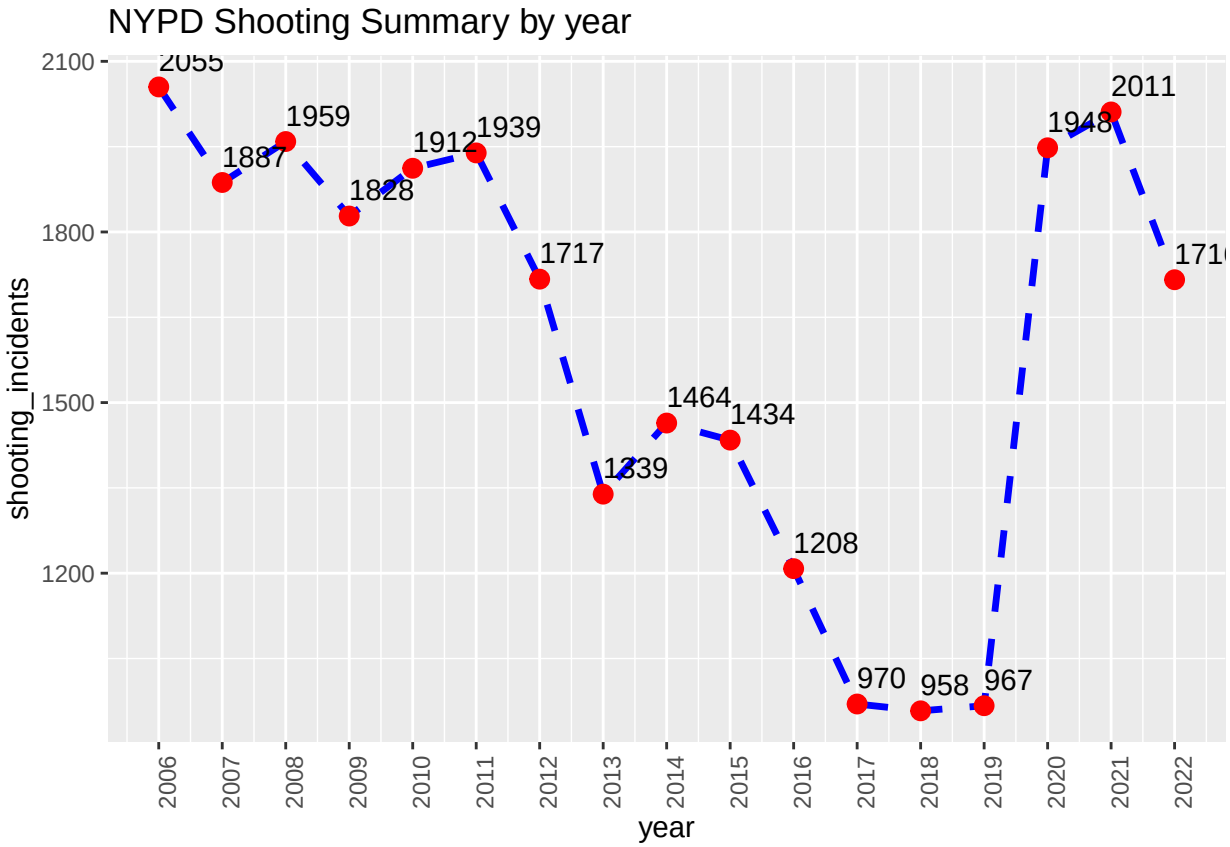
```
# Sample records from dataframe nypd_shooting_sum_by_vic_age_group
head(nypd_shooting_sum_by_vic_age_group)
```

```
## # A tibble: 6 x 3
## # Groups:   occur_year [1]
##   occur_year vic_age_group shooting_incidents
##        <dbl> <chr>                       <int>
## 1       2006 18-24                         849
## 2       2006 25-44                         813
## 3       2006 45-64                         111
## 4       2006 65+                            13
## 5       2006 <18                           264
## 6       2006 UNKNOWN                         5
```

Step 4 - Visualizing Data

1. Trend of shooting incident per year. R code chunk visualize_nypd_shooting_sum_by_year to visualize shooting incident per year trend

```
year=unique(nypd_shooting_sum_by_year$occur_year)
nypd_shooting_sum_by_year %>%
  ggplot(aes(x = occur_year, y = shooting_incidents)) +
  geom_line(linetype="dashed", color="blue", size=1.2)+
  geom_point(color="red", size=3) +
  geom_text(aes(label=shooting_incidents),hjust=0,vjust=-0.75) +
  scale_x_continuous("year", labels = as.character(year), breaks = year)+
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 90)) +
    labs(title = "NYPD Shooting Summary by year" )
```



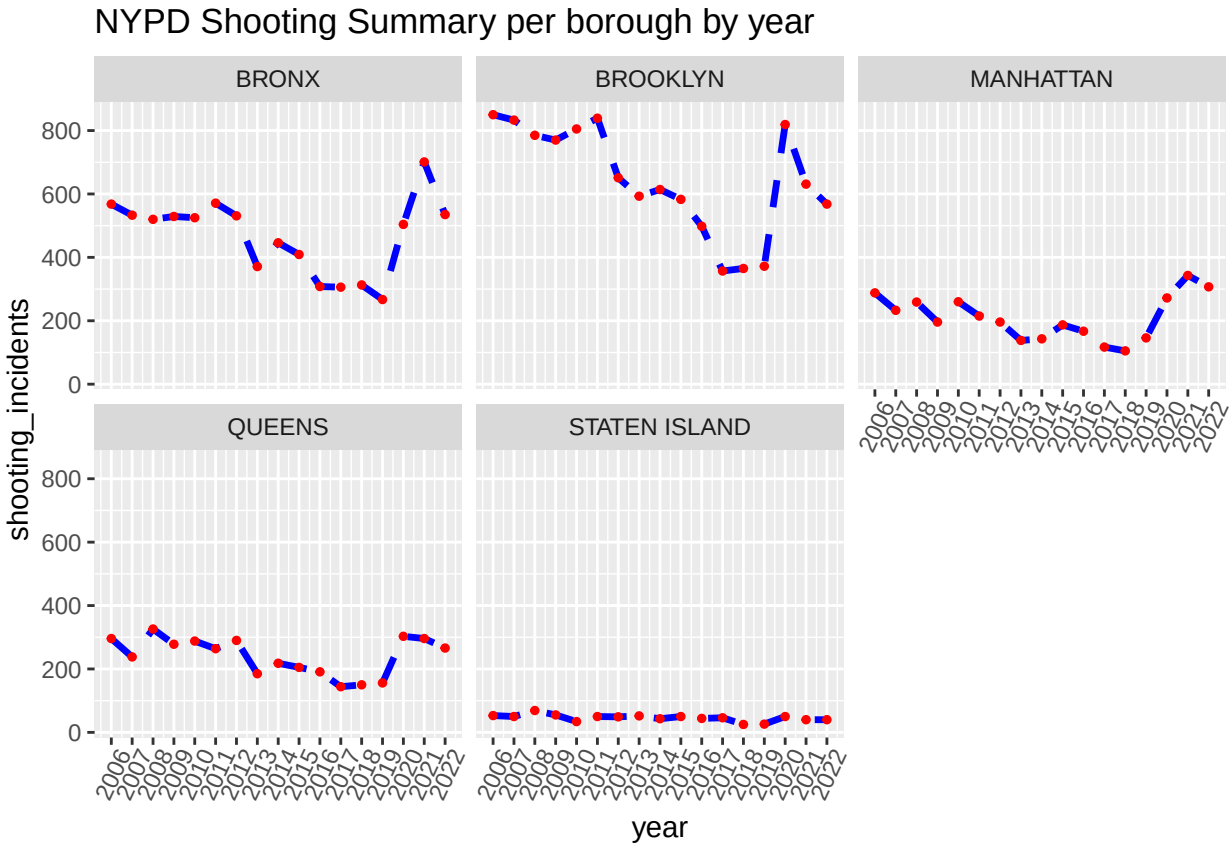
Based on above graph, below can be concluded:

- New York shooting trend was drastically reducing till year 2017.

- Year 2017 to 2019 was most stable period.
- After 2019 to 2021, there is rising trend of shooting trend. This might be related to crimes due to hardships because of Covid-19 pandemic.
- From year 2022, again there is downward trend which might be due to economic recovery and stability after Covid 19.

2. Trend of shooting per borough by year. R code chunk visualize_nypd_shooting_sum_by_boro to visualize trend of shooting per borough by yea

```
year=unique(nypd_shooting_sum_by_boro$occur_year)
nypd_shooting_sum_by_boro  %>%
  ggplot(aes(x = occur_year , y = shooting_incidents)) +
  geom_line(linetype="dashed", color="blue", size=1.2)+
  geom_point(color="red", size=1) +
  facet_wrap(~boro) +
  scale_x_continuous("year", labels = as.character(year), breaks = year)+
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 90)) +
    labs(title = "NYPD Shooting Summary per borough by year" ) +
  theme(axis.text.x = element_text(angle=65, vjust=0.6))
```

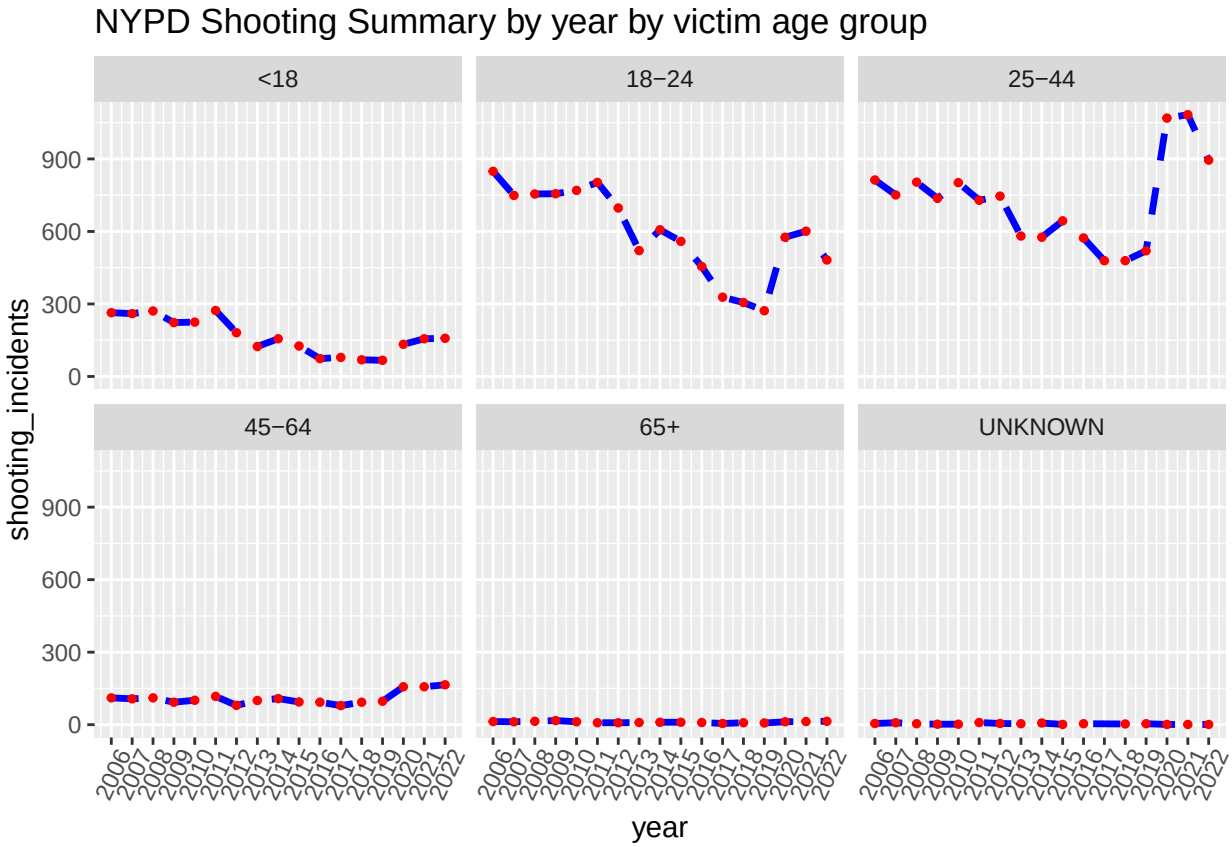


Based on above graph, below can be concluded:

- Bronx and Brooklyn are always most impacted borough. This might be due to population.
- Trends for all borough are aligned with trend of overall New Work.
- During Covid-19 there was rise in shooting incidents in all borough.

3. Trend of shooting per vicitim age group by year. **R code chunk visualize_nypd_shooting_sum_by_vic_age_gr to visualize trend of shooting per victim age group by year*

```
year=unique(nypd_shooting_sum_by_vic_age_group$occur_year)
nypd_shooting_sum_by_vic_age_group  %>%
  ggplot(aes(x = occur_year , y = shooting_incidents)) +
  geom_line(linetype="dashed", color="blue", size=1.2)+
  geom_point(color="red", size=1) +
  facet_wrap(~vic_age_group) +
  scale_x_continuous("year", labels = as.character(year), breaks = year)+
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 90)) +
    labs(title = "NYPD Shooting Summary by year by victim age group" ) +
  theme(axis.text.x = element_text(angle=65, vjust=0.6))
```



Based on above graph, below can be concluded:

- Most impacted and at risk age-group is of 18-24 and 25-44. That means young adult are at most risk of shooting incident. This might be because of money matters or failed relationships or drug/substance abuse .
- If we combine, age group 18-24 and 25-44, the age group (18-44) "adults" are at higher risk of being victim of the shooting incidents.
- There are also good amount of victim from Age group <18. These might be related to school shooting incidents.

Step 5 - Modelling Data

Based on the trend of victim age group, lets validate id there is any correlation between shooting incident and combined victim age group of 18-24 and 24-44. We will validate this using linear model.

Let's first create dataframe having year, summary of total shooting incident and shooting incident involving the victim of age group 18-44.

R code analyze_nypd_shooting_for_modelling to create dataframe for modelling

```
# Create dataframe combining "18-24" and "25-44" age groups.
nypd_victim_of_age_between_18_to_44 <-nypd_shooting_sum_by_vic_age_group %>%
  filter(vic_age_group %in% c("18-24", "25-44")) %>%
  group_by(occur_year) %>%
  summarise(victim_of_age_between_18_to_44 = sum(shooting_incidents) ) %>%
```

```
  select(occur_year, victim_of_age_between_18_to_44)

# Create dataframe for linear modelling having summary of total incident and incident with victim of ag
nypd_shooting_dataset_for_modelling  <- nypd_shooting_sum_by_year %>%
      full_join(nypd_victim_of_age_between_18_to_44)
```

```
## Joining with `by = join_by(occur_year)`
```

```
# Sample records from dataframe nypd_shooting_dataset_for_modelling
head(nypd_shooting_dataset_for_modelling)
```

```
## # A tibble: 6 x 3
##   occur_year shooting_incidents victim_of_age_between_18_to_44
##        <dbl>              <int>                          <int>
## 1       2006               2055                           1662
## 2       2007               1887                           1500
## 3       2008               1959                           1559
## 4       2009               1828                           1493
## 5       2010               1912                           1572
## 6       2011               1939                           1532
```

Linear model and summary of the linear model.

R code chunk model_linear_victim_of_age_between_18_to_44 for Linear model

```
mod <- lm(victim_of_age_between_18_to_44 ~ shooting_incidents, data= nypd_shooting_dataset_for_modellin
summary(mod)
```

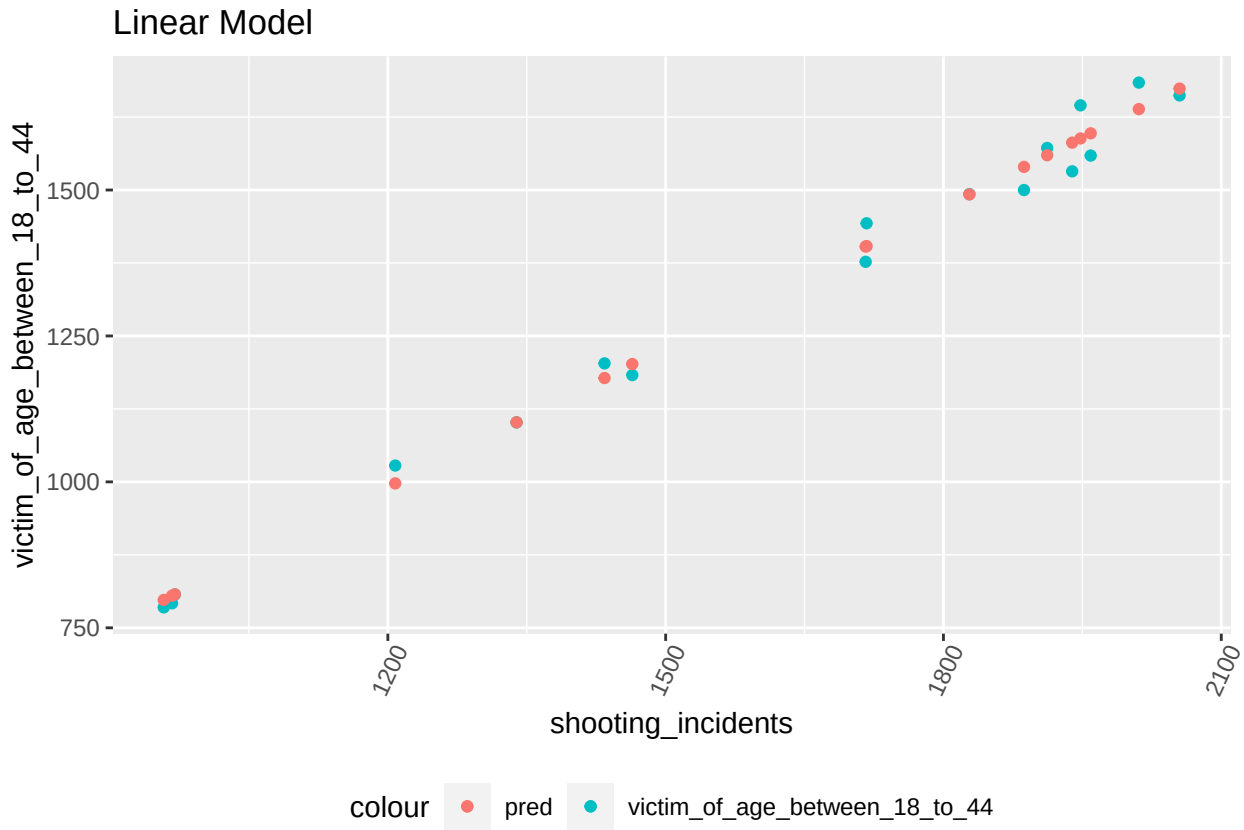
```
##
## Call:
## lm(formula = victim_of_age_between_18_to_44 ~ shooting_incidents,
##     data = nypd_shooting_dataset_for_modelling)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.133 -18.851  -0.397  25.104  56.681
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         32.8628    33.6908   0.975    0.345
## shooting_incidents   0.7985     0.0204  39.137   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.11 on 15 degrees of freedom
## Multiple R-squared:  0.9903, Adjusted R-squared:  0.9897
## F-statistic:  1532 on 1 and 15 DF,  p-value: < 2.2e-16
```

```
nypd_shooting_dataset_for_model_with_pred <- nypd_shooting_dataset_for_modelling %>%
  mutate(pred = predict(mod))
```

Let's visualize the linear model using prediction field and summary of victim of age group between 18 to 44

R code chunk visualize_model to visualise the liner model

```
nypd_shooting_dataset_for_model_with_pred  %>%
  ggplot() +
  geom_point(aes(x = shooting_incidents, y = victim_of_age_between_18_to_44, color = "victim_of_age_betw
  geom_point(aes(x = shooting_incidents, y = pred, color = "pred")) +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 90)) +
    labs(title = "Linear Model" ) +
  theme(axis.text.x = element_text(angle=65, vjust=0.6))
```



Based on above graph, below can be concluded:

- Prediction is linear.

- Model is reasonably correct in predicting in both lower and higher end.
- This also proves with current trend, age group of 18-44 are at higher risk of being victim of shooting incidents.

3. Final Summary

After analyzing NYPD shooting dataset, we can conclude below results of the questions stated at the start of the analysis :-

1. Is shooting rate increasing or decreasing year over year ? How much effective is law and order? Before Covid-19 Pandemic, shooting incident was drastically reducing. Year 2019 to 2021 was worse year for shooting incident might be due to hardship because of Covid-19. Last year's 2022 trend shows the decreasing trend, which proves great improvement and effectiveness of law enforcement team.
2. Are there any specific Boroughs in New York, which are more impacted by these shooting events ? With this analysis, we found Bronx and Brooklyn are always most impacted borough.
3. Are there any specific age groups which are more victimized due to shooting incidents ? This analysis clearing shows Age group of 18-44 are at higher risk of being victim of shooting incidents.

Bias Indentification

Personal Bias These are the two personal bias for this analysis:

- Shooting incident should have drastically decreased from 2013 due to strict police and law enforcement presence.
- Age group of 18 and below might have been be more engaged in shooting incidents as victim or Perpetrator.

Bias in Data There can be chances that all shooting incidents are not reported to the police. Missing data can give wrong interpretation.

To overcome bias, I trusted dataset and started analysis without any pre-judgment.

R code chunk to display session information

```
sessionInfo()
```

```
## R version 4.3.1 (2023-06-16)
## Platform: x86_64-apple-darwin20 (64-bit)
## Running under: macOS Ventura 13.5.2
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.3-x86_64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.3-x86_64/Resources/lib/libRlapack.dylib;  LAPACK v
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## time zone: Asia/Kolkata
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] lubridate_1.9.3 forcats_1.0.0   stringr_1.5.0   dplyr_1.1.3
##  [5] purrr_1.0.2     readr_2.1.4     tidyr_1.3.0     tibble_3.2.1
##  [9] ggplot2_3.4.4   tidyverse_2.0.0
##
## loaded via a namespace (and not attached):
##  [1] bit_4.0.5         gtable_0.3.4      crayon_1.5.2      compiler_4.3.1
##  [5] tidyselect_1.2.0  parallel_4.3.1    scales_1.2.1      yaml_2.3.7
```

```
##  [9] fastmap_1.1.1     R6_2.5.1          labeling_0.4.3    generics_0.1.3
## [13] curl_5.1.0        knitr_1.44        munsell_0.5.0     pillar_1.9.0
## [17] tzdb_0.4.0        rlang_1.1.1       utf8_1.2.4        stringi_1.7.12
## [21] xfun_0.40         bit64_4.0.5       timechange_0.2.0  cli_3.6.1
## [25] withr_2.5.1       magrittr_2.0.3    digest_0.6.33     grid_4.3.1
## [29] vroom_1.6.4       rstudioapi_0.15.0 hms_1.1.3         lifecycle_1.0.3
## [33] vctrs_0.6.4       evaluate_0.22     glue_1.6.2        farver_2.1.1
## [37] fansi_1.0.5       colorspace_2.1-0  rmarkdown_2.25    tools_4.3.1
## [41] pkgconfig_2.0.3   htmltools_0.5.6.1
```