Final Project 1: NYPD Shooting Incident Data Report

Tapas

11/8/2021

NYPD Shooting Incident Data Report

Summary: This project analyzes NYPD shooting dataset between year 2006 and 2020. My analysis focused on the number of shooting incident trend over the past 15 years and age group of the victims. Also, I tried to build a linear regression model to show the correlation between shooting incidents and deaths. Based on that analysis, my predicted death count was close to the actual count.

Load the NYPD Shooting data

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
url_in <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
```

Read the data

```
nypd_data<-read_csv(url_in)
```

```
## Rows: 23585 Columns: 19

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (10): OCCUR_DATE, BORO, LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX, PERP_R...
## 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.
```

Analyze the data

select few columns and add a year column:

```
nypd_select1 <- nypd_data %>% select (OCCUR_DATE,BORO,PERP_AGE_GROUP, PERP_RACE,VIC_AGE_GROUP,VIC_SEX,S
  mutate(death=case_when(STATISTICAL_MURDER_FLAG=="TRUE"~1,STATISTICAL_MURDER_FLAG=="FALSE"~0))
nypd_select2 <- nypd_select1 %>% mutate(year = str_sub(OCCUR_DATE,-4,-1))
nypd_select3 <- nypd_select2  %>% filter(PERP_AGE_GROUP != 'NA')
```

Number of Shootings per year

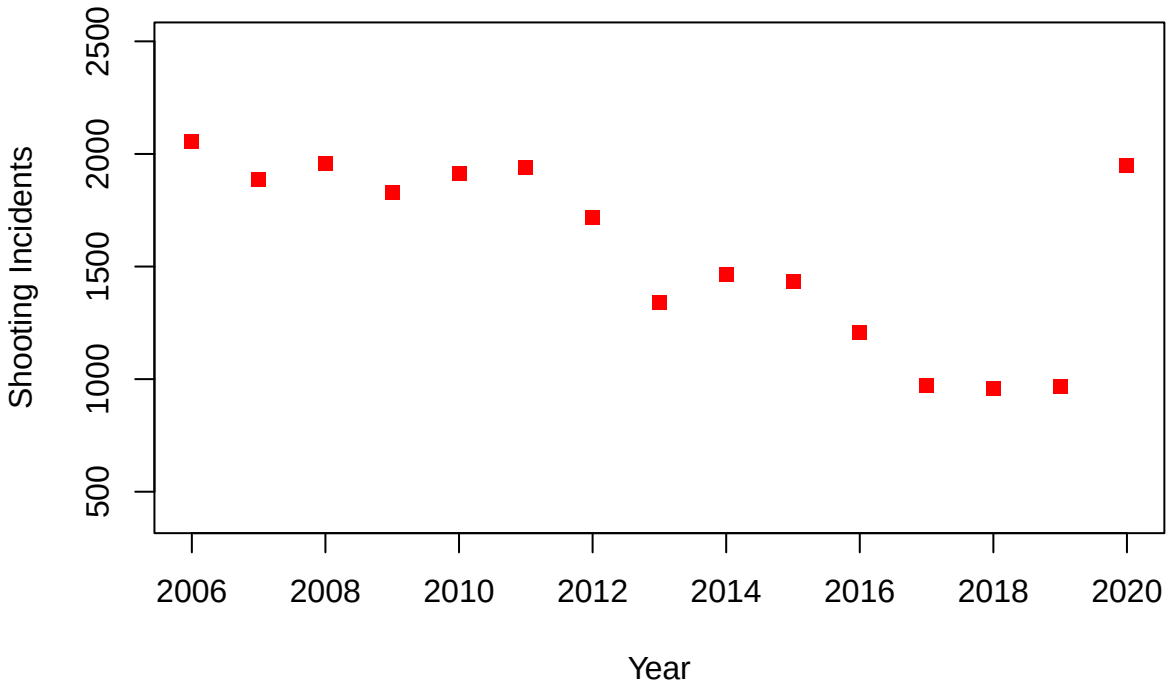
```
shooting_per_year <- nypd_select2 %>%
    group_by(year) %>%
    summarise(count_shootings = n())
```

```
victims_by_age <- nypd_select3 %>%
    group_by(VIC_AGE_GROUP) %>%
    summarise(count_victims = n())
```

Visualize the data

Shootings per year embed plot:

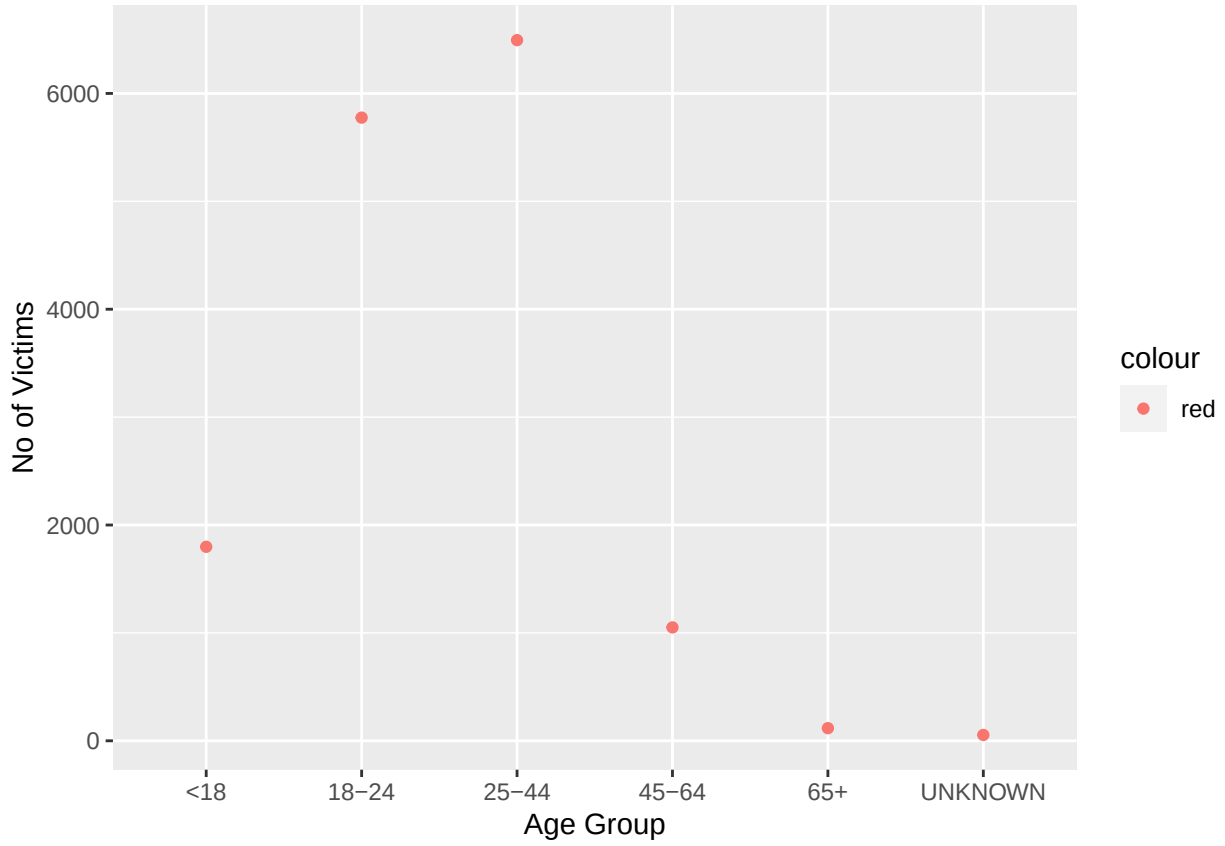
```
plot(shooting_per_year,col="red",ylim = c(400,2500),pch = 15,xlab = "Year",ylab = "Shooting Incidents")
```



```
##ggplot(shooting_per_year, aes(x=year,y=count_shootings)) + geom_bar(stat="identity")
```

Victims by age plot:

```
library(ggplot2)
qplot(VIC_AGE_GROUP,count_victims, data = victims_by_age,col="red",xlab = "Age Group",ylab = "No of Vic
```



Model the data: Linear Model

This model predicts the yearly murder count based on shooting data. As shown below the predicted death count (red dots) is close to the actual counts (blue dot)

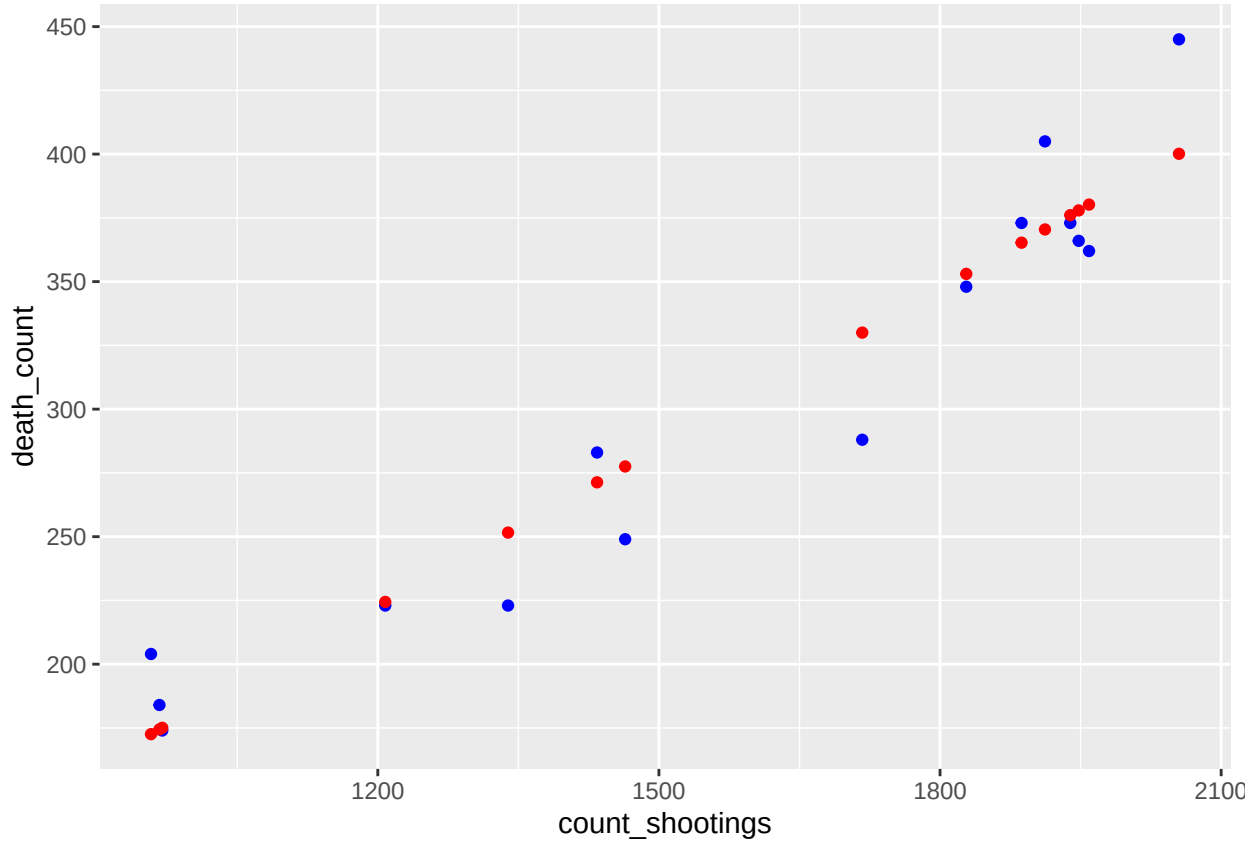
```
death_per_year <- nypd_select2 %>%
    group_by(year) %>%
    summarise(death_count=sum(death))

nypd_combined<-shooting_per_year%>%
  full_join(death_per_year,
            by=c("year"))

mod<-lm(death_count ~ count_shootings,data=nypd_combined)
summary(mod)
```

```
##
## Call:
## lm(formula = death_count ~ count_shootings, data = nypd_combined)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.010 -15.070  -1.422  10.634  44.875
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -26.16775   27.23773  -0.961    0.354
## count_shootings   0.20744    0.01681  12.338  1.5e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.41 on 13 degrees of freedom
## Multiple R-squared:  0.9213, Adjusted R-squared:  0.9153
## F-statistic: 152.2 on 1 and 13 DF,  p-value: 1.497e-08
```

```r
## create new dataset NYPD shhoting death with prediction
x_grid <- seq(1,2500)
new_df <- tibble(count_shootings=x_grid)
nypd_combined_w_pred <- nypd_combined %>% mutate(pred=predict(mod))
nypd_combined_w_pred %>% ggplot() +
  geom_point(aes(x=count_shootings,y=death_count),color="blue")+
  geom_point(aes(x=count_shootings,y=pred),color="red")
```



Conclusion

My analysis focused on the number of shooting incident trend over the past 15 years and age group of the victims. The linear regression model predicted murder count was close to the actual count based on the shooting incidents. I may have bias to analyze data based on victim age thinking that number of incidents and age may be a factor in shooting incidents.