

NYPD data project

2022-11-03

```
knitr::opts_chunk$set(echo = TRUE)
```

```
library(tidyverse)
```

```
## -- Attaching packages ----- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr  1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
##
## The following objects are masked from 'package:base':
##
##   date, intersect, setdiff, union
```

```
library(tibble)
```

Getting data

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

Reading the data

```
nypd_data <- read.csv(url_data)
```

Tidying the data

```
summary(nypd_data)
```

```
## INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO
## Min. : 9953245 Length:25596 Length:25596 Length:25596
## 1st Qu.: 61593633 Class :character Class :character Class :character
## Median : 86437258 Mode :character Mode :character Mode :character
## Mean :112382648
## 3rd Qu.:166660833
## Max. :238490103
##
## PRECINCT JURISDICTION_CODE LOCATION_DESC STATISTICAL_MURDER_FLAG
## Min. : 1.00 Min. :0.0000 Length:25596 Length:25596
## 1st Qu.: 44.00 1st Qu.:0.0000 Class :character Class :character
## Median : 69.00 Median :0.0000 Mode :character Mode :character
## Mean : 65.87 Mean :0.3316
## 3rd Qu.: 81.00 3rd Qu.:0.0000
## Max. :123.00 Max. :2.0000
## NA's :2
## PERP_AGE_GROUP PERP_SEX PERP_RACE VIC_AGE_GROUP
## Length:25596 Length:25596 Length:25596 Length:25596
## Class :character Class :character Class :character Class :character
## Mode :character Mode :character Mode :character Mode :character
##
##
##
## VIC_SEX VIC_RACE X_COORD_CD Y_COORD_CD
## Length:25596 Length:25596 Min. : 914928 Min. :125757
## Class :character Class :character 1st Qu.:1000011 1st Qu.:182782
## Mode :character Mode :character Median :1007715 Median :194038
## Mean :1009455 Mean :207894
## 3rd Qu.:1016838 3rd Qu.:239429
## Max. :1066815 Max. :271128
##
## Latitude Longitude Lon_Lat
## Min. :40.51 Min. : -74.25 Length:25596
## 1st Qu.:40.67 1st Qu.: -73.94 Class :character
## Median :40.70 Median : -73.92 Mode :character
## Mean :40.74 Mean : -73.91
## 3rd Qu.:40.82 3rd Qu.: -73.88
## Max. :40.91 Max. : -73.70
##
```

```
nypd_data <- nypd_data %>%
  select(OCCUR_DATE:VIC_RACE) %>%
  select(-c(PRECINCT:LOCATION_DESC)) %>%
  select(-c(PERP_SEX, PERP_RACE, VIC_SEX, VIC_RACE))
nypd_data <- nypd_data %>% filter(PERP_AGE_GROUP > 0)
nypd_data <- nypd_data %>% filter(VIC_AGE_GROUP > 0)
nypd_data$STATISTICAL_MURDER_FLAG <- replace(nypd_data$STATISTICAL_MURDER_FLAG, nypd_data$STATISTICAL_MURDER_FLAG == 0, 1)
nypd_data$STATISTICAL_MURDER_FLAG <- replace(nypd_data$STATISTICAL_MURDER_FLAG, nypd_data$STATISTICAL_MURDER_FLAG == 1, 0)
nypd_data <- nypd_data %>%
```

```

  rename(death = 'STATISTICAL_MURDER_FLAG',
         date = 'OCCUR_DATE',
         time = 'OCCUR_TIME')
nypd_data$death <- as.double(nypd_data$death)
nypd_data <- nypd_data %>%
  mutate(date = mdy(date)) %>%
  mutate(shooting = 1)
nypd_data <- nypd_data[(nypd_data$PERP_AGE_GROUP != "1020" & nypd_data$PERP_AGE_GROUP != "224" & nypd_data$VIC_AGE_GROUP != "UNKNOWN"), ]
nypd_data <- nypd_data %>%
  mutate(month = month(date), year = year(date))
summary(nypd_data)

```

```

##      date              time      BORO      death
##  Min.   :2006-01-01   Length:10579   Length:10579   Min.   :0.0000
##  1st Qu.:2009-04-01   Class :character   Class :character   1st Qu.:0.0000
##  Median :2013-01-23   Mode  :character   Mode  :character   Median :0.0000
##  Mean   :2013-07-23                                     Mean   :0.2495
##  3rd Qu.:2017-09-10                                     3rd Qu.:0.0000
##  Max.   :2021-12-31                                     Max.   :1.0000
##  PERP_AGE_GROUP    VIC_AGE_GROUP      shooting      month
##  Length:10579      Length:10579      Min.   :1   Min.   : 1.000
##  Class :character   Class :character   1st Qu.:1   1st Qu.: 4.000
##  Mode  :character   Mode  :character   Median :1   Median : 7.000
##                                     Mean   :1   Mean   : 6.693
##                                     3rd Qu.:1   3rd Qu.: 9.000
##                                     Max.   :1   Max.   :12.000
##      year
##  Min.   :2006
##  1st Qu.:2009
##  Median :2013
##  Mean   :2013
##  3rd Qu.:2017
##  Max.   :2021

```

```

all_shootings = sum(nypd_data$shooting)
all_death = sum(nypd_data$death)
partial = all_death / all_shootings
partial

```

```
## [1] 0.2494565
```

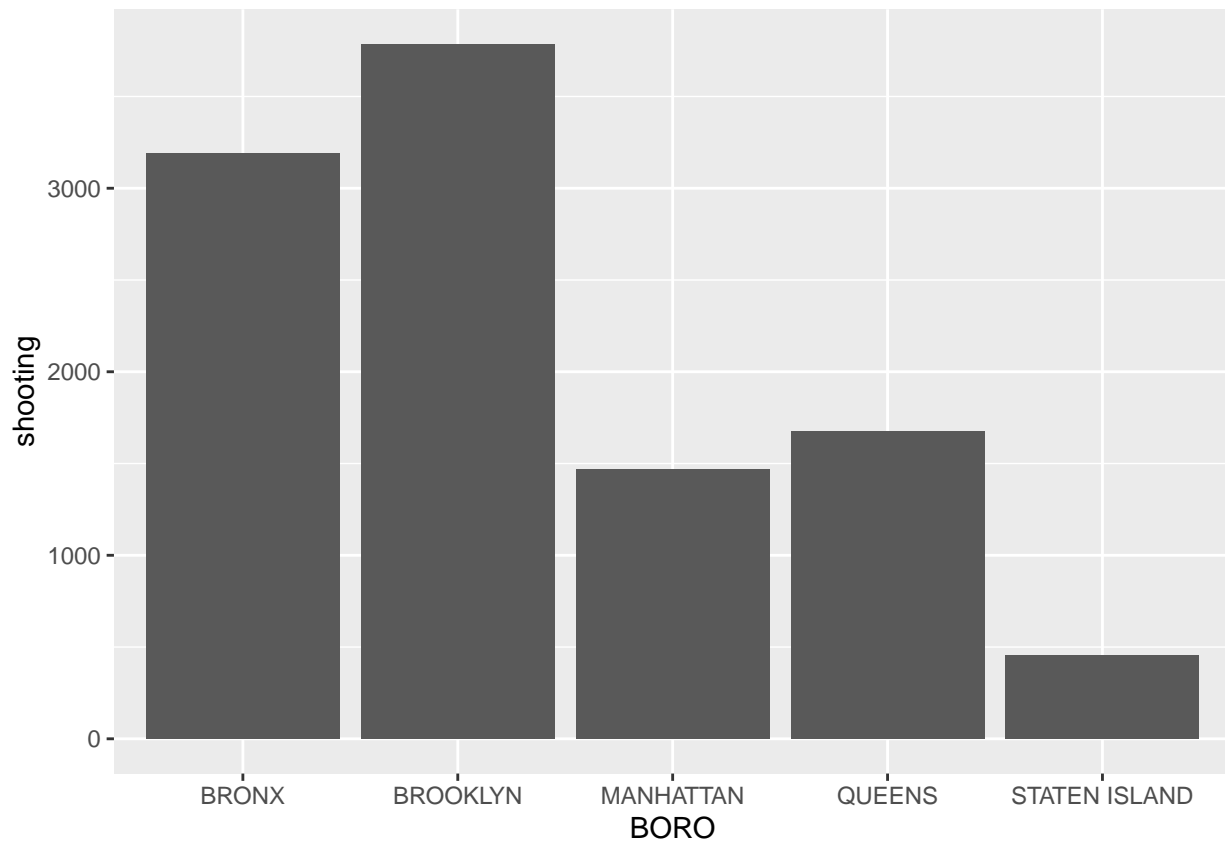
Visualizing and Analyzing the data

```

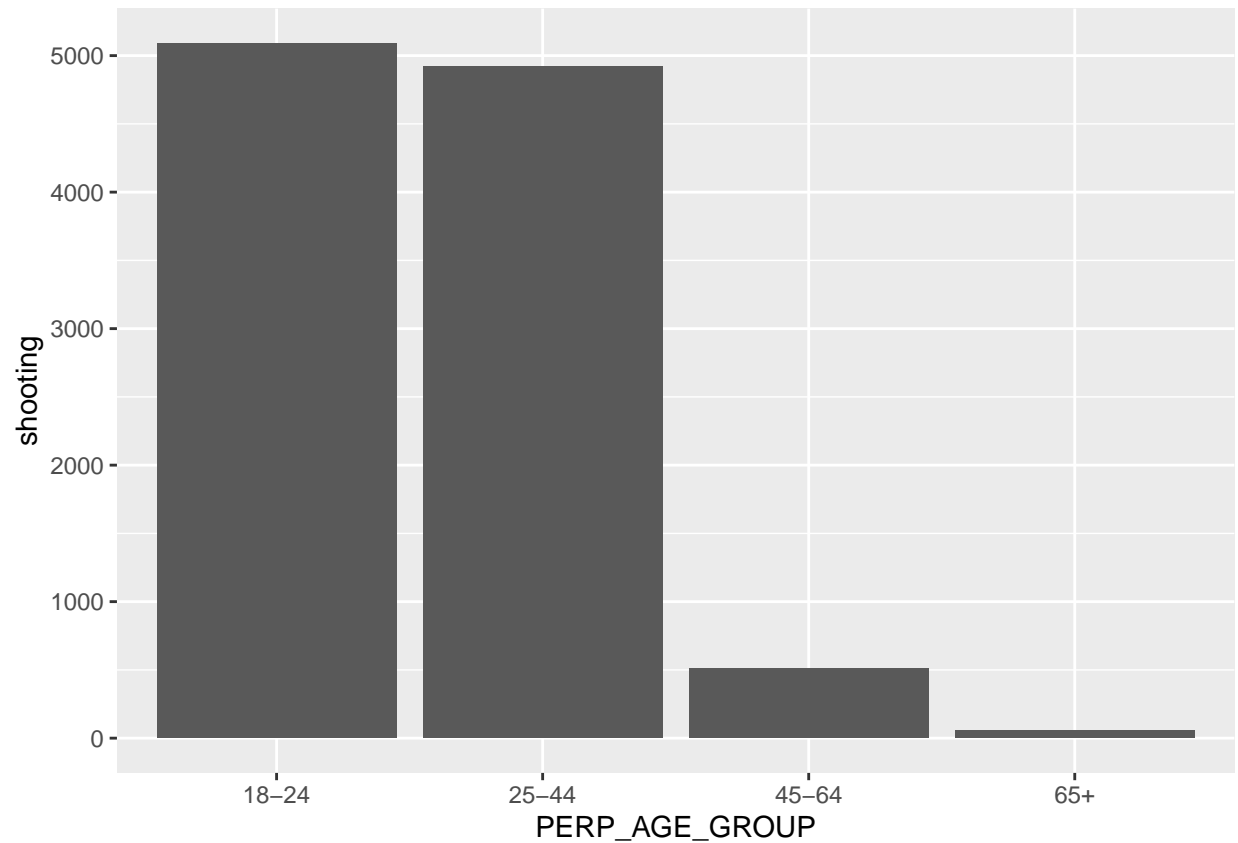
shootings_by_BORO <- nypd_data %>%
  group_by(BORO) %>%
  summarize(death = sum(death), shooting = sum(shooting)) %>%
  mutate(percentage = (death / shooting) * 100) %>%
  ungroup()

```

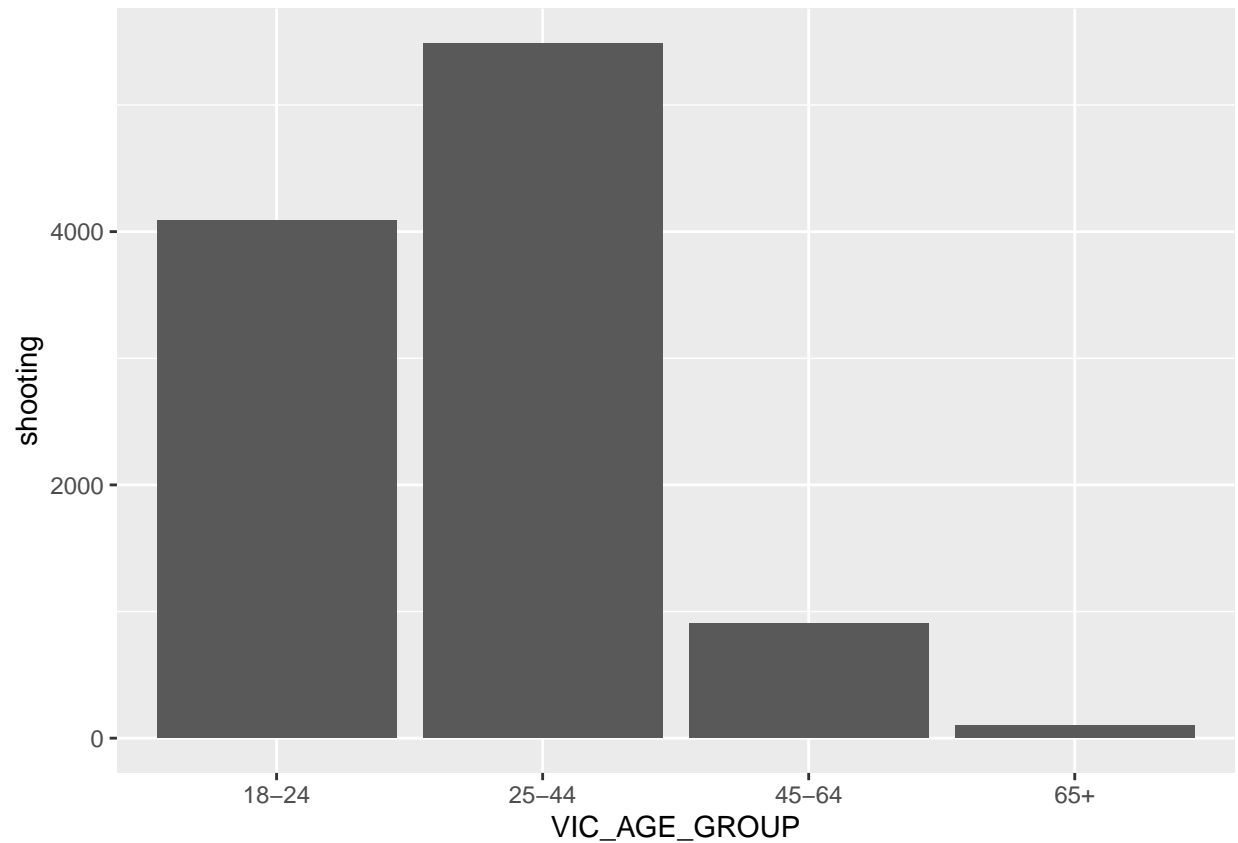
```
shootings_by_BORO %>%
  ggplot(aes(x = BORO, y = shooting)) + geom_bar(stat = "identity")
```



```
shootings_by_perp_age <- nypd_data %>%
  group_by(PERP_AGE_GROUP) %>%
  summarize(death = sum(death), shooting = sum(shooting)) %>%
  mutate(percentage = (death / shooting) * 100) %>%
  ungroup()
shootings_by_perp_age %>%
  ggplot(aes(x = PERP_AGE_GROUP, y = shooting)) + geom_bar(stat = "identity")
```

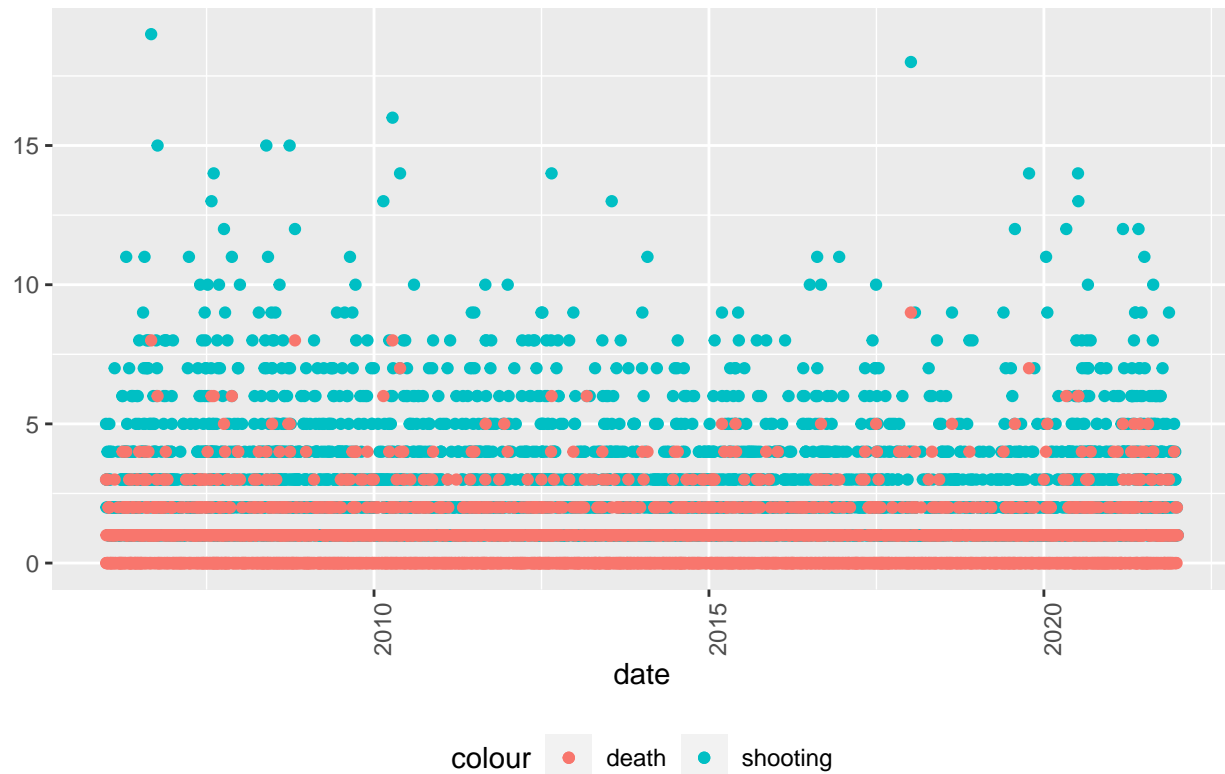


```
shootings_by_vic_age<- nypd_data %>%
  group_by(VIC_AGE_GROUP) %>%
  summarize(death = sum(death), shooting = sum(shooting)) %>%
  mutate(percentage = (death / shooting) * 100) %>%
  ungroup()
shootings_by_vic_age %>%
  ggplot(aes(x = VIC_AGE_GROUP, y = shooting)) + geom_bar(stat = "identity")
```



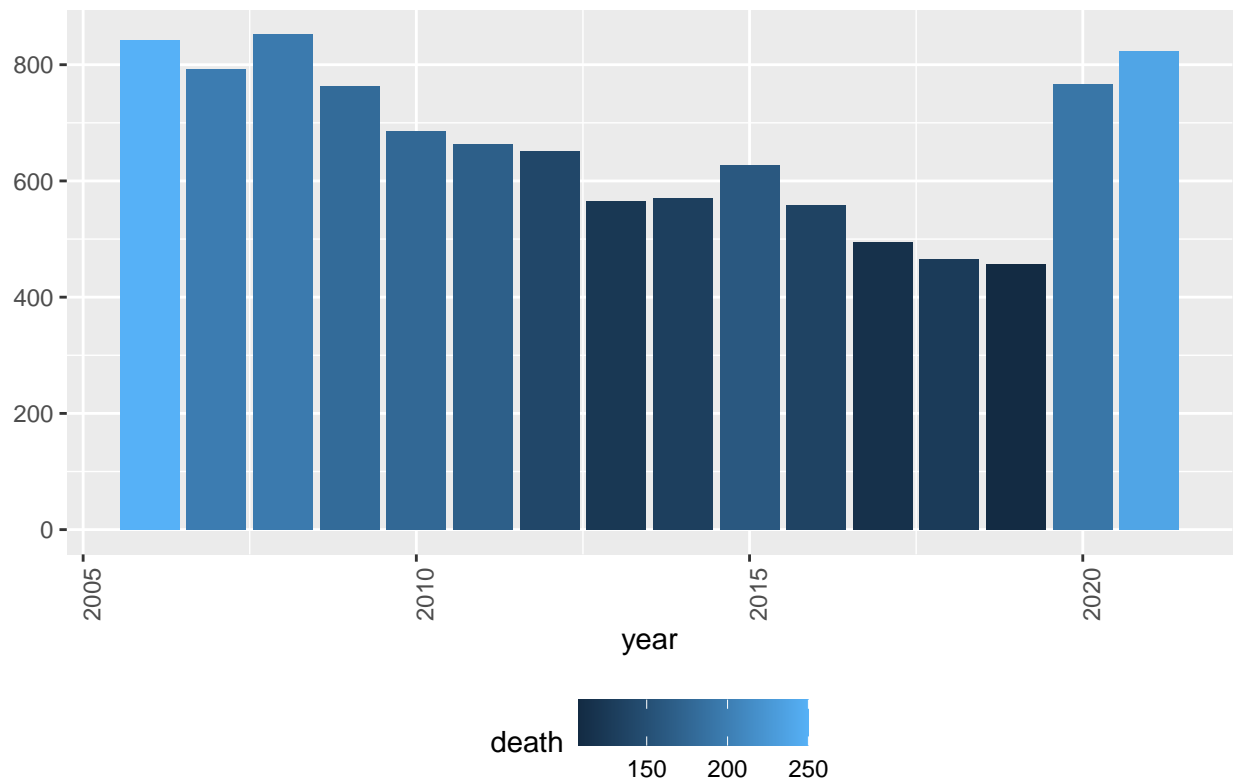
```
shootings_by_date <- nypd_data %>%
  group_by(date) %>%
  summarize(death = sum(death), shooting = sum(shooting)) %>%
  mutate(percentage = (death / shooting) * 100) %>%
  ungroup()
shootings_by_date %>%
  ggplot(aes(x = date, y = shooting)) + geom_point(aes(color = "shooting")) + geom_point(aes(y = death,
  theme(legend.position = "bottom", axis.text.x = element_text(angle = 90)) +
  labs(title = "NYPD data", y = NULL)
```

NYPD data



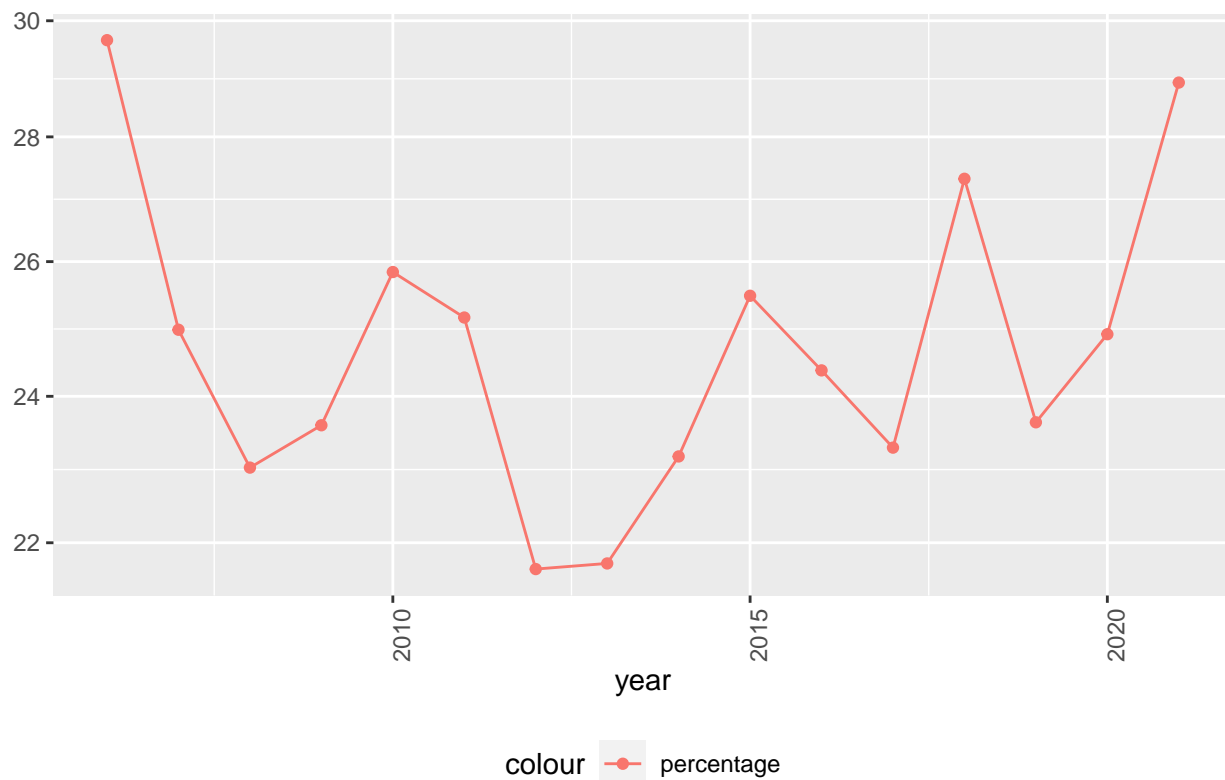
```
shootings_by_year <- nypd_data %>%
  group_by(year) %>%
  summarize(death = sum(death), shooting = sum(shooting)) %>%
  mutate(percentage = (death / shooting) * 100) %>%
  ungroup()
shootings_by_year %>%
  ggplot(aes(x = year, y = shooting, fill = death)) + geom_bar(stat = "identity") +
  theme(legend.position = "bottom", axis.text.x = element_text(angle = 90)) +
  labs(title = "NYPD data", y = NULL)
```

NYPD data



```
shootings_by_year %>%
  ggplot(aes(x = year, y = percentage)) + geom_line(aes(color = "percentage")) + geom_point(aes(color =
    scale_y_log10() +
    theme(legend.position = "bottom", axis.text.x = element_text(angle = 90)) +
    labs(title = "NYPD data", y = NULL)
```

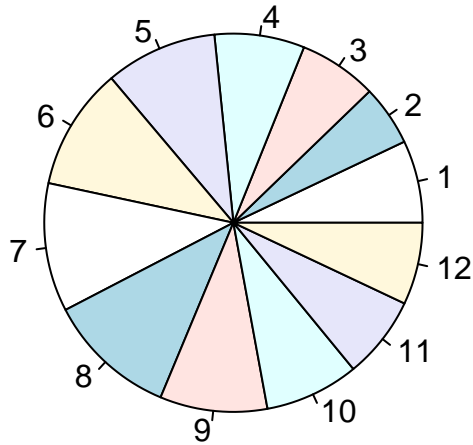

NYPD data



```
shootings_by_month <- nypd_data %>%
  group_by(month) %>%
  summarize(death = sum(death), shooting = sum(shooting)) %>%
  mutate(percentage = (death / shooting) * 100) %>%
  ungroup()
all_shootings = sum(shootings_by_month$shooting)
shootings_by_month <- shootings_by_month %>%
  mutate(part_of_whole = shooting / all_shootings)
shootings_by_month
```

```
## # A tibble: 12 x 5
##   month death shooting percentage part_of_whole
##   <dbl> <dbl>   <dbl>      <dbl>      <dbl>
## 1     1    191    746      25.6      0.0705
## 2     2    128    547      23.4      0.0517
## 3     3    166    708      23.4      0.0669
## 4     4    214    814      26.3      0.0769
## 5     5    273   1011      27.0      0.0956
## 6     6    252   1108      22.7      0.105
## 7     7    270   1162      23.2      0.110
## 8     8    275   1173      23.4      0.111
## 9     9    269    971      27.7      0.0918
## 10    10    209    851      24.6      0.0804
## 11    11    187    746      25.1      0.0705
## 12    12    205    742      27.6      0.0701
```

```
pie(shootings_by_month$part_of_whole)
```



From the analysis, months May, June, July and August have higher number of shootings than other months. How and why this is the case is an interesting question.

Modeling the data

Predicting number of deaths with month and shooting number as independent variables.

```
mod1 <- lm(death ~ BORO + PERP_AGE_GROUP + VIC_AGE_GROUP + year + month, data = nypd_data)
mod2 <- lm(death ~ BORO + PERP_AGE_GROUP + VIC_AGE_GROUP, data = nypd_data)
summary(mod1)
```

```
##
## Call:
## lm(formula = death ~ BORO + PERP_AGE_GROUP + VIC_AGE_GROUP +
##     year + month, data = nypd_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.5115 -0.2676 -0.2183 -0.1785  0.8247
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)          2.1126369  1.7755009   1.190  0.23412
## BOROBROOKLYN        -0.0249237  0.0103808  -2.401  0.01637 *
## BOROMANHATTAN        -0.0293857  0.0135917  -2.162  0.03064 *
## BOROQUEENS           -0.0186006  0.0130026  -1.431  0.15259
## BOROSTATEN ISLAND    -0.0101094  0.0216275  -0.467  0.64020
## PERP_AGE_GROUP25-44  0.0588642  0.0088803   6.629 3.55e-11 ***
## PERP_AGE_GROUP45-64  0.1360379  0.0205414   6.623 3.70e-11 ***
## PERP_AGE_GROUP65+    0.1870450  0.0582256   3.212  0.00132 **
## VIC_AGE_GROUP25-44   0.0164761  0.0092067   1.790  0.07355 .
## VIC_AGE_GROUP45-64   0.0279200  0.0163164   1.711  0.08708 .
## VIC_AGE_GROUP65+     0.0957144  0.0445115   2.150  0.03155 *
## year                 -0.0009447  0.0008823  -1.071  0.28432
## month                0.0012021  0.0013208   0.910  0.36280
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4307 on 10566 degrees of freedom
## Multiple R-squared:  0.01044,    Adjusted R-squared:  0.009319
## F-statistic: 9.292 on 12 and 10566 DF,  p-value: < 2.2e-16
```

```
summary(mod2)
```

```
##
## Call:
## lm(formula = death ~ BORO + PERP_AGE_GROUP + VIC_AGE_GROUP, data = nypd_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.5016 -0.2692 -0.2197 -0.1902  0.8098
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.219722   0.009684  22.689 < 2e-16 ***
## BOROBROOKLYN   -0.024089   0.010355  -2.326  0.02002 *
## BOROMANHATTAN  -0.029505   0.013585  -2.172  0.02989 *
## BOROQUEENS      -0.018504   0.013002  -1.423  0.15473
## BOROSTATEN ISLAND -0.009191   0.021618  -0.425  0.67074
## PERP_AGE_GROUP25-44 0.057998   0.008842   6.559 5.66e-11 ***
## PERP_AGE_GROUP45-64 0.134707   0.020495   6.573 5.18e-11 ***
## PERP_AGE_GROUP65+  0.185851   0.058209   3.193  0.00141 **
## VIC_AGE_GROUP25-44  0.015610   0.009166   1.703  0.08859 .
## VIC_AGE_GROUP45-64  0.026935   0.016285   1.654  0.09817 .
## VIC_AGE_GROUP65+   0.096008   0.044505   2.157  0.03101 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4307 on 10568 degrees of freedom
## Multiple R-squared:  0.01025,    Adjusted R-squared:  0.009318
## F-statistic: 10.95 on 10 and 10568 DF,  p-value: < 2.2e-16
```

```
mod3 <- glm(death ~ BORO + PERP_AGE_GROUP + VIC_AGE_GROUP + year + month, data = nypd_data, family = "b")
summary(mod3)
```

```
##
## Call:
## glm(formula = death ~ BORO + PERP_AGE_GROUP + VIC_AGE_GROUP +
##      year + month, family = "binomial", data = nypd_data)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.2318  -0.7883  -0.7002  -0.6355   1.8511
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      8.948512   9.555536   0.936  0.34903
## BOROBROOKLYN    -0.133683   0.055722  -2.399  0.01643 *
## BOROMANHATTAN   -0.158668   0.073728  -2.152  0.03139 *
## BOROQUEENS      -0.099146   0.069754  -1.421  0.15521
## BOROSTATEN ISLAND -0.053502   0.115371  -0.464  0.64283
## PERP_AGE_GROUP25-44 0.321362   0.048301   6.653 2.87e-11 ***
## PERP_AGE_GROUP45-64 0.676135   0.101646   6.652 2.89e-11 ***
## PERP_AGE_GROUP65+   0.884321   0.275997   3.204  0.00135 **
## VIC_AGE_GROUP25-44  0.091378   0.050276   1.818  0.06914 .
## VIC_AGE_GROUP45-64  0.150504   0.085992   1.750  0.08008 .
## VIC_AGE_GROUP65+   0.468290   0.218890   2.139  0.03240 *
## year             -0.005102   0.004748  -1.074  0.28262
## month             0.006541   0.007128   0.918  0.35882
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 11885  on 10578  degrees of freedom
## Residual deviance: 11777  on 10566  degrees of freedom
## AIC: 11803
##
## Number of Fisher Scoring iterations: 4
```

```
mod4 <- glm(death ~ BORO + PERP_AGE_GROUP + VIC_AGE_GROUP, data = nypd_data, family = "binomial")
summary(mod4)
```

```
##
## Call:
## glm(formula = death ~ BORO + PERP_AGE_GROUP + VIC_AGE_GROUP,
##      family = "binomial", data = nypd_data)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.2087  -0.7915  -0.7023  -0.6542   1.8150
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.27408   0.05306 -24.012 < 2e-16 ***
## BOROBROOKLYN   -0.12908   0.05557  -2.323  0.02020 *
## BOROMANHATTAN  -0.15903   0.07368  -2.158  0.03090 *
## BOROQUEENS     -0.09859   0.06975  -1.414  0.15750
## BOROSTATEN ISLAND -0.04865   0.11532  -0.422  0.67313
```

```

## PERP_AGE_GROUP25-44  0.31658      0.04810      6.582 4.63e-11 ***
## PERP_AGE_GROUP45-64  0.66883      0.10136      6.598 4.16e-11 ***
## PERP_AGE_GROUP65+    0.87767      0.27586      3.182  0.00146 **
## VIC_AGE_GROUP25-44   0.08655      0.05005      1.729  0.08377 .
## VIC_AGE_GROUP45-64   0.14485      0.08582      1.688  0.09146 .
## VIC_AGE_GROUP65+     0.46978      0.21879      2.147  0.03178 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 11885  on 10578  degrees of freedom
## Residual deviance: 11779  on 10568  degrees of freedom
## AIC: 11801
##
## Number of Fisher Scoring iterations: 4

```

From the first regression, I see that the coefficients for year and month are not statistically significant, so I drop them and do another regression with the other variables. Then I do a logistic regression. I conclude that boroughs, perpetrator age groups and victim age groups are statistically significant variables when predicting if there is a death that results from a shooting incident in NYC. There may be a bias because there could be other important variables that have been omitted in this analysis.