NYPD Shooting Incident Data (Historic) analysis

Alexey Sokolov

2022-07-04

Read the data and show summary

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

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:25596       Length:25596      Length:25596
##  1st Qu.: 61593633   Class :character   Class1:hms        Class :character
##  Median : 86437258   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :112382648                      Mode  :numeric
##  3rd Qu.:166660833
##  Max.   :238490103
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.0000    Length:25596       Mode :logical
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character   FALSE:20668
##  Median : 69.00   Median :0.0000    Mode  :character   TRUE :4928
##  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
##
```

Transform the data - remove unnecessesary columns and convert OCCUR_DATE to date format

```
data <- data %>% select(-c(INCIDENT_KEY, PRECINCT, JURISDICTION_CODE,
                           LOCATION_DESC, STATISTICAL_MURDER_FLAG, X_COORD_CD,
                           Y_COORD_CD, Latitude, Longitude, Lon_Lat))
data <- data %>% mutate(OCCUR_DATE = mdy(OCCUR_DATE))
```

Summary of data

```
summary(data)
```

```
##    OCCUR_DATE          OCCUR_TIME           BORO           PERP_AGE_GROUP
##  Min.   :2006-01-01   Length:25596      Length:25596       Length:25596
##  1st Qu.:2009-05-10   Class1:hms        Class :character   Class :character
##  Median :2012-08-26   Class2:difftime   Mode  :character   Mode  :character
##  Mean   :2013-06-13   Mode  :numeric
##  3rd Qu.:2017-07-01
##  Max.   :2021-12-31
##    PERP_SEX          PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  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_RACE
##  Length:25596
##  Class :character
##  Mode  :character
##
##
##
```

Check for missing values

```
sum(is.na(data$OCCUR_DATE))
```

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

```
sum(is.na(data$OCCUR_TIME))
```

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

```
sum(is.na(data$BORO))
```

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

```
sum(is.na(data$PERP_AGE_GROUP))
```

```
## [1] 9344
```

```
sum(is.na(data$PERP_SEX))
```

```
## [1] 9310
```

```
sum(is.na(data$PERP_RACE))
```

```
## [1] 9310
```

```
sum(is.na(data$VIC_AGE_GROUP))
```

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

```
sum(is.na(data$VIC_SEX))
```

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

```
sum(is.na(data$VIC_RACE))
```

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

From the command above it's clear that we have missing values in PERP_AGE_GROUP, PERP_SEX and PERP_RACE. There are several ways to deal with missing values:

1. Remove rows with missing values
2. Do an Imputation (fill in the missing values with some number), for example we can use average values
3. Imputation with extension. We can add additional column that will have TRUE value if this row has imputed value and FALSE otherwise. This way any model we want to build will include imputation fact in it and it will be more correct

For this I would suggest just to remove rows with missing values for every column with missing values - PERP_AGE_GROUP, PERP_SEX and PERP_RACE and have three additional datasets, this way we can save non missing values in other columns.

Remove missing values

```
data_perp_age_group <- data %>% drop_na(PERP_AGE_GROUP)
sum(is.na(data_perp_age_group$PERP_AGE_GROUP))
```

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

```
data_perp_sex <- data %>% drop_na(PERP_SEX)
sum(is.na(data_perp_sex$PERP_SEX))
```

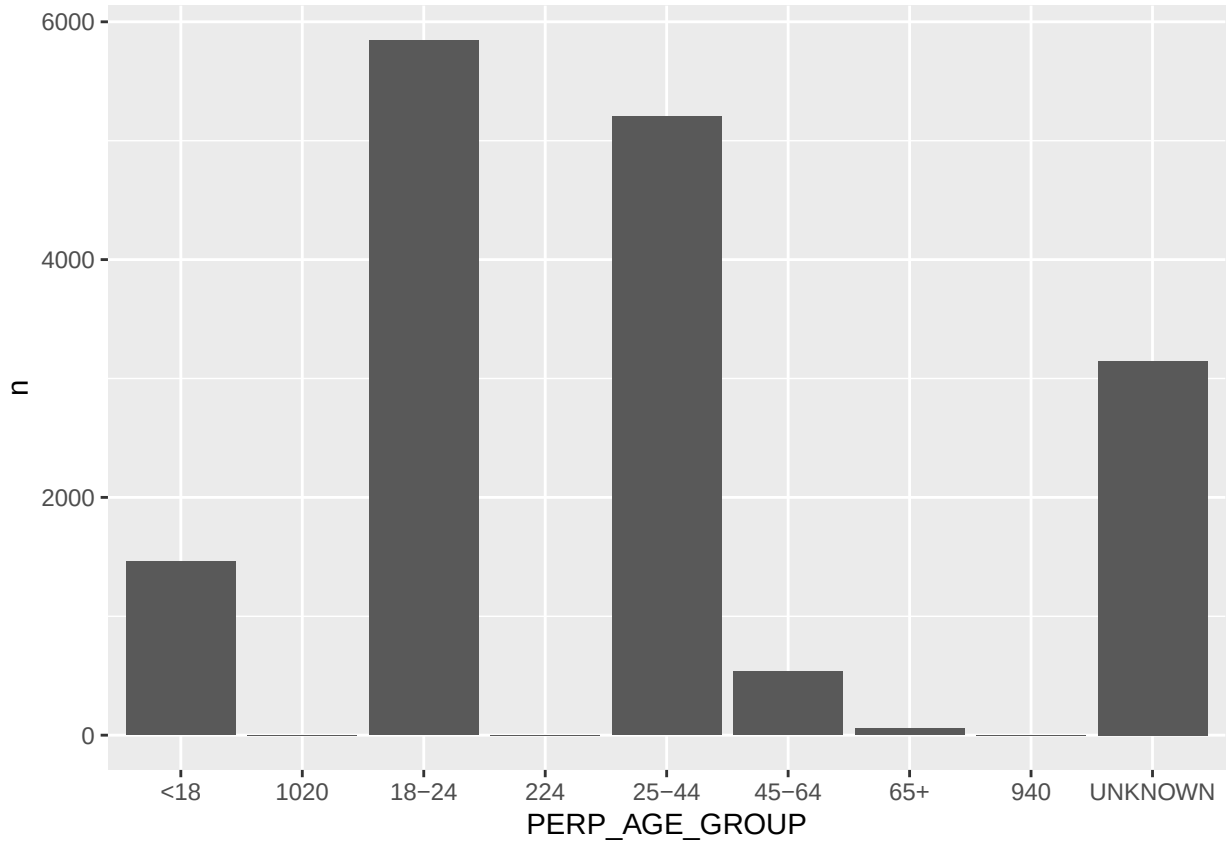
```
## [1] 0
```

```
data_perp_race <- data %>% drop_na(PERP_RACE)
sum(is.na(data_perp_race$PERP_RACE))
```

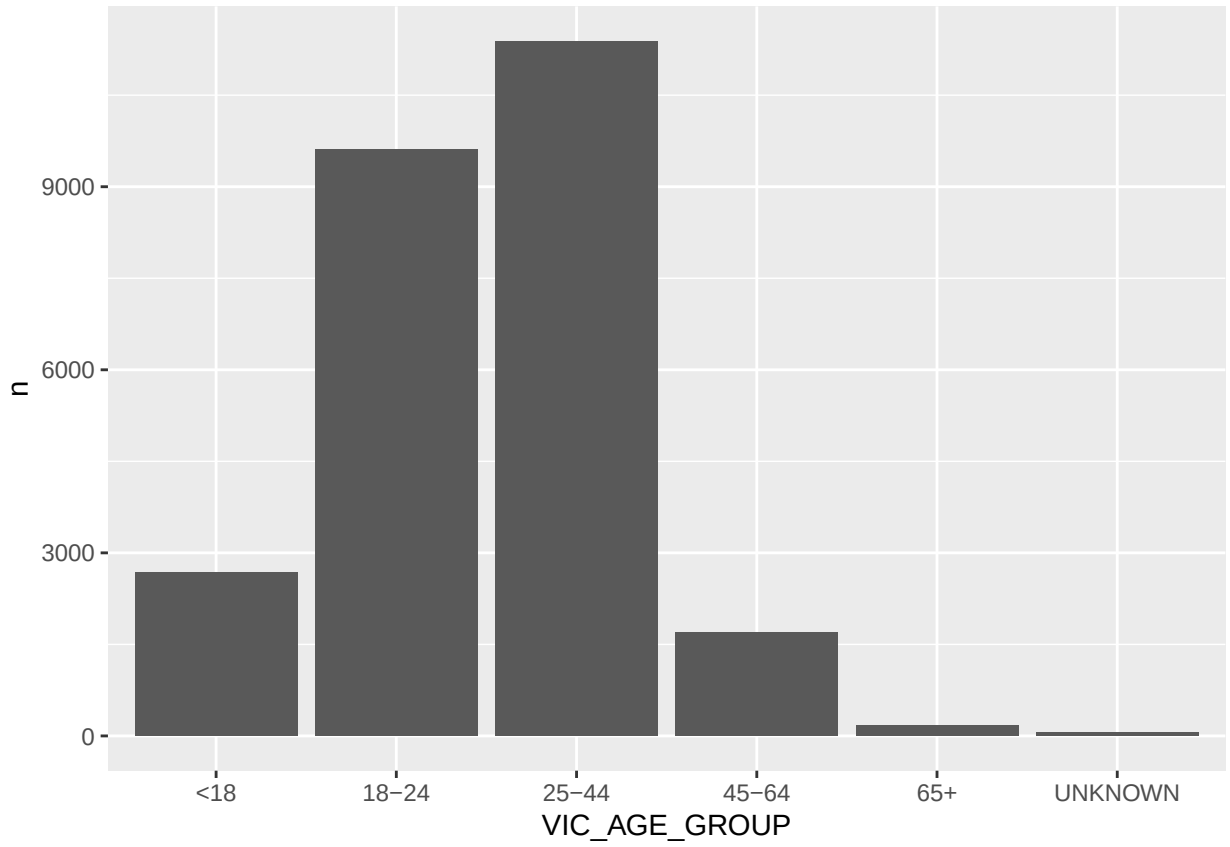
```
## [1] 0
```

Visualize data - Perpetrator and Vistim age groups

```
perp_age_group <- data_perp_age_group %>% count(PERP_AGE_GROUP)
victim_age_group <- data %>% count(VIC_AGE_GROUP)
perp_age_group %>% ggplot(aes(x=PERP_AGE_GROUP, y=n)) + geom_bar(stat="identity")
```



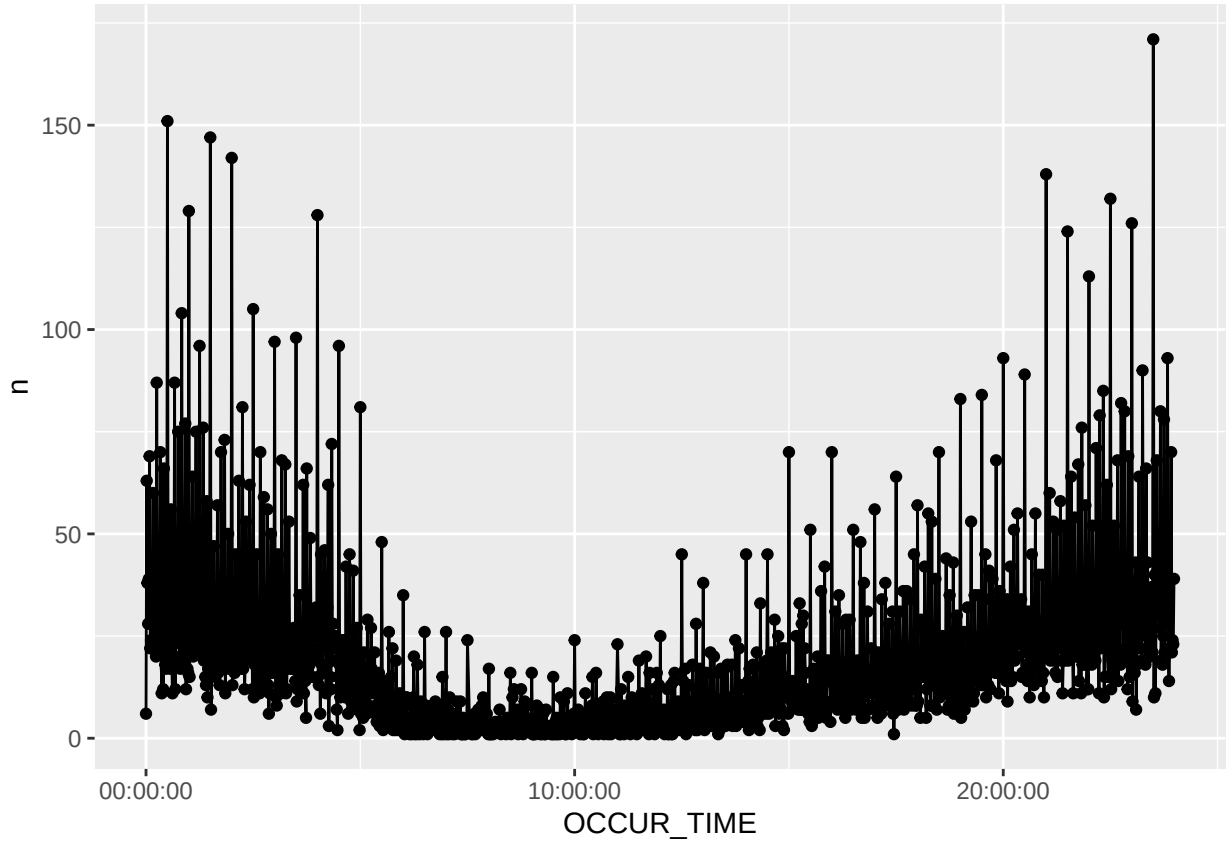
```
victim_age_group %>% ggplot(aes(x=VIC_AGE_GROUP, y=n)) + geom_bar(stat="identity")
```



From this plot it's clear that two age groups (18-24 and 25-44) have majority of cases. Also we have some weird values in PERP_AGE_GROUP column: 1020, 224 and 940 (this could be an error)

Visualize data - Occurrence Time

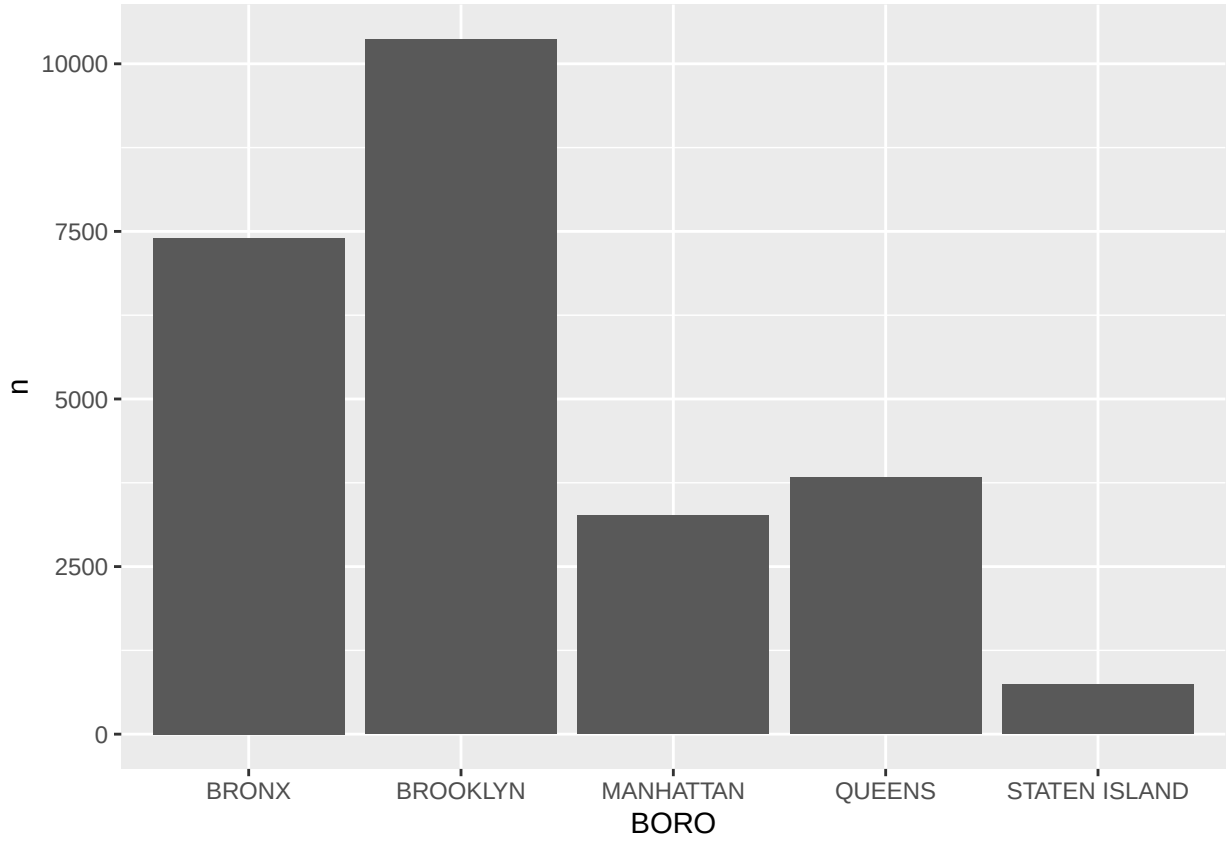
```
occur_time <- data %>% count(OCCUR_TIME)
occur_time %>% ggplot(aes(x=OCCUR_TIME, y=n)) + geom_line() + geom_point()
```



from this plot it's clear that majority of cases happen in night time

Visualize data - Borough Cases

```
borough_cases <- data %>% count(BORO)
borough_cases %>% ggplot(aes(x=BORO, y=n)) + geom_bar(stat="identity")
```



From this plot it's clear that majority of cases happen in Brooklyn

Model data

```
model <- lm(n ~ BORO, data = borough_cases)
summary(model)
```

```
##
## Call:
## lm(formula = n ~ BORO, data = borough_cases)
##
## Residuals:
## ALL 5 residuals are 0: no residual degrees of freedom!
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)           7402        NaN     NaN      NaN
## BOROBROOKLYN          2963        NaN     NaN      NaN
## BOROMANHATTAN        -4137        NaN     NaN      NaN
## BOROQUEENS           -3574        NaN     NaN      NaN
## BOROSTATEN ISLAND    -6666        NaN     NaN      NaN
##
## Residual standard error: NaN on 0 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:    NaN
```

```
## F-statistic:   NaN on 4 and 0 DF,  p-value: NA
```

```
borough_cases %>% mutate(pred = predict(model))
```

```
## # A tibble: 5 x 3
##   BORO              n   pred
##   <chr>         <int>  <dbl>
## 1 BRONX          7402  7402
## 2 BROOKLYN      10365 10365
## 3 MANHATTAN      3265  3265
## 4 QUEENS         3828  3828
## 5 STATEN ISLAND   736   736.
```

Conclusion

1. people in 18-24 and 25-44 have majority of cases
2. majority of cases happen in night time
3. majority of cases happen in Brooklyn

As any human I have lots of biases including perpetrator sex and race for example. I think there are two possible solutions to mitigate that: 1. Analyze data in all possible combinations with all possible types of visualizations. That's ideal solution but quite often it's just not feasible 2. Work only with unpersonalized data, for example instead of Categorical race and sex columns it's possible to use some numbers instead (for example 0 - Male and 1 - Female)