NY_Shootings

CarlosLorena

2022-10-01

Defining the setup

Loading Libraries

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readr)
library(lubridate)
```

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

```
library(dplyr)
```

```
# for fitting the gradient boosting model
#install.packages('gbm')
# for general data preparation and model fitting
#install.packages('caret')

library(gbm)
```

```
## Loaded gbm 2.1.8.1
```

```
library(caret)
```

```
## Carregando pacotes exigidos: lattice

##
## Attaching package: ’caret’

## The following object is masked from ’package:purrr’:
##
##     lift
```

Defining URL and uploading database

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

```
## Rows: 25596 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.
```

Step 0:

After a first look at the database it is possible to see some aspects, such as some missing data, specially regarding perpetrator information (sex, race, age). It is expected, since not all shootings incidents are solved, but may lead to some bias if not carefully analyzed before used.

Cleaning, formating data

Step 1:

To be allowed to proceed with the data analysis, first let’s take care of properly cleaning and formatting the data that will be used.

```
#fixing date
NYPD_shooting_incidents <- NYPD_shooting_incidents %>% mutate(OCCUR_DATE=mdy(OCCUR_DATE))
#fixing time
NYPD_shooting_incidents <- NYPD_shooting_incidents %>% mutate(OCCUR_TIME=hms(OCCUR_TIME))
```

Analysis of shootings by year and by borough

Step 2:

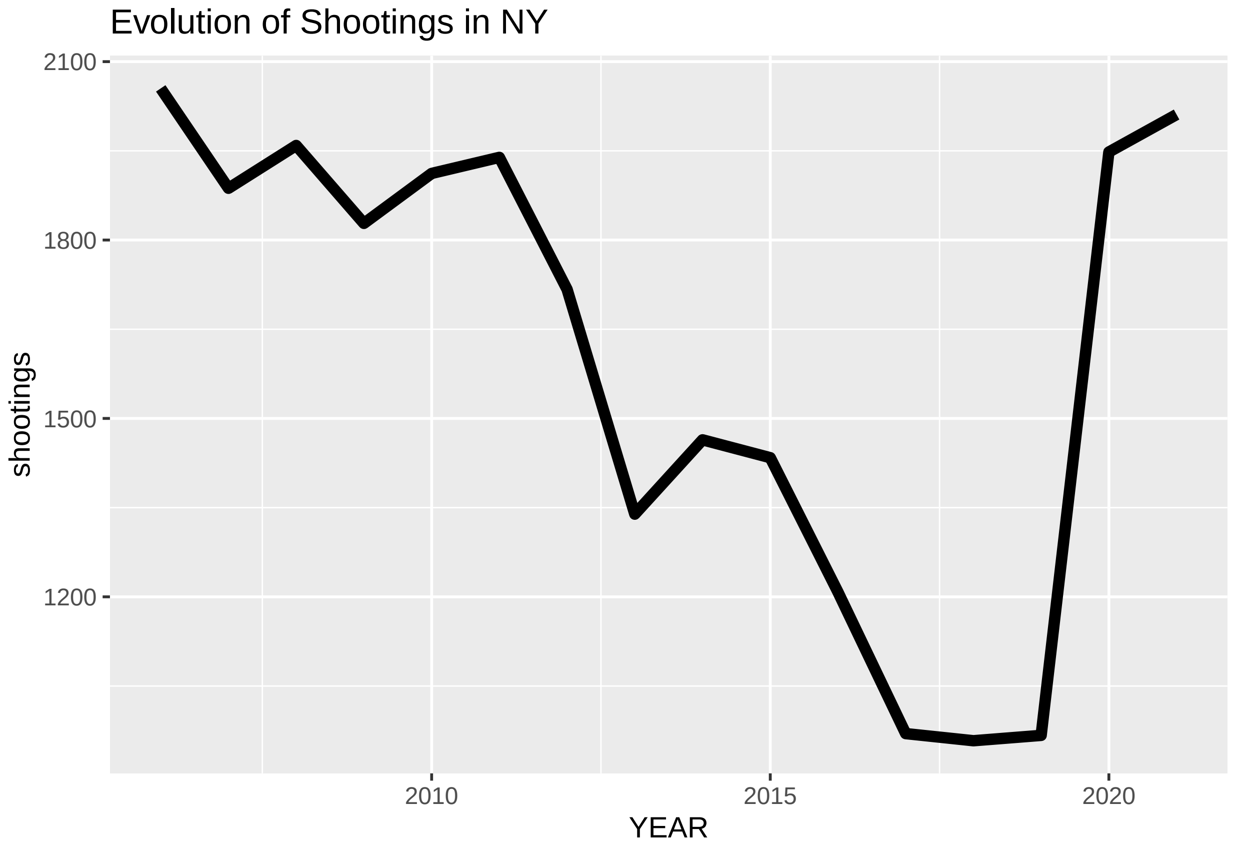
An initial exploration of the data got my attention in two points. The first one is the evolution of the shootings in time and the second is the difference between the neighborhoods. Our focus will be in the shootings incidents assuming that those that murders have a strong correlation with shootings (check the final section: model and correlation). A posterior deeper analysis can try to disentangle the data in order to evaluate if there are significant difference between the correlation shootings x murders for different inputs (area, period, etc).

```
#Create YEAR Column
NYPD_shooting_incidents$YEAR <- year(NYPD_shooting_incidents$OCCUR_DATE)
#Group incidents by year and by borough
NYPD_shooting_incidents_by_year <- NYPD_shooting_incidents %>% group_by(YEAR) %>%
  summarise(shootings = n())
NYPD_shooting_incidents_by_boro <- NYPD_shooting_incidents %>% group_by(BORO) %>%
  summarise(shootings = n())
#Separate incidents by borough by year
NYPD_shooting_incidents_by_boro_and_year <-
  NYPD_shooting_incidents %>% group_by(BORO,YEAR) %>% summarise(shootings = n())
```

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

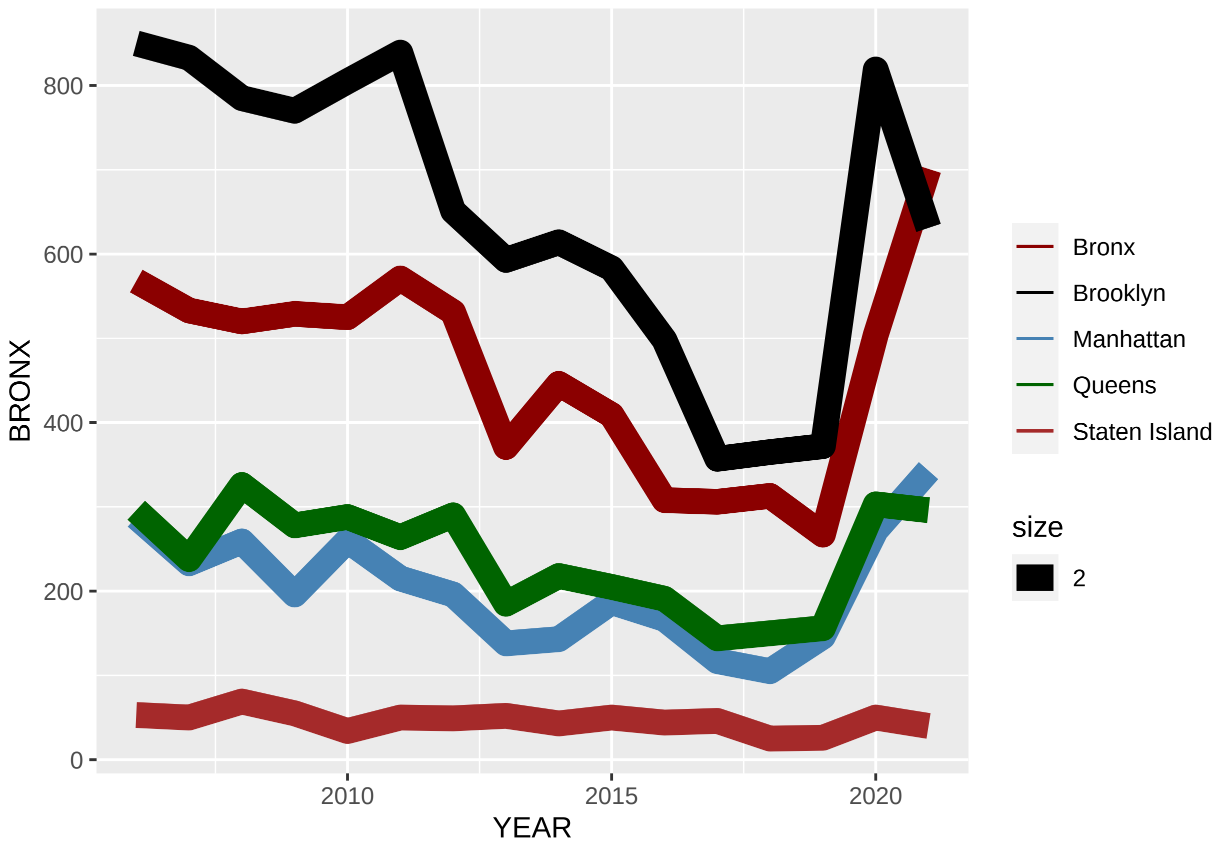
```
NYPD_shooting_incidents_by_boro_and_year_wide <-
  NYPD_shooting_incidents_by_boro_and_year %>%
  pivot_wider(names_from = BORO, values_from = shootings)
NYPD_shooting_incidents_by_boro_and_year_wide <-
  NYPD_shooting_incidents_by_boro_and_year_wide %>%
  rename(STATEN_ISLAND = "STATEN ISLAND")
```

```
#Plot incidents by year
NYPD_shooting_incidents_by_year %>% ggplot(aes(x = YEAR, y = shootings))+
  geom_line(size = 2)+
  ggtitle("Evolution of Shootings in NY")
```



It is possible to notice that the shootings had consistently dropped from 2006 to 2017 with an accentuate drop starting in 2012, a stable low plateau during the years 2017, 2018, and 2019, and a sudden rise in the last two years, that may be related to Covid-19 (the new high coincides with the Covid-19 period). Further development following this line may be interesting. Trying to find what policies were adopted that resulted in the drop of the incidents from 2012 and if the recent rise are really realated to the Covid-19 or there is another unseen potential cause.

```
#plot occurrences by year by borough
NYPD_shooting_incidents_by_boro_and_year_wide %>% ggplot(aes(x=YEAR))+
  geom_line(aes(y = BRONX, color = "Bronx", size=2))+
  geom_line(aes(y = BROOKLYN, color="Brooklyn", size=2))+
  geom_line(aes(y = MANHATTAN, color="Manhattan", size=2))+
  geom_line(aes(y = QUEENS, color="Queens", size=2))+
  geom_line(aes(y = STATEN_ISLAND, color="Staten Island", size=2))+
  scale_colour_manual("",
                      breaks = c("Bronx", "Brooklyn", "Manhattan", "Queens", "Staten Island"),
                      values = c("darkred", "black", "steelblue", "darkgreen", "brown"))
```



We can notice that there are some boroughs that contributes more to the total than others, however it is necessary to take into account the total population of each region.

```
#order the dataframe by borough
NYPD_shooting_incidents_by_boro <-
  NYPD_shooting_incidents_by_boro[order(NYPD_shooting_incidents_by_boro$BORO),]

#add population of each borough
NYPD_shooting_incidents_by_boro$population <-
  c(1424948,2641052,1576876,2331143,493494)

#calculate incidents by 100k people by borough (total period)
NYPD_shooting_incidents_by_boro$Shootings_per_100k <-
  NYPD_shooting_incidents_by_boro$shootings /
  (NYPD_shooting_incidents_by_boro$population/100000)

NYPD_shooting_incidents_by_boro
```

```
## # A tibble: 5 x 4
##   BORO          shootings population Shootings_per_100k
##   <chr>             <int>      <dbl>              <dbl>
## 1 BRONX              7402    1424948               519.
## 2 BROOKLYN          10365    2641052               392.
## 3 MANHATTAN          3265    1576876               207.
## 4 QUEENS             3828    2331143               164.
## 5 STATEN ISLAND       736     493494               149.
```

Source(*):

https://censusreporter.org/profiles/06000US3600508510-bronx-borough-bronx-county-ny/
https://censusreporter.org/profiles/06000US3604710022-brooklyn-borough-kings-county-ny/
https://censusreporter.org/profiles/06000US3606144919-manhattan-borough-new-york-county-ny/
https://censusreporter.org/profiles/06000US3608160323-queens-borough-queens-county-ny/
https://censusreporter.org/profiles/06000US3608570915-staten-island-borough-richmond-county-ny/

and compared to (check if the data above seems ok):

https://en.wikipedia.org/wiki/The_Bronx
https://en.wikipedia.org/wiki/Brooklyn
https://en.wikipedia.org/wiki/Manhattan
https://en.wikipedia.org/wiki/Queens
https://en.wikipedia.org/wiki/Staten_Island

(*) Note: this is an exercise, in real life it would be necessary to validate these data (I'm not from NY nor USA, and I'm not pretty sure that the names of the neighboroughs corresponds exactly on both sources).

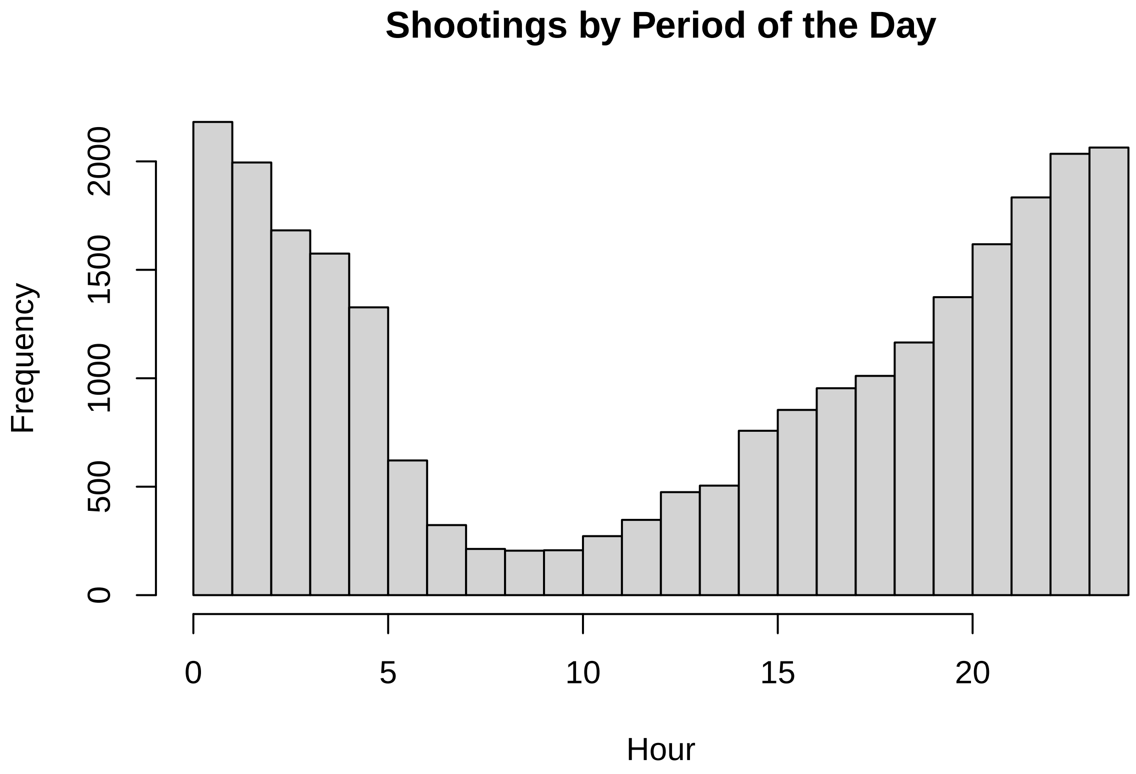
Considering the data about population we can see that there is a difference between the incidents per capita of each neighborhood.

Analysis of shootings by hour of the day

Step 3:

Another analysis that may help to have a general picture is breaking and grouping the data by different aspects. It also may help in future modeling, when trying to predict shooting cases.

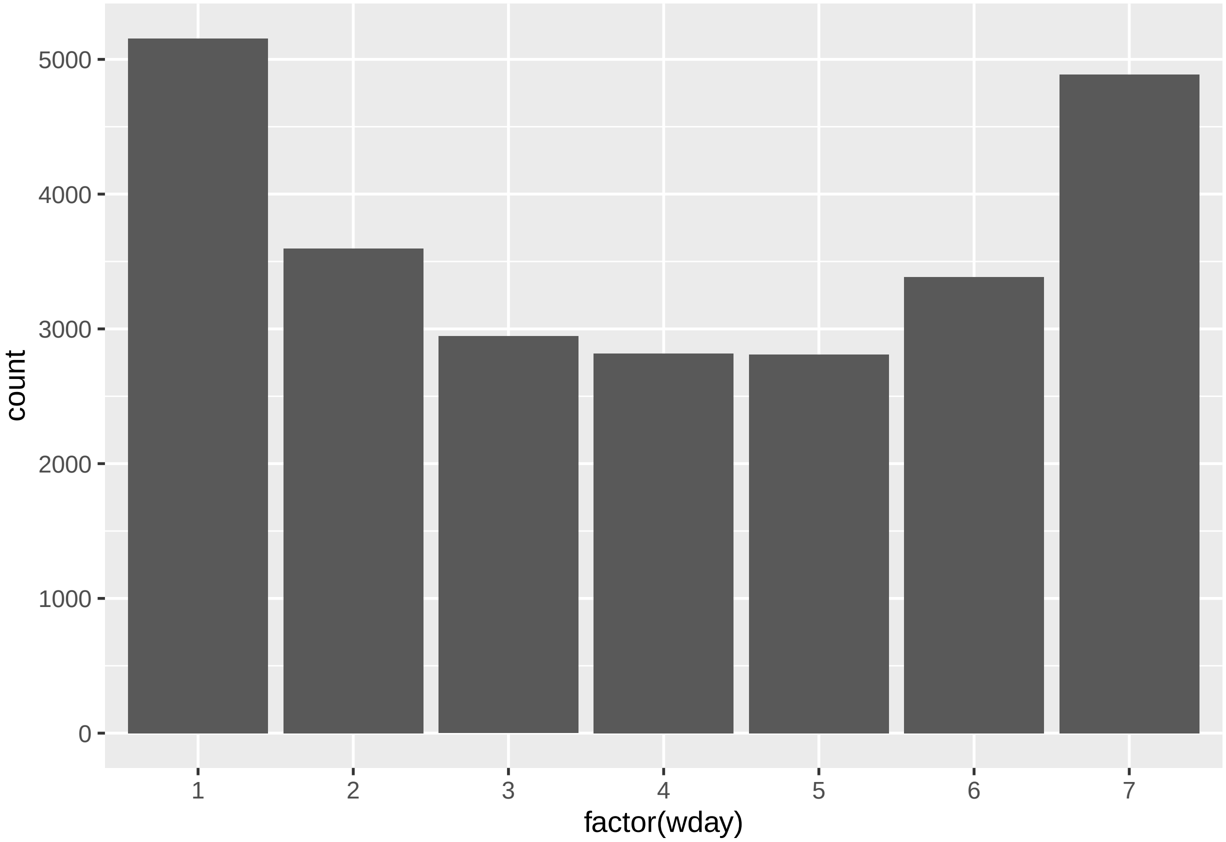
```
#Histogram of shootings by time of the day
hist(seconds(NYPD_shooting_incidents$OCCUR_TIME)/
       3600,breaks=24,main="Shootings by Period of the Day", xlab="Hour")
```



As expected, there are a concentration of shootings during the night and few cases during the daylight. This aspect will be relevant in the future, if we decide to build a model of the distributions of the shootings.

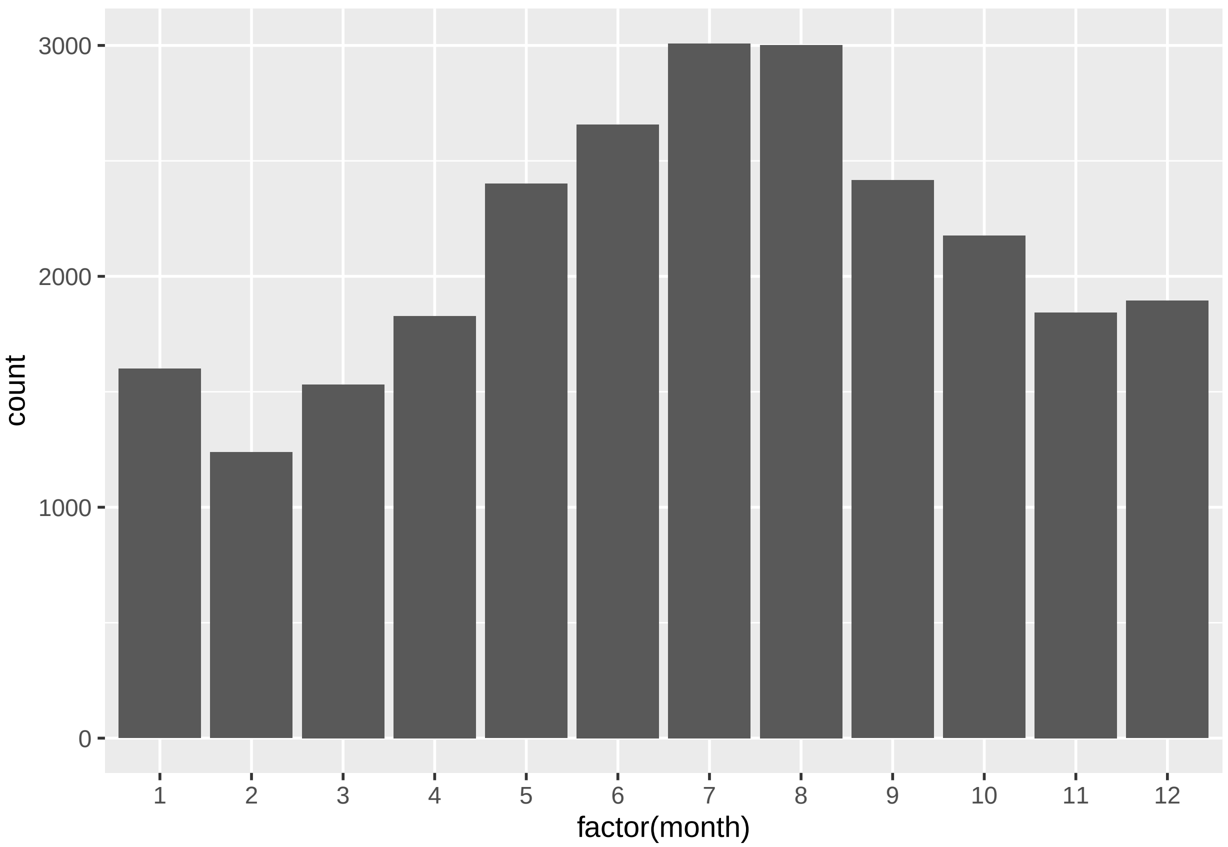
Analysis of shootings by the day of the week and by month

```
#Extract the month
NYPD_shooting_incidents$wday <- NYPD_shooting_incidents$OCCUR_DATE %>% wday()

#Histogram of shootings by day of the week
ggplot(NYPD_shooting_incidents, aes(x = factor(wday)))+
  geom_bar()
```



The incidents start to increase on Friday (day 6) and peaks on Sunday (day 1), droping during the weekdays. This is another fact that may be relevant to the model.

```
#Extract the month of the year
NYPD_shooting_incidents$month <- NYPD_shooting_incidents$OCCUR_DATE %>% month()

#Histogram of shootings by month
ggplot(NYPD_shooting_incidents, aes(x = factor(month))) +
  geom_bar()
```



Apparently the same can be said about the variation through months.

Model

Step 4:

Now exploring the relationship between shootings and deaths, we will model the number of deaths in a month as function of the total shootings in that month.

```
NYPD_shooting_incidents_by_month_year <-
  NYPD_shooting_incidents %>% group_by(YEAR,month) %>%
  summarise(shootings = n(), deaths = sum(STATISTICAL_MURDER_FLAG))
```

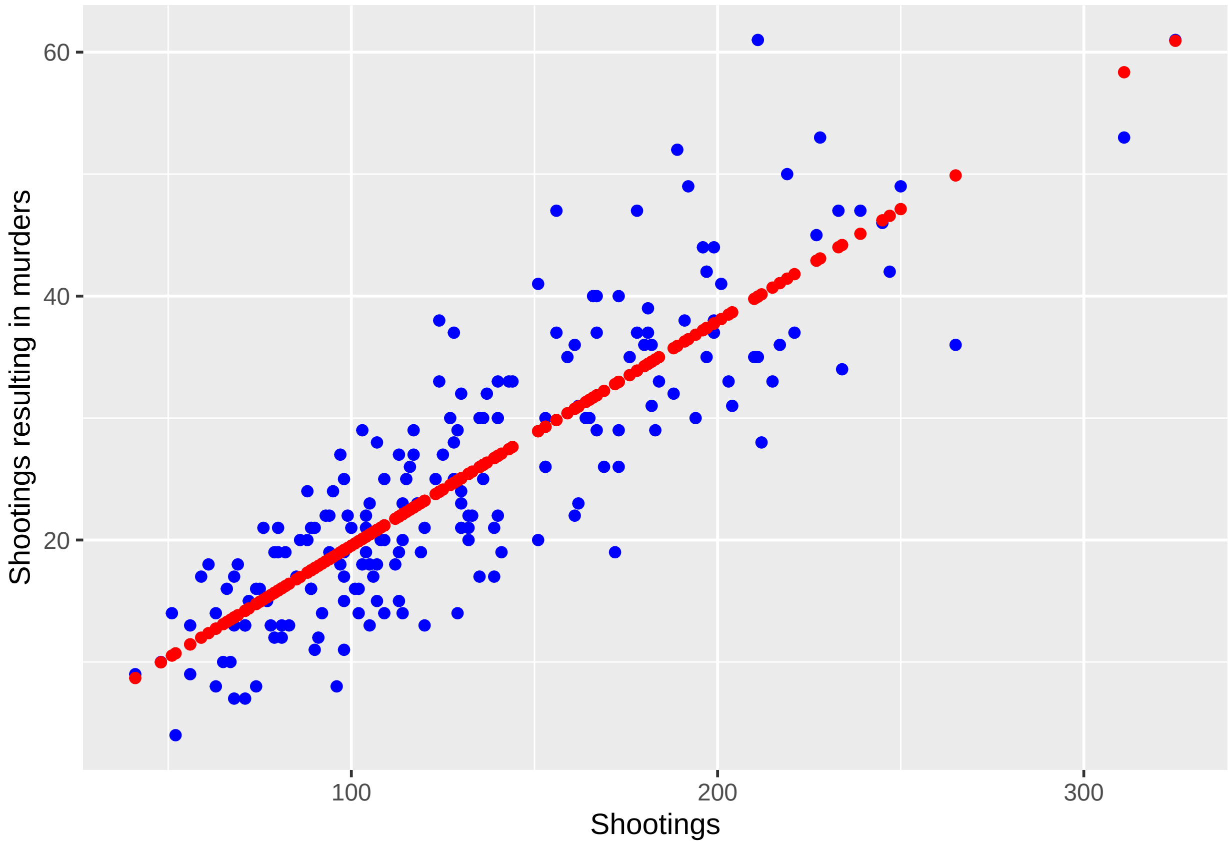
```
## ‘summarise()‘ has grouped output by ’YEAR’. You can override using the
## ‘.groups‘ argument.
```

```
#mod <- lm(deaths ~ shootings, data = NYPD_shooting_incidents_by_month_year)
#summary(mod)

X <- NYPD_shooting_incidents_by_month_year$shootings
Y <- NYPD_shooting_incidents_by_month_year$deaths
mod <- lm(Y ~ X)
summary(mod)
```

```
##
```

```
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.8922  -4.0361   0.0418   3.6113  21.0418
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.142231   1.150882   0.992    0.322
## X           0.183962   0.008021  22.936   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 5.898 on 190 degrees of freedom
## Multiple R-squared:  0.7347, Adjusted R-squared:  0.7333
## F-statistic:   526 on 1 and 190 DF,  p-value: < 2.2e-16
```

```
Y_pred <- predict(mod)

ggplot() +
  geom_point(aes(x = X, y = Y),color = "blue")+
  geom_point(aes(x = X, y = Y_pred),color = "red")+
  xlab("Shootings")+
  ylab("Shootings resulting in murders")
```



```
cor(X,Y)
```

```
## [1] 0.8571197
```

In blue the data X axis = shootings and Y axis = Deaths in a particular month. In red the prediction from a linear model. As per graph, it seems that shootings and shootings that resulted in murders to have a strong correlation, confirmed by the calculation (correlation = 0.857).