

assignment

student

2/24/2022

Read data from web url

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

```
## 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.
```

Tidy data

```
### get rid of
### INCIDENT_KEY, X_COORD_CD, X_COORD_CD, Latitude, Longitude, Lon_Lat
nypd_data %>% select(-c(INCIDENT_KEY, X_COORD_CD, X_COORD_CD, Latitude, Longitude, Lon_Lat))
```

```
## # A tibble: 25,596 x 14
##   OCCUR_DATE OCCUR_TIME BORO      PRECINCT JURISDICTION_CODE LOCATION_DESC
##   <chr>      <time>    <chr>      <dbl>      <dbl> <chr>
## 1 11/11/2021 15:04    BROOKLYN    79          0 <NA>
## 2 07/16/2021 22:05    BROOKLYN    72          0 <NA>
## 3 07/11/2021 01:09    BROOKLYN    79          0 <NA>
## 4 12/11/2021 13:42    BROOKLYN    81          0 <NA>
## 5 02/16/2021 20:00    QUEENS      113         0 <NA>
## 6 05/15/2021 04:13    QUEENS      113         0 <NA>
## 7 04/14/2021 21:08    BRONX       42          0 COMMERCIAL BLDG
## 8 12/10/2021 19:30    BRONX       52          0 <NA>
## 9 02/22/2021 00:18    MANHATTAN   34          0 <NA>
## 10 03/07/2021 06:15    BROOKLYN    75          0 <NA>
## # ... with 25,586 more rows, and 8 more variables:
## #   STATISTICAL_MURDER_FLAG <lgl>, PERP_AGE_GROUP <chr>, PERP_SEX <chr>,
```

```
## # PERP_RACE <chr>, VIC_AGE_GROUP <chr>, VIC_SEX <chr>, VIC_RACE <chr>,
## # Y_COORD_CD <dbl>
```

```
summary(nypd_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
##
```

Analysis

```
### see the number of shooting incident each district
```

```
district_incident <- nypd_data %>% group_by(BORO) %>% summarise(count=n())
```

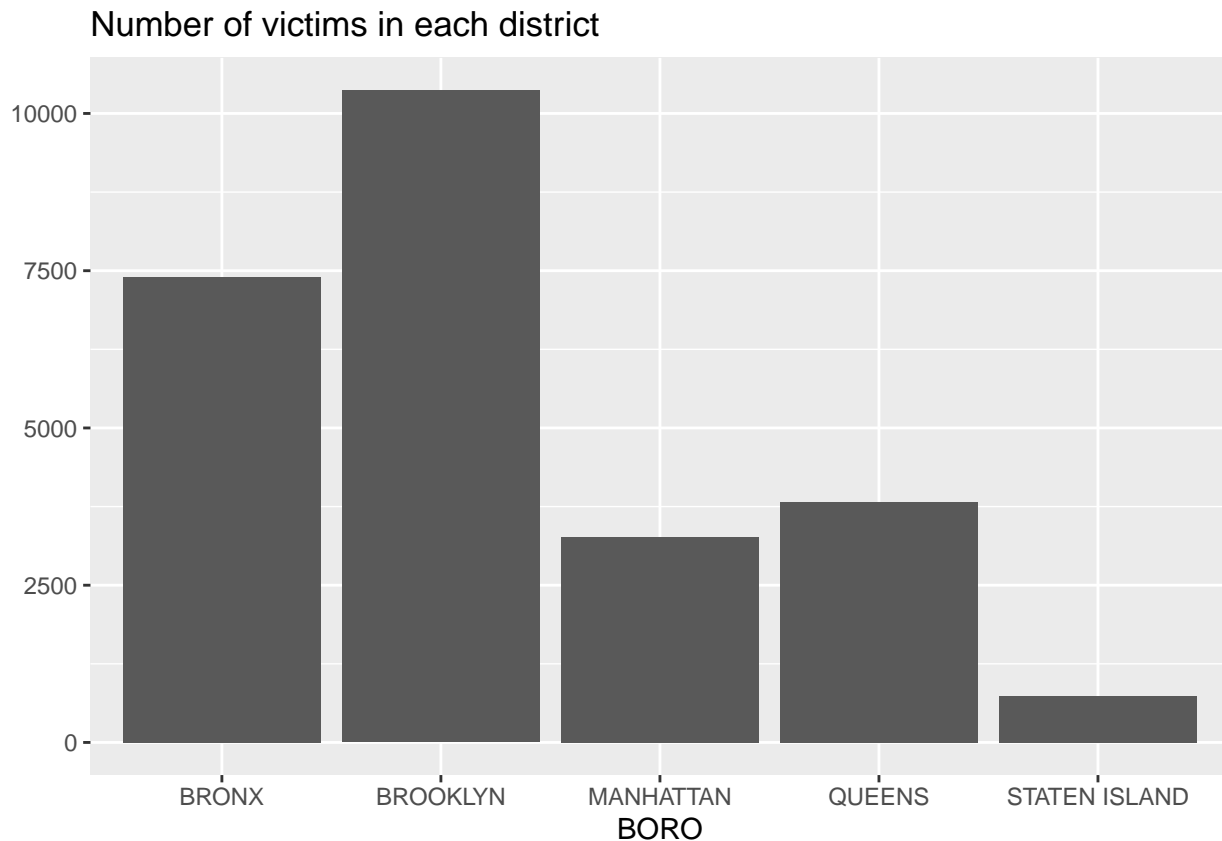
```
### see the max number
max(district_incident$count)
```

```
## [1] 10365
```

```
### plot the data
```

```
### Number of victims in each district, Brooklyn has the most number of victims.
```

```
ggplot(data = district_incident)+geom_bar(mapping = aes(x=BORO,y=count),stat="identity") + labs(title =
```



```
### see how many victims are female?
```

```
vic_female <- nypd_data %>% filter(VIC_SEX=="F") %>% select(c(VIC_RACE,VIC_AGE_GROUP))
```

```
summary(vic_female)
```

```
##      VIC_RACE      VIC_AGE_GROUP
## Length:2403      Length:2403
## Class :character  Class :character
## Mode  :character  Mode  :character
```

```
### see how many victims are male?
```

```
vic_male <- nypd_data %>% filter(VIC_SEX=="M") %>% select(c(VIC_RACE,VIC_AGE_GROUP))
```

```
summary(vic_male)
```

```
##      VIC_RACE      VIC_AGE_GROUP
## Length:23182      Length:23182
## Class :character  Class :character
## Mode  :character  Mode  :character
```

```
### how many victims group by sex
```

```
nypd_data %>% group_by(VIC_SEX) %>% summarise(count=n())
```

```
## # A tibble: 3 x 2
##   VIC_SEX count
##   <chr>   <int>
## 1 F      2403
## 2 M     23182
## 3 U       11
```

```
### the totals of vic_male is 21370, and the totals of vic_female is 2204.
```

```
### how many perps group by sex
```

```
nypd_data %>% group_by(PERP_SEX) %>% summarise(count=n())
```

```
## # A tibble: 4 x 2
##   PERP_SEX count
##   <chr>   <int>
## 1 F       371
## 2 M     14416
## 3 U       1499
## 4 <NA>    9310
```

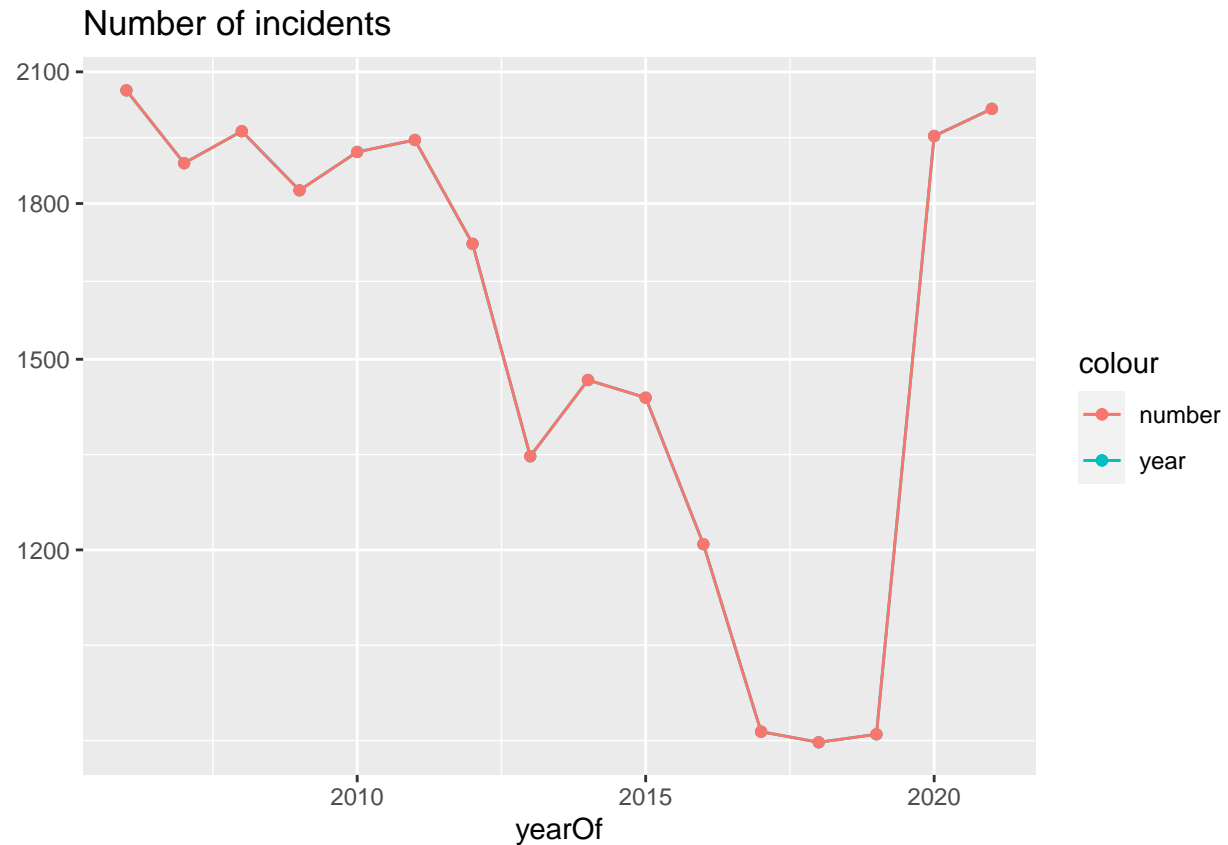
```
### the number of incident of every year
```

```
nypd_byyear = nypd_data %>% mutate(yearOf=year(mdy(OCCUR_DATE))) %>% group_by(yearOf) %>% summarise(number=
summary(nypd_byyear)
```

```
##      yearOf      number
## Min.   :2006   Min.    : 958
## 1st Qu.:2010   1st Qu.:1306
## Median :2014   Median :1772
## Mean   :2014   Mean    :1600
## 3rd Qu.:2017   3rd Qu.:1941
## Max.   :2021   Max.    :2055
```

```
### plot the incident by year
```

```
nypd_byyear %>% ggplot(aes(x=yearOf,y=number))+geom_line(aes(color="year")) + geom_point(aes(color="year"))
  scale_y_log10() +
  labs(title = "Number of incidents",y=NULL)
```



Mode

```
mod <- lm(yearOf~number,data=nypd_byyear)
summary(mod)
```

```
##
## Call:
## lm(formula = yearOf ~ number, data = nypd_byyear)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.7756 -2.3315 -0.3892  0.5406  9.9688
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.023e+03  4.445e+00  455.127  <2e-16 ***
## number       -6.003e-03  2.699e-03  -2.225   0.0431 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.236 on 14 degrees of freedom
## Multiple R-squared:  0.2612, Adjusted R-squared:  0.2084
## F-statistic: 4.949 on 1 and 14 DF,  p-value: 0.04307
```

Bias

The dataset has many variables, and in the report I only used some of them. Didn't use variables like PERP_RACE, VIC_RACE, PERCINT etc. The Analysis and mode is simple. Some of variables are NA values. I thinks this is also bias in dataset.