Project1_NYPD

CAPSA

17/7/2022

```
library(tidyverse)
```

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

## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

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

```
library(tidyr)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

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

```
url_in <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?"
#updating name of original dataset file to "NYPD_Shooting_Incidents.csv"
file_names <- c("NYPD_Shooting_Incidents.csv")
#just following the class in here with the covid example, I'm placing the link in a single vector
urls <- str_c(url_in,file_names)
```

```
library(tidyverse)
library(tidyr)
library(lubridate)
#importing original dataset
NYPD_dataset_original <- read_csv(urls[1])
```

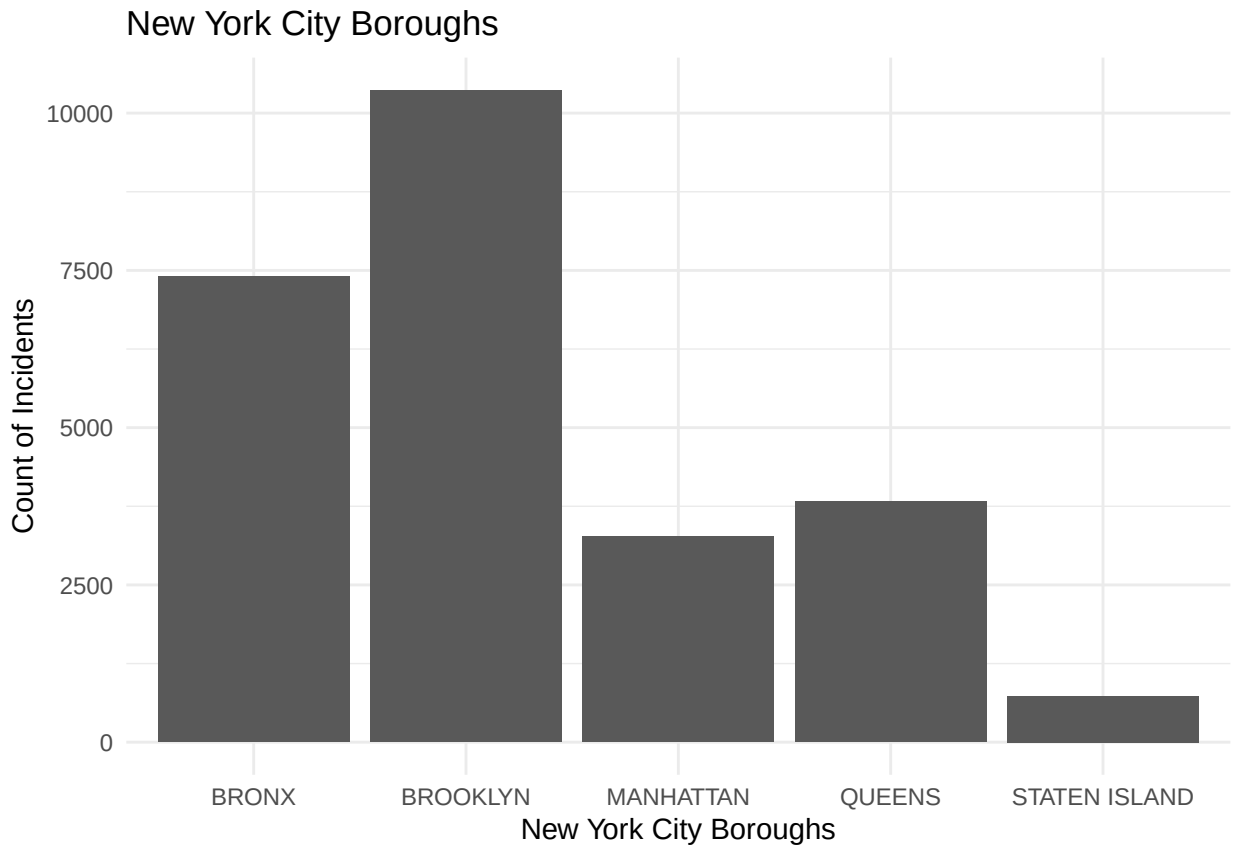
```
library(tidyverse)
library(tidyr)
library(lubridate)
```

```r
#deleting columns related with jurisdiction code & coordinates
NYPD_dataset_withoutColumns <-subset(NYPD_dataset_original,select=-c(JURISDICTION_CODE,X_COORD_CD, Y_COO


#changing format of columns as follows: "OCCUR_DATE" from <chr> to <date> ; from <char> to <factor> col
NYPD_dataset <-NYPD_dataset_withoutColumns%>%mutate(OCCUR_DATE =mdy(OCCUR_DATE))%>%
  mutate(BORO =as.factor(BORO))%>%
  mutate(PERP_AGE_GROUP =as.factor(PERP_AGE_GROUP))%>%
  mutate(PERP_SEX =as.factor(PERP_SEX))%>%
  mutate(PERP_RACE =as.factor(PERP_RACE))%>%
  mutate(VIC_AGE_GROUP =as.factor(VIC_AGE_GROUP))%>%
  mutate(VIC_SEX =as.factor(VIC_SEX))%>%
  mutate(VIC_RACE =as.factor(VIC_RACE))%>%
  #Transforming time format to hours to generate third plot below
  mutate(OCCUR_TIME =hour(hms(as.character(OCCUR_TIME))))


#remove from filter values "224", "940" & "1020" as do not correspond to an age value
NYPD_dataset <-NYPD_dataset%>%filter(PERP_AGE_GROUP%in%c("<18","18-24","25-44","45-64","65+","UNKNOWN",N

#showing summary of data after deleting columns not needed & converting data type as required
summary(NYPD_dataset)
```

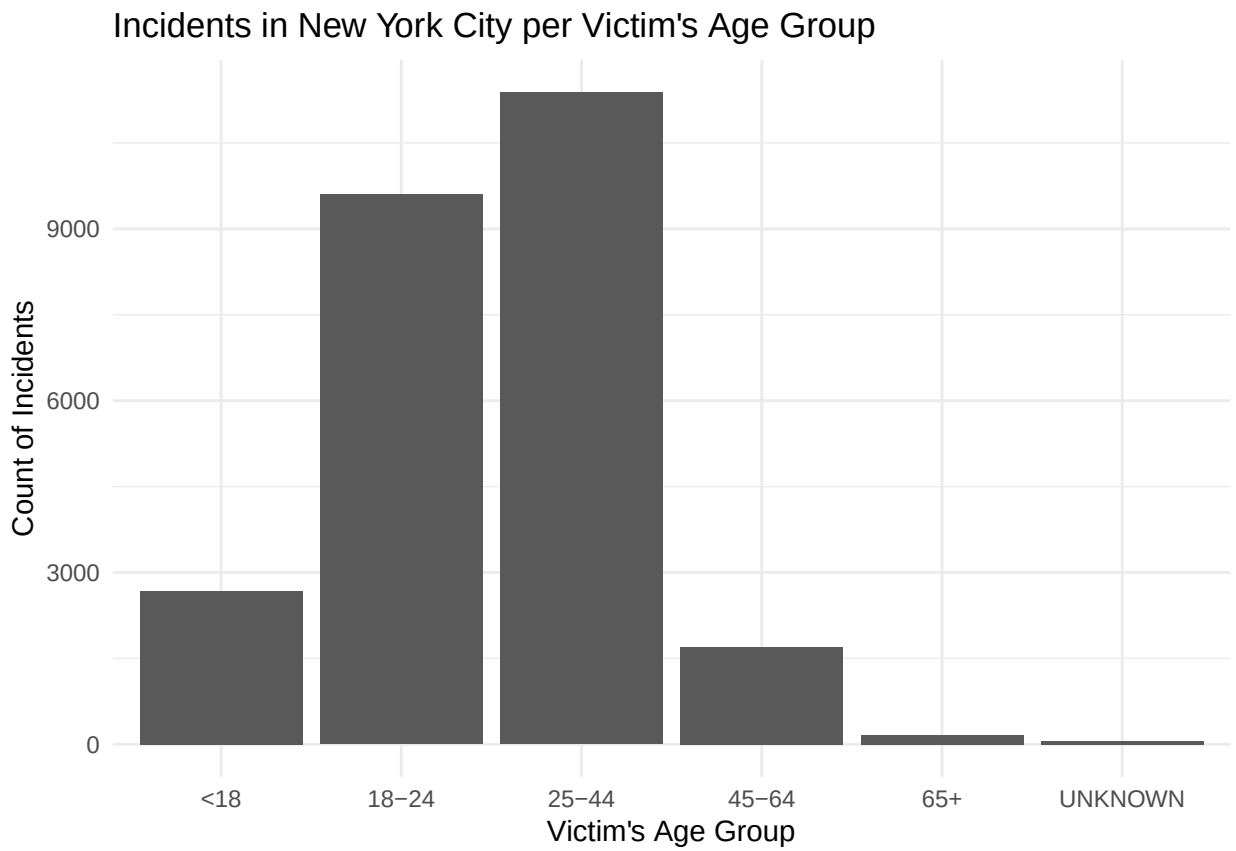
```
##   INCIDENT_KEY         OCCUR_DATE           OCCUR_TIME                BORO
##  Min.   :  9953245   Min.   :2006-01-01   Min.   : 0.00   BRONX        : 7400
##  1st Qu.: 61593632   1st Qu.:2009-05-10   1st Qu.: 3.00   BROOKLYN     :10364
##  Median : 86437258   Median :2012-08-26   Median :15.00   MANHATTAN    : 3265
##  Mean   :112383964   Mean   :2013-06-13   Mean   :12.19   QUEENS       : 3828
##  3rd Qu.:166660833   3rd Qu.:2017-07-01   3rd Qu.:20.00   STATEN ISLAND:  736
##  Max.   :238490103   Max.   :2021-12-31   Max.   :23.00
##
##     PRECINCT      LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Min.   :  1.00   Length:25593       Mode :logical           18-24  :5844
##  1st Qu.: 44.00   Class :character   FALSE:20665             25-44  :5202
##  Median : 69.00   Mode  :character   TRUE :4928              UNKNOWN:3148
##  Mean   : 65.87                                            <18    :1463
##  3rd Qu.: 81.00                                            45-64  : 535
##  Max.   :123.00                                            (Other):  57
##                                                            NA's   :9344
##  PERP_SEX              PERP_RACE     VIC_AGE_GROUP   VIC_SEX
##  F   :  371   BLACK          :10667   <18    : 2681   F: 2403
##  M   :14413   WHITE HISPANIC: 2162   18-24  : 9603   M:23179
##  U   : 1499   UNKNOWN        : 1836   25-44  :11384   U:   11
##  NA's: 9310   BLACK HISPANIC: 1203   45-64  : 1698
##               WHITE          :  272   65+    :  167
##               (Other)        :  143   UNKNOWN:   60
##               NA's           : 9310
##                           VIC_RACE
##  AMERICAN INDIAN/ALASKAN NATIVE:    9
##  ASIAN / PACIFIC ISLANDER      :  354
##  BLACK                         :18280
##  BLACK HISPANIC                : 2485
##  UNKNOWN                       :   65
```

```
##  WHITE                          :  660
##  WHITE HISPANIC                 : 3740
```

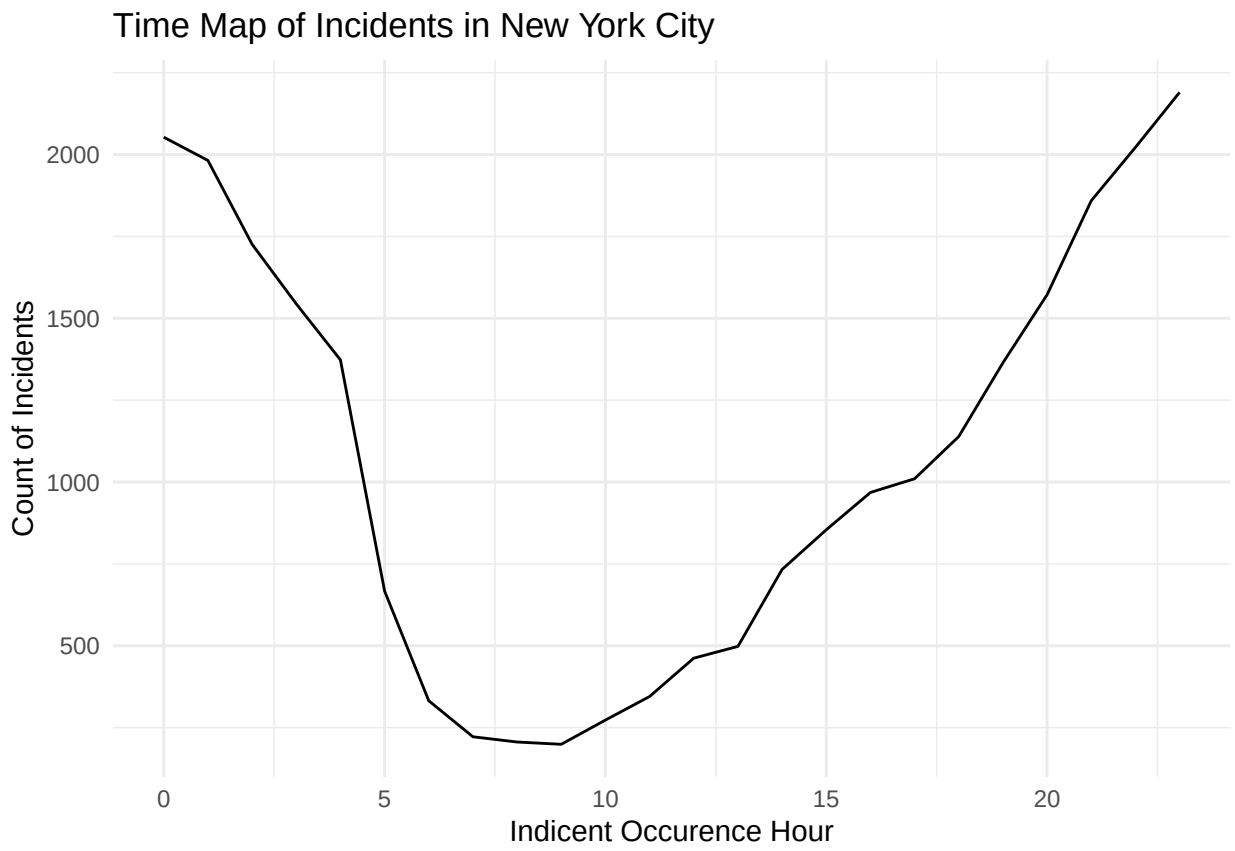
```
#The first plot I'm going to do is to understand which borough of New York has the most number of incid


plot_first <-ggplot(NYPD_dataset,aes(x=BORO))+
  geom_bar() +
  labs(title= "New York City Boroughs",
       x= "New York City Boroughs",
       y= "Count of Incidents") +
  theme_minimal()

plot_first
```




```
#The second plot I'm going to do is to have an idea of the incidents per victim's age group. Its result

plot_second <-ggplot(NYPD_dataset,aes(x=VIC_AGE_GROUP))+
  geom_bar() +
  labs(title= "Incidents in New York City per Victim's Age Group",
       x= "Victim's Age Group",
       y= "Count of Incidents") +
  theme_minimal()

plot_second
```




```
#I'll generate a third plot with only hour data to understand what time is more risky for citizens in N

NYPD_dataset_hour<-NYPD_dataset%>%
  group_by(OCCUR_TIME)%>%
  count()


plot_third <-ggplot(NYPD_dataset_hour,aes(x=OCCUR_TIME,y=n))+
  geom_line()+
  labs(title= "Time Map of Incidents in New York City",
       x= "Indicent Occurence Hour",
       y= "Count of Incidents") +
  theme_minimal()

plot_third
```




```
#Generate tables with Perp's vs Victim's race, sex and group age in order to perform some analysis


table_race <-table(NYPD_dataset$PERP_RACE,NYPD_dataset$VIC_RACE)
table_sex <- table(NYPD_dataset$PERP_SEX,NYPD_dataset$VIC_SEX)
table_age<-table(NYPD_dataset$PERP_AGE_GROUP,NYPD_dataset$VIC_AGE_GROUP)

table_race
```

```
##
##                                AMERICAN INDIAN/ALASKAN NATIVE
##   AMERICAN INDIAN/ALASKAN NATIVE                              0
##   ASIAN / PACIFIC ISLANDER                                    0
##   BLACK                                                       4
##   BLACK HISPANIC                                              0
##   UNKNOWN                                                     3
##   WHITE                                                       0
##   WHITE HISPANIC                                              0
##
##                                ASIAN / PACIFIC ISLANDER BLACK BLACK HISPANIC
##   AMERICAN INDIAN/ALASKAN NATIVE                        0     2              0
##   ASIAN / PACIFIC ISLANDER                             43    51             13
##   BLACK                                               135  8470            749
##   BLACK HISPANIC                                       17   481            320
##   UNKNOWN                                              16  1359            155
##   WHITE                                                11    34             21
```

```
##   WHITE HISPANIC                                    35   719           383
##
##                                  UNKNOWN WHITE WHITE HISPANIC
##   AMERICAN INDIAN/ALASKAN NATIVE       0     0              0
##   ASIAN / PACIFIC ISLANDER             0    11             23
##   BLACK                               24   183           1102
##   BLACK HISPANIC                       5    34            346
##   UNKNOWN                              6    42            255
##   WHITE                                1   156             49
##   WHITE HISPANIC                      11    89            925
```

```
table_sex
```

```
##
##         F     M     U
##   F    58   312     1
##   M  1540 12867     6
##   U   112  1386     1
```

```
table_age
```

```
##
##           <18 18-24 25-44 45-64  65+ UNKNOWN
##   <18      445   584   353    70    9       2
##   1020       0     0     0     0    0       0
##   18-24    742  2607  2141   305   37      12
##   224        0     0     0     0    0       0
##   25-44    247  1417  3033   431   40      34
##   45-64     19    62   290   148   11       5
##   65+        0     1    23    23   10       0
##   940        0     0     0     0    0       0
##   UNKNOWN  416  1364  1202   148   16       2
```

```
glm.fit<-glm(NYPD_dataset$STATISTICAL_MURDER_FLAG ~ NYPD_dataset$PERP_RACE  + NYPD_dataset$PERP_SEX + N
summary(glm.fit)
```

```
##
## Call:
## glm(formula = NYPD_dataset$STATISTICAL_MURDER_FLAG ~ NYPD_dataset$PERP_RACE +
##     NYPD_dataset$PERP_SEX + NYPD_dataset$PERP_AGE_GROUP + NYPD_dataset$OCCUR_TIME)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.51490  -0.24329  -0.20352  -0.01873   1.03488
##
## Coefficients:
##                                              Estimate Std. Error t value
## (Intercept)                                -0.0367305  0.2748130  -0.134
## NYPD_dataset$PERP_RACEASIAN / PACIFIC ISLANDER  0.3402816  0.2764666   1.231
## NYPD_dataset$PERP_RACEBLACK                  0.2530057  0.2745628   0.921
## NYPD_dataset$PERP_RACEBLACK HISPANIC         0.2302582  0.2747667   0.838
## NYPD_dataset$PERP_RACEUNKNOWN                0.2006449  0.2753399   0.729
```

```
## NYPD_dataset$PERP_RACEWHITE                      0.3841368  0.2755588   1.394
## NYPD_dataset$PERP_RACEWHITE HISPANIC             0.2751629  0.2746533   1.002
## NYPD_dataset$PERP_SEXM                          -0.0355676  0.0204468  -1.740
## NYPD_dataset$PERP_SEXU                           0.0551599  0.0302501   1.823
## NYPD_dataset$PERP_AGE_GROUP18-24                 0.0269071  0.0113564   2.369
## NYPD_dataset$PERP_AGE_GROUP25-44                 0.0865744  0.0115181   7.516
## NYPD_dataset$PERP_AGE_GROUP45-64                 0.1564839  0.0197409   7.927
## NYPD_dataset$PERP_AGE_GROUP65+                   0.2030565  0.0528654   3.841
## NYPD_dataset$PERP_AGE_GROUPUNKNOWN              -0.1591315  0.0140682 -11.311
## NYPD_dataset$OCCUR_TIME                         -0.0001781  0.0003674  -0.485
##                                                 Pr(>|t|)
## (Intercept)                                     0.893676
## NYPD_dataset$PERP_RACEASIAN / PACIFIC ISLANDER 0.218407
## NYPD_dataset$PERP_RACEBLACK                     0.356811
## NYPD_dataset$PERP_RACEBLACK HISPANIC            0.402035
## NYPD_dataset$PERP_RACEUNKNOWN                   0.466185
## NYPD_dataset$PERP_RACEWHITE                     0.163328
## NYPD_dataset$PERP_RACEWHITE HISPANIC            0.316428
## NYPD_dataset$PERP_SEXM                          0.081963 .
## NYPD_dataset$PERP_SEXU                          0.068252 .
## NYPD_dataset$PERP_AGE_GROUP18-24                0.017832 *
## NYPD_dataset$PERP_AGE_GROUP25-44                5.92e-14 ***
## NYPD_dataset$PERP_AGE_GROUP45-64                2.39e-15 ***
## NYPD_dataset$PERP_AGE_GROUP65+                  0.000123 ***
## NYPD_dataset$PERP_AGE_GROUPUNKNOWN               < 2e-16 ***
## NYPD_dataset$OCCUR_TIME                         0.627848
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1505454)
##
##     Null deviance: 2583.1  on 16248  degrees of freedom
## Residual deviance: 2444.0  on 16234  degrees of freedom
##   (9344 observations deleted due to missingness)
## AIC: 15362
##
## Number of Fisher Scoring iterations: 2
```

Analysis, Conclusions & Bias

Based on the results, we can come to the conclusion that the top 3 dangerous boroughs in NY are Brooklyn, Bronx and Queens. In order to minimize the number of incidents, the advice is to stay at home from 20:00 to 00:00 as this is the time frame that shows more incident numbers.

As we can see in the second plot and also using the information from the tables, the highest numbers of incidents come from the group of 25-44, more related with Male than with Female and with white hispanic and black races.

In regards of the model, I found in some literature and web pages that a logistic regression could be used to predict qualitative responses and because we have some qualitative fields like STATISTICAL_MURDER_FLAG I considered it could work. The idea is to relate a murder case with specific groups and incident time.

The bias I have from what I read on the news and media is that the Bronx is the top one of the most dangerous boroughs in NYC and would think that women would be more affected than men.

Now that I did this data driven analysis, I would like to compare it with the news from today trying to minimize or delete at all the bias I had related with specific groups or populations.