NYPD Shooting Incident Data Report

Data Loading

The data for this project comes from the NYPD Shooting Incident Data (Historic) dataset. First, we are going to download the data and take a preliminary look at it.

```
data_url = "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
data = read.csv(data_url, na.strings=c("NA",""))
```

```
head(data)
```

```
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME     BORO LOC_OF_OCCUR_DESC PRECINCT
## 1    228798151 05/27/2021   21:30:00   QUEENS              <NA>      105
## 2    137471050 06/27/2014   17:40:00    BRONX              <NA>       40
## 3    147998800 11/21/2015   03:56:00   QUEENS              <NA>      108
## 4    146837977 10/09/2015   18:30:00    BRONX              <NA>       44
## 5     58921844 02/19/2009   22:58:00    BRONX              <NA>       47
## 6    219559682 10/21/2020   21:36:00 BROOKLYN              <NA>       81
##   JURISDICTION_CODE LOC_CLASSFCTN_DESC LOCATION_DESC STATISTICAL_MURDER_FLAG
## 1                 0               <NA>          <NA>                  false
## 2                 0               <NA>          <NA>                  false
## 3                 0               <NA>          <NA>                   true
## 4                 0               <NA>          <NA>                  false
## 5                 0               <NA>          <NA>                   true
## 6                 0               <NA>          <NA>                   true
##   PERP_AGE_GROUP PERP_SEX PERP_RACE VIC_AGE_GROUP VIC_SEX       VIC_RACE
## 1           <NA>     <NA>      <NA>         18-24       M          BLACK
## 2           <NA>     <NA>      <NA>         18-24       M          BLACK
## 3           <NA>     <NA>      <NA>         25-44       M          WHITE
## 4           <NA>     <NA>      <NA>           <18       M WHITE HISPANIC
## 5          25-44        M     BLACK         45-64       M          BLACK
## 6           <NA>     <NA>      <NA>         25-44       M          BLACK
##   X_COORD_CD Y_COORD_CD Latitude Longitude
## 1    1058925   180924.0 40.66296 -73.73084
## 2    1005028   234516.0 40.81035 -73.92494
## 3    1007668   209836.5 40.74261 -73.91549
## 4    1006537   244511.1 40.83778 -73.91946
## 5    1024922   262189.4 40.88624 -73.85291
## 6    1004234   186461.7 40.67846 -73.92795
##                                      Lon_Lat
## 1 POINT (-73.73083868899994 40.662964620000025)
## 2  POINT (-73.92494232599995 40.81035186300006)
## 3  POINT (-73.91549174199997 40.74260663300004)
## 4  POINT (-73.91945661499994 40.83778200300003)
## 5  POINT (-73.85290950899997 40.88623791800006)
## 6 POINT (-73.92795224099996 40.678456718000064)
```

```
summary(data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME            BORO
##  Min.   :  9953245   Length:27312       Length:27312       Length:27312
##  1st Qu.: 63860880   Class :character   Class :character   Class :character
##  Median : 90372218   Mode  :character   Mode  :character   Mode  :character
##  Mean   :120860536
##  3rd Qu.:188810230
##  Max.   :261190187
##
##  LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:27312       Min.   :  1.00   Min.   :0.0000    Length:27312
##  Class :character   1st Qu.: 44.00   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 68.00   Median :0.0000    Mode  :character
##                     Mean   : 65.64   Mean   :0.3269
##                     3rd Qu.: 81.00   3rd Qu.:0.0000
##                     Max.   :123.00   Max.   :2.0000
##                                      NA's   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:27312       Length:27312            Length:27312
##  Class :character   Class :character        Class :character
##  Mode  :character   Mode  :character        Mode  :character
##
##
##
##
##    PERP_SEX           PERP_RACE          VIC_AGE_GROUP        VIC_SEX
##  Length:27312       Length:27312       Length:27312       Length:27312
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_RACE            X_COORD_CD        Y_COORD_CD        Latitude
##  Length:27312       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000028   1st Qu.:182834   1st Qu.:40.67
##  Mode  :character   Median :1007731   Median :194487   Median :40.70
##                     Mean   :1009449   Mean   :208127   Mean   :40.74
##                     3rd Qu.:1016838   3rd Qu.:239518   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA's   :10
##    Longitude         Lon_Lat
##  Min.   :-74.25   Length:27312
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :10
```

The dataset contains 27312 rows and 21 columns. Some of the location columns are showing the same information or have different levels of precision. The only column which has an unclear interpretation is STATISTICAL_MURDER_FLAG. To see what it means we can read the footnotes of the dataset on its official page.

The footnotes say that STATISTICAL_MURDER_FLAG is a boolean that indicated whether a shooting

incident resulted in a murder. Additionally, the footnotes contain some additional information of interest. For example, if a shooting incident resulted in multiple victims, the dataset contains a row for each of the victims and those rows have the same INCIDENT_KEY. Also, the dataset contains shooting incidents only with victims, i.e. the ones resulting in an injury or death.

For our use, we need to clean the dataset, which will include removing some columns and casting others to correct types. First of all, we are going to leave only the columns, which might use. Thus, we will remove all the columns connected with location except for the borough since this feature should be representive enough of the location without being too precise or verbose. Secondly, we need to correct the types of OCCUR_DATE and OCCUR_TIME from strings to date/time. Finally, all the other columns need to be converted to a factor.

```
data = data %>%
  mutate(
    OCCUR_DATE=mdy(OCCUR_DATE),
    OCCUR_TIME=hms(OCCUR_TIME),
    INCIDENT_KEY=factor(INCIDENT_KEY),
    BORO=factor(BORO),
    STATISTICAL_MURDER_FLAG=factor(STATISTICAL_MURDER_FLAG),
    PERP_AGE_GROUP=factor(PERP_AGE_GROUP),
    PERP_SEX=factor(PERP_SEX),
    PERP_RACE=factor(PERP_RACE),
    VIC_AGE_GROUP=factor(VIC_AGE_GROUP),
    VIC_SEX=factor(VIC_SEX),
    VIC_RACE=factor(VIC_RACE)
  ) %>%
  select(
    -c(Lon_Lat, X_COORD_CD, Y_COORD_CD, PRECINCT, JURISDICTION_CODE, LOCATION_DESC, LOC_OF_OCCUR_DESC, L
  )
head(data)
```

```
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME     BORO STATISTICAL_MURDER_FLAG
## 1    228798151 2021-05-27 21H 30M 0S   QUEENS                  false
## 2    137471050 2014-06-27 17H 40M 0S    BRONX                  false
## 3    147998800 2015-11-21  3H 56M 0S   QUEENS                   true
## 4    146837977 2015-10-09 18H 30M 0S    BRONX                  false
## 5     58921844 2009-02-19 22H 58M 0S    BRONX                   true
## 6    219559682 2020-10-21 21H 36M 0S BROOKLYN                   true
##   PERP_AGE_GROUP PERP_SEX PERP_RACE VIC_AGE_GROUP VIC_SEX       VIC_RACE
## 1           <NA>     <NA>      <NA>         18-24       M          BLACK
## 2           <NA>     <NA>      <NA>         18-24       M          BLACK
## 3           <NA>     <NA>      <NA>         25-44       M          WHITE
## 4           <NA>     <NA>      <NA>           <18       M WHITE HISPANIC
## 5          25-44        M     BLACK         45-64       M          BLACK
## 6           <NA>     <NA>      <NA>         25-44       M          BLACK
```

```
summary(data)
```

```
##     INCIDENT_KEY     OCCUR_DATE           OCCUR_TIME
##  173354054:   18   Min.   :2006-01-01   Min.   :0S
##  23749375 :   12   1st Qu.:2009-07-18   1st Qu.:3H 27M 0S
##  24717013 :   12   Median :2013-04-29   Median :15H 11M 0S
##  33478089 :   12   Mean   :2014-01-06   Mean   :12H 41M 31.7091388400731S
##  33706902 :   12   3rd Qu.:2018-10-15   3rd Qu.:20H 45M 0S
##  35803777 :   12   Max.   :2022-12-31   Max.   :23H 59M 0S
##  (Other)  :27234
```

```
##             BORO       STATISTICAL_MURDER_FLAG PERP_AGE_GROUP   PERP_SEX
##  BRONX        : 7937   false:22046             18-24  :6222   (null):  640
##  BROOKLYN     :10933   true : 5266             25-44  :5687   F     :  424
##  MANHATTAN    : 3572                           UNKNOWN:3148   M     :15439
##  QUEENS       : 4094                           <18    :1591   U     : 1499
##  STATEN ISLAND:  776                           (null) : 640   NA's  : 9310
##                                                (Other): 680
##                                                NA's   :9344
##           PERP_RACE     VIC_AGE_GROUP   VIC_SEX
##  BLACK         :11432   <18    : 2839   F: 2615
##  WHITE HISPANIC: 2341   1022   :    1   M:24686
##  UNKNOWN       : 1836   18-24  :10086   U:   11
##  BLACK HISPANIC: 1314   25-44  :12281
##  (null)        :  640   45-64  : 1863
##  (Other)       :  439   65+    :  181
##  NA's          : 9310   UNKNOWN:   61
##                            VIC_RACE
##  AMERICAN INDIAN/ALASKAN NATIVE:   10
##  ASIAN / PACIFIC ISLANDER      :  404
##  BLACK                         :19439
##  BLACK HISPANIC                : 2646
##  UNKNOWN                       :   66
##  WHITE                         :  698
##  WHITE HISPANIC                : 4049
```

After the transformation, we have a couple of problems to look into. The columns PERP_AGE_GROUP, PERP_SEX and PERP_RACE have missing or other factors. While the null should be replaced with unknown, we should take a look what are other factors since they might be erroneous.

```
levels(data$PERP_AGE_GROUP)
```

```
##  [1] "(null)"  "<18"     "1020"    "18-24"   "224"     "25-44"   "45-64"
##  [8] "65+"     "940"     "UNKNOWN"
```

Here, we can see three unusual factors for age groups: 1020, 224 and 940. Those are mostly likely to be mistakes, and I think the best course of action is to replace them with UNKNOWNs.

```
levels(data$PERP_RACE)
```

```
## [1] "(null)"                         "AMERICAN INDIAN/ALASKAN NATIVE"
## [3] "ASIAN / PACIFIC ISLANDER"       "BLACK"
## [5] "BLACK HISPANIC"                 "UNKNOWN"
## [7] "WHITE"                          "WHITE HISPANIC"
```

With PERP_RACE there are no problems.

```
data = data %>%
  mutate(
    PERP_RACE=recode(PERP_RACE,"(null)"="UNKNOWN"),
    PERP_SEX=recode(PERP_SEX,"(null)"="U"),
    PERP_AGE_GROUP=recode(
      PERP_AGE_GROUP,
      "(null)"="UNKNOWN",
      "1020"="UNKNOWN",
      "224"="UNKNOWN",
      "940"="UNKNOWN"
    ),
```

```
    VIC_AGE_GROUP=recode(VIC_AGE_GROUP, "1022"="UNKNOWN")
  ) %>%
  mutate(
    PERP_RACE=replace_na(PERP_RACE, "UNKNOWN"),
    PERP_SEX=replace_na(PERP_SEX, "U"),
    PERP_AGE_GROUP=replace_na(PERP_AGE_GROUP, "UNKNOWN")
  ) %>%
  mutate(
    PERP_SEX=factor(PERP_SEX, levels=levels(data$VIC_SEX))
  )

head(data)
```

```
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME     BORO STATISTICAL_MURDER_FLAG
## 1    228798151 2021-05-27 21H 30M 0S   QUEENS                  false
## 2    137471050 2014-06-27 17H 40M 0S    BRONX                  false
## 3    147998800 2015-11-21  3H 56M 0S   QUEENS                   true
## 4    146837977 2015-10-09 18H 30M 0S    BRONX                  false
## 5     58921844 2009-02-19 22H 58M 0S    BRONX                   true
## 6    219559682 2020-10-21 21H 36M 0S BROOKLYN                   true
##   PERP_AGE_GROUP PERP_SEX PERP_RACE VIC_AGE_GROUP VIC_SEX       VIC_RACE
## 1        UNKNOWN        U   UNKNOWN         18-24       M          BLACK
## 2        UNKNOWN        U   UNKNOWN         18-24       M          BLACK
## 3        UNKNOWN        U   UNKNOWN         25-44       M          WHITE
## 4        UNKNOWN        U   UNKNOWN           <18       M WHITE HISPANIC
## 5          25-44        M     BLACK         45-64       M          BLACK
## 6        UNKNOWN        U   UNKNOWN         25-44       M          BLACK
```

```
summary(data)
```

```
##     INCIDENT_KEY     OCCUR_DATE           OCCUR_TIME
##  173354054:   18   Min.   :2006-01-01   Min.   :0S
##  23749375 :   12   1st Qu.:2009-07-18   1st Qu.:3H 27M 0S
##  24717013 :   12   Median :2013-04-29   Median :15H 11M 0S
##  33478089 :   12   Mean   :2014-01-06   Mean   :12H 41M 31.7091388400731S
##  33706902 :   12   3rd Qu.:2018-10-15   3rd Qu.:20H 45M 0S
##  35803777 :   12   Max.   :2022-12-31   Max.   :23H 59M 0S
##  (Other)  :27234
##             BORO       STATISTICAL_MURDER_FLAG PERP_AGE_GROUP  PERP_SEX
##  BRONX        : 7937   false:22046             UNKNOWN:13135   F:  424
##  BROOKLYN     :10933   true : 5266             <18    : 1591   M:15439
##  MANHATTAN    : 3572                           18-24  : 6222   U:11449
##  QUEENS       : 4094                           25-44  : 5687
##  STATEN ISLAND:  776                           45-64  :  617
##                                                65+    :   60
##
##                            PERP_RACE     VIC_AGE_GROUP   VIC_SEX
##  UNKNOWN                         :11786   <18    : 2839   F: 2615
##  AMERICAN INDIAN/ALASKAN NATIVE:    2   UNKNOWN:   62   M:24686
##  ASIAN / PACIFIC ISLANDER       :  154   18-24  :10086   U:   11
##  BLACK                          :11432   25-44  :12281
##  BLACK HISPANIC                 : 1314   45-64  : 1863
##  WHITE                          :  283   65+    :  181
##  WHITE HISPANIC                 : 2341
```

```
##                              VIC_RACE
##  AMERICAN INDIAN/ALASKAN NATIVE:   10
##  ASIAN / PACIFIC ISLANDER      :  404
##  BLACK                         :19439
##  BLACK HISPANIC                : 2646
##  UNKNOWN                       :   66
##  WHITE                         :  698
##  WHITE HISPANIC                : 4049
```

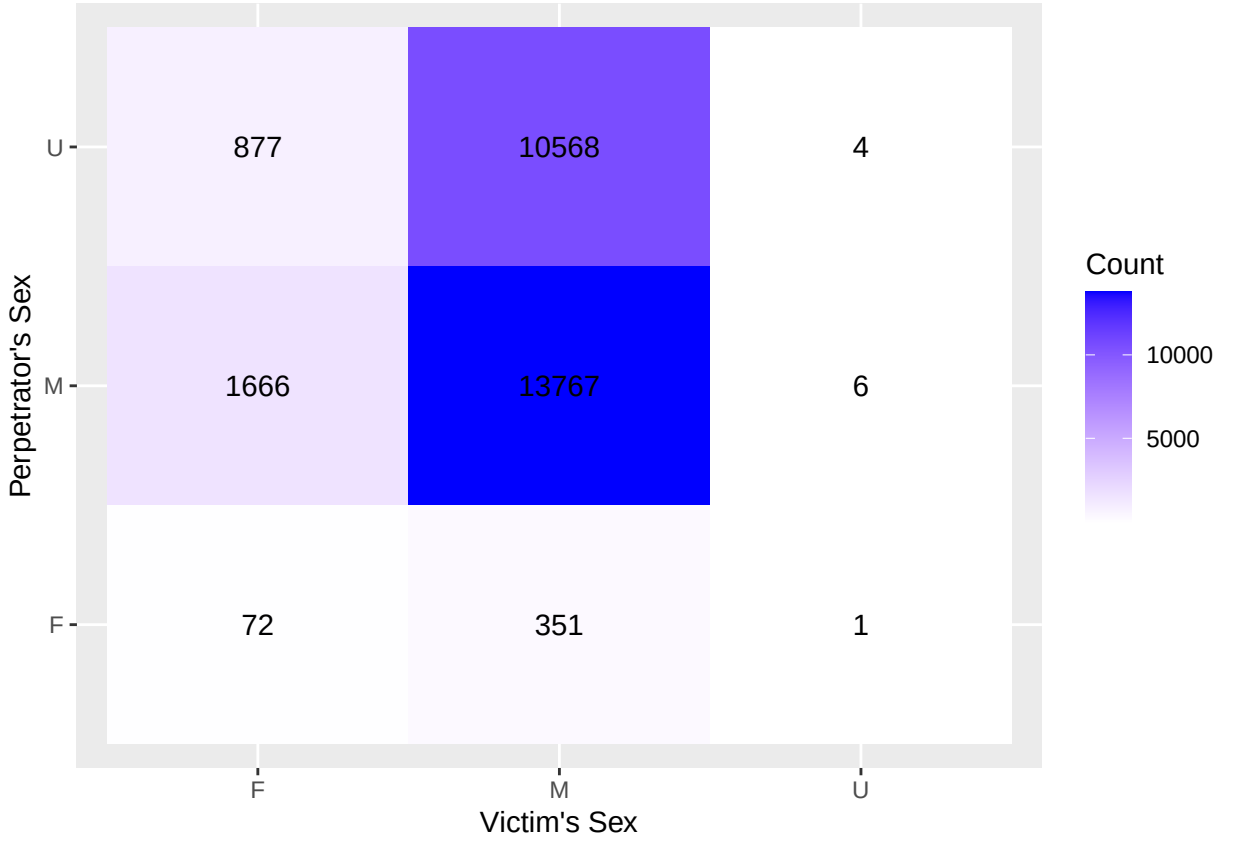
After the last pass the data looks good. There does not seem to be any problems with it and we can move on.

First, let's take a look at amounts of total shooting incidents and those resulting in a death by sexes of a victim and a perpetrator.

```
incidents_by_sex = data %>% count(PERP_SEX, VIC_SEX)
```

```
murders_by_sex = data %>%
  filter(STATISTICAL_MURDER_FLAG == "true") %>%
  count(PERP_SEX, VIC_SEX)
```

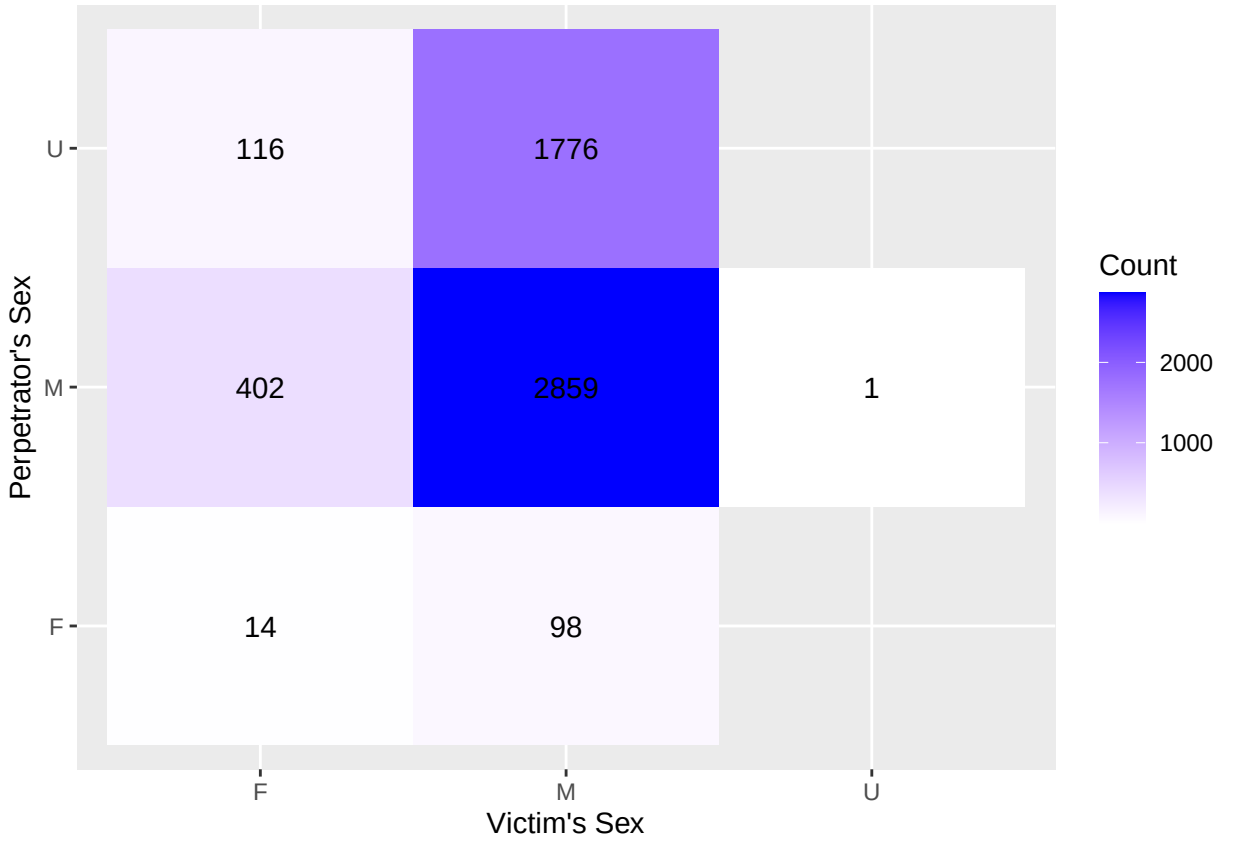
```
ggplot(incidents_by_sex, aes(x=VIC_SEX, y=PERP_SEX, fill=n, label=n)) +
  geom_tile() +
  geom_text() +
  scale_fill_gradient(low="white", high="blue") +
  labs(x="Victim's Sex", y="Perpetrator's Sex", fill="Count")
```



For shooting incidents in general, there are clear differences based on the sexes. The largest amount of incidents occur when both the victim and the perpetrator are male, the least out of known occur when both are female.

There is a significant amount of incidents where the perpetrator's sex is unknown. However, I don't think that can skew the difference for known sexes since it is too large compared to the overall number of incidents. Another interesting thing is that there are shooting incidents where the victim's sex is unknown.

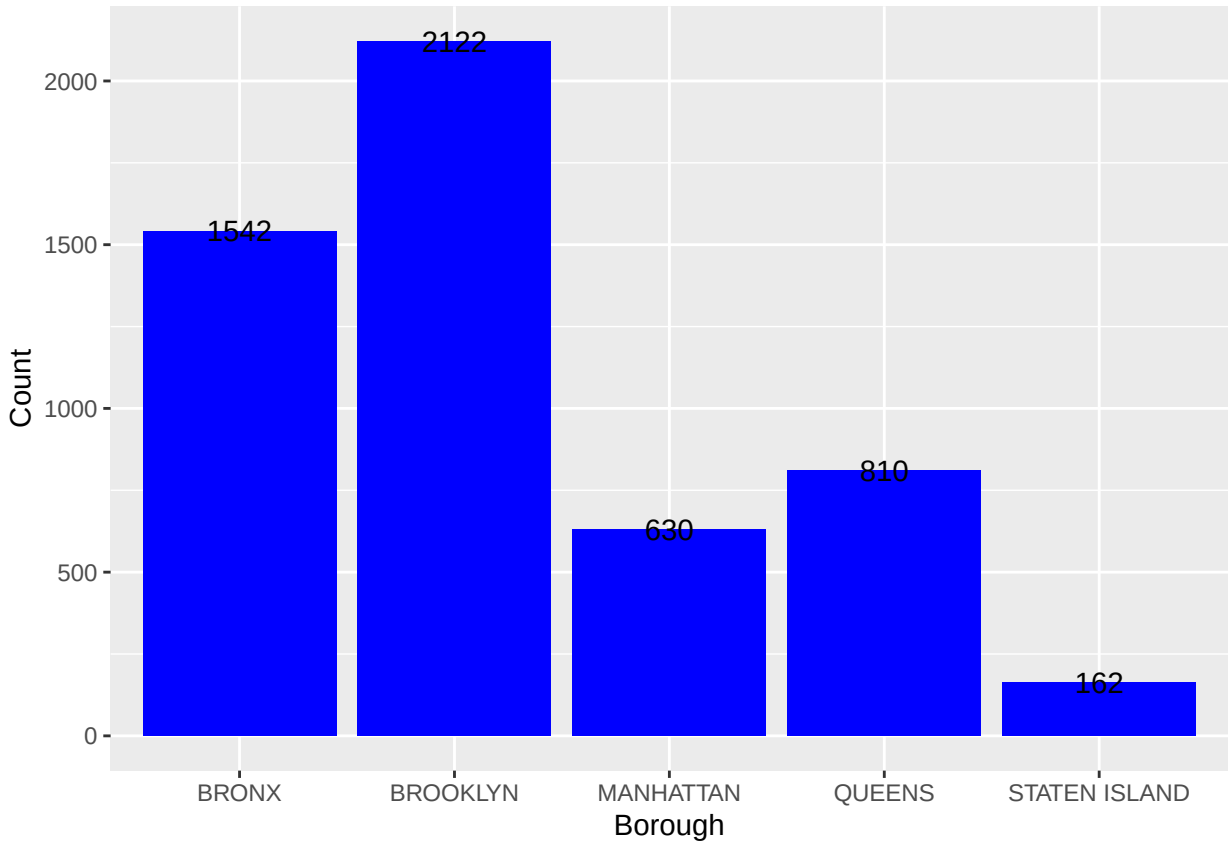
```
ggplot(murders_by_sex, aes(x=VIC_SEX, y=PERP_SEX, fill=n, label=n)) +
  geom_tile() +
  geom_text() +
  scale_fill_gradient(low="white", high="blue") +
  labs(x="Victim's Sex", y="Perpetrator's Sex", fill="Count")
```



For shooting incidents resulting in a death, we see similar discrepancies with slight differences in proportions, which might or might not be significant. There is also a victim with unknown sex, which is even more bizarre than previously, since it raises the question of how it was determined that the incident resulted in a death.

In any case, now, we are going to take a deeper look at how different factors affect the distribution of incidents resulting in deaths starting with boroughs.

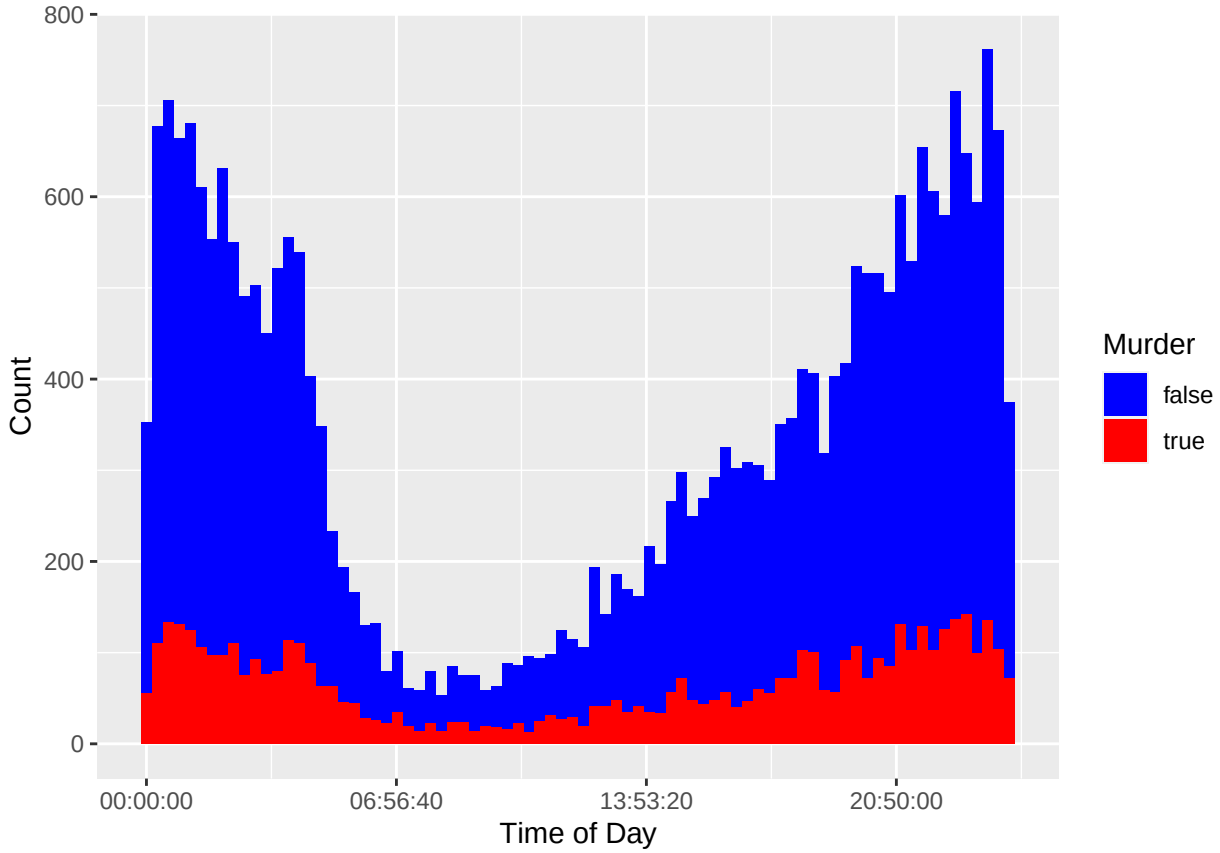
```
murders_by_boro = data %>%
  filter(STATISTICAL_MURDER_FLAG == "true") %>%
  count(BORO)

ggplot(murders_by_boro, aes(x=BORO, y=n, label=n)) +
  geom_bar(stat="identity", fill="blue") +
  geom_text() +
  scale_fill_gradient(low="white", high="blue") +
  labs(x="Borough", y="Count")
```



The graph shows that there are clear differences between the amounts of lethal incidents in each of them. However, there is a chance that the differences can be explained by the populations and sizes of those boroughs. But if that were the case, we wouldn't expect to see a difference between proportions of non-lethal to lethal incidents in those boroughs. We are going to take a look at that later when we will be fitting the data.

Now, we are going to take a look at how the time of day affects shooting incidents.

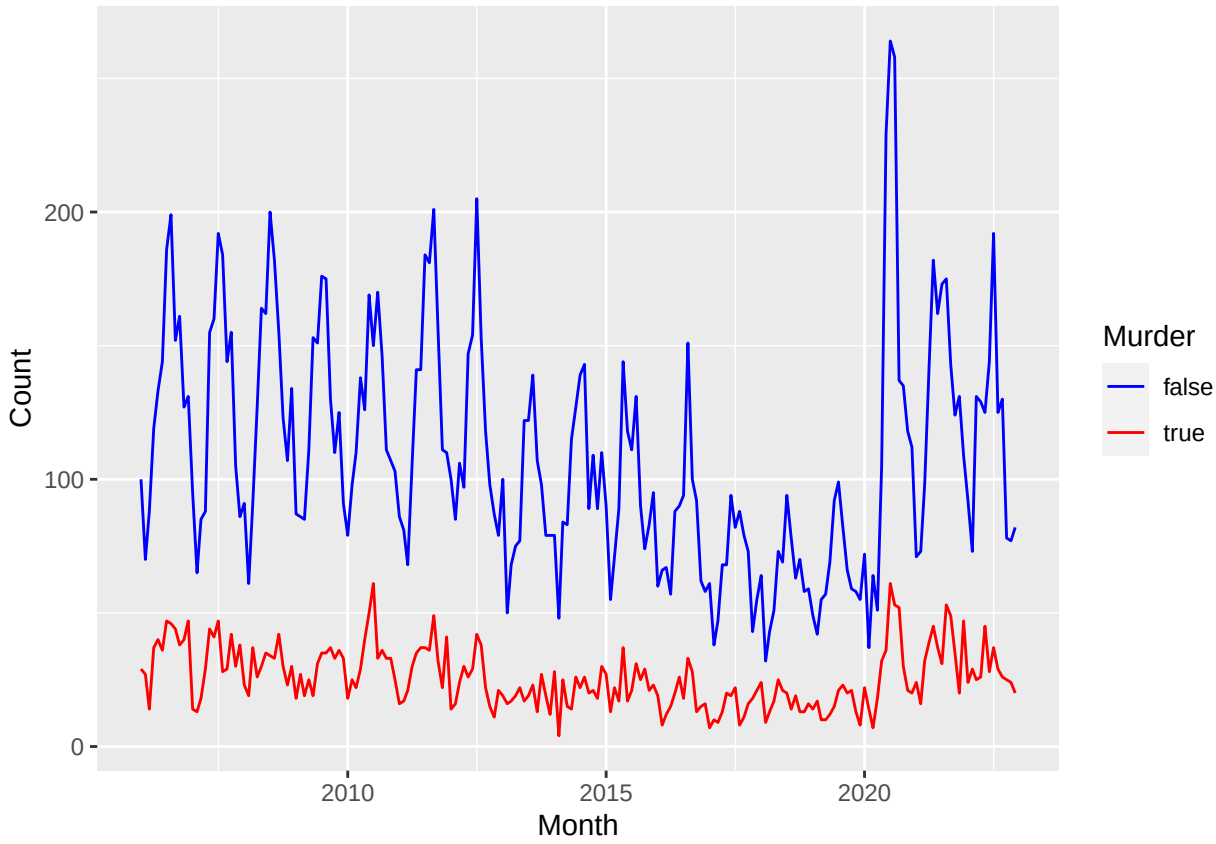
```
ggplot(
  data %>% mutate(OCCUR_TIME=as.numeric(OCCUR_TIME)),
  aes(x=OCCUR_TIME, fill=STATISTICAL_MURDER_FLAG)
) +
  geom_histogram(bins=80) +
  scale_fill_manual(values=c("blue", "red")) +
  scale_x_continuous(labels=function(x) format(as.POSIXct(x, origin="2022-01-01", tz="UTC"), "%H:%M:%S"
  labs(x="Time of Day", y="Count", fill="Murder")
```



As expected, the number of shooting incidents increases sharply during the night and evening hours. However, the proportion of murders during the day might be slightly higher than during other hours, which would be quite interesting. This also should be investigated.

The next step is to take at historic distribution of shooting incidents to see if there are any trends. For that we are going to plot monthly shooting incidents.

```
monthly = data %>%
  mutate(month=floor_date(OCCUR_DATE, "month")) %>%
  count(month, STATISTICAL_MURDER_FLAG)

ggplot(monthly, aes(x=month, y=n, color=STATISTICAL_MURDER_FLAG)) +
  geom_line() +
  labs(x="Month", y="Count", color="Murder") +
  scale_color_manual(values=c("blue", "red"))
```



From the plot, we can see that there is a clear correlation between lethal and non-lethal shooting incidents. Additionally, winter seems to be the period where there is the least amount of incidents of both types. There is also a general trend until 2020 where the amount of non-lethal incidents decreases faster than lethal incidents, but after that, at the start of the pandemic, both rise sharply.

With the data looked at, we can see that there is quite a few patterns that can be explored further in the dataset. We are going to focus on lethal incidents. For that we are going to fit a logistic regression model in order to predict if an incident results in a murder.

The first model is going to predict the probability that the incident is lethal from the sex of the victim only. Here, we wouldn't expect the used feature to be statistically significant.

```
model_sex = glm(STATISTICAL_MURDER_FLAG ~ VIC_SEX, data=data, family=binomial)
summary(model_sex)
```

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ VIC_SEX, family = binomial,
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6745  -0.6525  -0.6525  -0.6525   2.1899
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.36492    0.04858 -28.098   <2e-16 ***
## VIC_SEXM    -0.07390    0.05120  -1.443    0.149
## VIC_SEXU    -0.93766    1.04972  -0.893    0.372
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 26781  on 27311  degrees of freedom
## Residual deviance: 26778  on 27309  degrees of freedom
## AIC: 26784
##
## Number of Fisher Scoring iterations: 4
```

```
predict_data = expand.grid(VIC_SEX=levels(data$VIC_SEX))
predict_data$prob = predict(model_sex, newdata=predict_data, type="response")
predict_data
```

```
##   VIC_SEX      prob
## 1       F 0.2034417
## 2       M 0.1917281
## 3       U 0.0909091
```

Looking at both the model summary and predicted probabilities, we see that the difference is insignificant. Now, we will fit a model with all the features about the victim, the location, the time of day, the month and the year.

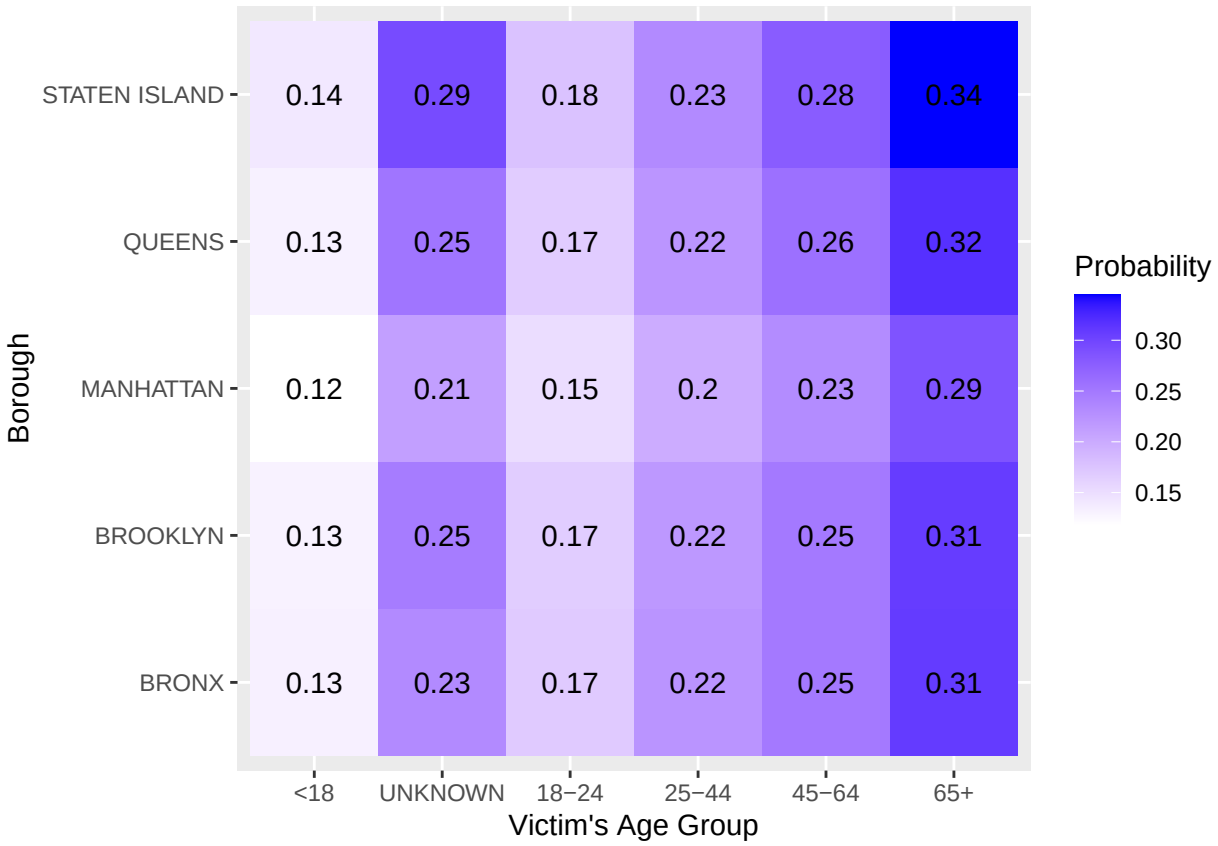
```
data = data %>%
  mutate(
    time_num=as.numeric(OCCUR_TIME),
    month=month(OCCUR_DATE),
    year=year(OCCUR_DATE)
  )
model = glm(STATISTICAL_MURDER_FLAG ~ VIC_SEX + BORO + time_num + VIC_RACE + VIC_AGE_GROUP + month + ye
summary(model)
```

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ VIC_SEX + BORO + time_num +
##     VIC_RACE + VIC_AGE_GROUP + month + year, family = binomial,
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9818  -0.6973  -0.6121  -0.5325   2.3172
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -9.982e-02  1.024e+02  -0.001   0.9992
## VIC_SEXM                        -5.173e-02  5.211e-02  -0.993   0.3208
## VIC_SEXU                        -5.788e-01  1.083e+00  -0.534   0.5932
## BOROBROOKLYN                    -2.540e-02  3.882e-02  -0.654   0.5130
## BOROMANHATTAN                   -1.355e-01  5.265e-02  -2.575   0.0100 *
## BOROQUEENS                      -2.419e-02  4.951e-02  -0.489   0.6251
## BOROSTATEN ISLAND                2.850e-02  9.422e-02   0.302   0.7623
## time_num                         3.841e-07  5.077e-07   0.757   0.4493
## VIC_RACEASIAN / PACIFIC ISLANDER  1.131e+01  1.022e+02   0.111   0.9119
## VIC_RACEBLACK                    1.102e+01  1.022e+02   0.108   0.9141
## VIC_RACEBLACK HISPANIC           1.085e+01  1.022e+02   0.106   0.9155
```

```
## VIC_RACEUNKNOWN                   1.027e+01  1.022e+02   0.100   0.9200
## VIC_RACEWHITE                     1.135e+01  1.022e+02   0.111   0.9116
## VIC_RACEWHITE HISPANIC            1.114e+01  1.022e+02   0.109   0.9132
## VIC_AGE_GROUPUNKNOWN              8.488e-01  3.153e-01   2.692   0.0071 **
## VIC_AGE_GROUP18-24                2.910e-01  6.212e-02   4.685 2.80e-06 ***
## VIC_AGE_GROUP25-44                6.275e-01  6.046e-02  10.378  < 2e-16 ***
## VIC_AGE_GROUP45-64                7.753e-01  7.816e-02   9.919  < 2e-16 ***
## VIC_AGE_GROUP65+                  1.025e+00  1.716e-01   5.973 2.33e-09 ***
## month                             3.374e-03  4.923e-03   0.685   0.4931
## year                             -6.358e-03  3.031e-03  -2.098   0.0359 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 26781  on 27311  degrees of freedom
## Residual deviance: 26491  on 27291  degrees of freedom
## AIC: 26533
##
## Number of Fisher Scoring iterations: 11
```

For this model we have a few significant predictors. First of all, the age group of the victim, which makes sense since we would expected different chances of a successful recovery based on the overall health of the victim and age is a good predictor for that. Another significant predictor is the year. We saw from the daily incidents plot that there might have been a difference and this reinforces that suspicion. Finally, there is a statistically significant difference based on the borough, which is a bit unusual. So, let's take a look at the predicted probabilities based on the borough and the age group of the victim.

```
predictions = copy(data)
predictions$prob = predict(model, newdata=predictions, type="response")
predictions = predictions %>%
  group_by(BORO, VIC_AGE_GROUP) %>%
  summarize(prob=mean(prob)) %>%
  ungroup()

ggplot(predictions, aes(y=BORO, x=VIC_AGE_GROUP, fill=prob, label=prob)) +
  geom_tile() +
  geom_text(aes(label=round(prob,2))) +
  scale_fill_gradient(low="white", high="blue") +
  labs(x="Victim's Age Group", y="Borough", fill="Probability")
```



From the predicted probabilities, we can see that the probability of a lethal incidents increase with the the age for all boroughs. However, the base probability for Manhattan is a couple percent lower for all groups. This is pretty strange and I have difficulties coming up with a solid explanation for that. In any case, I believe this would need further investigation, but that goes out of scope of this basic analysis of the dataset.

With the analysis done, I want to highlight a few biases both in the data and of my own, which might have affected the results. First of all, the data contains only the incidents which resulted in an injury. Secondly, due to economic, geographic and population difference between boroughs, the emergency services availability might be different between them, which can affect how the data is collected or reported. These factors are important to consider since they can affect the data quite dramatically. Finally, I only have a cursory knowledge of New York and shooting incidents in general. This means I might be looking only into a surface level connections, which might be missing underlying reasons, in particular the ones connected with boroughs.