- This is a list of every shooting incident that occurred in NYC going back to 2006 through the end of 2020. This data is manually extracted every quarter and reviewed by the Office of Management Analysis and Planning before being posted on the NYPD website. Each record represents a shooting incident in NYC and includes information about the event, the location and time of occurrence. In addition, information related to suspect and victim demographics is also included.
- Data was downloaded from https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic

Step 1: Start an Rmd Document

In [47]:
```
# install libraries
library(tidyverse)
library(lubridate)
library(ggplot2)
```

In [2]:
```
#install.packages("plotrix")
library(plotrix)
```

In [3]:
```
# read data
data <- read.csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType
```

In [4]:
```
# view some first rows of data
head(data)
```

INCIDENT_KEY	OCCUR_DATE	OCCUR_TIME	BORO	PRECINCT	JURISDICTION_CODE	LOCATION_DES
24050482	08/27/2006	05:35:00	BRONX	52	0	
77673979	03/11/2011	12:03:00	QUEENS	106	0	
203350417	10/06/2019	01:09:00	BROOKLYN	77	0	
80584527	09/04/2011	03:35:00	BRONX	40	0	
90843766	05/27/2013	21:16:00	QUEENS	100	0	
92393427	09/01/2013	04:17:00	BROOKLYN	67	0	

Step 2: Tidy and Transform Data

In [5]:
```
# remove some unused columns of data and view some first rows of data
nypd <- subset (data, select = -c(X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat)
head(nypd)
```

INCIDENT_KEY	OCCUR_DATE	OCCUR_TIME	BORO	PRECINCT	JURISDICTION_CODE	LOCATION_DES
24050482	08/27/2006	05:35:00	BRONX	52	0	
77673979	03/11/2011	12:03:00	QUEENS	106	0	
203350417	10/06/2019	01:09:00	BROOKLYN	77	0	
80584527	09/04/2011	03:35:00	BRONX	40	0	
90843766	05/27/2013	21:16:00	QUEENS	100	0	
92393427	09/01/2013	04:17:00	BROOKLYN	67	0	

In [6]:
```
# summary data
summary(nypd)
```

```
 INCIDENT_KEY            OCCUR_DATE       OCCUR_TIME               BORO
Min.   :  9953245   07/05/2020:   47   23:30:00:  159   BRONX        :6701
1st Qu.: 55322804   09/04/2011:   31   01:30:00:  141   BROOKLYN     :9734
Median : 83435362   07/26/2020:   29   00:30:00:  136   MANHATTAN    :2922
Mean   :102280741   08/11/2007:   26   02:00:00:  129   QUEENS       :3532
3rd Qu.:150911774   09/04/2006:   25   21:00:00:  128   STATEN ISLAND: 696
Max.   :230611229   08/15/2020:   24   22:30:00:  126
                    (Other)   :23403   (Other) :22766
   PRECINCT      JURISDICTION_CODE                    LOCATION_DESC
Min.   :  1.00   Min.   :0.000                              :13581
1st Qu.: 44.00   1st Qu.:0.000     MULTI DWELL - PUBLIC HOUS: 4240
Median : 69.00   Median :0.000     MULTI DWELL - APT BUILD  : 2553
Mean   : 66.21   Mean   :0.333     PVT HOUSE                :  857
3rd Qu.: 81.00   3rd Qu.:0.000     GROCERY/BODEGA           :  574
Max.   :123.00   Max.   :2.000     BAR/NIGHT CLUB           :  562
                 NA's   :2         (Other)                  : 1218
STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX          PERP_RACE
false:19085                    :8295    : 8261   BLACK         :10025
true : 4500             18-24  :5508   F:  335                 : 8261
                        25-44  :4714   M:13490   WHITE HISPANIC: 1988
                        UNKNOWN:3148   U: 1499   UNKNOWN       : 1836
                        <18    :1368             BLACK HISPANIC: 1096
                        45-64  : 495             WHITE         :  255
                        (Other):  57             (Other)       :  124
VIC_AGE_GROUP   VIC_SEX                              VIC_RACE
<18    : 2525   F: 2204   AMERICAN INDIAN/ALASKAN NATIVE:    9
18-24  : 9003   M:21370   ASIAN / PACIFIC ISLANDER      :  327
25-44  :10303   U:   11   BLACK                         :16869
45-64  : 1541             BLACK HISPANIC                : 2245
65+    :  154             UNKNOWN                       :   65
```

```
UNKNOWN:    59            WHITE                       :  620
                          WHITE HISPANIC              : 3450
```

In [7]:
```
# Quick glimpse data also tells us the number of rows (observations), columns (variable
glimpse(nypd)
```

```
Observations: 23,585
Variables: 14
$ INCIDENT_KEY           <int> 24050482, 77673979, 203350417, 80584527, 90...
$ OCCUR_DATE             <fct> 08/27/2006, 03/11/2011, 10/06/2019, 09/04/2...
$ OCCUR_TIME             <fct> 05:35:00, 12:03:00, 01:09:00, 03:35:00, 21:...
$ BORO                   <fct> BRONX, QUEENS, BROOKLYN, BRONX, QUEENS, BRO...
$ PRECINCT               <int> 52, 106, 77, 40, 100, 67, 77, 81, 101, 106,...
$ JURISDICTION_CODE      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
$ LOCATION_DESC          <fct> , , , , , , , , , , , , , , , , , , , , , ,...
$ STATISTICAL_MURDER_FLAG <fct> true, false, false, false, false, false, fa...
$ PERP_AGE_GROUP         <fct> , , , , , , , , , , , , , , , , , 18-24, , ...
$ PERP_SEX               <fct> , , , , , , , , , , , , , , , , , M, , , , ...
$ PERP_RACE              <fct> , , , , , , , , , , , , , , , , , BLACK, , ...
$ VIC_AGE_GROUP          <fct> 25-44, 65+, 18-24, <18, 18-24, <18, <18, 25...
$ VIC_SEX                <fct> F, M, F, M, M, M, M, M, M, M, F, M, M, M, M...
$ VIC_RACE               <fct> BLACK HISPANIC, WHITE, BLACK, BLACK, BLACK,...
```

- As we can see, there are some blank values in the data. Therefore, to deal with this, first, I'll set blank to NA and then I'll check for missing data.

In [8]:
```
# set blank to NA
nypd[nypd == ""] <- NA
```

In [9]:
```
#Checking for missing data
sapply(nypd,function(x) sum(is.na(x)))
```

INCIDENT_KEY	0
OCCUR_DATE	0
OCCUR_TIME	0
BORO	0
PRECINCT	0
JURISDICTION_CODE	2
LOCATION_DESC	13581
STATISTICAL_MURDER_FLAG	0
PERP_AGE_GROUP	8295
PERP_SEX	8261
PERP_RACE	8261
VIC_AGE_GROUP	0
VIC_SEX	0
VIC_RACE	0

- As we see above, there are 2 missing values of JURISDICTION_CODE and a lot of missing values of LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX and PERP_RACE. To deal with this, I'll delete 2

missing values of JURISDICTION_CODE, fill "NONE" for the missing values of LOCATION_DESC and fill "UNKNOWN" for the missing values of PERP_AGE_GROUP, PERP_SEX and PERP_RACE.

In [10]:
```
# delete missing values in JURISDICTION_CODE
nypd <- nypd %>% drop_na(JURISDICTION_CODE)
```

In [11]:
```
# replace NA with NONE in column LOCATION_DESC
nypd$LOCATION_DESC[is.na(nypd$LOCATION_DESC)] <- "NONE"
```

In [12]:
```
# replace NA with UNKNOWN in column PERP_AGE_GROUP, PERP_SEX and PERP_RACE
nypd$PERP_SEX <- sapply(nypd$PERP_SEX, as.character) # since our values are `factor`
nypd[is.na(nypd)] <- "UNKNOWN"
```

In [13]:
```
# check missing values again
sapply(nypd,function(x) sum(is.na(x)))
```

INCIDENT_KEY	0
OCCUR_DATE	0
OCCUR_TIME	0
BORO	0
PRECINCT	0
JURISDICTION_CODE	0
LOCATION_DESC	0
STATISTICAL_MURDER_FLAG	0
PERP_AGE_GROUP	0
PERP_SEX	0
PERP_RACE	0
VIC_AGE_GROUP	0
VIC_SEX	0
VIC_RACE	0

In [14]:
```
# summary data again
summary(nypd)
```

```
 INCIDENT_KEY           OCCUR_DATE       OCCUR_TIME               BORO
Min.   :  9953245   07/05/2020:   47   23:30:00:  159   BRONX        :6701
1st Qu.: 55322804   09/04/2011:   31   01:30:00:  141   BROOKLYN     :9734
Median : 83435362   07/26/2020:   29   00:30:00:  136   MANHATTAN    :2921
Mean   :102279763   08/11/2007:   26   02:00:00:  129   QUEENS       :3531
3rd Qu.:150895962   09/04/2006:   25   21:00:00:  128   STATEN ISLAND: 696
Max.   :230611229   08/15/2020:   24   22:30:00:  126
                    (Other)   :23401   (Other) :22764
   PRECINCT      JURISDICTION_CODE                  LOCATION_DESC
Min.   :  1.00   Min.   :0.000     NONE                     :13755
1st Qu.: 44.00   1st Qu.:0.000     MULTI DWELL - PUBLIC HOUS: 4240
Median : 69.00   Median :0.000     MULTI DWELL - APT BUILD  : 2553
Mean   : 66.22   Mean   :0.333     PVT HOUSE                :  857
3rd Qu.: 81.00   3rd Qu.:0.000     GROCERY/BODEGA           :  574
```

```
Max.   :123.00   Max.   :2.000     BAR/NIGHT CLUB           :  562
                                   (Other)                  : 1042
STATISTICAL_MURDER_FLAG PERP_AGE_GROUP    PERP_SEX
false:19083             UNKNOWN:11442   Length:23583
true : 4500             18-24  : 5508   Class :character
                        25-44  : 4714   Mode  :character
                        <18    : 1367
                        45-64  :  495
                        65+    :   54
                        (Other):    3
                   PERP_RACE     VIC_AGE_GROUP   VIC_SEX
UNKNOWN                 :10097   <18    : 2525   F: 2204
BLACK                   :10024   18-24  : 9002   M:21368
WHITE HISPANIC          : 1987   25-44  :10302   U:   11
BLACK HISPANIC          : 1096   45-64  : 1541
WHITE                   :  255   65+    :  154
ASIAN / PACIFIC ISLANDER:  122   UNKNOWN:   59
(Other)                 :    2
                          VIC_RACE
AMERICAN INDIAN/ALASKAN NATIVE:    9
ASIAN / PACIFIC ISLANDER      :  327
BLACK                         :16868
BLACK HISPANIC                : 2245
UNKNOWN                       :   65
WHITE                         :  620
WHITE HISPANIC                : 3449
```

Next, in the summary we can see there are 3 other values in the PERP_AGE_GROUP, I think it might be outliers there, so now I'm looking into it.

In [15]:

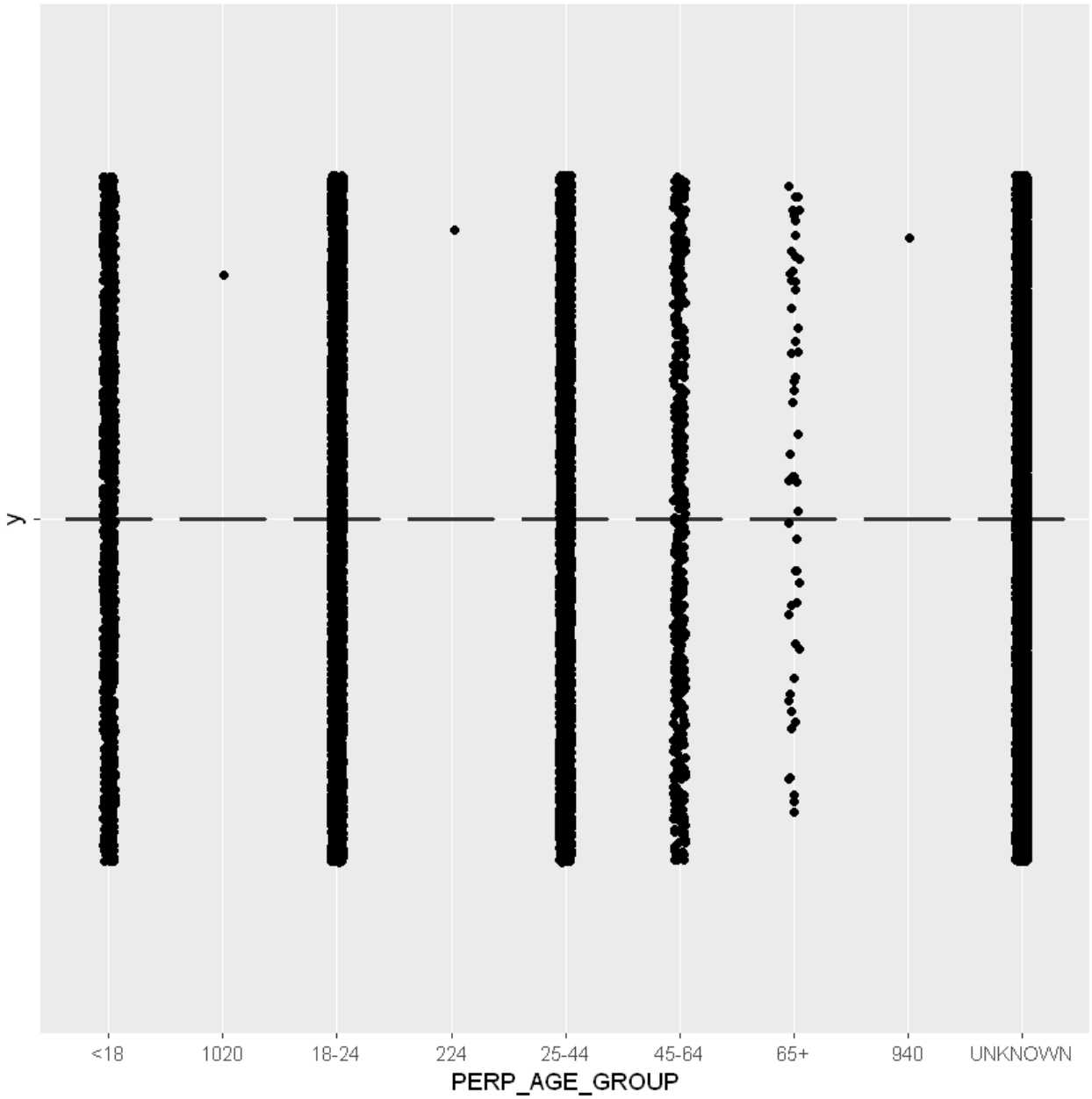
```
# summary PERP_AGE_GROUP column
summary(nypd$PERP_AGE_GROUP)
```

1	0
<18	1367
1020	1
18-24	5508
224	1
25-44	4714
45-64	495
65+	54
940	1
UNKNOWN	11442

As we see, there are 3 outlier values in PERP_AGE_GROUP column: 1020, 224, 940. Now, I'll plot PERP_AGE_GROUP for seeing those outliers more clearly.

In [16]:

```
# plot PERP_AGE_GROUP column
ggplot(nypd, aes(x=PERP_AGE_GROUP, y="")) +
  geom_boxplot()+
  geom_jitter(position=position_jitter(0.05))
```



To solve this problem, I'll remove rows of outlier value out of our data.

In [17]:

```
# delete rows with PERP_AGE_GROUP as 1020 or 224 or 940
nypd <- nypd[!(nypd$PERP_AGE_GROUP=="1020" | nypd$PERP_AGE_GROUP=="224"| nypd$PERP_AGE_

# summary to check PERP_AGE_GROUP column without outlier anymore
summary(nypd$PERP_AGE_GROUP)
```

1	0
<18	1367
1020	0
18-24	5508
224	0
25-44	4714
45-64	495
65+	54

940	0
UNKNOWN	11442

In [18]:
```
#converting dates to standard MM-DD-YYYY format
#nypd$OCCUR_DATE <- mdy(nypd$OCCUR_DATE)
```

In [19]:
```
#Sorting data by dates and view some last rows of data
nypd<- nypd[order(nypd$OCCUR_DATE),]
tail(nypd)
```

	INCIDENT_KEY	OCCUR_DATE	OCCUR_TIME	BORO	PRECINCT	JURISDICTION_CODE	LOCA
6345	206890929	12/31/2019	23:15:00	MANHATTAN	28	0	HC
12563	206891917	12/31/2019	20:14:00	BROOKLYN	73	0	
6997	222446417	12/31/2020	00:42:00	BRONX	44	0	
11512	222468112	12/31/2020	14:59:00	QUEENS	103	0	
13916	222466833	12/31/2020	19:27:00	QUEENS	113	0	
22541	222473262	12/31/2020	23:45:00	MANHATTAN	33	0	

Step 3: Add Visualizations and Analysis

- Now, after cleaning up and check there is no missing data, I'll analyze and visualize data. ### Question 1: How many victims are female, male and unisex?

In [20]:
```
# Number of cases where the victims are female, male and unisex
number_of_victim_female = nrow(filter(nypd, VIC_SEX == "F"))
number_of_victim_male = nrow(filter(nypd, VIC_SEX == "M"))
number_of_victim_unisex = nrow(filter(nypd, VIC_SEX == "U"))
```

In [21]:
```
print(paste("The number of female victims is: ",number_of_victim_female,"."))
print(paste("The number of male victims is: ",number_of_victim_male,"."))
print(paste("The number of unisex victims is: ",number_of_victim_unisex,"."))
```

```
[1] "The number of female victims is:  2204 ."
[1] "The number of male victims is:  21365 ."
[1] "The number of unisex victims is:  11 ."
```

In [22]:
```
# Create the data for the chart
chem <- c("F", "M", "U")
vol <- c(number_of_victim_female, number_of_victim_male, number_of_victim_unisex)
```

In [23]:
```
# create a dataframe x with the catagories of sex and the number of each kind
```

```
x <- list(col1 = chem, col2 = vol)
as.data.frame(x)
x <- rep( c("Female","Male","Unisex"), c(number_of_victim_female,number_of_victim_male,
```

col1	col2
F	2204
M	21365
U	11

Since pie charts are especially useful for proportions, let's have a look on the proportions of our victim's sex, than we will report on the graph in this case:

In [24]:
```
# the proportions of our victim's sex
paste(prop.table(table(x))*100, "%", sep = "")
```

1. '9.34690415606446%'
2. '90.6064461407973%'
3. '0.0466497031382528%'

In [25]:
```
# visualize Victim's sex pie chart
pie3D(table(x), labels = paste(round(prop.table(table(x))*100), "%", sep = ""),
col = heat.colors(3), explode = 0.1, main = "Victim's Sex")
legend("topright", legend = c("Female","Male","Unisex"),
fill = heat.colors(3), title = "Categories", cex = 0.5)
```

Base on the Victim's Sex chart, we can see that the most of victims are male (91%), the least victims are unisex (0.047%), and the remaining is female (9%).

Question 2: How many victim in each range of age?

In [26]:
```
# Number of cases where the victims are female, male and unisex
number_of_victim_U18 = nrow(filter(nypd, VIC_AGE_GROUP == "<18"))
number_of_victim_U24 = nrow(filter(nypd, VIC_AGE_GROUP == "18-24"))
number_of_victim_U44 = nrow(filter(nypd, VIC_AGE_GROUP == "25-44"))
number_of_victim_U64 = nrow(filter(nypd, VIC_AGE_GROUP == "45-64"))
number_of_victim_O65 = nrow(filter(nypd, VIC_AGE_GROUP == "65+"))
number_of_victim_UN = nrow(filter(nypd, VIC_AGE_GROUP == "UNKNOWN"))
```

In [27]:
```
print(paste("The number of under 18 victims is: ",number_of_victim_U18,"."))
print(paste("The number of 18-24 victims is: ",number_of_victim_U24,"."))
print(paste("The number of 25-44 victims is: ",number_of_victim_U44,"."))
print(paste("The number of 45-64 victims is: ",number_of_victim_U64,"."))
```

```
print(paste("The number of 65+ victims is: ",number_of_victim_O65,"."))
print(paste("The number of unknown age victims is: ",number_of_victim_UN,"."))
```

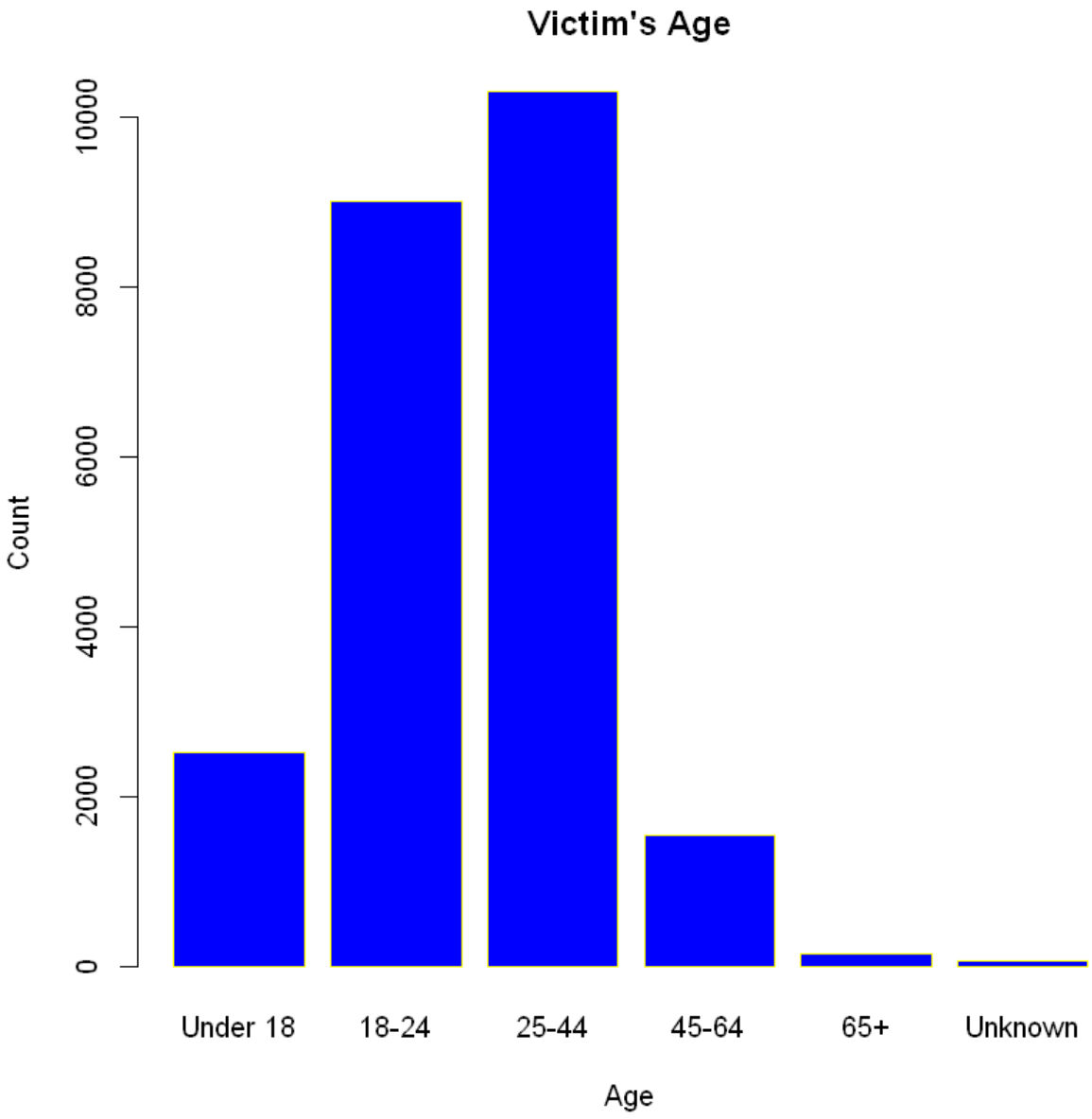
```
[1] "The number of under 18 victims is:  2525 ."
[1] "The number of 18-24 victims is:  9001 ."
[1] "The number of 25-44 victims is:  10300 ."
[1] "The number of 45-64 victims is:  1541 ."
[1] "The number of 65+ victims is:  154 ."
[1] "The number of unknown age victims is:  59 ."
```

In [28]:

```
# Create the data for the chart
kind <- c("Under 18","18-24","25-44","45-64","65+","Unknown")
val <- c(number_of_victim_U18,number_of_victim_U24,number_of_victim_U44,number_of_victi
```

In [29]:

```
# Visualize the number of Victim's Age
b<-barplot(val,names.arg=kind,xlab="Age",ylab="Count",col="blue",main="Victim's Age",bo
```



The histogram above tells us that the victim's age are most at 25-44 and the least are at 65+ and

unknown.

Question 3: How many cases are investigated by each jurisdiction code, occurred at each city where the shooting incident happened?

In [30]:
```
number_of_BOROX = nrow(filter(nypd, BORO == "BRONX"))
number_of_QUEENS = nrow(filter(nypd, BORO == "QUEENS"))
number_of_BROOKLYN = nrow(filter(nypd, BORO == "BROOKLYN"))
number_of_MANHATTAN = nrow(filter(nypd, BORO == "MANHATTAN"))
number_of_S_ISLAND = nrow(filter(nypd, BORO == "STATEN ISLAND"))
```

In [31]:
```
print(paste("The number of shooting incident occurred in Borox is: ",number_of_BOROX,".
print(paste("The number of shooting incident occurred in Queens is: ",number_of_QUEENS,
print(paste("The number of shooting incident occurred in Brooklyn is: ",number_of_BROOK
print(paste("The number of shooting incident occurred in Manhattan is: ",number_of_MANH
print(paste("The number of shooting incident occurred in Staten Island is: ",number_of_
```

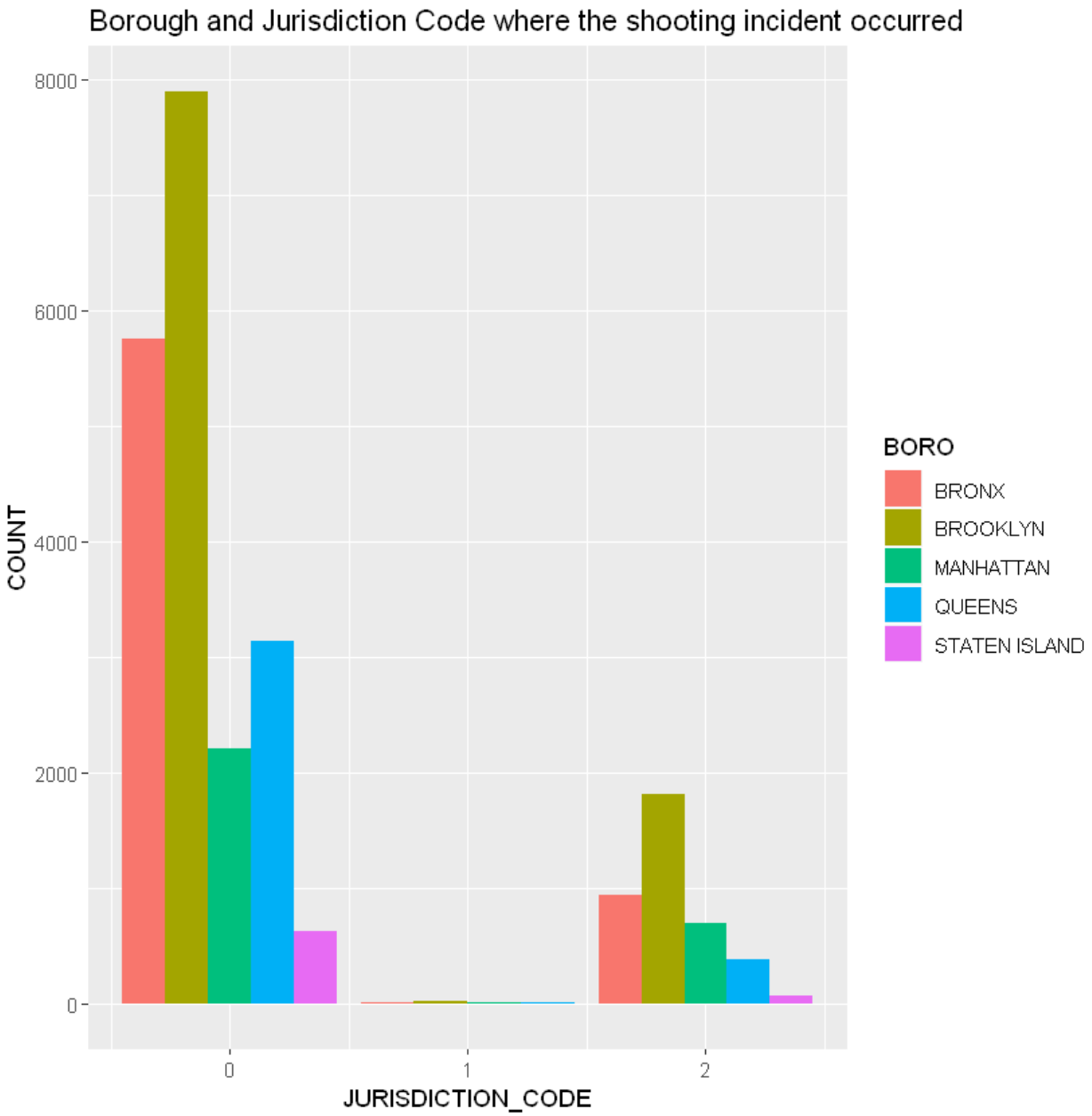
```
[1] "The number of shooting incident occurred in Borox is:  6699 ."
[1] "The number of shooting incident occurred in Queens is:  3531 ."
[1] "The number of shooting incident occurred in Brooklyn is:  9733 ."
[1] "The number of shooting incident occurred in Manhattan is:  2921 ."
[1] "The number of shooting incident occurred in Staten Island is:  696 ."
```

In [32]:
```
# create a table of counting cases were investigated by each jurisdiction_code at each
tbl1 <- nypd %>% group_by(JURISDICTION_CODE,BORO) %>% summarise(COUNT = n())
as.data.frame(tbl1)
```

JURISDICTION_CODE	BORO	COUNT
0	BRONX	5751
0	BROOKLYN	7895
0	MANHATTAN	2214
0	QUEENS	3145
0	STATEN ISLAND	622
1	BRONX	12
1	BROOKLYN	21
1	MANHATTAN	14
1	QUEENS	7
2	BRONX	936
2	BROOKLYN	1817
2	MANHATTAN	693
2	QUEENS	379
2	STATEN ISLAND	74

In [33]:
```
# plot the chart
```

```
ggplot(tbl1, aes(JURISDICTION_CODE, COUNT, fill = BORO))+ geom_col(position = "dodge")
    ggtitle("Borough and Jurisdiction Code where the shooting incident occurred")
```



As chart above, we can see that:

- The most shooting incidents were investigated by Patrol (jurisdiction_code = 0) and occurred in Brooklyn.
- Bronx is the second most place where the shooting incidents occurred.
- The place where the least shooting incident occurred is State Island.
- There are least of shooting incidents that were investigated by Transit (jurisdiction_code = 1).

Question 4: How many Perpetrator in each range of Race?

In [34]:
```
number_of_AMERICAN = nrow(filter(nypd, PERP_RACE == "AMERICAN INDIAN/ALASKAN NATIVE"))
number_of_ASIAN = nrow(filter(nypd, PERP_RACE == "ASIAN / PACIFIC ISLANDER"))
number_of_BLACK = nrow(filter(nypd, PERP_RACE == "BLACK"))
number_of_B_HIS = nrow(filter(nypd, PERP_RACE == "BLACK HISPANIC"))
```

```
number_of_WHITE = nrow(filter(nypd, PERP_RACE == "WHITE"))
number_of_W_HIS = nrow(filter(nypd, PERP_RACE == "WHITE HISPANIC"))
number_of_UNKN= nrow(filter(nypd, PERP_RACE == "UNKNOWN"))
```

In [35]:

```
print(paste("The number of American Indian / Alaskan Native Perpetrators is: ",number_o
print(paste("The number of Asian / Pacific Islander Perpetrators is: ",number_of_ASIAN,
print(paste("The number of Black Perpetrators is: ",number_of_BLACK,"."))
print(paste("The number of Black Hispanic Perpetrators is: ",number_of_B_HIS,"."))
print(paste("The number of White Perpetrators is: ",number_of_WHITE,"."))
print(paste("The number of White Hispanic Perpetrators is: ",number_of_W_HIS,"."))
print(paste("The number of Unknown Perpetrators is: ",number_of_UNKN,"."))
```

```
[1] "The number of American Indian / Alaskan Native Perpetrators is:  2 ."
[1] "The number of Asian / Pacific Islander Perpetrators is:  122 ."
[1] "The number of Black Perpetrators is:  10023 ."
[1] "The number of Black Hispanic Perpetrators is:  1096 ."
[1] "The number of White Perpetrators is:  255 ."
[1] "The number of White Hispanic Perpetrators is:  1985 ."
[1] "The number of Unknown Perpetrators is:  10097 ."
```

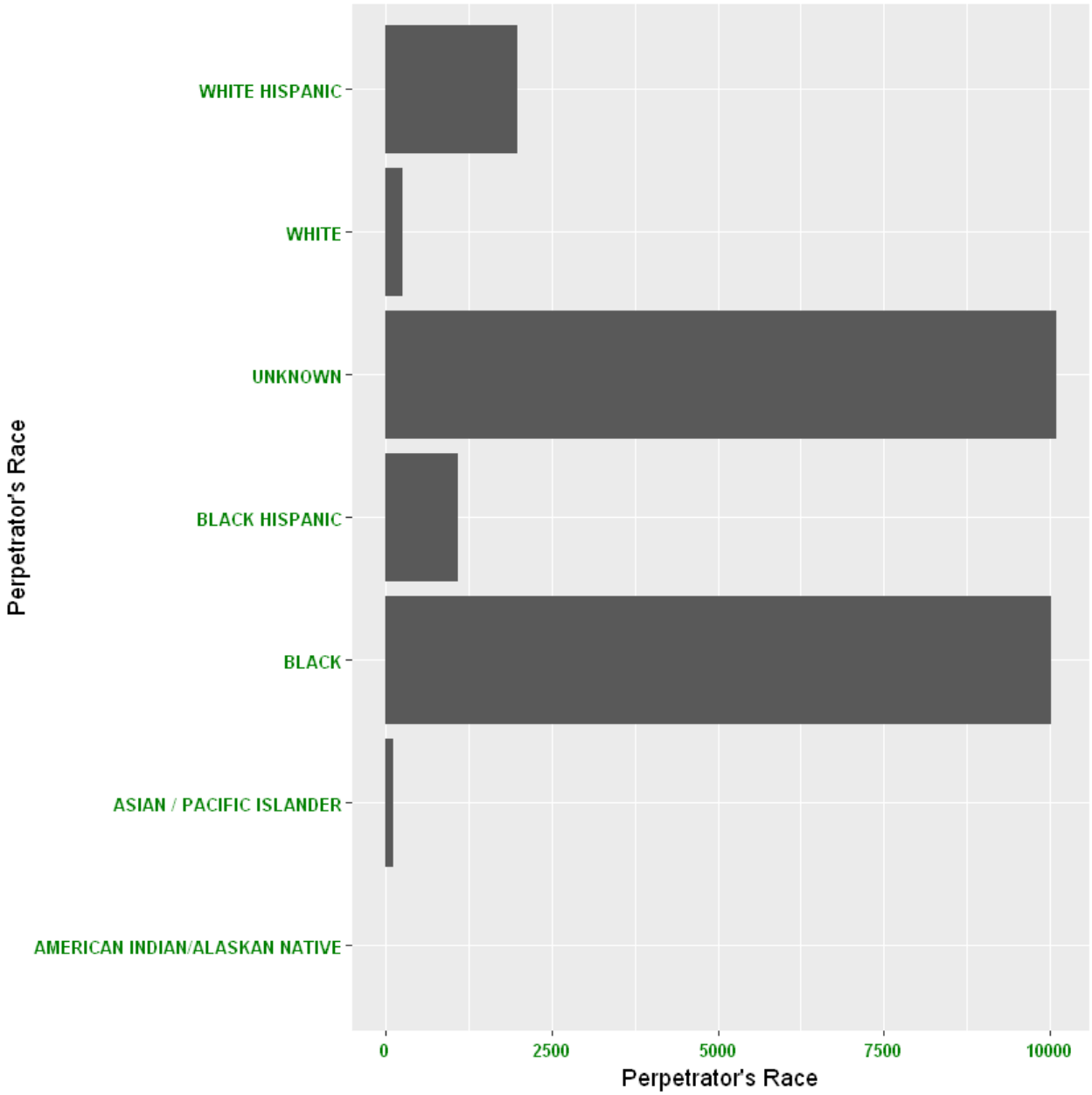
In [36]:

```
# create a dataframe y with the catagories of Perpetrator's Race and the number of each
y <- nypd %>% group_by(PERP_RACE) %>% summarise(COUNT = n())
as.data.frame(y)
```

PERP_RACE	COUNT
AMERICAN INDIAN/ALASKAN NATIVE	2
ASIAN / PACIFIC ISLANDER	122
BLACK	10023
BLACK HISPANIC	1096
UNKNOWN	10097
WHITE	255
WHITE HISPANIC	1985

In [37]:

```
# plot the bar chart for Perpetrator's Race
ggplot(y, aes(x = PERP_RACE, y = COUNT)) +
         geom_bar(stat = "identity") +
         coord_flip() + scale_y_continuous(name="Perpetrator's Race") +
  scale_x_discrete(name="Perpetrator's Race") +
theme(axis.text.x = element_text(face="bold", color="#008000",
                           size=8, angle=0),
          axis.text.y = element_text(face="bold", color="#008000",
                           size=8, angle=0))
```



As histogram above, we see that:

- There were still have many cases that lacked of information about Perpetrator's Race.
- Besides that, the most Perpetrator's Race is Black.
- The least Perpetrator's Race is American Indian / Alaskan Native.

Question 5: How many Perpetrator at each level of Age in different level of Sex?

In [38]:

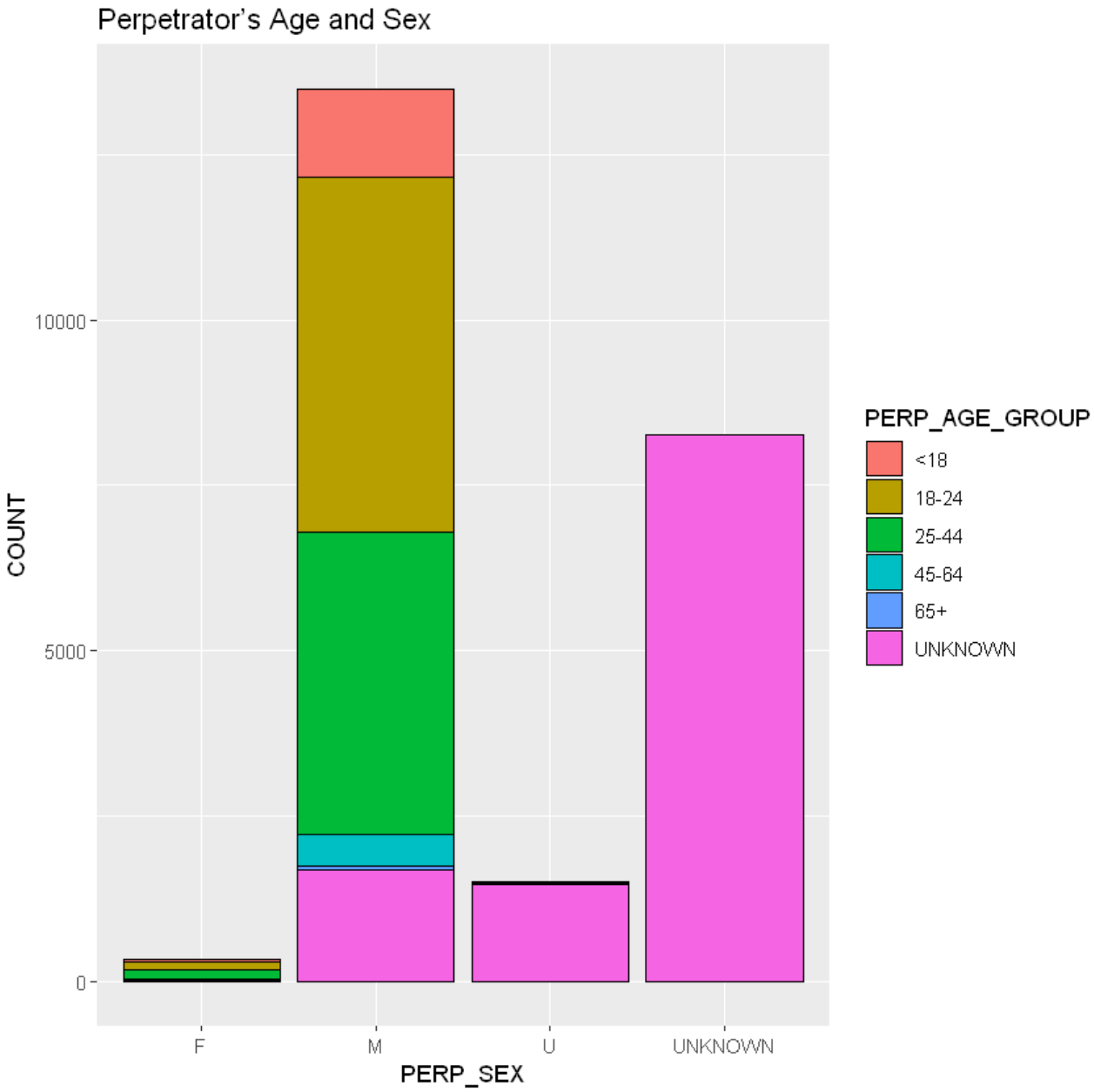
```
# create a table of counting cases with each level of Perpatrator's age and Sex
tbl2 <- nypd %>% group_by(PERP_SEX,PERP_AGE_GROUP) %>% summarise(COUNT = n())
as.data.frame(tbl2)
```

PERP_SEX	PERP_AGE_GROUP	COUNT
F	<18	34

PERP_SEX	PERP_AGE_GROUP	COUNT
F	18-24	126
F	25-44	139
F	45-64	17
F	65+	1
F	UNKNOWN	18
M	<18	1330
M	18-24	5366
M	25-44	4568
M	45-64	478
M	65+	53
M	UNKNOWN	1690
U	<18	3
U	18-24	16
U	25-44	7
U	UNKNOWN	1473
UNKNOWN	UNKNOWN	8261

In [39]:

```
# plot the chart
ggplot(data=tbl2, aes(x=PERP_SEX, y=COUNT, fill=PERP_AGE_GROUP)) +
     geom_bar(stat="identity", colour="black")+
     ggtitle("Perpetrator's Age and Sex")
```



The chart above tells us that:

- There are a lot of cases that were missing information about Perpetrator's sex and age yet.
- Perpetrator concentrated at age of 18-24 and 25-44.
- There are least Perpetrator at age of 65+.
- There are more male perpetrator than female perpetrator.

Question 6: What time did every shooting incident occur?

In [40]:

```
# create the data by hour
nypd_by_hour <- nypd %>%
  mutate(HOUR = hour(strptime(OCCUR_TIME, '%H')) %>% as.integer() ) %>%
  group_by(HOUR) %>%
  summarise(COUNT = n())%>%
  mutate(FREQ = round(COUNT / sum(COUNT), 4))
nypd_by_hour
```

HOUR	COUNT	FREQ
0	1908	0.0809
1	1864	0.0791
2	1620	0.0687
3	1464	0.0621
4	1291	0.0547
5	636	0.0270
6	301	0.0128
7	198	0.0084
8	190	0.0081
9	177	0.0075
10	248	0.0105
11	315	0.0134
12	415	0.0176
13	442	0.0187
14	685	0.0291
15	770	0.0327
16	874	0.0371
17	909	0.0385
18	1054	0.0447
19	1235	0.0524
20	1417	0.0601
21	1717	0.0728
22	1854	0.0786
23	1996	0.0846

In [41]:
```
max(nypd_by_hour$COUNT)
```

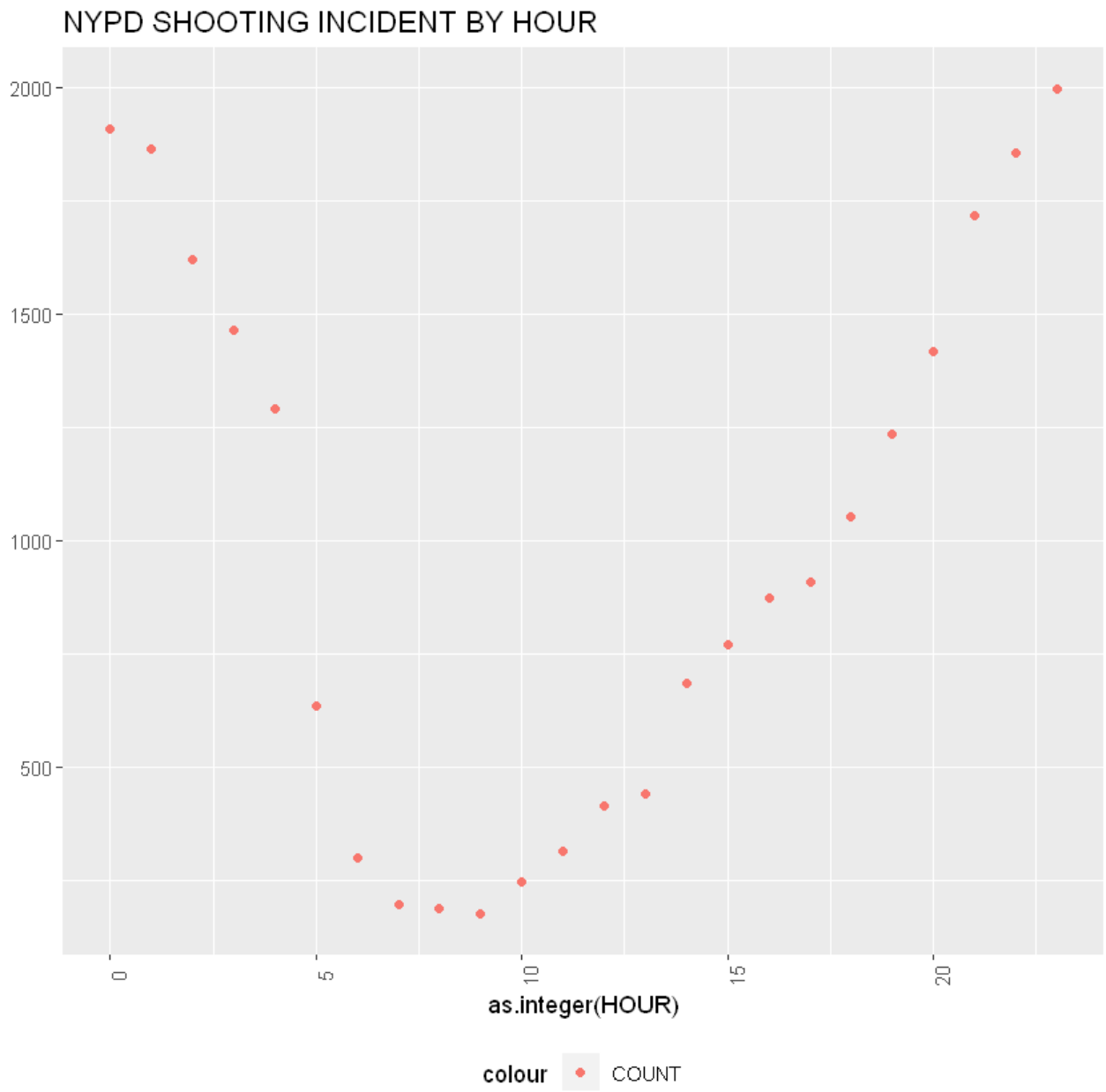
1996

In [42]:
```
min(nypd_by_hour$COUNT)
```

177

In [43]:
```
# plot the data by hour
nypd_by_hour %>%
  filter(COUNT > 0) %>%
  ggplot(aes(x = as.integer(HOUR), y = COUNT))+
  geom_point(aes(color="COUNT"))+
```

```
  theme(legend.position = "bottom",axis.text.x = element_text(angle = 90))+
  labs(title = "NYPD SHOOTING INCIDENT BY HOUR",y=NULL)
```



The plot tells us that the maximum count of shooting incident (1996) occurred at hour 23 and the minimum count of shooting incident (177) occurred at hour 9.

In [44]:

```
# Use the lm() function to perform a polinomial regression with count as the response
# and hour as the predictor.
# Use the summary() function to print the results
mod <- lm(COUNT ~ poly(HOUR, 2, raw=TRUE), data = nypd_by_hour)
summary(mod)
```

```
Call:
lm(formula = COUNT ~ poly(HOUR, 2, raw = TRUE), data = nypd_by_hour)

Residuals:
    Min      1Q  Median      3Q     Max 
-375.49 -124.21   47.35  163.44  284.87 
```

```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                1991.7738   117.1570   17.00 9.40e-14 ***
poly(HOUR, 2, raw = TRUE)1 -300.8056    23.5949  -12.75 2.37e-11 ***
poly(HOUR, 2, raw = TRUE)2   13.5985     0.9908   13.72 5.89e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 207.5 on 21 degrees of freedom
Multiple R-squared:  0.9015,    Adjusted R-squared:  0.8921
F-statistic:  96.1 on 2 and 21 DF,  p-value: 2.698e-11
```

Looking at the summary of this model, we can see that our p-value is very small, this means that the predictor were statistically significant in determining the Count. And the count of shooting incident is 1991.7738 - 300.8056 X hour + 13.5985 X $hour^2$.

In [45]:

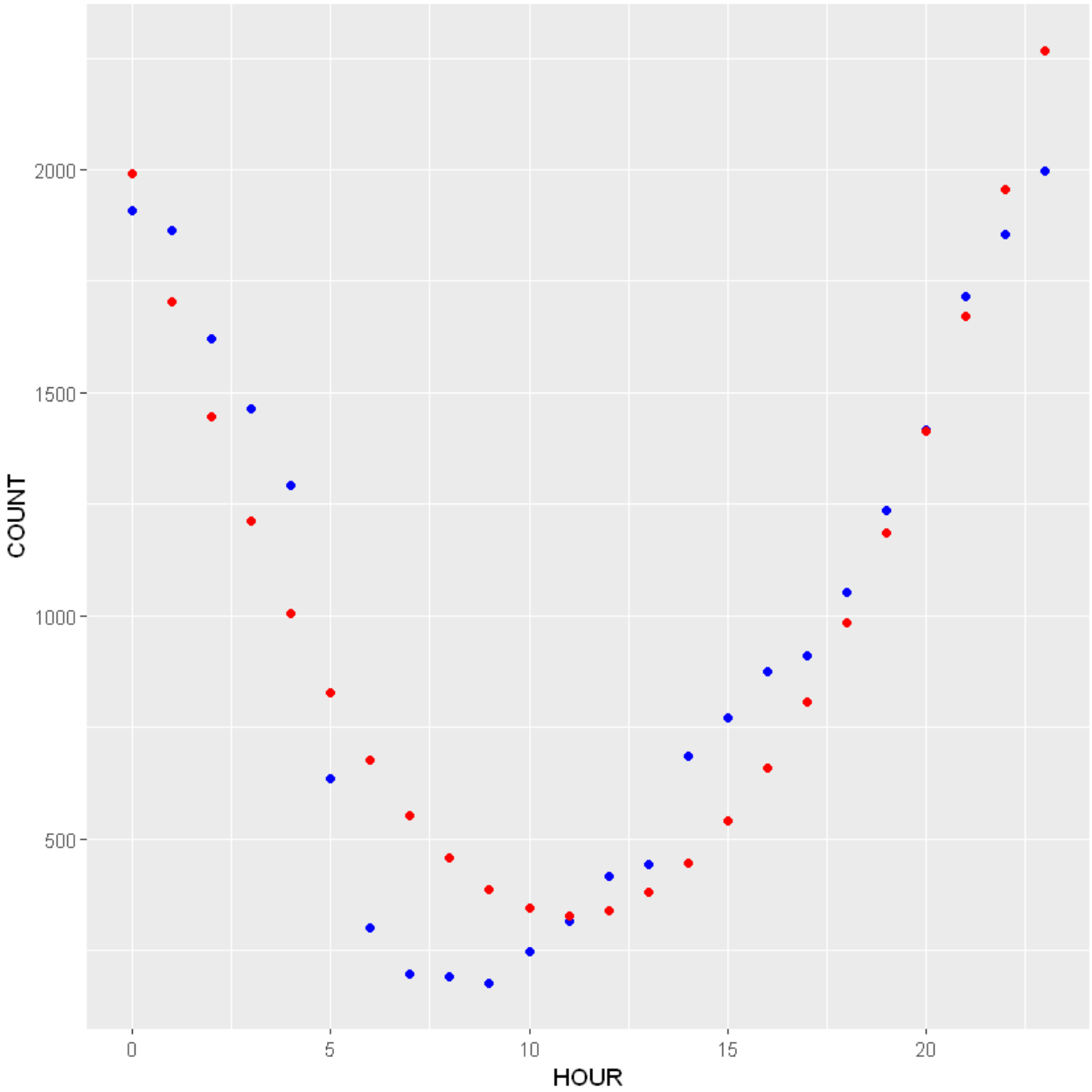
```
# create new data with predict the shooting incident by hour
nypd_by_hour_w_pred <- nypd_by_hour %>% mutate(PRED = round(predict(mod)))
nypd_by_hour_w_pred
```

HOUR	COUNT	FREQ	PRED
0	1908	0.0809	1992
1	1864	0.0791	1705
2	1620	0.0687	1445
3	1464	0.0621	1212
4	1291	0.0547	1006
5	636	0.0270	828
6	301	0.0128	676
7	198	0.0084	552
8	190	0.0081	456
9	177	0.0075	386
10	248	0.0105	344
11	315	0.0134	328
12	415	0.0176	340
13	442	0.0187	379
14	685	0.0291	446
15	770	0.0327	539
16	874	0.0371	660
17	909	0.0385	808
18	1054	0.0447	983
19	1235	0.0524	1186
20	1417	0.0601	1415

HOUR	COUNT	FREQ	PRED
21	1717	0.0728	1672
22	1854	0.0786	1956
23	1996	0.0846	2267

In [46]:

```
# plot the new data
nypd_by_hour_w_pred %>% ggplot() + geom_point(aes(x = HOUR, y = COUNT), color = "blue") +
  geom_point(aes(x = HOUR, y = PRED), color = "red")
```



In the plot above, our predictions are in red and our actuals are in blue. So we can see the model does a reasonably good job of predicting at the lower hour (0-5) and at the higher hour (17-23).

Step 4: Conclusion and add bias identification

In conclusion, base on NYPD Shooting Incident Data :

- First, while cleaning up the data, I recognized that there are three outliers in Perpetrator's age. They are 1020, 224 and 940. A common cause of bias is caused by data outliers that differ greatly from other samples. Outlier biases should be removed from the survey population to achieve a more accurate result. Hence, I deleted those three outliers out of the data.
- Second, there are two missing values of jurisdiction code and I solved this problem by deleting those 2 missing values.
- Third, there are a lot of missing values about age, sex and race of Perpetrators. Since this is a huge number, deleting the instances with missing observations can result in biased parameters and estimates and reduce the statistical power of the analysis. So I thought that I should not remove or delete them out of the data. And, to deal with this, I filled those missing values as "UNKNOWN". There are many missing values about location of the shooting incident as well and I filled them with "NONE".
- Next, base on the Victim's Sex chart and the Victim's Age plot, we can see that the most of victims are male and at age of 25-44, the least victims are unisex and at age of 65+.
- As "Borough and Jurisdiction Code where the shooting incident occurred" histogram, we see that the most shooting incidents were investigated by Patrol and occurred in Brooklyn. Bronx is the second most place where the shooting incidents occurred. The place where the least shooting incident occurred is State Island. And Transit investigated least of shooting incidents.
- Perpetrator' Rage histogram tells us that besides Unknown values, the most Perpetrator's Race is Black and the least Perpetrator's Race is American Indian / Alaskan Native.
- Perpetrator's Sex and Age chart shows that the most Perpetrator were at age of 18-24 and 25-44. The least Perpetrator at age of 65+. And there are more male perpetrator than female perpetrator.
- The maximum count of shooting incident (1996) occurred at 23 o'clock and the minimum count of shooting incident (177) occurred at 9 o'clock.