Week 3 Project

P.M.

2024-04-05

Step 1: Source information

Welcome to the report on the NYPD Shooting incident report. This report looks at all the shootings that have happened in New York from 2006 to 2022.

Question to Answer: What are some of the demographic trends we can see in the shootings?

Step 1: Source information

The first thing we did was to find a reliable source of data of the historic Shooting data for New York and so decided to use data from a US government Data site: "https://catalog.data.gov/dataset", particularly in the data set titled: NYPD Shooting Incident Data (Historic).

Note: We used the tidyverse package to tidy in our data

Importing Data:

We imported our data as a CSV file from this link: https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD

```
library(readr)
Shooting_Data_main <- read_csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=D
```

```
## Rows: 28562 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (12): OCCUR_DATE, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION...
## 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.
```

```
summary(Shooting_Data_main)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:28562       Length:28562       Length:28562
##  1st Qu.: 65439914   Class :character   Class1:hms         Class :character
##  Median : 92711254   Mode  :character   Class2:difftime    Mode  :character
```

```
##  Mean   :127405824                      Mode  :numeric
##  3rd Qu.:203131993
##  Max.   :279758069
##
##  LOC_OF_OCCUR_DESC     PRECINCT     JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:28562       Min.   :  1.0   Min.   :0.0000    Length:28562
##  Class :character   1st Qu.: 44.0   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 67.0   Median :0.0000    Mode  :character
##                     Mean   : 65.5   Mean   :0.3219
##                     3rd Qu.: 81.0   3rd Qu.:0.0000
##                     Max.   :123.0   Max.   :2.0000
##                                     NA's   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:28562       Mode :logical           Length:28562
##  Class :character   FALSE:23036             Class :character
##  Mode  :character   TRUE :5526              Mode  :character
##
##
##
##
##    PERP_SEX          PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  Length:28562       Length:28562       Length:28562       Length:28562
##  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:28562       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000068   1st Qu.:182912   1st Qu.:40.67
##  Mode  :character   Median :1007772   Median :194901   Median :40.70
##                     Mean   :1009424   Mean   :208380   Mean   :40.74
##                     3rd Qu.:1016807   3rd Qu.:239814   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA's   :59
##    Longitude        Lon_Lat
##  Min.   :-74.25   Length:28562
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :59
```

Step 2: Tidying the Data:

After importing the data, I read through it to see if it needs any tidying or if it's missing any information:

1. **Changing the date to a date object**:

```r
#installing needed libraries
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

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

```r
library(lubridate)
```

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

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

```r
Shooting_Data_main <- Shooting_Data_main %>% mutate(OCCUR_DATE = mdy(OCCUR_DATE))
```

2. **Removing columns we will not use**

   We view all the columns and identified 9 columns we won't need for our analysis, so we removed columns like, Jurisdiction_code, statistical_murder_flag, latitude as seen below:

   ```r
   to_remove <- c("JURISDICTION_CODE","STATISTICAL_MURDER_FLAG","Latitude","X_COORD_CD","Longitude","Y

   Shooting_Data_main <- Shooting_Data_main[,!(names(Shooting_Data_main) %in% to_remove)]
   ```

   We also filtered out columns with empty values like LOC_CLASSFCTN_DESC and LOC_OF_OCCUR_DESC.

3. Filtering out empty values such as NA and UNKOWN in the Race and age groups catergories:

   Removing NA values:

   ```r
   Shooting_Data_main <- na.omit(Shooting_Data_main)
   ```

   Removing all "UNKOWN" values

   The PERP_AGE_GROUP column had many unknown values so we filtered them out:

   ```r
   Shooting_Data_main <- Shooting_Data_main %>% filter(PERP_AGE_GROUP != "UNKNOWN")
   ```

   Missing and incorrect values

   At the end this is how the data looks:

```r
summary(Shooting_Data_main)
```

```
##   INCIDENT_KEY         OCCUR_DATE          OCCUR_TIME           BORO          
##  Min.   :  9953245   Min.   :2006-01-01   Length:8842       Length:8842       
##  1st Qu.: 71514503   1st Qu.:2010-02-27   Class1:hms        Class :character  
##  Median :150339140   Median :2016-02-12   Class2:difftime   Mode  :character  
##  Mean   :152522271   Mean   :2016-01-13   Mode  :numeric                      
##  3rd Qu.:246530273   3rd Qu.:2022-06-12                                       
##  Max.   :279758069   Max.   :2023-12-29                                       
##     PRECINCT      LOCATION_DESC      PERP_AGE_GROUP       PERP_SEX        
##  Min.   :  1.00   Length:8842        Length:8842        Length:8842       
##  1st Qu.: 43.00   Class :character   Class :character   Class :character  
##  Median : 67.00   Mode  :character   Mode  :character   Mode  :character  
##  Mean   : 64.18                                                           
##  3rd Qu.: 81.00                                                           
##  Max.   :123.00                                                           
##   PERP_RACE         VIC_AGE_GROUP        VIC_SEX            VIC_RACE        
##  Length:8842        Length:8842        Length:8842        Length:8842       
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
## 
```

For the rest of the missing values in some of the columns for example in perpetrator age groups there were some unknown values in some rows, as we go along in doing our analysis, we will check to see if there are any missing values and adjust appropriately.

Step 3: Analyzing the Data:

After cleaning and importing the data we will now analyze the data, particularly looking at the demographic statistics within the data.

1. Looking at the Boroughs affected over the years:

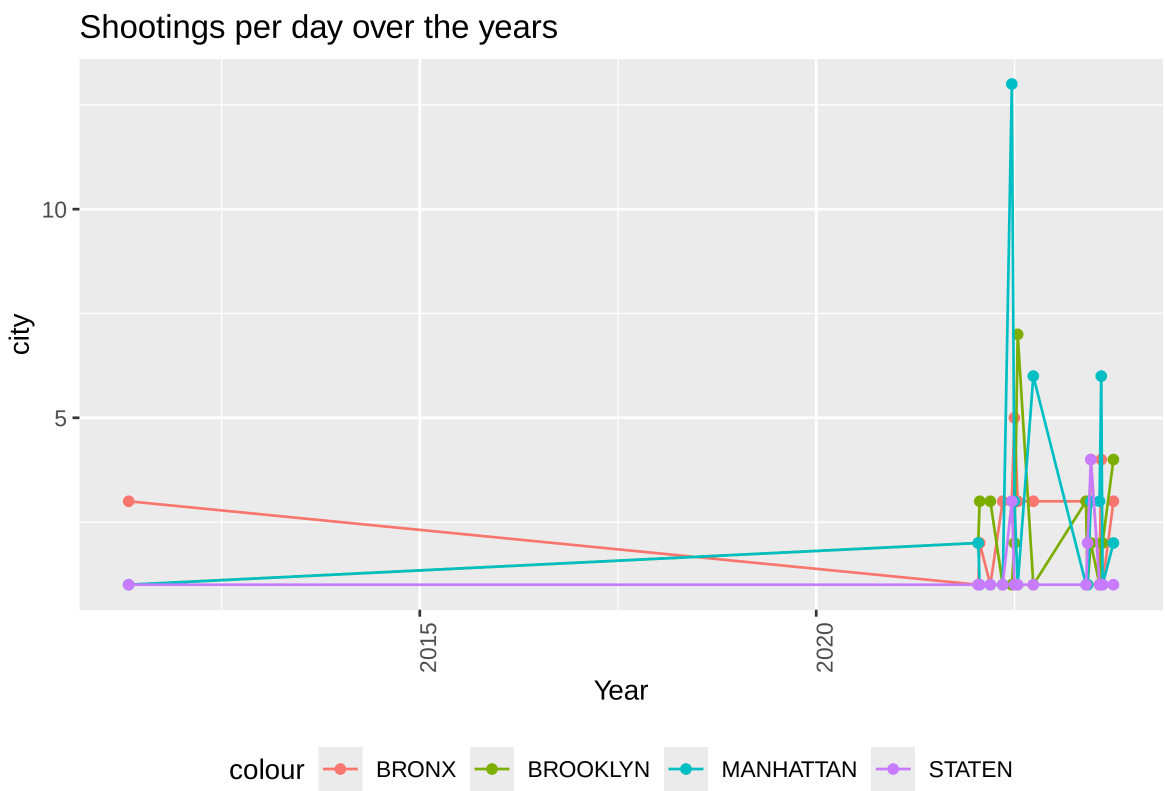
- First we start by looking at the total number of shootings in the different boroughs to see the highest and lowest:

```r
Boro_Sum <- Shooting_Data_main %>% count(BORO)
print(Boro_Sum)
```

```
## # A tibble: 5 x 2
##   BORO              n
##   <chr>         <int>
## 1 BRONX          2661
## 2 BROOKLYN       3201
## 3 MANHATTAN      1378
## 4 QUEENS         1286
## 5 STATEN ISLAND   316
```

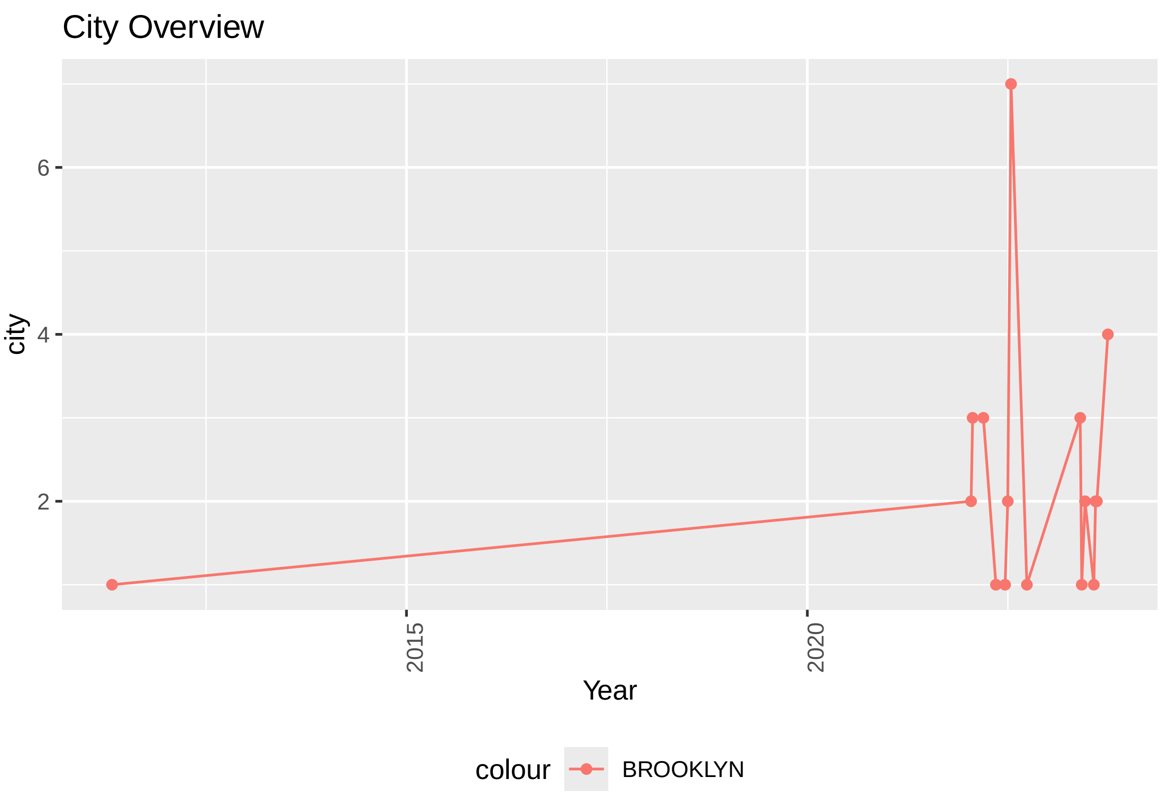
from this we see Brooklyn as the highest and State Island as the lowest, we then did an analysis of the number of deaths between 2006 to 2022 in the graph below:

```r
library(ggplot2)
# Organising the data by Borough and by the date
Bronx <- Shooting_Data_main %>% filter(BORO == "BRONX")
Bronx_date <- Bronx %>% count(OCCUR_DATE)
Bronx_date <- Bronx_date %>% rename(BRONX = n)

BROOKLYN <- Shooting_Data_main %>% filter(BORO == "BROOKLYN")
BROOKLYN_date <- BROOKLYN %>% count(OCCUR_DATE)
BROOKLYN_date <- BROOKLYN_date %>% rename(BROOKLYN = n)

MANHATTAN <- Shooting_Data_main %>% filter(BORO == "MANHATTAN")
MANHATTAN_date <- MANHATTAN %>% count(OCCUR_DATE)
MANHATTAN_date <- MANHATTAN_date %>% rename(MANHATTAN = n)

QUEENS <- Shooting_Data_main %>% filter(BORO == "QUEENS")
QUEENS_date <- QUEENS %>% count(OCCUR_DATE)
QUEENS_date <- QUEENS_date %>% rename(QUEENS = n)

STATEN_ISLAND <- Shooting_Data_main %>% filter(BORO == "STATEN ISLAND")
STATEN_ISLAND_date <- STATEN_ISLAND %>% count(OCCUR_DATE)
STATEN_ISLAND_date <- STATEN_ISLAND_date %>% rename(STATEN = n)

#Merging them into one Dataframe and the number of deaths for each Borough at each date:
BOROS <- merge(merge(merge(Bronx_date,BROOKLYN_date, by = "OCCUR_DATE"), MANHATTAN_date, by = "OCCU

#Plotting it on a graph

ggplot(BOROS, aes(x = OCCUR_DATE)) + geom_line(aes(y = BRONX, color="BRONX")) + geom_point(aes(y =
```



- From the graph of shootings per day above we see that Staten Island has been relatively constant over the years but had a high growth from around 2022. We also see Bronx has had a steady decline over the years. We see also that 2022 has had high growth for all Boroughs. Below we investigate Brooklyn a little further as it had the highest shootings across all Boroughs:

```
ggplot(BOROS, aes(x = OCCUR_DATE)) + geom_line(aes(y=BROOKLYN, color="BROOKLYN"))+ geom_point(aes(y
```



From above we see that Brooklyn has had a stead rise over the years. In 2022, its been unstable dropping and rising.

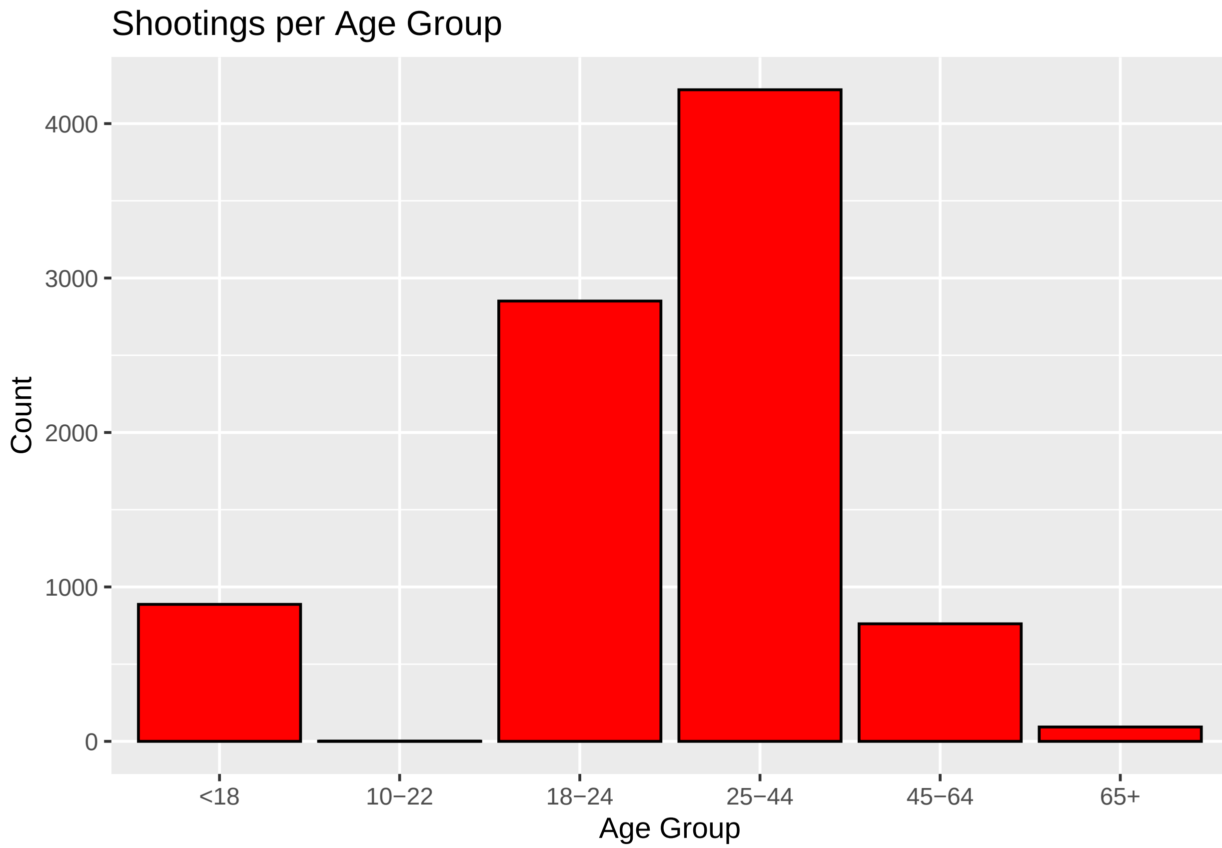
Additional Questions:

1. One question raised would be the population between cities, is the high deaths in Brooklyn a result of a higher population or higher crime rate?
2. Another would be what caused the spike in 2022 and above, would it be a new government policy, or the immigration crisis, or something else?

2. Age Groups We also looked at what age groups are the victims and the perpetrators of the crime.

We started of with the **victims** as seen below in the graph:

```
#Organising by age groups
histgm <- Shooting_Data_main %>% count(VIC_AGE_GROUP)
histgm[1,1] <- "10-22"
histgm1 <- histgm[-7,]
ggplot(histgm1, aes(x=VIC_AGE_GROUP,y=n)) + geom_bar(stat="identity",fill = "red",color = "black")+ labs
```



We also did some cleaning of the data, removing "UNKNOWN" values and editing the age group text from "**1022**" to "**10-22**".

Analysis: From the graph above we see that the largest groups are between 18-24, and and 25-44. This makes sense in that these are the most active groups in any society.

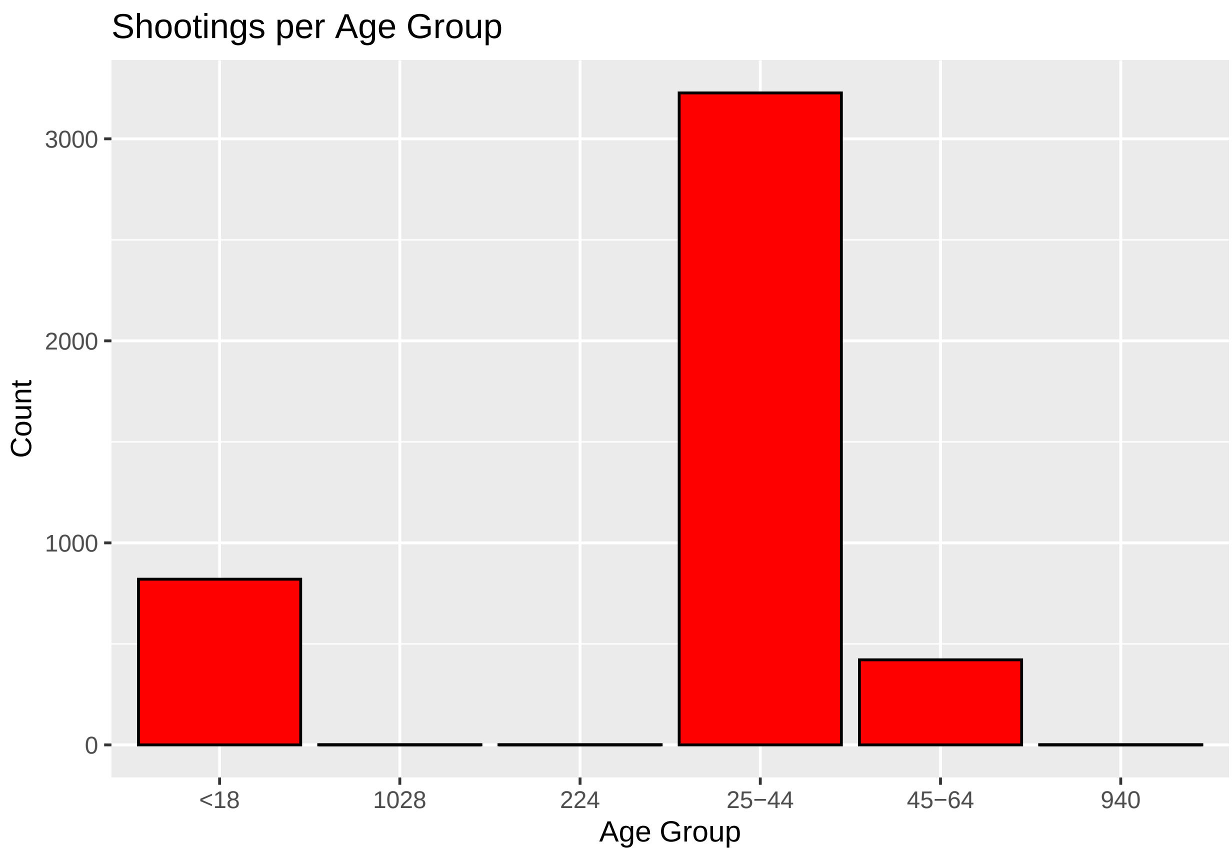
We also looked at the **prepetrators** age group:

We started by first cleaning the data, removing Null values and wrong values that don't fit in any age group:

```
#Organising by perp age groups
prepbar <- Shooting_Data_main %>% count(PERP_AGE_GROUP)
print(prepbar)
```

```
## # A tibble: 10 x 2
##    PERP_AGE_GROUP     n
##    <chr>          <int>
##  1 (null)          1141
##  2 1020               1
##  3 1028               1
##  4 18-24           3182
##  5 224                1
##  6 25-44           3227
##  7 45-64            421
##  8 65+               47
##  9 940                1
## 10 <18              820
```

```
prepbar <- prepbar[-c(1,2,4,8),]
ggplot(prepbar, aes(x=PERP_AGE_GROUP,y=n)) + geom_bar(stat="identity",fill = "red",color = "black")+ lal
```



Over here we see a larger amount between 18-24, and and 25-44 as before but this time a higher number from the 18-24 age range, which despite being the smallest in terms of years part (that is from 18 to 24 is just 6 years) has the highest number of crime.

Additional Questions raised:

1. What is the population between the different age groups from the victims to the perpetrator. This might raise question as to whether the rates of crime committed at the different age groups is a result of population and not other factors such as economic status and so forth.
2. What was the reason for the shooting, was it a robbery related crime, a gang violence related crime, domestic or civil case. This would help in understanding what factors cause shootings the most.

3. Gender We also look at the gender demographics of the shooters and the victims.

We first started with the **perpetrators**:

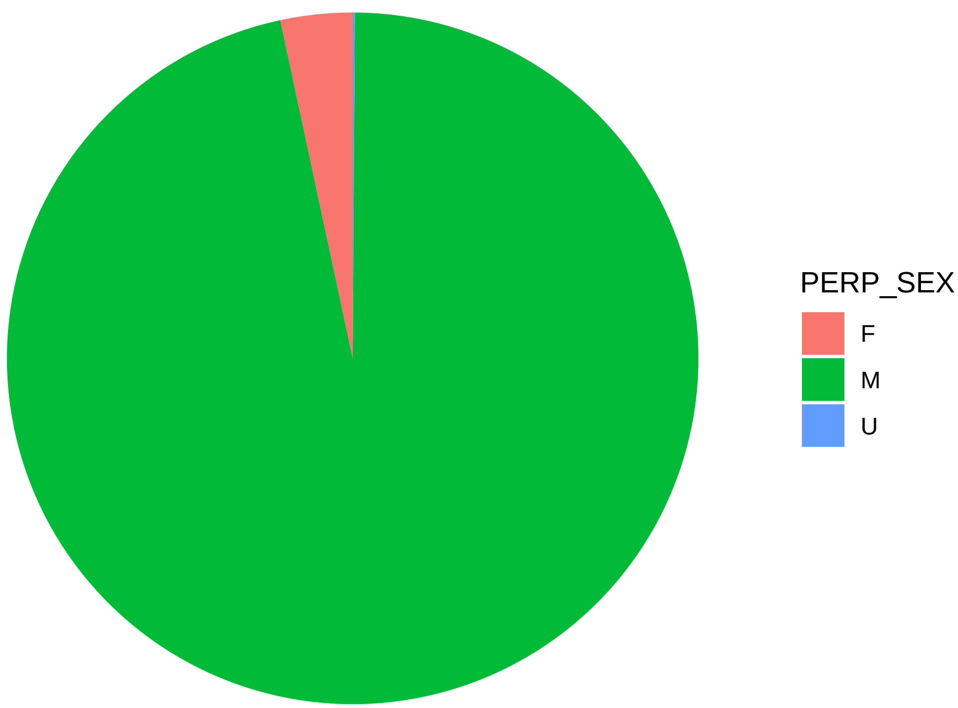
```
race_pie <- Shooting_Data_main %>% count(PERP_SEX)
print(race_pie)
```

```
## # A tibble: 4 x 2
##   PERP_SEX     n
```

```
##   <chr>    <int>
## 1 (null)    1141
## 2 F          259
## 3 M         7434
## 4 U            8
```

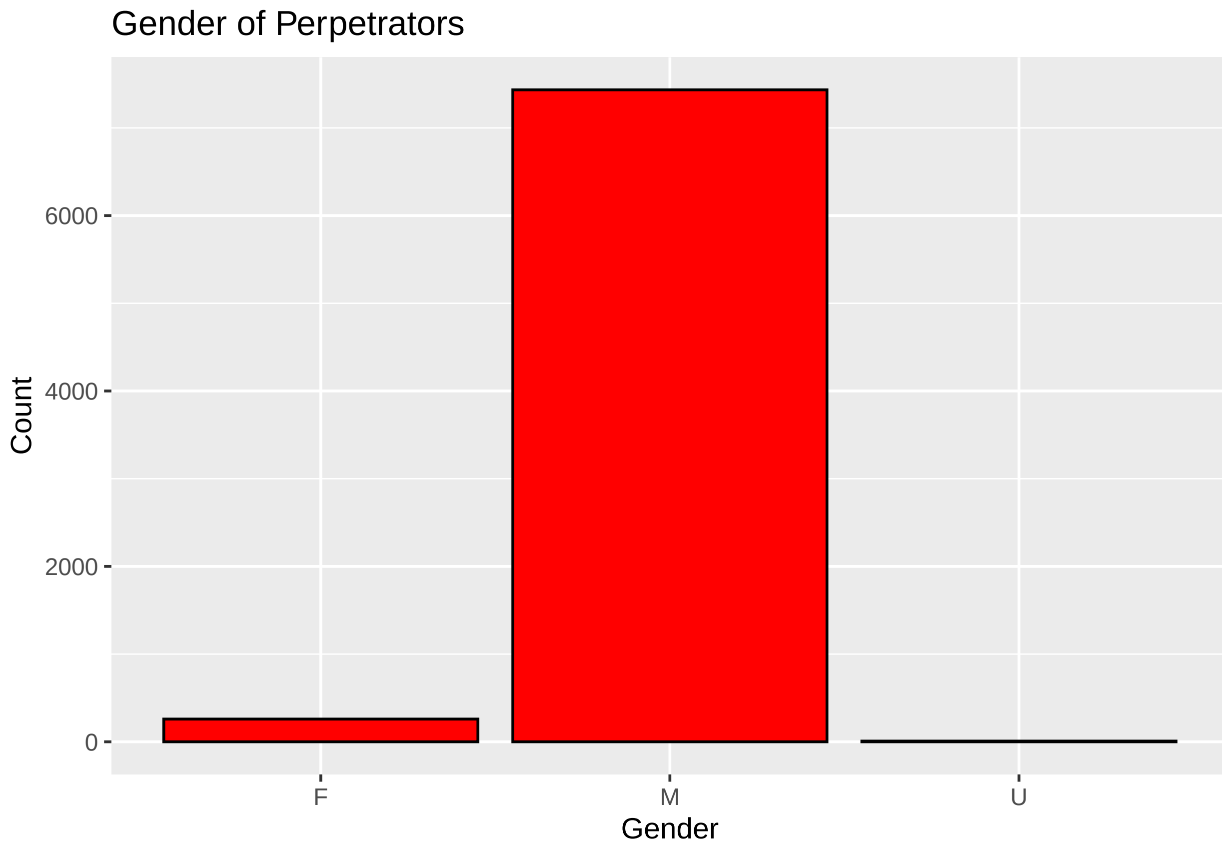
After this we clean the data to remove null values:

```r
#removing empty/null values
race_pie <- race_pie[-1,]
ggplot(race_pie,aes(x = "", y = n, fill = PERP_SEX))+ geom_bar(stat="identity",width = 1) + coord_polar
```



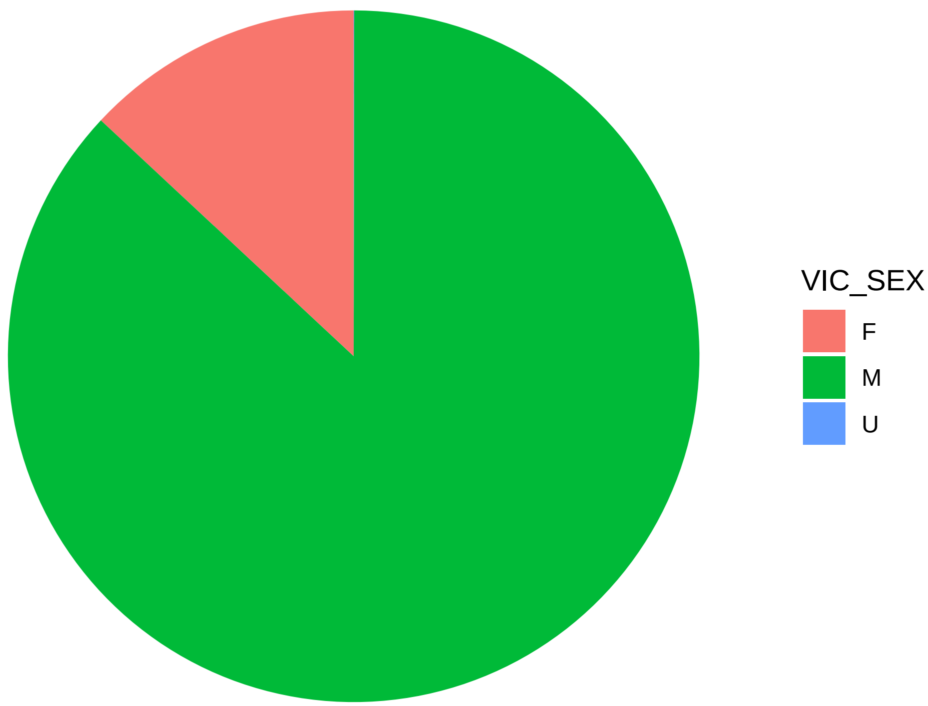
From here we can see there was a high ratio of male perpetrators (96%). We see it more in detail below in a bar graph that there were over 6000 shootings by male perpetrators and less than 500 for the rest:

```r
ggplot(race_pie, aes(x=PERP_SEX,y=n)) + geom_bar(stat="identity",fill = "red",color = "black")+ labs(x=
```



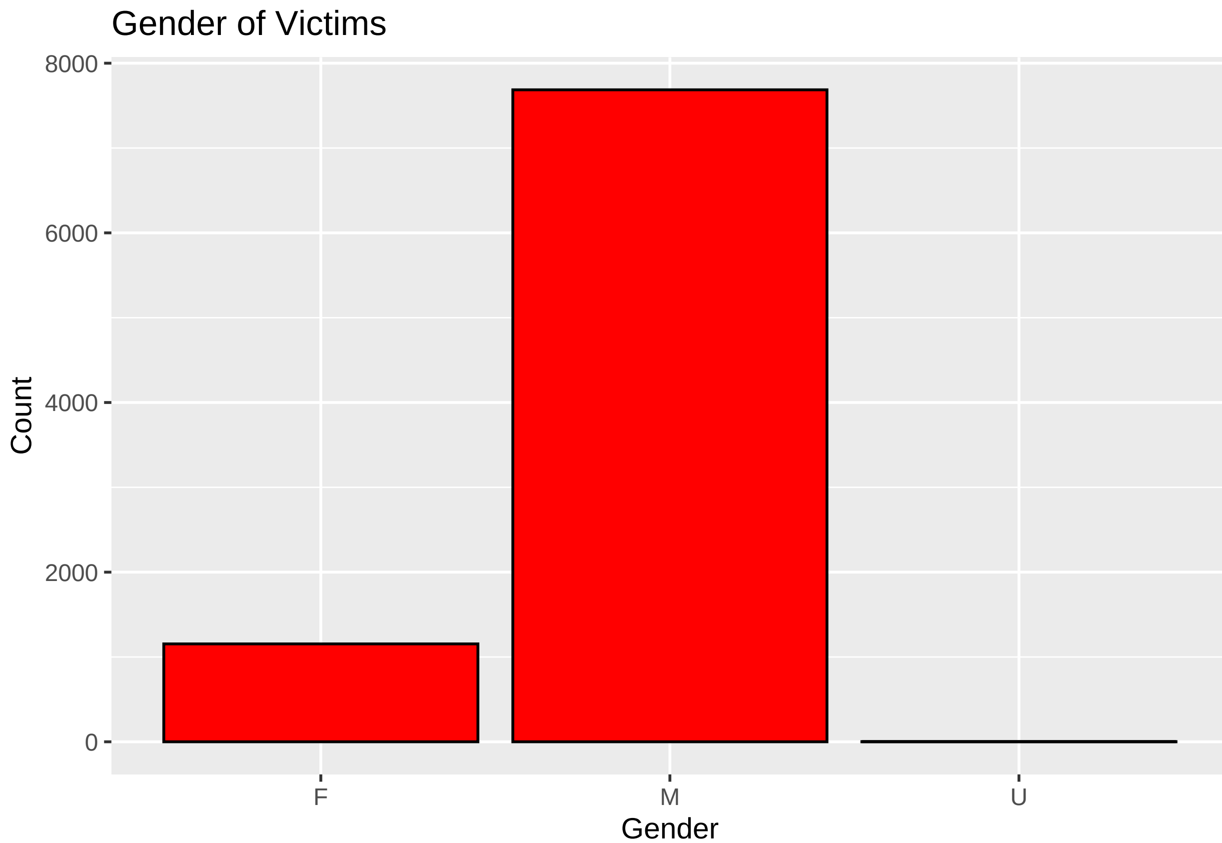
Then looking at the **victims** as shown below:

```
race_pie_vic <- Shooting_Data_main %>% count(VIC_SEX)
ggplot(race_pie_vic,aes(x = "", y = n, fill = VIC_SEX))+ geom_bar(stat="identity",width = 1) + coord_pol
```



From here we see similar trend of having a significantly higher number of male vitctims, however it's lesser here by by about 10% (86%) than for perpetrators. We see it more in detail below:

```
ggplot(race_pie_vic, aes(x=VIC_SEX,y=n)) + geom_bar(stat="identity",fill = "red",color = "black")+ labs
```



From here we see that male victims were over 6000 but female victims were just about 1000. This raises more questions:

1. What was the reason/cause of the shooting, were the high male to male shootings a result of gang violence or other causes?
2. What is the population distribution of males, is this the cause of the high numbers?

4. Predicting Model of Age Group VS Race I also looked at the relation between race and age group for perpetrators, to understand if it race was a factor that affected the distribution of shootings between age groups.

To start it of, we first looked at shootings from between 2006 to 2011 (5 year period), looking at the percentages between the age groups and we used this to train our model. We then looked test the model with data from 2011 to 2016 (5 year period) to see if our model accurately represented the data.

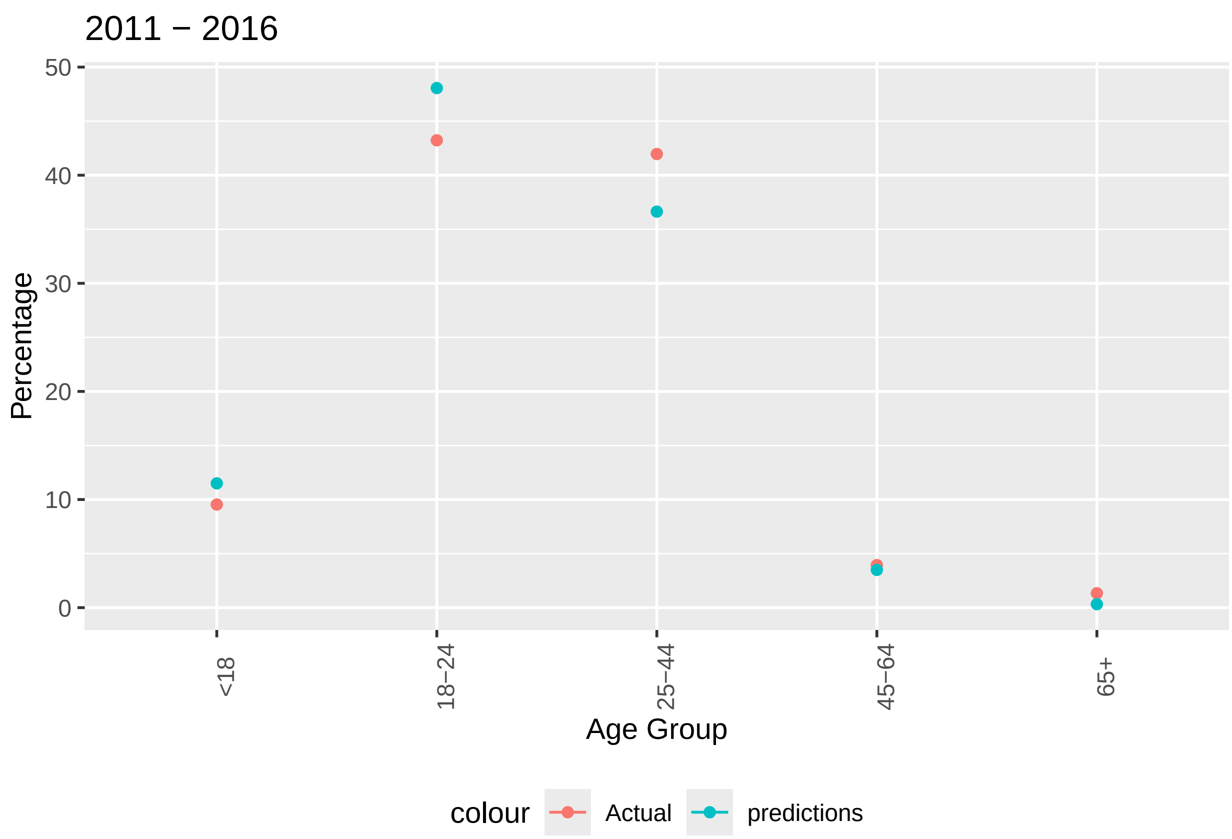
```
mod_used <- Shooting_Data_main %>% filter(year(OCCUR_DATE) <= 2011) %>% count(PERP_AGE_GROUP)
print(mod_used)
```

```
## # A tibble: 6 x 2
##   PERP_AGE_GROUP     n
##   <chr>          <int>
## 1 18-24           1501
## 2 224                1
## 3 25-44           1144
## 4 45-64            109
```

```
## 5 65+                10
## 6 <18               359
```

```
#Cleaning it to remove unwanted data:
mod_used <- mod_used[-2,]

#Getting the percentage distribution of the different age groups:
total_count <- sum(mod_used$n)
mod_used <- mod_used %>% mutate(percentage = n / total_count *100)

#making a prediction of the data:
preditions <- lm(percentage ~ PERP_AGE_GROUP, data=mod_used)

#data we will use to test our model:
test_mod <- Shooting_Data_main %>% filter(year(OCCUR_DATE) > 2011 & year(OCCUR_DATE) <= 2016) %>% count

#cleaning up the data and getting percentage ratios:
test_mod <- test_mod[-c(1,6),]
total_count <- sum(test_mod$n)
test_mod <- test_mod %>% mutate(percentage = n / total_count *100)

#predicting the data:
test_mod$predictions <- predict(preditions, newdata = test_mod)

ggplot(test_mod, aes(x = PERP_AGE_GROUP)) + geom_line(aes(y=percentage, color="Actual"))+ geom_point(aes
```

```
## `geom_line()`: Each group consists of only one observation.
## i Do you need to adjust the group aesthetic?
## `geom_line()`: Each group consists of only one observation.
## i Do you need to adjust the group aesthetic?
```

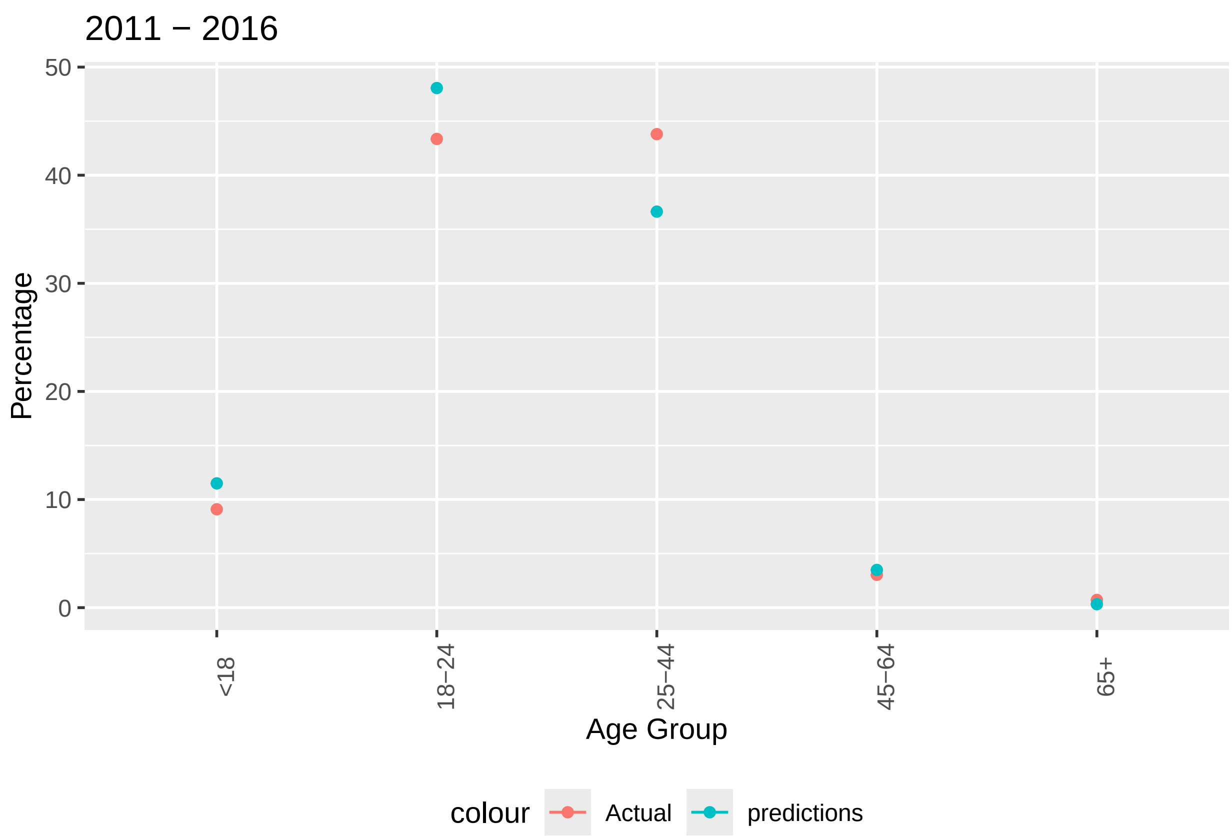


From below, we see that our predicted values closely imitate the actual values. now to test our theory, we looked at data from the 2 races with the highest shootings, that is Black and White Hispanic.

Starting with **Black**

```
Black <- test_mod <- Shooting_Data_main %>% filter(year(OCCUR_DATE) > 2011 & year(OCCUR_DATE) <= 2016) %

Black <- Black[-1,]
total_count <- sum(Black$n)
Black <- Black %>% mutate(percentage = n / total_count *100)
Black$predictions <- predict(preditions, newdata = Black)
ggplot(Black, aes(x = PERP_AGE_GROUP)) + geom_line(aes(y=percentage, color="Actual"))+ geom_point(aes(y=
```

```
## ‘geom_line()‘: Each group consists of only one observation.
## i Do you need to adjust the group aesthetic?
## ‘geom_line()‘: Each group consists of only one observation.
## i Do you need to adjust the group aesthetic?
```

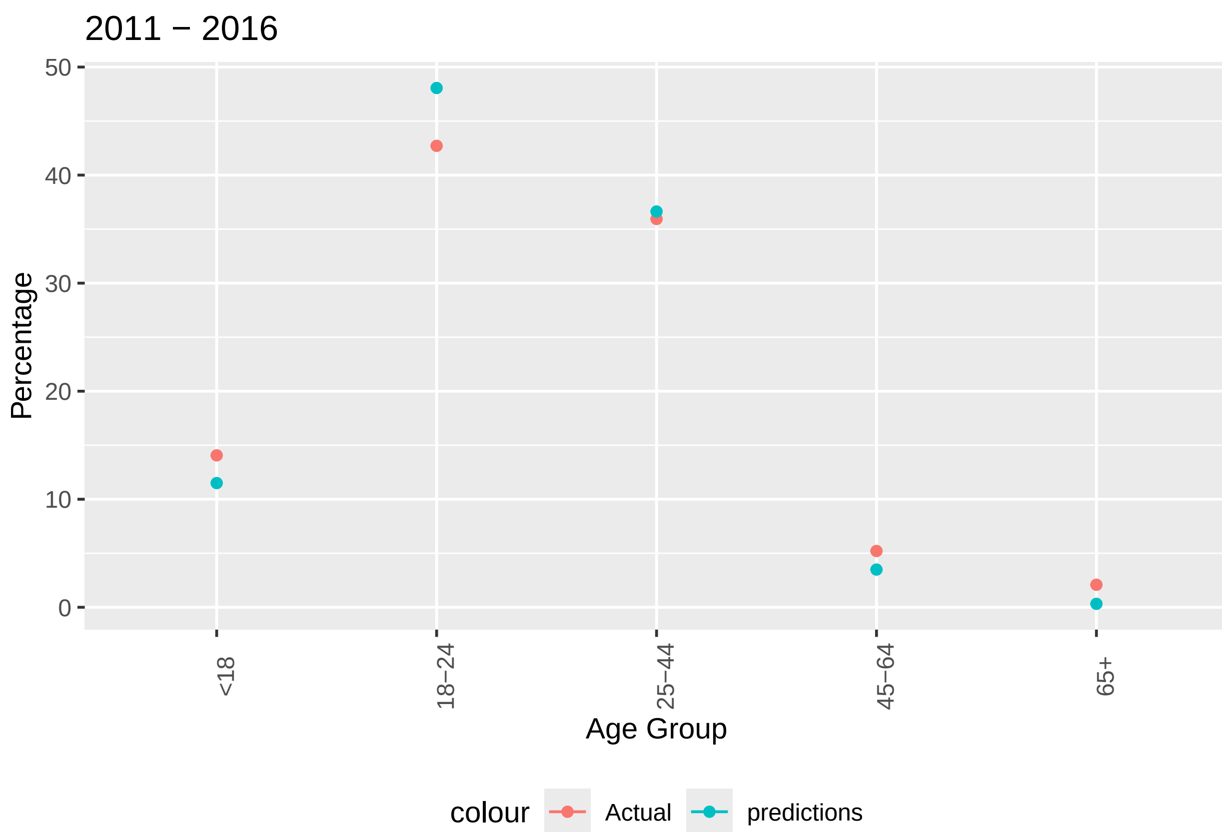


We can see it's following a similar pattern , however in this case our predictions are sometimes lower and sometimes higher than the actual values, but nothing significant showing a difference here.

I did the same for **White-Hispanic**

```
White <- Shooting_Data_main %>% filter(year(OCCUR_DATE) > 2011 & year(OCCUR_DATE) <= 2016) %>% filter(PE
White <- White[-5,]
total_count <- sum(White$n)
White <- White %>% mutate(percentage = n / total_count *100)
White$predictions <- predict(preditions, newdata = White)

ggplot(White, aes(x = PERP_AGE_GROUP)) + geom_line(aes(y=percentage, color="Actual"))+ geom_point(aes(y=
```

```
## 'geom_line()': Each group consists of only one observation.
## i Do you need to adjust the group aesthetic?
## 'geom_line()': Each group consists of only one observation.
## i Do you need to adjust the group aesthetic?
```



Here we see a similar trend where the predictions are not too far off from the actual values. This therefore concludes that race does not affect age group.

5. Conclusion and possible bias: In conclusion as a summary we first got our data from a US government website and after cleaning it up, we looked at 4 areas:

1. which Boroughs are affected the most and how have the murders increased over the years. We determined that Brooklyn had the highest murders and Staten Island had the lowest. We also that found the for the most part the shooting incidents had seen a significant growth around the year 2022.
2. Which age groups are affected the most both victim and perpetrator. we found that between the ages of 25-44 had the highest number of victims and 18-24 had the highest number of perpetrators
3. We also looked at the gender distribution of the shootings and found that there was a significantly high number of male perpetrators and male victims.
4. We also looked at the possibility, if there was a link between race and age group and developed a model to predict what was most likely the outcome of each race. it was concluded that race didn't affect the distribution of age group substantially.

Possible Sources of Bias: From my own knowledge and understanding of New York and especially some of the challenges facing youths I already came into the research expecting a certain criteria and demographic of perpetrators of gun violence. However to mitigate this, I looked at the overall age groups to understand what were their numbers and let the data speak for itself and only analyzed it from the insights I was getting from the data and not what I already now.

Another source of Bias is personal interest, particularly in the analysis, I was interested in just knowing one aspect, which was the race divisions of the shootings, however to mitigate this, I looked at many aspects other than just race, looked at location, gender and age group to have a balance of insights.