DataScienceWeek3Assignment

Megan Arnold

Import Data

```
url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
nypdData <- read.csv(url,header=TRUE,sep=",")
```

Initial Analysis

The imported data comes from the NYPD database of shooting incidences from 2006 until present. In the data, it includes the data and time, location (Precinct, jurisdiction, longitude/latitude, borough), the demographics (Race, Age Group, Sex) of the perpetrator and victim, and whether or not the victim was murdered. This seems to be a good source for violent crimes involving a firearm.

Using this data, I'm going to determine if the location (borough) has any correlation to the age group of the perpetrator. This information could allow for a better understanding of the age groups that are committing the shootings. There are many things that could be done with this information (additional job training, anger management help, gun control, etc). Special consideration must be taking to prevent predictive policing causing additional racial injustices.

```
summary(nypdData)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME            BORO
##  Min.   :  9953245   Length:25596       Length:25596       Length:25596
##  1st Qu.: 61593633   Class :character   Class :character   Class :character
##  Median : 86437258   Mode  :character   Mode  :character   Mode  :character
##  Mean   :112382648
##  3rd Qu.:166660833
##  Max.   :238490103
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.0000    Length:25596       Length:25596
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character   Class :character
##  Median : 69.00   Median :0.0000    Mode  :character   Mode  :character
##  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
##
```

Cleaning Data for Analysis

```
columns = c("BORO","PERP_AGE_GROUP","PERP_SEX")
df = select(nypdData,all_of(columns))

#Data tabulated
table(df$BORO,df$PERP_AGE_GROUP)
```

```
##
##                      <18 1020 18-24  224 25-44 45-64  65+  940 UNKNOWN
##   BRONX          2512  473    1  1847    1  1529   182    8    0     849
##   BROOKLYN       4291  556    0  2107    0  1852   176   23    1    1359
##   MANHATTAN      1030  224    0   776    0   769    63    7    0     396
##   QUEENS         1366  159    0   864    0   838    89   13    0     499
##   STATEN ISLAND  145   51    0   250    0   214    25    6    0      45
```

Table Analysis

The above table is grouped by the BORO column and displaying the summary of the PERP_AGE_GROUP. There is a significant portion of missing data. I plan of keeping that information and displaying it as unknown in the charts. This will help to determine if there are causes for the unknown data or if it is randomly unknown.

Also, there are two data points where the age value is “1020”, “940”, and “224”. Since these are errors in the data, I will remove those values before I begin my analysis.

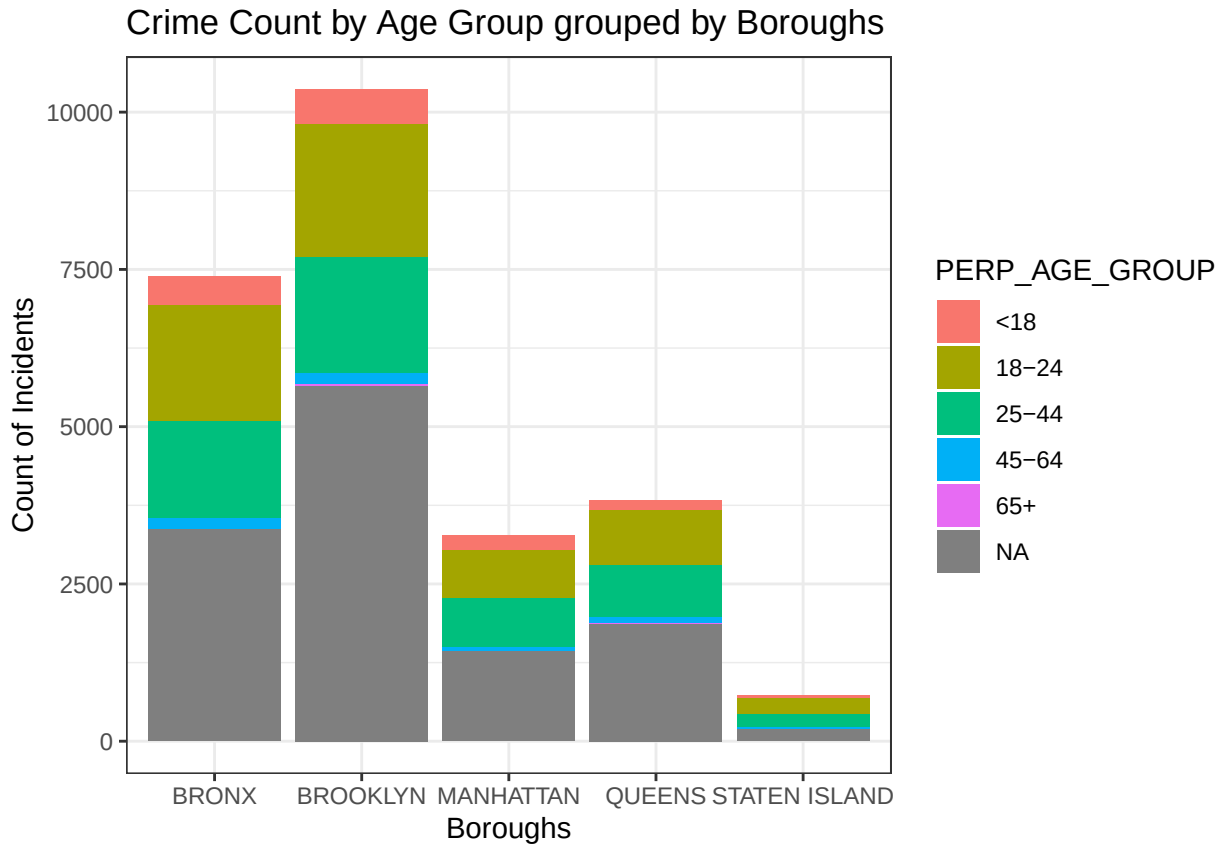
Also, I will combine the values that were labeled as blank or “unknown” as NA

```
df = df[!(df$PERP_AGE_GROUP==224 | df$PERP_AGE_GROUP==940 | df$PERP_AGE_GROUP==1020),] #Remove the inco
df$PERP_AGE_GROUP[(df$PERP_AGE_GROUP=="" | df$PERP_AGE_GROUP=="UNKNOWN")] <-NA #Make blank and unknown
```

Visualizations

Visualization 1

```
ggplot(data = df, aes(x = BORO, fill = PERP_AGE_GROUP)) +
  geom_bar() +
  ggtitle("Crime Count by Age Group grouped by Boroughs") +
  xlab("Boroughs") +
  ylab("Count of Incidents") +
  theme_bw()
```



Visualization 1 Analysis

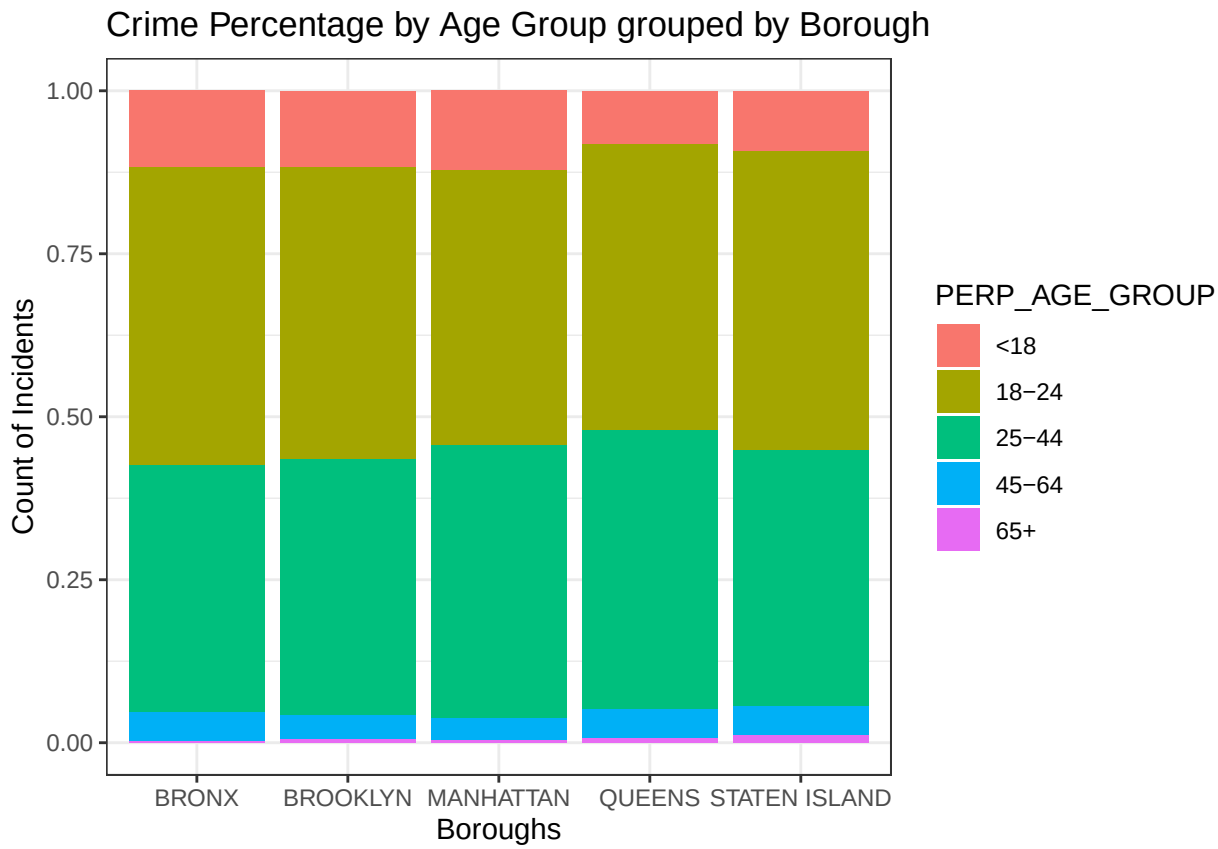
Looking at the above data, it appears that the distribution in crimes per age group is similar amongst the boroughs. However, the unknown data is skewing the results. I'm going to create stacked percentage chart without the NA data.

Visualization 2

```
dfKnown = df[!(is.na(df$PERP_AGE_GROUP)),]
table(df$BORO,df$PERP_AGE_GROUP)
```

```
##
##                 <18 18-24 25-44 45-64  65+
##   BRONX          473  1847  1529   182    8
##   BROOKLYN       556  2107  1852   176   23
##   MANHATTAN      224   776   769    63    7
##   QUEENS         159   864   838    89   13
##   STATEN ISLAND   51   250   214    25    6
```

```
ggplot(data = dfKnown, aes(x = BORO, fill = PERP_AGE_GROUP)) +
  geom_bar(position = "fill") +
  ggtitle("Crime Percentage by Age Group grouped by Borough") +
  xlab("Boroughs") +
  ylab("Count of Incidents")+
  theme_bw()
```



Visualization 2 Analysis

As we can see in the Percentile chart, it appears that the borough where the crime was committed is independent of the age group. A chi squared test for independence will tell us our p-value.

Additional analysis can be done on based on the rate of crime per 100,000 people in the borough. Also, the crime rate within the age group could be determined. This information could help make predictions as age demographics shift with time.

Model/Analysis

Since this data is categorized data and something like a linear model isn't valid, I'm going to do a statistical analysis to determine if my initial assumption based on the visualization was correct. I will do a chi-squared test for independence to determine if the age of the perpetrator is independent of the borough.

```
chisqTest = chisq.test(table(df$BORO,df$PERP_AGE_GROUP))
```

```
## Warning in chisq.test(table(df$BORO, df$PERP_AGE_GROUP)): Chi-squared
## approximation may be incorrect
```

```
print(chisqTest)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(df$BORO, df$PERP_AGE_GROUP)
## X-squared = 56.582, df = 16, p-value = 1.95e-06
```

```
print(chisq.residuals(table(df$BORO,df$PERP_AGE_GROUP),std=FALSE, raw=FALSE))
```

```
## Warning in stats::chisq.test(tab): Chi-squared approximation may be incorrect
```

```
##
##                    <18 18-24 25-44 45-64   65+
##   BRONX          1.03  1.07 -1.87  1.33 -2.28
##   BROOKLYN       1.29  0.09 -0.46 -1.19  0.55
##   MANHATTAN      1.30 -1.55  1.44 -1.40 -0.35
##   QUEENS        -4.07 -0.39  2.10  0.99  1.53
##   STATEN ISLAND -1.28  0.41 -0.19  0.57  2.35
```

Model Analysis

In the above test, it shows a p value of 0.00000195. Looking at the residuals of the test, we can see that the "Queens, <18" jumps out as a large residual.

Note: I duplicated the calculation outside of R due to the warning: "Chi-squared approximation may be incorrect". The p value was the same.

Bias

A potential cause of bias in the data could be caused by the patrol patterns of the police. If they suspect a certain age group to be committing crimes, then it is possible they would patrol the area more frequently; thus, catch them committing the crimes with an increased frequency compared to the areas they were not patrolling.

Also, there could be a source of bias from who reports the crimes. If there is a non random cause for people to not report crimes, then that cause could show up in the data.

There was also a significant amount of unknown data in the "Age Group" category. If the cause of that missing data isn't random, then that would create a bias in the data source.