Final Project Part 1

11/16/2021

Get Current Data

```
url_in <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
```

Read in the Data Set

```
NYPD_Shooting_incident_data <- read_csv(url_in)
```

```
## Parsed with column specification:
## cols(
##   INCIDENT_KEY = col_double(),
##   OCCUR_DATE = col_character(),
##   OCCUR_TIME = col_time(format = ""),
##   BORO = col_character(),
##   PRECINCT = col_double(),
##   JURISDICTION_CODE = col_double(),
##   LOCATION_DESC = col_character(),
##   STATISTICAL_MURDER_FLAG = col_logical(),
##   PERP_AGE_GROUP = col_character(),
##   PERP_SEX = col_character(),
##   PERP_RACE = col_character(),
##   VIC_AGE_GROUP = col_character(),
##   VIC_SEX = col_character(),
##   VIC_RACE = col_character(),
##   X_COORD_CD = col_number(),
##   Y_COORD_CD = col_number(),
##   Latitude = col_double(),
##   Longitude = col_double(),
##   Lon_Lat = col_character()
## )
```

Tidy the Data Set and show the summary

I drop all the columns related to pinpoint location data, and also the location description column. There were too many NA's in the location description column, so the data would not give us very accurate insight if used.

```
Shooting_incident <- NYPD_Shooting_incident_data %>%
  select(-c(Lon_Lat, X_COORD_CD, Y_COORD_CD, Longitude, Latitude, LOCATION_DESC)) %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE)) %>%
  mutate(OCCUR_TIME = hour(hms(OCCUR_TIME)))

Shooting_incident$OCCUR_TIME <- factor(Shooting_incident$OCCUR_TIME)
Shooting_incident$PERP_SEX <- factor(Shooting_incident$PERP_SEX)
Shooting_incident$PERP_AGE_GROUP <- factor(Shooting_incident$PERP_AGE_GROUP)
Shooting_incident$PERP_RACE <- factor(Shooting_incident$PERP_RACE)

Shooting_incident$BORO <- factor(Shooting_incident$BORO)
Shooting_incident$PRECINCT <- factor(Shooting_incident$PRECINCT)
Shooting_incident$JURISDICTION_CODE <- factor(Shooting_incident$JURISDICTION_CODE)

Shooting_incident$VIC_SEX <- factor(Shooting_incident$VIC_SEX)
Shooting_incident$VIC_AGE_GROUP <- factor(Shooting_incident$VIC_AGE_GROUP)
Shooting_incident$VIC_RACE <- factor(Shooting_incident$VIC_RACE)

summary(Shooting_incident)
```

```
##   INCIDENT_KEY         OCCUR_DATE           OCCUR_TIME               BORO
##  Min.   :  9953245   Min.   :2006-01-01   23     : 1994   BRONX        :6700
##  1st Qu.: 55317014   1st Qu.:2008-12-30   0      : 1908   BROOKLYN     :9722
##  Median : 83365370   Median :2012-02-26   1      : 1864   MANHATTAN    :2921
##  Mean   :102218616   Mean   :2012-10-03   22     : 1854   QUEENS       :3527
##  3rd Qu.:150772442   3rd Qu.:2016-02-28   21     : 1708   STATEN ISLAND: 698
##  Max.   :222473262   Max.   :2020-12-31   2      : 1622
##                                           (Other):12618
##     PRECINCT     JURISDICTION_CODE STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  75     : 1367   0   :19624        Mode :logical           18-24  :5448
##  73     : 1282   1   :   54        FALSE:19080             25-44  :4613
##  67     : 1102   2   : 3888        TRUE :4488              UNKNOWN:3156
##  79     :  920   NA's:    2                                <18    :1354
##  44     :  842                                            45-64  : 481
##  47     :  815                                            (Other):  57
##  (Other):17240                                            NA's   :8459
##  PERP_SEX              PERP_RACE    VIC_AGE_GROUP   VIC_SEX
##  F   :  334   BLACK         :9855   <18    : 2525   F: 2195
##  M   :13305   WHITE HISPANIC:1961   18-24  : 9000   M:21353
##  U   : 1504   UNKNOWN       :1869   25-44  :10287   U:   20
##  NA's: 8425   BLACK HISPANIC:1081   45-64  : 1536
##               WHITE         : 255   65+    :  155
##               (Other)       : 122   UNKNOWN:   65
##               NA's          :8425
##                            VIC_RACE
##  AMERICAN INDIAN/ALASKAN NATIVE:    9
##  ASIAN / PACIFIC ISLANDER      :  320
##  BLACK                         :16846
##  BLACK HISPANIC                : 2244
##  UNKNOWN                       :  102
##  WHITE                         :  615
##  WHITE HISPANIC                : 3432
```

I noticed there were data entry errors in the PERP_AGE_GROUP column, so I will remove the 3 anomalies to preserve the good data.

Column before cleaning:

```
summary(Shooting_incident$PERP_AGE_GROUP)
```

```
##     <18    1020   18-24     224   25-44   45-64     65+     940 UNKNOWN    NA's
##    1354       1    5448       1    4613     481      54       1    3156    8459
```

Column after cleaning:

```
Shooting_incident_cleaned <- Shooting_incident %>%
  slice (-c(1407, 19669, 2915))

summary(Shooting_incident_cleaned$PERP_AGE_GROUP)
```

```
##     <18    1020   18-24     224   25-44   45-64     65+     940 UNKNOWN    NA's
##    1354       0    5448       0    4613     481      54       0    3156    8459
```

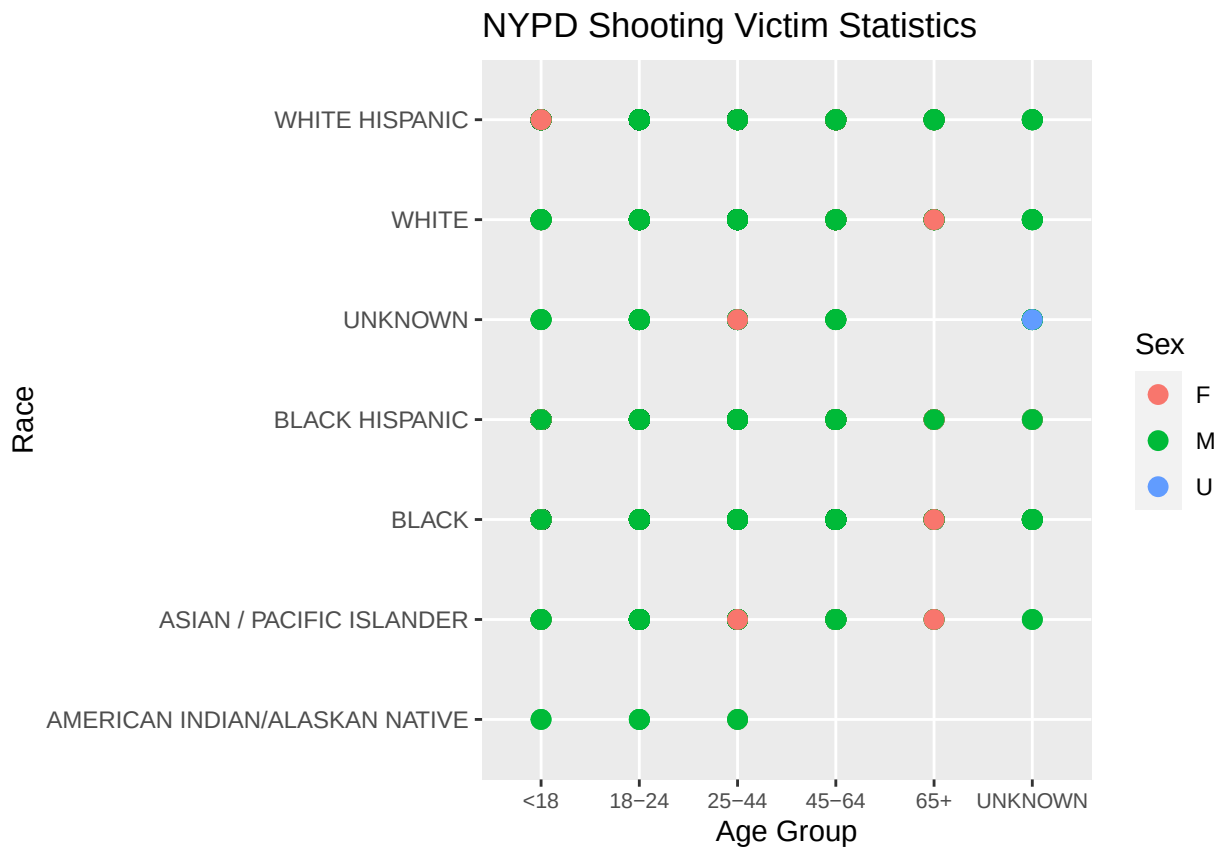
Explanation of missing values:

For missing values under PERP_AGE_GROUP, PERP_SEX, and PERP_RACE I will be dropping the rows that are missing the fields when using the columns as factors in order not to create unintentional bias.

Visualizations

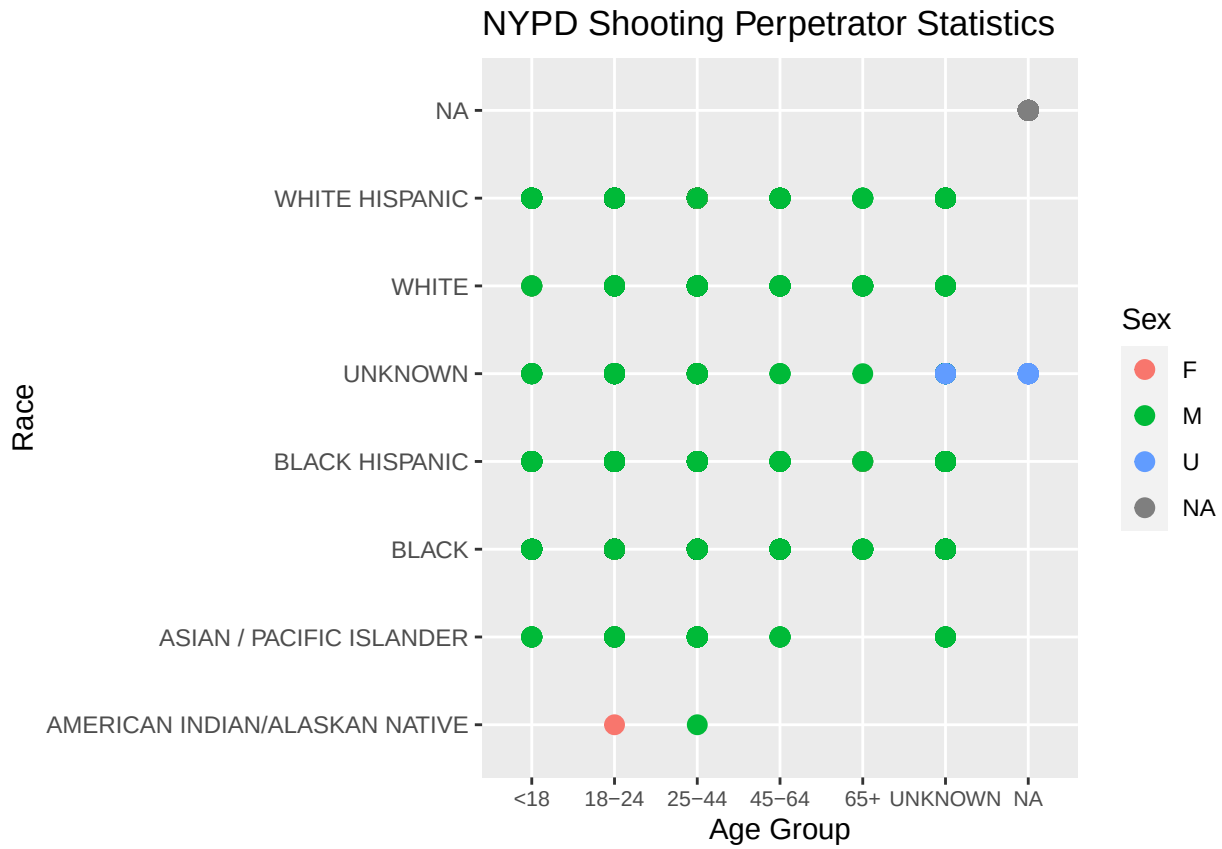
To start with, I will create a graph plotting which sex occurs more often based on race and age group.

```
ggplot(data = Shooting_incident_cleaned) +
  geom_point(size = 3, mapping = aes(x = VIC_AGE_GROUP, y = VIC_RACE, color = VIC_SEX)) +
  theme(axis.text.x = element_text(size = 8)) +
  ggtitle("NYPD Shooting Victim Statistics") +
  labs(x = "Age Group", y = "Race", color = "Sex")
```



I will now do the same for the shooting perpetrator group.

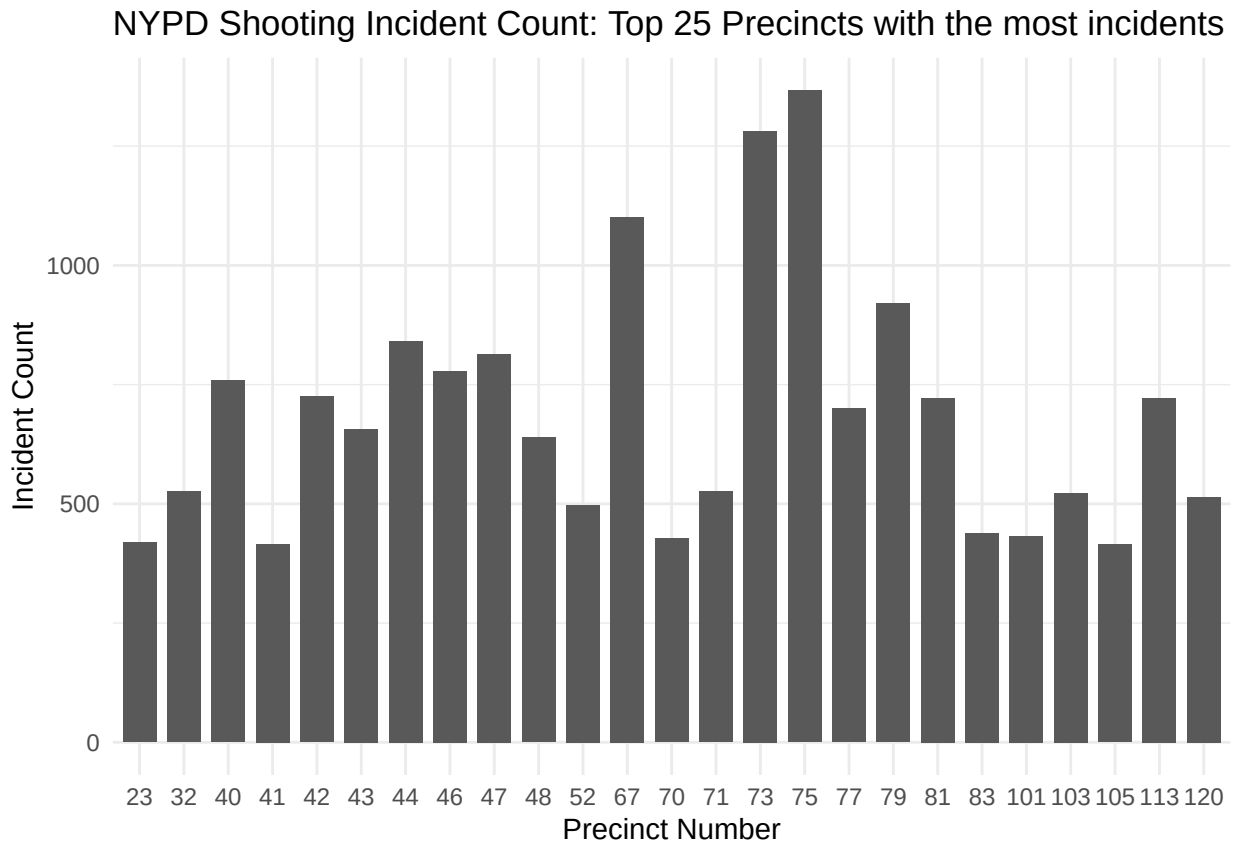
```
ggplot(data = Shooting_incident_cleaned) +
  geom_point(size = 3, mapping = aes(x = PERP_AGE_GROUP, y = PERP_RACE, color = PERP_SEX)) +
  theme(axis.text.x = element_text(size = 8)) +
  ggtitle("NYPD Shooting Perpetrator Statistics") +
  labs(x = "Age Group", y = "Race", color = "Sex")
```



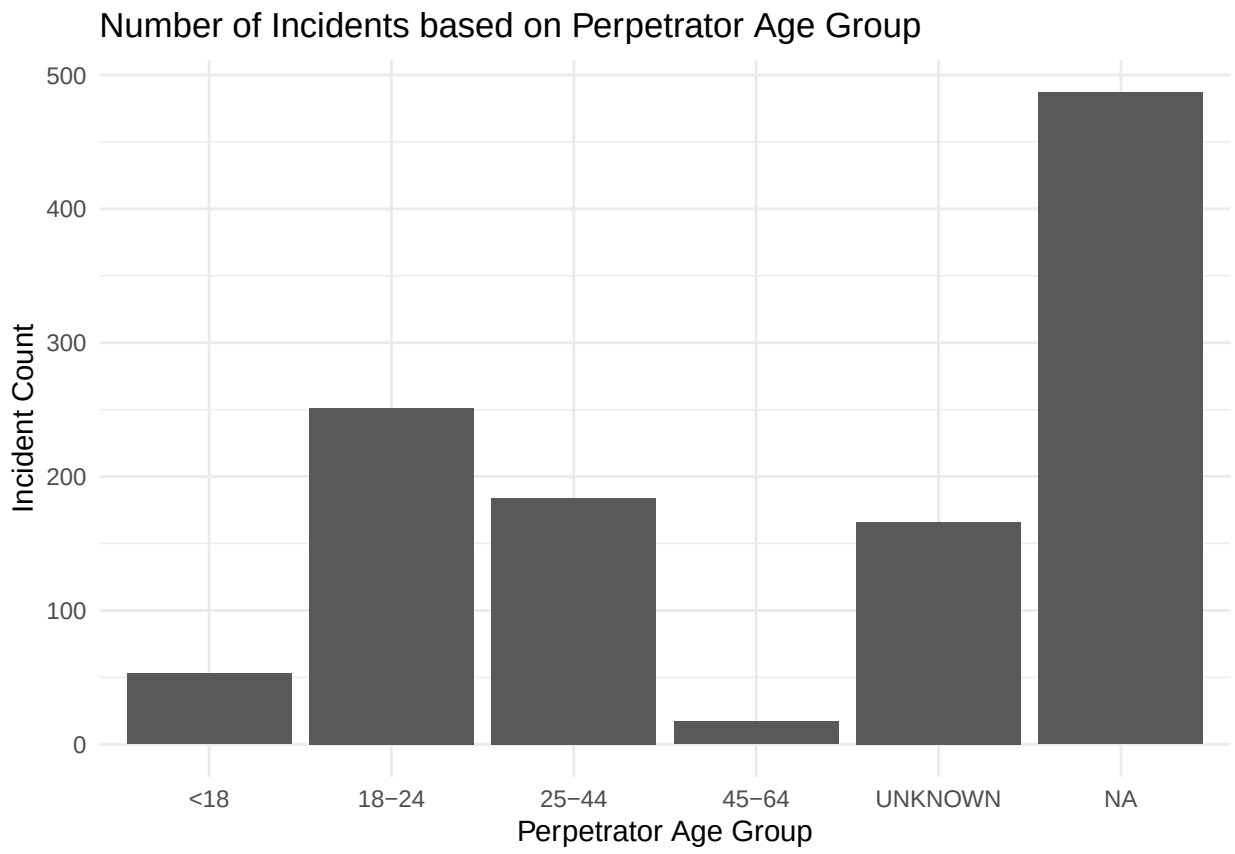
Looking at the data, we can see some interesting statistics regarding both groups. Specifically, that the majority of perpertrators and victims are both overwhelmingly male. Some interesting subsets where the majority are female are: White hispanic victims under the age of 18, asian/pacific islander victims from the ages of 25-44, and white, black, and asian/pacific islander victims over the age of 65.

Precincts with the most and least amount of shooting incidents visualized:

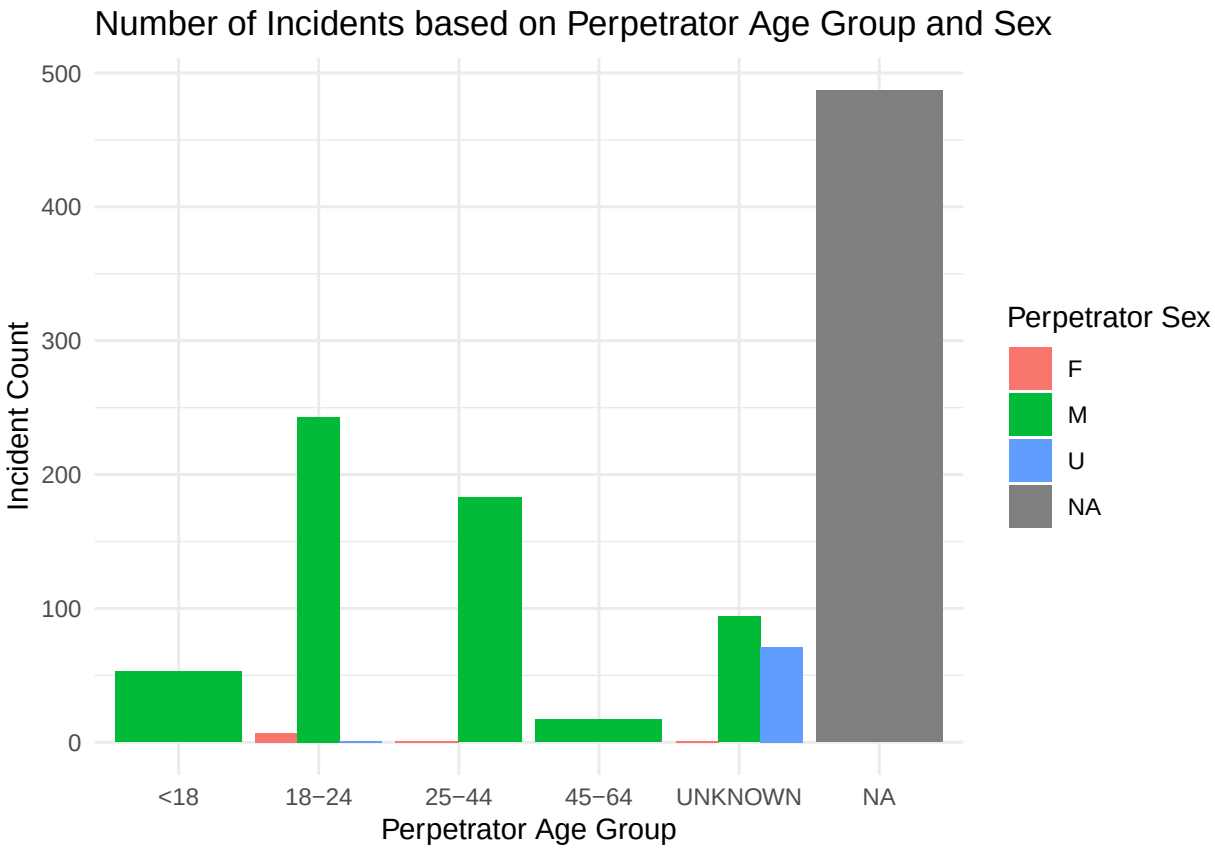
```
precinct_occurences  <- Shooting_incident_cleaned %>%
  group_by(PRECINCT) %>%
  summarize(count = n())

precinct_occurences <- precinct_occurences[with(precinct_occurences, order(-count)),]
ggplot(precinct_occurences[1:25,], aes(x=PRECINCT, y=count)) +
  geom_bar(stat="identity", width = 0.75) +
  labs(x = "Precinct Number", y = "Incident Count") +
  ggtitle("NYPD Shooting Incident Count: Top 25 Precincts with the most incidents") +
  theme_minimal()
```



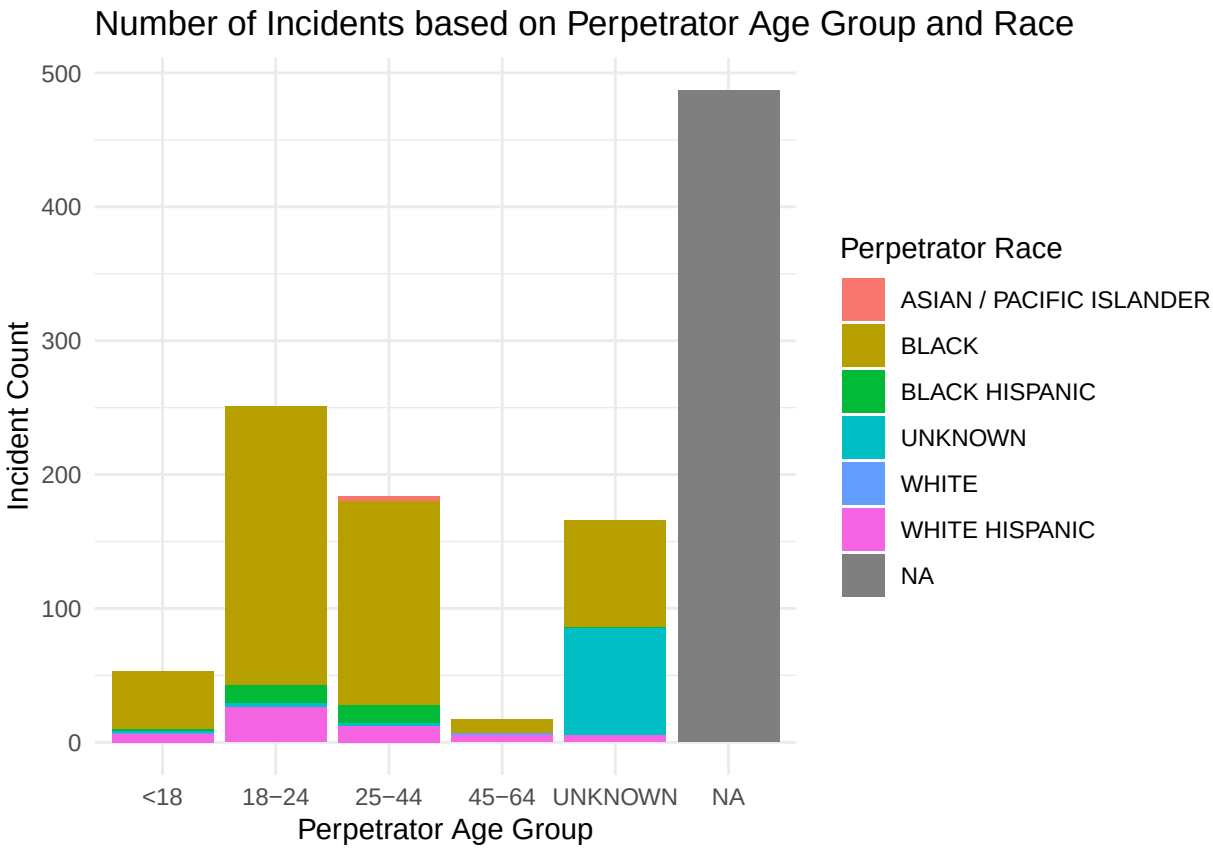

```
top_5_prec_count <- precinct_occurences[1:5,]
top_5_prec <- top_5_prec_count$PRECINCT[1:5]

top_prec <- Shooting_incident_cleaned[Shooting_incident_cleaned$PRECINCT == top_5_prec,]

ggplot(top_prec, aes(factor(PERP_AGE_GROUP),)) +
  geom_bar(stat="count", position="dodge") +
  labs(x = "Perpetrator Age Group", y = "Incident Count") +
  ggtitle("Number of Incidents based on Perpetrator Age Group ") +
  theme_minimal()
```



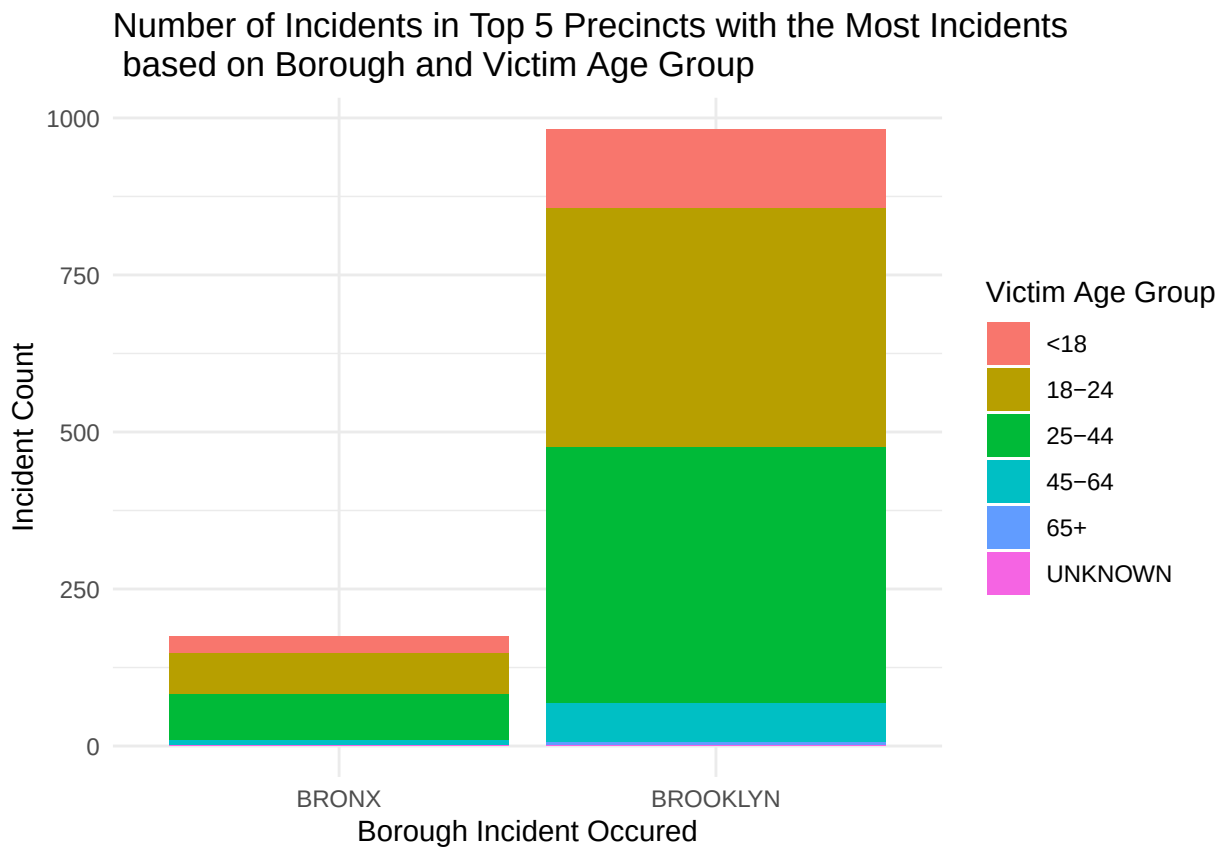

```
ggplot(top_prec, aes(factor(PERP_AGE_GROUP), fill=PERP_SEX)) +
  geom_bar(stat="count", position="dodge") +
  labs(x = "Perpetrator Age Group", y = "Incident Count", fill = "Perpetrator Sex") +
  ggtitle("Number of Incidents based on Perpetrator Age Group and Sex") +
  theme_minimal()
```



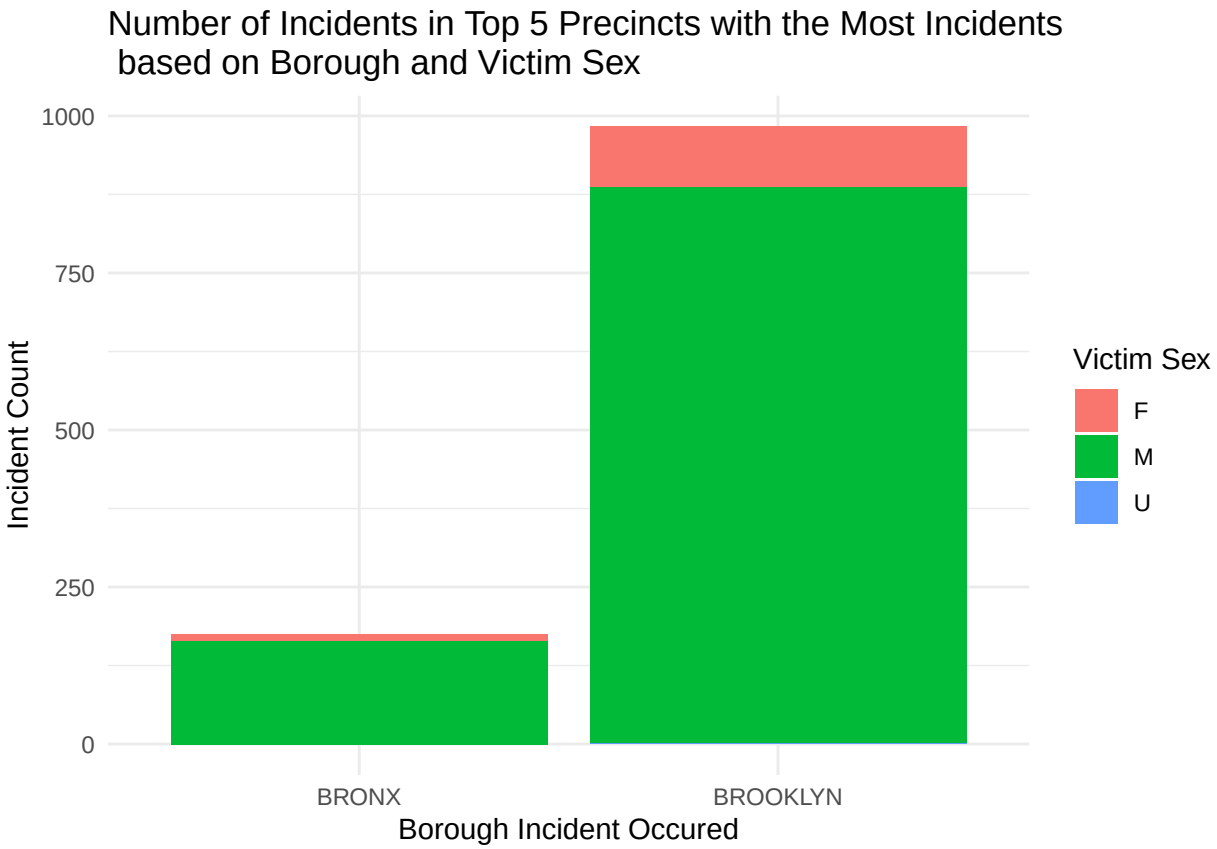

```
ggplot(top_prec, aes(factor(PERP_AGE_GROUP), fill=PERP_RACE)) +
  geom_bar(stat="count") +
  labs(x = "Perpetrator Age Group", y = "Incident Count", fill = "Perpetrator Race") +
  ggtitle("Number of Incidents based on Perpetrator Age Group and Race") +
  theme_minimal()
```



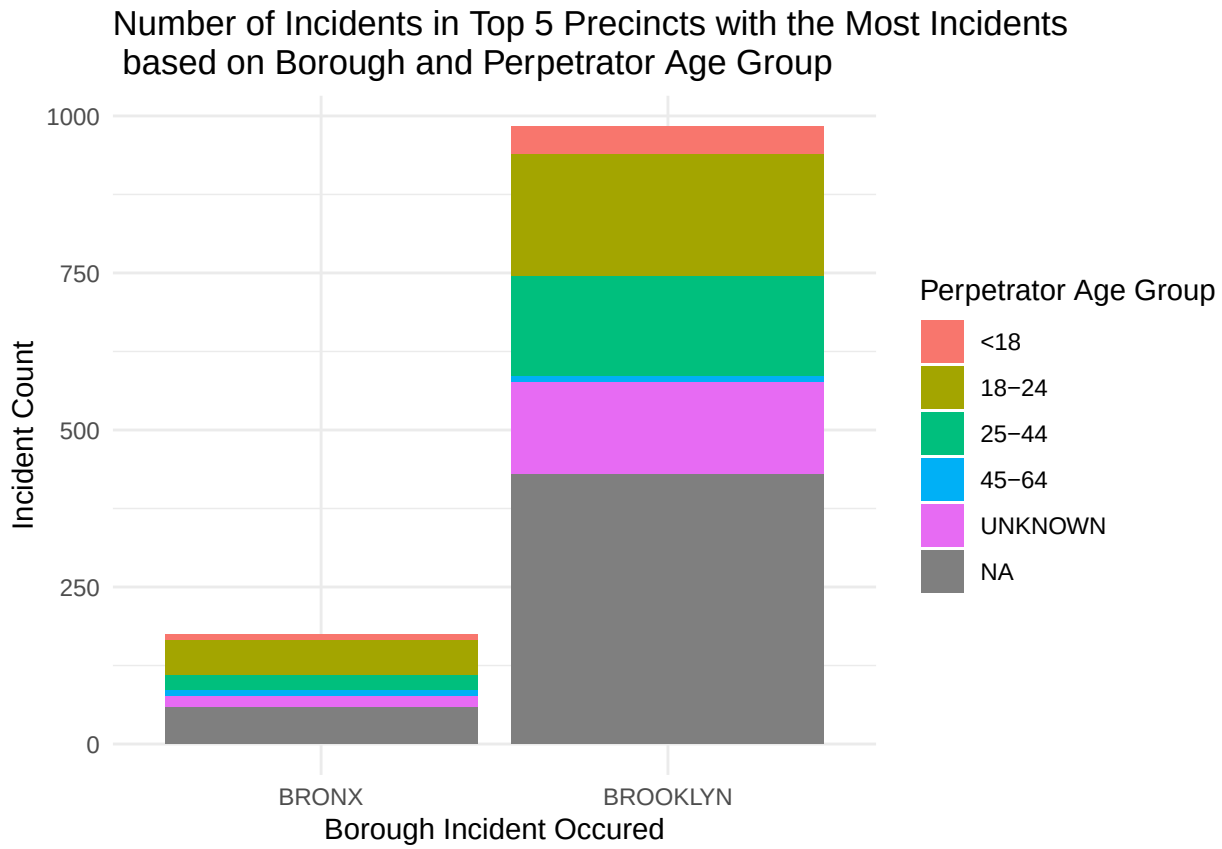

```
ggplot(top_prec, aes(factor(BORO), fill=VIC_AGE_GROUP)) +
  geom_bar(stat="count") +
  labs(x = "Borough Incident Occured", y = "Incident Count", fill = "Victim Age Group") +
  ggtitle("Number of Incidents in Top 5 Precincts with the Most Incidents \n based on Borough and Victi
  theme_minimal()
```



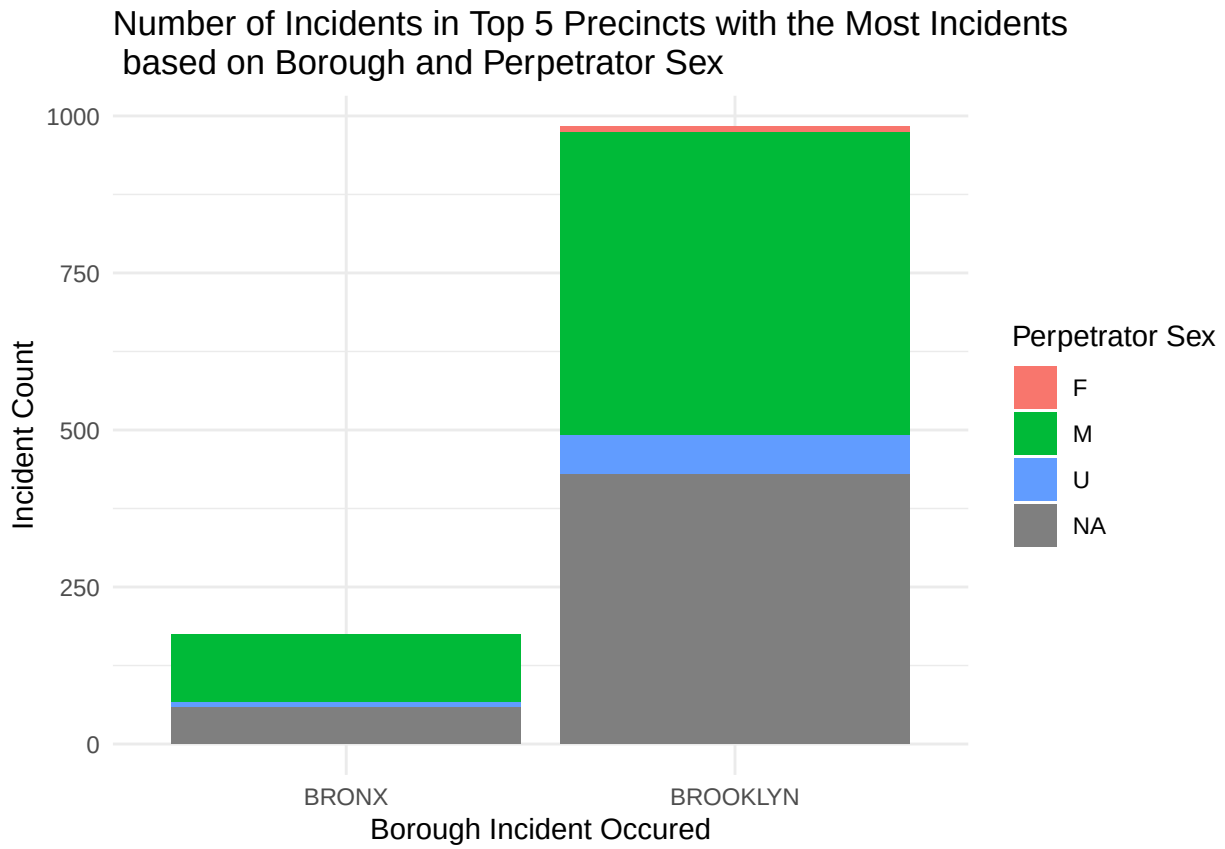

```
ggplot(top_prec, aes(factor(BORO), fill=VIC_SEX)) +
  geom_bar(stat="count") +
  labs(x = "Borough Incident Occured", y = "Incident Count", fill = "Victim Sex") +
  ggtitle("Number of Incidents in Top 5 Precincts with the Most Incidents \n based on Borough and Victi
  theme_minimal()
```



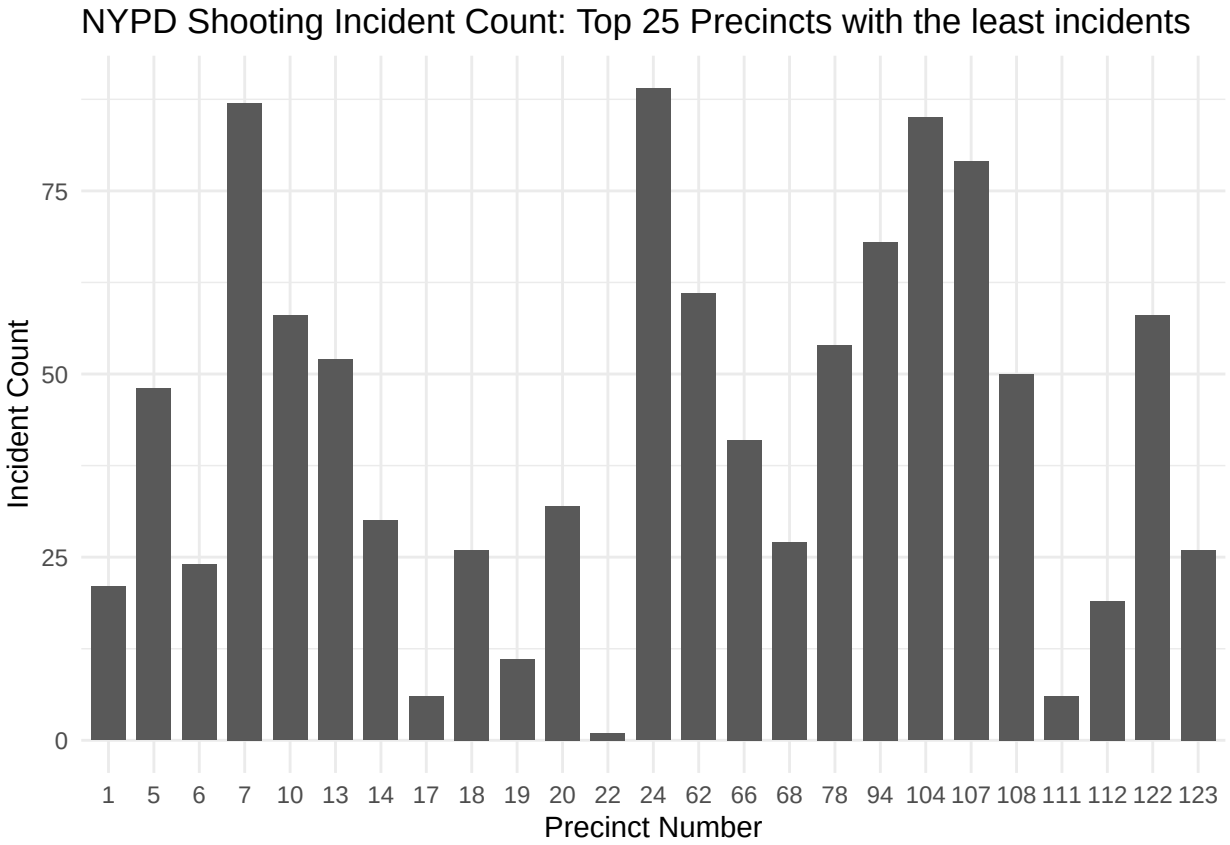

```
ggplot(top_prec, aes(factor(BORO), fill=PERP_AGE_GROUP)) +
  geom_bar(stat="count") +
  labs(x = "Borough Incident Occured", y = "Incident Count", fill = "Perpetrator Age Group") +
  ggtitle("Number of Incidents in Top 5 Precincts with the Most Incidents \n based on Borough and Perpe
  theme_minimal()
```




```
ggplot(top_prec, aes(factor(BORO), fill=PERP_SEX)) +
  geom_bar(stat="count") +
  labs(x = "Borough Incident Occured", y = "Incident Count", fill = "Perpetrator Sex") +
  ggtitle("Number of Incidents in Top 5 Precincts with the Most Incidents \n based on Borough and Perpe
  theme_minimal()
```




```
precinct_occurences  <- Shooting_incident_cleaned %>%
  group_by(PRECINCT) %>%
  summarize(count = n())

precinct_occurences <- precinct_occurences[with(precinct_occurences, order(count)),]
ggplot(precinct_occurences[1:25,], aes(x=PRECINCT, y=count)) +
  geom_bar(stat="identity", width = 0.75) +
  labs(x = "Precinct Number", y = "Incident Count") +
  ggtitle("NYPD Shooting Incident Count: Top 25 Precincts with the least incidents") +
  theme_minimal()
```



These are the top 25 precincts with the least shooting incidents in New York. This data may be useful for folks that are looking to find a safe place to live – although this is just one of the factors of many.

```
m2 = lm(STATISTICAL_MURDER_FLAG~OCCUR_TIME, data = Shooting_incident_cleaned)
m1 = glm(STATISTICAL_MURDER_FLAG~OCCUR_TIME, family="poisson", data = Shooting_incident_cleaned)
summary(m2)
```

```
##
## Call:
## lm(formula = STATISTICAL_MURDER_FLAG ~ OCCUR_TIME, data = Shooting_incident_cleaned)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3384 -0.1951 -0.1738 -0.1590  0.8557
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.168763   0.008967  18.820  < 2e-16 ***
## OCCUR_TIME1   0.004520   0.012756   0.354 0.723078
## OCCUR_TIME2   0.006437   0.013231   0.487 0.626588
## OCCUR_TIME3   0.008756   0.013622   0.643 0.520391
## OCCUR_TIME4   0.026283   0.014112   1.862 0.062554 .
## OCCUR_TIME5   0.083205   0.017945   4.637 3.56e-06 ***
## OCCUR_TIME6   0.063795   0.024292   2.626 0.008642 **
## OCCUR_TIME7   0.169621   0.029245   5.800 6.72e-09 ***
## OCCUR_TIME8   0.099658   0.029798   3.344 0.000826 ***
```

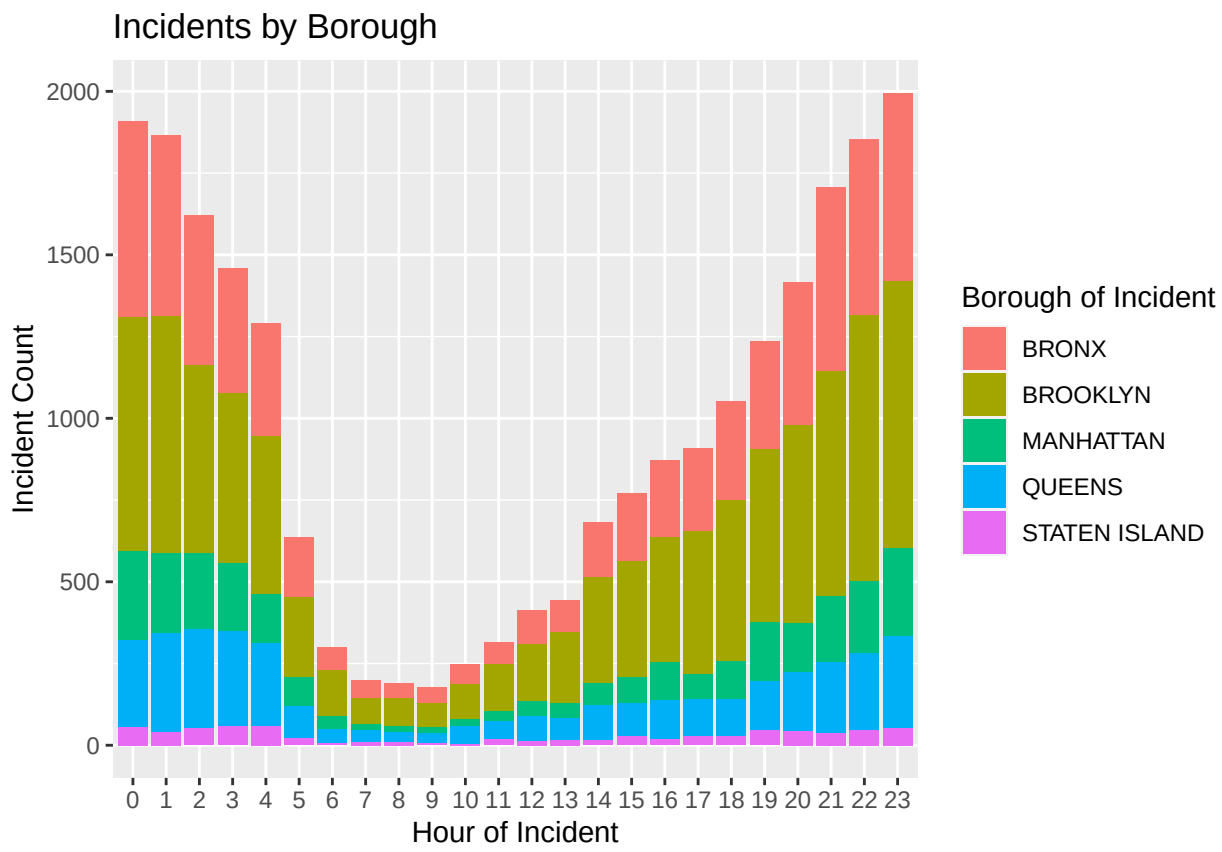
```
## OCCUR_TIME9   0.085474   0.030777   2.777 0.005487 **
## OCCUR_TIME10  0.069140   0.026440   2.615 0.008928 **
## OCCUR_TIME11  0.087566   0.023789   3.681 0.000233 ***
## OCCUR_TIME12  0.069791   0.021216   3.290 0.001005 **
## OCCUR_TIME13  0.052456   0.020658   2.539 0.011114 *
## OCCUR_TIME14  0.039143   0.017465   2.241 0.025023 *
## OCCUR_TIME15  0.005037   0.016715   0.301 0.763150
## OCCUR_TIME16 -0.024433   0.016005  -1.527 0.126868
## OCCUR_TIME17  0.036957   0.015786   2.341 0.019231 *
## OCCUR_TIME18  0.062006   0.015037   4.124 3.74e-05 ***
## OCCUR_TIME19  0.018281   0.014305   1.278 0.201277
## OCCUR_TIME20  0.002605   0.013733   0.190 0.849558
## OCCUR_TIME21  0.018591   0.013047   1.425 0.154217
## OCCUR_TIME22  0.033502   0.012773   2.623 0.008727 **
## OCCUR_TIME23 -0.009786   0.012544  -0.780 0.435312
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3917 on 23541 degrees of freedom
## Multiple R-squared:  0.005927,   Adjusted R-squared:  0.004956
## F-statistic: 6.102 on 23 and 23541 DF,  p-value: < 2.2e-16
```

```
summary(m1)
```

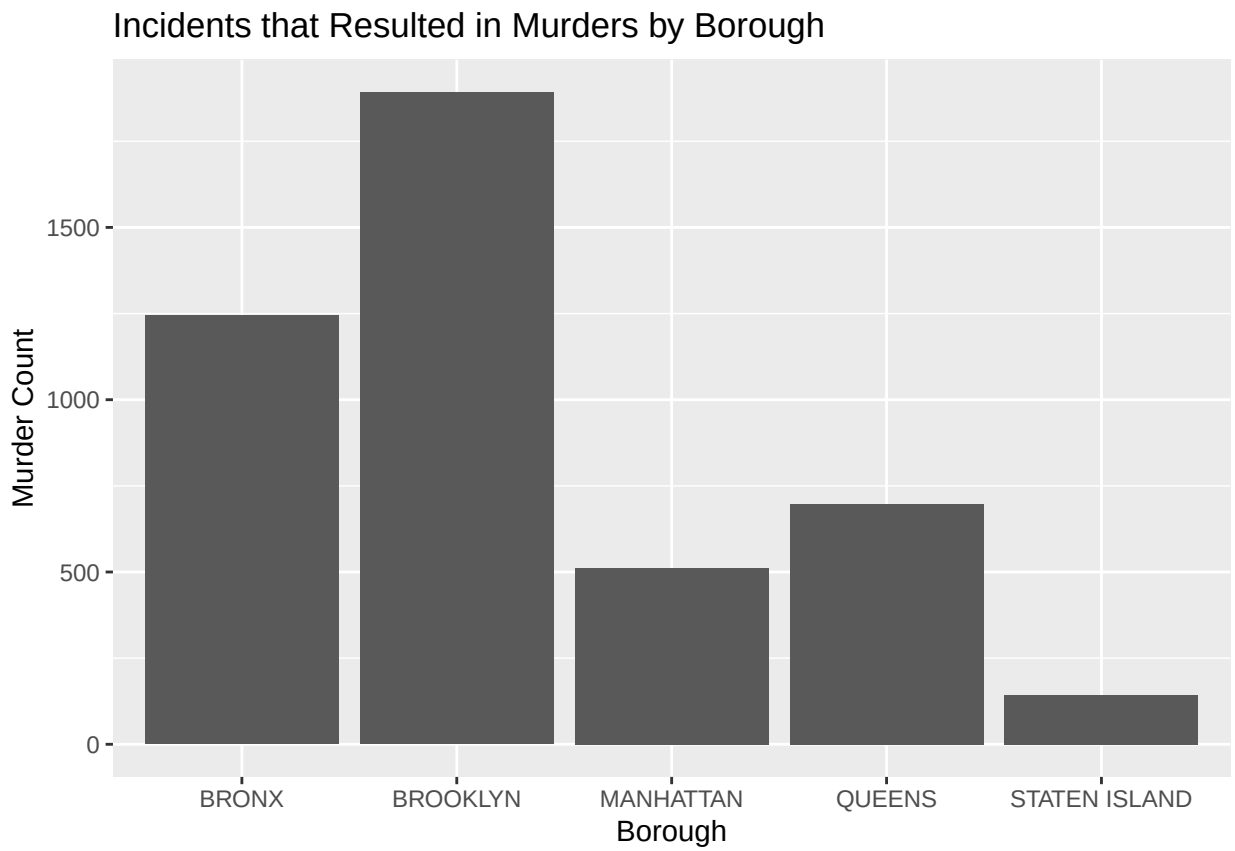
```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ OCCUR_TIME, family = "poisson",
##     data = Shooting_incident_cleaned)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8227  -0.6246  -0.5896  -0.5639   1.4697
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.77926    0.05573 -31.928  < 2e-16 ***
## OCCUR_TIME1   0.02643    0.07875   0.336 0.737143
## OCCUR_TIME2   0.03744    0.08140   0.460 0.645614
## OCCUR_TIME3   0.05058    0.08347   0.606 0.544512
## OCCUR_TIME4   0.14474    0.08411   1.721 0.085262 .
## OCCUR_TIME5   0.40081    0.09672   4.144 3.42e-05 ***
## OCCUR_TIME6   0.32064    0.13188   2.431 0.015040 *
## OCCUR_TIME7   0.69568    0.13428   5.181 2.21e-07 ***
## OCCUR_TIME8   0.46406    0.15071   3.079 0.002076 **
## OCCUR_TIME9   0.40977    0.15915   2.575 0.010030 *
## OCCUR_TIME10  0.34337    0.14161   2.425 0.015323 *
## OCCUR_TIME11  0.41797    0.12430   3.362 0.000772 ***
## OCCUR_TIME12  0.34610    0.11492   3.012 0.002598 **
## OCCUR_TIME13  0.27066    0.11537   2.346 0.018974 *
## OCCUR_TIME14  0.20859    0.10074   2.071 0.038390 *
## OCCUR_TIME15  0.02941    0.10280   0.286 0.774809
## OCCUR_TIME16 -0.15639    0.10508  -1.488 0.136667
## OCCUR_TIME17  0.19802    0.09194   2.154 0.031256 *
## OCCUR_TIME18  0.31292    0.08498   3.683 0.000231 ***
```

```
## OCCUR_TIME19  0.10285    0.08622   1.193 0.232936
## OCCUR_TIME20  0.01532    0.08498   0.180 0.856945
## OCCUR_TIME21  0.10450    0.07893   1.324 0.185532
## OCCUR_TIME22  0.18108    0.07598   2.383 0.017151 *
## OCCUR_TIME23 -0.05974    0.07912  -0.755 0.450246
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 14885  on 23564  degrees of freedom
## Residual deviance: 14779  on 23541  degrees of freedom
## AIC: 23803
##
## Number of Fisher Scoring iterations: 6
```

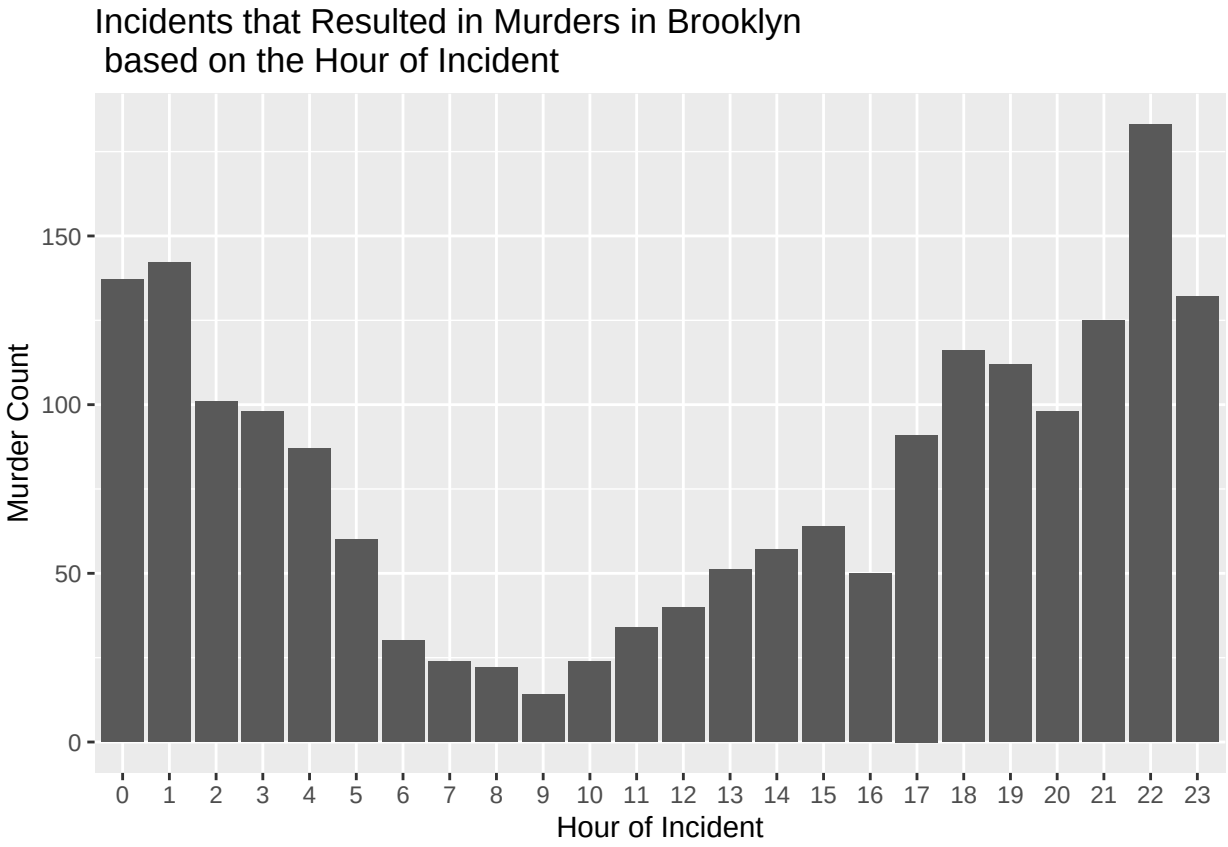
```r
ggplot(Shooting_incident_cleaned, aes(factor(OCCUR_TIME), fill=BORO)) +
  geom_bar(stat="count") +
  labs(x = "Hour of Incident", y = "Incident Count", fill = "Borough of Incident") +
  ggtitle("Incidents by Borough")
```




```r
ggplot(Shooting_incident_cleaned[Shooting_incident_cleaned$STATISTICAL_MURDER_FLAG == TRUE,], aes(facto
  geom_bar(stat="count") +
  labs(x = "Borough", y = "Murder Count") +
  ggtitle("Incidents that Resulted in Murders by Borough")
```




```
murder_set = Shooting_incident_cleaned[Shooting_incident_cleaned$STATISTICAL_MURDER_FLAG == TRUE,]

ggplot(murder_set[murder_set$BORO == 'BROOKLYN',], aes(factor(OCCUR_TIME))) +
  geom_bar(stat="count") +
  labs(x = "Hour of Incident", y = "Murder Count") +
  ggtitle("Incidents that Resulted in Murders in Brooklyn \n based on the Hour of Incident")
```



Bias and Conclusion

In this specific report, I have tried to mitigate any bias by specifically only showing generic data/statistics. I am not too passionate about this subject/dataset and thus do not want to make any hard conclusions or findings based off these statistics. What I can conclude based off my limited analysis on this dataset, is that the majority of perpetrators and victims are both male. There are few subsets in which they are are specifically more females, but overall there is an overwhelming majority of both male perpetrators and victims according to this dataset.

In conclusion, the one concrete thing we can get out of this analysis is that the overwhelming majority of perpetrators and victims are male. Some interesting things I learned while analyzing the dataset is that Brooklyn is the borough with the most incidents, and consequently the borough where the most incidents result in murder. This could be due to many factors like population, economic factors, or something similar so without these important aspects, we can't really come to a concrete conclusion. To conclude, this dataset and analysis may be useful to individuals that are worried about the safety in their specific borough or precinct compared to others.