NYPD DATA

First, we want to get the data and read it, so that we can begin our analysis.

```
police = read.csv('https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD')
```

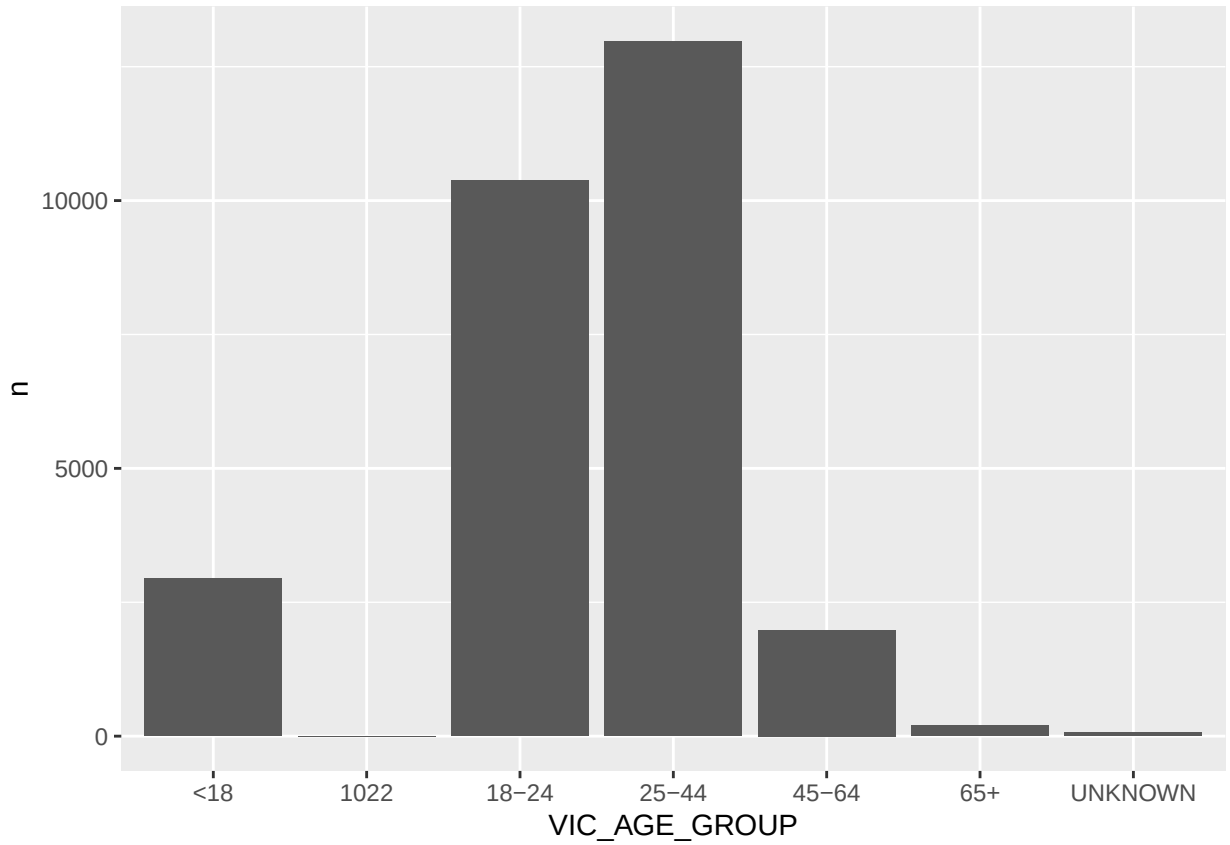
Once it is read, we can see that there are a lot of columns with different values. For the sake of simplicity and readability, I am going to remove some of the columns, so that I can focus my analysis on a few different factors at a time. I chose to remove all the columns which were mostly empty or columns that had no meaning to the average person (such as INCIDENT_KEY).

```
police = police %>%
  select(-c(INCIDENT_KEY, LOC_OF_OCCUR_DESC, PRECINCT, JURISDICTION_CODE, LOC_CLASSFCTN_DESC, LOCATION_
```

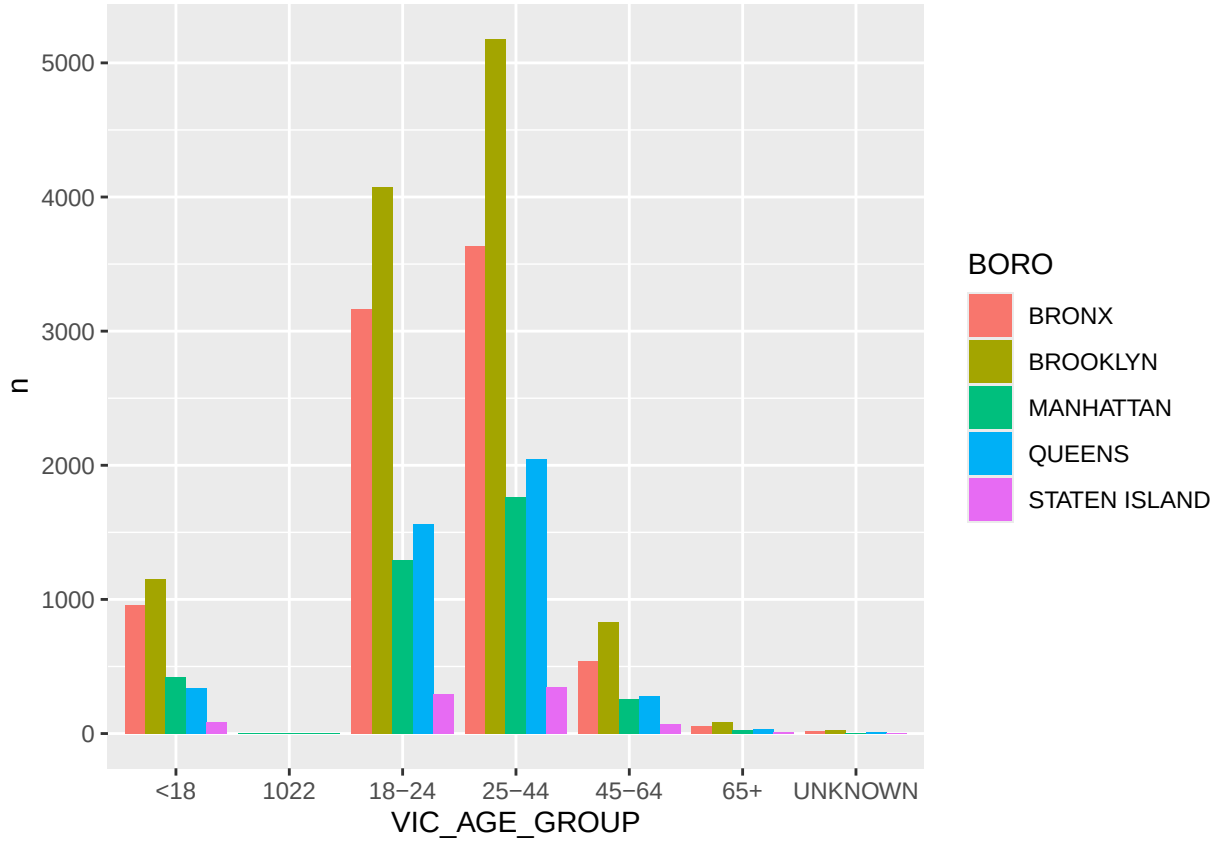
Once the number of columns were reduced, I was left with date, time, boro, victim age group, victim sex, and victim race. Unlike the COVID data example, there were no other data sets, so I did not need to combine anything, meaning that the data I had was the data I needed to do an analysis on.

Boro (location) vs Victim Age Group

```
boro = count(police, BORO, VIC_AGE_GROUP)
boro %>%
  ggplot(aes(x=VIC_AGE_GROUP, y=n)) + geom_bar(stat="identity")
```



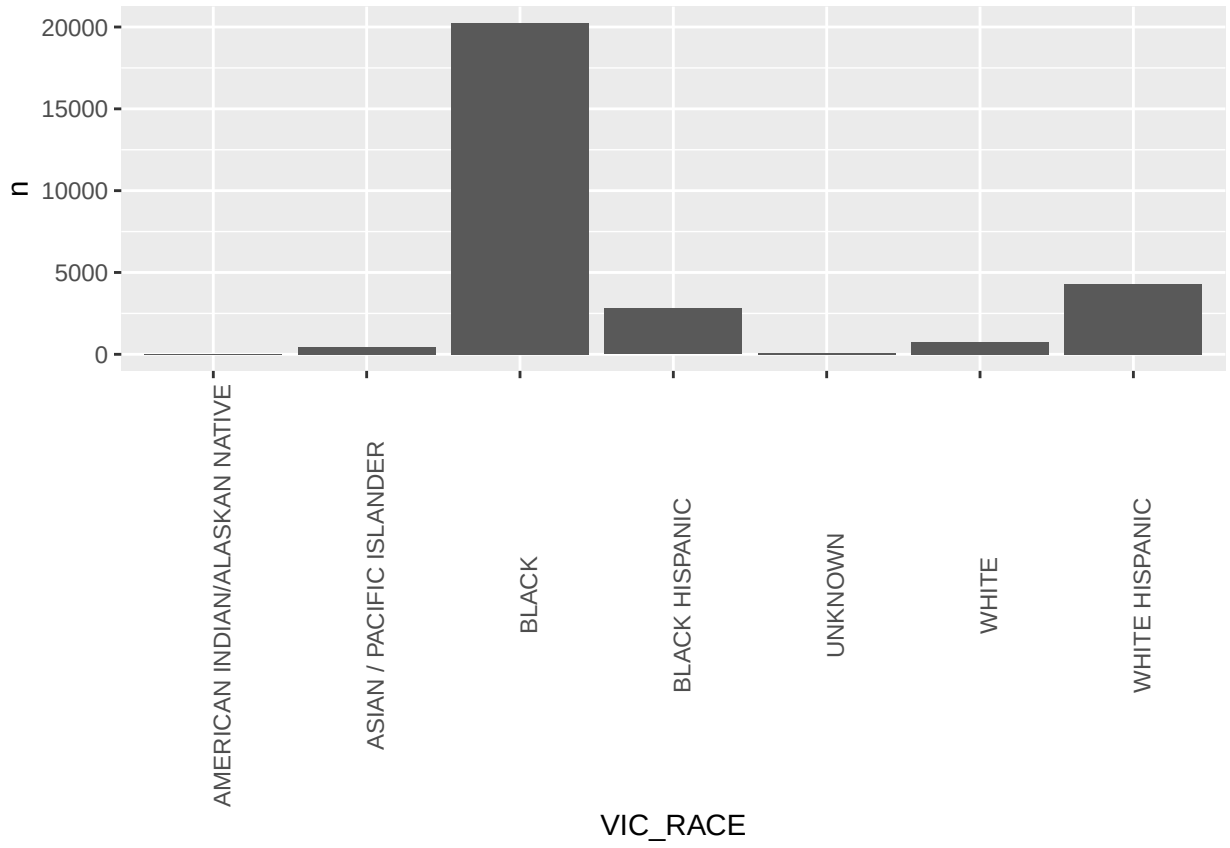
```
boro %>%
  ggplot(aes(x=VIC_AGE_GROUP, y=n, fill=BORO)) + geom_bar(stat="identity", position=position_dodge())
```



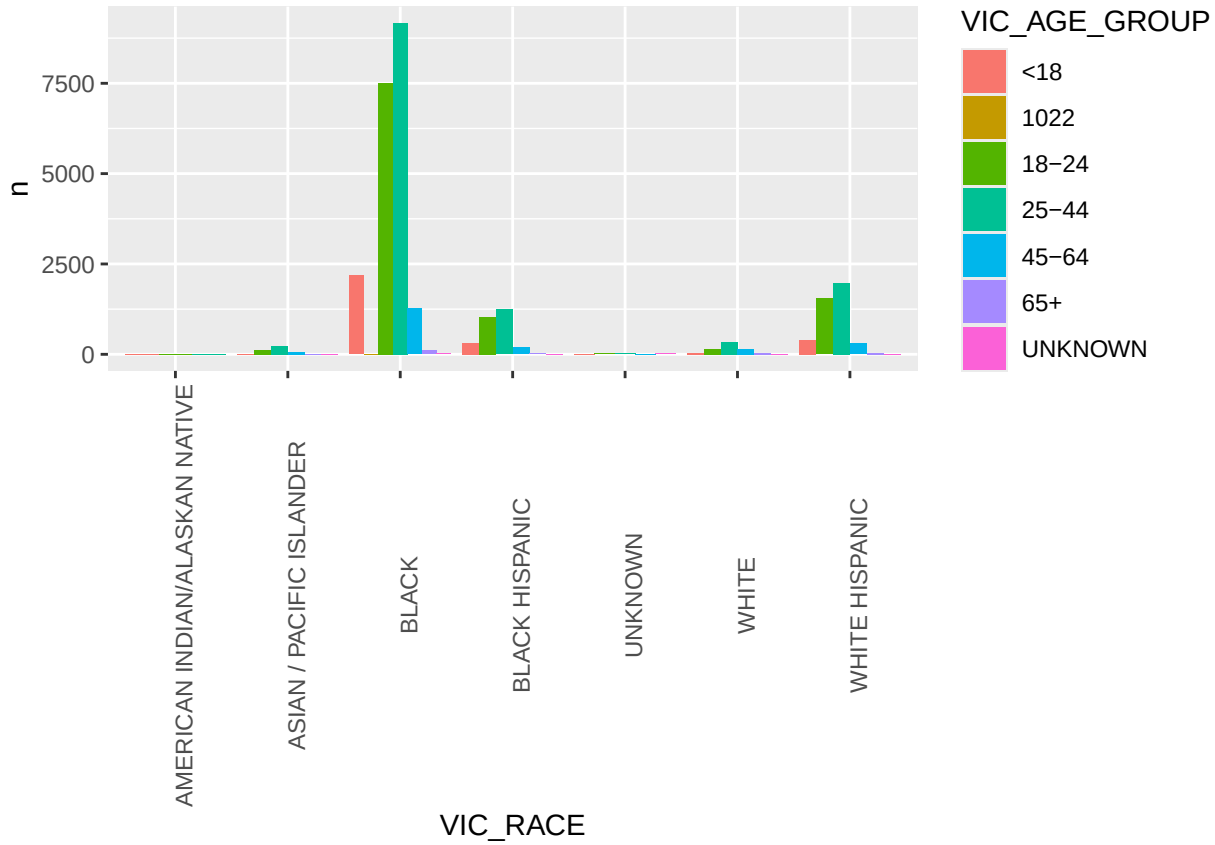
For the first graphs, I looked at the victims' ages and then broke it down by boro. This allowed me to try to see if there was a connection between boro (and therefore location) and the age of the victim. However, this was not very helpful as there was no clear connection and because we do not know the population numbers of each boro. Without these numbers, it is impossible to know if the higher numbers in Brooklyn are due to higher violence or if there is simply more people, leading to higher numbers of crime.

Victim Race vs Victim Age

```
race = count(police, VIC_RACE, VIC_AGE_GROUP)
race %>%
  ggplot(aes(x=VIC_RACE, y=n)) + geom_bar(stat="identity")  + theme(axis.text.x = element_text(angle = 9
```




```
race %>%
  ggplot(aes(x=VIC_RACE, y=n, fill=VIC_AGE_GROUP)) + geom_bar(stat="identity", position=position_dodge()
```

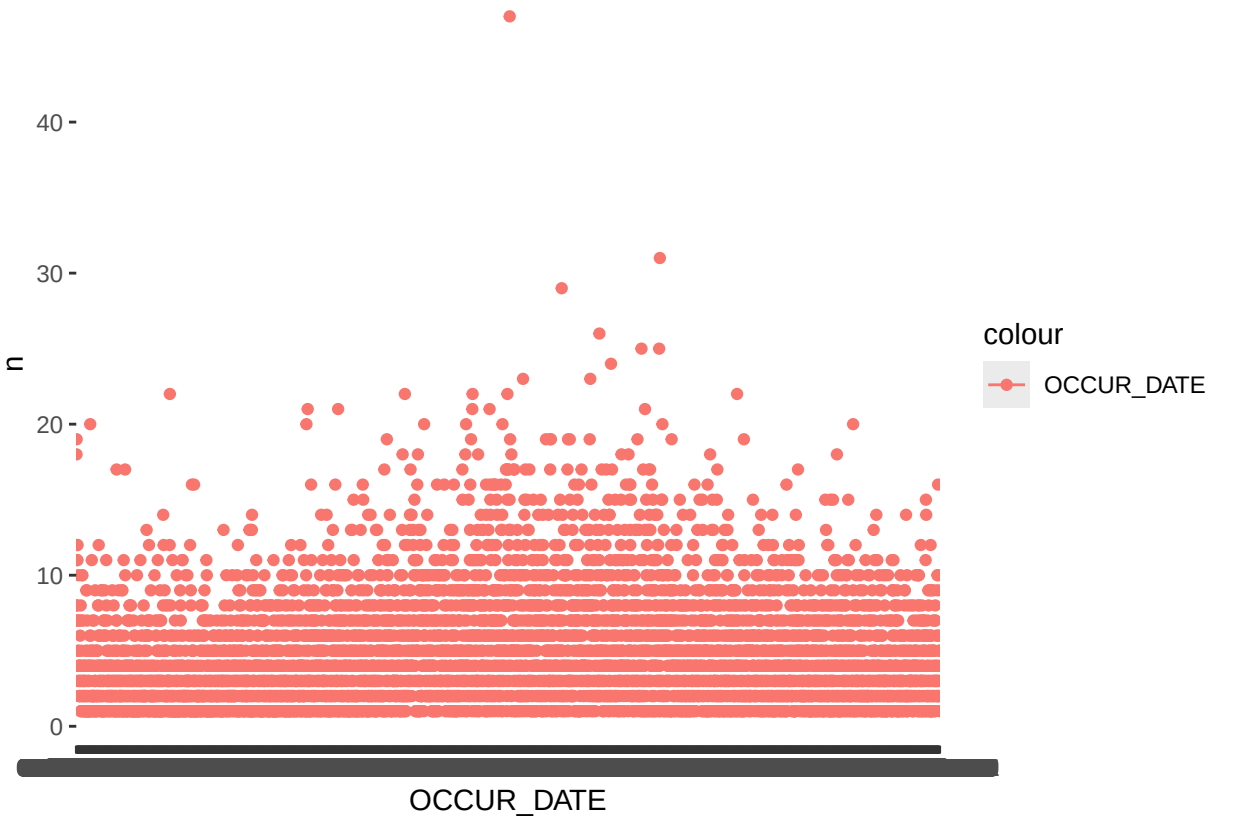


Because we did not know population of the boros, I next looked at the distribution of race and age of the victims. However, I once again realized that this is not necessarily a helpful breakdown as we do not know the population or distribution of race in New York City, so we cannot comment on high levels or low levels of crime by age or race.

Time of Year

```
crime = count(police, OCCUR_DATE)
crime %>%
  ggplot(aes(x=OCCUR_DATE, y=n))  +
  geom_line(aes(color = "OCCUR_DATE")) +
  geom_point(aes(color = "OCCUR_DATE"))
```

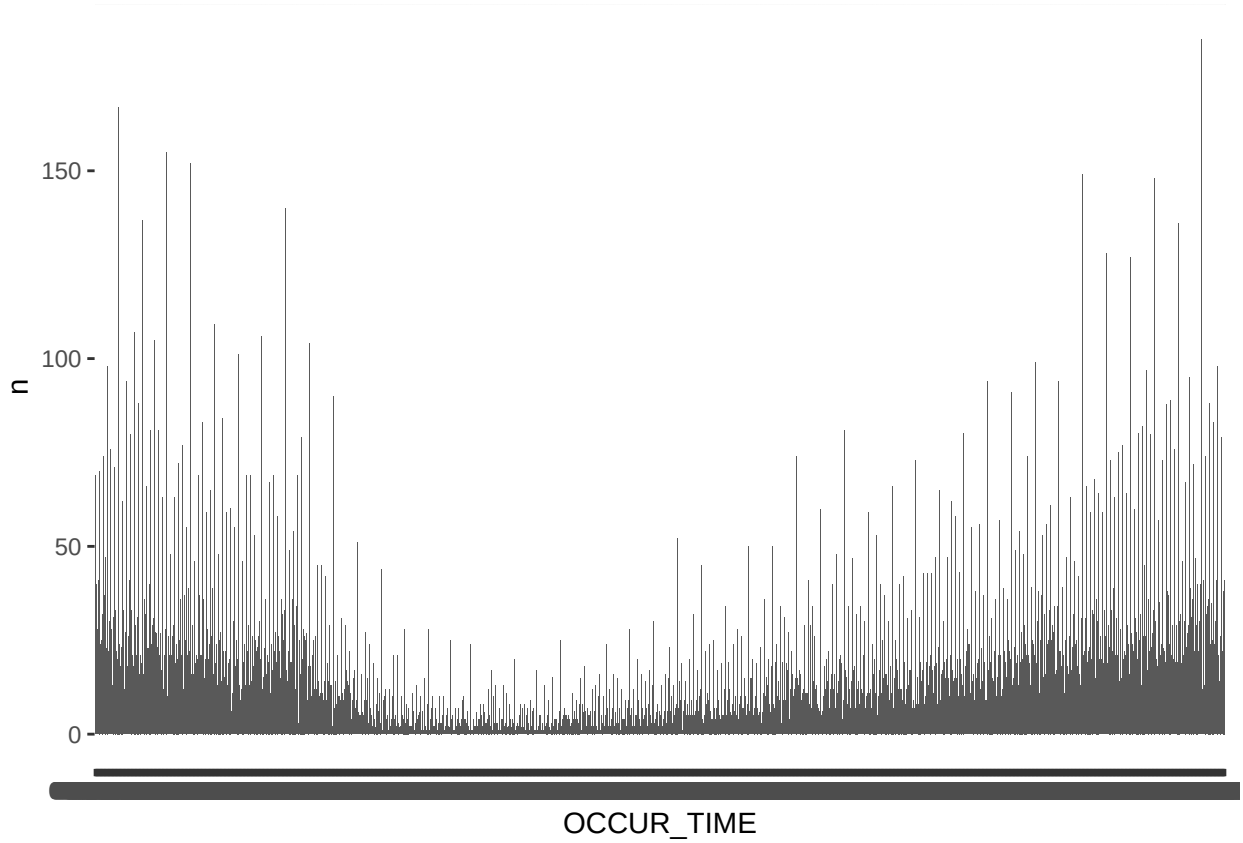
```
## `geom_line()`: Each group consists of only one observation.
## i Do you need to adjust the group aesthetic?
```



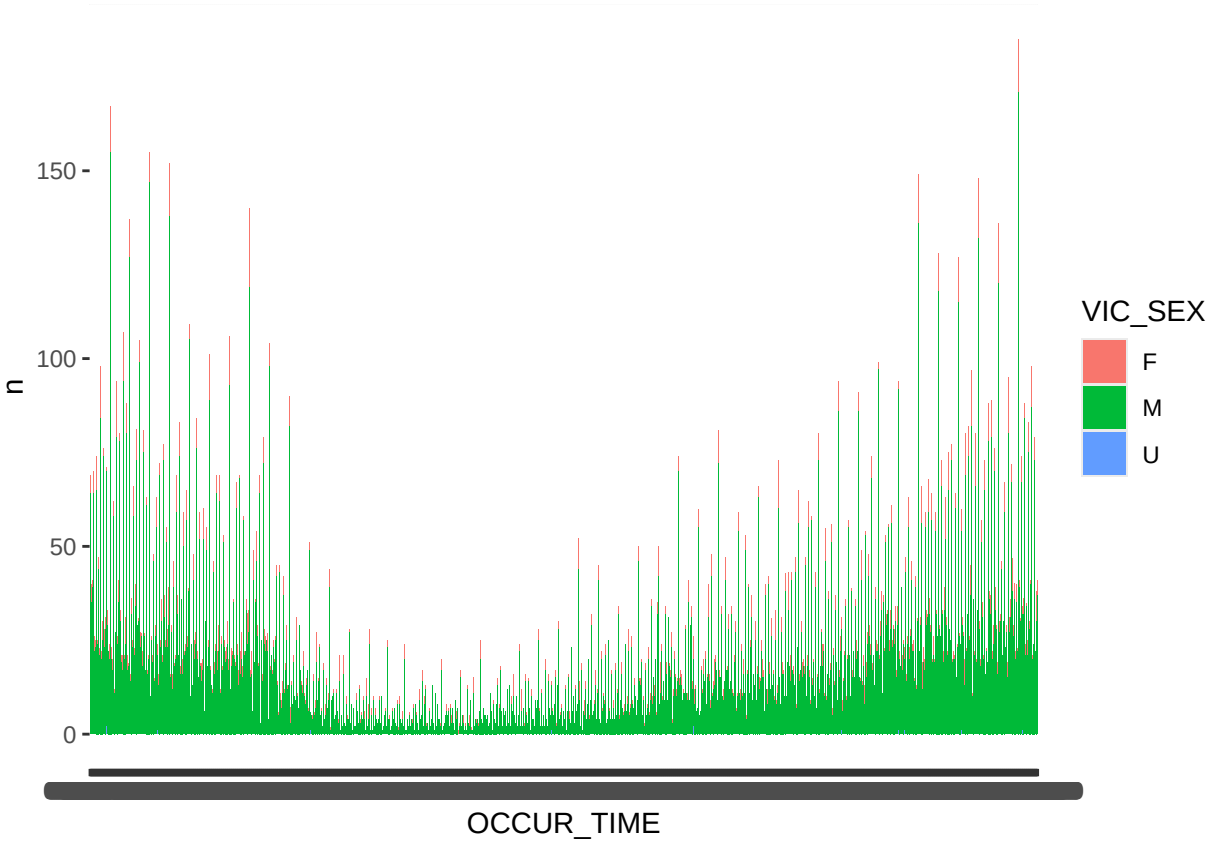
I then tried to plot the number of crimes by the date to see if we could see any trends of number of crime per dya and over time; however, the analysis did not show any trends. If we were trying to analysis this, we would need to try a different approach. I believe that the way I did the graphing ending up graphing the points based on the time of year rather than in chronological order. Ordering by year may have given a better analysis of how crime numbers changed between years.

Victim Sex vs Occur Time

```
sex = count(police, VIC_SEX, OCCUR_TIME)
sex %>%
  ggplot(aes(x=OCCUR_TIME, y=n)) + geom_bar(stat="identity")
```




```
sex %>%
  ggplot(aes(x=OCCUR_TIME, y=n, fill=VIC_SEX)) + geom_bar(stat="identity")
```



For my final analysis, I wanted to see if the time of the crime affected the sex of the victim. However, once again, there seemed to be little relation between the sex and time. However, it was interesting to see how the crimes were more likely to occur at night with a large dip during the day.

Bias

Overall, I there there could have been several instances of bias. To begin with, I chose the variables that I understood and were of interest to me. At the beginning, I greatly reduced the data, so many important elements might have been removed from my analysis. Furthermore, there could have been other ways to analysis the data than how I did it. My understanding and usage of the tools I knew and had could have limited and affected my analysis as there could have been better graphing or comparisons that would have made the analysis better or less biased.