NYPD Shooting Incidence Data

Introduction

The following dataset has been chosen as part of the Week 3 assignment to look at the NYPD Shooting Incidence data from 2006 to 2021. The following description comes directly from the metadata of the dataset. This will provide some understanding on the usage and definitions of the data. I am going to specifically look at the timing of when these incidences occur.

> List of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year.This is a breakdown of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year. 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. This data can be used by the public to explore the nature of shooting/criminal activity. Please refer to the attached data footnotes for additional information about this dataset.

The data dictionary for these fields are as follows:

- INCIDENT_KEY - Randomly generated persistent ID for each arrest
- OCCUR_DATE - Exact date of the shooting incident
- OCCUR_TIME - Exact time of the shooting incident
- BORO - Borough where the shooting incident occurred
- PRECINCT - Precinct where the shooting incident occurred
- JURISDICTION_CODE - Jurisdiction where the shooting incident occurred. Jurisdiction codes 0 (Patrol), 1 (Transit) and 2 (Housing) represent NYPD whilst codes 3 and more represent non NYPD jurisdictions
- LOCATION_DESC - Location of the shooting incident
- STATISTICAL_MURDER_FLAG - Shooting resulted in the victim's death which would be counted as a murder
- PERP_AGE_GROUP - Perpetrator's age within a category
- PER_SEX - Perpetrator's sex description
- PERP_RACE - Perpetrator's race description
- VIC_AGE_GROUP - Victim's age within a category
- VIC_SEX - Victim's sex description
- VIC_RACE - Victim's race description
- X_COORD - Midblock X-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)
- Y_COORD - Midblock Y-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)
- Latitude - Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
- Longitude - Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
- Lon_Lat - Longitude and Latitude Coordinates for mapping

Library Import

```
library(tidyverse)
library(readr)
library(lubridate)
library(reshape2)
```

Importing and Data Cleaning

The first thing we need to do is import data.

Data Import

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

```
## Rows: 25596 Columns: 19
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (10): OCCUR_DATE, BORO, LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX, PERP_R...
## 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.
```

Data Cleaning

There are a number of data variables that need to be factored and the data type has to be changed.

```
nypd_data$OCCUR_DATE <- mdy(nypd_data$OCCUR_DATE)
nypd_data$OCCUR_TIME <- hms(nypd_data$OCCUR_DATE)

col_names <- c('BORO','PRECINCT','JURISDICTION_CODE','LOCATION_DESC','PERP_AGE_GROUP','PERP_SEX','PERP_
nypd_data[col_names] <- lapply(nypd_data[col_names] ,factor)
```

I find that it is also much easier to break out some of the time information to get a better sense of what the data can tell us.

Since I am interested in seeing some of the trends throughout time, I've added on a few columns such as `Year`, `WeekDay` and `Month` which will help understand the time of the data.

```
# Remove
 nypd_data <- nypd_data %>% filter(!PERP_AGE_GROUP %in% c( "1020", "224", "940"))
# Add additional columns that help provide a little more color
nypd_data <- nypd_data %>% mutate(Month = month(OCCUR_DATE, label=TRUE), Year = year(OCCUR_DATE), WeekDa
```

Selecting data

There are a few attributes that may not have much significance to the scope of this analysis:

- Location Attributes
  - **LOCATION_DESC** - one of the main reasons that it is hard to understand exactly what is happening is based on the 59% null values..
  - **X_COORD_CD** - I'd assume that the mapping software used by the NYPD or the data set find this information useful. However with the same map it make is difficult to use
  - **Y_COORD_CD** - This would account for the same information as above
  - **Lon_Lat** - This has been already provided in the Latitude and Longitude attributes
- Table elements
  - **STATISTICAL_MURDER_FLAG** - This attribute seems to be named a little strange... I know that there are deaths involved in this data and to call it statistical is probably just a rubric to help identify cases.

```
drop_col <- c('X_COORD_CD', 'Y_COORD_CD', 'Lon_Lat', 'LOCATION_DESC','STATISTICAL_MURDER_FLAG')
nypd_data_subset <- nypd_data[,!(names(nypd_data) %in% drop_col)]
```

```
nypd_data_subset %>% summary()
```

```
##   INCIDENT_KEY         OCCUR_DATE
##  Min.   :  9953245   Min.   :2006-01-01
##  1st Qu.: 61593632   1st Qu.:2009-05-10
##  Median : 86437258   Median :2012-08-26
##  Mean   :112383964   Mean   :2013-06-13
##  3rd Qu.:166660833   3rd Qu.:2017-07-01
##  Max.   :238490103   Max.   :2021-12-31
##
##    OCCUR_TIME                                      BORO          PRECINCT
##  Min.   :83d 13H 47M 29S                BRONX        : 7400   75     : 1470
##  1st Qu.:83d 16H 50M 39S                BROOKLYN     :10364   73     : 1372
##  Median :83d 19H 54M 51S                MANHATTAN    : 3265   67     : 1160
##  Mean   :83d 20H 48M 12.3960067220032S  QUEENS       : 3828   79     :  982
##  3rd Qu.:84d 0H 52M 45S                 STATEN ISLAND:  736   44     :  949
##  Max.   :84d 4H 58M 59S                                      47     :  902
##                                                              (Other):18758
##  JURISDICTION_CODE PERP_AGE_GROUP PERP_SEX             PERP_RACE
##  0   :21319        18-24  :5844   F   :  371   BLACK         :10667
##  1   :   59        25-44  :5202   M   :14413   WHITE HISPANIC: 2162
##  2   : 4213        UNKNOWN:3148   U   : 1499   UNKNOWN       : 1836
##  NA's:    2        <18    :1463   NA's: 9310   BLACK HISPANIC: 1203
##                    45-64  : 535                WHITE         :  272
##                    (Other):  57                (Other)       :  143
##                    NA's   :9344                NA's          : 9310
##  VIC_AGE_GROUP   VIC_SEX                              VIC_RACE
##  <18    : 2681   F: 2403   AMERICAN INDIAN/ALASKAN NATIVE:    9
##  18-24  : 9603   M:23179   ASIAN / PACIFIC ISLANDER      :  354
##  25-44  :11384   U:   11   BLACK                         :18280
##  45-64  : 1698             BLACK HISPANIC                : 2485
##  65+    :  167             UNKNOWN                       :   65
```

```
##  UNKNOWN:   60             WHITE                          :  660
##                            WHITE HISPANIC                 : 3740
##     Latitude       Longitude          Month           Year      WeekDay
##  Min.   :40.51   Min.   :-74.25   Jul    :3009   Min.   :2006   Sun:5155
##  1st Qu.:40.67   1st Qu.:-73.94   Aug    :3002   1st Qu.:2009   Mon:3597
##  Median :40.70   Median :-73.92   Jun    :2657   Median :2012   Tue:2944
##  Mean   :40.74   Mean   :-73.91   Sep    :2416   Mean   :2013   Wed:2818
##  3rd Qu.:40.82   3rd Qu.:-73.88   May    :2401   3rd Qu.:2017   Thu:2809
##  Max.   :40.91   Max.   :-73.70   Oct    :2176   Max.   :2021   Fri:3384
##                                   (Other):9932                  Sat:4886
```

Missing Data

As for missing data, there are a few things to consider. Most of the data is available, however there are a few insights:

- JURISDICTION_CODE has 2 NAs, most likely this was due to incorrect transcription or input error. They probably could be giving a 0 as a large majority of these are coded as 0 for Patrol jurisdiction
- For the **PERP_** attributes, there are two types of what at first glance appears to be missing data. UNKNOWN / NAs could fall into a few different cases:
  - The unknowns could be cases where the perpetrator has fled the scene.
  - The unknowns could be cases where individuals shot themselves
  - There could be missing data though when you look at the victim information, there are only 65 unknown cases which would mean that the number of bad data entry is a small percentage in these cases.

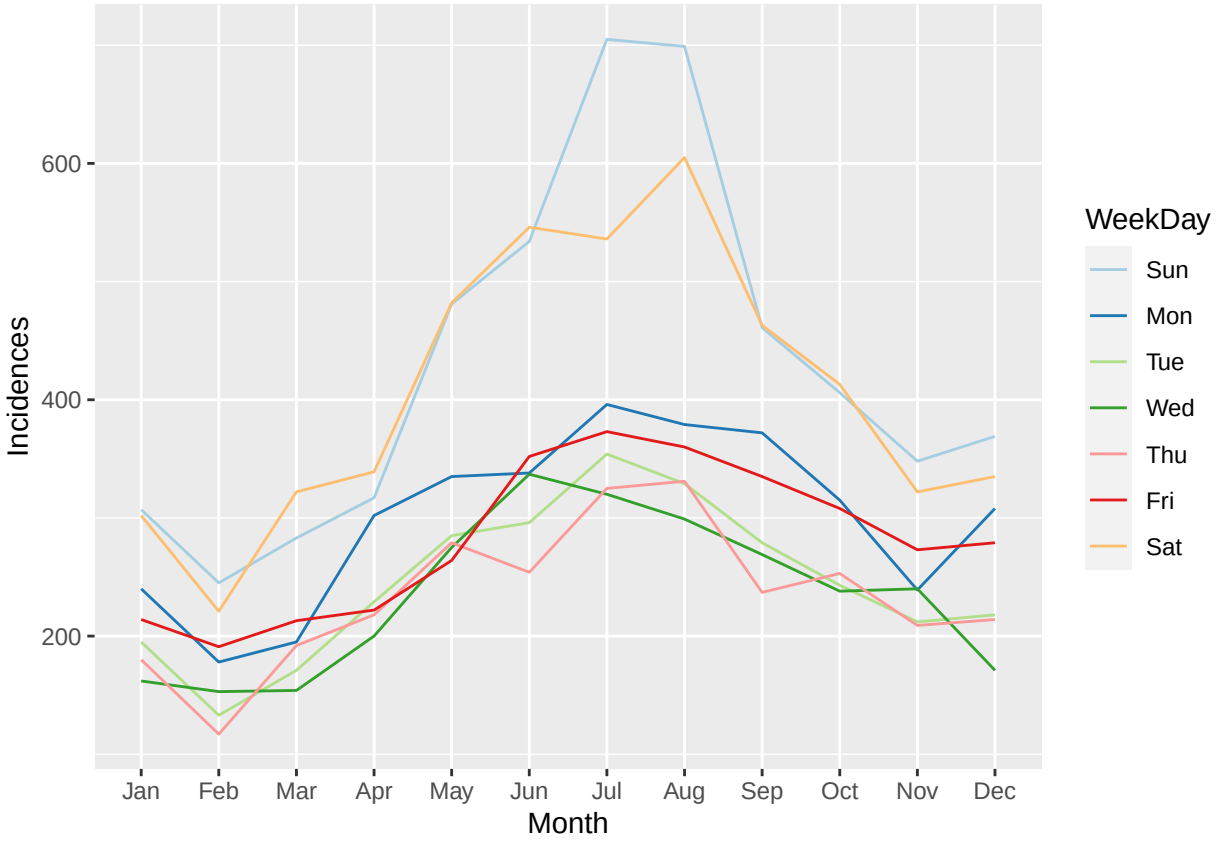
To handle this information, I find that keeping the unknowns and NAs in the data are the most likely scenario and better to have some understand of what they are.

Analysis

Are there patterns in time that affect the incidental behavior?

```
nypd_data_subset %>%
  group_by( Month, WeekDay) %>%
  summarize(Incidences = n()) %>%
  ggplot(aes(x=Month, y=Incidences, color=WeekDay, group=WeekDay)) + geom_line() +
scale_color_brewer(palette="Paired")
```

```
## `summarise()` has grouped output by 'Month'. You can override using the
## `.groups` argument.
```



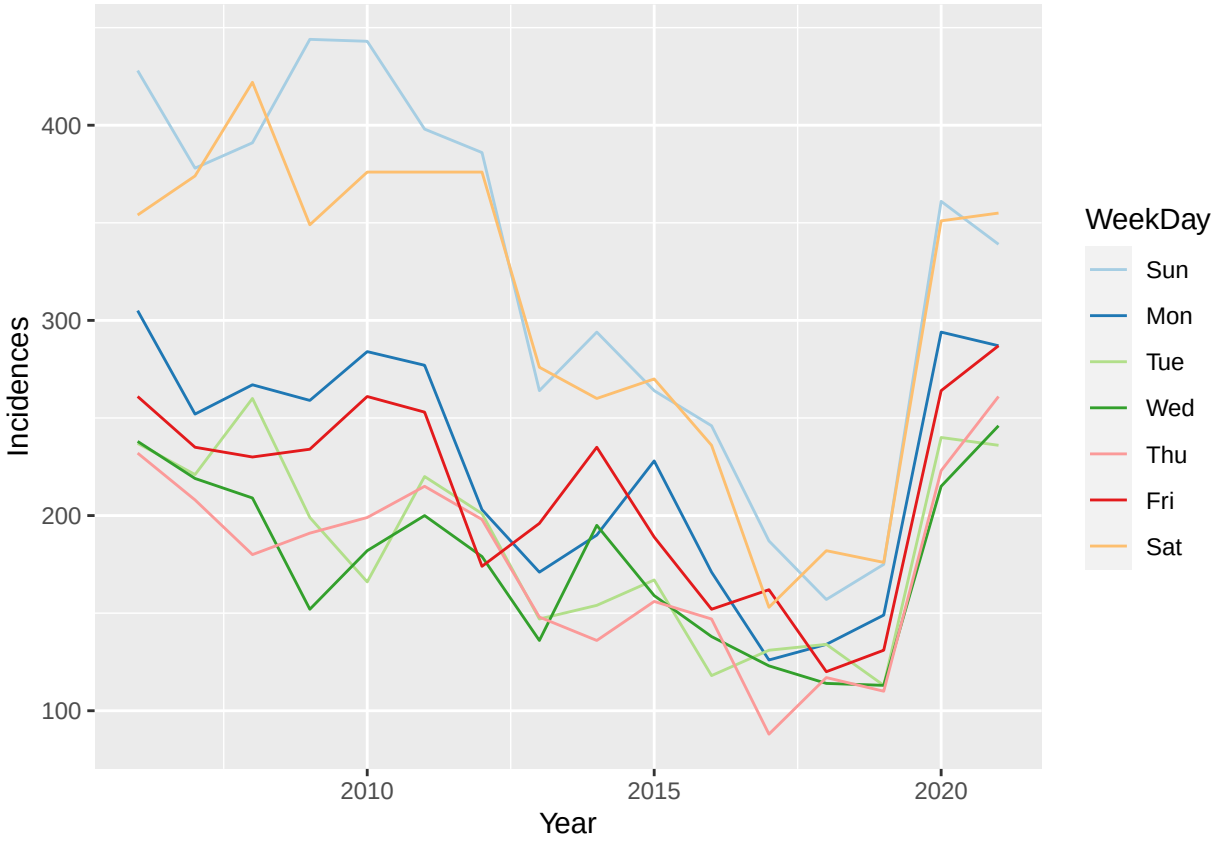
I was curious if the week of the day had any effect on the incidences; and we can see that there are a few observations.

1. We see that over the summer between May and September there is an uptick to incidences
2. There is a visual difference where the trends on Saturday and Sunday are higher
3. The trends compound where we see that on Saturday and Sunday has a much higher incidence rate than the rest of the year.

This could suggest that it could be linked to the following: * School aged children have a high incidence rate on the weekends and between school sessions * Weekends provide more opportunities outside traditional "work days" * Summers provide more opportunities for people to be in closer proximity

Lets just look at how over the years what the incidents were reported..

```
## `summarise()` has grouped output by 'Year'. You can override using the
## `.groups` argument.
```



We see that earlier 2006 to around 2013 there had been a larger incidence of incidents that occurred on the weekends on Saturday and Sundays. After 2013, we see that the weekends start converging which could be due to a number of variables that are regulatory, or satisfaction levels. This would be hard to find an answer without any further sources.

What is the age distribution?

Here we take the data and cast the information in such a way to understand the perpetrator's age group and sex.

```
nypd_data_subset %>%
  filter(!PERP_AGE_GROUP %in% c(NA)) %>%
  group_by(PERP_AGE_GROUP, PERP_SEX) %>%
  summarize(Incidences = n()) %>%
  #mutate(IncidencePct = round(Incidences / sum(Incidences),3)) %>%
  dcast(PERP_AGE_GROUP  ~ PERP_SEX)
```

```
## `summarise()` has grouped output by 'PERP_AGE_GROUP'. You can override using
## the `.groups` argument.
## Using Incidences as value column: use value.var to override.
```

```
##   PERP_AGE_GROUP   F    M    U
## 1            <18  37 1423    3
## 2          18-24 141 5687   16
## 3          25-44 157 5038    7
```

```
## 4          45-64  17  518   NA
## 5            65+   1   56   NA
## 6        UNKNOWN  18 1691 1439
```

We see that the perpetrators in this data is predominantly male and between the ages of 18 and 44 which span two age groups. So we can at least rule out a dominant effect that school aged children are part of the incidents that have been recorded. We do need to highlight that the unknown column would be cases in which an incident was reported and the victim had not been able to identify or remember the perpetrator.

Linear Model

```
nypd_data_lm <- nypd_data_subset %>%
  group_by( Month) %>%
  summarize(Incidences = n())


mod <- lm(Incidences ~ Month, data = nypd_data_lm )

summary(mod)
```

```
##
## Call:
## lm(formula = Incidences ~ Month, data = nypd_data_lm)
##
## Residuals:
## ALL 12 residuals are 0: no residual degrees of freedom!
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2132.75        NaN     NaN      NaN
## Month.L       765.20        NaN     NaN      NaN
## Month.Q     -1407.19        NaN     NaN      NaN
## Month.C      -477.87        NaN     NaN      NaN
## Month^4       873.22        NaN     NaN      NaN
## Month^5        96.84        NaN     NaN      NaN
## Month^6       -29.02        NaN     NaN      NaN
## Month^7      -132.93        NaN     NaN      NaN
## Month^8        48.99        NaN     NaN      NaN
## Month^9        92.07        NaN     NaN      NaN
## Month^10      193.95        NaN     NaN      NaN
## Month^11      -38.47        NaN     NaN      NaN
##
## Residual standard error: NaN on 0 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:    NaN
## F-statistic:   NaN on 11 and 0 DF,  p-value: NA
```

Showing the output for a linear model here but not sure how to interpret these results as we haven't taken linear models in any class at this point.

What about the victims?

Here we take the information above and try to display in a slightly different way to get an understanding of the information.

```
nypd_data_subset %>%
  filter(!VIC_AGE_GROUP %in% c(NA)) %>%
  group_by(VIC_AGE_GROUP, VIC_SEX) %>%
  summarize(Incidences = n()) %>%
  dcast(VIC_AGE_GROUP  ~ VIC_SEX)
```

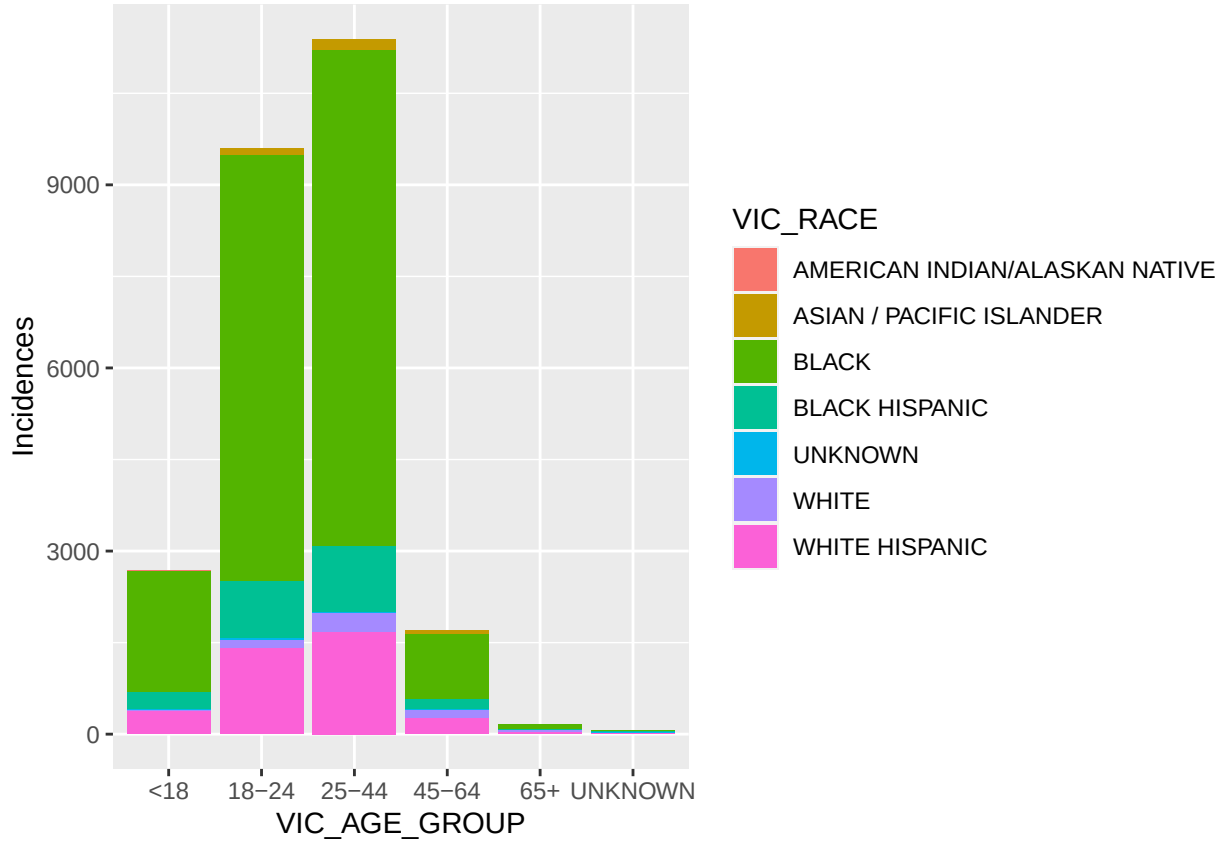
```
## `summarise()` has grouped output by 'VIC_AGE_GROUP'. You can override using the
## `.groups` argument.
## Using Incidences as value column: use value.var to override.
```

```
##   VIC_AGE_GROUP   F     M  U
## 1           <18 376  2305 NA
## 2         18-24 732  8867  4
## 3         25-44 914 10468  2
## 4         45-64 322  1376 NA
## 5           65+  54   113 NA
## 6       UNKNOWN   5    50  5
```

We see that most of the incidences are against a majority of the 18-44 male population. Which does display that there seems to be a similar distribution of age groups where these incidences occur. So it doesn't give us enough information so lets consider race as a potential factor.

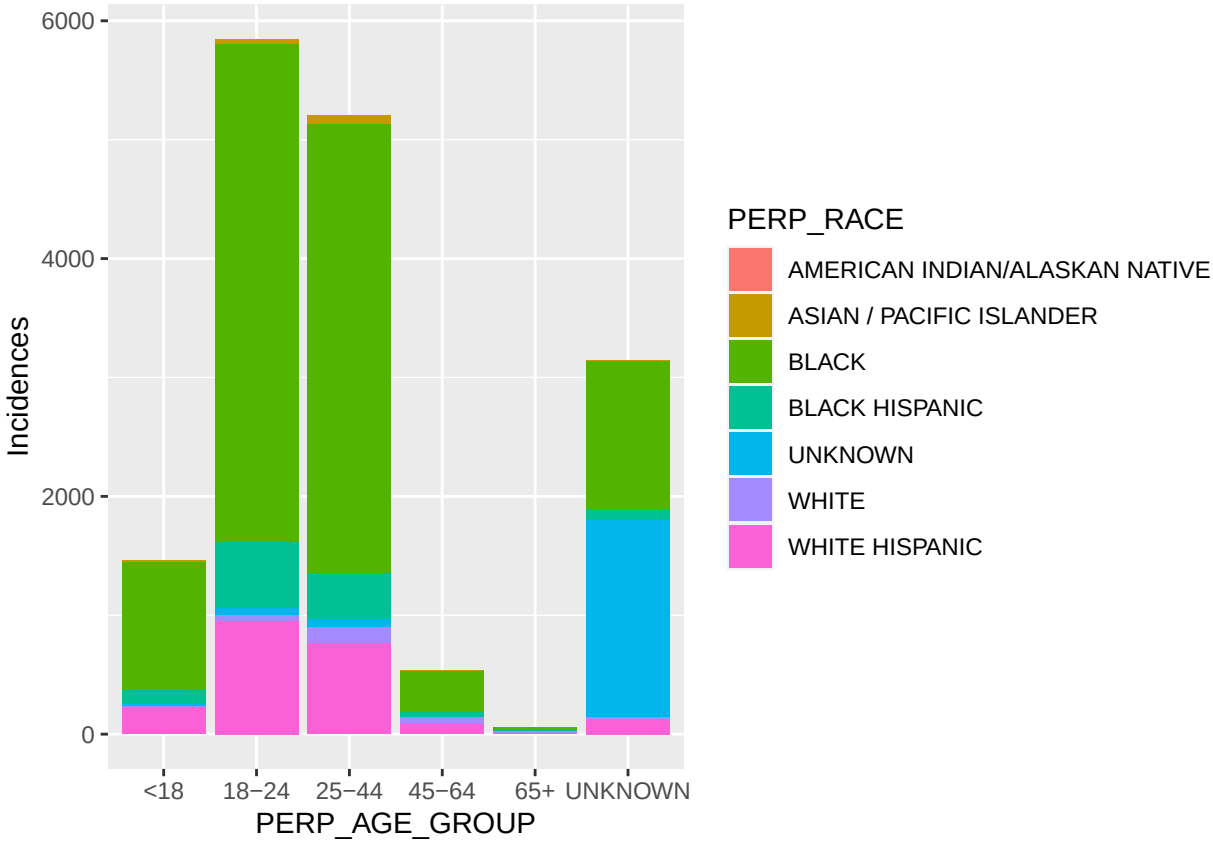
```
nypd_data_subset %>%
  filter(!VIC_AGE_GROUP %in% c(NA)) %>%
  group_by(VIC_AGE_GROUP, VIC_RACE) %>%
  summarize(Incidences = n()) %>%
  ggplot(aes(x=VIC_AGE_GROUP, fill=VIC_RACE, group=VIC_RACE,  y=Incidences)) + geom_col() +
scale_color_brewer(palette="Paired")
```

```
## `summarise()` has grouped output by 'VIC_AGE_GROUP'. You can override using the
## `.groups` argument.
```



There is a high proportion of Black victims, lets see if the proportions are similar for perpetrators within the dataset.

```
## `summarise()` has grouped output by 'PERP_AGE_GROUP'. You can override using
## the `.groups` argument.
```



It does appear that the only difference we see between the age distribution is that there more 18-24 perpetrators than there are 18-24 victims. One might say there is a bit of a signal that the perpetrators would be younger and more violent against older victims. However this is a rather rough way to really assess this especially in aggregate.

This information can be sliced and diced in many more various ways.. I would have liked to take some of the location information (LONG/LAT) and taken a further look at certain hot spots where these incidences occurred. One might be able to further assess if these locations are confined in certain places and thus could alternative solutions be done to help the violence subside.

If we could have access to more demographic information data, I think that there would be a different story about the data. Things like population density by block or socioeconomic prosperity data could help provide an alternative image into why the insights we just looked at may provide a new different story. That gives us the opportunity to look at bias in this dataset.

Bias

If we wanted to be Fox News in this case, we'd stop and say that there is a clear connection between the Black community (and other minority groups) and gun violence according to this data set. There are so many other questions that need to be asked and answered before one might want to make any sort of suggestion about a particular signal in the data. We'd need to ask many other questions such as:

- Are the precincts located in communities where the demographics are a higher proportion Black? If that is the case, is there crime that occurs in other places that isn't recorded due to not being around in those areas.
- Are different jurisdictions that do not fall into this data that could tell a different story?

- Perhaps there are equivalent incidences that occur that might not be gun shootings but rather knifes, and other instruments of violence that are not taken into consideration of this information?
- As mentioned earlier, is there a connection between socioeconomic status in these areas that may contribute to these incidence but unrelated to race.

I can see how easy to look at this information and automatically have prejudice. It is hard to overlook what appears to be hard fact. **Even throughout this assignment, you have internal moral questions on if you should even analyze data with race** as part of the analysis. Is there not a more appropriate way of displaying the same information?

```
sessionInfo()
```

```
## R version 4.2.0 (2022-04-22)
## Platform: aarch64-apple-darwin20 (64-bit)
## Running under: macOS Monterey 12.5.1
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.2-arm64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.2-arm64/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] reshape2_1.4.4  lubridate_1.8.0 forcats_0.5.1   stringr_1.4.0
##  [5] dplyr_1.0.9     purrr_0.3.4     readr_2.1.2     tidyr_1.2.0
##  [9] tibble_3.1.7    ggplot2_3.3.6   tidyverse_1.3.2
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.8.3        assertthat_0.2.1    digest_0.6.29
##  [4] utf8_1.2.2          R6_2.5.1            cellranger_1.1.0
##  [7] plyr_1.8.7          backports_1.4.1     reprex_2.0.1
## [10] evaluate_0.15       highr_0.9           httr_1.4.3
## [13] pillar_1.7.0        rlang_1.0.4         curl_4.3.2
## [16] googlesheets4_1.0.0 readxl_1.4.0        rstudioapi_0.13
## [19] rmarkdown_2.14      labeling_0.4.2      googledrive_2.0.0
## [22] bit_4.0.4           munsell_0.5.0       broom_1.0.0
## [25] compiler_4.2.0      modelr_0.1.8        xfun_0.31
## [28] pkgconfig_2.0.3     htmltools_0.5.2     tidyselect_1.1.2
## [31] fansi_1.0.3         crayon_1.5.1        tzdb_0.3.0
## [34] dbplyr_2.2.1        withr_2.5.0         grid_4.2.0
## [37] jsonlite_1.8.0      gtable_0.3.0        lifecycle_1.0.1
## [40] DBI_1.1.3           magrittr_2.0.3      scales_1.2.0
## [43] cli_3.3.0           stringi_1.7.6       vroom_1.5.7
## [46] farver_2.1.0        fs_1.5.2            xml2_1.3.3
## [49] ellipsis_0.3.2      generics_0.1.3      vctrs_0.4.1
## [52] RColorBrewer_1.1-3  tools_4.2.0         bit64_4.0.5
## [55] glue_1.6.2          hms_1.1.1           parallel_4.2.0
## [58] fastmap_1.1.0       yaml_2.3.5          colorspace_2.0-3
## [61] gargle_1.2.0        rvest_1.0.2         knitr_1.39
## [64] haven_2.5.0
```