NYPD Shooting Report

These following libraries should be imported prior to importing data and running report.

```
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
library(dplyr)
library(rgdal)
library(tmap)
library(tmaptools)
library(tigris)
```

Importing Data

I will start by reading and importing the data from the csv file. This data set is the NYPD Shooting data set from the data.gov catalog. We will start by looking at a summary of the data set.

```
## here is the file
url_in <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
nypd_shootings <- read_csv(url_in)
summary(nypd_shootings)
```

```
##   INCIDENT_KEY        OCCUR_DATE          OCCUR_TIME           BORO
##  Min.   :  9953245   Length:23568       Length:23568      Length:23568
##  1st Qu.: 55317014   Class :character   Class1:hms        Class :character
##  Median : 83365370   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :102218616                      Mode  :numeric
##  3rd Qu.:150772442
##  Max.   :222473262
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.0000    Length:23568       Mode :logical
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character   FALSE:19080
##  Median : 69.00   Median :0.0000    Mode  :character   TRUE :4488
##  Mean   : 66.21   Mean   :0.3323
##  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:23568       Length:23568       Length:23568       Length:23568
##  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:23568       Length:23568       Min.   : 914928   Min.   :125757
##  Class :character   Class :character   1st Qu.: 999900   1st Qu.:182565
##  Mode  :character   Mode  :character   Median :1007645   Median :193482
##                                        Mean   :1009363   Mean   :207312
##                                        3rd Qu.:1016807   3rd Qu.:239163
```

```
##                                        Max.   :1066815   Max.   :271128
##
##     Latitude       Longitude        Lon_Lat
##  Min.   :40.51   Min.   :-74.25   Length:23568
##  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
##
```

Tidying up Data / Transforming Data

Now let's tidy up the dataset to remove all NA's and Unknown data points. We'll do this by using the filter function.

```
nypd_shootings <- nypd_shootings %>% filter(PERP_SEX != "U")
nypd_shootings <- nypd_shootings %>% filter(VIC_SEX != "U")
nypd_shootings <- nypd_shootings %>% filter(LOCATION_DESC != "NA")
nypd_shootings <- na.omit(nypd_shootings)
nypd_shootings
```

```
## # A tibble: 6,246 x 19
##    INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO          PRECINCT JURISDICTION_CODE
##           <dbl> <chr>      <time>     <chr>            <dbl>             <dbl>
##  1    204192600 10/24/2019 00:52      STATEN ISLAND      121                 0
##  2    193694863 02/17/2019 03:00      QUEENS             114                 2
##  3    201436772 08/21/2019 23:34      STATEN ISLAND      120                 0
##  4    201852654 08/31/2019 07:42      BRONX               45                 0
##  5    193939359 02/24/2019 23:20      BRONX               44                 2
##  6    199247701 07/03/2019 00:04      QUEENS             114                 2
##  7    199134406 06/29/2019 05:48      BROOKLYN            69                 0
##  8    204971625 11/10/2019 14:03      BROOKLYN            63                 0
##  9    200365034 07/28/2019 14:35      MANHATTAN           30                 2
## 10    199422329 07/07/2019 10:50      BROOKLYN            60                 0
## # ... with 6,236 more rows, and 13 more variables: LOCATION_DESC <chr>,
## #   STATISTICAL_MURDER_FLAG <lgl>, PERP_AGE_GROUP <chr>, PERP_SEX <chr>,
## #   PERP_RACE <chr>, VIC_AGE_GROUP <chr>, VIC_SEX <chr>, VIC_RACE <chr>,
## #   X_COORD_CD <dbl>, Y_COORD_CD <dbl>, Latitude <dbl>, Longitude <dbl>,
## #   Lon_Lat <chr>
```

At this point, we have removed a lot of data that may be missing in the data set by removing all the incidents with unknown sex and locations. The way I handled this is that I filtered out the data and completely wiped out it from the data set so that it doesn't have an impact at all. This could heavily skew the correlations in the data. However, I had to do it in order to produce robust visualizations to gain insights on the majority of the data set.

Here we can see an aggregated percentage of perpetrators and victims by race and age group.

```
perp_aggr = nypd_shootings %>% group_by(PERP_RACE) %>% count()
perp_aggr$percentage = (perp_aggr$n / sum(perp_aggr$n) ) * 100
perp_aggr
```

```
## # A tibble: 7 x 3
## # Groups:   PERP_RACE [7]
##   PERP_RACE                          n percentage
##   <chr>                          <int>      <dbl>
## 1 AMERICAN INDIAN/ALASKAN NATIVE     1     0.0160
## 2 ASIAN / PACIFIC ISLANDER          58     0.929
## 3 BLACK                           4627    74.1
## 4 BLACK HISPANIC                   448     7.17
## 5 UNKNOWN                          170     2.72
## 6 WHITE                            140     2.24
## 7 WHITE HISPANIC                   802    12.8
```

```
vic_aggr = nypd_shootings %>% group_by(VIC_RACE) %>% count()
vic_aggr$percentage = (vic_aggr$n / sum(vic_aggr$n) ) * 100
vic_aggr
```

```
## # A tibble: 7 x 3
## # Groups:   VIC_RACE [7]
##   VIC_RACE                           n percentage
##   <chr>                          <int>      <dbl>
## 1 AMERICAN INDIAN/ALASKAN NATIVE     4     0.0640
## 2 ASIAN / PACIFIC ISLANDER         109     1.75
## 3 BLACK                           4260    68.2
## 4 BLACK HISPANIC                   602     9.64
## 5 UNKNOWN                           26     0.416
## 6 WHITE                            229     3.67
## 7 WHITE HISPANIC                  1016    16.3
```

```
perp_age_group = nypd_shootings %>% group_by(PERP_AGE_GROUP) %>% count()
perp_age_group$percentage = (perp_age_group$n / sum(perp_age_group$n) ) * 100
perp_age_group
```

```
## # A tibble: 9 x 3
## # Groups:   PERP_AGE_GROUP [9]
##   PERP_AGE_GROUP     n percentage
##   <chr>          <int>      <dbl>
## 1 <18              558     8.93
## 2 1020               1     0.0160
## 3 18-24           2444    39.1
## 4 224                1     0.0160
## 5 25-44           2178    34.9
## 6 45-64            239     3.83
## 7 65+               37     0.592
## 8 940                1     0.0160
## 9 UNKNOWN          787    12.6
```

```
vic_age_group = nypd_shootings %>% group_by(VIC_AGE_GROUP) %>% count()
vic_age_group$percentage = (vic_age_group$n / sum(vic_age_group$n) ) * 100
vic_age_group
```

```
## # A tibble: 6 x 3
## # Groups:   VIC_AGE_GROUP [6]
```

```
##   VIC_AGE_GROUP     n percentage
##   <chr>         <int>      <dbl>
## 1 <18             671     10.7
## 2 18-24          2266     36.3
## 3 25-44          2735     43.8
## 4 45-64           491      7.86
## 5 65+              58      0.929
## 6 UNKNOWN          25      0.400
```

Let's dive deeper into the data set

During the transformation phase, I realized that there were commonalities in the race and age groups for both perpetrators and victims with the ages 18-24 and 25-44 being the most dense percentages in the data set. Additionally, the highest percentage of race for both perpetrators and victims were "Black" with 74% and 68%, with "White Hispanic" coming second for both with 12% and 16% for both perpetrators and victims. With that data, I wanted to dive deeper into the top percentage of locations that the incidents happened in the entire data set. With this information, we could get more information to a come to a correlation.

```
incident_location = nypd_shootings %>% group_by(LOCATION_DESC) %>% count()
incident_location$percentage = (incident_location$n / sum(incident_location$n) ) * 100
incident_location <- incident_location[order(-incident_location$percentage),]
incident_location
```

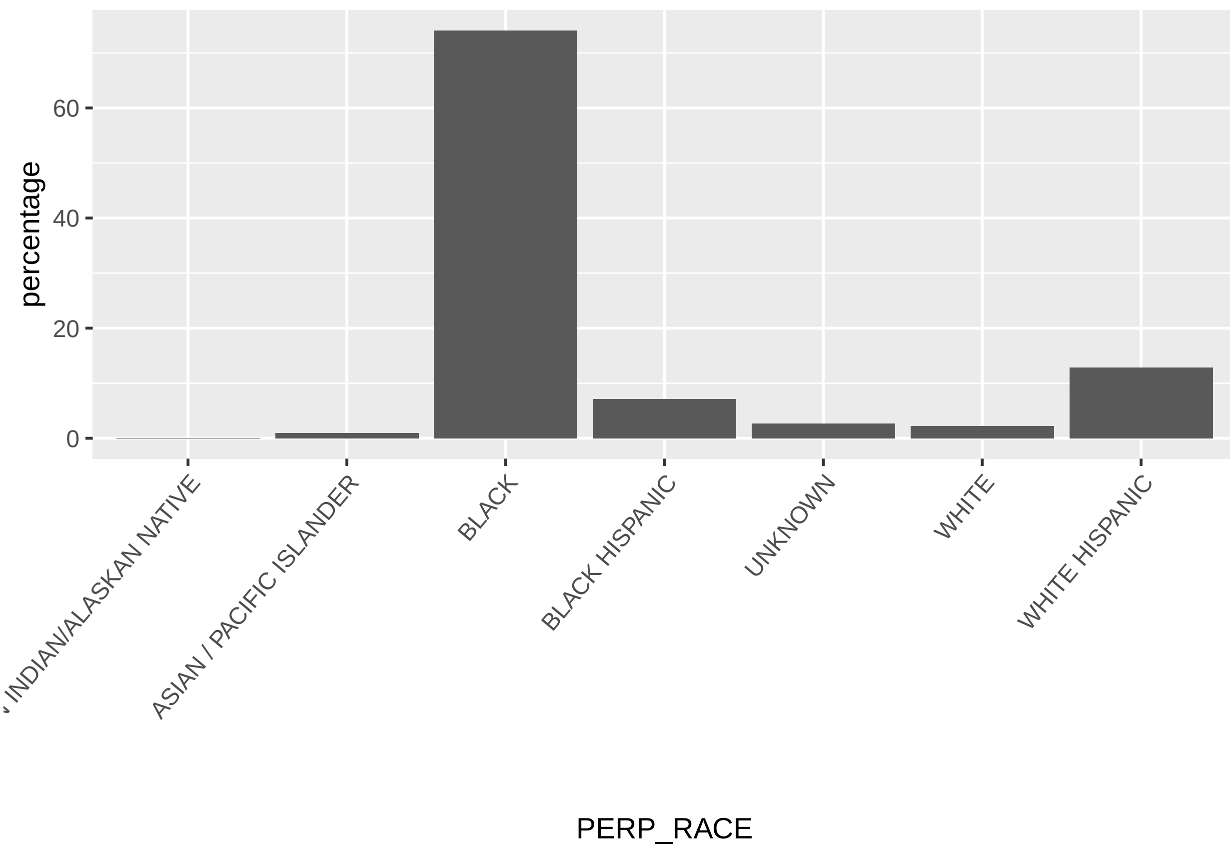
```
## # A tibble: 36 x 3
## # Groups:   LOCATION_DESC [36]
##    LOCATION_DESC                 n percentage
##    <chr>                     <int>      <dbl>
##  1 MULTI DWELL - PUBLIC HOUS  2372     38.0
##  2 MULTI DWELL - APT BUILD    1792     28.7
##  3 PVT HOUSE                   547      8.76
##  4 BAR/NIGHT CLUB              370      5.92
##  5 GROCERY/BODEGA              367      5.88
##  6 NONE                        144      2.31
##  7 COMMERCIAL BLDG             127      2.03
##  8 RESTAURANT/DINER            125      2.00
##  9 BEAUTY/NAIL SALON            69      1.10
## 10 FAST FOOD                    55      0.881
## # ... with 26 more rows
```

Here we can see the top 10 locations for all incidents with Multi Dwell - Public Housing and Multi Dwell - Apt Buildings being respectively 1 and 2. My hypothesis at first was that the bar / night club would be the #1 top location. However, now my intuition tells me that the high percentage rates are in low-income housing neighborhoods.

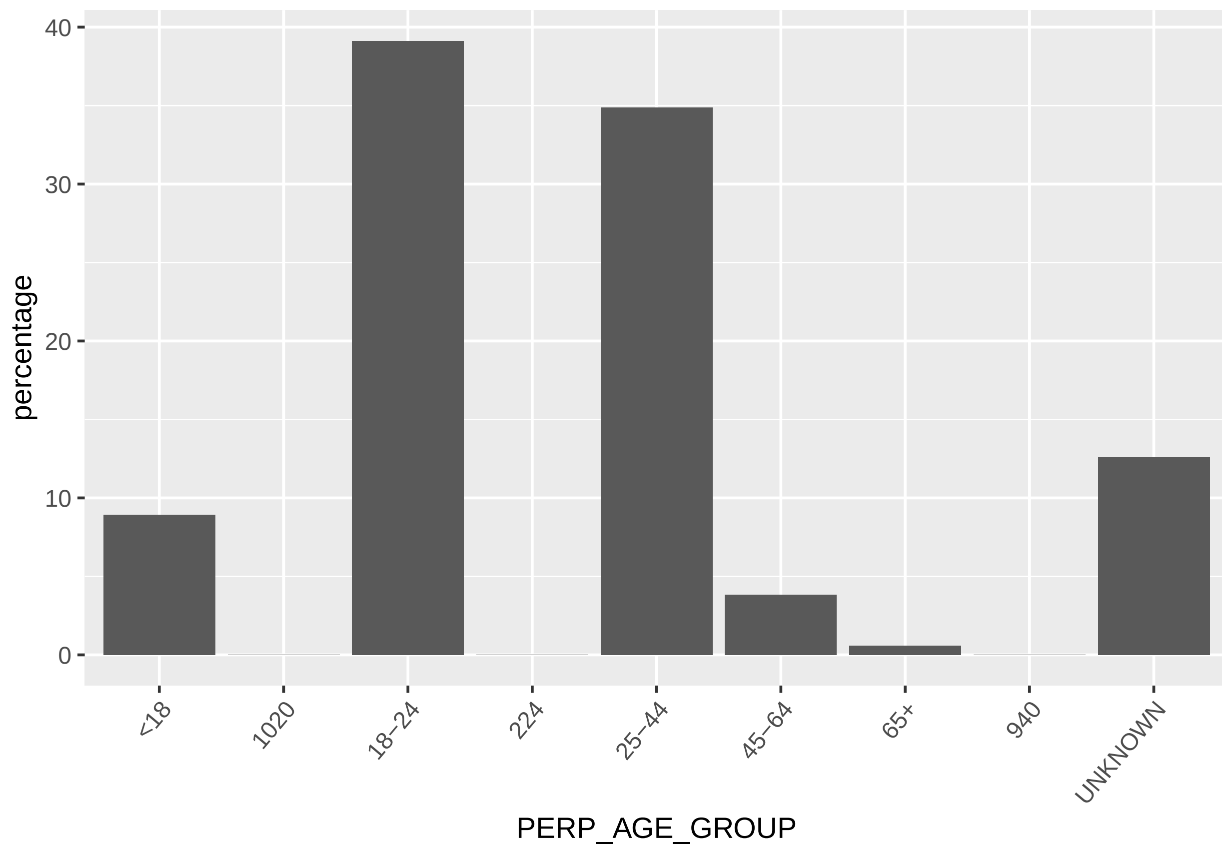
Visualization and Analysis - Modeling the Data

Now we will visualize all the data in graphs. During the exploration phase, I realized that both perpetrators and victims biggest age groups are 18-24 and 25-44 and the biggest races are Black and White Hispanic. With that being said, I decided to look further into the location and not look at graphs for victims as I wanted to just dive deeper into statistics for perpetrators. Because both groups have similar statistics, we could identify that both races and age groups could correlate to a specific location.

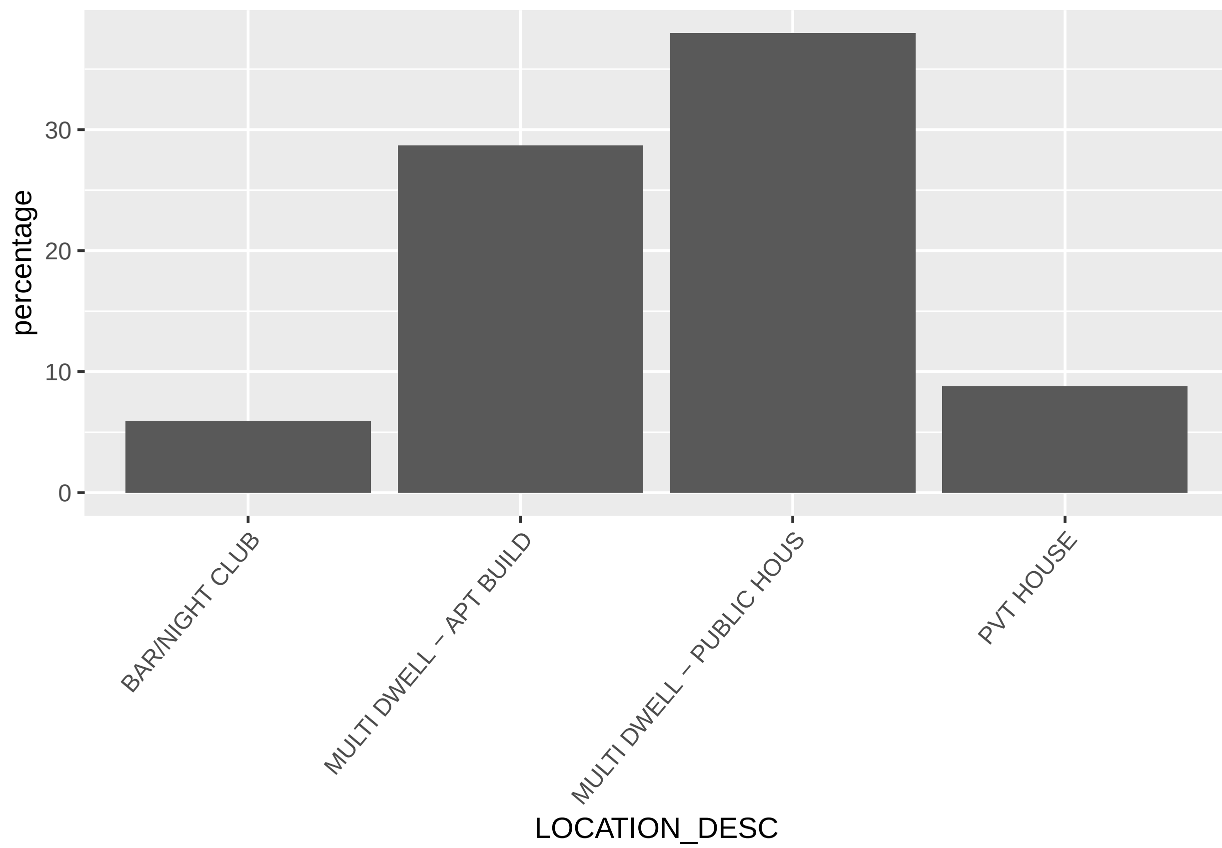
```r
perp_aggr %>%
  ggplot(aes(x = PERP_RACE, y=percentage)) +
  geom_bar(stat="identity") +
  theme(axis.text.x = element_text(angle = 50, hjust = 1))
```




```r
perp_age_group %>%
  ggplot(aes(x = PERP_AGE_GROUP, y=percentage)) +
  geom_bar(stat="identity") +
  theme(axis.text.x = element_text(angle = 50, hjust = 1))
```




```
incident_location[1:4,] %>%
  ggplot(aes(x = LOCATION_DESC, y=percentage)) +
  geom_bar(stat="identity") +
  theme(axis.text.x = element_text(angle = 50, hjust = 1))
```



Analyzing the data

Looking into the data, some questions that I asked myself were:

1. Does the missing unknown data have a great impact on the data? Or would it have a linear regression?
2. Are there any important locations that are missing and not in this data set?

Being able to ask myself these questions during the analysis phase, I was able to uncover that while this data may show good enough correlations to be able to make an inference that the *"Black"* and *"White Hispanic"* races who tend to live in low-income housing may be the majority of victims and perpetrators in NY. However, upon further analysis, I would love to investigate if the data set includes all the different locations and ensure the missing data in the data set indeed falls under a **linear regression**.

Bias Identification and Conclusion

The potential biases in the data:

- skewed data
- missing locations / non-updated entries

Skewed data is a potenial bias because the missing data points could potentially not be in a linear regression and could change the average of the correlations. For example, every unknown or missing race could have been "Asian" or "White" or every unknown age group could have been 65+ - which would have skewed the data by a lot. Another potential bias could be that there are missing locations or some columns are not up-to-date. In future iterations, the way that I would mitigate the skewed data and non-updated entries is by using a more complex transformation like imputing the data and taking each missing input and making it the average of the data set.