DS_Week3_NYPD

2023-08-08

Start an Rmd Document

We can start this project by loading in a few libraries and then by loading in our data. The dataset comes from the cityofnewyork.us website as provided by the course and contains a record of information on shootings in NYC. The goal of this project will be to look into which people are most likely to be victims in NYC shootings.

```
library(ggplot2)
library(lessR)
library(lubridate)
library("tidyverse")
```

```
file_url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/"
file_names <- c("rows.csv")
urls <- str_c(file_url, file_names)
nypd_data <- read_csv(urls)
```

```
summary(nypd_data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:27312       Length:27312       Length:27312
##  1st Qu.: 63860880   Class :character   Class1:hms         Class :character
##  Median : 90372218   Mode  :character   Class2:difftime    Mode  :character
##  Mean   :120860536                      Mode  :numeric
##  3rd Qu.:188810230
##  Max.   :261190187
##
##  LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:27312       Min.   :  1.00   Min.   :0.0000    Length:27312
##  Class :character   1st Qu.: 44.00   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 68.00   Median :0.0000    Mode  :character
##                     Mean   : 65.64   Mean   :0.3269
##                     3rd Qu.: 81.00   3rd Qu.:0.0000
##                     Max.   :123.00   Max.   :2.0000
##                                      NA's   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:27312       Mode :logical           Length:27312
##  Class :character   FALSE:22046             Class :character
##  Mode  :character   TRUE :5266              Mode  :character
##
##
##
```

```
##
##    PERP_SEX          PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  Length:27312       Length:27312       Length:27312       Length:27312
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_RACE           X_COORD_CD        Y_COORD_CD        Latitude
##  Length:27312       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000029   1st Qu.:182834   1st Qu.:40.67
##  Mode  :character   Median :1007731   Median :194487   Median :40.70
##                     Mean   :1009449   Mean   :208127   Mean   :40.74
##                     3rd Qu.:1016838   3rd Qu.:239518   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA's   :10
##    Longitude        Lon_Lat
##  Min.   :-74.25   Length:27312
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :10
```

Tidy and Transform Data

I can remove a good of columns that contain information I will not need. Mainly fields such as Longitude and Latitude, I can also remove columns with specific NYPD codes such as Jurisdiction Codes and Precinct. Having information about which Borough the crime happened in is enough for this project. We can also see from the summary above that the Occur Date is a string, we can transform that into a proper Date object.

```
nypd_data <- nypd_data %>% select(-c(Latitude, Longitude, Lon_Lat, X_COORD_CD,
                                     Y_COORD_CD, INCIDENT_KEY))
nypd_data <- nypd_data %>% select(-c(PRECINCT, JURISDICTION_CODE, LOCATION_DESC))
nypd_data$OCCUR_DATE <- mdy(nypd_data$OCCUR_DATE)
```

```
summary(nypd_data)
```

```
##    OCCUR_DATE          OCCUR_TIME           BORO           LOC_OF_OCCUR_DESC
##  Min.   :2006-01-01   Length:27312      Length:27312       Length:27312
##  1st Qu.:2009-07-18   Class1:hms        Class :character   Class :character
##  Median :2013-04-29   Class2:difftime   Mode  :character   Mode  :character
##  Mean   :2014-01-06   Mode  :numeric
##  3rd Qu.:2018-10-15
##  Max.   :2022-12-31
##  LOC_CLASSFCTN_DESC STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:27312       Mode :logical           Length:27312
##  Class :character   FALSE:22046             Class :character
##  Mode  :character   TRUE :5266              Mode  :character
##
```

```
##
##
##    PERP_SEX          PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  Length:27312       Length:27312       Length:27312       Length:27312
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    VIC_RACE
##  Length:27312
##  Class :character
##  Mode  :character
##
##
##
```

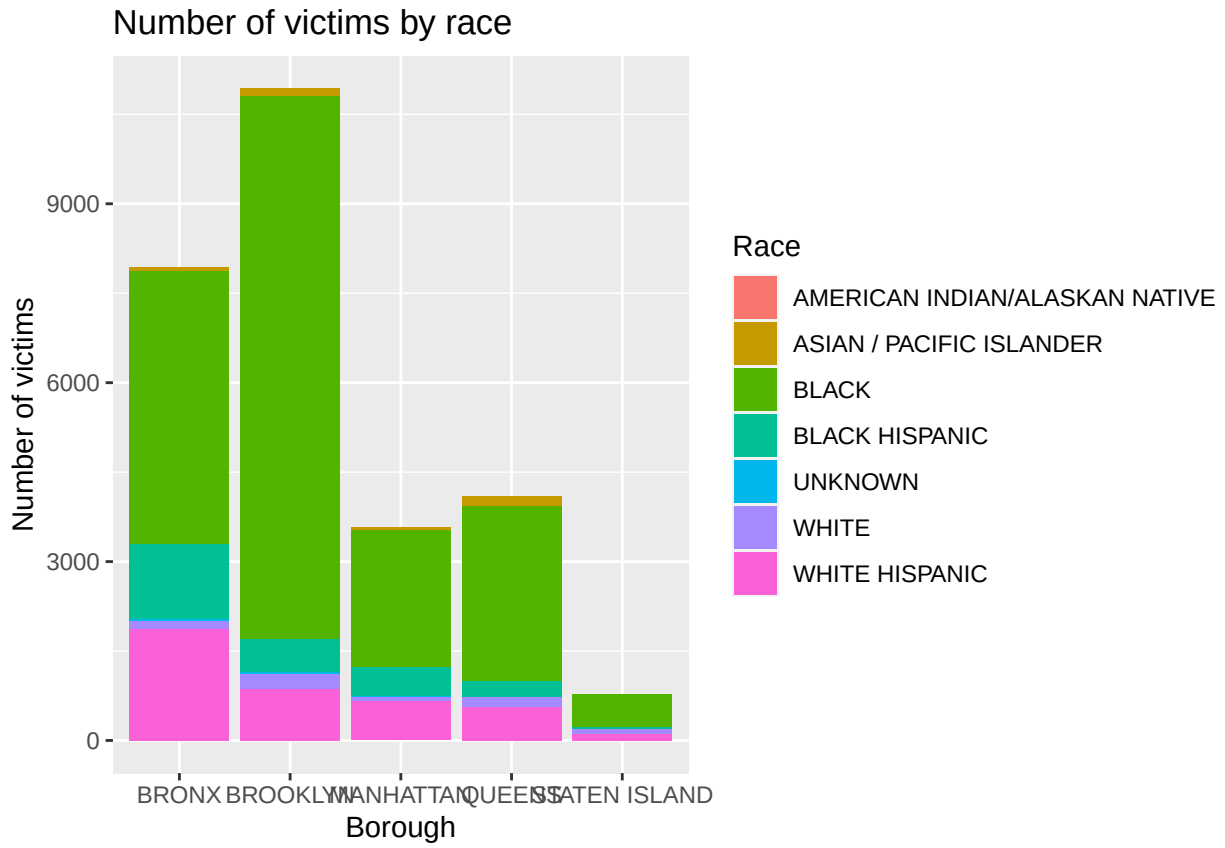
```
colMeans_df <- stack((colMeans(is.na(nypd_data)))*100)
plot1 <- ggplot(colMeans_df, aes(y=ind,x=values)) + geom_col() +
  labs(title="Missing Values", y="Column", x="% Missing Values")
plot1
```

We can also see that over 75% of values are missing from LOC_CLASSFCTN_DESC and LOC_OF_OCCUR_DESC, so we can remove those too. We can also see that information is missing for over 25% of perpetrators, however we will leave the missing values and will not be augmenting any data in this project.

```
nypd_data <- nypd_data %>% select(-c(LOC_CLASSFCTN_DESC, LOC_OF_OCCUR_DESC))
```

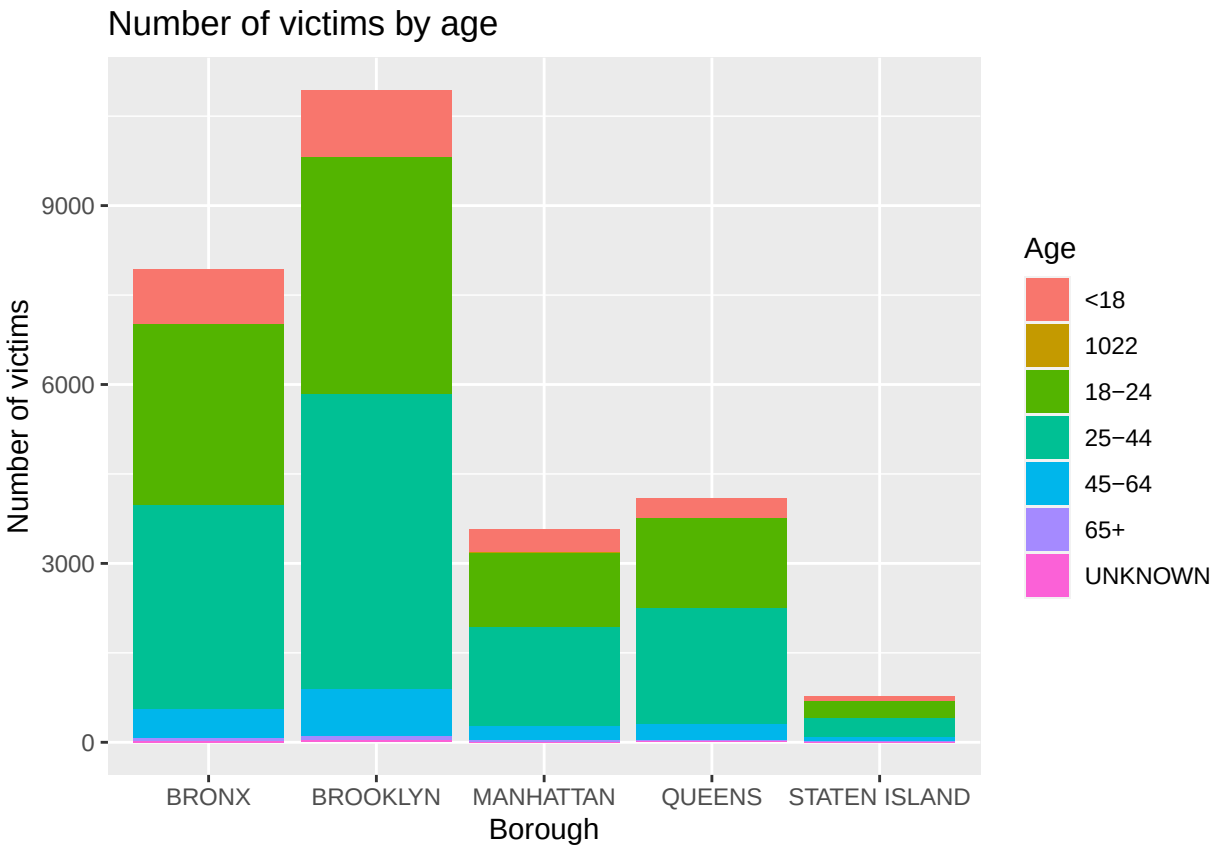
Visualization and Analysis

```
ggplot(nypd_data, aes(x=BORO, fill=VIC_RACE)) + geom_bar() +
  labs(title='Number of victims by race', x='Borough', y='Number of victims', fill='Race')
```



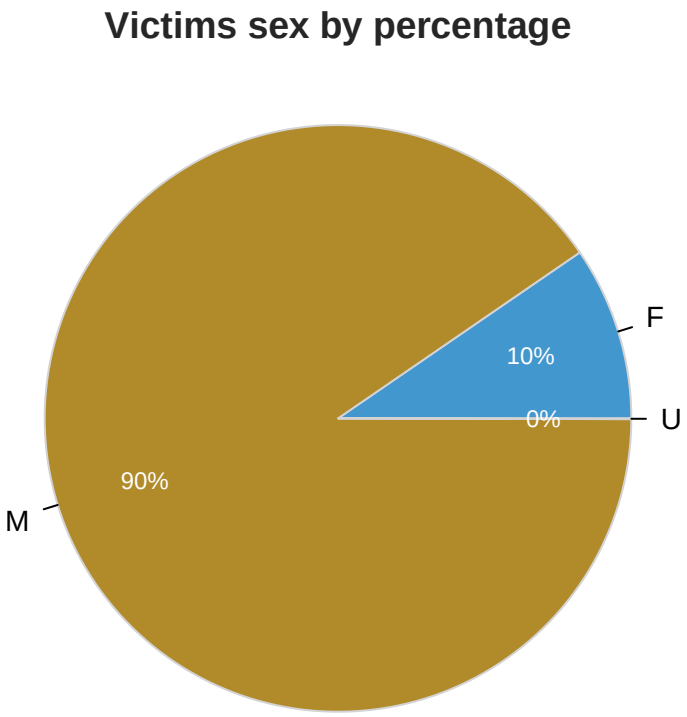
In this graph we can see black victims are the most common. It would be nice if this dataset contained information about the population in NYC so we could compare the percentage of black citizens in these areas to the percent of black victims, as well as the other races.

```
ggplot(nypd_data, aes(x=BORO, fill=VIC_AGE_GROUP)) + geom_bar() +
  labs(title='Number of victims by age', x='Borough', y='Number of victims', fill='Age')
```



This graph shows the number of victims by their age group. Victims aged 25 to 44 appear to be the most common.

```
VicSex_tb <- table(nypd_data$VIC_SEX)
PieChart(VicSex_tb, hole=0, values="%", main="Victims sex by percentage")
```



This pie chart shows the percentage of victims by their sex. Assuming NYC has a 50/50 split between males and females; male victims are overwhelmingly more likely than female victims.

```
grouped_tb <- nypd_data %>%
  group_by(VIC_RACE, VIC_AGE_GROUP) %>%
  summarise(total_count=n(),.groups = 'drop')

model <- lm(total_count ~ VIC_RACE + VIC_AGE_GROUP, data=grouped_tb)
summary(model)
```

```
##
## Call:
## lm(formula = total_count ~ VIC_RACE + VIC_AGE_GROUP, data = grouped_tb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2502.2  -642.0    23.3   452.2  4268.1
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       -791.4      944.6  -0.838 0.409784
## VIC_RACEASIAN / PACIFIC ISLANDER   734.1     1049.4   0.700 0.490402
## VIC_RACEBLACK                     3906.4     1049.4   3.723 0.000959 ***
## VIC_RACEBLACK HISPANIC            1107.8     1049.4   1.056 0.300829
## VIC_RACEUNKNOWN                    505.6     1075.5   0.470 0.642163
## VIC_RACEWHITE                      783.1     1049.4   0.746 0.462196
```

```
## VIC_RACEWHITE HISPANIC            1341.6     1049.4   1.279 0.212366
## VIC_AGE_GROUP1022                -3114.1     1639.3  -1.900 0.068634 .
## VIC_AGE_GROUP18-24                1035.3      771.3   1.342 0.191097
## VIC_AGE_GROUP25-44                1348.9      771.3   1.749 0.092109 .
## VIC_AGE_GROUP45-64                -294.6      813.0  -0.362 0.720039
## VIC_AGE_GROUP65+                  -747.0      862.7  -0.866 0.394450
## VIC_AGE_GROUPUNKNOWN              -594.9      813.0  -0.732 0.470870
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1443 on 26 degrees of freedom
## Multiple R-squared:  0.5664, Adjusted R-squared:  0.3663
## F-statistic: 2.831 on 12 and 26 DF,  p-value: 0.01279
```

Here we try to use the victims race and age group as predictors for how likely a person is to be a victim of a shooting. From this we can see that black victims between the age of 25 and 44 are the most likely to be victims. This matches up with the visualizations we saw earlier.

Bias Identification

Addressing bias is very important in any data science project, especially a project with political implications such as this one. As someone who lives in the NYC area it can be very easy for me to feel like I should represent this data in a more positive light, leading to bias. However, I corrected this bias by treating it like I would any other data science project and by not focusing on where the data came from and just focusing on how to represent what the data is showing us. I believe I handled and presented this data in an unbiased way, letting the data speak for itself.