NYPD Shooting Incidents

Saurabh Sant

13/07/2021

Data Source: https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic

We will working with NYPD Shooting Incident dating back to 2006. It contains information about the timing, Boro, Precinct, perp and victim information. Each record represents a shooting incident in New York City.

We can use this data to answer the following questions: Which Boro had the most incidents? Which age group of a perp were most involved in the shooting?

Data Cleaning and Transformation

Import the necessary libraries

```
library(stringr)
library(readr)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.7
## v tidyr   1.1.3     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

import the data

```
NYPD_Shooting <-read_csv("~/Downloads/NYPD_Shooting_Incident_Data__Historic_.csv")
```

```
##
## -- 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()
## )
```

```
head(NYPD_Shooting)
```

```
## # A tibble: 6 x 19
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO          PRECINCT JURISDICTION_CODE
##          <dbl> <chr>      <time>     <chr>            <dbl>             <dbl>
## 1    201575314 08/23/2019 22:10      QUEENS             103                 0
## 2    205748546 11/27/2019 15:54      BRONX               40                 0
## 3    193118596 02/02/2019 19:40      MANHATTAN           23                 0
## 4    204192600 10/24/2019 00:52      STATEN ISLAND      121                 0
## 5    201483468 08/22/2019 18:03      BRONX               46                 0
## 6    198255460 06/07/2019 17:50      BROOKLYN            73                 0
## # ... with 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>
```

Data cleaning

After manually looking through the data, I would like to tidy up the dataset. I don't need lat and long for the analysis I am planning, so I will get rid of those and rename Region and State to be more R friendly.

```
NYPD_Shooting <- NYPD_Shooting %>% mutate(OCCUR_DATE = mdy(OCCUR_DATE))%>%
  select(-c(X_COORD_CD, Y_COORD_CD,Latitude, Longitude))
```

```
head(NYPD_Shooting)
```

```
## # A tibble: 6 x 15
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO          PRECINCT JURISDICTION_CODE
##          <dbl> <date>     <time>     <chr>            <dbl>             <dbl>
## 1    201575314 2019-08-23 22:10      QUEENS             103                 0
## 2    205748546 2019-11-27 15:54      BRONX               40                 0
## 3    193118596 2019-02-02 19:40      MANHATTAN           23                 0
## 4    204192600 2019-10-24 00:52      STATEN ISLAND      121                 0
## 5    201483468 2019-08-22 18:03      BRONX               46                 0
## 6    198255460 2019-06-07 17:50      BROOKLYN            73                 0
## # ... with 9 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>,
## #   Lon_Lat <chr>
```

Lets find out how many missing values can be found in our data:

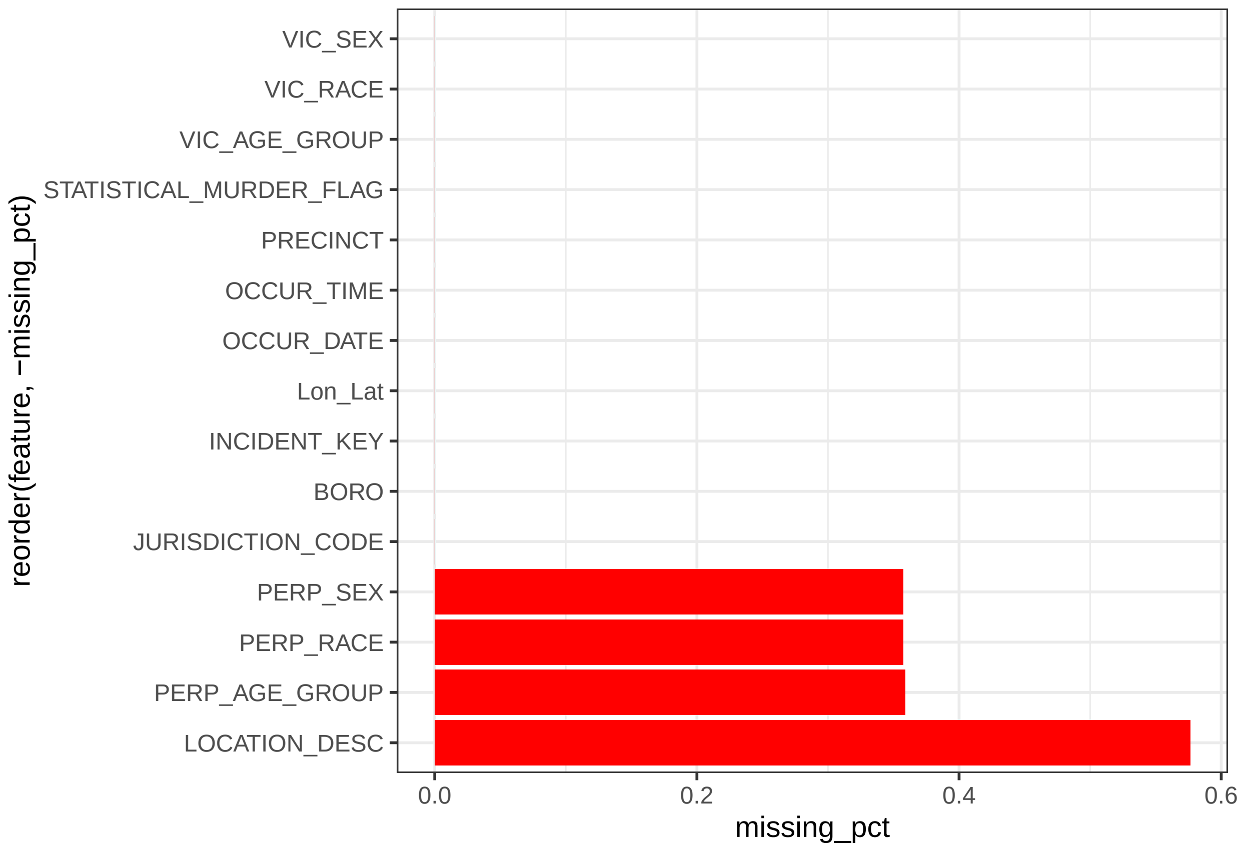
```
missing_values <- NYPD_Shooting %>% summarize_each(funs(sum(is.na(.))/n()))
```

```
## Warning: ‘summarise_each_()‘ was deprecated in dplyr 0.7.0.
## Please use ‘across()‘ instead.
```

```
## Warning: ‘funs()‘ was deprecated in dplyr 0.8.0.
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with ‘tibble::lst()‘:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
```

It is a good idea to see how many values are missing to see what kind of analysis we can do. As you can see we don't have much for location_desc so we will drop this as well.

```
missing_values <- gather(missing_values, key="feature", value="missing_pct")
missing_values %>%
ggplot(aes(x=reorder(feature,-missing_pct),y=missing_pct)) +
geom_bar(stat="identity",fill="red")+
coord_flip()+theme_bw()
```



```
NYPD_Shooting <- NYPD_Shooting %>% mutate(OCCUR_DATE = mdy(OCCUR_DATE))%>%
  select(-c(LOCATION_DESC))
```

```
## Warning: All formats failed to parse. No formats found.
```

Data Analysis & Visualization

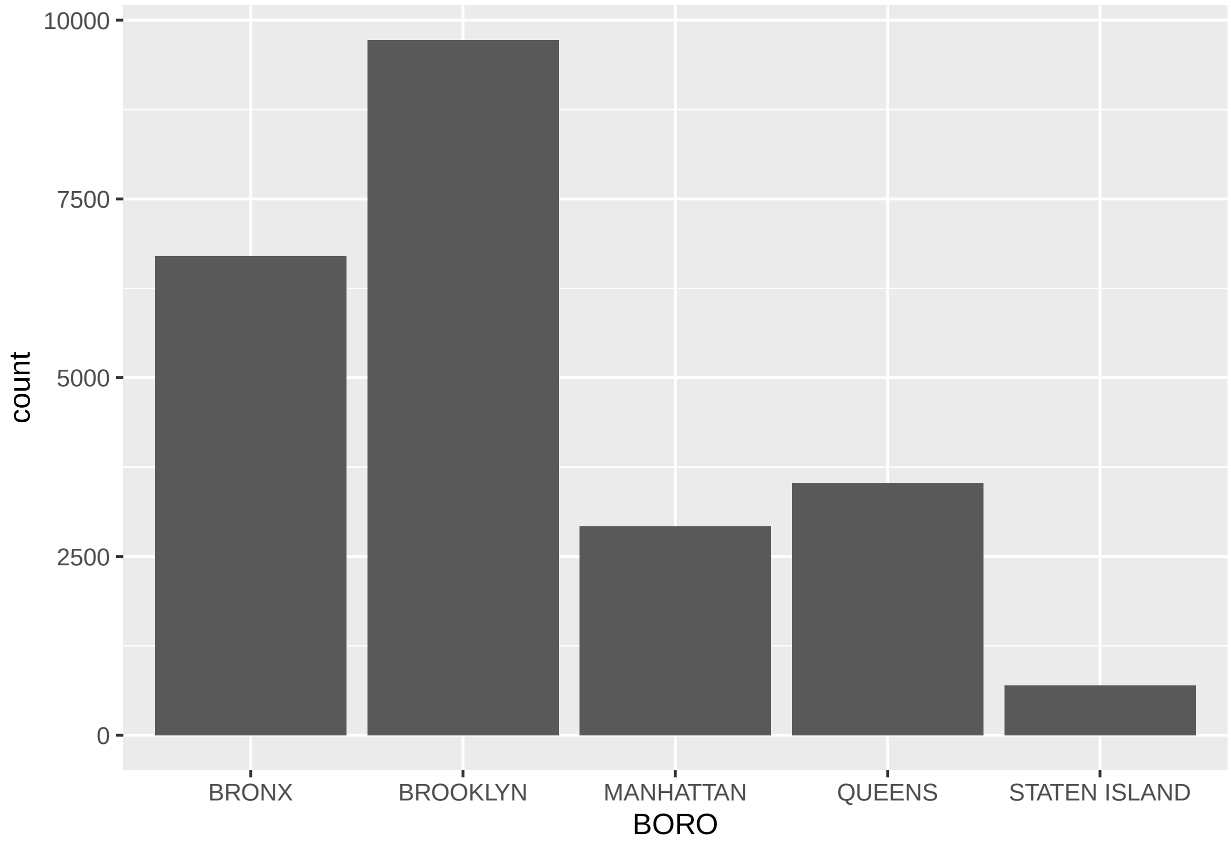
Let us find out which Boro in New York City had the most shootings. From the graph below, we can see that Brooklyn has the most shootings with 9722 and second most are in Bronx with a count of 6700.

```
summary_shooting <- NYPD_Shooting %>% count(BORO)
```

```
head(summary_shooting)
```

```
## # A tibble: 5 x 2
##   BORO              n
##   <chr>         <int>
## 1 BRONX          6700
## 2 BROOKLYN       9722
## 3 MANHATTAN      2921
## 4 QUEENS         3527
## 5 STATEN ISLAND   698
```

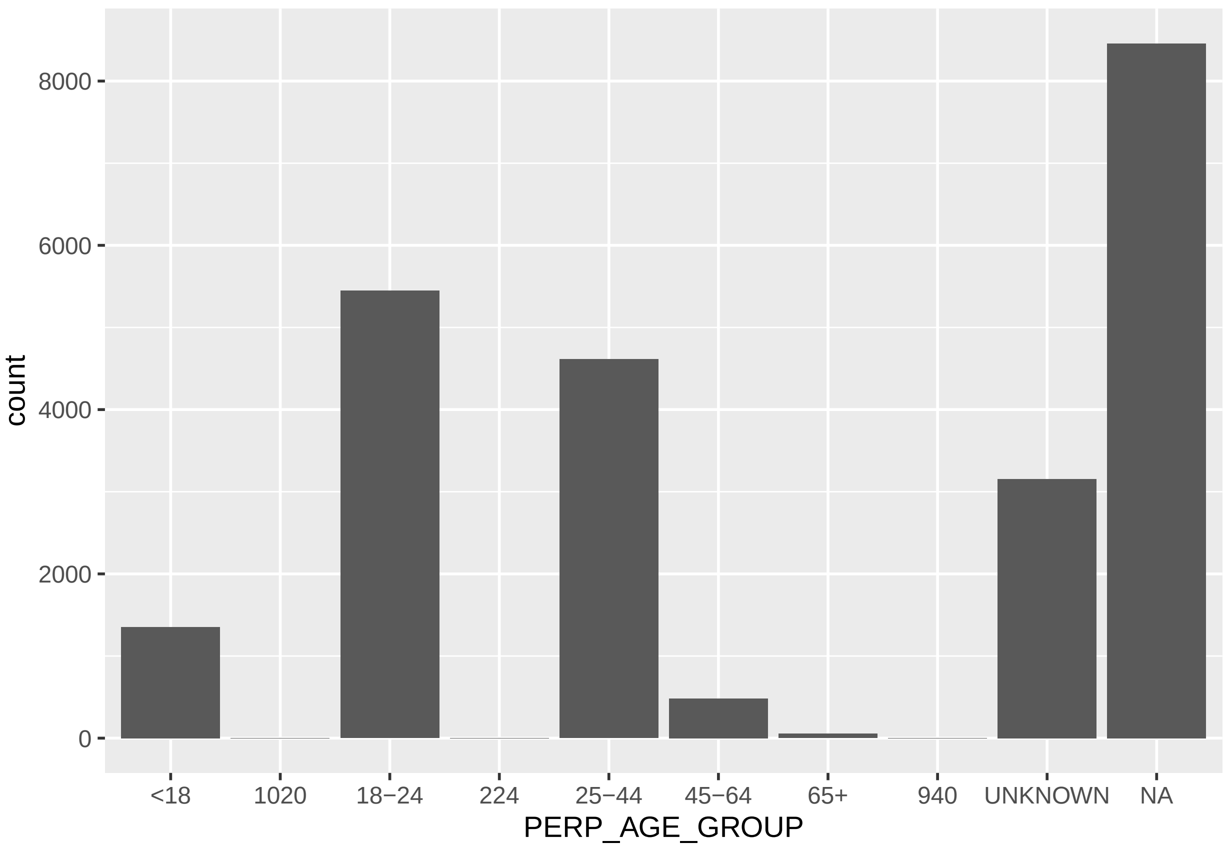
```
ggplot(data = NYPD_Shooting, aes(x = BORO), color = blues9) +
  geom_bar()
```



We can find out what the age group of the people involved in the shootings are. We can see that most of the perps are young in the age group of 18-24 with a count of 5,448 however most of the values for age_group is missing.

```
summary_age <- NYPD_Shooting %>% count(PERP_AGE_GROUP)
```

```
ggplot(data = NYPD_Shooting, aes(x = PERP_AGE_GROUP), color = blues9) +
  geom_bar()
```



Bias

There is a lot of missing values for victim and perp race so I will be dropping those values and using only records that have the full information so my data model is more accurate.

Data Modelling

We will model the data using Decision Tree to determine the liklihood of a crime happening in each Boro of New York City. I will remove missing values and use values with enough data for perp race and victim race.

```
modelling_data <- NYPD_Shooting %>% filter(VIC_RACE %in% c('BLACK','BLACK HISPANIC', 'WHITE', 'WHITE HIS
```

```
modelling_data <- modelling_data %>% filter(PERP_RACE %in% c('BLACK','BLACK HISPANIC', 'WHITE', 'WHITE
```

```
summary_sample <- modelling_data %>% count(VIC_RACE)
```

```
summary_sample
```

```
## # A tibble: 5 x 2
##   VIC_RACE                     n
##   <chr>                    <int>
## 1 ASIAN / PACIFIC ISLANDER   221
## 2 BLACK                     8966
## 3 BLACK HISPANIC            1332
## 4 WHITE                      436
## 5 WHITE HISPANIC            2257
```

```
library(rpart)
```

```
## Warning: package 'rpart' was built under R version 4.1.2
```

```
library(rpart.plot)
```

```
## Warning: package 'rpart.plot' was built under R version 4.1.2
```

```
library(tree)
```

```
## Warning: package 'tree' was built under R version 4.1.2
```

```
## Registered S3 method overwritten by 'tree':
##   method     from
##   print.tree cli
```

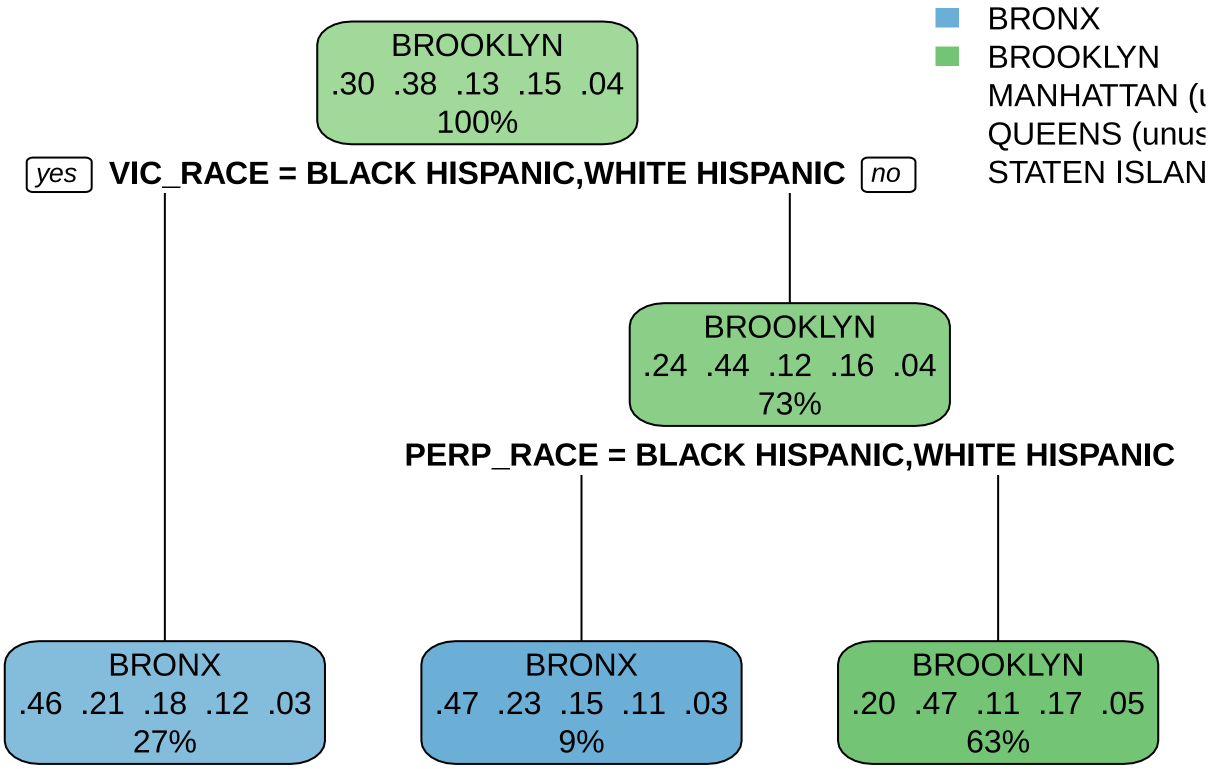
```
head(NYPD_Shooting)
```

```
## # A tibble: 6 x 14
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO          PRECINCT JURISDICTION_CODE
##          <dbl> <date>     <time>     <chr>            <dbl>             <dbl>
## 1    201575314 NA         22:10      QUEENS             103                 0
## 2    205748546 NA         15:54      BRONX               40                 0
## 3    193118596 NA         19:40      MANHATTAN           23                 0
## 4    204192600 NA         00:52      STATEN ISLAND      121                 0
## 5    201483468 NA         18:03      BRONX               46                 0
## 6    198255460 NA         17:50      BROOKLYN            73                 0
## # ... with 8 more variables: STATISTICAL_MURDER_FLAG <lgl>,
## #   PERP_AGE_GROUP <chr>, PERP_SEX <chr>, PERP_RACE <chr>, VIC_AGE_GROUP <chr>,
## #   VIC_SEX <chr>, VIC_RACE <chr>, Lon_Lat <chr>
```

```
model = rpart(BORO ~ PERP_RACE + VIC_RACE + VIC_AGE_GROUP + PERP_AGE_GROUP, data = modelling_data)
```

From the diagram below, we observe that most crimes happen in Bronx and Brooklyn. If the victim race is black hispanic or white hispanic then the probability of the shooting being in Bronx is 27%.

```
rpart.plot(model)
```



Conclusion

Based on the modelling, we can see that most shootings occur in Bronx and Brooklyn. This information can useful when coordinating police officers to patrol certain areas or for areas to target to reduce crime.