Producing Reproducible Rmd Document Analyzing NYPD Shooting Incident Data (Historical)

According to Data.Gov, this is a: >List of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year.

> [...] 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.

Libaries Used

```
library(readr)
library(reshape)
library(dplyr)
library(ggplot2)
library(lubridate)
```

Reading in CSV data from City of New York

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

```
## 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.
```

Summary of Data

There are 25596 rows and 19 columns. Many of the columns will not be used in the visualization and analysis so I will later remove them from the data frame.

```
summary(historical_data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:25596       Length:25596       Length:25596
##  1st Qu.: 61593633   Class :character   Class1:hms         Class :character
##  Median : 86437258   Mode  :character   Class2:difftime    Mode  :character
##  Mean   :112382648                      Mode  :numeric
##  3rd Qu.:166660833
##  Max.   :238490103
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.0000    Length:25596       Mode :logical
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character   FALSE:20668
##  Median : 69.00   Median :0.0000    Mode  :character   TRUE :4928
##  Mean   : 65.87   Mean   :0.3316
##  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:25596       Length:25596       Length:25596       Length:25596
##  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:25596       Length:25596       Min.   : 914928   Min.   :125757
##  Class :character   Class :character   1st Qu.:1000011   1st Qu.:182782
##  Mode  :character   Mode  :character   Median :1007715   Median :194038
##                                        Mean   :1009455   Mean   :207894
##                                        3rd Qu.:1016838   3rd Qu.:239429
##                                        Max.   :1066815   Max.   :271128
##
##     Latitude       Longitude        Lon_Lat
##  Min.   :40.51   Min.   :-74.25   Length:25596
##  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
##
```

Removing columns that we aren't using

For further visualization and analysis I will be keeping OCCUR_DATE and BORO. The type of OCCUR_DATE is currently so before any visualizations I will mutate it to a data object first using lubridate.

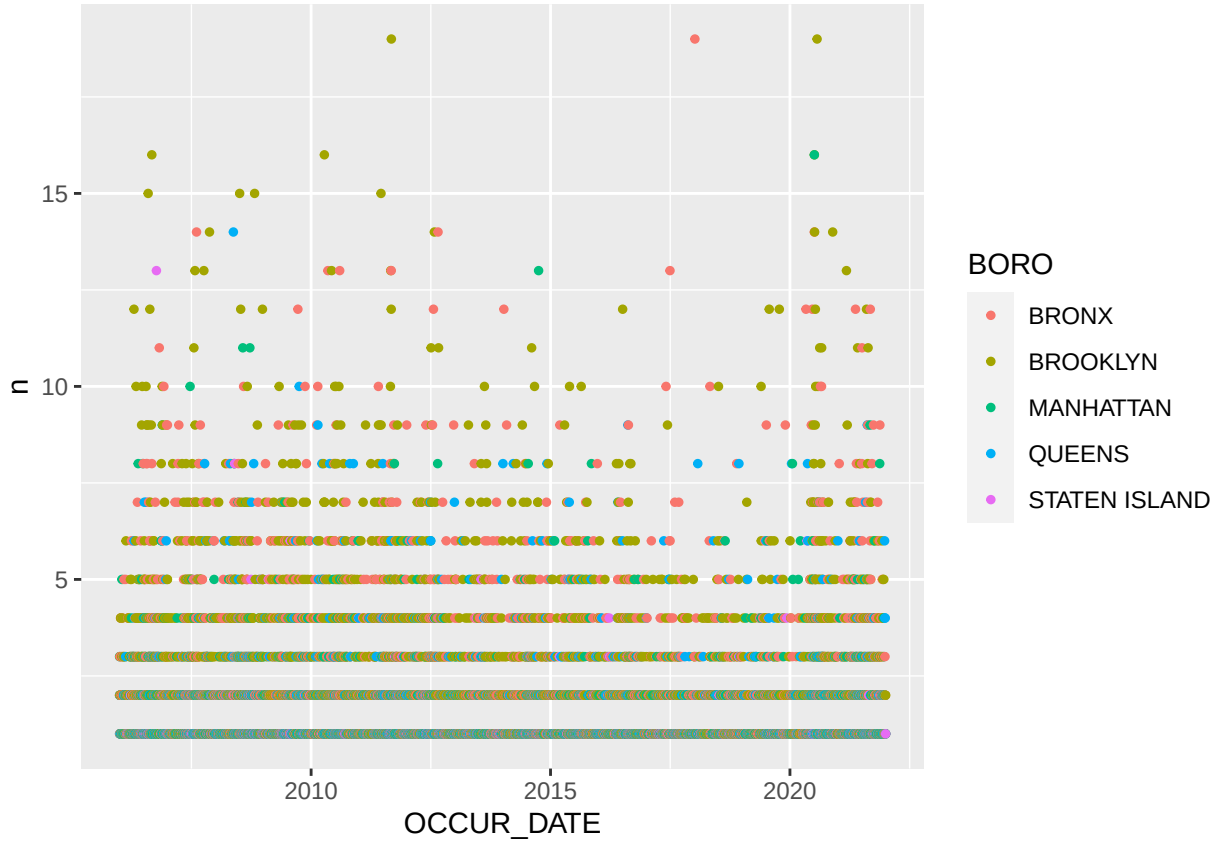
```
historical_data_by_location <- subset(historical_data, select=-c(INCIDENT_KEY, JURISDICTION_CODE, LOCATI

historical_data_by_location <- historical_data_by_location %>% mutate(OCCUR_DATE = mdy(OCCUR_DATE))
```

I will then aggregate the data, or counting each instances of each shooting group by OCCUR_DATE and BORO. My hope is to get the count of shootings each borough has on a particular day.

```
count <- historical_data_by_location %>%  count(OCCUR_DATE, BORO)
```

Plotting count of shootings by date and by borough

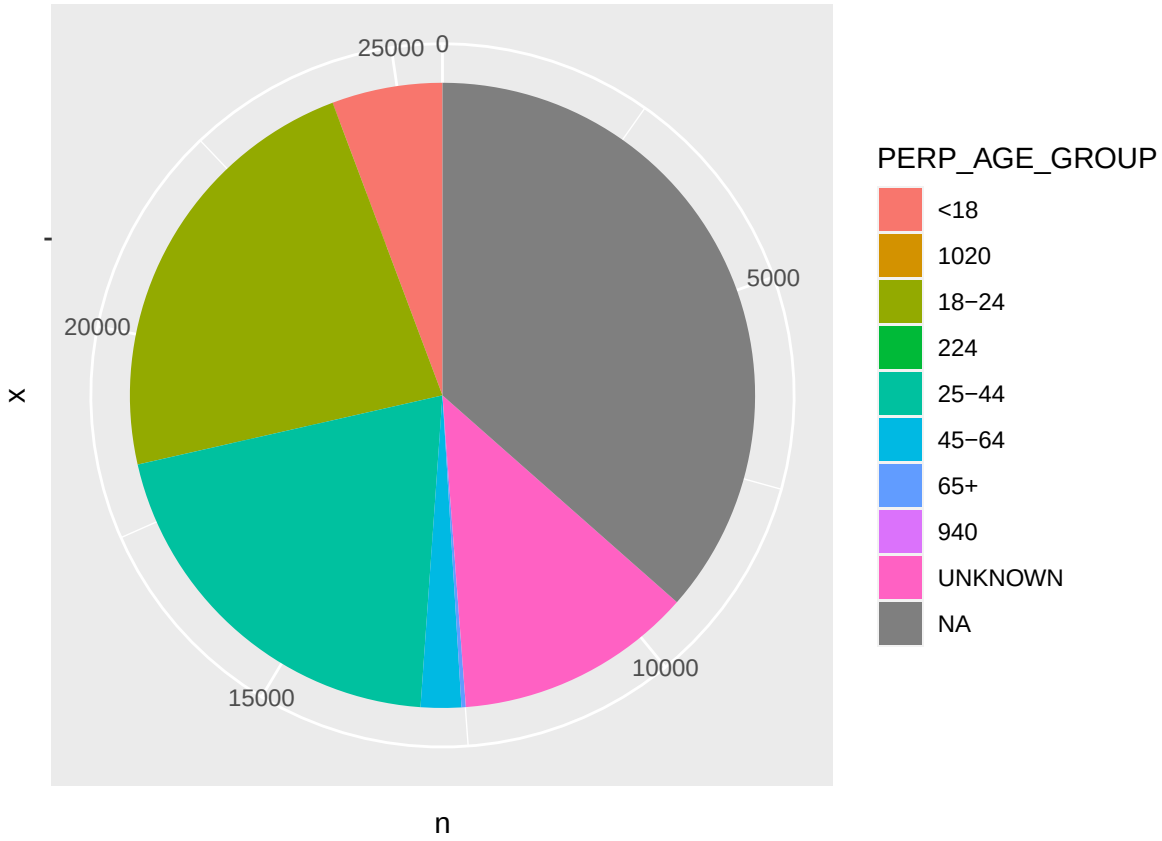
```
ggplot(count, aes(x = OCCUR_DATE, y = n, color = BORO)) +geom_point(size=1)
```



According to the scatter plot above, a majority of boroughs have shooting count of five or less on a given day. Also with a cursory glance, it appears boroughs like Queens and Manhattan have fewer shootings than the other boroughs.

Creating Pie Chart of Perpetrator Age Groups of the shootings

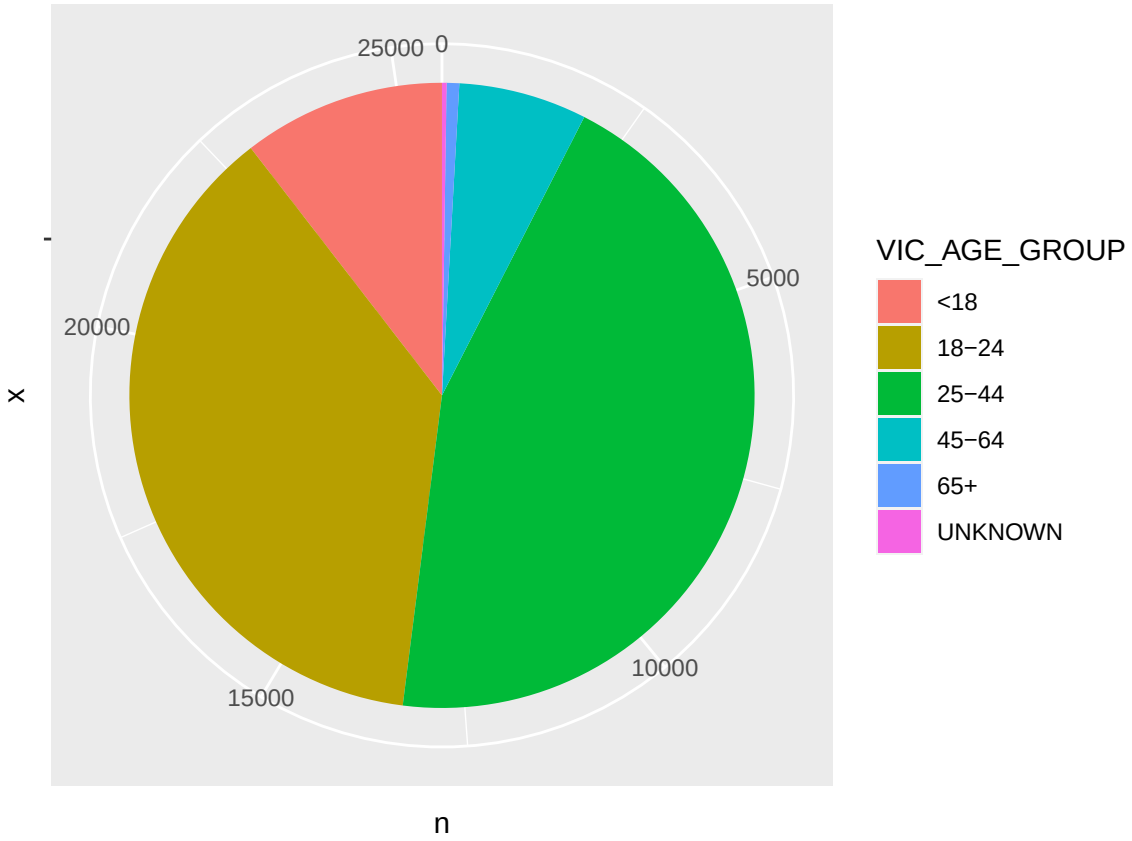
```
historical_data_by_person_demo <- subset(historical_data, select=-c(INCIDENT_KEY, JURISDICTION_CODE, LO

historical_data_by_person_demo_count <- historical_data_by_person_demo %>%  count(PERP_AGE_GROUP)

ggplot(historical_data_by_person_demo_count, aes(x = "", y = n , fill = PERP_AGE_GROUP)) +
  geom_col() +
  coord_polar(theta = "y")
```



The above pie charts shows the distributions of PERP_AGE_GROUP. I kept the unknown or NA records because the data would be inappropriately portray as a result of removing these records. The largest two age groups are 18-24 and 25-44, meaning these two age groups are the majority of perpetrator.

Creating Pie Chart of Victims Age Groups of the shootings

```
historical_data_by_person_demo_count <- historical_data_by_person_demo %>%  count(VIC_AGE_GROUP)
ggplot(historical_data_by_person_demo_count, aes(x = "", y = n , fill = VIC_AGE_GROUP)) +
  geom_col() +
  coord_polar(theta = "y")
```



The above pie charts shows the distributions of VIC_AGE_GROUP. The largest two age groups are 18-24 and 25-44, meaning these two age groups were the majority of shooting victims.

Possible Questions

Given my brief analysis, I would want to further research the shooting count relative to the population of each borough. Similarly, I would also further see the population density of each age group and compare it to the PERP_AGE_GROUP and VIC_AGE_GROUP distribution.