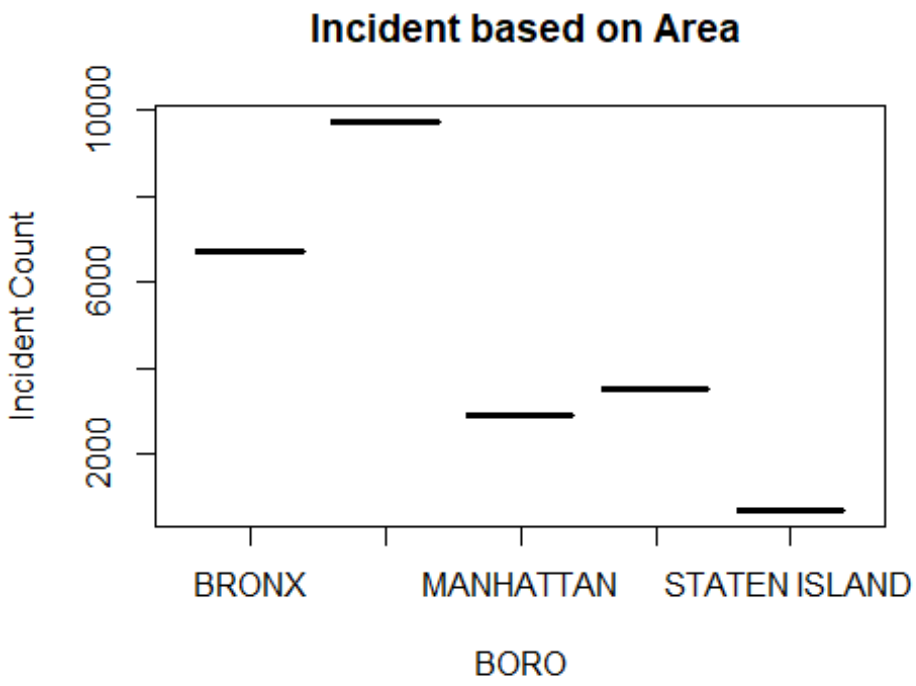
NYPD_shooting_assingment

11/7/2021

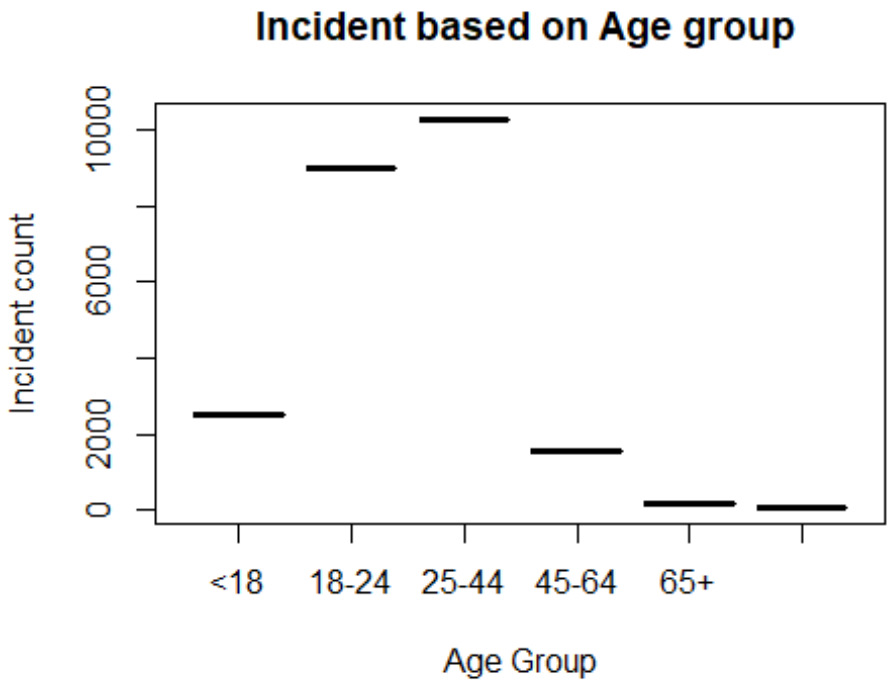
```
## At beginning we will load few libraries which will be useful for our
further data cleaning and visualization.
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

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

library(ggplot2)
library(tinytex)

# Here we are reading the data frame to dt1 variable from our source. The
source can be a local data file or from any remote location. In our use case
we are using a file located remotely.
url_in <- "https://data.cityofnewyork.us/api/views/833y-
fsy8/rows.csv?accessType=DOWNLOAD"
dt1 <- read.csv(url_in)

# Here we are cleaning the data frame to filter out extra attributes from the
data frame for further  analysis
dt1 <- dt1 %>% select(-c(Lon_Lat, X_COORD_CD,Y_COORD_CD,Latitude,
Longitude,LOCATION_DESC,PERP_SEX,PERP_AGE_GROUP,PERP_SEX,PERP_RACE))


# After the cleaning and restructing of data we will use the below command to
verify show the structure of data frame.
str(dt1)

## 'data.frame':    23568 obs. of  10 variables:
##  $ INCIDENT_KEY           : int  201575314 205748546 193118596 204192600
201483468 198255460 194570529 203211777 193694863 199582060 ...
##  $ OCCUR_DATE             : chr  "08/23/2019" "11/27/2019" "02/02/2019"
"10/24/2019" ...
##  $ OCCUR_TIME             : chr  "22:10:00" "15:54:00" "19:40:00"
"00:52:00" ...
##  $ BORO                   : chr  "QUEENS" "BRONX" "MANHATTAN" "STATEN
ISLAND" ...
```

```
##  $ PRECINCT               : int  103 40 23 121 46 73 81 67 114 69 ...
##  $ JURISDICTION_CODE      : int  0 0 0 0 0 0 0 0 2 0 ...
##  $ STATISTICAL_MURDER_FLAG: chr  "false" "false" "false" "true" ...
##  $ VIC_AGE_GROUP          : chr  "25-44" "25-44" "18-24" "25-44" ...
##  $ VIC_SEX                : chr  "M" "F" "M" "F" ...
##  $ VIC_RACE               : chr  "BLACK" "BLACK" "BLACK HISPANIC" "BLACK"
...

# In below command we are retrieving our first summary from the data frame to
review numerical min/max/mean of the data frame
summary(dt1)

##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME            BORO
##  Min.   :  9953245   Length:23568       Length:23568       Length:23568
##  1st Qu.: 55317014   Class :character   Class :character   Class
:character
##  Median : 83365370   Mode  :character   Mode  :character   Mode
:character
##  Mean   :102218616
##  3rd Qu.:150772442
##  Max.   :222473262
##
##     PRECINCT      JURISDICTION_CODE STATISTICAL_MURDER_FLAG VIC_AGE_GROUP
##  Min.   :  1.00   Min.   :0.0000    Length:23568            Length:23568
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character        Class
:character
##  Median : 69.00   Median :0.0000    Mode  :character        Mode
:character
##  Mean   : 66.21   Mean   :0.3323
##  3rd Qu.: 81.00   3rd Qu.:0.0000
##  Max.   :123.00   Max.   :2.0000
##                   NA's   :2
##    VIC_SEX           VIC_RACE
##  Length:23568       Length:23568
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##

#Here we are changing the format of Occur date from char to Date for our
analysis.
dt1$OCCUR_DATE <- as.Date(dt1$OCCUR_DATE, "%m/%d/%Y")

# In this section we are building a data set  on incident based on area
(BORO). This helped to visualize which BORO had more or less incidents for
entire duration.
dt2 <- data.frame(table(dt1$BORO))
```

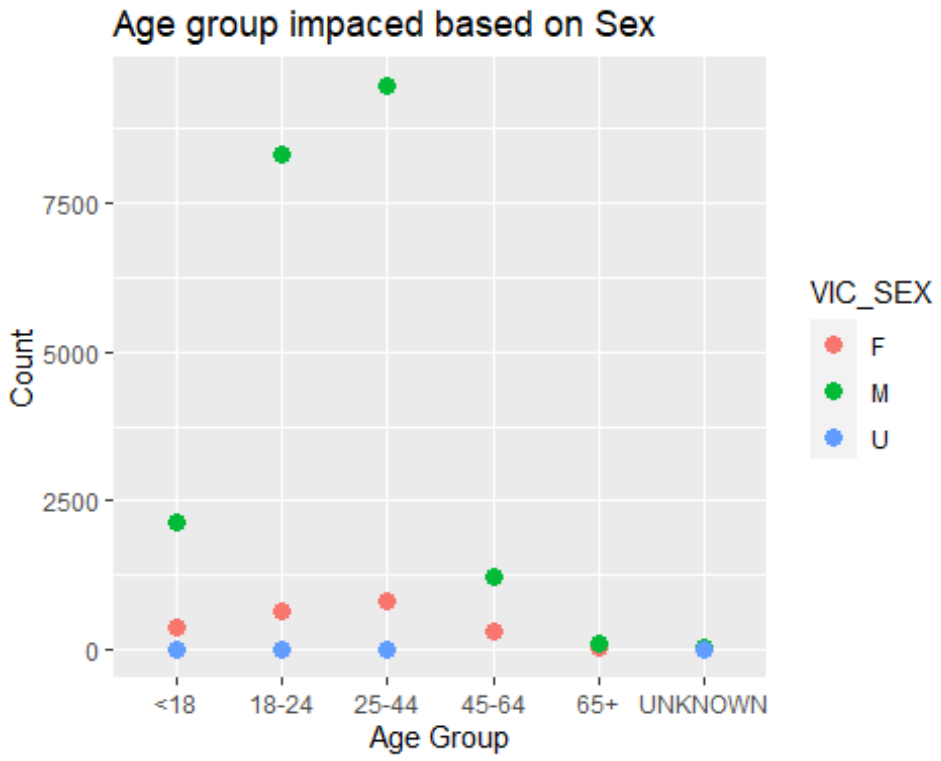
```
plot(dt2,main = " Incident based on Area", xlab = "BORO", ylab="Incident
Count")
```




```
# In this section we are building a data set  on incident based on age group.
This helped to visualize which age group is more vulnerable during entire
duration.
data.frame(table(dt1$VIC_AGE_GROUP))

##      Var1  Freq
## 1     <18  2525
## 2   18-24  9000
## 3   25-44 10287
## 4   45-64  1536
## 5     65+   155
## 6 UNKNOWN    65

plot(data.frame(table(dt1$VIC_AGE_GROUP)),main = " Incident based on Age
group", xlab = "Age Group", ylab="Incident count")
```




```
# We are plotting another graph here to analyse incident pattern based on age group and sex.
dt11 <- dt1 %>% group_by(VIC_AGE_GROUP) %>% count(VIC_SEX)
ggplot(dt11,aes(x=VIC_AGE_GROUP,y=n)) + geom_point(aes(col=VIC_SEX), size=3)+
labs(title="Age group impaced based on Sex", y="Count", x="Age Group")
```




```
# Analysis for bias --> We analysed this data from the point of age group, location & sex. There could be many other ways this data could be analysed.After carefully analyzing and verifying our area of visualization, I don't see any bias. In my view,we need more data to establish or identify bias in this visualization around my scope of analysis.
```