Source: https://images.app.goo.gl/YStTSgPUfhG5JeYn7

College of Professional Studies, Northeastern University
ALY6015 80625 Intermediate Analytics SEC 04
Spring 2021 CPS [BOS-A-HY]
Team Members: Aswin Kumar Rajendran, Chaya Kotturesh and Neil Mascarenhas

Introduction

This paper summarizes our preliminary (Exploratory Data Analysis - EDA) analysis on the selected dataset, the NYPD shooting incident dataset for this group project. It will share some insightful results from the EDA and descriptive statistics. Further, this paper identifies a set of methods used to answer the business questions on the dataset and justifies those methods using statistics and analytics concepts.

First, let us create the environment by importing the required packages to analyze and create a visualization on the dataset.

Business Questions

1. Is there a relationship between the Race of the victim and the Perpetrator who died in the shooting incident? *by Chaya Kotturesh*
2. Is there any noticeable relationship between the Perpetrator's Race to the victim's sex and age with a motive to recognize the Perpetrator's race pattern? *by Neil Mascarenhas*
3. Recognize the pattern/relationship between the victim's age and the Location, specifically Bar/Night Club, to recognize the trend throughout the years and increase the patrolling. *by Aswin Kumar Rajendran*

Analysis and Interpretations

```
NYPD = data.frame(read_csv("Data\\NYPD_Shooting_Incident_Data__Historic_.csv", skip_empty_rows =
        TRUE), stringsAsFactors = T)
```

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

```
### Removing unwanted variables/ columns
###

NYPD <- NYPD %>%
    select(-INCIDENT_KEY, -X_COORD_CD, -Y_COORD_CD, -Latitude, -Longitude, -Lon_Lat)


NYPD[duplicated(NYPD$INCIDENT_KEY),]
```

```
##  [1] OCCUR_DATE              OCCUR_TIME              BORO
##  [4] PRECINCT                JURISDICTION_CODE       LOCATION_DESC
##  [7] STATISTICAL_MURDER_FLAG PERP_AGE_GROUP          PERP_SEX
## [10] PERP_RACE               VIC_AGE_GROUP           VIC_SEX
## [13] VIC_RACE
## <0 rows> (or 0-length row.names)
```

```
Datatypelist<-as.data.frame(skimr::skim(NYPD)[[1]],skimr::skim(NYPD)[[2]])

knitr::kable(Datatypelist, "pipe", col.names = "Data Type")
```

	Data Type
OCCUR_DATE	character
BORO	character
LOCATION_DESC	character
PERP_AGE_GROUP	character
PERP_SEX	character
PERP_RACE	character

	Data Type
VIC_AGE_GROUP	character
VIC_SEX	character
VIC_RACE	character
OCCUR_TIME	difftime
STATISTICAL_MURDER_FLAG	logical
PRECINCT	numeric
JURISDICTION_CODE	numeric

```
#$`skimr::skim(NYPD)[[1]]`=="character"
```

```
#char_vector
        c("OCCUR_DATE","BORO","LOCATION_DESC","PERP_AGE_GROUP","PERP_SEX","PERP_RACE","VIC_AGE_GROUP","VIC_SEX","VIC


# Putting unknown for missing string values
NYPD[Datatypelist$`skimr::skim(NYPD)[[1]]`=="character"]
        [is.na(NYPD[Datatypelist$`skimr::skim(NYPD)[[1]]`=="character"])] <- "UNKNOWN"


# Putting 0 for missing numeric values
NYPD[Datatypelist$`skimr::skim(NYPD)[[1]]`=="numeric"][is.na(NYPD[Datatypelist$`skimr::skim(NYPD)
        [[1]]`=="numeric"])] <- 0
```

We have taken care of missing values. Lets us begin with our analysis and understand the dataset. Here we go through each variable/feature column and get meaningful insights from them. As we proceed, we will keep updating our dataset so that by the end, we get a complete dataset with variables and values we need for our further analysis.

We can see above that the frequency of the Boro is plotted. For this dataset, we have five Boros. Among them, Brooklyn has the highest records, while Staten Island has the lowest record of shootouts. Can we say that Brooklyn is more dangerous than the rest of Boros? Let us find it out further in our report.

Methodology

Here, we proceed with identifying the methods we will be using, along with justification for those methods.

Business Questions

1. Is there a relationship between the Race of the victim and the Perpetrator who died in the shooting incident? *Major contribution by Chaya Kotturesh*

Here, we will use the Generalized regression model to solve this question. We know that Black people have a higher count of shootings when compared with others. Here are

predictable variable is the Perpetrator's race, and the independent variable will be Victim race. We will create a model for those who are dead in the shootings.

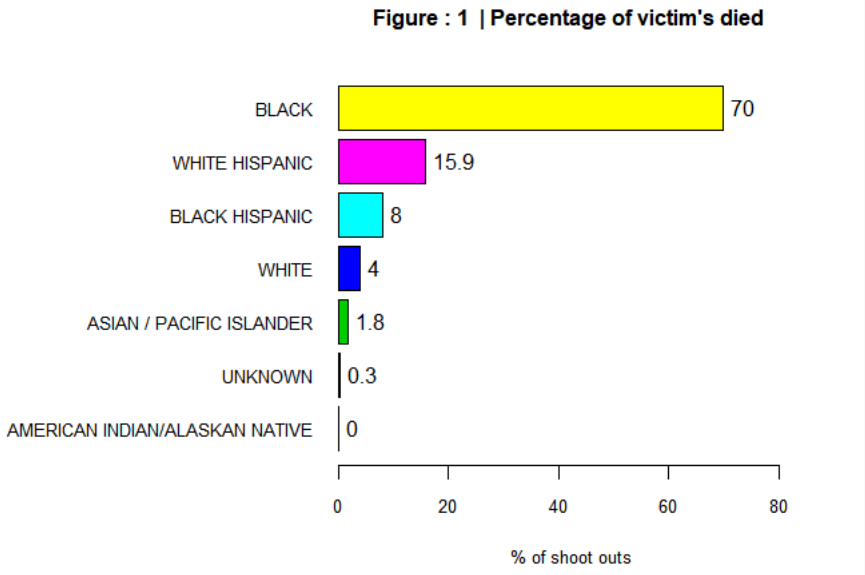
```
nypd_df <- read.csv("Data\\NYPD_Shooting_Incident_Data__Historic_.csv")
char_vector
        c("OCCUR_DATE","BORO","LOCATION_DESC","PERP_AGE_GROUP","PERP_SEX","PERP_RACE","VIC_AGE_GROUP","VIC_SEX","VIC

nypd_df[char_vector][is.na(nypd_df[char_vector])] <- "UNKNOWN"
```

Reading the NYPD data set and performed the cleaning of character columns by filling UNKNOWN for NA's

```
death_data<- nypd_df %>% filter(nypd_df$STATISTICAL_MURDER_FLAG== TRUE)

tab1(death_data$VIC_RACE,  cex.main=1,  cex.name=0.8,  cex.axis=0.8,  cex.lab=0.8,  sort.group
        ="decreasing", bar.values ="percent", main = paste(GetFigureCount()," | Percentage of
        victim's died"), xlab=" % of shoot outs", ylab="Race of Victims", horiz = T)
```




```
## death_data$VIC_RACE :
##                                Frequency Percent Cum. percent
## BLACK                               2888    70.0         70.0
## WHITE HISPANIC                       656    15.9         85.9
## BLACK HISPANIC                       329     8.0         93.8
## WHITE                                166     4.0         97.9
## ASIAN / PACIFIC ISLANDER              74     1.8         99.7
## UNKNOWN                               14     0.3        100.0
## AMERICAN INDIAN/ALASKAN NATIVE         0     0.0        100.0
##   Total                             4127   100.0        100.0
```

As the business question is to find the relationship between the race of people that lead to the victim's death, I filtered the data set based on the victim's death. As we can see, there are around 4,127 records of shootings that lead to the victim's deaths. In that, the highest number of victims who died were of Black Race with whopping 70%, followed by White Hispanics with 15.9%

```
death_data$PERP_RACE <- sub("^$", "UNKNOWN", death_data$PERP_RACE)#replacing empty cells with
        UNKNOWN
death_data$VIC_RACE <- sub("^$", "UNKNOWN", death_data$VIC_RACE)#replacing empty cells with
        UNKNOWN

crosstable2<- table(death_data$PERP_RACE,death_data$VIC_RACE)
crosstable2 <- round(prop.table(crosstable2)*100, 2)
crosstable2<-kable(crosstable2, caption = paste(GetFigureCount(), " | Percentage of Perpetrator's
        race killing victim's race"), xlab="test") %>%
    kableExtra::kable_styling(., position = "float_left")
crosstable2
```

Figure : 2 | Percentage of Perpetrator's race killing victim's race

	ASIAN / PACIFIC ISLANDER	BLACK	BLACK HISPANIC	UNKNOWN	WHITE	WHITE HISPANIC
ASIAN / PACIFIC ISLANDER	0.44	0.17	0.02	0.00	0.07	0.10
BLACK	0.63	35.52	2.64	0.12	0.90	4.92
BLACK HISPANIC	0.07	1.84	0.99	0.02	0.02	1.57
UNKNOWN	0.41	29.20	2.57	0.12	0.94	3.95
WHITE	0.07	0.19	0.07	0.00	1.62	0.34
WHITE HISPANIC	0.17	3.05	1.67	0.07	0.46	5.02

Intending to find the relationship between Perpetrator and victim's race, we created the contingency table.

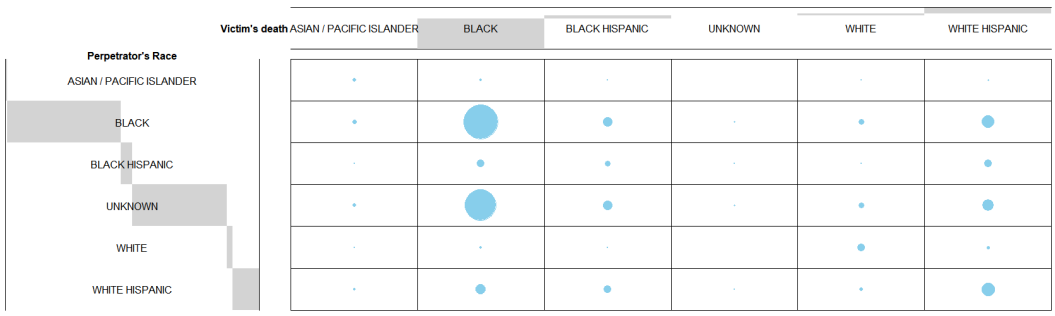
We first convert the data as a table.

```
#library("gplots")
# 1. convert the data as a table
test<- table(death_data$PERP_RACE,
             death_data$VIC_RACE)
dt <- as.table(as.matrix(test))
```

Now, we use the converted data to visualize the Perpetrator's race killing the victim's race.

```
# 2. Graph
balloonplot(t(dt),
            main =paste(GetFigureCount(), " | Perpetrator's race killing victim's race"),
            xlab ="Victim's death",
            ylab="Perpetrator's Race",
            label = FALSE,
            show.margins = FALSE,
            cex.main = 0.9  )
```

Figure : 3 | Perpetrator's race killing victim's race



From the above graph, we can observe that Black and Unknown or white-Hispanic races murdered most Blacks. Unfortunately, the following significant pattern can be found where White-Hispanic were majorly murdered by White Hispanic themselves and next highest by BLACK people.

Even though contingency tables show some relationship between, Perpetrator and the victim's race, let us identify by Performing a Chi-square test at 0.05 significant level to identify the relationship between two categorical variables in the contingency table.

`df=(r-1)(c-1)`* degrees of freedom and *`p = 0.05`* *("Chi-Square Test of Independence in R - Easy Guides - Wiki - STHDA", 2020).

Here,

r is the number of rows in the contingency table

c is the number of column in the contingency table

Step a: State the hypothesis and identify the claim

Null hypothesis $\it{H_{0}}$: There is no relationship exists between the race of perpetrator and victim.
Alternative hypothesis $\it{H_{1}}$: There is relationship between the race of perpetrator and victim. **Step b: Find the critical value**

```
#here rows=6 columns=6 so df= 5*5

Critical_Val = qchisq(.05, df=25)

cat("The Critical value is ",Critical_Val)
```

```
## The Critical value is  14.61141
```

Step c: Performing χ^2 Test or find the p-value

```
output_task4 <-chisq.test(dt)
```

```
## Warning in chisq.test(dt): Chi-squared approximation may be incorrect
```

```
output_task4
```

```
##
##  Pearson's Chi-squared test
##
## data:  dt
## X-squared = 2301.6, df = 25, p-value < 2.2e-16
```

```
cat("Our Decision mustbe :", ifelse(output_task4$p.value < 0.05, "\nReject the null Hypothesis",
         "\nFail to reject the null hypothesis" ))
```

```
## Our Decision mustbe :
## Reject the null Hypothesis
```

Step d: Make the decision
The decision rule is that if p-value is lesser than α we must reject the null hypothesis. p-value < α i.e. 0.00000000000000022 is lesser than 0.05 ,so we must reject the $\it{H_{0}}$ **Step e: Summary**

Interpretation: The evidence leads us to reject the null hypothesis. Therefore, there is a relationship between the race of Perpetrator and victim.

Another approach to figure out the relationship between categorical variables in R is by using Cramer's V coefficient

```
##library(vcd)

assocstats(table(death_data$PERP_RACE, death_data$VIC_RACE))
```

```
##                     X^2 df P(> X^2)
## Likelihood Ratio 1043.4 25        0
## Pearson          2301.6 25        0
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.598
## Cramer's V        : 0.334
```

interpretation of this function:

1. Cramer's V varies from 0 to 1, a value of 1 indicates perfect association.In this scenario Cramer's V co-efficient is 0.334 which shows a moderate effect due to association.A value between *0.30 - < 0.50* is considered to be moderate.

2. Contingency coefficient values also vary between 0 to 1. The larger the contingency coefficient, the stronger the association. 0.5 again shows the moderate association between Perpetrator and victim's race.

The discipline, the relevant data, and the analyst's goals all influence how measures of association are interpreted. There are sometimes guidelines for "small," "medium," and "large" impact. A smaller effect size may be considered "large," but in physical science, such as chemistry, it may be considered very small in psychology or behavioral science. The unique circumstances of the study are necessary as well.

---

2. Is there any noticeable relationship between the victim's Sex to the Perpetrator's Race and age with a motive to recognize the Perpetrator's race pattern? *Major contribution by Neil Mascarenhas*

The NYPD is facing difficulties in identifying the reason behind the shooting. They are interested in seeing the relation between the Perpetrator's Race with Victim's sex and Age Group.

Here, we will use several modeling techniques to determine the pattern. First, we need to know the relationship between the Perpetrator's Race to the victim's Sex. We will create the model for these variables and understand what the variable that determines the relationship is. Moreover, which are the variables that have the highest relationship. Knowing this will help us predict and recognize the Perpetrator's Race.

```
BQ2 <- select(NYPD, c(PERP_RACE, VIC_SEX, VIC_AGE_GROUP)) %>% na.omit() %>% filter(PERP_RACE !=
        "UNKNOWN", VIC_SEX != "U", VIC_AGE_GROUP != "UNKNOWN", PERP_RACE != "AMERICAN
        INDIAN/ALASKAN NATIVE")

set.seed(11521)
BQ2 <- BQ2[sample(nrow(BQ2)), ] ### Shuffles
```

```
All_males<-BQ2 %>% filter(VIC_SEX=="M" ) %>% head(table(BQ2$VIC_SEX)[1])
All_females<-BQ2 %>% filter(VIC_SEX=="F") %>% head(table(BQ2$VIC_SEX)[1])

males_females <-  rbind(All_males, All_females)

set.seed(11521)
BQ2 <- males_females[sample(nrow(males_females)), ] ### Shuffles


#describe(BQ2)

cat("Any missing or NA values? \n(0 = No)\nResult = ",sum(!complete.cases(BQ2)))
```

```
## Any missing or NA values?
## (0 = No)
## Result =  0
```

```
##
```

Even though the tables show some sort of relationship between the variables, let us identify by Performing the Chi-square test of Independence at 0.05 significant level to identify the relationship between three categorical variables in the table above.

```
alpha = 0.05

LoSig   = 1-alpha
DegFre  = (nrow(BQ2) -1)*(ncol(BQ2) -1 )
```

Step 1: State the hypotheses and identify the claims

Claim 1

- Null hypothesis H_0 : The Victim Sex is independent of Perpetrator Race
- Alternative hypothesis H_1 : The Victim Sex is dependent upon Perpetrator Race

Claim 2

- Null hypothesis H_0 : The Victim Age Group is independent of Perpetrator Race
- Alternative hypothesis H_1 : The Victim Age Group is dependent upon Perpetrator Race

Step 2: Find the critical value.

```
Critical_Val <-qchisq(p=LoSig, DegFre)

cat("The Critical value for both claims is ", Critical_Val)
```

```
## The Critical value for both claims is  5565.973
```

Step 3: Compute the test value.

```
Table_PR_VS <- table(BQ2$VIC_SEX,BQ2$PERP_RACE) # Vic Sex and Prep Race are stored in TAB
Table_PR_VA <- table(BQ2$PERP_RACE,BQ2$VIC_AGE_GROUP) # Vic Age Group and Vic Sex are stored in
        TAB


ChT_PR_VS <- chisq.test(Table_PR_VS, correct = T)
ChT_PR_VA <-chisq.test(Table_PR_VA, correct = T)

# ChT_PR_VS$p.value
# ChT_PR_VG$p.value
```

Step 4: Make the decision.

```
cat(("Determine Dependancy between and Victim Sex "),
ifelse(ChT_PR_VS$statistic < Critical_Val,
      "Exists",
       "Does Not Exist" ), " and the Test Score is ", ChT_PR_VS$statistic )
```

```
## Determine Dependancy between and Victim Sex  Exists  and the Test Score is  22.91352
```

```
cat("\n \n \n")
```

```
cat(("Determine Dependancy between Perpetrator Race and Victim Age Group "),
ifelse(ChT_PR_VA$statistic < Critical_Val,
       "Exists",
       "Does Not Exist" ), " and the Test Score is ", ChT_PR_VA$statistic)
```

```
## Determine Dependancy between Perpetrator Race and Victim Age Group  Exists  and the Test Score is
```

From the above tests, we computed the Chi-Square statistic and the critical value. We see that the Chi-Square statistic is Greater than the Critical value. Therefore we have enough

evidence to Reject the null hypothesis H0 for 0.05 degree of freedom.

Now that we have made the decision, let us cross-verify using the p-value

```
cat(("Determine Dependancy between and Victim Sex "),
ifelse(ChT_PR_VS$p.value < alpha,
      "Exists",
       "Does Not Exist" ), " and the P Value is ", ChT_PR_VS$statistic )
```

```
## Determine Dependancy between and Victim Sex  Exists  and the P Value is  22.91352
```

```
cat("\n \n \n")
```

```
cat(("Determine Dependancy between Perpetrator Race and Victim Age Group "),
ifelse(ChT_PR_VA$p.value < alpha,
       "Exists",
       "Does Not Exist" ), " and the P Value is ", ChT_PR_VA$p.value)
```

```
## Determine Dependancy between Perpetrator Race and Victim Age Group  Exists  and the P Value is  4
```

We can see that the p-value is significantly less than the decided alpha. Therefore we have made the right decision to reject the null hypothesis.

Step 5: Summarize the results.

As we have enough evidence to reject the null hypothesis H0 for both the claims, Which states The Victim Sex is dependent upon Perpetrator Race and Victim Age Group

Now, we proceed with preparing the data for modeling. Here we split the data between the test and train dataset.

```
set.seed(11521)
trainIndex <- createDataPartition(BQ2$VIC_SEX, p = 0.7, list = FALSE, times = 1)



BQ2_train <- BQ2[ trainIndex,]
BQ2_test <- BQ2[-trainIndex,]


table(NYPD$VIC_SEX)  %>% t()%>%
  kbl(caption = paste(GetFigureCount()," | Male and Female count in Entire dataset"),
      align = 'c') %>%
        kable_classic_2(full_width = T,
                        position = "center",
                        fixed_thead = T)
```

Figure : 4 | Male and Female count in Entire dataset

F	M	U
1999	19615	12

```
table(BQ2_train$VIC_SEX)  %>% t()  %>%
  kbl(caption = "Male and Female count in Train dataset",
      align = 'c') %>%
        kable_classic_2(full_width = T,
                        position = "center",
                        fixed_thead = T)
```

Male and Female count in Train dataset

F	M
945	945

```
table(BQ2_test$VIC_SEX)%>% t()  %>%
  kbl(caption = "Male and Female count in Test dataset",
      align = 'c') %>%
        kable_classic_2(full_width = T,
                        position = "center",
                        fixed_thead = T)
```

Male and Female count in Test dataset

F	M
404	404

We have successfully split the dataset into two different and unique sets. We have 70% of the dataset for train and 30% of the dataset for the test. Let us have a glimpse of those two sets.

```
headTail(BQ2_train,top = 5, bottom = 0) %>%
  kbl(caption = paste(GetFigureCount()," | NYPD Train set : 70%"),
      align = 'c') %>%
        kable_classic_2(full_width = T,
                        position = "center",
                        fixed_thead = T)
```

Figure : 5 | NYPD Train set : 70%

	PERP_RACE	VIC_SEX	VIC_AGE_GROUP
2510	BLACK	F	<18
1652	BLACK	F	18-24
2692	BLACK	F	25-44
366	WHITE	M	45-64
948	BLACK	M	25-44
...	NA	NA	NA

```
headTail(BQ2_test,top = 5, bottom = 0)%>%
  kbl(caption = paste(GetFigureCount()," |NYPD Test set : 30%"),
      align = 'c') %>%
        kable_classic_2(full_width = T,
                        position = "center",
                        fixed_thead = T)
```

Figure : 6 |NYPD Test set : 30%

	PERP_RACE	VIC_SEX	VIC_AGE_GROUP
659	BLACK	M	18-24
345	BLACK	M	25-44
446	BLACK	M	18-24
804	BLACK HISPANIC	M	<18
2278	WHITE HISPANIC	F	<18
...	NA	NA	NA

Fitting the Models.

As mentioned earlier, we will fit several models. Below are the models our team came up with and agreed upon.

1) Support Vector Machine

2) Decision Trees

3) Logistic regression

4) Generalized Linear Model

5) Random Forest

6) Generalized Boosted Regression Modelling

Let us now fit them and compare their performance later.

```
###  VIC_SEX ~ VIC_AGE_GROUP * PERP_RACE


# prepare training scheme
control <- trainControl(method="repeatedcv", number=5, repeats=2, sampling = "up",
        allowParallel=TRUE)

# train the SVM model
set.seed(15)
modelSVM <- train(VIC_SEX~., data=BQ2_train, method="svmRadial", trControl=control)

# train the GLM model
set.seed(15)
modelGLM <- train(VIC_SEX~., data=BQ2_train, method="glm", trControl=control, family =
        "binomial",na.action = "na.exclude")

# train the RF model
set.seed(15)
```

```
modelRF <- train(VIC_SEX~., data=BQ2_train, method="rf", trControl=control)

# train the Log Reg model
set.seed(15)
modelLOGR    <-    train(VIC_SEX~.,    data=BQ2_train,    method="multinom",    trControl=control,
         verbose=FALSE, trace=FALSE)

# train the gbm model
set.seed(15)
modelGBM <- train(VIC_SEX~., data=BQ2_train, method="gbm", trControl=control, verbose=FALSE)

# train the Decison trees model
set.seed(15)
modelDTREE <- train(VIC_SEX~., data=BQ2_train, method="rpart", trControl=control)
```

We have generated several models. Let us now cumulate the results and compare the efficiency.

```
# collect resamples
results <- resamples(list(SVM=modelSVM,
                          RF=modelRF,
                          GLM=modelGLM,
                          MULTM=modelLOGR,
                          GBM=modelGBM,
                          DTREE=modelDTREE))

summary(results)
```

```
##
## Call:
## summary.resamples(object = results)
##
## Models: SVM, RF, GLM, MULTM, GBM, DTREE
## Number of resamples: 10
##
## Accuracy
##            Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## SVM   0.5608466 0.5773810 0.5952381 0.5910053 0.6018519 0.6269841    0
## RF    0.5687831 0.5773810 0.5925926 0.5928571 0.6011905 0.6375661    0
## GLM   0.5661376 0.5800265 0.5939153 0.5939153 0.5992063 0.6375661    0
## MULTM 0.5661376 0.5820106 0.5952381 0.5944444 0.5992063 0.6375661    0
## GBM   0.5661376 0.5800265 0.5925926 0.5933862 0.6005291 0.6322751    0
## DTREE 0.5608466 0.5780423 0.5886243 0.5923280 0.6078042 0.6322751    0
##
## Kappa
##            Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## SVM   0.1216931 0.1547619 0.1904762 0.1820106 0.2037037 0.2539683    0
## RF    0.1375661 0.1547619 0.1851852 0.1857143 0.2023810 0.2751323    0
## GLM   0.1322751 0.1600529 0.1878307 0.1878307 0.1984127 0.2751323    0
## MULTM 0.1322751 0.1640212 0.1904762 0.1888889 0.1984127 0.2751323    0
## GBM   0.1322751 0.1600529 0.1851852 0.1867725 0.2010582 0.2645503    0
## DTREE 0.1216931 0.1560847 0.1772487 0.1846561 0.2156085 0.2645503    0
```

The above results show us the summary of the model's performance. It has the Accuracy and Kappa metrics. We will use both to determine which one did best.

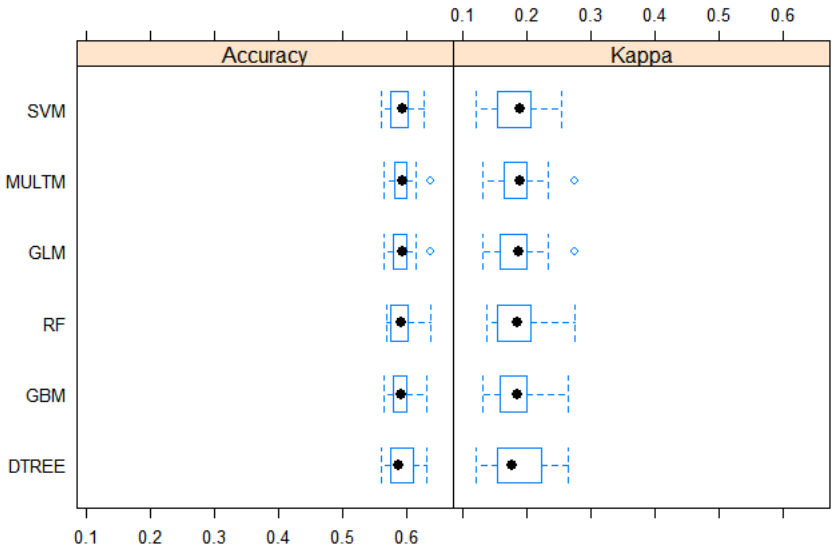
We can see that Random forest *RF*, Generalised Linear model *GLM* and logistic regression model *MULTM* did best with the accuracy of 63.2% by seeing the maximum accuracy.

It is difficult to compare as the accuracy is the same for all three.

Therefore, we try to visualize it, hoping it might help to compare the results.

```
# boxplots of results
bwplot(results, main=paste(GetFigureCount()," | Kappa and Accuracy Matrics"))
```

Figure : 7 | Kappa and Accuracy Matrics



From the above figure, we can see that the kappa max value was highest among all the other models for Random forest. We also see that for GLM and Logistic Regression model, the max score was not in the third quartile. This shows that they performed well a few times compared to random forest.

We choose to go ahead with Random Forest as our desired model for further analysis.

Before we fit the model, we needed to find out which parameters were best and use them in our final models

We decided to go with cross-validation and perform the grid algorithm to determine the features.

```
# Define the control
trControl <- trainControl(method = "cv",
    number = 10,
    search = "grid")

set.seed(1234)
# Run the model
rf_default <- train(VIC_SEX~.,
    data = BQ2_train,
    method = "rf",
    metric = "Accuracy",
    trControl = trControl)
# Print the results
print(rf_default)
```

```
## Random Forest
##
## 1890 samples
##    2 predictor
```

```
##    2 classes: 'F', 'M'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1700, 1702, 1700, 1700, 1700, 1702, ...
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##   2     0.5879171  0.1758343
##   5     0.5884434  0.1768869
##   8     0.5847312  0.1694625
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 5.
```

After doing Grid and Cross-validation, we found out that our model accuracy got decreased. It was 63.3% and now looking at the above result we see it it is at 58.8%

To figure out why this shift was caused. We then decided to plot the results and visually see the error rate.

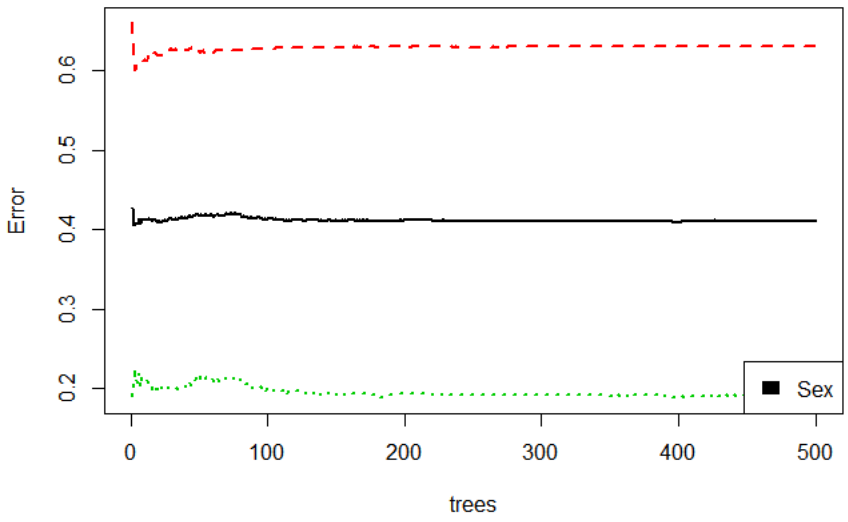
```
num.var     <- ncol(BQ2_train)

BQ2_train_rf = BQ2_train %>% mutate_if(is.character, as.factor)
#BQ2_train_rf$VIC_SEX = as.numeric(BQ2_train$VIC_SEX)

RF_Mod<-randomForest::randomForest(VIC_SEX~.,data = BQ2_train_rf,
                      mtry = num.var - 1, # try all variables at each split, except the response
         variable
                    proximity = TRUE,
                    importance = TRUE,
                    ntree = 500)


plot(RF_Mod, main=paste(GetFigureCount()," | Random Forest Trees and error"),lwd=2)
legend("bottomright",
c("Sex"),
fill=c("black")
)
```

Figure : 8 | Random Forest Trees and error



As we can see above for Victim Race, the error was very high. However, it was not high as in the initial iterations. However, as the iterations increased, the rate increased and remained constant throughout the Age group. The error rate had spiked, and then it reduced. It was then constant throughout. Lastly, we see for Victim Sex that the error was moderate and kept constant throughout. Entirely we can see that the error rate fluctuated initially but later kept constant throughout the entire 500 trees.

Now, we create a confusion matrix and report the results of your model predictions on the train set. We will interpret and discuss the confusion matrix.

A confusion matrix is a handy tool for calibrating the output of a model and examining all possible outcomes of our predictions. We see the values such as true positive, true negative, false positive, false negative.

We First need to have a prediction matrix to get the confusion matrix. Let us generate and use it.

```
# c("link", "response", "terms", vector, prob, class, raw )
#
predicted <- predict(rf_default, BQ2_train)  # predicted scores
```

We have created the prediction matrix for the training dataset. Now we will use these results to get the confusion matrix. Which will show us how well did the model predict the values. This is to see for values that are known. Later, we will use a test dataset for values unknown to the model and see how well it does on unseen data.

```
table(Pred = predicted, True =BQ2_train$VIC_SEX)%>%
  kbl(caption = paste(GetFigureCount()," | Confussion matrix for Train"),
      align = 'c') %>%
        kable_classic_2(full_width = T,
                        position = "center",
                        fixed_thead = T)
```

Figure : 8 | Confussion matrix for Train

	F	M
F	355	170
M	590	775

```
BQ2_train$VIC_SEX = factor(BQ2_train$VIC_SEX)

levels(BQ2_train$VIC_SEX) = c("F", "M")


caret::confusionMatrix(predicted, BQ2_train$VIC_SEX)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   F   M
##          F 355 170
```

```
##          M 590 775
##
##                Accuracy : 0.5979
##                  95% CI : (0.5754, 0.6201)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.1958
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.3757
##             Specificity : 0.8201
##          Pos Pred Value : 0.6762
##          Neg Pred Value : 0.5678
##              Prevalence : 0.5000
##          Detection Rate : 0.1878
##    Detection Prevalence : 0.2778
##       Balanced Accuracy : 0.5979
##
##        'Positive' Class : F
##
```

We generated the confusion matrics. As we can see above, To understand the confusion matrix and interpret it, we need to base it on assumptions.

We see that our prediction model did not work very well. We see that there is a false positive present.

From our analysis, we can say that False Positives are higher and are more damaging for the analysis. We feel this is not good. The type of error depends on which use case we need to have. What are we trying to predict, and what values we expect to get.

Reporting and interpreting metrics for Accuracy, Recall-Precision, and Specificity.

```
pred    <-    prediction(as.numeric(as.character(factor(predicted,    labels    =    c(0,1)))),
        as.numeric(as.character(factor(BQ2_train$VIC_SEX, labels = c(0,1)))))
```

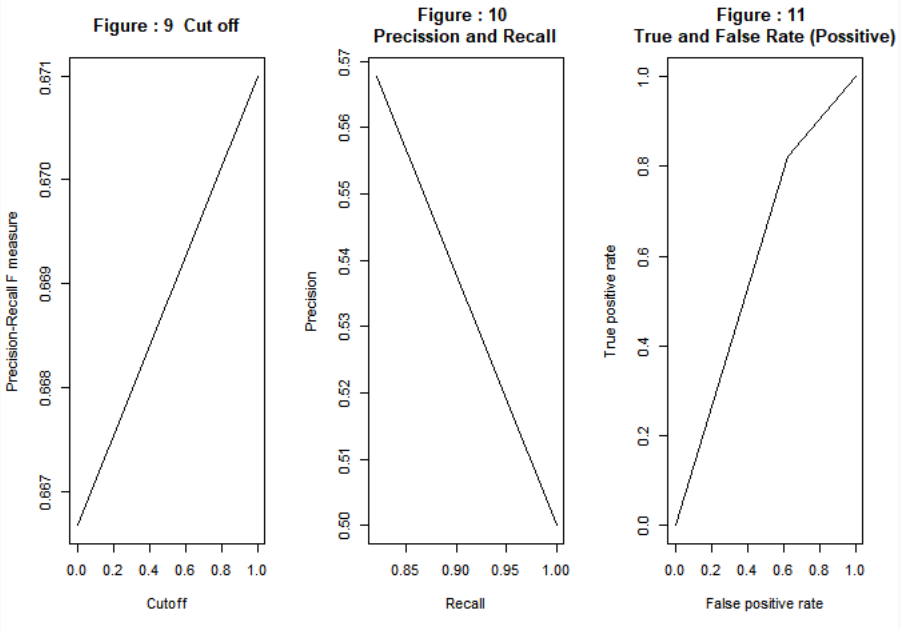
We have clreated the prediction matrics let us use it to dettermine the performances.

```
## ROC curve
ROC.perf_tf <- performance(pred, "tpr", "fpr")


## Recall-Precision curve
RP.perf_pr <- performance(pred, "prec", "rec")


## Cutoff
RP.F_measure <- performance(pred, "f", x.measure="cutoff")
```

```
#plot the curves
par(mfrow=c(1,3))
plot (RP.F_measure,main= paste(GetFigureCount()," Cut off"))
plot (RP.perf_pr, main=paste(GetFigureCount(),"\nPrecission and Recall"))
plot (ROC.perf_tf, main=paste(GetFigureCount(),"\nTrue and False Rate (Possitive)"))
```




```
##
```

```
# Specificity and Sensitivity

c_mat  = caret::confusionMatrix(predicted, BQ2_train$VIC_SEX)

# c_mat$table
# c_mat$table[1]
# c_mat$table[2]
# c_mat$table[3]
# c_mat$table[4]


sensitivity_Train <- c_mat$table[1]/(c_mat$table[1]+c_mat$table[2])

specificity_Train <- c_mat$table[4]/(c_mat$table[4]+c_mat$table[3])

cat("The sensitivity is ", sensitivity_Train, " and Specificity is ", specificity_Train)
```

```
## The sensitivity is  0.3756614  and Specificity is  0.8201058
```

Sensitivity is the percentage of `1`'s (actuals) correctly predicted by the model. It is also called True Positive Rate. Here our Sensitivity of the model is 0.376

Sensitivity= # Actual 1's and Predicted as 1's / # of Actual 1's

Specificity is the percentage of `0`'s (actuals) correctly predicted by the model. It is also called True Negative Rate. Here our Specificity of the model is 0.82. Specificity can also be calculated as 1 − False Positive Rate.

Specificity=# Actual 0's and Predicted as 0's / # of Actual 0's

Creating a confusion matrix and reporting the results of our model for the test set.

```
# c("link", "response", "terms", vector, prob, class, raw )
#
predicted_Test <- predict(rf_default, BQ2_test)  # predicted scores


BQ2_test$VIC_SEX = factor(BQ2_test$VIC_SEX)
levels(BQ2_test$VIC_SEX) = c("F",'M')

caret::confusionMatrix(BQ2_test$VIC_SEX, predicted_Test)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   F   M
##          F 137 267
##          M  72 332
##
##                Accuracy : 0.5804
##                  95% CI : (0.5456, 0.6147)
##     No Information Rate : 0.7413
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1609
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.6555
##             Specificity : 0.5543
##          Pos Pred Value : 0.3391
##          Neg Pred Value : 0.8218
##              Prevalence : 0.2587
##          Detection Rate : 0.1696
##    Detection Prevalence : 0.5000
##       Balanced Accuracy : 0.6049
##
##        'Positive' Class : F
##
```

Here, we again generated confusion matrics. As we can see above, it is the default threshold.

We see that our prediction model did not work well. We see that there is a false positive present.

Calculating and interpret the AUC

```
## Accuracy

auc.tmp <- performance(pred,"auc");
```

```
auc <- as.numeric(auc.tmp@y.values)
cat("ROC area under the curve for train is = ", auc, "\n")
```

```
## ROC area under the curve for train is =  0.5978836
```

```
predicted_Test <- predict(rf_default, BQ2_test)  # predicted scores

pred    <-    prediction(as.numeric(as.character(factor(predicted_Test,    labels    =    c(0,1)))),
         as.numeric(as.character(factor(BQ2_test$VIC_SEX, labels = c(0,1)))))


auc.tmp <- performance(pred,"auc");
auc <- as.numeric(auc.tmp@y.values)
cat("ROC area under the curve for test is = ", auc, "\n")
```

```
## ROC area under the curve for test is =  0.5804455
```

Let us dig deep and understand what the value means. We see above that the ROC area under the curve is = 0.598. This value is a measurement of the efficiency of the model. Receiver Operating Characteristic (ROC) curves are a popular way to visualize the tradeoffs between Sensitivity and Specificity in a binary classifier. The probabilistic interpretation is that if we randomly choose an optimistic case and a negative case, the probability that the positive case outranks the negative case according to the classifier is given by the AUC. The matrix cells enumerate all possible combinations of positive and negative cases, and the fraction under the curve comprises the cells where the positive case outranks the negative one.

---

3. Recognize the pattern/relationship between the victim's age and the Location, specifically Bar/Night Club, to recognize the trend throughout the years and increase the patrolling. *Major contribution by Aswin Kumar Rajendran*

Lastly, we will use the Generalized Linear regression and Linear regression model after finding out the relationship between the victim's age and the location. We have selected a couple of prime locations. Using these locations, we will create a model and recognize the trend throughout the years and increase the patrolling for those locations. We will create a model to predict which Location will need high patrolling and the driving factors for getting these results.

Preparing the data to be used here.

```
df <- subset(NYPD, LOCATION_DESC!="")
```

Seeing below, we see the unique values in each column. This is important to understand and then work on it.

```
unique(df$LOCATION_DESC)
```

```
##  [1] "UNKNOWN"                   "PVT HOUSE"
##  [3] "MULTI DWELL - APT BUILD"   "MULTI DWELL - PUBLIC HOUS"
##  [5] "GROCERY/BODEGA"            "GAS STATION"
##  [7] "FAST FOOD"                 "BAR/NIGHT CLUB"
##  [9] "COMMERCIAL BLDG"           "SOCIAL CLUB/POLICY LOCATI"
## [11] "HOSPITAL"                  "SUPERMARKET"
## [13] "LIQUOR STORE"              "HOTEL/MOTEL"
## [15] "RESTAURANT/DINER"          "SHOE STORE"
## [17] "DRUG STORE"                "NONE"
## [19] "DRY CLEANER/LAUNDRY"       "BEAUTY/NAIL SALON"
## [21] "STORE UNCLASSIFIED"        "SMALL MERCHANT"
## [23] "DEPT STORE"                "FACTORY/WAREHOUSE"
## [25] "CLOTHING BOUTIQUE"         "VARIETY STORE"
## [27] "JEWELRY STORE"             "TELECOMM. STORE"
## [29] "CHAIN STORE"               "CANDY STORE"
## [31] "VIDEO STORE"               "GYM/FITNESS FACILITY"
## [33] "ATM"                       "SCHOOL"
## [35] "PHOTO/COPY STORE"          "BANK"
## [37] "LOAN COMPANY"              "STORAGE FACILITY"
## [39] "CHECK CASH"                "DOCTOR/DENTIST"
```

```
unique(df$VIC_AGE_GROUP)
```

```
## [1] "25-44"   "18-24"   "45-64"   "<18"     "65+"     "UNKNOWN"
```

We had to clean the data, to avoid getting the incorrect details. As you saw above, the *Age Group* had *Unknown* values. We remove those rows and work with known Age groups.

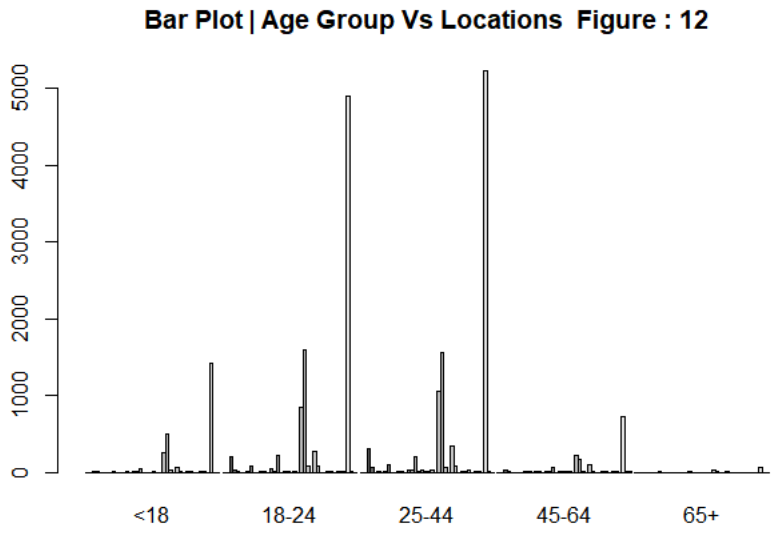
```
df <- subset(df, VIC_AGE_GROUP!="UNKNOWN")
#everything from the df variable excluding the victim age group unknown is stored in the variable
        df.



TAB <- table(df$LOCATION_DESC,df$VIC_AGE_GROUP) # age group and location are stored in TAB
```

We use a barplot to see the frequency of the victims. We see in all the age groups for all the locations. We see in the below figure that specific Age groups have a higher count in a particular location. We look into this more and use these findings for our further analysis.

```
barplot(TAB, beside = T, legend = F, main = paste("Bar Plot | Age Group Vs Locations
        ",GetFigureCount())) # bar plot shows the frequency of the the victims in all age groups
        in all the locations.
```



Before we dive in more, we must first confirm a significant relationship between the two features. Knowing this will ensure that we are not working with features that do not have sufficient dependency among them.

To determine the dependency, we use the Chi-Square Test of Independence

Step 1: State the hypothesis:

H_0: Location and Age group of the victim are independent. H_a: Location and Age group of the victim are dependent.

We have decided to go with alpha = 0.05 as the best value to determine the relationship.

```
alpha = 0.05
```

Step 2 and 3: Use the Chi-square test.

```
#Step 2: Use the Chi-square test.
CHI = chisq.test(TAB, correct = T)
CHI
```

```
##
##  Pearson's Chi-squared test
##
## data:  TAB
## X-squared = 664.06, df = 156, p-value < 2.2e-16
```

```
cat(ifelse(CHI$p.value < alpha, "**Reject the null Hypothesis**", "**Fail to reject the null
        hypothesis**" ))
```

```
## **Reject the null Hypothesis**
```

Step 4: Make the decision.

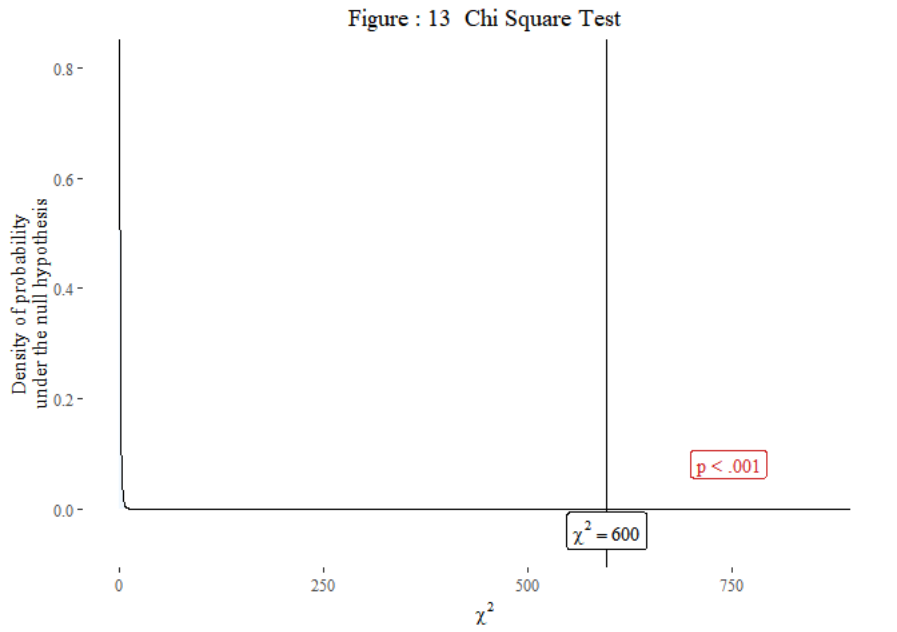
From the test, we know that the p-value is 7.6575464^{-64}, which is less than 0.05 . Since the p-value is less than 0.05, we would reject the null hypothesis.

Step 5: Summarize the results

Since the p-value is less than 0.05. We will reject the null hypothesis. This gives us enough evidence to conclude that the location and age group of the victim are dependent.

```
# attributes(CHI)
#
# CHI$expected

#Library(nhstplot) ## Used to plot Chi-Square
#Tittext = parse(expression(chi^2 ~ "Test"))
#Making a chi-squared plot with Chi-squared of 8 and df of 4
plotchisqtest(chisq = 597.16, df = 1, title = paste(GetFigureCount()," Chi Square Test"),
         signifdigitschisq=0.05)
```



We are good to proceed with our analysis and prediction modeling.

```
#For location Bar/Night Club


new_df <- as.data.frame.matrix(TAB)

new_df [3,] %>% t() %>%
kbl(caption = "location Bar/Night Club with Age Group",
      align = 'c') %>%
        kable_classic_2(full_width = T,
                        position = "center",
                        fixed_thead = T)
```

location Bar/Night Club with Age Group

	BAR/NIGHT CLUB

	BAR/NIGHT CLUB
<18	11
18-24	208
25-44	305
45-64	22
65+	0

From the above table, we see that in Bar/Night Club, 25-44 is the age group with the highest number of victims with 305, and the 18-24 age group is the second-highest number of victims with 208. Furthermore, the 65+ age group has no victims in this location.

This shows that there are a total of 546 victims in all five age groups for Bar/ Night Club.

```
# This shows that there are a total of 546 victims in all five age groups for Bar/ Night Club.
cat("The total number of insidents at the bar were ", sum(new_df[3,]))
```

```
## The total number of insidents at the bar were  546
```

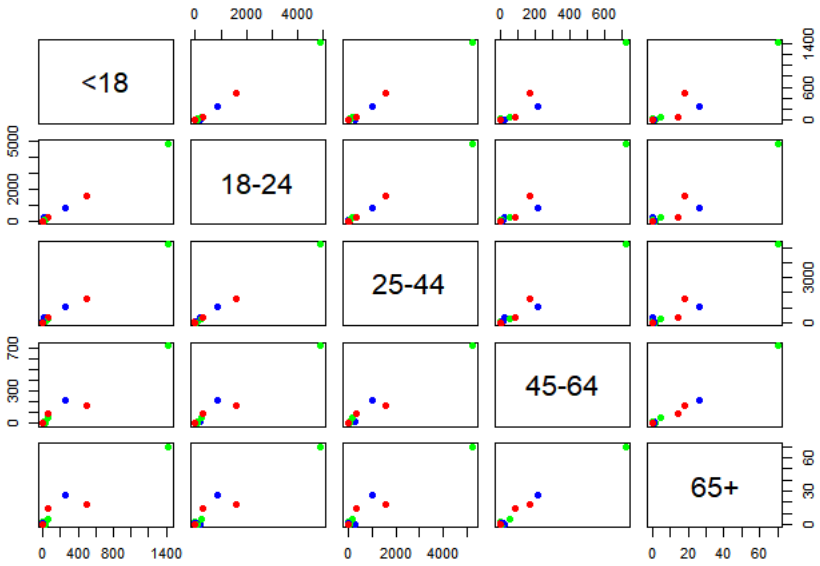
Displays the maximum number of the victim in an age group among all locations. Since we already know, 305 belongs to the 25-44 age group in Bar/ Night Club. Now we can say that bar/night club has the maximum victim in an age group.

```
#Max incidents in all the age groups
cat("The maximum number of insidnets that took place by any age group were",max(new_df[3,]))
```

```
## The maximum number of insidnets that took place by any age group were 305
```

Figure : 14 The Scatter plot for Age group and count of shooting

```
#scatterplot matrix is used to display how many victims are there in each age group
pairs(new_df, pch=21, #pch symbol
      bg = rainbow(3),  # Background color of the symbol (pch 21 to 25)
      col = rainbow(3)) #Border color of the symbol
```



From the above scatterplot, we can easily see the red dot. It shows the count of shooing by age groups at the bar.

```
#Filter for Bar/Night Club
df1 <- subset(df, LOCATION_DESC == "BAR/NIGHT CLUB")
df1$OCCUR_DATE <- as.character(df1$OCCUR_DATE)
df1$OCCUR_DATE <- substr(df1$OCCUR_DATE, nchar(df1$OCCUR_DATE)-3, nchar(df1$OCCUR_DATE) )
df1$OCCUR_DATE <- as.numeric(df1$OCCUR_DATE) #converted the dates and months to display in the
        format YYYY


#summary(df1[]) # displays the summary of the data only with bar/night club, and it has 549
        records
```

Now, we have made the data ready for modeling. We will first use the generalized linear model as discussed earlier for creating a model and fit it.

```
mdl <- glm(df1$OCCUR_DATE ~ df1$VIC_AGE_GROUP ,data = df1)
mdl
```

```
##
## Call:  glm(formula = df1$OCCUR_DATE ~ df1$VIC_AGE_GROUP, data = df1)
##
## Coefficients:
##            (Intercept)  df1$VIC_AGE_GROUP18-24  df1$VIC_AGE_GROUP25-44
##               2007.364                   2.218                   3.230
## df1$VIC_AGE_GROUP45-64
##                  5.591
##
## Degrees of Freedom: 545 Total (i.e. Null);  542 Residual
## Null Deviance:       7371
## Residual Deviance: 6990  AIC: 2952
```

```
summary(mdl)
```

```
##
## Call:
## glm(formula = df1$OCCUR_DATE ~ df1$VIC_AGE_GROUP, data = df1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.9545  -2.5934  -0.5934   1.4183   9.4183
##
## Coefficients:
##                         Estimate Std. Error  t value Pr(>|t|)
## (Intercept)             2007.364      1.083 1853.928  < 2e-16 ***
## df1$VIC_AGE_GROUP18-24    2.218      1.111    1.996  0.04639 *
## df1$VIC_AGE_GROUP25-44    3.230      1.102    2.931  0.00353 **
## df1$VIC_AGE_GROUP45-64    5.591      1.326    4.216 2.91e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 12.89612)
##
##     Null deviance: 7371.0  on 545  degrees of freedom
## Residual deviance: 6989.7  on 542  degrees of freedom
## AIC: 2951.5
##
## Number of Fisher Scoring iterations: 2
```

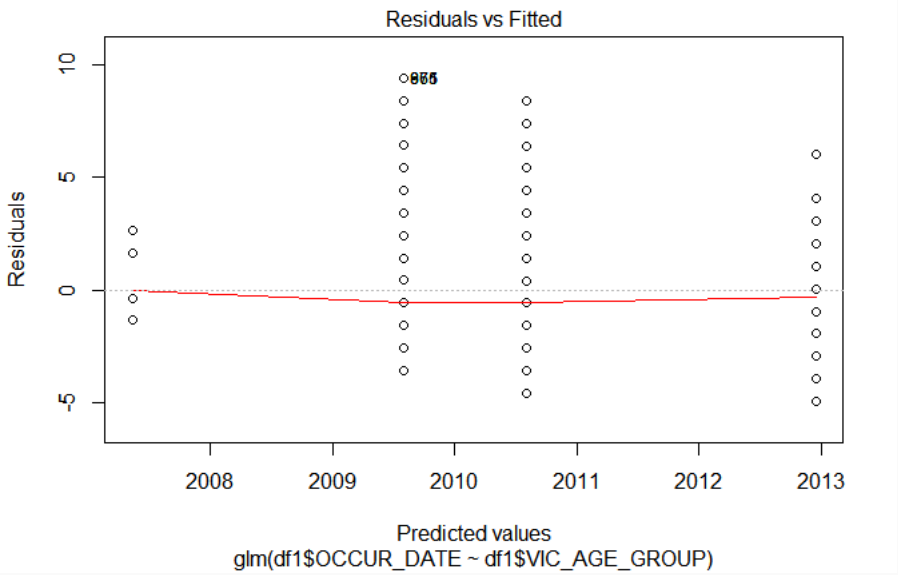
```
#hist(mdl$residuals, breaks = 100)
#which = 1:2)
```

We can see that there is very high significance in the variables. To determine the efficiency of the model, we used the Fisher Scoring and AIC value. We see that Fisher Scoring is two, and the AIC value is 2952. This is not a good score as our model is not performing as we expected.

From the p-value of the glm model, we can tell that the age group between 45-64 has a high chance of involving in an incident occurring. The age group, 25-44, is the second-highest age group involved in the incident. Furthermore, the age group between 18-24 is the third-highest age group involved in an incident and could be a victim.

Figure : 15 Plot the Residuals with Predicted Values (fitted)

```
plot(mdl, which = 1:1)
```



Now, we proceed with our next model. We will fit this model using the traditional linear regression algorithm for a model.

```
lmmdl <- lm(df1$OCCUR_DATE ~ df1$VIC_AGE_GROUP ,data = df1)
lmmdl
```

```
##
## Call:
## lm(formula = df1$OCCUR_DATE ~ df1$VIC_AGE_GROUP, data = df1)
##
## Coefficients:
##            (Intercept)  df1$VIC_AGE_GROUP18-24  df1$VIC_AGE_GROUP25-44
##               2007.364                   2.218                   3.230
## df1$VIC_AGE_GROUP45-64
##                  5.591
```

```
summary(lmmdl)
```

```
##
## Call:
## lm(formula = df1$OCCUR_DATE ~ df1$VIC_AGE_GROUP, data = df1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9545 -2.5934 -0.5934  1.4183  9.4183
##
## Coefficients:
##                       Estimate Std. Error  t value Pr(>|t|)
## (Intercept)           2007.364      1.083 1853.928  < 2e-16 ***
## df1$VIC_AGE_GROUP18-24    2.218      1.111    1.996  0.04639 *
## df1$VIC_AGE_GROUP25-44    3.230      1.102    2.931  0.00353 **
## df1$VIC_AGE_GROUP45-64    5.591      1.326    4.216 2.91e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.591 on 542 degrees of freedom
```

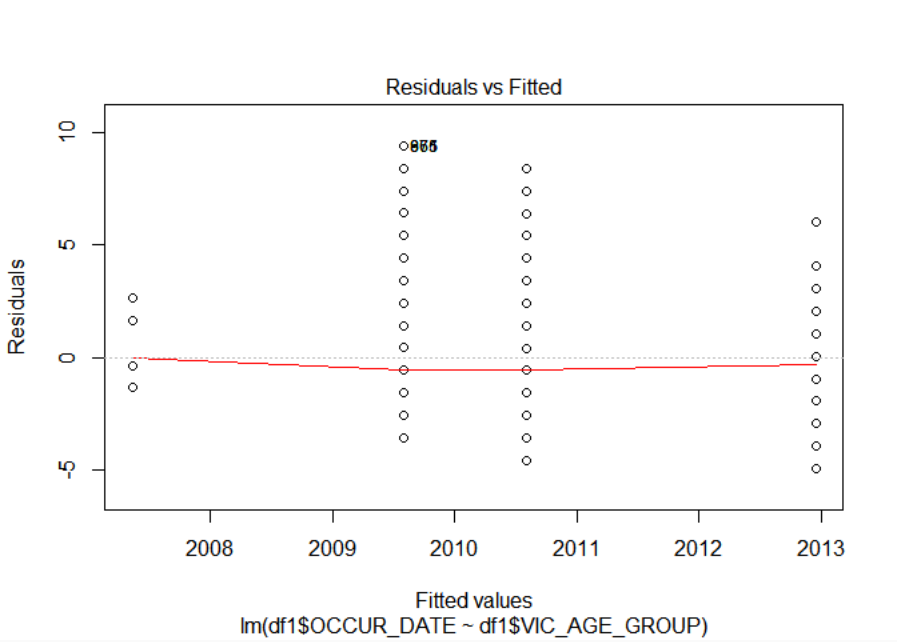
```
## Multiple R-squared:  0.05174,    Adjusted R-squared:  0.04649
## F-statistic: 9.857 on 3 and 542 DF,  p-value: 2.445e-06
```

We can see that there is very high significance in the variables. To determine the efficiency of the model, we used the Adjusted R-square. We see that the Adjusted R-square is 0.0465. This is not a good score as our model is not performing as we expected.

The occurrence date is the dependent response variable, and age is the independent variable. Linear Regression is done between occurrence data and victim age group.

Figure : 16 Plot the Residuals with Predicted Values (fitted)

```
plot(lmmdl, which = 1:1)
```



From the above plot, we see that the residuals are not evenly spread across the line. We can assume this is model will overfit the results.

From the p-value, we can tell that the age group between 45-64 has a high chance of an incident occurring. The age group, 25-44, is the second-highest age group # involved in the incident.

However, we move ahead and compare the performance of the models we just created. We will use ANOVA to compare the model's efficiency.

```
anova <- aov(df1$OCCUR_DATE~df1$VIC_AGE_GROUP ,data = df1)
anova
```

```
## Call:
##    aov(formula = df1$OCCUR_DATE ~ df1$VIC_AGE_GROUP, data = df1)
##
## Terms:
##                 df1$VIC_AGE_GROUP Residuals
## Sum of Squares            381.350  6989.697
## Deg. of Freedom                 3       542
##
```

```
## Residual standard error: 3.591117
## Estimated effects may be unbalanced
```

```
summary(anova)
```

```
##                    Df Sum Sq Mean Sq F value   Pr(>F)
## df1$VIC_AGE_GROUP   3    381   127.1   9.857 2.44e-06 ***
## Residuals         542   6990    12.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the p-value of ANOVA, we can say that there is a significant relationship between the occurrence date and victim age group.

From the above test results, we can say that there are more victims involved in this location irrespective of race, and I would suggest NYPD increase the patrolling to reduce crime.

Conclusion

To conclude this report, we have completed the preliminary (Exploratory Data Analysis - EDA) analysis. Furthermore, we Identified the methods we will be using to answer the questions and justify those methods. Out interpretation of the dataset is much more precise now than before. However, it was difficult to conclude the appropriate methods as we do not know if we may need to perform extra pre procedures to get the data ready for modeling. We also needed to know how we will handle missing or unknown values in the whole dataset creating the model. Will it affect the performance of the model or not. We performed the Chi-square test and implemented it. It was used to determine that a significant relationship exists between the victim's race—furthermore, the Perpetrator who died in the shooting incident. There is a relationship between the victim's Sex to the Perpetrator's Race and Age. There is a relationship between location and age group. Next, we implemented the glm method to model the relationship and predict the results. We concluded that BAR/NIGHT CLUB has a significant relationship with the age group between 45-64 getting shot in an incident.

####Lastly, we choose Random Forest as the best model after comparing several models. We implemented this model by using the Cross-validation and Grid to get the best parameters. This helped us create a model of the sex of the victim using the Perpetrator's age group and race. Model performance was better in Random forest. However, it reduced reasonably. We feel more in-depth analysis is to be done to determine why it happened and use different methods to determine if they performed similarly. Maybe the parameters affected the performances, or running it several folds did. Overall the analysis was fantastic. It gave us a glimpse of real-world problems and showed us where we lie tacking such business questions.

References

1. The City of New York. (2020, July 15). NYPD Shooting Incident Data (Historic) | NYC Open Data. NYC Open Data. https://data.cityofnewyork.us/Public-Safety/NYPD-ShootingIncident-Data-Historic-/833y-fsy8

2. Winston, A., & Winston, A. (2018, January 27). Transparency Advocates Win Release of NYPD "Predictive Policing" Documents. The Intercept. https://theintercept.com/2018/01/27/nypd-predictive-policing-documents-lawsuitcrime-forecasting-brennan/

3. Advanced Modeling. (n.d.). Retrieved April 29, 2021, from https://datascienceplus.com/category/advanced-modeling/O'H, V. A. P. B. B. (2012, April 23). Why Simpler

Models are Better. Retrieved April 29, 2021, from https://methodsblog.com/2012/04/23/simple-models-ftw/

4. Wenger, S. J. (2012, April 1). Assessing transferability of ecological models: an underappreciated aspect of statistical validation. Retrieved April 29, 2021, from https://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/j.2041-210X.2011.00170.x
5. What factors should I consider when choosing a predictive model technique? (2021, April 26). Retrieved April 29, 2021, from https://sebastianraschka.com/faq/docs/choosing-technique.html
6. Chi-Square Test of Independence in R - Easy Guides - Wiki - STHDA. Sthda.com. (2020). Retrieved May 16, 2021, from http://www.sthda.com/english/wiki/chi-square-test-of-independence-in-r.
7. Chapter 11 Categorical Predictors and Interactions | Applied Statistics with R. (2020, October 30). NA. https://daviddalpiaz.github.io/appliedstats/categorical-predictors-and-interactions.html
8. Coding for Categorical Variables in Regression Models | R Learning Modules. (n.d.). NA. Retrieved May 16, 2021, from https://stats.idre.ucla.edu/r/modules/coding-for-categorical-variables-in-regression-models/
9. GeeksforGeeks. (2020, October 12). Regression with Categorical Variables in R Programming. https://www.geeksforgeeks.org/regression-with-categorical-variables-in-r-programming/
10. Logit Regression | R Data Analysis Examples. (n.d.). Are UCLA. Retrieved May 10, 2021, from https://stats.idre.ucla.edu/r/dae/logit-regression/
11. Quick-R: Generalized Linear Models. (n.d.). Statmethods. Retrieved May 16, 2021, from https://www.statmethods.net/advstats/glm.html
12. Rungta, K. (2021, April 8). R Random Forest Tutorial with Example. Rungta Blog. https://www.guru99.com/r-random-forest-tutorial.html

Appendix

```
## in the appendix

headTail(NYPD)
```

```
##       OCCUR_DATE OCCUR_TIME          BORO PRECINCT JURISDICTION_CODE
## 1      8/23/2019   22:10:00        QUEENS      103                 0
## 2     11/27/2019   15:54:00         BRONX       40                 0
## 3       2/2/2019   19:40:00     MANHATTAN       23                 0
## 4     10/24/2019   00:52:00 STATEN ISLAND      121                 0
## ...         <NA>         NA          <NA>      ...               <NA>
## 21623 12/28/2018   21:52:00         BRONX       40                 0
## 21624   7/4/2009   16:00:00      BROOKLYN       67                 0
## 21625  9/17/2013   08:08:00     MANHATTAN       24                 0
## 21626  4/15/2007   01:05:00     MANHATTAN       33                 0
##                   LOCATION_DESC STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX
## 1                       UNKNOWN                   FALSE        UNKNOWN  UNKNOWN
## 2                       UNKNOWN                   FALSE            <18        M
## 3                       UNKNOWN                   FALSE          18-24        M
## 4                     PVT HOUSE                    TRUE          25-44        M
## ...                         <NA>                    <NA>          <NA>     <NA>
## 21623                   UNKNOWN                   FALSE        UNKNOWN  UNKNOWN
## 21624                   UNKNOWN                   FALSE          25-44        M
## 21625 MULTI DWELL - PUBLIC HOUS                    TRUE        UNKNOWN  UNKNOWN
## 21626                   UNKNOWN                   FALSE            <18        M
##            PERP_RACE VIC_AGE_GROUP VIC_SEX       VIC_RACE
## 1               <NA>         25-44       M          BLACK
## 2              BLACK         25-44       F          BLACK
## 3     WHITE HISPANIC         18-24       M BLACK HISPANIC
## 4              BLACK         25-44       F          BLACK
```

```
## ...              <NA>          <NA>    <NA>           <NA>
## 21623            <NA>         45-64       M          BLACK
## 21624           BLACK         25-44       M          BLACK
## 21625            <NA>         45-64       M WHITE HISPANIC
## 21626 WHITE HISPANIC           <18>       M BLACK HISPANIC
```

```
glimpse(NYPD)
```

```
## Rows: 21,626
## Columns: 13
## $ OCCUR_DATE              <chr> "8/23/2019", "11/27/2019", "2/2/2019", "10/...
## $ OCCUR_TIME              <time> 22:10:00, 15:54:00, 19:40:00, 00:52:00, 18...
## $ BORO                    <chr> "QUEENS", "BRONX", "MANHATTAN", "STATEN ISL...
## $ PRECINCT                <dbl> 103, 40, 23, 121, 46, 73, 81, 67, 114, 69, ...
## $ JURISDICTION_CODE       <chr> "0", "0", "0", "0", "0", "0", "0", "0", "2"...
## $ LOCATION_DESC           <chr> "UNKNOWN", "UNKNOWN", "UNKNOWN", "PVT HOUSE...
## $ STATISTICAL_MURDER_FLAG <lgl> FALSE, FALSE, FALSE, TRUE, FALSE, FALSE, FA...
## $ PERP_AGE_GROUP          <chr> "UNKNOWN", "<18", "18-24", "25-44", "25-44"...
## $ PERP_SEX                <chr> "UNKNOWN", "M", "M", "M", "M", "M", "M", "U...
## $ PERP_RACE               <chr> NA, "BLACK", "WHITE HISPANIC", "BLACK", "BL...
## $ VIC_AGE_GROUP           <chr> "25-44", "25-44", "18-24", "25-44", "18-24"...
## $ VIC_SEX                 <chr> "M", "F", "M", "F", "M", "M", "M", "M", "M"...
## $ VIC_RACE                <chr> "BLACK", "BLACK", "BLACK HISPANIC", "BLACK"...
```

```
str(NYPD)
```

```
## 'data.frame':    21626 obs. of  13 variables:
##  $ OCCUR_DATE             : chr  "8/23/2019" "11/27/2019" "2/2/2019" "10/24/2019" ...
##  $ OCCUR_TIME             : 'hms' num  22:10:00 15:54:00 19:40:00 00:52:00 ...
##   ..- attr(*, "units")= chr "secs"
##  $ BORO                   : chr  "QUEENS" "BRONX" "MANHATTAN" "STATEN ISLAND" ...
##  $ PRECINCT               : num  103 40 23 121 46 73 81 67 114 69 ...
##  $ JURISDICTION_CODE      : chr  "0" "0" "0" "0" ...
##  $ LOCATION_DESC          : chr  "UNKNOWN" "UNKNOWN" "UNKNOWN" "PVT HOUSE" ...
##  $ STATISTICAL_MURDER_FLAG: logi  FALSE FALSE FALSE TRUE FALSE FALSE ...
##  $ PERP_AGE_GROUP         : chr  "UNKNOWN" "<18" "18-24" "25-44" ...
##  $ PERP_SEX               : chr  "UNKNOWN" "M" "M" "M" ...
##  $ PERP_RACE              : chr  NA "BLACK" "WHITE HISPANIC" "BLACK" ...
##  $ 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" ...
```

```
### Removing unwanted variables/ columns
###
```

We have taken care of missing values. Lets us begin with our analysis and understand the dataset. Here we go through each variable/feature column and get meaningful insights from them. As we proceed, we will keep updating our dataset so that by the end, we get a complete dataset with variables and values we need for our further analysis.

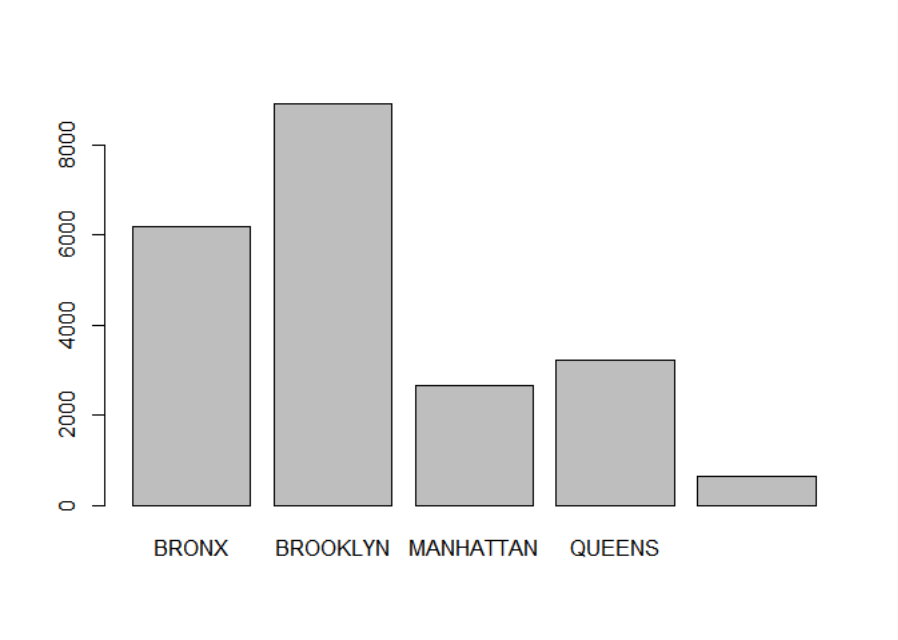
```
knitr::kable(freq(NYPD$BORO,na.rm = T), "pipe")
```

```
## Warning in plot.window(xlim, ylim, log = log, ...): "na.rm" is not a graphical
## parameter
```

```
## Warning in axis(if (horiz) 2 else 1, at = at.l, labels = names.arg, lty =
## axis.lty, : "na.rm" is not a graphical parameter
```

```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...): "na.rm"
## is not a graphical parameter
```

```
## Warning in axis(if (horiz) 1 else 2, cex.axis = cex.axis, ...): "na.rm" is not a
## graphical parameter
```



	Frequency	Percent
BRONX	6195	28.646074
BROOKLYN	8913	41.214279
MANHATTAN	2647	12.239896
QUEENS	3225	14.912605
STATEN ISLAND	646	2.987145
Total	21626	100.000000

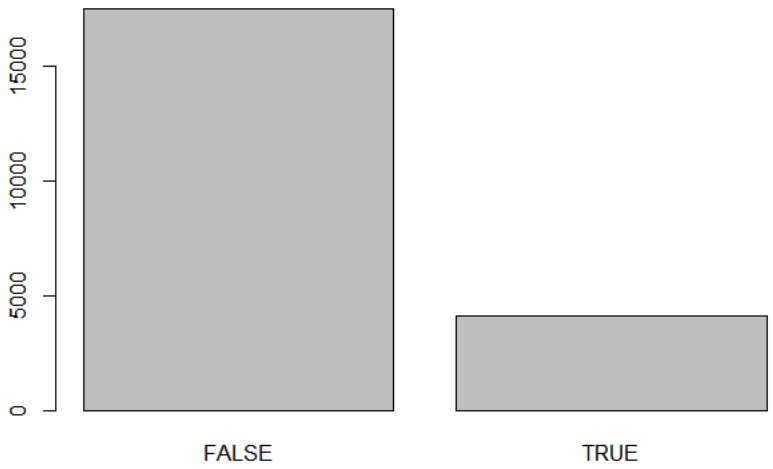
```
## We are interested in locations that have 100 and above shooting. We do this to avoid lengthy
        reports and outliers.

NYPD %>%
    select(LOCATION_DESC)%>%
     filter(LOCATION_DESC == c("BAR/NIGHT CLUB","COMMERCIAL BLDG" ,"GROCERY/BODEGA", "MULTI DWELL
         - APT BUILD","MULTI DWELL - PUBLIC HOUS","PVT HOUSE","RESTAURANT/DINER"))%>%
    freq(na.rm = T) %>% knitr::kable(., "pipe")
```

	Frequency	Percent	Valid Percent
c("MULTI DWELL - APT BUILD", "BAR/NIGHT CLUB", "PVT HOUSE", "MULTI DWELL - PUBLIC HOUS", "GROCERY/BODEGA", "COMMERCIAL BLDG", "RESTAURANT/DINER")	0	0	NaN
NA's	1	100	NA
Total	1	100	0

Now, we plot the location and see how the crime rate is spread across different locations. Here we are interested in seeing those locations which have shootings higher than 100. We see that a shocking figure of 45% of shooting was done at MULTI DWELL - PUBLIC HOUS We see that it correlates to 553 shooting in total. Later in the report, we will see the location details on Race and age group.

```
knitr::kable(freq(as.character(NYPD$STATISTICAL_MURDER_FLAG)), "pipe")
```



	Frequency	Percent
FALSE	17499	80.91649
TRUE	4127	19.08351
Total	21626	100.00000

Moving ahead, we try to visualize the deaths in the event of a shootout. We see fortunate there has not been a large number of deaths in this event. We see that in the event of a shootout, only about 19% of the time, there was a victim's death. In the further report, we will see race separation and understand the pattern with it.

```
### In our analysis, we found that there were garbage values in the age group. We remove them and
        do our analysis.


NYPD <-subset(NYPD, NYPD$PERP_AGE_GROUP %ni% c("940","1020","224"))   # Apply subset function
```

```
# freq(NYPD$PERP_AGE_GROUP,na.rm = T,plot=T)
# freq(NYPD$VIC_AGE_GROUP,na.rm = T,plot=T)


knitr::kable(freq(NYPD$PERP_AGE_GROUP,na.rm = T,plot=F), "pipe")
```

	Frequency	Percent
<18	1286	5.947371
18-24	5182	23.965222
25-44	4230	19.562503
45-64	414	1.914628
65+	51	0.235860

	Frequency	Percent
UNKNOWN	10460	48.374416
Total	21623	100.000000

```
knitr::kable(freq(NYPD$VIC_AGE_GROUP,na.rm = T,plot=F), "pipe")
```

	Frequency	Percent
<18	2391	11.0576701
18-24	8425	38.9631411
25-44	9226	42.6675299
45-64	1381	6.3867178
65+	142	0.6567081
UNKNOWN	58	0.2682329
Total	21623	100.0000000

Once we clean the variable and get only the cleaned data, we see that a stunning figure of 42.67% of shoutouts was done by people in the age group of 24-44, whereas the least number of shootouts were carried out by age group 65 and above. We also see that there are anonymous data as well. We see somewhat a pattern here. The younger generation is more indulged, and the older generation is less when compared.

```
knitr::kable(freq(NYPD$PERP_RACE,na.rm = T), "pipe")
```

```
## Warning in plot.window(xlim, ylim, log = log, ...): "na.rm" is not a graphical
## parameter
```

```
## Warning in axis(if (horiz) 2 else 1, at = at.l, labels = names.arg, lty =
## axis.lty, : "na.rm" is not a graphical parameter
```

```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...): "na.rm"
## is not a graphical parameter
```

```
## Warning in axis(if (horiz) 1 else 2, cex.axis = cex.axis, ...): "na.rm" is not a
## graphical parameter
```

	Frequency	Percent	Valid Percent
AMERICAN INDIAN/ALASKAN NATIVE	2	0.0092494	0.0139353
ASIAN / PACIFIC ISLANDER	105	0.4855940	0.7316054
BLACK	9335	43.1716228	65.0431996

	Frequency	Percent	Valid Percent
BLACK HISPANIC	1007	4.6570781	7.0164437
UNKNOWN	1839	8.5048328	12.8135452
WHITE	239	1.1053045	1.6652731
WHITE HISPANIC	1825	8.4400869	12.7159978
NA's	7271	33.6262313	NA
Total	21623	100.0000000	100.0000000

```
knitr::kable(freq(NYPD$VIC_RACE,na.rm = T), "pipe")
```

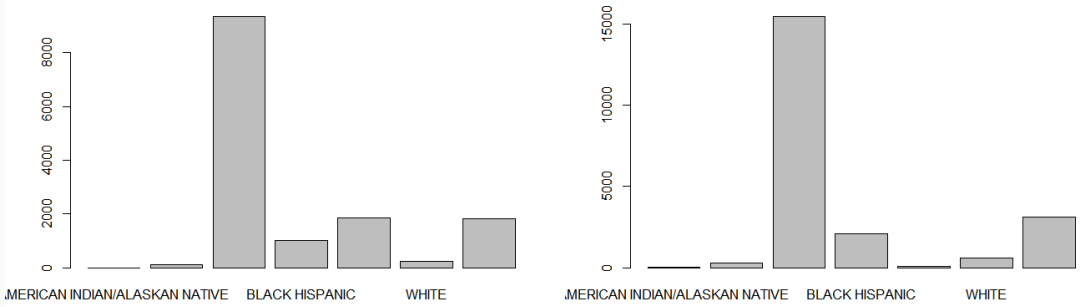
```
## Warning in plot.window(xlim, ylim, log = log, ...): "na.rm" is not a graphical
## parameter
```

```
## Warning in axis(if (horiz) 2 else 1, at = at.l, labels = names.arg, lty =
## axis.lty, : "na.rm" is not a graphical parameter
```

```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...): "na.rm"
## is not a graphical parameter
```

```
## Warning in axis(if (horiz) 1 else 2, cex.axis = cex.axis, ...): "na.rm" is not a
## graphical parameter
```

	Frequency	Percent
AMERICAN INDIAN/ALASKAN NATIVE	9	0.0416223
ASIAN / PACIFIC ISLANDER	286	1.3226657
BLACK	15469	71.5395644
BLACK HISPANIC	2085	9.6425103
UNKNOWN	93	0.4300976
WHITE	578	2.6730796
WHITE HISPANIC	3103	14.3504602
Total	21623	100.0000000



Now we see the shooting records as per the Race. We are not shocked to see that Black Race dominates this proportion. We also see that American Indian and Alaskan Native does the least number of shootouts. We can see that there is a surge in shooting by back Race as the second-highest Race to have indulged in shootouts is White Hispanic. However, the slope is very steep. That means there is a vast difference.

```
NYPD$PERP_SEX[NYPD$PERP_SEX=="UNKNOWN"] = "U"

knitr::kable(freq(NYPD$PERP_SEX,na.rm = T), "pipe")
```

```
## Warning in plot.window(xlim, ylim, log = log, ...): "na.rm" is not a graphical
## parameter
```

```
## Warning in axis(if (horiz) 2 else 1, at = at.l, labels = names.arg, lty =
## axis.lty, : "na.rm" is not a graphical parameter
```

```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...): "na.rm"
## is not a graphical parameter
```

```
## Warning in axis(if (horiz) 1 else 2, cex.axis = cex.axis, ...): "na.rm" is not a
## graphical parameter
```

	Frequency	Percent
F	306	1.41516
M	12544	58.01230
U	8773	40.57254
Total	21623	100.00000

```
knitr::kable(freq(NYPD$VIC_SEX,na.rm = T), "pipe")
```

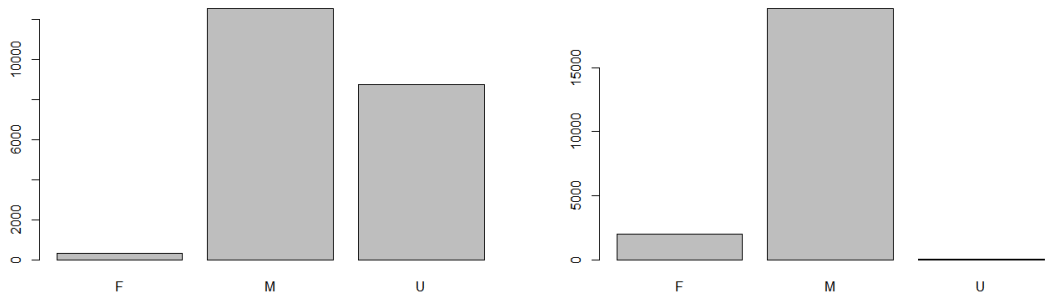
```
## Warning in plot.window(xlim, ylim, log = log, ...): "na.rm" is not a graphical
## parameter
```

```
## Warning in axis(if (horiz) 2 else 1, at = at.1, labels = names.arg, lty =
## axis.lty, : "na.rm" is not a graphical parameter
```

```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...): "na.rm"
## is not a graphical parameter
```

```
## Warning in axis(if (horiz) 1 else 2, cex.axis = cex.axis, ...): "na.rm" is not a
## graphical parameter
```

	Frequency	Percent
F	1999	9.2447856
M	19612	90.6997179
U	12	0.0554965
Total	21623	100.0000000



At last, we come toward the SexSex of perpetrators and victims. We see that Men are involved in shooting, and Men are the most victims in the shooting. However, we see that there are 1% of women (females) involved in the shooting. Unfortunate, the data is flawed. As we can see, 40% of the records for perpetrators are unknown. This record makes the data incomplete. We will have to exclude these rows when building models over this variable.

```
CT_PERP_RACE_SFlag <- round(prop.table(table(NYPD$PERP_RACE, NYPD$STATISTICAL_MURDER_FLAG))*100,
        2)

colnames(CT_PERP_RACE_SFlag) <- c("% of Victim's alive ","% of Victim's death")
kable(CT_PERP_RACE_SFlag, caption = "Perpertrator's race versus murder attempted") %>%
        kableExtra::kable_styling(.,position = "float_left")
```

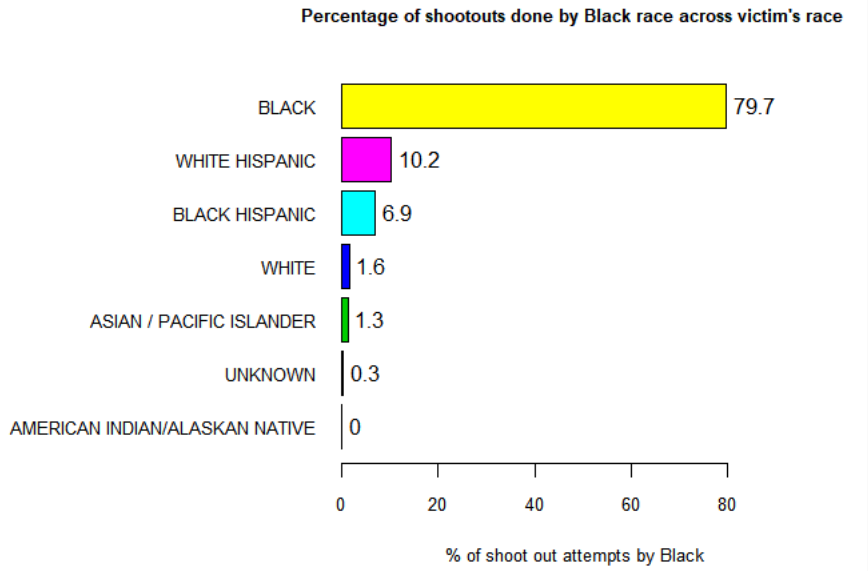
Perpertrator's race versus murder attempted

	% of Victim's alive	% of Victim's death
AMERICAN INDIAN/ALASKAN NATIVE	0.01	0.00
ASIAN / PACIFIC ISLANDER	0.50	0.23
BLACK	52.18	12.86

	% of Victim's alive	% of Victim's death
BLACK HISPANIC	5.71	1.30
UNKNOWN	11.98	0.84
WHITE	1.00	0.66
WHITE HISPANIC	9.71	3.00

```
subset_black_perp<- NYPD %>% filter(NYPD$PERP_RACE == "BLACK")

tab1(subset_black_perp$VIC_RACE,    cex.main=0.8,    cex.name=0.8,    cex.axis=0.8,    cex.lab=0.8,
         sort.group ="decreasing", bar.values ="percent", main = "Percentage of shootouts done by
         Black race across victim's race", xlab=" % of shoot out attempts by Black", ylab="Race
         of Victims") %>% knitr::kable(., "pipe")
```



x

subset_black_perp$VIC_RACE :

	Frequency	Percent	Cum. percent
BLACK	7439	79.7	79.7
WHITE HISPANIC	952	10.2	89.9
BLACK HISPANIC	640	6.9	96.7
WHITE	151	1.6	98.4
ASIAN / PACIFIC ISLANDER	117	1.3	99.6
UNKNOWN	32	0.3	100.0
AMERICAN INDIAN/ALASKAN NATIVE	4	0.0	100.0
Total	9335	100.0	100.0

As we can notice, the highest percentage of victims were killed by the Black Race, followed by White Hispanics. We see that Black perpetrators have killed the highest victims belonging to the Black Race itself on further analysis. Furthermore, second, the highest Race of victims targeted by black perpetrators is white Hispanic.