Data Mining Project 1

Paichana Kularb 57090015

**Business Understanding**

**1.1 What is the purpose of stop and frisk program?**

The purpose of the program is to eliminate weapons that are carried illegally and eventually reduce the number of crime rates.

**1.2 How do you define and measure effectiveness for such a program?**

I would measure the effectiveness of the program by averaging the numbers of crimes occur in a certain period of time during the program and when the program is on hold. The reason is to try and estimate the number of crime that are prevented by the program.

**1.3 What data would be needed to judge its effectiveness?**

The number of crimes occur in New York when the program is running and when the program is on hold.

**Data Understanding**

Type of attributes:

|  |  |  |
| --- | --- | --- |
| **Variable** | **Label** | **Data Type** |
| year | YEAR OF STOP (CCYY) | **Interval** |
| pct | PRECINCT OF STOP (FROM 1 TO 123) | **Nominal** |
| ser\_num | UF250 SERIAL NUMBER | **Nominal** |
| datestop | DATE OF STOP (MM-DD-YYYY) | **Interval** |
| timestop | TIME OF STOP (HH:MM) | **Interval** |
| recstat | RECORD STATUS | **Nominal** |
| inout | WAS STOP INSIDE OR OUTSIDE ? | **Nominal** |
| trhsloc | WAS LOCATION HOUSING OR TRANSIT AUTHORITY ? | **Nominal** |
| perobs | PERIOD OF OBSERVATION (MINUTES) | **ratio** |
| crimsusp | CRIME SUSPECTED | **Nominal** |
| perstop | PERIOD OF STOP (MINUTES) | **ratio** |
| typeofid | STOPPED PERSON'S IDENTIFICATION TYPE | **Nominal** |
| explnstp | DID OFFICER EXPLAIN REASON FOR STOP ? | **Nominal** |
| othpers | WERE OTHER PERSONS STOPPED, QUESTIONED OR FRISKED ? | **Nominal** |
| arstmade | WAS AN ARREST MADE ? | **Nominal** |
| arstoffn | OFFENSE SUSPECT ARRESTED FOR | **Nominal** |
| sumissue | WAS A SUMMONS ISSUED ? | **Nominal** |
| sumoffen | OFFENSE SUSPECT WAS SUMMONSED FOR | **Nominal** |
| compyear | COMPLAINT YEAR (IF COMPLAINT REPORT PREPARED) | **Interval** |
| comppct | COMPLAINT PRECINCT (IF COMPLAINT REPORT PREPARED) | **Nominal** |
| offunif | WAS OFFICER IN UNIFORM ? | **Nominal** |
| officrid | ID CARD PROVIDED BY OFFICER (IF NOT IN UNIFORM) | **Nominal** |
| frisked | WAS SUSPECT FRISKED ? | **Nominal** |
| searched | WAS SUSPECT SEARCHED ? | **Nominal** |
| contrabn | WAS CONTRABAND FOUND ON SUSPECT ? | **Nominal** |
| adtlrept | WERE ADDITIONAL REPORTS PREPARED ? | **Nominal** |
| pistol | WAS A PISTOL FOUND ON SUSPECT ? | **Nominal** |
| riflshot | WAS A RIFLE FOUND ON SUSPECT ? | **Nominal** |
| asltweap | WAS AN ASSAULT WEAPON FOUND ON SUSPECT ? | **Nominal** |
| knifcuti | WAS A KNIFE OR CUTTING INSTRUMENT FOUND ON SUSPECT ? | **Nominal** |
| machgun | WAS A MACHINE GUN FOUND ON SUSPECT ? | **Nominal** |
| othrweap | WAS ANOTHER TYPE OF WEAPON FOUND ON SUSPECT | **Nominal** |
| pf\_hands | PHYSICAL FORCE USED BY OFFICER - HANDS | **Nominal** |
| pf\_wall | PHYSICAL FORCE USED BY OFFICER - SUSPECT AGAINST WALL | **Nominal** |
| pf\_grnd | PHYSICAL FORCE USED BY OFFICER - SUSPECT ON GROUND | **Nominal** |
| pf\_drwep | PHYSICAL FORCE USED BY OFFICER - WEAPON DRAWN | **Nominal** |
| pf\_ptwep | PHYSICAL FORCE USED BY OFFICER - WEAPON POINTED | **Nominal** |
| pf\_baton | PHYSICAL FORCE USED BY OFFICER - BATON | **Nominal** |
| pf\_hcuff | PHYSICAL FORCE USED BY OFFICER - HANDCUFFS | **Nominal** |
| pf\_pepsp | PHYSICAL FORCE USED BY OFFICER - PEPPER SPRAY | **Nominal** |
| pf\_other | PHYSICAL FORCE USED BY OFFICER - OTHER | **Nominal** |
| radio | RADIO RUN | **Nominal** |
| ac\_rept | ADDITIONAL CIRCUMSTANCES - REPORT BY VICTIM/WITNESS/OFFICER | **Nominal** |
| ac\_inves | ADDITIONAL CIRCUMSTANCES - ONGOING INVESTIGATION | **Nominal** |
| rf\_vcrim | REASON FOR FRISK - VIOLENT CRIME SUSPECTED | **Nominal** |
| rf\_othsw | REASON FOR FRISK - OTHER SUSPICION OF WEAPONS | **Nominal** |
| ac\_proxm | ADDITIONAL CIRCUMSTANCES - PROXIMITY TO SCENE OF OFFENSE | **Nominal** |
| rf\_attir | REASON FOR FRISK - INAPPROPRIATE ATTIRE FOR SEASON | **Nominal** |
| cs\_objcs | REASON FOR STOP - CARRYING SUSPICIOUS OBJECT | **Nominal** |
| cs\_descr | REASON FOR STOP - FITS A RELEVANT DESCRIPTION | **Nominal** |
| cs\_casng | REASON FOR STOP - CASING A VICTIM OR LOCATION | **Nominal** |
| cs\_lkout | REASON FOR STOP - SUSPECT ACTING AS A LOOKOUT | **Nominal** |
| rf\_vcact | REASON FOR FRISK- ACTIONS OF ENGAGING IN A VIOLENT CRIME | **Nominal** |
| cs\_cloth | REASON FOR STOP - WEARING CLOTHES COMMONLY USED IN A CRIME | **Nominal** |
| cs\_drgtr | REASON FOR STOP - ACTIONS INDICATIVE OF A DRUG TRANSACTION | **Nominal** |
| ac\_evasv | ADDITIONAL CIRCUMSTANCES - EVASIVE RESPONSE TO QUESTIONING | **Nominal** |
| ac\_assoc | ADDITIONAL CIRCUMSTANCES - ASSOCIATING WITH KNOWN CRIMINALS | **Nominal** |
| cs\_furtv | REASON FOR STOP - FURTIVE MOVEMENTS | **Nominal** |
| rf\_rfcmp | REASON FOR FRISK - REFUSE TO COMPLY W OFFICER'S DIRECTIONS | **Nominal** |
| ac\_cgdir | ADDITIONAL CIRCUMSTANCES - CHANGE DIRECTION AT SIGHT OF OFFICER | **Nominal** |
| rf\_verbl | REASON FOR FRISK - VERBAL THREATS BY SUSPECT | **Nominal** |
| cs\_vcrim | REASON FOR STOP - ACTIONS OF ENGAGING IN A VIOLENT CRIME | **Nominal** |
| cs\_bulge | REASON FOR STOP - SUSPICIOUS BULGE | **Nominal** |
| cs\_other | REASON FOR STOP - OTHER | **Nominal** |
| ac\_incid | ADDITIONAL CIRCUMSTANCES - AREA HAS HIGH CRIME INCIDENCE | **Nominal** |
| ac\_time | ADDITIONAL CIRCUMSTANCES - TIME OF DAY FITS CRIME INCIDENCE | **Nominal** |
| rf\_knowl | REASON FOR FRISK - KNOWLEDGE OF SUSPECT'S PRIOR CRIM BEHAV | **Nominal** |
| ac\_stsnd | ADDITIONAL CIRCUMSTANCES - SIGHTS OR SOUNDS OF CRIMINAL ACTIVITY | **Nominal** |
| ac\_other | ADDITIONAL CIRCUMSTANCES - OTHER | **Nominal** |
| sb\_hdobj | BASIS OF SEARCH - HARD OBJECT | **Nominal** |
| sb\_outln | BASIS OF SEARCH - OUTLINE OF WEAPON | **Nominal** |
| sb\_admis | BASIS OF SEARCH - ADMISSION BY SUSPECT | **Nominal** |
| sb\_other | BASIS OF SEARCH - OTHER | **Nominal** |
| repcmd | REPORTING OFFICER'S COMMAND (1 TO 999) | **Nominal** |
| revcmd | REVIEWING OFFICER'S COMMAND (1 TO 999) | **Nominal** |
| rf\_furt | REASON FOR FRISK - FURTIVE MOVEMENTS | **Nominal** |
| rf\_bulg | REASON FOR FRISK - SUSPICIOUS BULGE | **Nominal** |
| offverb | VERBAL STATEMENT PROVIDED BY OFFICER (IF NOT IN UNIFORM) | **Nominal** |
| offshld | SHIELD PROVIDED BY OFFICER (IF NOT IN UNIFORM) | **Nominal** |
| forceuse | REASON FORCE USED | **Nominal** |
| sex | SUSPECT'S SEX | **Nominal** |
| race | SUSPECT'S RACE | **Nominal** |
| dob | SUSPECT'S DATE OF BIRTH (CCYY-MM-DD) | **Interval** |
| age | SUSPECT'S AGE | **Ratio** |
| ht\_feet | SUSPECT'S HEIGHT (FEET) | **Ratio** |
| ht\_inch | SUSPECT'S HEIGHT (INCHES) | **Ratio** |
| weight | SUSPECT'S WEIGHT | **Ratio** |
| haircolr | SUSPECT'S HAIRCOLOR | **Nominal** |
| eyecolor | SUSPECT'S EYE COLOR | **Nominal** |
| build | SUSPECT'S BUILD | **Ordinal** |
| addrtyp | LOCATION OF STOP ADDRESS TYPE | **Nominal** |
| rescode | LOCATION OF STOP RESIDENT CODE | **Nominal** |
| premtype | LOCATION OF STOP PREMISE TYPE | **Nominal** |
| premname | LOCATION OF STOP PREMISE NAME | **Nominal** |
| addrnum | LOCATION OF STOP ADDRESS NUMBER | **Ordinal** |
| stname | LOCATION OF STOP STREET NAME | **Nominal** |
| stinter | LOCATION OF STOP INTERSECTION | **Nominal** |
| crossst | LOCATION OF STOP CROSS STREET | **Nominal** |
| aptnum | LOCATION OF STOP APT NUMBER | **Ordinal** |
| city | LOCATION OF STOP CITY | **Nominal** |
| state | LOCATION OF STOP STATE | **Nominal** |
| zip | LOCATION OF STOP ZIP CODE | **Nominal** |
| addrpct | LOCATION OF STOP ADDRESS PRECINCT | **Nominal** |
| sector | LOCATION OF STOP SECTOR | **Nominal** |
| xcoord | LOCATION OF STOP X COORD | **Interval** |
| ycoord | LOCATION OF STOP Y COORD | **Interval** |
| detailCM | CRIME CODE DESCRIPTION | **Nominal** |

**R Mark Down**

The Data is first loaded into RStudio:

SQF = read.csv("/Users/boom/Desktop/Data mining/SQF\_2016.csv",na.strings=" ",stringsAsFactors = TRUE)

Remove last row which is all NA values by

df = SQF[0:(nrow(SQF)-1),]

## Verify data quality

### Locate missing values

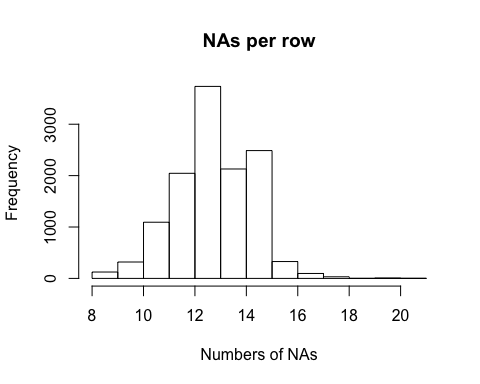
Columns with over 1 missing value can be seen by running

colNA = colSums(is.na(df))  
colNA[colNA>0]

## arstoffn sumoffen officrid offverb offshld forceuse dob addrtyp   
## 9762 12037 12236 9376 8401 9464 12404 12404   
## rescode premtype premname addrnum stname stinter crossst aptnum   
## 12404 12404 1295 7004 6989 43 43 12404   
## state zip sector xcoord ycoord   
## 12404 12404 120 351 351

There are too many rows that contains NA values so i have decided to plot a historgram show the number of NA values in each row

rowNA = rowSums(is.na(df))  
hist(rowNA,main = "NAs per row", xlab = "Numbers of NAs")



### Removing missing values

##### **arstoffn**

The arstoffn columns indicates why a suspect get arrested, and if the suspect didn't get arrested the values will be NA. Any row where no arrest is made (arst=='N'), arstoffn will be filled with 'NOARREST'

levels(df$arstoffn) <- c(levels(df$arstoffn),"NOARREST")  
df$arstoffn[df$arstmade=='N'] = 'NOARREST'   
summary(is.na(df$arstoffn))

## Mode FALSE TRUE   
## logical 12403 1

There is still 1 NA value when the arrested was made but no reason was given, the mode(except NOARREST) of the reason for arrest will use to fill in 1 NA.

df$arstoffn[is.na(df$arstoffn)] = "CPW"

##### **sumoffen**

The sumoffen column indicates why a suspect is summoned, and if no summoned is issued the values will be NA. Any row where the summon is not made(sumissue=='N'), sumoffen will be filled with 'NOSUMMON'

levels(df$sumoffen) <- c(levels(df$sumoffen),"NOSUMMON")  
df$sumoffen[df$sumissue=='N'] = 'NOSUMMON'  
summary(is.na(df$sumissue))

## Mode FALSE   
## logical 12404

##### **offcrid**

The offcrid column indicates whether an ID card is provided by the officer when not in uniform, and if the officer is in uniform the value will be null. Any row where officer is in uniform(offunif=='Y'), offcrid will be filled with 'INUNIFORM'

levels(df$officrid) <- c(levels(df$officrid),"INUNIFORM")  
df$officrid[df$offunif=='Y'] = 'INUNIFORM'  
summary(is.na(df$officrid))

## Mode FALSE TRUE   
## logical 8418 3986

There is still 3986 NA values after 'INUNIFORM' is filled. The NA values will be assume that officer didn't provide ID card even not in uniform. So NA values will be filled with 'N'

levels(df$officrid) <- c(levels(df$offunif),'N')  
df$officrid[df$offunif=='N' & is.na(df$officrid)] = 'N'

##### **offverb**

The offverb column indicates whether verbal statement is provided by the officer when not in uniform, and if the officer is in uniform the value will be null. Any row where officer is in uniform(offunif=='Y'), offverb will be filled with 'INUNIFORM'

levels(df$offverb) <- c(levels(df$offverb),"INUNIFORM")  
df$offverb[df$offunif=='Y'] = 'INUNIFORM'  
summary(df$offverb)

## V INUNIFORM NA's   
## 3028 8250 1126

There is still 1126 NA values after 'INUNIFORM' is filled. The NA values will be assume that officer didn't provide verbal statement even not in uniform. So NA values will be filled with 'N'

levels(df$offverb) <- c(levels(df$offverb),"N")  
df$offverb[df$offunif=='N' & is.na(df$offverb)] = 'N'

##### **offshld**

The offshld column indicates whether shield is provided by the officer when not in uniform, and if the officer is in uniform the value will be null. Any row where officer is in uniform(offunif=='Y'), offshld will be filled with 'INUNIFORM'

levels(df$offshld) <- c(levels(df$offshld),"INUNIFORM")  
df$offshld[df$offunif=='Y'] = 'INUNIFORM'  
summary(df$offshld)

## S INUNIFORM NA's   
## 4003 8250 151

There is still 151 NA values after 'INUNIFORM' is filled. The NA values will be assume that officer didn't show the shield even not in uniform. So NA values will be filled with 'N'

levels(df$offshld) <- c(levels(df$offshld),"N")  
df$offshld[df$offunif=='N' & is.na(df$offshld)] = 'N'

##### **stinter**

NA value in intersection will be assume that the frisk doesn't occur in the intersection and will be fill with NOTINTERSECTION.

levels(df$stinter) <- c(levels(df$stinter),"NOTINTERSECTION")  
df$stinter[is.na(df$stinter)] = "NOTINTERSECTION"

##### **crossst**

NA value in cross street will be assume that the frisk doesn't occur in the cross street and will be fill with NOTCROSSSTREET

levels(df$crossst) <- c(levels(df$crossst),"NOTCROSSSTREET")  
df$crossst[is.na(df$crossst)] = "NOTCROSSSTREET"

##### **xcoord and ycoord**

The xcoord and ycoord is the coordinates of the stop. This will be filled in based on the coordinate of each city. The average cooridinate of each city is found by

noCoorNull= df[!is.na(df$xcoord),c('xcoord','ycoord','city')]  
meanQUEENSx = mean(noCoorNull[noCoorNull$city=='QUEENS',1])  
meanQUEENSy = mean(noCoorNull[noCoorNull$city=='QUEENS',2])  
meanSTATENx = mean(noCoorNull[noCoorNull$city=='STATEN IS',1])  
meanSTATENy = mean(noCoorNull[noCoorNull$city=='STATEN IS',2])  
meanBROOKLYNx = mean(noCoorNull[noCoorNull$city=='BROOKLYN',1])  
meanBROOKLYNy = mean(noCoorNull[noCoorNull$city=='BROOKLYN',2])  
meanBRONXx = mean(noCoorNull[noCoorNull$city=='BRONX',1])  
meanBRONXy = mean(noCoorNull[noCoorNull$city=='BRONX',2])  
meanMANHATTANx = mean(noCoorNull[noCoorNull$city=='MANHATTAN',1])  
meanMANHATTANy = mean(noCoorNull[noCoorNull$city=='MANHATTAN',2])

The missing coordinate of each city is filled in with the average coordinate of each city

df$xcoord[is.na(df$xcoord) & df$city=='QUEENS'] = meanQUEENSx   
df$ycoord[is.na(df$ycoord) & df$city=='QUEENS'] = meanQUEENSy  
df$xcoord[is.na(df$xcoord) & df$city=='STATEN IS'] = meanSTATENx   
df$ycoord[is.na(df$ycoord) & df$city=='STATEN IS'] = meanSTATENy   
df$xcoord[is.na(df$xcoord) & df$city=='BROOKLYN'] = meanBROOKLYNx   
df$ycoord[is.na(df$ycoord) & df$city=='BROOKLYN'] = meanBROOKLYNy   
df$xcoord[is.na(df$xcoord) & df$city=='BRONX'] = meanBRONXx   
df$ycoord[is.na(df$ycoord) & df$city=='BRONX'] = meanBRONXy   
df$xcoord[is.na(df$xcoord) & df$city=='MANHATTAN'] = meanMANHATTANx   
df$ycoord[is.na(df$ycoord) & df$city=='MANHATTAN'] = meanMANHATTANy

##### **isforceuse (new variable)**

There are many types of physical force used by officer in the dataset, which are pf\_baton, pf\_drwep, pf\_grnd, pf\_hands, pf\_hcuff, pf\_other, pf\_pepsp, pf\_ptwep and pf\_wall. A column will be added to indicate whether force is used.

tempForce = (df$pf\_baton=='Y' | df$pf\_drwep == 'Y' | df$pf\_grnd =='Y' | df$pf\_hands =='Y' | df$pf\_hcuff=='Y' | df$pf\_other=='Y'| df$pf\_pepsp=='Y' | df$pf\_ptwep == 'Y' | df$pf\_wall=='Y')  
df$isforceuse[tempForce] ='Y'  
df$isforceuse[!tempForce] = 'N'  
df$isforceuse = factor(df$isforceuse)  
summary(df$isforceuse)

## N Y   
## 8531 3873

##### **forceuse**

The forceuse column indicates the reason why force is used by the officer , and if the no force is used the value will be null. Any row where no force is used(isforceuse=='N'), forceuse will be filled with 'NOFORCE'

levels(df$forceuse) <- c(levels(df$forceuse),"NOFORCE")  
df$forceuse[df$isforceuse=='N'] = 'NOFORCE'  
summary(is.na(df$forceuse))

## Mode FALSE TRUE   
## logical 11471 933

There is still 933 null values after 'NOFORCE' is filled. This will be dealth with later.

#### Further NA removal

recalulate NAs in each column

colNA = colSums(is.na(df))  
colNA[colNA>0]

## forceuse dob addrtyp rescode premtype premname addrnum stname   
## 933 12404 12404 12404 12404 1295 7004 6989   
## aptnum state zip sector   
## 12404 12404 12404 120

Columns that all its value is NA will be remove. stname and addrnum will also be remove because there is too much NAs and no promising way to fill in the values

df = df[,colNA!=nrow(df)]  
df = subset(df, select = -c(stname,addrnum))

##### **Remove rows**

Check how many rows contain more than 1 NA values

rowNA = rowSums(is.na(df))  
nrow(df[rowNA>1,])

## [1] 121

Only 121 rows so these will be removed by invoking

df = df[rowNA<=1,]

##### **Filling in remaing column with mode**

**Recalulate NAs in each column**

colNA = colSums(is.na(df))  
colNA[colNA>0]

## forceuse premname sector   
## 833 1182 88

There is only 3 columns with relatively low number of NA values and will be filled which the mode of each column

**premname**

modePremname = names(summary(df$premname)[(summary(df$premname)==max(summary(df$premname)))])  
df$premname[is.na(df$premname)] = modePremname  
modePremname

## [1] "STREET"

##### **sector**

modeSector = names(summary(df$sector)[(summary(df$sector)==max(summary(df$sector)))])  
df$sector[is.na(df$sector)] = modeSector  
modeSector

## [1] "B"

##### **forceuse**

modeForceuse = names(summary(df$forceuse[df$isforceuse=='Y'])[(summary(df$forceuse[df$isforceuse=='Y'])==max(summary(df$forceuse[df$isforceuse=='Y'])))])  
df$forceuse[is.na(df$forceuse)] = modeForceuse  
modeForceuse

## [1] "OT"

Check whether NA still exist in the dataset

colNA = colSums(is.na(df))  
colNA[colNA>0]

## named numeric(0)

### Adding new variable

##### **weaponFound**

In the dataset the weapons is categorize into many types. It will be simpified by creating a column to indicate whether any type of weapon is found. If weapon is found, weaponFound will be set as 'Y' otherwise 'N'

df$weaponfound = df$weaponfound = (df$pistol=='Y' | df$riflshot == 'Y' | df$asltweap =='Y' | df$knifcuti =='Y' | df$machgun=='Y' | df$othrweap=='Y')  
df$weaponfound[df$weaponfound] = 'Y'  
df$weaponfound[df$weaponfound=='FALSE'] = 'N'  
df$weaponfound = factor(df$weaponfound)

##### **day**

Add days of the week, Monday to Sunday.

df$datestop = as.Date(as.character(df$datestop),format="%m/%d/%Y")  
df$day = factor(weekdays(df$datestop),levels = c('Monday','Tuesday','Wednesday','Thursday','Friday','Saturday','Sunday'))

##### **height**

Add height in cm using the ht\_feet and ht\_inch column

df$height = ((df$ht\_feet\*12) + df$ht\_inch) \* 2.54

##### **bmi**

bmi column will store the body mass index of the suspect. First the weight needs to be converted to kilograms. The the formula for finding bmi can be applied.

df$weight = df$weight/2.2046  
df$bmi = df$weight/((df$height/100)\*\*2)

#### hours

Indicate time in hours with fractions. Hours + Minute/60

library(stringr)  
minutes = as.numeric(str\_sub(as.character(df$timestop),-2,-1))/60  
hours = as.numeric(substr(df$timestop,1,nchar(df$timestop)-2))  
hours[is.na(hours)] = 0  
df$hours = round(hours+minutes,2)

### Find duplicates

The condition of duplicate is flagged if age, height, datestop, weight and race are the same, and is likely to be the same person.

dupes = duplicated(df[c('age','height','datestop','weight','race')])  
df = df[!dupes,]  
summary(dupes)

## Mode FALSE TRUE   
## logical 12166 117

117 rows are found to be duplicate and only the duplicated row is remove from the dataset.

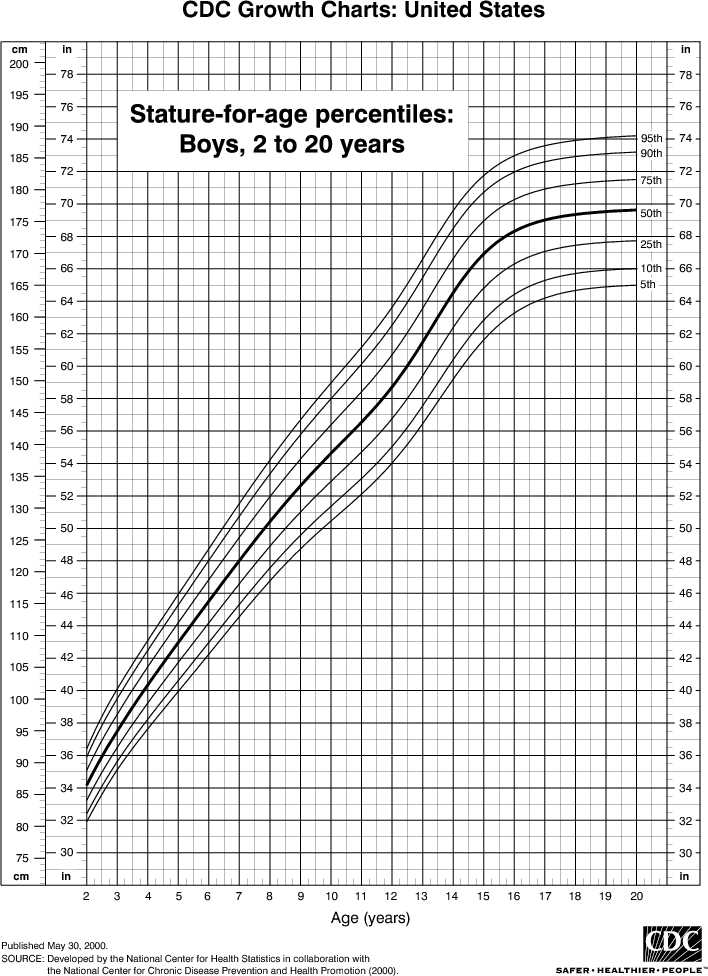
### Outliers

##### **Age**

Convert age and height in to numeric

df$age = as.numeric(df$age)  
df$height = as.numeric(df$height)

The height vs age chart is provided by CDC National Center for Health and will be use to help indicating outliers.

image: 

It seems unreasonble to stop and frisk a 5 years old child and the age will be replace with the average age

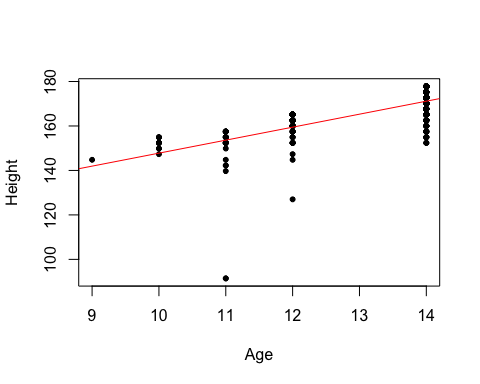
df$age[df$age<=5] = mean(df$age)

By looking at the height vs age chart suspect that is alot higher than usual is replaced with the mean age

df$age[df$age==6 & df$height >110] = mean(df$age)  
df$age[df$age==7 & df$height >135] = mean(df$age)  
df$age[df$age==8 & df$height >145] = mean(df$age)  
df$age[df$age==9 & df$height >150] = mean(df$age)  
df$age[df$age==10 & df$height >155] = mean(df$age)  
df$age[df$age==11 & df$height >160] = mean(df$age)  
df$age[df$age==12 & df$height >167] = mean(df$age)  
df$age[df$age==13 & df$height >175] = mean(df$age)  
df$age[df$age==14 & df$height >180] = mean(df$age)  
df$age[df$age==15 & df$height >185] = mean(df$age)  
df$age = round(df$age)

Plot age vs height to compare with the CDC data

plot(df$age[df$age<15],df$height[df$age<15],pch=20,xlab = "Age",ylab="Height")  
abline(lm(df$height[df$age<15]~ df$age[df$age<15]), col="red") # regression line (y~x)



Statistics of the age

summary(df$age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 9.0 19.0 24.0 26.2 30.0 82.0

The statistics of age seems possible and nothing else will be done to age.

##### **Weight and Height**

Looking at some statistics the invalid data can be seen

summary(df$weight)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.4536 68.0396 74.8435 76.8966 83.9154 453.1434

summary(df$height)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 91.44 170.18 175.26 174.55 180.34 210.82

Weight below 30 and above 200 will be consider as incorrect data and will be replace with average weight

df$weight[df$weight<30] = mean(df$weight)  
df$weight[df$weight>200] = mean(df$weight)

Height below 100 and above 250 will be consider as incorrect data and will be replace with the average height

df$height[df$height<30] = mean(df$height)  
df$height[df$height>200] = mean(df$height)

BMI will be help to use to indicate abnormal ratios between weight and height. [patient.info](https://patient.info/doctor/anorexia-nervosa-pro) stated that doctors consider BMI below 17.5 as 'Anorexia Nervosa'. BMI below 15 will be replaced with average weight and height. On the otherhand BMI above 60 will also be replaced with average weight and height.

#Recalculate BMI  
df$bmi = df$weight/((df$height/100)\*\*2)  
df$weight[df$bmi>50|df$bmi<15] = mean(df$weight)  
df$height[df$bmi>50|df$bmi<15] = mean(df$height)  
#Recalculate bmi  
df$bmi = df$weight/((df$height/100)\*\*2)

Statistic of BMI and Weight

summary(df$bmi)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 15.12 22.60 24.41 25.19 27.12 50.00

summary(df$weight)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 34.02 68.04 74.84 76.88 83.92 181.44

The statistic of bmi and weight looks to be possible and can't be consider as an invalid data.

## Appropriate Statistics of 10 attributes

##### **1.Age**

Statistical analysis of age can be shown by

summary(df$age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 9.0 19.0 24.0 26.2 30.0 82.0

paste("SD = ",round(sd(df$age),2)," ,IQR = ",round(IQR(df$age),2))

## [1] "SD = 10.06 ,IQR = 11"

The distribution of age looks right-skewed this could be because police believe younger pedestrians are likely to be have illegal weapons. Or it might be that most pedesdrians are on the younger side.

##### **2.Height(cm)**

Statistical analysis of height can be shown by

summary(df$height)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 121.9 170.2 175.3 174.6 180.3 198.1

paste("SD = ",round(sd(df$height),2)," ,IQR = ",round(IQR(df$height),2))

## [1] "SD = 8.29 ,IQR = 10.16"

The distribution of height of pedestrians that are stopped is almost symetric because the mean and median is very close. The SD and IQR is low and it appear to be the the distribution of the height doesn't spread out much.

##### **3.Body Mass Index**

Statistical analysis of BMI can be shown by

summary(df$bmi)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 15.12 22.60 24.41 25.19 27.12 50.00

paste("SD = ",round(sd(df$bmi),2)," ,IQR = ",round(IQR(df$bmi),2))

## [1] "SD = 4.13 ,IQR = 4.53"

BMI of the suspect looks almost symetric because the mean and median is close together. The standard deviation and IQR is quite low and shows that the BMI doesn't spread out much.

##### **4.Day of the Week**

The percentage of day of the week can be found by

round(summary(df$day)\*100/nrow(df),2)

## Monday Tuesday Wednesday Thursday Friday Saturday Sunday   
## 9.62 16.08 17.13 15.03 16.26 14.94 10.95

At first I thought that Friday and Saturday will have the most frequency because those are the days when people go out at night. But in fact the numbers of occurances on Tuesday to Saturday are similiar. While Sunday and Monday are relatively lower.

##### **5.Arrest made**

The percentage of each attributes in arstmade can be found by

round(summary(df$arstmade)\*100/nrow(df),2)

## N Y   
## 78.71 21.29

This shows that most of the people that are stopped doesn't not get arrested.

##### **6.Is force used**

The percentage of each attributes in forceused can be found by

round(summary(df$isforceuse)\*100/nrow(df),2)

## N Y   
## 69.33 30.67

This is interesting the percentage of force used is more than percentage of arrest made. Which means that officers uses force to stop pedestrians which some are innocent.

##### **7.Race**

The percentage of each attributes in race can be found by

round(summary(df$race)\*100/nrow(df),2)

## A B I P Q U W Z   
## 6.02 52.15 0.31 7.10 22.23 0.77 10.30 1.12

Source from [Wikipedia](https://en.wikipedia.org/wiki/Demographics_of_New_York_City) White: 44.6%, Black:25.1%, Hispanic:27.5% and Asian: 11.8% is the percentage of race in New York City. This clearly shows that police are bias into suspecting Blacks and flavour the Whites.

##### **8.Weapon Found**

The percentage of each attributes in weaponfound can be found by

round(summary(df$weaponfound)\*100/nrow(df),2)

## N Y   
## 94.25 5.75

It is surprising that the percentage of weapons way lower than that of the percentage of arrested. This means than many pedestrians get arrested while not having any weapon in hand.

##### **9.City**

The percentage of each attributes in city can be found by

round(summary(df$city)\*100/nrow(df),2)

## BRONX BROOKLYN MANHATTAN QUEENS STATEN IS   
## 19.77 28.74 20.14 26.08 5.27

Source from [Wikipedia](https://en.wikipedia.org/wiki/Demographics_of_New_York_City) Manhattan:19.25%, Bronx:17.06% ,Brooklyn:30.79% ,Queens:27.33% ,Staten Island:5.58%. The distribution of seems to be similiar. This tells that the police didn't specifically choose one city over the other.

##### **10.Sex**

The percentage of each attributes in sex can be found by

round(summary(df$sex)\*100/nrow(df),2)

## F M Z   
## 7.26 92.29 0.45

This is suprising that over 90% of the suspect are males. It clearly shows that polices are bias towards stopping males over female

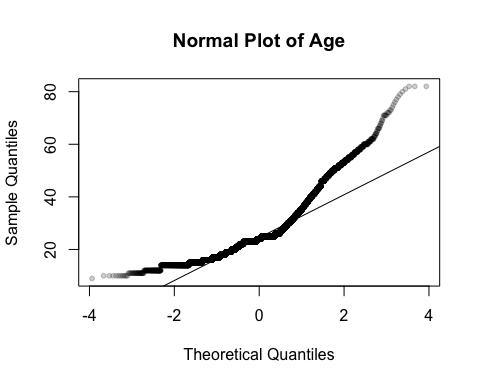
## Visualize 10 attributes

library(ggplot2)

##### **1.Age**

Normal plot of height can be shown by

qqnorm(df$age,pch=20,main ="Normal Plot of Age",col = rgb(0, 0, 0, 0.2))  
qqline(df$age)

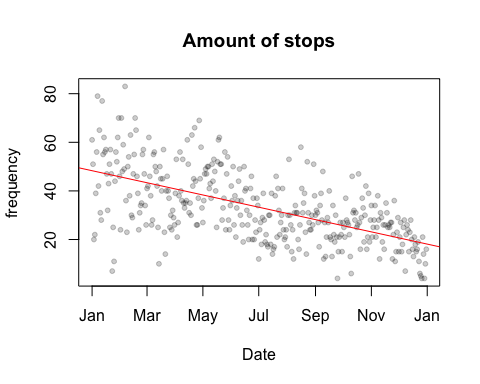


The plot is not in a straight line and shows that it is doesn't follow the normal distribution. The age distribution is skewed to the right.

##### **2.Amount of stops**

The number of stops per day can be plot by

a = aggregate(df$datestop, by=list(df$datestop),FUN =length)  
plot(a$Group.1,a$x,pch=20,main = "Amount of stops",xlab="Date",ylab="frequency",col = rgb(0, 0, 0, 0.2))  
abline(lm(a$x~ a$Group.1), col="red")

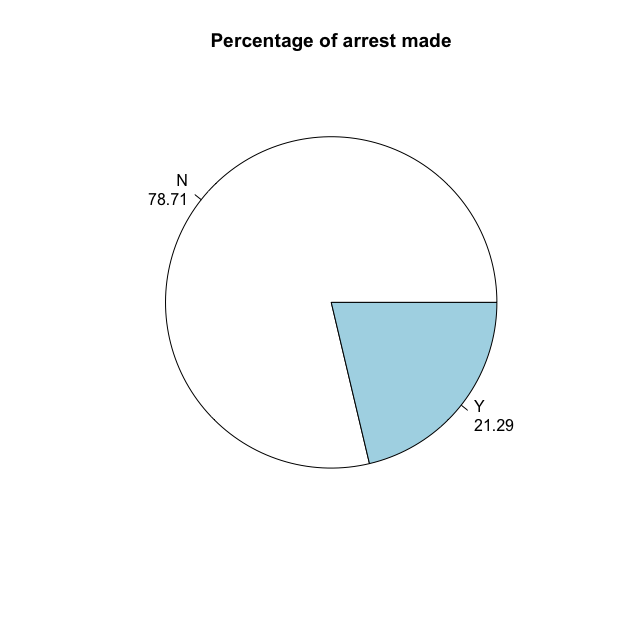


The fitted line shows that the number of stop decreases qucikly as time pass by. This could be because the officers are less active in the job? Or that they are more experience which leads to less random stops?

##### **3.Arrest made**

The piechart of whether the arrest is made is shown by

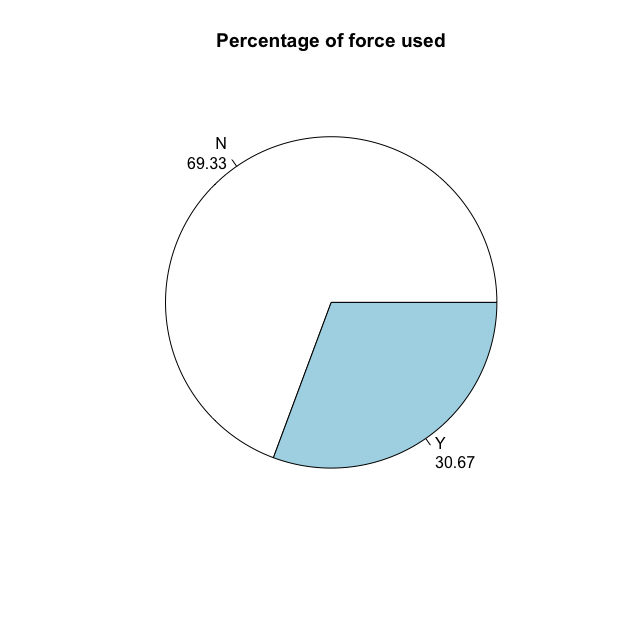
tb <- round(table(df$arstmade)/nrow(df),4)\*100  
lbls <- paste(names(tb), "\n", tb, sep="")  
pie(tb, labels = lbls, main="Percentage of arrest made")

 The piechart shows that most of the pedestrians that get stopped is innocent.

##### **4.Is force used**

The piechart of whether the force is used is shown by

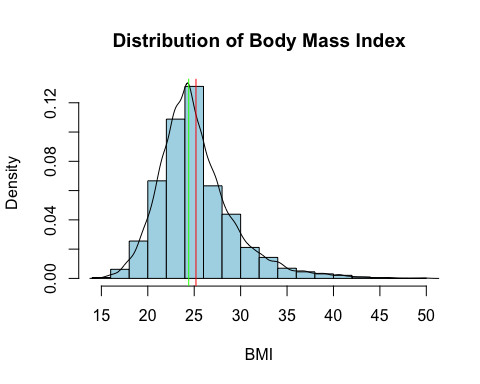
tb <- round(table(df$isforceuse)/nrow(df),4)\*100  
lbls <- paste(names(tb), "\n", tb, sep="")  
pie(tb, labels = lbls, main="Percentage of force used")

 The piechart shows that most of the pedestrians that get stopped is innocent.

##### **5.Body Mass Index**

The distribution of Body Mass Index can be shown by

hist(df$bmi, main = "Distribution of Body Mass Index",xlab ="BMI",col = "lightblue", prob = TRUE)  
lines(density(df$bmi))  
abline(v = mean(df$bmi), col = "red")  
abline(v = median(df$bmi), col = "green")

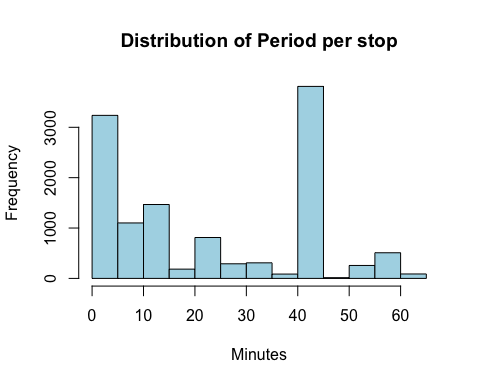


The read line shows the mean of BMI and the green line shows the median of BMI. The distribution of BMI is slightly right-skewed which means that there are more younger populations that are caught.

##### **6.Period of stop**

The histogram of the period of stop distribution can be shown by

df$perstop =as.numeric(df$perstop)  
hist(df$perstop, main = "Distribution of Period per stop",xlab ="Minutes",col = "lightblue",breaks = 20)

 By looking at the histogram of the period stop, the distribution doesn't follows the normal distribution. Most stop takes around 45-50 minutes and 0-5 minutes. The 40-45 minutes peak seems abnormal unless there is a reason for this.

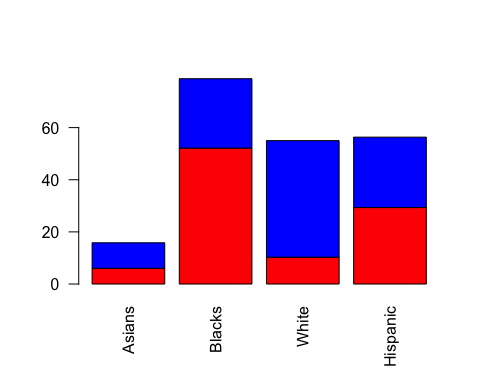
##### **7.Races**

The barchartshows the population of race distribution in New York City and the distribution of the race that are stopped. The red area shows in percentage of the race that are stopped. The blue area shows the percentage of race in New York City. Only 4 the main race are choosen.

temp = df$race  
levels(temp) <- c(levels(temp),"H")  
#combine black and white hispanics  
temp[temp =="P"|temp=="Q"] = "H"  
tb <- round(table(temp)/length(temp),4)\*100  
tb = tb[tb>2]  
tb

## temp  
## A B W H   
## 6.02 52.15 10.30 29.34

names(tb) = c("Asians","Blacks","White","Hispanic")  
tb = cbind(tb,c(9.8,26.6,44.7,27))  
tb = t(tb)  
barplot(tb,las=2,col = c("red","blue"))

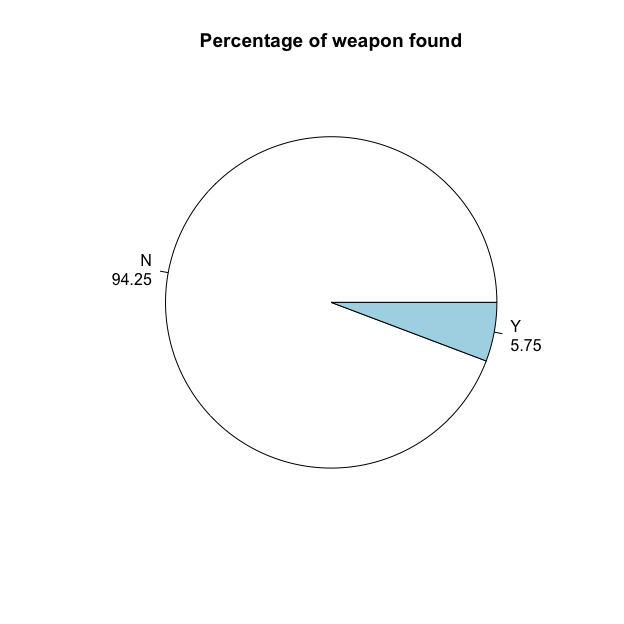


If the person stopped didn't depend on the race the area of each barchart should be the same. However, the percentage of whites that are stopped is far less than the percentage of whites in NYC. Moreover, The percentage the blacks are stopped is far more than the percentage of blacks in NYC. This shows that officers are bias in stopping blacks and not stopping whites. [Source for population in NYC](https://en.wikipedia.org/wiki/Demographics_of_New_York_City)

##### **8.Weapon Found**

Piechart of percentage to shows whether the weapons are found can be shown by

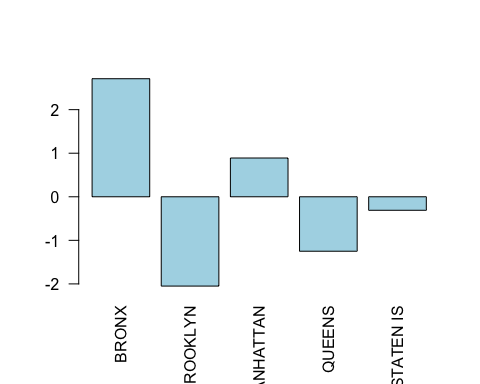
tb <- round(table(df$weaponfound)/nrow(df),4)\*100  
lbls <- paste(names(tb), "\n", tb, sep="")  
pie(tb, labels = lbls, main="Percentage of weapon found")



##### **9.City**

The barplot shows the difference between the percentage of the city that pedestrians were stop and the percentage population of each city.

tb <- round(table(df$city)/nrow(df),4)\*100  
tb = tb - c(17.06,30.79,19.25,27.33,5.58)  
barplot(tb,las=2,col = c("lightblue"))

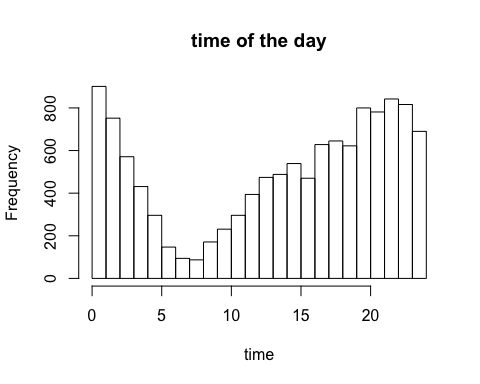


[Source](https://en.wikipedia.org/wiki/Demographics_of_New_York_City) The positive barchart shows that officers stop more pedestrian in this city. However, the values is so little that no conclusion can be made in whether pedestrian in which city is more likely to be stopped.

##### **10.Hours**

The histogram that shows what time of the day pedestrians are likely to get stopped are plotted by

hist(df$hours,breaks = 24,main = "time of the day",xlab = 'time')

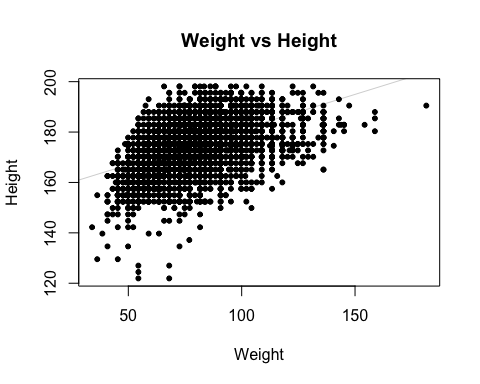


The stop mostly occur at night from 19:00 to 02:00 and least in the morning from 05:00 to 10:00. The explanation for this could be that people are hurry to work in the morning and officers doesn't want to interfere. While at night, crimes rate usually rise.

## 10 Relationship Between Attributes

##### **1.Weight and height**

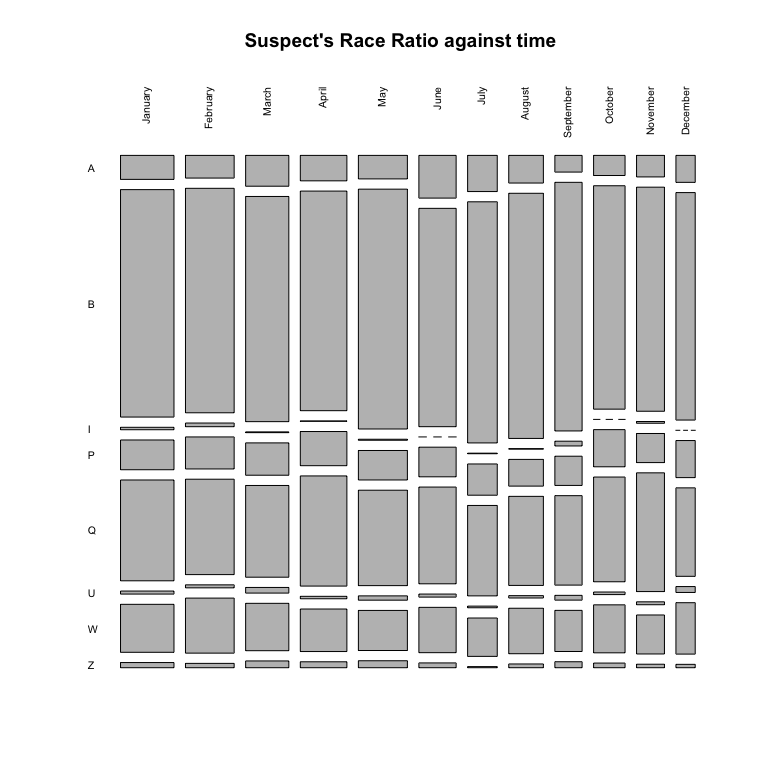
plot(df$weight,df$height,pch=20,main="Weight vs Height",xlab = "Weight",ylab = "Height")  
abline(lm(df$height~ df$weight),col = rgb(0, 0, 0, 0.2))



##### **2.Date and Race**

Plot to show how officer stop pedestrian base on race changes over time.

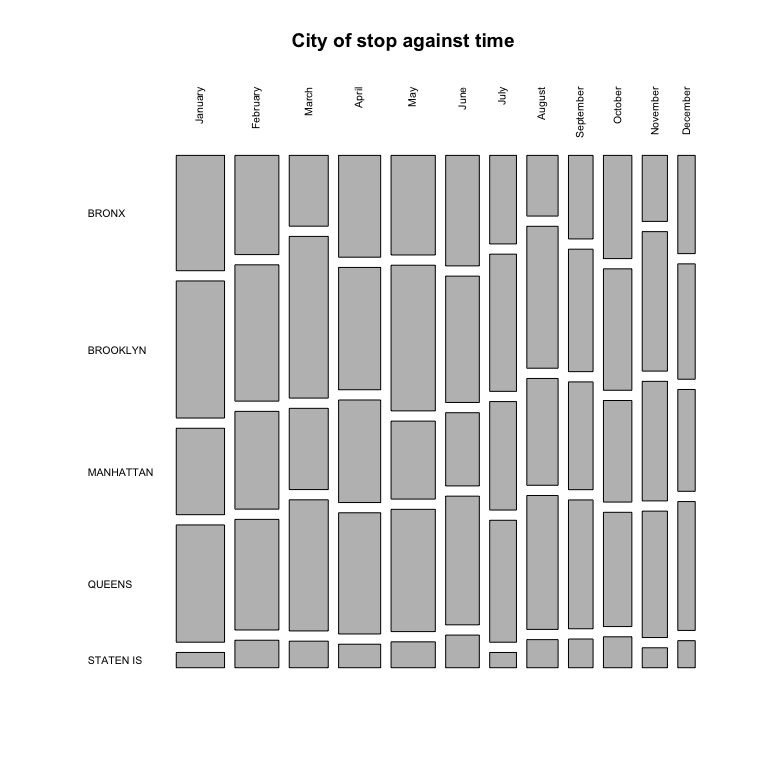
df$month = months(as.Date(df$datestop))  
df$month = factor(df$month,levels = c("January","February","March","April","May","June","July","August","September","October","November","December"))  
  
plot(table(df$month,df$race),las = 2,main = "Suspect's Race Ratio against time ")

 From the plot the chance of each race getting stop stay relatively constant. Which means that police still suspect the blacks more than the white throughout the year 2016.

##### **3.Date and city**

Plot to show how the amount of pedestrian stops change over time.

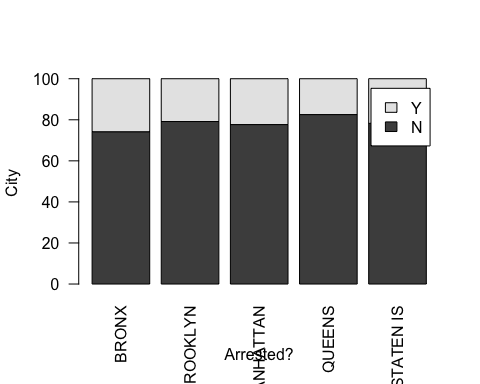
plot(table(df$month,df$city),las = 2 ,main = "City of stop against time")

 From the plot the width of the plot gets smaller which means that less stops occur. However, the amount of stops in each city stays about the same.

##### **4.Arrest made and City**

To show the ratio whether which city is likely to get arrested the graph can be plot as follows

tb = table(df$arstmade,df$city)  
for (x in 1:ncol(tb)){  
 tb[,x] = tb[,x]\*100/sum(tb[,x])  
}  
  
barplot(tb,las = 2,xlab = "Arrested?",ylab="City",legend = rownames(tb))



The height of the barchart is 100% and the darker area represents no arrest have been made. From the chart chances of getting arrest in different cities is very similiar.

##### **5.Arrest made and Race**

To show which race is likely to be arrested can be shown as follows

tb = table(df$arstmade,df$race)  
for (x in 1:ncol(tb)){  
 tb[,x] = tb[,x]\*100/sum(tb[,x])  
}  
barplot(tb,xlab = "Arrest?",ylab = "Race",legend = rownames(tb))

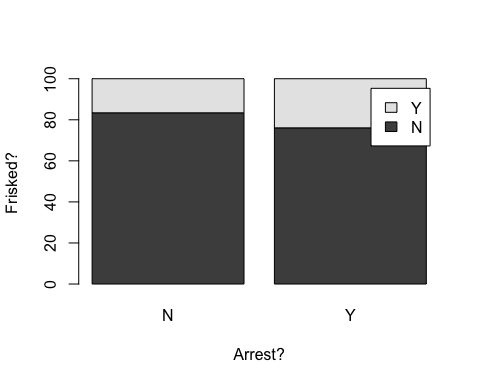


This doesn't not support the reason why blacks are getting stop more often and whites getting stop less. There doesn't not seem to be any relationship between getting arrested and race.

##### **6.Arrest made and Frisked**

To show the relationships between arrest made and frisked a chart can be plot by

tb = table(df$arstmade,df$frisked)  
for (x in 1:ncol(tb)){  
 tb[,x] = tb[,x]\*100/sum(tb[,x])  
}  
barplot(tb,xlab = "Arrest?",ylab = "Frisked?",legend = rownames(tb))



The darker area shows that the no arrest is made. The percentage of arrest is more when the suspect have been frisked. This make sense because if the police didn't frisk the suspect illegal items may not have been found.

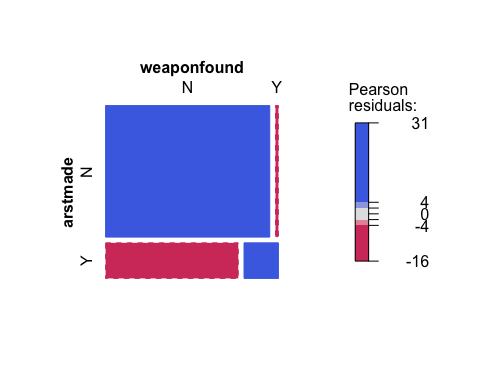
##### **7.Arrest made and Weapon found**

The relationship between weapon found and arrest made can be shown by

library(vcd)

## Loading required package: grid

mosaic(~arstmade+weaponfound,data = df,gp=shading\_Friendly2())

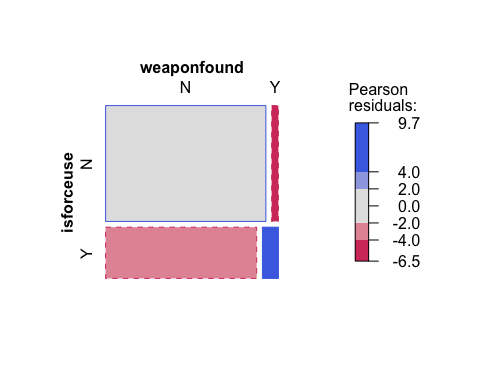


There seems to be a high correlation between weapon found and arrest mande which makes sense. If weapons are found on the suspect an arrest is likely to be made and vice versa.

##### **8.Is force used vs Weapon found**

The relationship between is force used and weapon found can be shown with a mosaic plot using 'vcd' library by

mosaic(~isforceuse+weaponfound,data = df,gp=shading\_Friendly2())

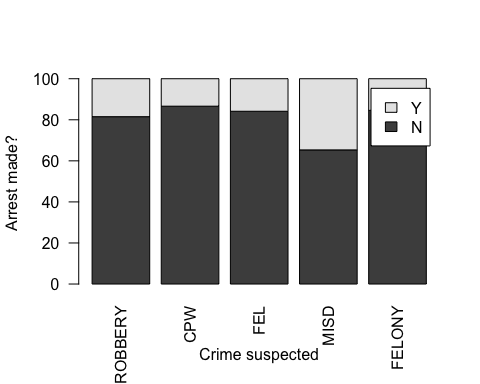


Looking at the plot there seems to be some relationship between this 2 attributes. If force is used it is likely that weapon have been found by the police.

##### **9.Crime suspected(top 5) and Arrest made**

The relationship between crime suspected and arrest made can be shown as follows

top = 5  
x = sort(summary(df$crimsusp)[names(summary(df$crimsusp))!="(Other)"])  
x = x[(length(x)-top+1):length(x)]  
  
tb = table(df$arstmade,df$crimsusp)  
tb = tb[,(names(x)[1:length(x)])]  
  
for (x in 1:ncol(tb)){  
 tb[,x] = tb[,x]\*100/sum(tb[,x])  
}  
barplot(tb,xlab = "Crime suspected",ylab = "Arrest made?",legend = rownames(tb),las = 2)

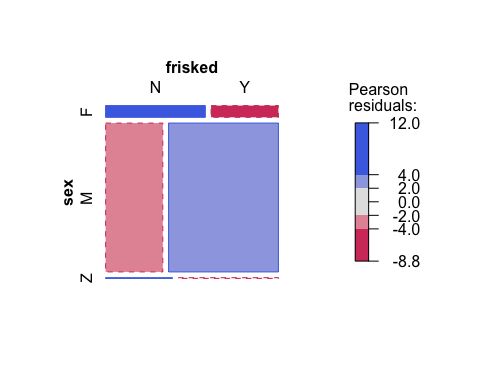


Looking at the graph, pedestrians that are suspected with MISD get arrested most often. This could be that officers suspect MISD crimes with higher accuracy.

##### **10. Sex and Frisked**

To see whether chance of getting frisked and sex have relationships among a mosaic plot can be created

mosaic(~sex+frisked,data = df,gp=shading\_Friendly2())

 There seems to be a relationship between frisked and sex. If the person is female the chance of getting frisked is low. However, if the person stopped is male the chance of getting frisked is slightly higher.