Capstone

install.packages("class") install.packages("gmodels") install.packages("ggplot2") library(ggplot2) library(class) library(gmodels) library(reshape2) library(dplyr)

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

Load the data to R

library(readr)  
#shooting\_data <- read\_csv("D:/datasets/police-KBP-2013-14-15-FULL-DFE.csv")  
shooting\_data <- read\_csv("https://www.figure-eight.com/wp-content/uploads/2016/03/police-KBP-2013-14-15-FULL-DFE.csv")

## Parsed with column specification:  
## cols(  
## Name = col\_character(),  
## `Cause of Death` = col\_character(),  
## `Date of Death` = col\_character(),  
## State = col\_character(),  
## City = col\_character(),  
## Gender = col\_character(),  
## Age = col\_integer(),  
## Race = col\_character(),  
## Photo = col\_character(),  
## `News Story` = col\_character(),  
## `Month of Death` = col\_character(),  
## `Year of Death` = col\_integer(),  
## `Was the deceased armed?` = col\_character(),  
## `Did the deceased have priors?` = col\_character(),  
## `Was the officer involved fired or suspended?` = col\_character()  
## )

Look at the data and understand the attributes

head(shooting\_data)

## # A tibble: 6 x 15  
## Name `Cause of Death` `Date of Death` State  
## <chr> <chr> <chr> <chr>  
## 1 Aaron Alexis Shooting 16/09/2013 Texas  
## 2 Aaron McDaniels Shooting 20/08/2013 Pennsylvania  
## 3 Abdul Kamal Shooting 11/11/13 New Jersey  
## 4 Adam Ignatz Bosch Shooting 7/7/13 California  
## 5 Adolfo Ramirez Shooting 23/11/2013 California  
## 6 Adolfo Vargas Tovar Shooting 24/05/2013 California  
## # ... with 11 more variables: City <chr>, Gender <chr>, Age <int>,  
## # Race <chr>, Photo <chr>, `News Story` <chr>, `Month of Death` <chr>,  
## # `Year of Death` <int>, `Was the deceased armed?` <chr>, `Did the  
## # deceased have priors?` <chr>, `Was the officer involved fired or  
## # suspended?` <chr>

tail(shooting\_data)

## # A tibble: 6 x 15  
## Name `Cause of Death` `Date of Death` State  
## <chr> <chr> <chr> <chr>  
## 1 William Dean Poole Shooting 16/3/15 North Carolina  
## 2 William J. Dick III Taser Gun 6/4/15 Washington  
## 3 William L. Chapman II Shooting 22/4/15 Virginia  
## 4 William Russell Smith Shooting 10/3/15 Alabama  
## 5 Yvette Henderson Shooting 3/2/15 California  
## 6 Zaki Shinwary Shooting 16/1/15 California  
## # ... with 11 more variables: City <chr>, Gender <chr>, Age <int>,  
## # Race <chr>, Photo <chr>, `News Story` <chr>, `Month of Death` <chr>,  
## # `Year of Death` <int>, `Was the deceased armed?` <chr>, `Did the  
## # deceased have priors?` <chr>, `Was the officer involved fired or  
## # suspended?` <chr>

str(shooting\_data)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 2355 obs. of 15 variables:  
## $ Name : chr "Aaron Alexis" "Aaron McDaniels" "Abdul Kamal" "Adam Ignatz Bosch" ...  
## $ Cause of Death : chr "Shooting" "Shooting" "Shooting" "Shooting" ...  
## $ Date of Death : chr "16/09/2013" "20/08/2013" "11/11/13" "7/7/13" ...  
## $ State : chr "Texas" "Pennsylvania" "New Jersey" "California" ...  
## $ City : chr "Fort Worth, Texas" "North Philadelphia, Pennsylvania" NA NA ...  
## $ Gender : chr "Male" "Male" "Male" "Male" ...  
## $ Age : int 34 19 30 29 37 47 27 22 21 52 ...  
## $ Race : chr "Black" "Unknown" "Black" "White" ...  
## $ Photo : chr "http://www.killedbypolice.net/victims/130475.jpg" NA "http://www.killedbypolice.net/victims/130638.jpg" "http://www.killedbypolice.net/victims/130255.jpg" ...  
## $ News Story : chr "http://www.washingtonpost.com/politics/aaron-alexis-34-is-dead-gunman-in-navy-yard-shooting-authorities-say/201"| \_\_truncated\_\_ "http://www.nbcphiladelphia.com/news/local/Man-Shot-by-Police-Officer-in-North-Philly-220436941.html" "http://www.dailyjournal.net/view/story/a4d38b31dfc440dbba89140724d52838/NJ--Irvington-Police-Shooting/" "http://www.pe.com/local-news/riverside-county/riverside/riverside-headlines-index/20130707-jurupa-valley-norco-"| \_\_truncated\_\_ ...  
## $ Month of Death : chr "Sep-13" "Aug-13" "Nov-13" "Jul-13" ...  
## $ Year of Death : int 2013 2013 2013 2013 2013 2013 2013 2013 2013 2013 ...  
## $ Was the deceased armed? : chr "Yes" "Yes" "Unclear" "Unclear" ...  
## $ Did the deceased have priors? : chr "Yes" "Yes" "Unclear" "Unclear" ...  
## $ Was the officer involved fired or suspended?: chr "Unclear" "Yes" "Unclear" "Unclear" ...  
## - attr(\*, "spec")=List of 2  
## ..$ cols :List of 15  
## .. ..$ Name : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## .. ..$ Cause of Death : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## .. ..$ Date of Death : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## .. ..$ State : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## .. ..$ City : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## .. ..$ Gender : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## .. ..$ Age : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_integer" "collector"  
## .. ..$ Race : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## .. ..$ Photo : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## .. ..$ News Story : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## .. ..$ Month of Death : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## .. ..$ Year of Death : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_integer" "collector"  
## .. ..$ Was the deceased armed? : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## .. ..$ Did the deceased have priors? : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## .. ..$ Was the officer involved fired or suspended?: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## ..$ default: list()  
## .. ..- attr(\*, "class")= chr "collector\_guess" "collector"  
## ..- attr(\*, "class")= chr "col\_spec"

summary(shooting\_data)

## Name Cause of Death Date of Death   
## Length:2355 Length:2355 Length:2355   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## State City Gender Age   
## Length:2355 Length:2355 Length:2355 Min. : 5.00   
## Class :character Class :character Class :character 1st Qu.: 26.00   
## Mode :character Mode :character Mode :character Median : 34.00   
## Mean : 36.91   
## 3rd Qu.: 46.00   
## Max. :107.00   
## NA's :76   
## Race Photo News Story   
## Length:2355 Length:2355 Length:2355   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## Month of Death Year of Death Was the deceased armed?  
## Length:2355 Min. :2013 Length:2355   
## Class :character 1st Qu.:2013 Class :character   
## Mode :character Median :2014 Mode :character   
## Mean :2014   
## 3rd Qu.:2014   
## Max. :2015   
##   
## Did the deceased have priors?  
## Length:2355   
## Class :character   
## Mode :character   
##   
##   
##   
##   
## Was the officer involved fired or suspended?  
## Length:2355   
## Class :character   
## Mode :character   
##   
##   
##   
##

Check the missing values of the "Cause of Death", "State", "Gender", "age", "Race", "Was the deceased armed?" and "Did the deceased have priors?" attributes.

sum(is.na(shooting\_data$`Cause of Death`) == TRUE)

## [1] 0

length(shooting\_data$`Cause of Death`)

## [1] 2355

sum(is.na(shooting\_data$`State`) == TRUE)

## [1] 0

length(shooting\_data$`State`)

## [1] 2355

sum(is.na(shooting\_data$`Gender`) == TRUE)

## [1] 1

length(shooting\_data$`Gender`)

## [1] 2355

sum(is.na(shooting\_data$`Age`) == TRUE)

## [1] 76

length(shooting\_data$`Age`)

## [1] 2355

sum(is.na(shooting\_data$Race) == T)

## [1] 2

length(shooting\_data$Race)

## [1] 2355

sum(is.na(shooting\_data$`Was the deceased armed?`) == TRUE)

## [1] 87

length(shooting\_data$`Was the deceased armed?`)

## [1] 2355

sum(is.na(shooting\_data$`Did the deceased have priors?`) == TRUE)

## [1] 195

length(shooting\_data$`Did the deceased have priors?`)

## [1] 2355

Replace "Gender" attribute missing value with "Male"

str(shooting\_data$Gender)

## chr [1:2355] "Male" "Male" "Male" "Male" "Male" "Male" "Male" "Male" ...

table(shooting\_data$Gender, useNA = "always")

##   
## Female male Male Unknown <NA>   
## 142 3 2207 2 1

shooting\_data$Gender<-replace(shooting\_data$Gender,shooting\_data$Gender == "", NA)  
shooting\_data$Gender[which(is.na(shooting\_data$Gender))] = 'Male'  
sum(is.na(shooting\_data$Gender))

## [1] 0

Replase "Age" attribute missing values with mean age.

str(shooting\_data$Age)

## int [1:2355] 34 19 30 29 37 47 27 22 21 52 ...

mean.age<- mean(shooting\_data$Age[!is.na(shooting\_data$Age)])  
  
mean.age <- round(mean.age, digits = 0)  
mean.age

## [1] 37

shooting\_data$Age[is.na(shooting\_data$Age)] = mean.age  
sum(is.na(shooting\_data$Age))

## [1] 0

Replace "Race" attribute missing value with "Unknown"

str(shooting\_data$Race)

## chr [1:2355] "Black" "Unknown" "Black" "White" "Latino" "Unknown" ...

table(shooting\_data$Race, useNA = "always")

##   
## Asian Black Latino Native American   
## 27 511 319 16   
## Other Pacific Islander Unknown White   
## 8 3 703 766   
## <NA>   
## 2

shooting\_data$Race<-replace(shooting\_data$Race,shooting\_data$Race == "", NA)  
shooting\_data$Race[which(is.na(shooting\_data$Race))] = 'Unknown'  
sum(is.na(shooting\_data$Race))

## [1] 0

Replace "Was the deceased armed?" attribute missing value with "Unknown".

str(shooting\_data$`Was the deceased armed?`)

## chr [1:2355] "Yes" "Yes" "Unclear" "Unclear" "Yes" "Yes" "Yes" "Yes" ...

table(shooting\_data$`Was the deceased armed?`, useNA = "always")

##   
## Disputed No Unclear Unknown Yes <NA>   
## 48 275 452 11 1482 87

shooting\_data$`Was the deceased armed?`<-replace(shooting\_data$`Was the deceased armed?`,shooting\_data$`Was the deceased armed?` == "", NA)  
shooting\_data$`Was the deceased armed?`[which(is.na(shooting\_data$`Was the deceased armed?`))] = 'Unknown'  
sum(is.na(shooting\_data$`Was the deceased armed?`))

## [1] 0

Replace "Did the deceased have priors?" attribute missing value with "Unclear".

str(shooting\_data$`Did the deceased have priors?`)

## chr [1:2355] "Yes" "Yes" "Unclear" "Unclear" "Unclear" "Unclear" ...

table(shooting\_data$`Did the deceased have priors?`, useNA = "always")

##   
## No Unclear Yes <NA>   
## 117 1608 435 195

shooting\_data$`Did the deceased have priors?`<-replace(shooting\_data$`Did the deceased have priors?`,shooting\_data$`Did the deceased have priors?` == "", NA)  
shooting\_data$`Did the deceased have priors?`[which(is.na(shooting\_data$`Did the deceased have priors?`))] = 'Unknown'  
sum(is.na(shooting\_data$`Did the deceased have priors?`))

## [1] 0

Remove unwanted attributes

testdata <- shooting\_data[c("Cause of Death", "State", "Gender", "Age", "Race", "Was the deceased armed?", "Did the deceased have priors?")]  
str(testdata)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 2355 obs. of 7 variables:  
## $ Cause of Death : chr "Shooting" "Shooting" "Shooting" "Shooting" ...  
## $ State : chr "Texas" "Pennsylvania" "New Jersey" "California" ...  
## $ Gender : chr "Male" "Male" "Male" "Male" ...  
## $ Age : num 34 19 30 29 37 47 27 22 21 52 ...  
## $ Race : chr "Black" "Unknown" "Black" "White" ...  
## $ Was the deceased armed? : chr "Yes" "Yes" "Unclear" "Unclear" ...  
## $ Did the deceased have priors?: chr "Yes" "Yes" "Unclear" "Unclear" ...

Analysis of each attributes and normaliztion

#Remove all rows that are not shooting death  
summary(testdata)

## Cause of Death State Gender Age   
## Length:2355 Length:2355 Length:2355 Min. : 5.00   
## Class :character Class :character Class :character 1st Qu.: 26.00   
## Mode :character Mode :character Mode :character Median : 35.00   
## Mean : 36.92   
## 3rd Qu.: 45.00   
## Max. :107.00   
## Race Was the deceased armed? Did the deceased have priors?  
## Length:2355 Length:2355 Length:2355   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##

testdata<-testdata[!(testdata$`Cause of Death`!="Shooting"),]  
summary(testdata)

## Cause of Death State Gender Age   
## Length:2076 Length:2076 Length:2076 Min. : 5.00   
## Class :character Class :character Class :character 1st Qu.: 26.00   
## Mode :character Mode :character Mode :character Median : 34.00   
## Mean : 36.36   
## 3rd Qu.: 45.00   
## Max. :107.00   
## Race Was the deceased armed? Did the deceased have priors?  
## Length:2076 Length:2076 Length:2076   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##

#Cause of Death  
str(testdata$`Cause of Death`)

## chr [1:2076] "Shooting" "Shooting" "Shooting" "Shooting" "Shooting" ...

testdata$`Cause of Death`<- gsub("car", "Car", testdata$`Cause of Death`)  
testdata$`Cause of Death`<-as.factor(testdata$`Cause of Death`)  
summary(testdata$`Cause of Death`)

## Shooting   
## 2076

#State  
str(testdata$State)

## chr [1:2076] "Texas" "Pennsylvania" "New Jersey" "California" ...

table(testdata$State)

##   
## Alabama Alaska Arizona   
## 37 5 94   
## Arkansas California Colorado   
## 17 359 45   
## Connecticut Delaware District of Columbia   
## 7 7 8   
## Florida Georgia Hawaii   
## 158 57 5   
## Idaho Illinois Indiana   
## 9 57 34   
## Iowa Kansas Kentucky   
## 13 25 29   
## Louisiana Maine Maryland   
## 38 12 38   
## Massachusetts Michigan Minnesota   
## 21 32 19   
## Mississippi Missouri Montana   
## 25 45 13   
## Nebraska Nevada New Hampshire   
## 13 33 4   
## New Jersey New Mexico New York   
## 32 32 44   
## North Carolina North Dakota Ohio   
## 58 1 56   
## Oklahoma Oregon Pennsylvania   
## 59 31 47   
## Rhode Island South Carolina South Dakota   
## 2 39 5   
## Tennessee Texas Utah   
## 38 212 26   
## Vermont Virgin Islands Virginia   
## 2 2 33   
## Washington West Virginia Wisconsin   
## 56 12 25   
## Wyoming   
## 5

#Gender  
str(testdata$Gender)

## chr [1:2076] "Male" "Male" "Male" "Male" "Male" "Male" "Male" "Male" ...

testdata$Gender<- gsub("male", "Male", testdata$Gender)  
testdata$Gender<-as.factor(testdata$Gender)  
summary(testdata$Gender)

## FeMale Male Unknown   
## 114 1960 2

#Age  
str(testdata$Age)

## num [1:2076] 34 19 30 29 37 47 27 22 21 52 ...

testdata$Age<-as.integer(testdata$Age)  
str(testdata$Age)

## int [1:2076] 34 19 30 29 37 47 27 22 21 52 ...

summary(testdata$Age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 5.00 26.00 34.00 36.36 45.00 107.00

#Race  
str(testdata$Race)

## chr [1:2076] "Black" "Unknown" "Black" "White" "Latino" "Unknown" ...

testdata$Race<-as.factor(testdata$Race)  
summary(testdata$Race)

## Asian Black Latino Native American   
## 24 430 289 15   
## Other Pacific Islander Unknown White   
## 7 2 635 674

str(testdata$Race)

## Factor w/ 8 levels "Asian","Black",..: 2 7 2 8 3 7 3 3 2 7 ...

#Was the deceased armed?  
str(testdata$`Was the deceased armed?`)

## chr [1:2076] "Yes" "Yes" "Unclear" "Unclear" "Yes" "Yes" "Yes" "Yes" ...

testdata$`Was the deceased armed?`<-as.factor(testdata$`Was the deceased armed?`)  
summary(testdata$`Was the deceased armed?`)

## Disputed No Unclear Unknown Yes   
## 45 133 335 88 1475

str(testdata$`Was the deceased armed?`)

## Factor w/ 5 levels "Disputed","No",..: 5 5 3 3 5 5 5 5 5 5 ...

#Did the deceased have priors?  
str(testdata$`Did the deceased have priors?`)

## chr [1:2076] "Yes" "Yes" "Unclear" "Unclear" "Unclear" "Unclear" ...

testdata$`Did the deceased have priors?`<- as.factor(testdata$`Did the deceased have priors?`)  
summary(testdata$`Did the deceased have priors?`)

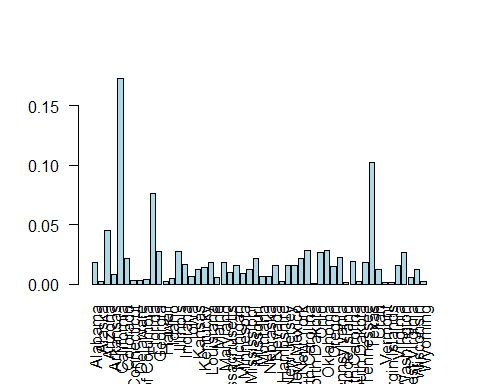
## No Unclear Unknown Yes   
## 98 1417 175 386

str(testdata$`Did the deceased have priors?`)

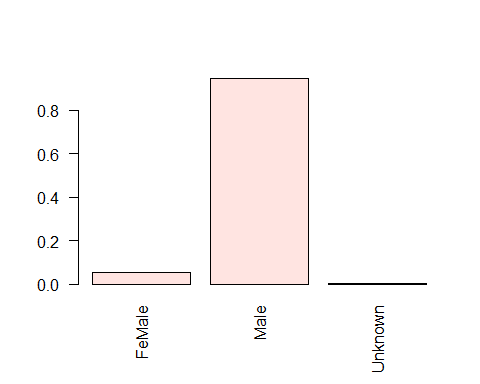
## Factor w/ 4 levels "No","Unclear",..: 4 4 2 2 2 2 4 2 2 2 ...

Visualization of key attributes

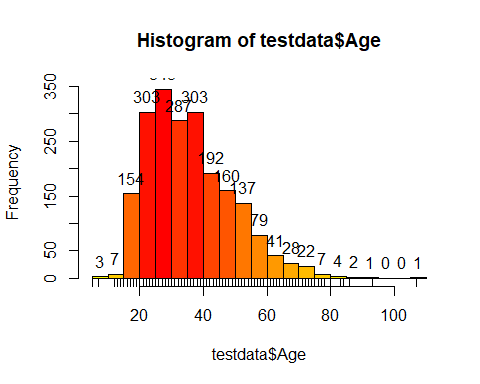
#State  
barplot(prop.table(table(testdata$State)), col = "lightblue", las=2)



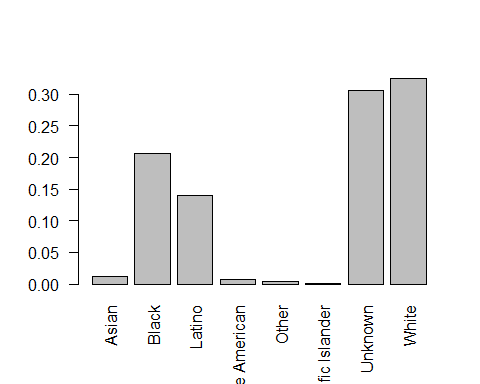
#Gender  
barplot(prop.table(table(testdata$Gender)), col = "mistyrose", las=2)



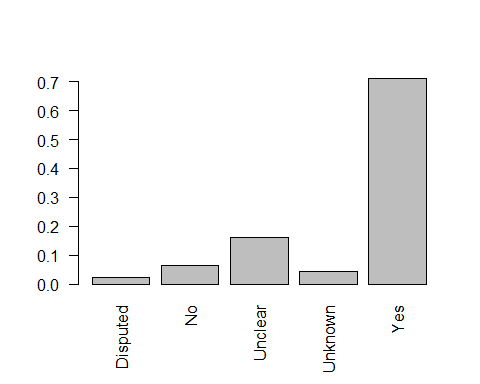
#Age  
h<-hist(testdata$Age, breaks = 20, plot = F)  
plot(h, col = heat.colors(length(h$mids))[length(h$count)-rank(h$count)+1], ylim = c(0, max(h$count)+5))  
rug(testdata$Age)  
text(h$mids, h$count, h$count, pos=3)



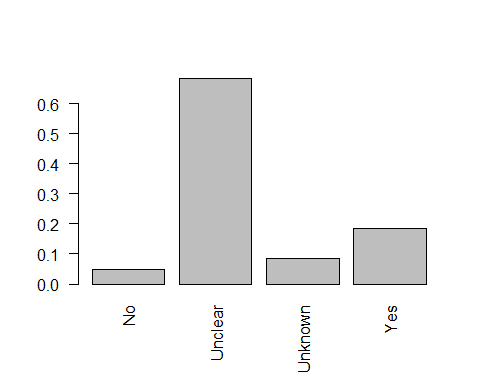
#Race  
barplot(prop.table(table(testdata$Race)), las=2)



#Was the deceased armed?  
barplot(prop.table(table(testdata$`Was the deceased armed?`)), las=2)



#Did the deceased have priors?  
barplot(prop.table(table(testdata$`Did the deceased have priors?`)), las=2)



Armed person is more likly to be shot dead irrespective of their race. Compared to other races unarmed black death percenatge was more.(More analysis needed)

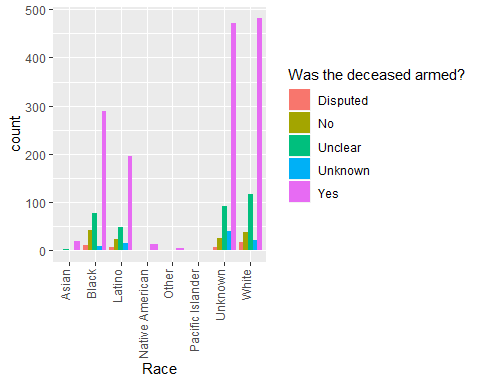
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.4.4

table(testdata$Race, testdata$`Was the deceased armed?`)

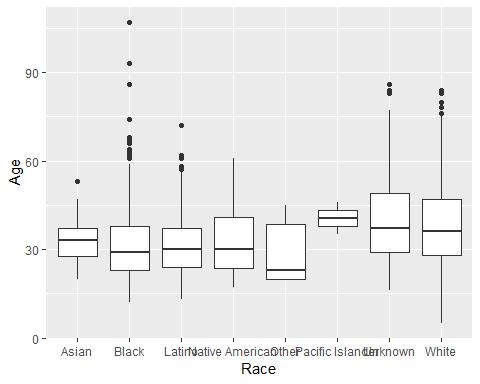
##   
## Disputed No Unclear Unknown Yes  
## Asian 0 1 2 1 20  
## Black 12 43 77 10 288  
## Latino 7 23 48 15 196  
## Native American 0 1 1 0 13  
## Other 0 1 0 1 5  
## Pacific Islander 1 0 0 0 1  
## Unknown 8 25 91 40 471  
## White 17 39 116 21 481

ggplot(testdata, aes(x = Race, fill = `Was the deceased armed?`)) + geom\_bar(position = "dodge")+theme(axis.text.x = element\_text(angle = 90, vjust = 0, hjust = 1))



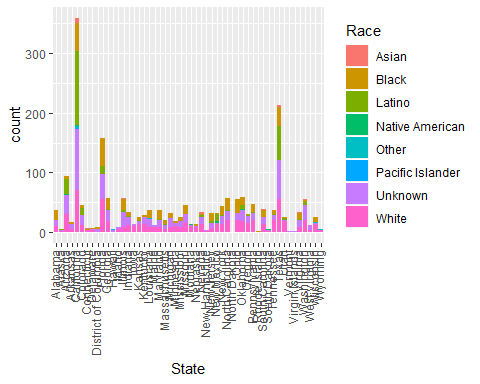
Compared to whites, other race victims are younger(ignoring "Unknown")

ggplot(testdata, aes(x = Race, y = Age)) + geom\_boxplot()



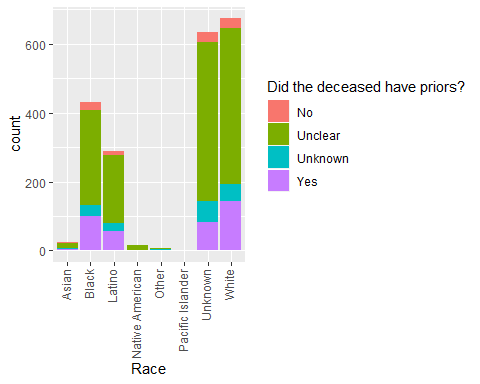
Police shooting death is high in Florida, Texas and California where as North Dakota had lowest death.

ggplot(testdata, aes(x = State, fill = Race)) + geom\_bar()+theme(axis.text.x = element\_text(angle = 90, vjust = 0.25, hjust = 1))



"Did the deceased have priors" didn't had much difference acrose races.

ggplot(testdata, aes(x = Race, fill = `Did the deceased have priors?`)) + geom\_bar()+theme(axis.text.x = element\_text(angle = 90, vjust = 0.25, hjust = 1))



Even though the mean age for police shooting death accross the state was 36 some states like Hawai, Vermount and Wyoming had higher age spectrum.

summary(testdata$Age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 5.00 26.00 34.00 36.36 45.00 107.00

ggplot(testdata, aes(x = State, y = Age)) + geom\_boxplot()+theme(axis.text.x = element\_text(angle = 90, vjust = 0.25, hjust = 1))

