**Project Idea:**

US data, for 1997-2002, from police-reported car crashes in which there is a harmful event (people or property), and from which at least one vehicle was towed. Data are restricted to front-seat occupants, include only a subset of the variables recorded, and are restricted in other ways also.

**Dataset Description:**

A data frame with 26217 observations on the following 15 variables.

dvcat

ordered factor with levels (estimated impact speeds) 1-9km/h, 10-24, 25-39, 40-54, 55+

weight

Observation weights, albeit of uncertain accuracy, designed to account for varying sampling probabilities.

dead

factor with levels alive dead

airbag

a factor with levels none airbag

seatbelt

a factor with levels none belted

frontal

a numeric vector; 0 = non-frontal, 1=frontal impact

sex

a factor with levels f m

ageOFocc

age of occupant in years

yearacc

year of accident

yearVeh

Year of model of vehicle; a numeric vector

abcat

Did one or more (driver or passenger) airbag(s) deploy? This factor has levels deploy nodeploy unavail

occRole

a factor with levels driver pass

deploy

a numeric vector: 0 if an airbag was unavailable or did not deploy; 1 if one or more bags deployed.

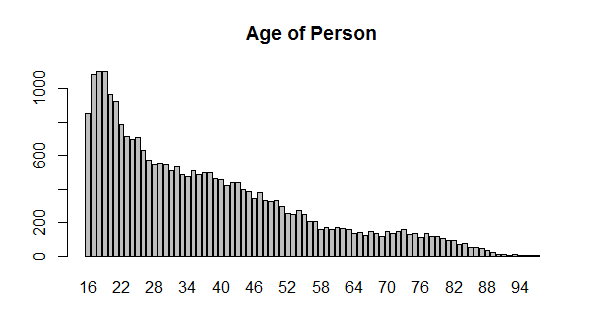
injSeverity

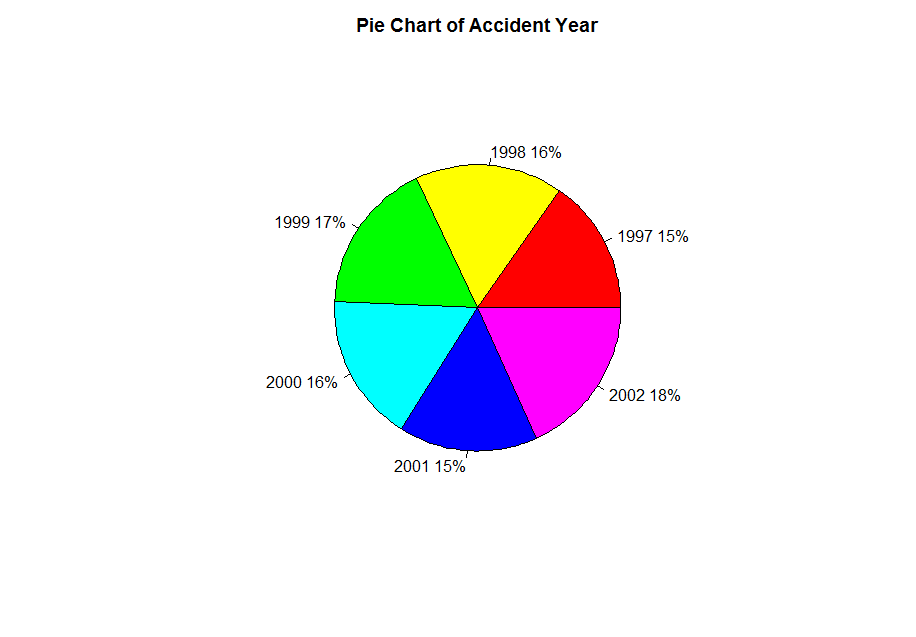
a numeric vector; 0:none, 1:possible injury, 2:no incapacity, 3:incapacity, 4:killed; 5:unknown, 6:prior death

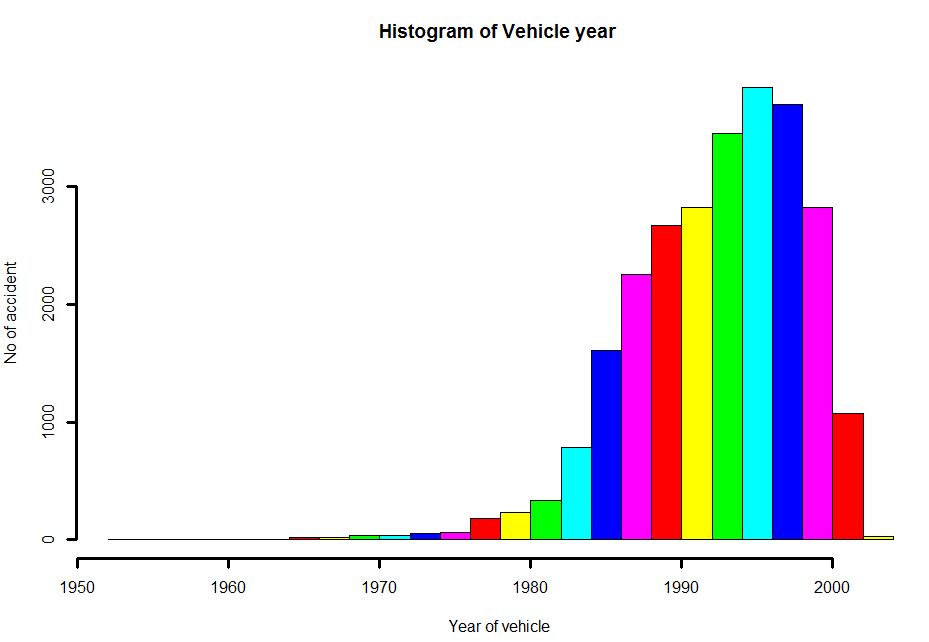
caseid

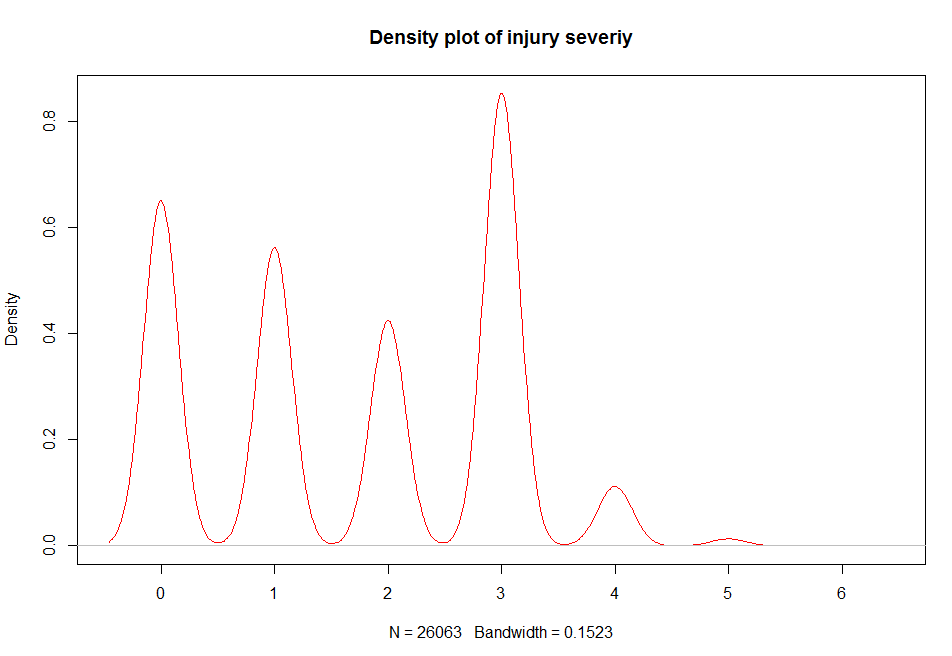
character, created by pasting together the populations sampling unit, the case number, and the vehicle number. Within each year, use this to uniquely identify the vehicle.

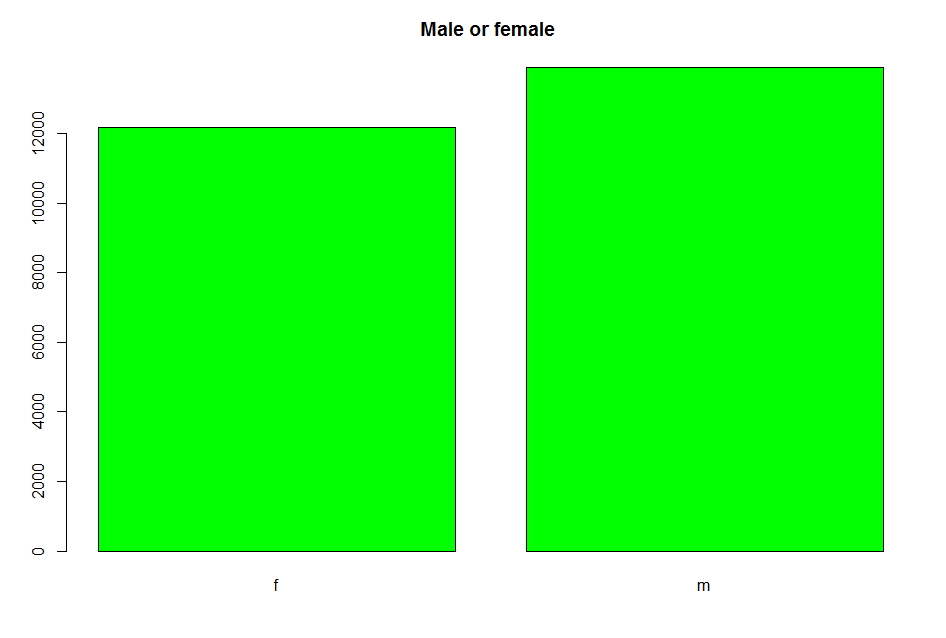
**Dataset Visualization & Analytics:**

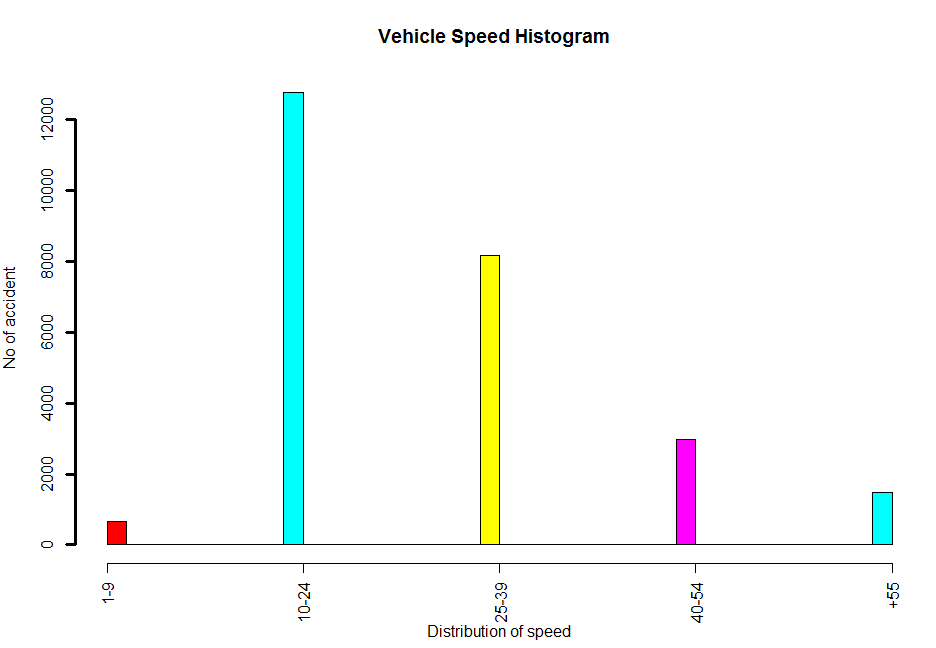
****

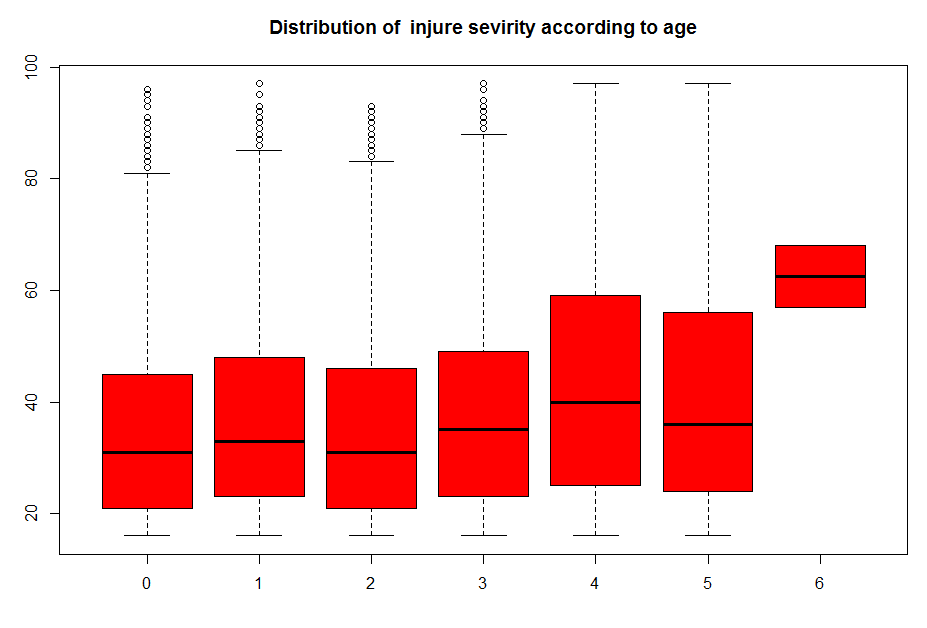
****

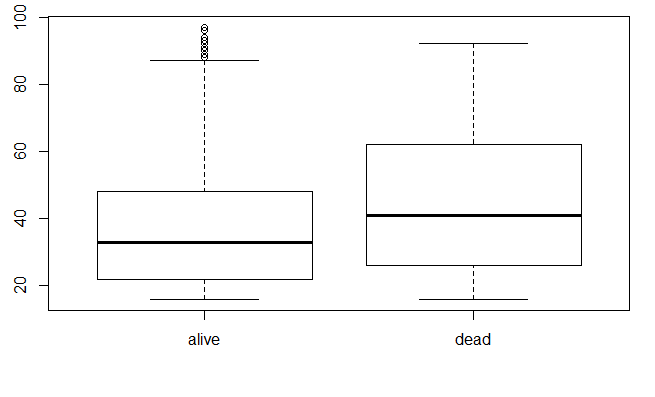
****

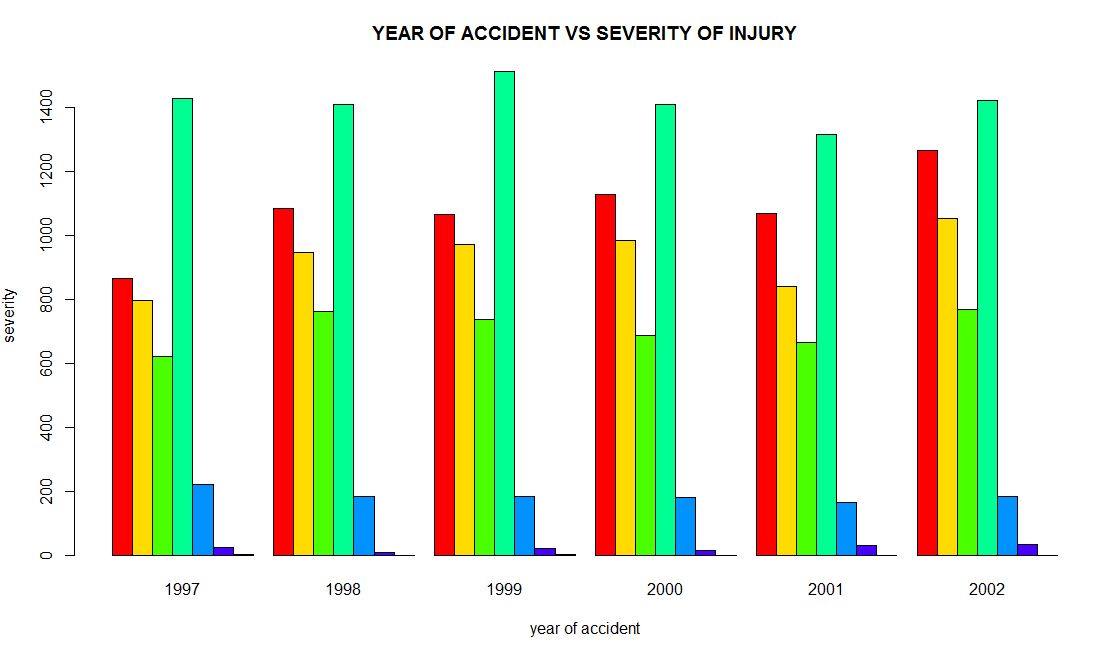
****

****

****

****

****

****

**Used tools:**

R Language.

**Project Code:**

#read dataset

accid <- read.csv(file="C:\\Users\\Mortada\\Documents\\R\\win-library\\3.5\\nassCDS.csv", header=TRUE, stringsAsFactors = FALSE)

#show name of coloums

names(accid)

#show details of each coloum

str(accid)

#detec missing values and remove it from dataset

NoOfNull<-apply(accid,2,function(x)sum(is.na(x)))

NoOfNull

accid<-na.omit(accid)

NoOfNull<-apply(accid,2,function(x)sum(is.na(x)))

NoOfNull

#delete wrong values of weight

accid$weight[accid$weight>100]<-0

accid$weight[accid$weight<10]<-0

accid$weight[accid$weight==0]<-NA

accid<-na.omit(accid)

accid

#barplot to show age of person in accident

Person\_Age<-table(accid$ageOFocc)

barplot(Person\_Age,main="Age of Person")

#piechart to show distribution of year of accident

#barplot of year accident

year<-table(accid$yearacc)

barplot(year,main="Year of accident")

slices <- c(4000,4400,4500,4400, 4100, 4800)

lbls <- c("1997","1998","1999","2000","2001","2002")

pct <- slices/26000\*100

pcct<-as.integer(pct)

lbls <- paste(lbls,pcct)

lbls <- paste(lbls,"%",sep="")

pie(slices, labels = lbls, col= rainbow (length(lbls)), main="Pie Chart of Accident Year")

#histogram to show the distribution model of cars

hist(accid$yearVeh, col = rainbow(6),main = "Vehicle year Histogram ", xlab = "Year of vehicle",ylab = "No of accident",breaks = 30,lwd=3)

#densityplot show the injury severity

plot(density(accid$injSeverity),main="Density plot of injury severiy",col="red")

#barplot to determine no of alive and dead people

d<-accid$dead

barplot(table(d),main = "Dead or Alive",col = "blue")

#barplot to determine no of male and female people

s<-accid$sex

barplot(table(s),main = "Male or female ",col = "Green")

#histogram to show the distribution speed of cars

accidtemp$dvcat<-as.numeric(as.factor(accidtemp$dvcat))

hist(accidtemp$dvcat,col = rainbow(6),main = "Vehicle Speed Histogram ", xaxt="n", ,xlab = "Distribution of speed",ylab = "No of accident",breaks = 30,lwd=3)

axis(1,labels=c("1-9","10-24","25-39","40-54","+55"),at=1:5)

#show details of age

summary(accid$ageOFocc)

#injury severity based on age

ggplot(aes(x = injSeverity), data = accid) +

geom\_bar(color = "black", fill = "#993366") +

ggtitle("Distribution of injure sevirity ")

boxplot(accid$ageOFocc ~ accid$injSeverity,main="Distribution of injure sevirity according to age",col="red")

#determine distribution of dead and alive based on their age

p1<-plot(accid$dead,accid$ageOFocc)

by(accid$ageOFocc, accid$dead, mean)

#injury severity based on year of accident

counts<-table(accid$injSeverity,accid$yearacc)

barplot(counts,main = "YEAR OF ACCIDENT VS SEVERITY OF INJURY",

xlab ="year of accident",ylab = "severity",col = rainbow(7),

legend=rownames(counts),beside = T)

# Classification with decision tree

accid$dead<-as.integer(accid$dead)

accid$airbag<-as.numeric(accid$airbag)

accid$seatbelt<-as.numeric(accid$seatbelt)

accid$dvcat<-as.numeric(accid$dvcat)

accid$sex<-as.numeric(accid$sex)

accid$occRole<-as.numeric(accid$occRole)

accid$injSeverity<-as.numeric(accid$injSeverity)

accid$ageOFocc<-as.numeric(accid$ageOFocc)

accid$yearacc<-as.numeric(accid$yearacc)

accid$yearVeh<-as.numeric(accid$yearVeh)

ind<-sample(2,nrow(accid),prob=c(0.7,0.3),replace=TRUE)

train.data<-accid[ind==1,]

test.data<-accid[ind==2,]

accid.tree <- ctree(dead~ageOFocc+weight+dvcat

+airbag+seatbelt+yearacc+yearVeh+injSeverity ,data = train.data)

table(predict(accid.tree),train.data$dead)

testPred <-predict(accid.tree,newdata = test.data)

table(testPred,test.data$dead)

plot(accid.tree,type="simple")

accuracy(predict(accid.tree,test.data),test.data[,"dead"])

# Classification with naive bayesian

accid$dead<-as.factor(accid$dead)

accid$airbag<-as.factor(accid$airbag)

accid$seatbelt<-as.factor(accid$seatbelt)

accid$dvcat<-as.factor(accid$dvcat)

accid$sex<-as.factor(accid$sex)

accid$occRole<-as.factor(accid$occRole)

accid$injSeverity<-as.factor(accid$injSeverity)

accid$caseid<-as.integer(accid$caseid)

nb\_mod<-naiveBayes(accid$dead~.,data = accid)

nb\_mod

prediction=predict(nb\_mod,accid)

table(prediction,accid$dead)

prop.table(table(prediction,accid$dead),1)

**Libraries:**

1-Metric

2- ggplot2

3- gridExtra

4- PerformanceAnalytics

5- klaR

6- e1071

7- party