HW4_Do_Quyen

Quyen Do

September 12, 2018

Problem 3

According to Roger Peng, what is the focus of the EDA stage of an analysis?

In his book, Roger Peng uses an analogy to film editing step of making a movie to the EDA stage of data analysis. It is a critical stage after the data is collected and serves many purposes. The focus of the stage is for the researchers to be aware of any problem with the data, determine if more data need collected, and to examine the relationships between variables in order to make importatnt decisions for later stages of the research.

Problem 4

```r
prob4_data1 <- read.xlsx("HW4_data.xlsx",sheetIndex = 1)
prob4_data2 <- read.xlsx("HW4_data.xlsx",sheetIndex = 2)
prob4_combined <- rbind(prob4_data1,prob4_data2)
```

1,2. Summary statistics and factor exploration

```r
#Summary table
knitr::kable(summary(prob4_combined),caption="Summary of Problem 4 Data")
```

Table 1: Summary of Problem 4 Data

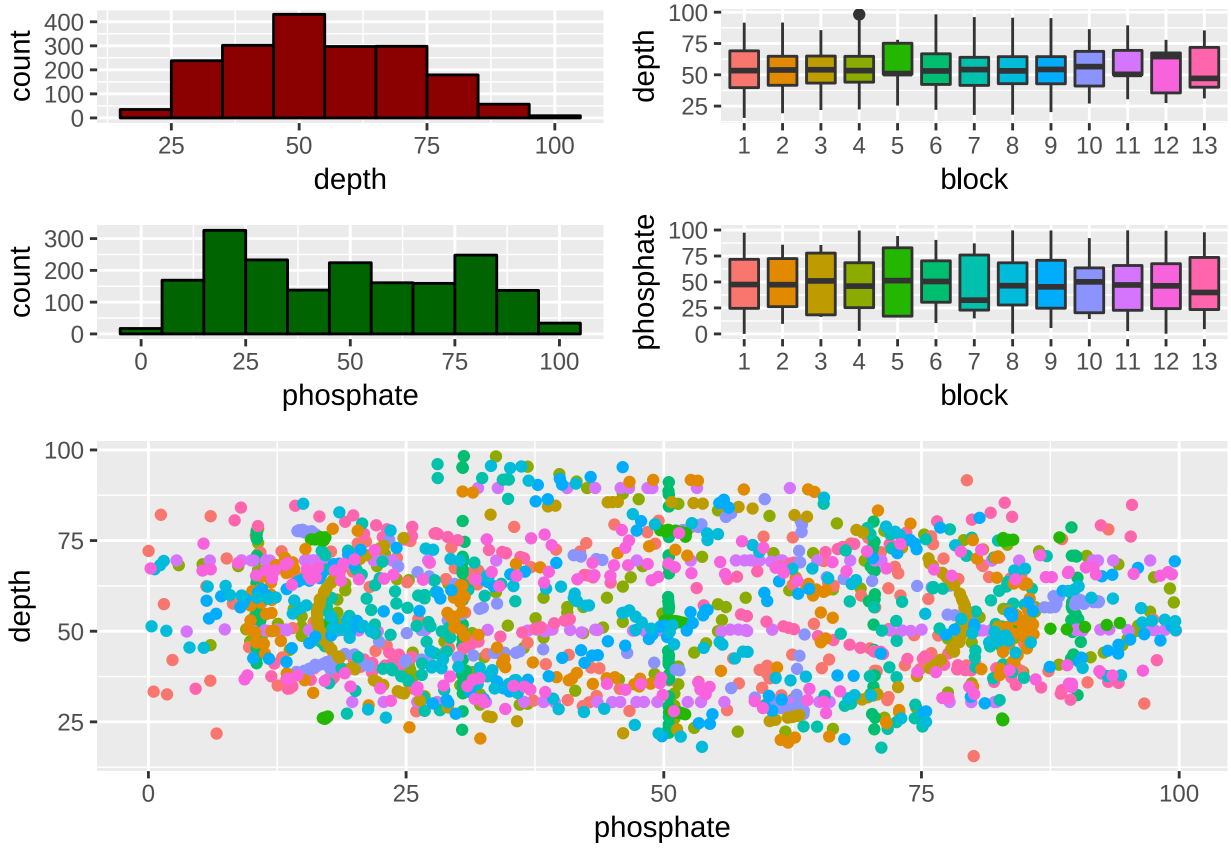
	block	depth	phosphate
	Min. : 1	Min. :15.56	Min. : 0.01512
	1st Qu.: 4	1st Qu.:41.07	1st Qu.:22.56107
	Median : 7	Median :52.59	Median :47.59445
	Mean : 7	Mean :54.27	Mean :47.83510
	3rd Qu.:10	3rd Qu.:67.28	3rd Qu.:71.81078
	Max. :13	Max. :98.29	Max. :99.69468
The d	ata has 3 var	iables named “bl	ock“,”depth" and “phosphate”. “Block” is a discrete variable from 1 to 13, whic

```r
#Factor
prob4_combined$block <- as.factor(prob4_combined$block)
```

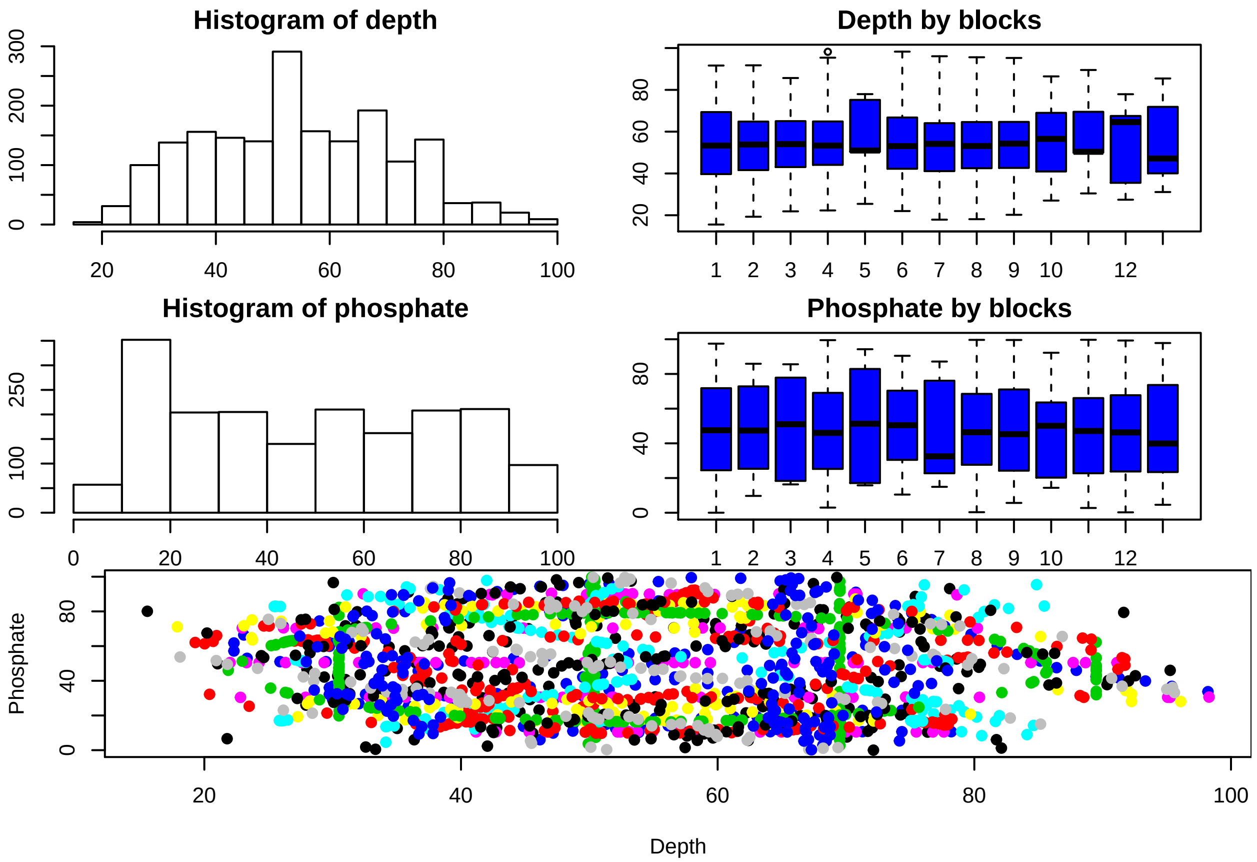
3. Multipanel plot

```r
#Multipane plot using ggplot and ggpubr
p1 <- ggplot(prob4_combined,aes(x=depth)) + geom_histogram(colour= "black",binwidth = 10,fill="darkred"

p2 <- ggplot(prob4_combined,aes(x=block,y=depth ,group=block,fill=block))
p2 <- p2 + geom_boxplot() + guides(fill=FALSE) + labs(x="block")

p3 <- ggplot(prob4_combined,aes(x=phosphate)) + geom_histogram(colour= "black",binwidth = 10,fill="darkg

p4 <- ggplot(prob4_combined,aes(x=block,y=phosphate ,group=block,fill=block))
```

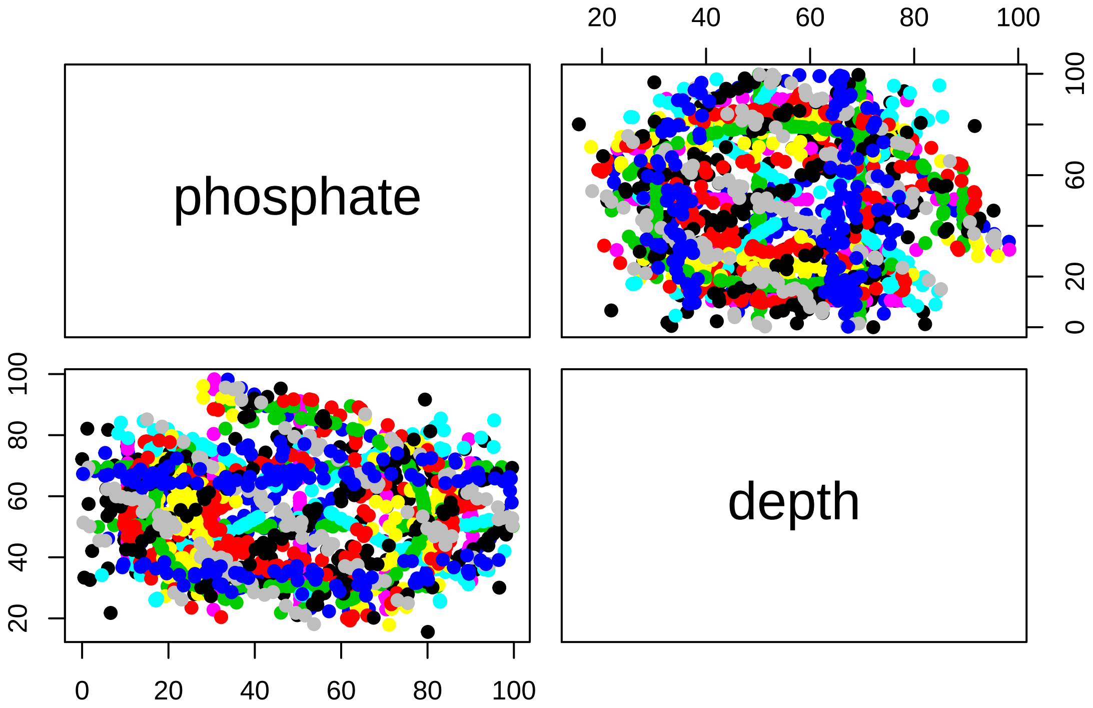
```r
p4 <- p4 + geom_boxplot() + guides(fill=FALSE) + labs(x="block")

p5 <- ggplot(prob4_combined,aes(phosphate,depth,colour=block)) + geom_point() + labs(x="phosphate",y="de

ggarrange(ggarrange(p1,p2,p3,p4,ncol = 2,nrow=2), p5, nrow = 2)
```



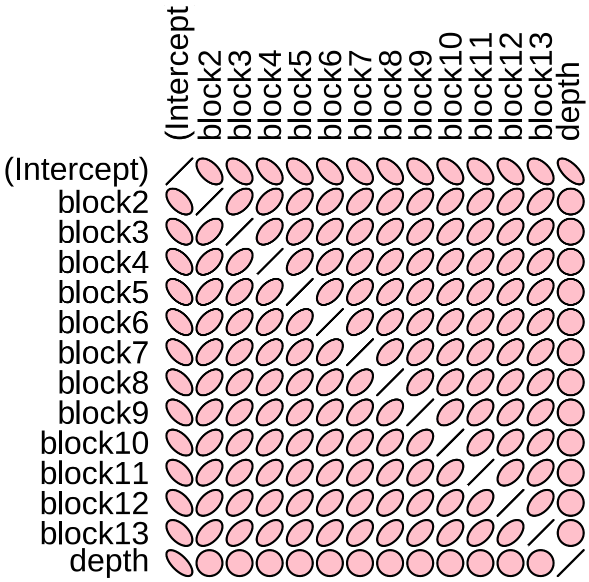
```r
layout(matrix(c(1,2,3,4,5,5), 3, 2, byrow = TRUE))
par(mar=rep(2,4), oma=rep(0, 4))
hist(prob4_combined$depth,main="Histogram of depth")
boxplot(prob4_combined$depth~prob4_combined$block,main="Depth by blocks",col=prob4_combined$block)
hist(prob4_combined$phosphate,main="Histogram of phosphate")
boxplot(prob4_combined$phosphate~prob4_combined$block,main="Phosphate by blocks",col=prob4_combined$bloc
par(mar=c(4,4,0,0))
plot(x=prob4_combined$depth,y=prob4_combined$phosphate,col=prob4_combined$block,
     pch=19,xlab="Depth",ylab="Phosphate")
```



```
par(mar=rep(0,4))
with(data=prob4_combined,pairs(phosphate~depth, col=block,pch=19))
```

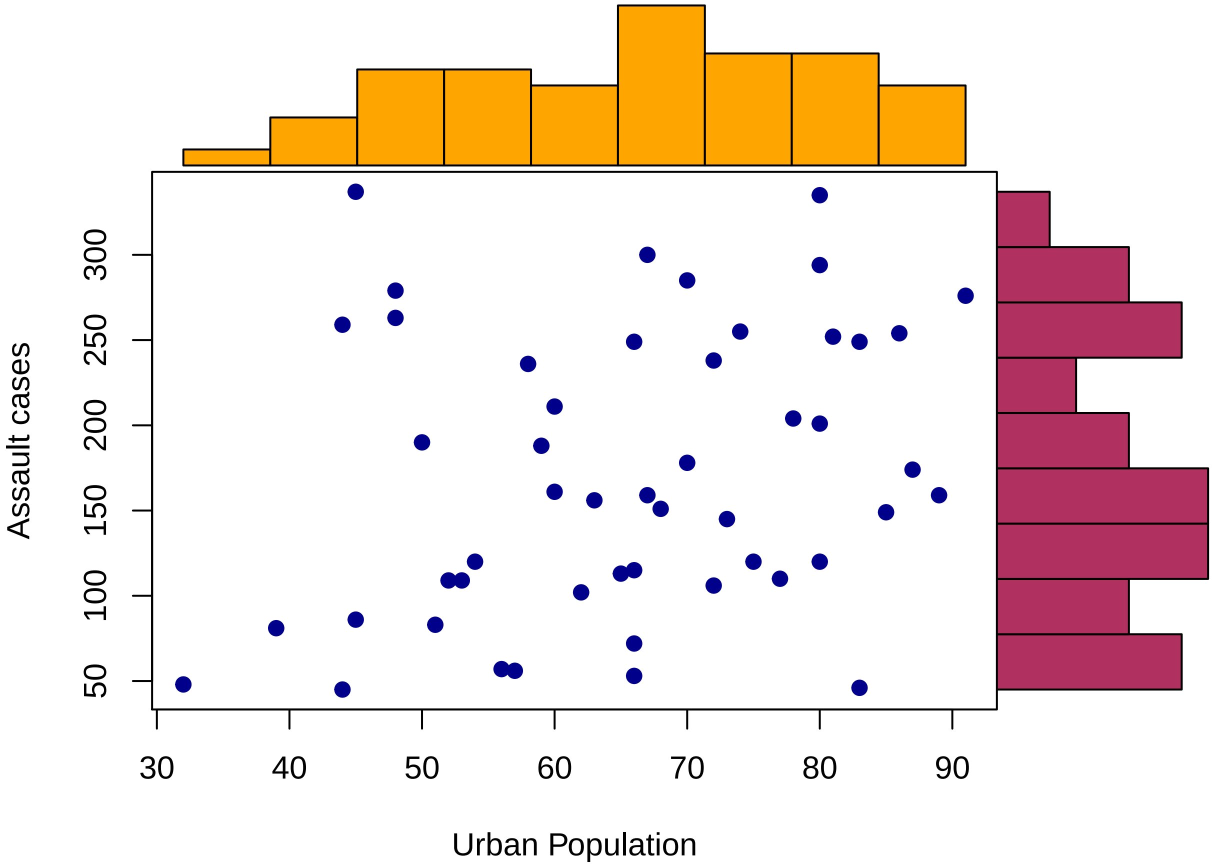


```
fit <- lm(phosphate~.,prob4_combined)
corr.fit <- summary(fit,correlation=T)$correlation
plotcorr(corr.fit,col="pink")
```



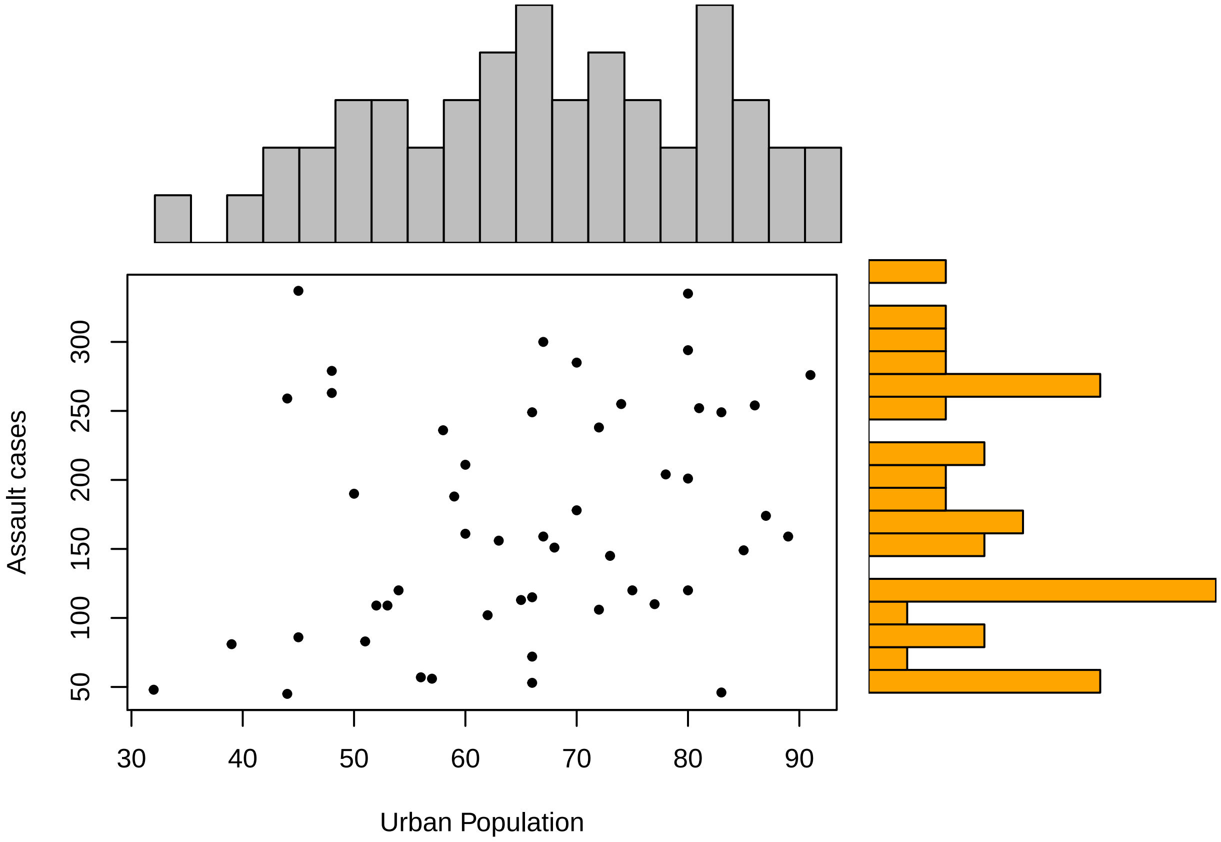
Problem 5

```r
# Add boxplots to a scatterplot

get_scatter_n_hist <- function(X,bin_num = 10, ycol = "red",xcol="blue",pcol = "green", ...) {

  ## check input
  stopifnot(ncol(X)==2)

  #Build the scatter plot
  par(fig=c(0,0.8,0,0.8),mar=c(4,4,0,0),oma=rep(0,4), new=FALSE)
  plot(X[,1], X[,2], col=pcol, ...)

  #Build marginal X histogram
  par(fig=c(0,0.8,0.8,1),mar=c(0,4,0,0),oma=rep(0,4), new=TRUE)
  hist(X[,1], axes=FALSE,col=xcol,main=NULL,ylab=NULL,breaks=seq(from=min(X[,1])-+sd(X[,1])/length(X[,1
                                              to=max(X[,1])+sd(X[,1])/length(X[,1]),
                                              length.out=bin_num))

  #Build marginal Y histogram
  par(fig=c(0.8,1,0,0.8),mar=c(4,0,0,1),oma=rep(0,4),new=TRUE)
  yhist <- hist(X[,2], plot=FALSE, breaks=seq(from=min(X[,2])-+sd(X[,2])/length(X[,2]),
                                              to=max(X[,2])+sd(X[,2])/length(X[,2]),
                                              length.out=bin_num))
  barplot(yhist$density, axes=FALSE, xlim=c(0, max(yhist$density)),
          space=0, horiz=TRUE, col = ycol)

}

crime.data <- USArrests

get_scatter_n_hist(data.frame(crime.data$UrbanPop,crime.data$Assault),xlab="Urban Population",
                ylab="Assault cases",pch=19,ycol="maroon",xcol="orange",pcol = "darkblue")
```



```
# par(fig=c(0,0.8,0,0.8),mar=c(4,4,0,0),oma=rep(0,4), new=FALSE)
# plot(mtcars$wt, mtcars$mpg, xlab="Car Weight", ylab="Miles Per Gallon",pch=19,col="darkgreen")
#
# par(fig=c(0,0.8,0.8,1),mar=c(0,4,0,0),oma=rep(0,4), new=TRUE)
# hist(mtcars$wt, axes=FALSE,col="maroon",main=NULL,ylab=NULL)
#
# par(fig=c(0.8,1,0,0.8),mar=c(4,0,0,1),oma=rep(0,4),new=TRUE)
# yhist <- hist(mtcars$mpg, plot=FALSE, breaks=seq(from=min(mtcars$mpg)-+sd(mtcars$mpg)/length(mtcars$m
#                                                 to=max(mtcars$mpg)+sd(mtcars$mpg)/length(mtcars$mpg),
#                                                 length.out=10))
# barplot(yhist$density, axes=FALSE, xlim=c(0, max(yhist$density)),
#           space=0, horiz=TRUE, col = "orange")
```

```
#Code reference: First answer of the question on Stackoverflow: https://stackoverflow.com/questions/110

scatterBarNorm <- function(x, dcol="blue", lhist=20, num.dnorm=5*lhist, ...){
  ## check input
  stopifnot(ncol(x)==2)
  ## set up layout and graphical parameters
  layMat <- matrix(c(2,0,1,3), ncol=2, byrow=TRUE)
  layout(layMat, widths=c(5/7, 2/7), heights=c(2/7, 5/7))
  ospc <- 0.5 # outer space
  pext <- 4 # par extension down and to the left
  bspc <- 1 # space between scatter plot and bar plots
  par. <- par(mar=c(pext, pext, bspc, bspc),
              oma=rep(ospc, 4)) # plot parameters
  ## scatter plot
  plot(x, xlim=range(x[,1]), ylim=range(x[,2]), pch=20,...)
  ## 3) determine barplot and height parameter
  ## histogram (for barplot-ting the density)
  xhist <- hist(x[,1], plot=FALSE, breaks=seq(from=min(x[,1])-+sd(x[,1])/length(x[,1]),
                                              to=max(x[,1])+sd(x[,1])/length(x[,1]),  length.out=lhist)
  yhist <- hist(x[,2], plot=FALSE, breaks=seq(from=min(x[,2])-+sd(x[,2])/length(x[,2]),
                                              to=max(x[,2])+sd(x[,2])/length(x[,2]),  length.out=lhist)
  ## determine the plot range and all the things needed for the barplots and lines
  xx <- seq(min(x[,1])-sd(x[,1])/length(x[,1]), max(x[,1])+sd(x[,1])/length(x[,1]), length.out=num.dnorm
  xy <- dnorm(xx, mean=mean(x[,1]), sd=sd(x[,1])) # density points
  yx <- seq(min(x[,2])-sd(x[,2])/length(x[,2]), max(x[,2])+sd(x[,2])/length(x[,2]), length.out=num.dnorm
  yy <- dnorm(yx, mean=mean(x[,2]), sd=sd(x[,2]))
  ## barplot and line for x (top)
  par(mar=c(0, 4, 0, 0)) #pext = 4 bottom, left, top, right
  barplot(xhist$density, axes=FALSE, ylim=c(0, max(xhist$density, xy)),
          space=0, col = "grey") # barplot
  #lines(seq(from=0, to=lhist-1, length.out=num.dnorm), xy, col=dcol) # line
  ## barplot and line for y (right)
  par(mar=c(4, 0, 0, 0))
  barplot(yhist$density, axes=FALSE, xlim=c(0, max(yhist$density, yy)),
          space=0, horiz=TRUE, col = "orange") # barplot
  #lines(yy, seq(from=0, to=lhist-1, length.out=num.dnorm), col=dcol) # line
  ## restore parameters
  par(par.)
}

#Using R dataset
crime.data <- USArrests

scatterBarNorm(data.frame(crime.data$UrbanPop,crime.data$Assault),lhist=20,xlab="Urban Population",
               ylab="Assault cases")
```



```
#Using ggplot
x <- data.frame(crime.data$UrbanPop,crime.data$Assault)
p <- ggplot(x, aes(x[,1], x[,2])) + geom_point() + theme_classic()
    ggMarginal(p, x, type = "histogram", yparams=list(colour="orange"))
p
```