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

```
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

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

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

Table 1: Summary of Problem 4 Data

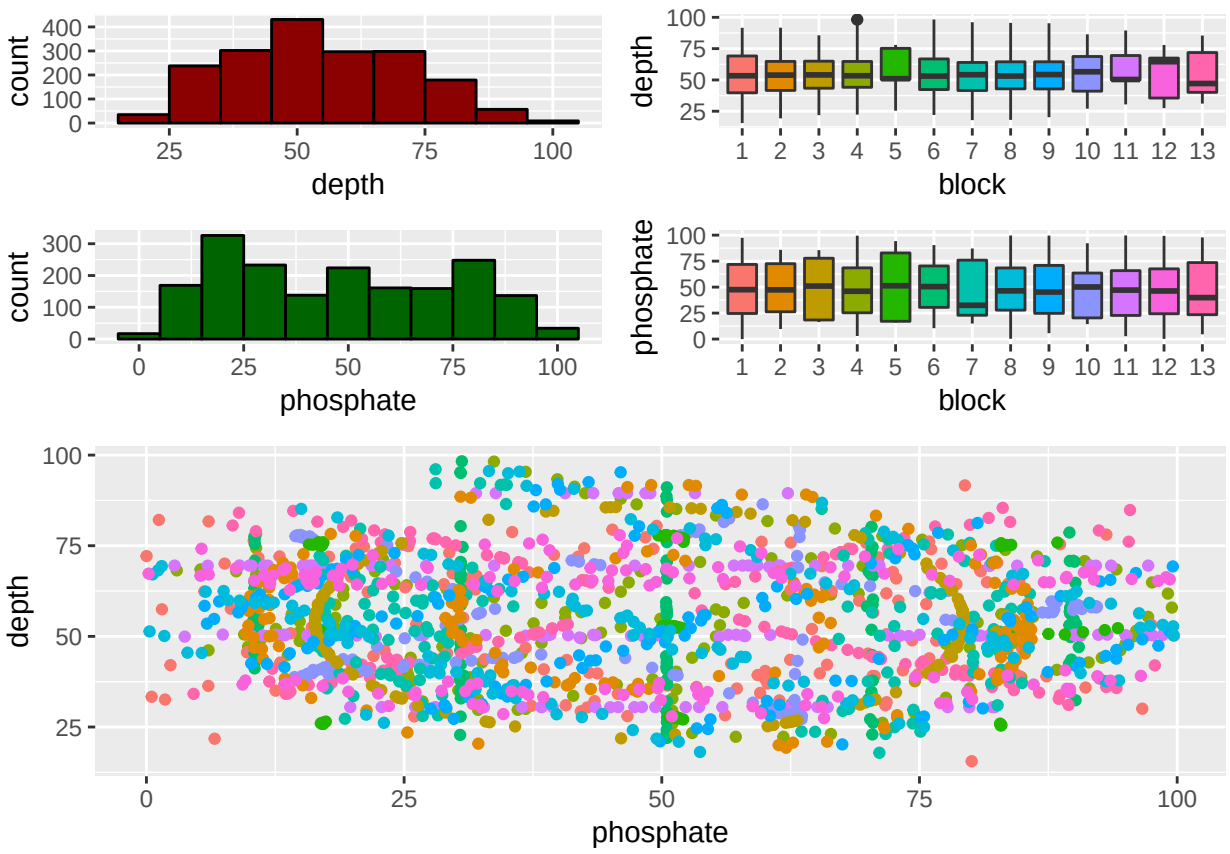
block	depth	phosphate
1 :142	Min. :15.56	Min. : 0.01512
2 :142	1st Qu.:41.07	1st Qu.:22.56107
3 :142	Median :52.59	Median :47.59445
4 :142	Mean :54.27	Mean :47.83510
5 :142	3rd Qu.:67.28	3rd Qu.:71.81078
6 :142	Max. :98.29	Max. :99.69468
(Other):994	NA	NA

The data has 3 variables named “block”, “depth” and “phosphate”. “Block” is a discrete variable from 1 to 13, which is likely a factor variable. Both “depth” and “block” are continuous variables.

3. Multipanel plot

```
#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="dark

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

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

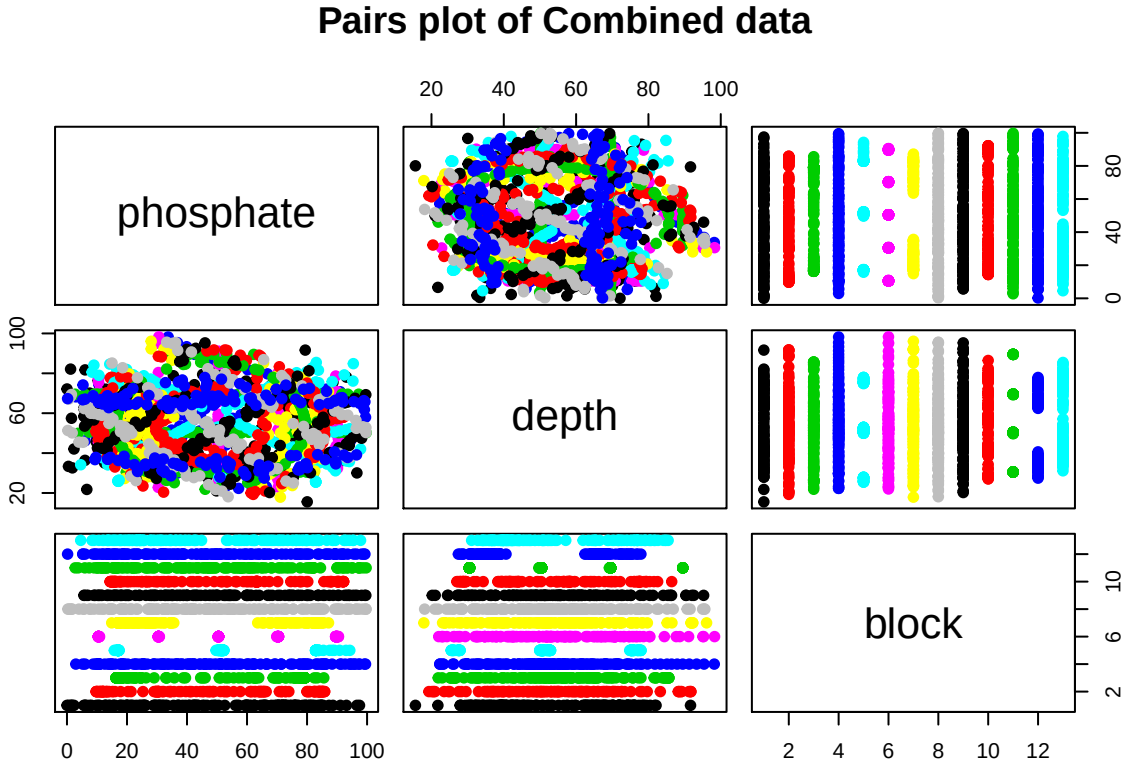


From the histograms, variable "depth" has a slightly symmetric and unimodal distribution. One could also say that there is a slight right skew with variable "depth". Variable "phosphate", however, is multimodal. Plotting the boxplot of these two variables by block shows that the distribution of "depth" among different blocks are not as similar as those from "phosphate". Specifically, "depth" from block 5, 11, 12 and 13 are very skewed and there is an outlier for "depth" variable of block 4. Among "phosphate" groups by block, the distributions are much more symmetric, with exception to block 7.

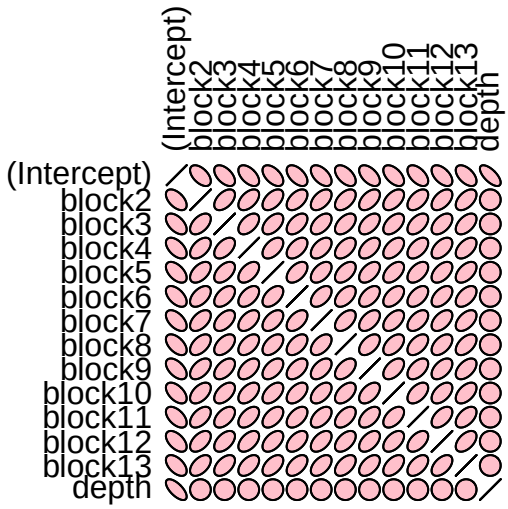
Scatter plot between "depth" and "phosphate" does not show any trend or correlation between the two variables. When adding color to each dot denoting the block that it comes from, not much information is gained since the colors seem equally scatter across.

4. Correlation plots

```
par(mar=rep(1,4), oma=rep(0, 4))
with(data=prob4_combined,pairs(phosphate~depth+block, col=block,pch=19,main="Pairs plot of Combined data
```




```
par(mar=rep(1,4), oma=rep(0, 4))
fit <- lm(phosphate~.,prob4_combined)
corr.fit <- summary(fit,correlation=T)$correlation
plotcorr(corr.fit,col="pink")
```



Using the pairs and correlation plot, we can notice some interesting patterns. The distribution of the phosphate value among block 5 and 6 seem concentrate on some specific values while the rest are scattered evenly across the range of "phosphate". Similary, block 5 and 12 also seem to be grouped by certain "depth" values. Both pairs and correlation plot confirm no linear relationship between "phosphate" and "depth".

5. Lesson when considering the summary statistics and the plots

In the last assignment, we built the boxplots of the mean by observer of both variables in the same graph. As shown in the graphs, this does not give us much information because of the scale of each variables are drastically different. This results in meaningless graphs. There, when it comes to different variables, their distributions should not be plotted on the same graph. Instead, to examine the relationship between variables, it is better to use scatter plot since it captures the two dimensions of each data point. We can also color each point on the plot to a value of a categorical variable. This will add another dimension to the scatterplot for consideration. The distribution of each variables can be compared when ploting them on different graphs.

Problem 5

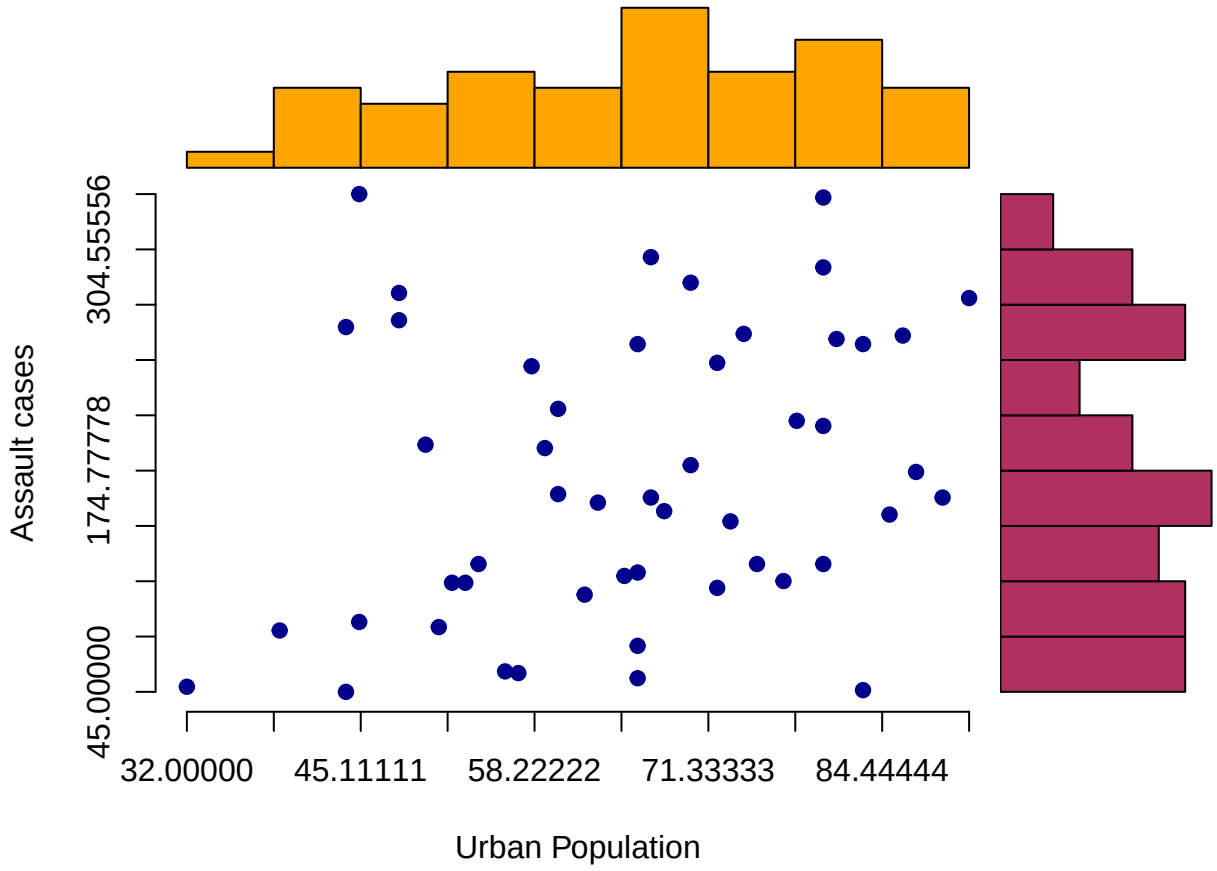
```
# Add boxplots to a scatterplot
draw_scatter_with_hist <- function(df,bin_num = 10,
                                   ycol = "red",xcol="blue",pcol = "green", ...) {
  # Create a scatter plot of two variable bordered with their histograms

  # Args:
  # df: a data frame with two variables
  # bin_num: the number of bin for the histogram. Default value is 10
  # ycol: color for the histogram of dependent variable
  # xcol: color for the histogram of independent variable
  # pcol: color for each points on the scatter plot
  # ... : arguments that can be paste into the scatter plot to govern its appearance
```

```
  # Returns:
  #  A plot containing a scatter plot with the independent variable being the first variable of df and
  #  Bordering the right side of the scatterplot is the histogram of X and sitting on top of the scatte
  #  The interval on the axes also reflect the bins of the histograms

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

  ## Get the sequence for plotting
  xseq <- seq(from=min(df[,1]), to=max(df[,1]), length.out=bin_num)
  yseq <- seq(from=min(df[,2]), to=max(df[,2]), length.out=bin_num)

  #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(df[,1], df[,2], col=pcol, axes = FALSE,...)
  axis(side=1, at=xseq)
  axis(side=2, at=yseq)

  #Build marginal df histogram
  par(fig=c(0,0.8,0.8,1),mar=c(0,4,0,0),oma=rep(0,4), new=TRUE)
  hist(df[,1], axes=FALSE, col=xcol, main=NULL, ylab=NULL, breaks=xseq,freq=FALSE)

  #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(df[,2], plot=FALSE, breaks=yseq)
  barplot(yhist$density, axes=FALSE, xlim=c(0, max(yhist$density)), space=0, horiz=TRUE, col = ycol)

}
```

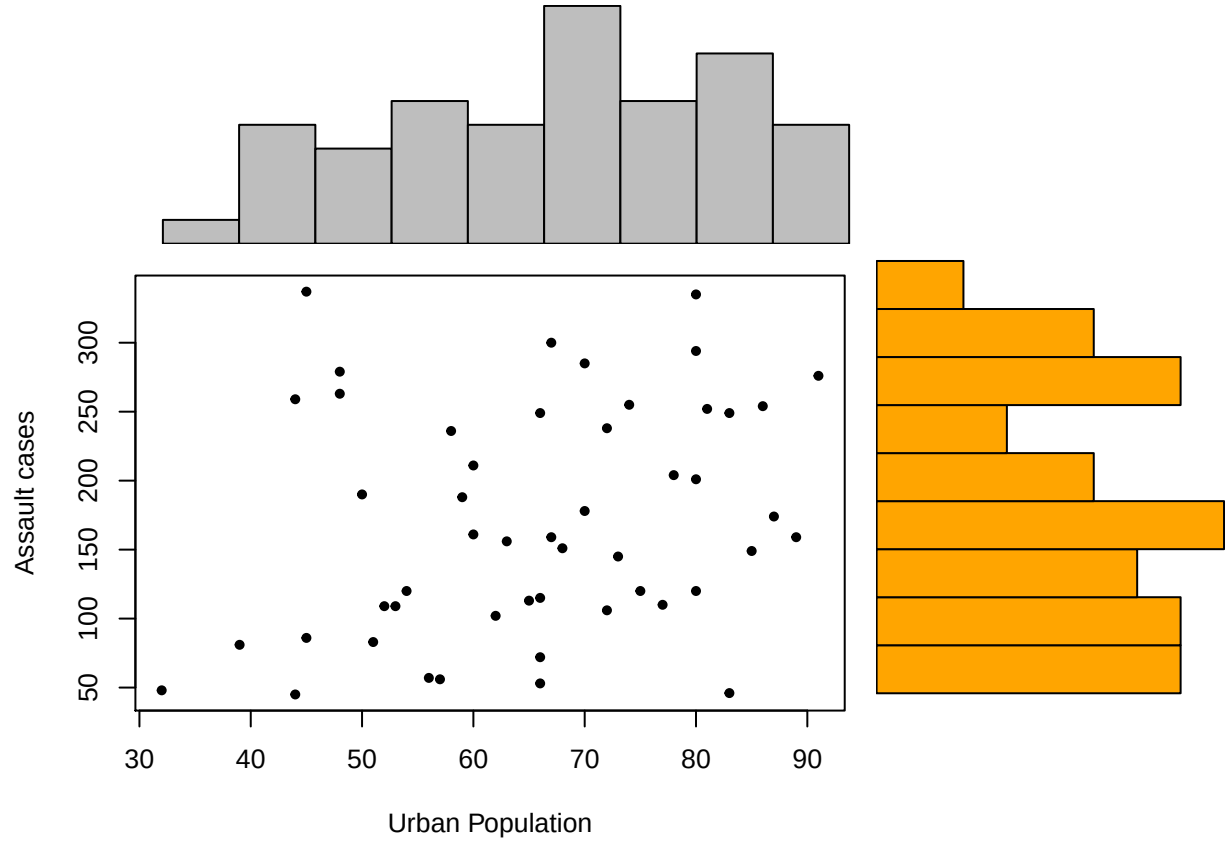
```
#Code reference: First answer of the question on Stackoverflow: https://stackoverflow.com/questions/110

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]),to=max(x[,1]),length.out=lhist))
  yhist <- hist(x[,2], plot=FALSE, breaks=seq(from=min(x[,2]), to=max(x[,2]),length.out=lhist)) # note:
  ## determine the plot range and all the things needed for the barplots and lines
  xx <- seq(min(x[,1]), max(x[,1]), length.out=num.dnorm) # evaluation points for the overlaid density
  xy <- dnorm(xx, mean=mean(x[,1]), sd=sd(x[,1])) # density points
  yx <- seq(min(x[,2]), max(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.)
}
```

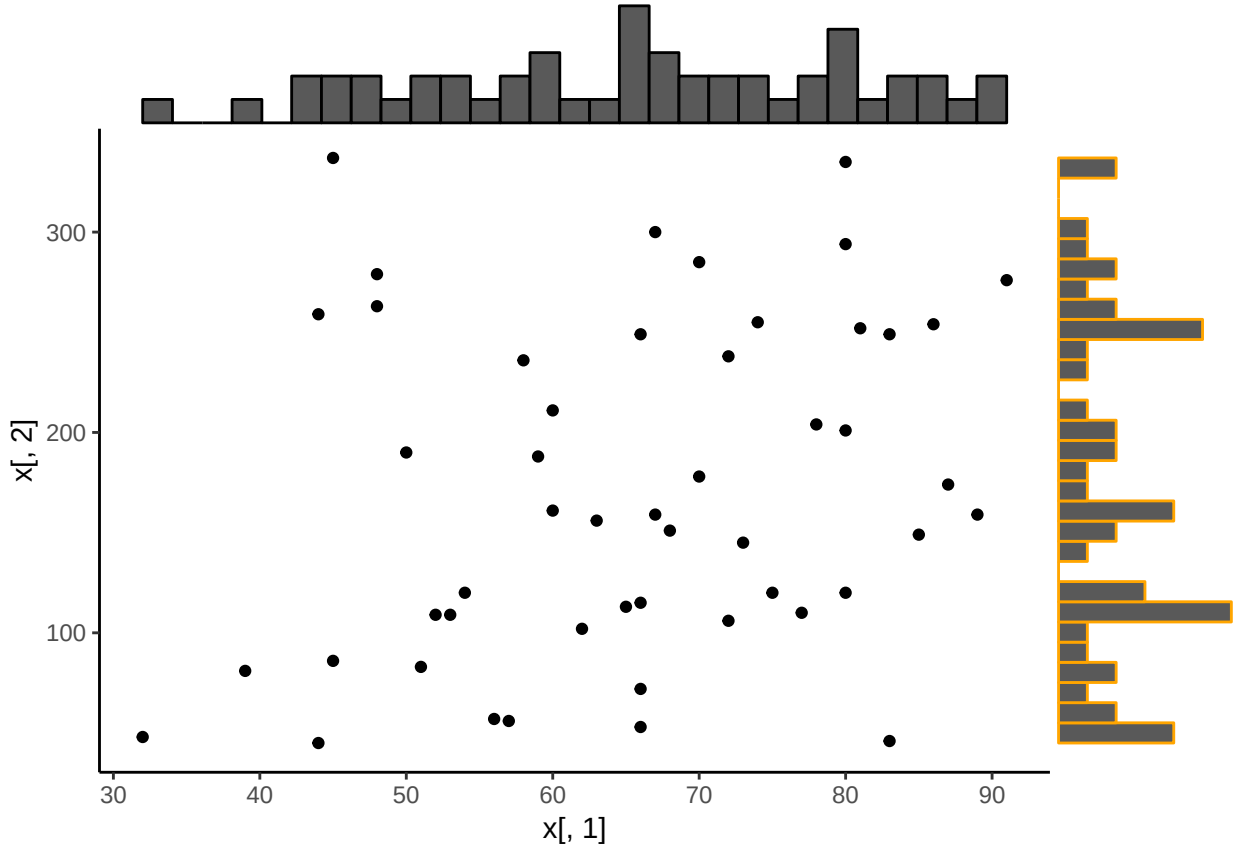
```
#Import US Arrests data set
crime.data <- USArrests

#Self-made plot
draw_scatter_with_hist(data.frame(crime.data$UrbanPop,crime.data$Assault),bin_num = 10,
                       xlab="Urban Population", ylab="Assault cases",
                       pch=19,ycol="maroon",xcol="orange",pcol = "darkblue")
```




```
#Plot using StackOverFlow answer
scatterBarNorm(data.frame(crime.data$UrbanPop,crime.data$Assault),lhist=10,
               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"))
```


300
200
100
x[, 2]
30
40
50
60
70
80
90
x[, 1]