HW4_Sauer_Annie

Annie Sauer

9/13/2018

Problem 3

According to Roger Peng, the goal of exploratory data analysis is to "show the data, summarize the evidence, and identify interesting patterns while eliminating ideas that likely won't pan out" (Peng, 2016).

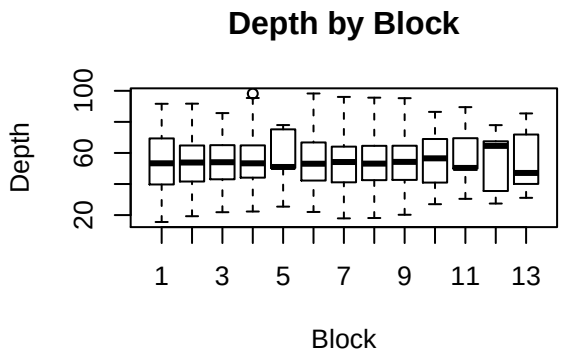
Peng, R. D. (2016). Exploratory Data Analysis with R. LeanPub.

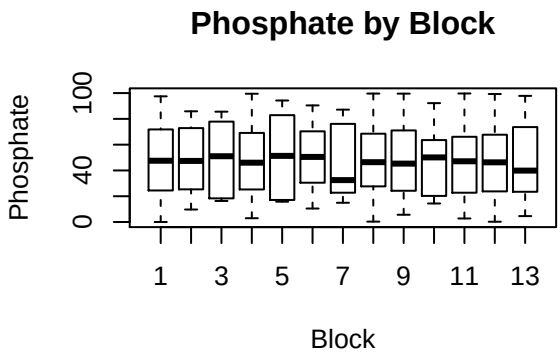
Problem 4

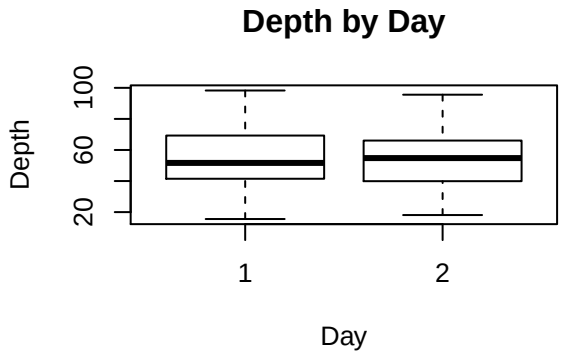
```
# Import data set
data_1 <- read_excel("04_projecting_knowledge_plots/HW4_data.xlsx", sheet = 1)
data_2 <- read_excel("04_projecting_knowledge_plots/HW4_data.xlsx", sheet = 2)

# Add day variable and combine sheets
data_1$day <- 1
data_2$day <- 2
data <- rbind(data_1, data_2)

# Summary statistics
summary(data)
```

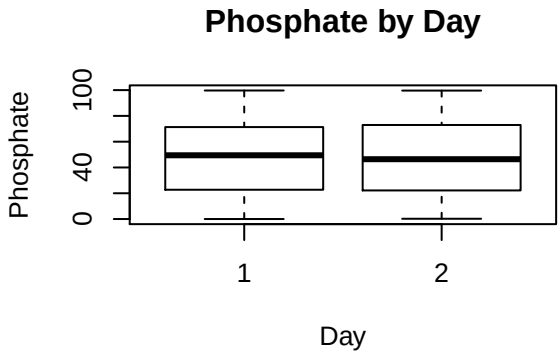
```
##      block        depth         phosphate             day
##  Min.   : 1   Min.   :15.56   Min.   : 0.01512   Min.   :1.000
##  1st Qu.: 4   1st Qu.:41.07   1st Qu.:22.56107   1st Qu.:1.000
##  Median : 7   Median :52.59   Median :47.59445   Median :1.000
##  Mean   : 7   Mean   :54.27   Mean   :47.83510   Mean   :1.385
##  3rd Qu.:10   3rd Qu.:67.28   3rd Qu.:71.81078   3rd Qu.:2.000
##  Max.   :13   Max.   :98.29   Max.   :99.69468   Max.   :2.000
```

```
# Factor Exploration with Multipanel plots
par(mfrow = c(2,2))
data$block <- as.factor(data$block)
boxplot(data$depth ~ data$block, main = 'Depth by Block', xlab = 'Block', ylab = 'Depth')
boxplot(data$phosphate ~ data$block, main = 'Phosphate by Block', xlab = 'Block',
        ylab = 'Phosphate')
data$day <- as.factor(data$day)
boxplot(data$depth ~ data$day, main = 'Depth by Day', xlab = 'Day', ylab = 'Depth')
boxplot(data$phosphate ~ data$day, main = 'Phosphate by Day', xlab = 'Day', ylab = 'Phosphate')
```

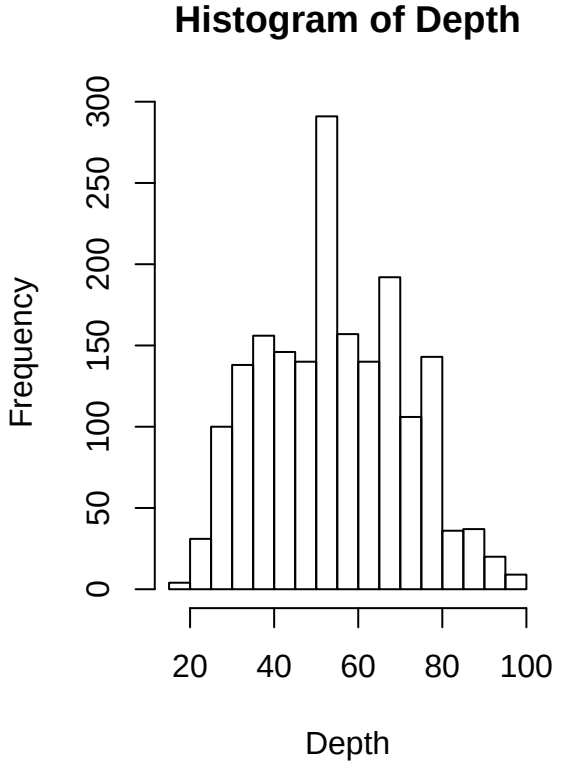


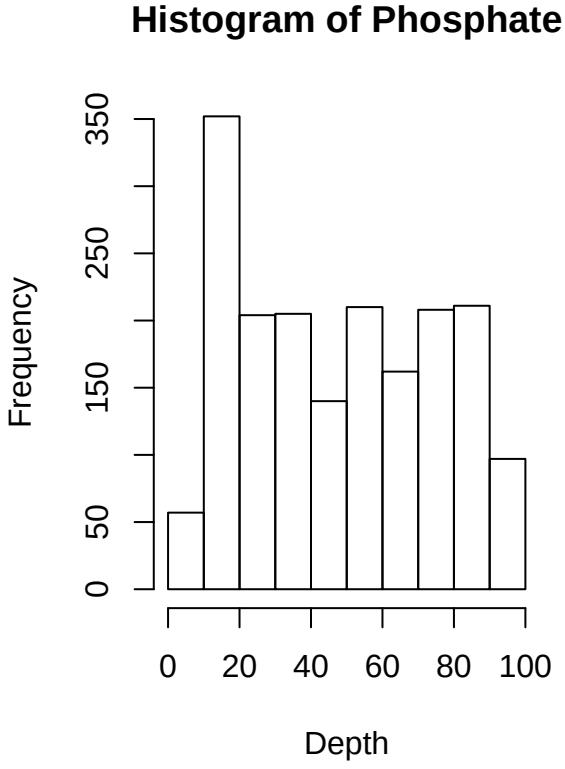




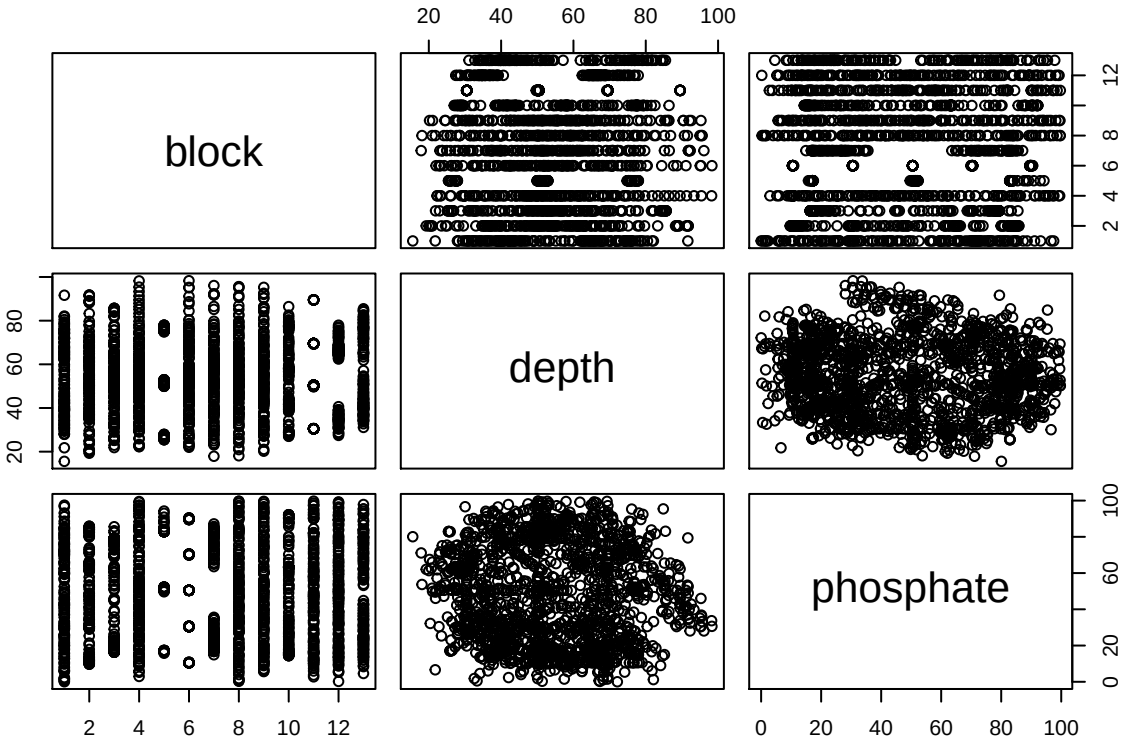



```
# Additional multipanel plots
par(mfrow = c(1,2))
hist(data$depth, main = 'Histogram of Depth', xlab = 'Depth', ylab = 'Frequency')
hist(data$phosphate, main = 'Histogram of Phosphate', xlab = 'Depth', ylab = 'Frequency')
```





```r
# Correlation Plots
pairs(data[1:3])
```



In our factor exploration, we grouped both continuous variables according to both categorical variables. There don't appear to be differences in the distributions of either depth or phosphate measurements across days or blocks. This is confirmed in our scatter plots - the greatest difference across blocks is simply the numer of measurements included in each. For example, block 6 includes fewer measurements than block 8 (which is easily visible in the paired scatter plots).

Problem 5

```r
# Define function
scatterhist = function(x, y, xlab="", ylab=""){
  zones=matrix(c(2,0,1,3), ncol=2, byrow=TRUE)
  layout(zones, widths=c(4/5,1/5), heights=c(1/5,4/5))
  xhist = hist(x, plot=FALSE)
  yhist = hist(y, plot=FALSE)
  top = max(c(xhist$counts, yhist$counts))
  par(mar=c(4,4,1,1))
  plot(x,y, xlab = ' ', ylab = ' ')
  par(mar=c(0,3,1,1))
  barplot(xhist$counts, axes=FALSE, ylim=c(0, top), space=0)
  par(mar=c(3,0,1,1))
  barplot(yhist$counts, axes=FALSE, xlim=c(0, top), space=0, horiz=TRUE)
  par(oma=c(3,3,0,0))
  mtext(xlab, side=1, line=1, outer=TRUE, adj=0,
    at=.8 * (mean(x) - min(x))/(max(x)-min(x)))
  mtext(ylab, side=2, line=1, outer=TRUE, adj=0,
    at=(.8 * (mean(y) - min(y))/(max(y) - min(y))))
```

```
}

# Use dataset 'iris'
# Call function for 'sepal.length' and 'sepal.width'
scatterhist(iris$Sepal.Length, iris$Sepal.Width, 'Sepal Length', 'Sepal Width')
```