HW4_Anwyll_Tessa

Tessa Anwyll

September 14, 2018

Problem 3

Roger Peng says that the main goals of exploratory data analysis include “identifying relationships between variables that are particularly interesting or unexpected, checking to see if there is evidence for or against a stated hypothesis, checking for problems with collected data (such as missing data or a measurement error), or identifying certain areas where more data need to be collected…It allows the investigator to make critical decisions about what to follow up on and what probabily isn’t worth pursuing because the data just don’t provide the evidence (and might never provide the evidence even with follow up). These kinds of decisions are important to make if a project is to move forward and remain within its budget.” (Peng, 1) Additionally, I think it is important so you can set your goals and decide what kind of analysis will need to be done, and it also helps you check your assumptions once you decide on a course of action.

Problem 4

1.

```r
library(xlsx)

# Read in data and combine both sheets into a single data frame
prob4_data1 <- read.xlsx("HW4_data.xlsx", sheetIndex = 1)
prob4_data2 <- read.xlsx("HW4_data.xlsx", sheetIndex = 2)
prob4Data <- rbind(prob4_data1, prob4_data2)


# Make vectors to simplify overall summary statistics computation
block <- prob4Data$block
depth <- prob4Data$depth
phos <- prob4Data$phosphate

# Put initial summary information into data frame and make label vector for
# columns and rows
dataSummary <- rbind.data.frame(summary(depth), summary(phos))
sumNames <- c("Min", "Q1", "Med", "Mean", "Q3", "Max")
rown <- c("Overall Depth", "Overall Phosphate")

# Find summary information by block for both phosphate and depth and add
# to data frame
for(i in 1:13){
  dataSummary <- rbind.data.frame(dataSummary, summary
                 (prob4Data[which(prob4Data$block == i), "depth"]),
                  summary(prob4Data[which(prob4Data$block == i), "phosphate"]))
  rown <- c(rown, paste("Block", i, "Depth"), paste("Block", i, "Phosphate"))
}

# Add names to data table
```

```r
colnames(dataSummary) <- sumNames
rownames(dataSummary) <- row

# Calculate IQRs and Ranges and add  to data summary table
dataIQR <- dataSummary[,"Q3"]-dataSummary[,"Q1"]
dataRange <- dataSummary[,"Max"]-dataSummary[,"Min"]
dataSummary <- cbind.data.frame(dataSummary, dataIQR, dataRange)

# Initialize sd vector and calculate standard deviation within each block
# for phosphate and depth
dataSD <- c(sd(depth), sd(phos))
for(i in 1:13){
  dataSD <- c(dataSD, sd(prob4Data[which(prob4Data$block == i), "depth"]),
              sd(prob4Data[which(prob4Data$block == i), "phosphate"]))
}
# Initialize variance vector and calculate variance within each block
# for phosphate and depth
dataVar <- c(var(depth), var(phos))
for(i in 1:13){
  dataVar <- c(dataVar, var(prob4Data[which(prob4Data$block == i), "depth"]),
               var(prob4Data[which(prob4Data$block == i), "phosphate"]))
}

# add remaining columns and column names and print summary data
dataSummary <- cbind.data.frame(dataSummary, dataVar, dataSD)
sumNames <- c(sumNames, "IQR", "Range", "Variance", "Standard Deviation")
colnames(dataSummary) <- sumNames
dataSummary
```

```
##                             Min       Q1      Med     Mean       Q3
## Overall Depth      15.56074952 41.07340 52.59127 54.26570 67.27784
## Overall Phosphate   0.01511933 22.56107 47.59445 47.83510 71.81078
## Block 1 Depth      15.56074952 39.72412 53.34030 54.26610 69.14660
## Block 1 Phosphate   0.01511933 24.62589 47.53527 47.83472 71.80315
## Block 2 Depth      19.28820474 41.62797 53.84209 54.26873 64.79890
## Block 2 Phosphate   9.69154713 26.24473 47.38294 47.83082 72.53285
## Block 3 Depth      21.86358128 43.37912 54.02321 54.26732 64.97267
## Block 3 Phosphate  16.32654637 18.34961 51.02502 47.83772 77.78238
## Block 4 Depth      22.30770000 44.10260 53.33330 54.26327 64.74360
## Block 4 Phosphate   2.94870000 25.28845 46.02560 47.83225 68.52567
## Block 5 Depth      25.44352570 50.35971 50.97677 54.26030 75.19736
## Block 5 Phosphate  15.77189199 17.10714 51.29929 47.83983 82.88159
## Block 6 Depth      22.00370914 42.29383 53.06968 54.26144 66.76827
## Block 6 Phosphate  10.46391519 30.47991 50.47353 47.83025 70.34947
## Block 7 Depth      17.89349871 41.53598 54.16869 54.26881 63.95267
## Block 7 Phosphate  14.91396246 22.92084 32.49920 47.83545 75.94002
## Block 8 Depth      18.10947229 42.89093 53.13516 54.26785 64.46999
## Block 8 Phosphate   0.30387242 27.84086 46.40131 47.83590 68.43943
## Block 9 Depth      20.20977816 42.81087 54.26135 54.26588 64.48801
## Block 9 Phosphate   5.64577748 24.75625 45.29224 47.83150 70.85584
## Block 10 Depth     27.02460324 41.03421 56.53473 54.26734 68.71149
## Block 10 Phosphate 14.36559047 20.37414 50.11055 47.83955 63.54858
## Block 11 Depth     30.44965384 49.96451 50.36289 54.26993 69.50407
## Block 11 Phosphate  2.73476017 22.75288 47.11362 47.83699 65.84539
## Block 12 Depth     27.43963221 35.52245 64.55023 54.26692 67.45367
## Block 12 Phosphate  0.21700627 24.34694 46.27933 47.83160 67.56813
## Block 13 Depth     31.10686656 40.09166 47.13646 54.26015 71.85692
## Block 13 Phosphate  4.57766135 23.47081 39.87621 47.83972 73.60963
##                          Max      IQR    Range Variance Standard Deviation
## Overall Depth      98.28812 26.20444 82.72737 279.3244           16.71300
## Overall Phosphate  99.69468 49.24971 99.67956 720.8026           26.84777
## Block 1 Depth      91.63996 29.42248 76.07921 281.2270           16.76982
## Block 1 Phosphate  97.47577 47.17726 97.46065 725.7498           26.93974
## Block 2 Depth      91.73554 23.17093 72.44733 281.2074           16.76924
## Block 2 Phosphate  85.87623 46.28812 76.18468 725.5334           26.93573
## Block 3 Depth      85.66476 21.59356 63.80118 280.8980           16.76001
## Block 3 Phosphate  85.57813 59.43277 69.25159 725.2268           26.93004
## Block 4 Depth      98.20510 20.64100 75.89740 281.0700           16.76514
## Block 4 Phosphate  99.48720 43.23722 96.53850 725.5160           26.93540
## Block 5 Depth      77.95444 24.83766 52.51091 281.1570           16.76774
## Block 5 Phosphate  94.24933 65.77445 78.47744 725.2352           26.93019
## Block 6 Depth      98.28812 24.47445 76.28441 281.0953           16.76590
## Block 6 Phosphate  90.45894 39.86956 79.99502 725.7569           26.93988
## Block 7 Depth      96.08052 22.41669 78.18702 281.1224           16.76670
## Block 7 Phosphate  87.15221 53.01918 72.23825 725.7635           26.94000
## Block 8 Depth      95.59342 21.57906 77.48394 281.1242           16.76676
## Block 8 Phosphate  99.64418 40.59857 99.34031 725.5537           26.93610
## Block 9 Depth      95.26053 21.67714 75.05075 281.1944           16.76885
## Block 9 Phosphate  99.57959 46.09960 93.93381 725.6886           26.93861
## Block 10 Depth     86.43590 27.67728 59.41129 281.1980           16.76896
## Block 10 Phosphate 92.21499 43.17445 77.84940 725.2397           26.93027
## Block 11 Depth     89.50485 19.53956 59.05520 281.2315           16.76996
## Block 11 Phosphate 99.69468 43.09251 96.95992 725.6388           26.93768
## Block 12 Depth     77.91587 31.93122 50.47624 281.2329           16.77000
## Block 12 Phosphate 99.28376 43.22119 99.06676 725.6506           26.93790
## Block 13 Depth     85.44619 31.76527 54.33932 281.2315           16.76996
## Block 13 Phosphate 97.83761 50.13882 93.25995 725.2250           26.93000
```

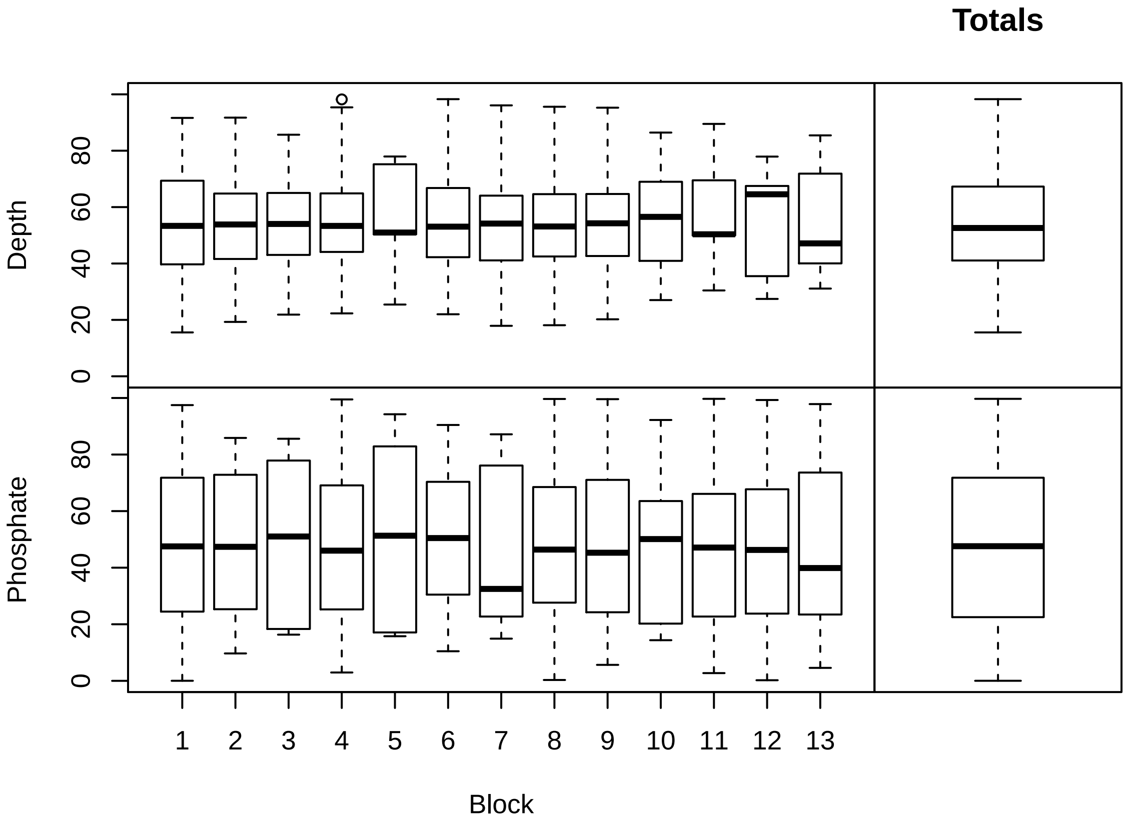
2. Factor exploration

The factors are block, depth and phosphate. Depth and phosphate are continuous numerical variables and block is categorical with levels 1-13. There are 142 observations for each block

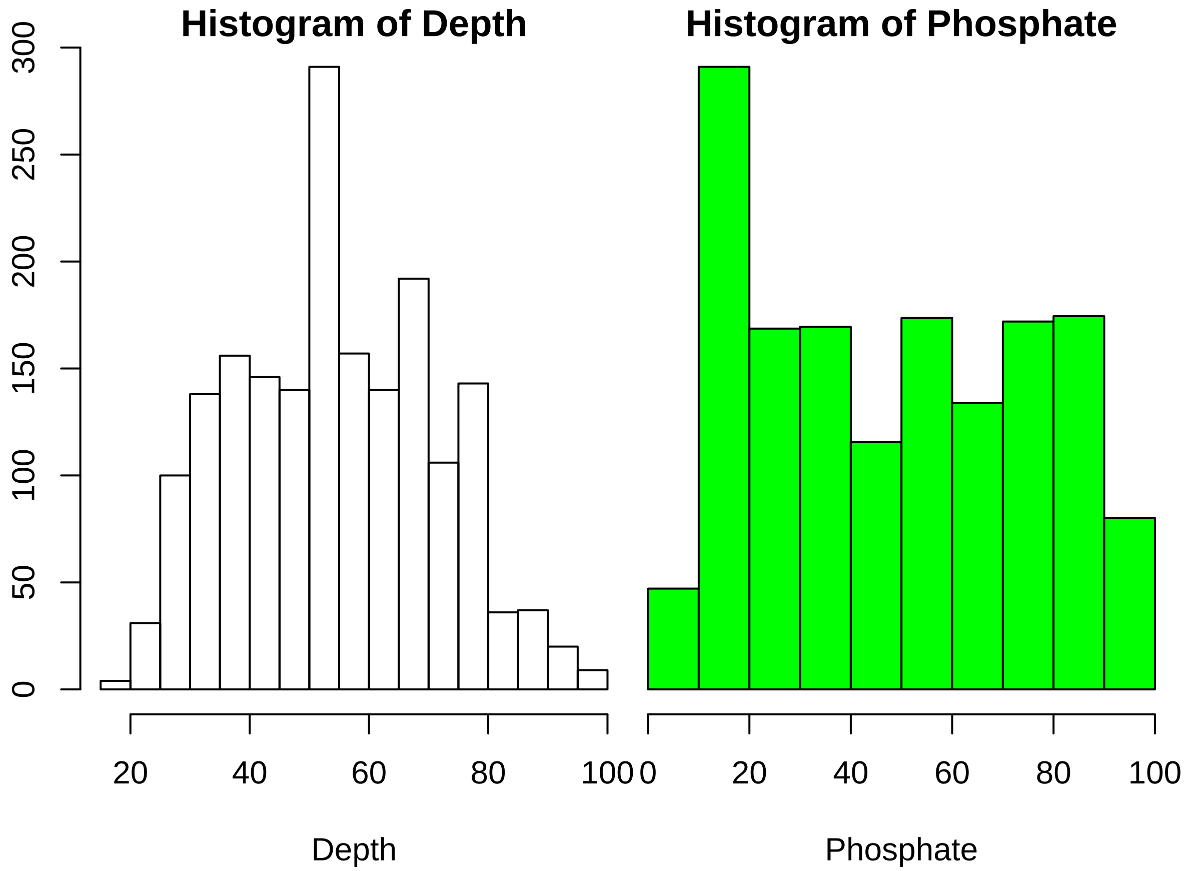
3. Multi Panel Plots

```r
# create matrix to hold each graph in multipanel plot
matrixLay <- matrix(c(1, 2, 3, 4), ncol=2, byrow=TRUE)
# set widths of each plot location
layout(matrixLay, widths=c(7/10, 3/10), heights=c(5/10, 5/10))
#set margins for first plot
par(mar = c(0, 4, 4, 0))
# make boxplot of depth vs block
boxplot(prob4Data$depth~prob4Data$block, ylim = c(0,100), xaxt = "n", ylab = "Depth")
# set margins for second graph
par(mar = c(0, 0, 4, 4))
# plot boxplot of all depth information
```

```r
boxplot(prob4Data$depth, ylim = c(0,100), main = "Totals", yaxt = "n")
# set margins for third plot
par(mar = c(4, 4, 0, 0))
# create boxplot of phosphate by block
boxplot(prob4Data$phosphate~prob4Data$block, xlab = "Block", ylab = "Phosphate")
# set margins for fourth plot
par(mar = c(4, 0, 0, 4))
# create boxplot of all phosphate data
boxplot(prob4Data$phosphate, yaxt = "n")
```

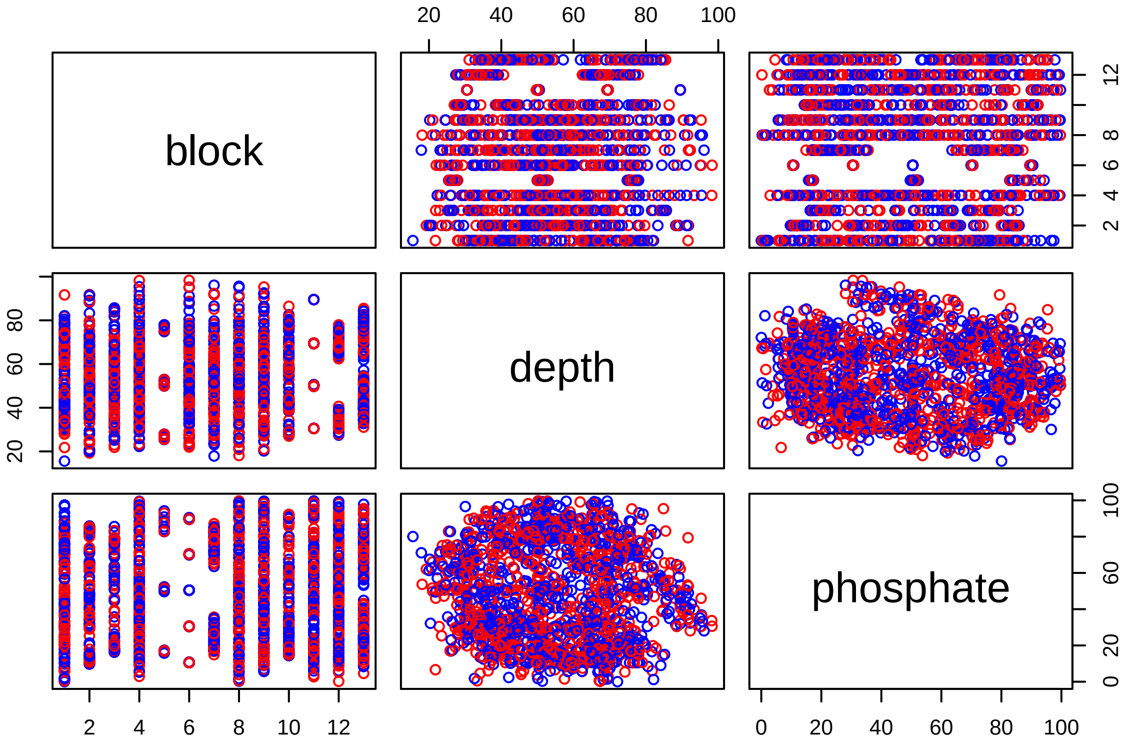


```r
# set layout for 2 graphs and set margin for first graph
par(mfrow = c(1,2), mar = c(4, 2, 1, 0))
# plot histogram in first cell created by par function
hist(prob4Data$depth, xlab = "Depth", main = "Histogram of Depth")
# set margins for 2nd histogram
par(mar = c(4, 0, 1, 2))
# plot second histogram in second cell created by par function
hist(prob4Data$phosphate, col = "green", xlab = "Phosphate",
     main = "Histogram of Phosphate", ylab = NULL, yaxt = "n")
```



### 4. Correlation plots

```
# use pairs function to create a correlation matrix
pairs(prob4Data, col = c("blue", "red"))
```

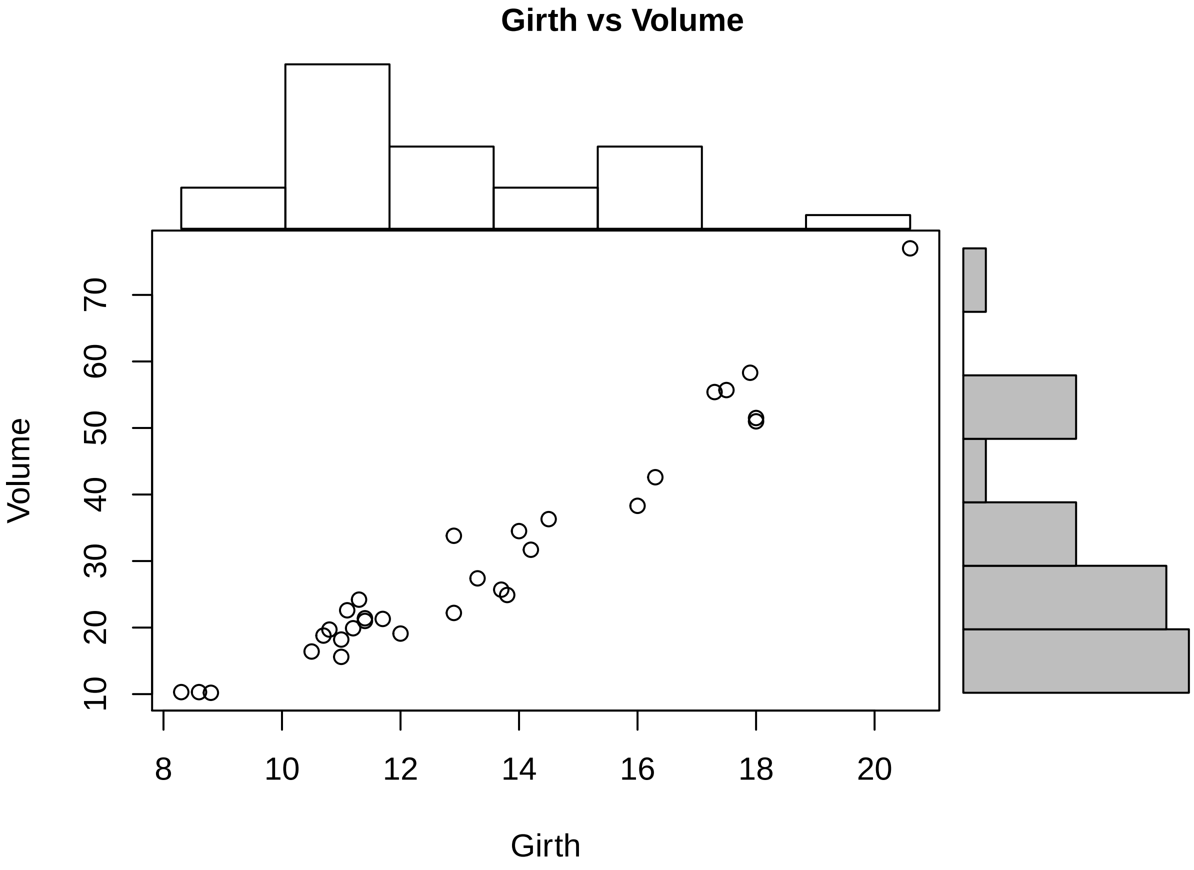


### 5. Just looking at the summary statistics and the plots, there appears to be nothing all that interesting to investigate. We can see that phosphate tends to have a little bit more spread than depth and that depth has a few blocks where the data is skewed to one end of the box, but other than that, there doesn't appear to be a lot that would be worth the time to perform data analysis on this data set.

Problem 5

```
# Make function to make multi-panel plots with inputs for data and labels
multiPlot <- function(data1, data2, xlabel = "x", ylabel = "y"){
# set margins and sizes for scatter plot
par(fig=c(0,0.8,0,0.8), mar = c(4, 4, 1.5, 1.5) )
# scatter plot of data1 vs data2
plot(data1, data2, xlab= xlabel, ylab= ylabel)
# set margins for top histogram
par(fig=c(0,0.8, .55 ,1), new = TRUE)
# plot histogram of x data
hist(data1, axes=FALSE, main = NULL, xlab = NULL, ylab = NULL)
# use hist function to generate the information for a histogram
# without plotting it
yhist <- hist(data2, plot = FALSE)
# set margins for y "histogram"
par(fig=c(0.65,1,0,.8), new = TRUE)
# use density information from the hist function in the barplot function to utilize
# horiz parameter
barplot(yhist$density, axes=FALSE, xlim=c(0, max(yhist$density)), space=0, horiz=TRUE)
# add title in margin
mtext(paste(xlabel, "vs", ylabel), side=3, outer=TRUE, line=-1, font = 2)
}

# make plot using trees data
multiPlot(trees[,1], trees[,3], colnames(trees)[1], colnames(trees)[3])
```



Sources

Peng, Roger D. **Exploratory Data Analysis with R**. 2016. Leanpub.com, Web. 14 Sep. 2018 https://www.statmethods.net/advgraphs/layout.html