STAT 5014 HW4

Max McGill

2017-09-25

Problem 3

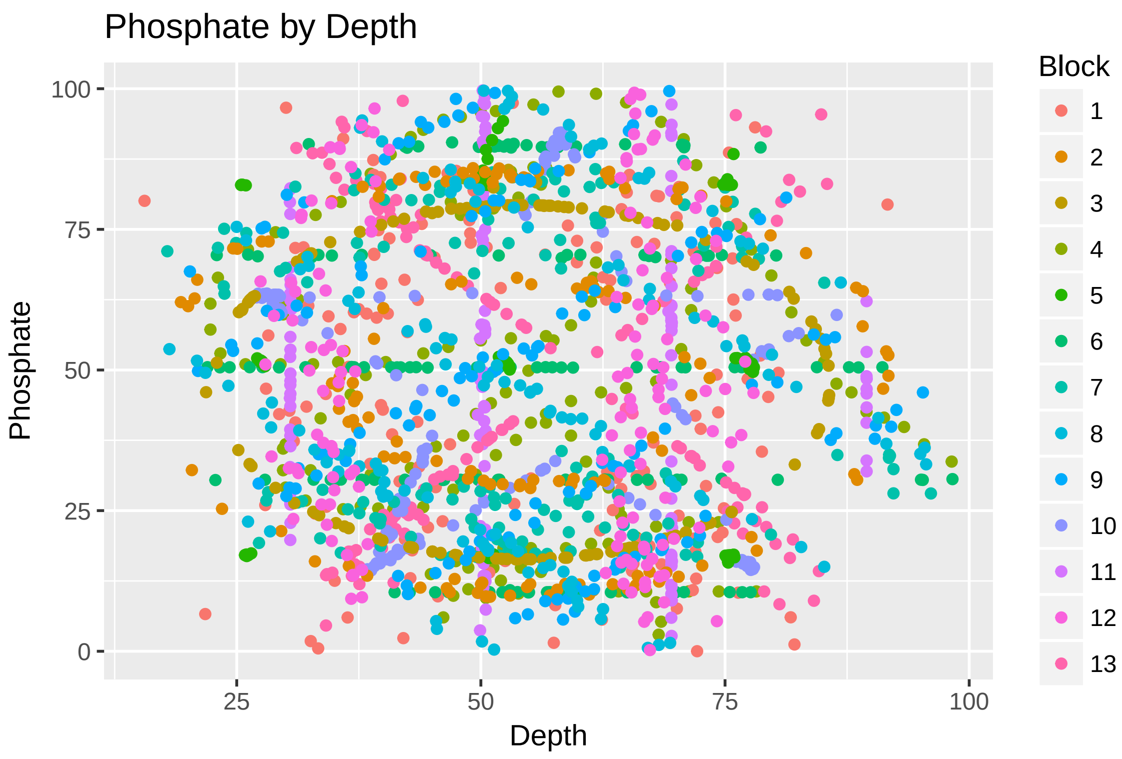
According to Roger Peng's *Exploratory Data Analysis with R*, the focus of EDA "is to show the data, summarize the evidence and identify interesting patterns while eliminating ideas that likely won't pan out" (2016, p. 2).

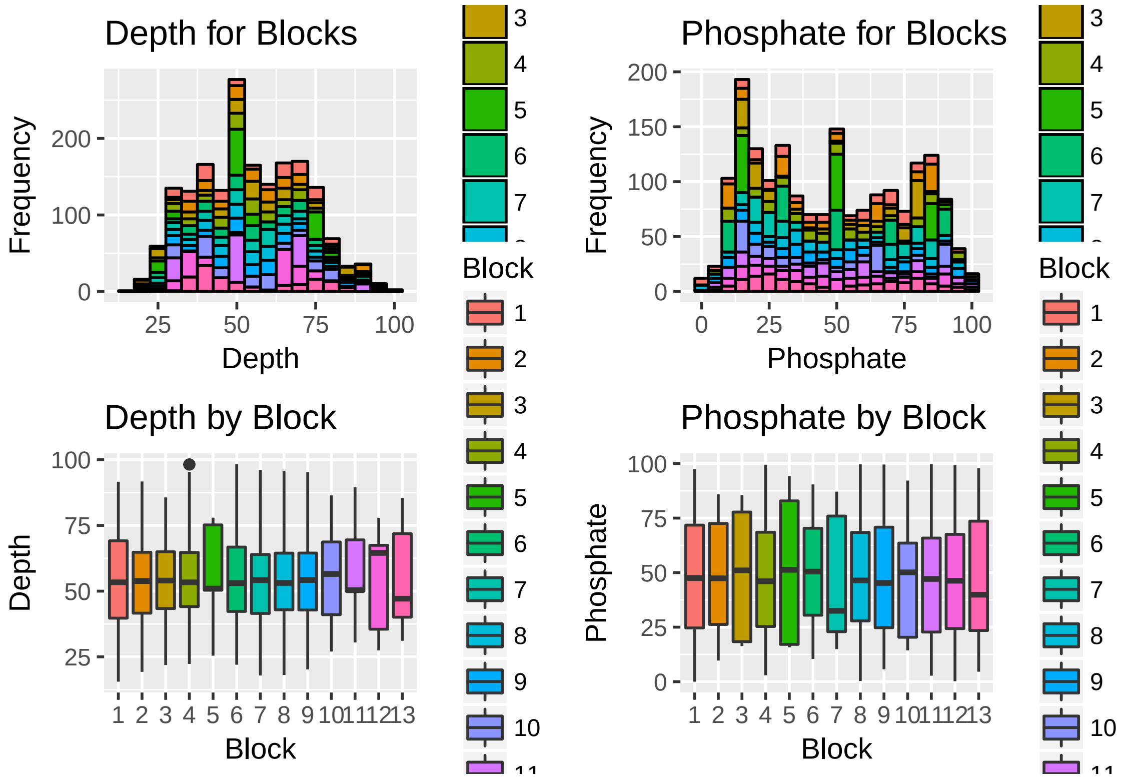
Problem 4

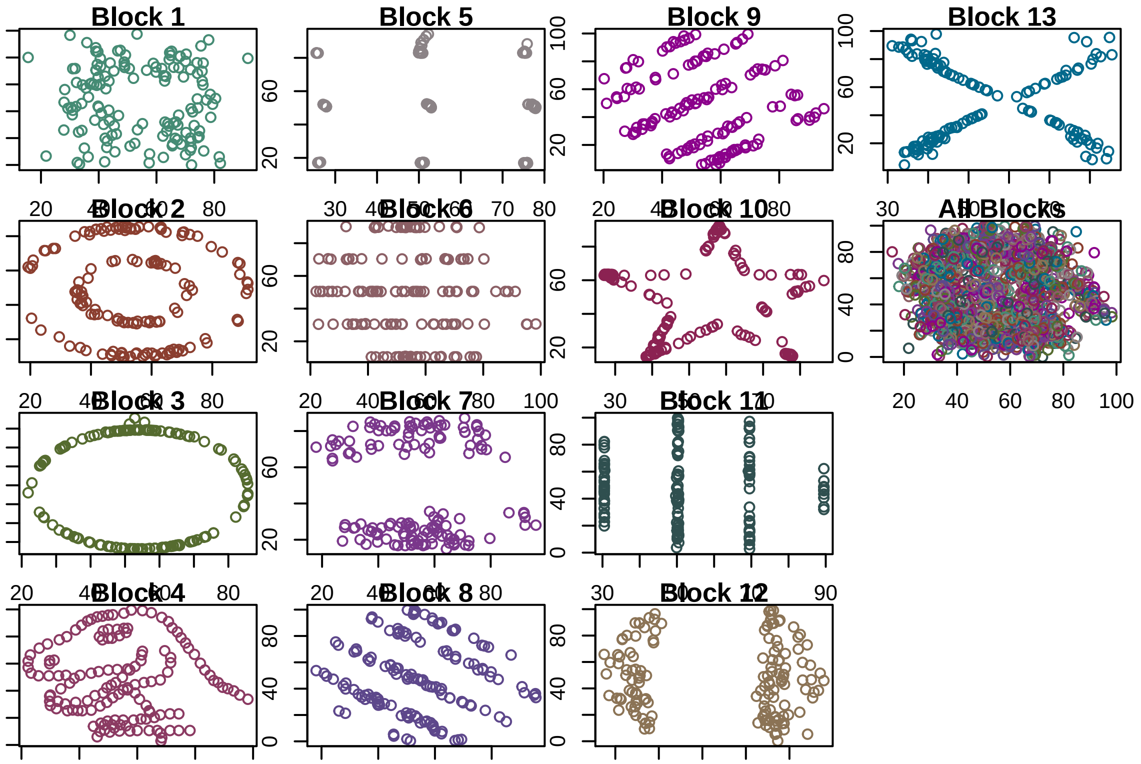
Factors within this dataset are limited to the numerical designation by block, while depth and phosphate are continuous, quantitative observations. The lesson from this exploration is to always loo over your data to ensure there are no lurking dinosaurs, cookie cutter shapes, or other obscuring fabrications.

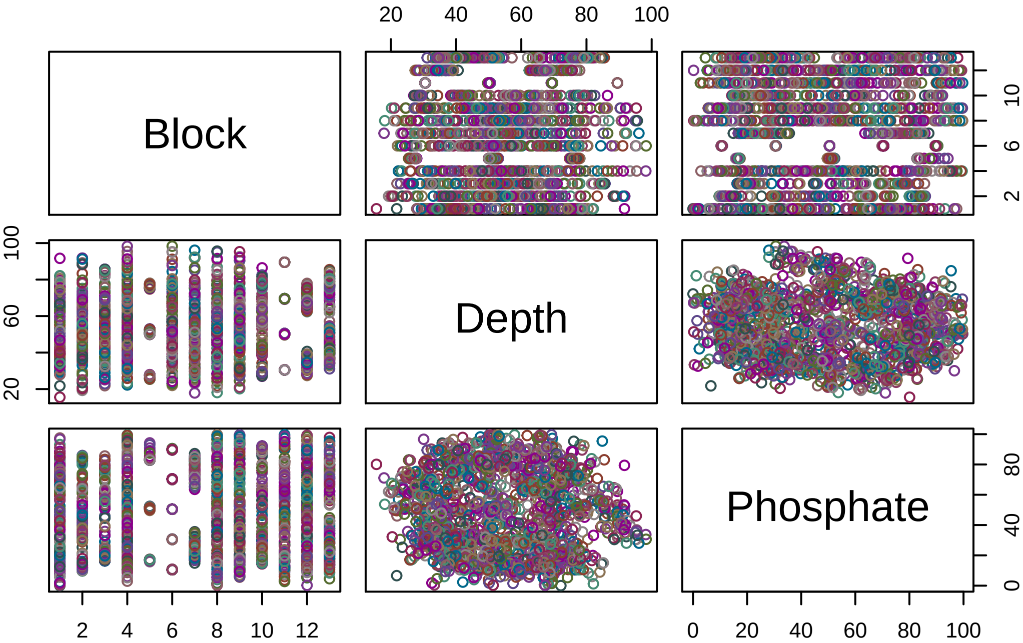
Table 1: Data Summary

block	depth	phosphate
Min. : 1	Min. :15.6	Min. : 0.0151
1st Qu.: 4	1st Qu.:41.1	1st Qu.:22.5611
Median : 7	Median :52.6	Median :47.5944
Mean : 7	Mean :54.3	Mean :47.8351
3rd Qu.:10	3rd Qu.:67.3	3rd Qu.:71.8108
Max. :13	Max. :98.3	Max. :99.6947


Phosphate by Depth
Phosphate
100
75
50
25
0
25
50
75
100
Depth
Block
1
2
3
4
5
6
7
8
9
10
11
12
13


Depth for Blocks
Frequency
200
100
0
25
50
75
100
Depth
3
4
5
6
7
Block
1
2
3
4
5
6
7
8
9
10
11
Phosphate for Blocks
Frequency
200
150
100
50
0
0
25
50
75
100
Phosphate
3
4
5
6
7
Block
1
2
3
4
5
6
7
8
9
10
11
Depth by Block
Depth
100
75
50
25
1 2 3 4 5 6 7 8 9 10 11 12 13
Block
Phosphate by Block
Phosphate
100
75
50
25
0
1 2 3 4 5 6 7 8 9 10 11 12 13
Block


Block 1
Block 5
Block 9
Block 13
Block 2
Block 6
Block 10
All Blocks
Block 3
Block 7
Block 11
Block 4
Block 8
Block 12
20
40
60
80
100
30
50
70
90
0


Block
Depth
Phosphate
20
40
60
80
100
2
6
10
4
8
12
0

Problem 5

The single most illuminating figure of this dataset is the multipanel figure of scatterplots that reveal the data of individual blocks.

Appendix 1: R Code

Problem 4

```r
# get data
library(xlsx)
prob4_data1 <- read.xlsx("HW4_data.xlsx", sheetIndex = 1)
prob4_data2 <- read.xlsx("HW4_data.xlsx", sheetIndex = 2)
# combine data
prob4data <- rbind(prob4_data1, prob4_data2)

# summarize data
knitr::kable(summary(prob4data), caption = "Data Summary")

library(ggplot2)

# from
# 'http://www.cookbook-r.com/Graphs/Multiple_graphs_on_one_page_(ggplot2)/'
multiplot <- function(..., plotlist = NULL, file, cols = 1,
    layout = NULL) {
    library(grid)

    # Make a list from the ... arguments and plotlist
    plots <- c(list(...), plotlist)

    numPlots = length(plots)

    # If layout is NULL, then use 'cols' to determine layout
    if (is.null(layout)) {
        # Make the panel ncol: Number of columns of plots nrow:
        # Number of rows needed, calculated from # of cols
        layout <- matrix(seq(1, cols * ceiling(numPlots/cols)),
            ncol = cols, nrow = ceiling(numPlots/cols))
    }

    if (numPlots == 1) {
        print(plots[[1]])

    } else {
        # Set up the page
        grid.newpage()
        pushViewport(viewport(layout = grid.layout(nrow(layout),
            ncol(layout))))

        # Make each plot, in the correct location
        for (i in 1:numPlots) {
            # Get the i,j matrix positions of the regions that
            # contain this subplot
            matchidx <- as.data.frame(which(layout == i,
                arr.ind = TRUE))

            print(plots[[i]], vp = viewport(layout.pos.row = matchidx$row,
                layout.pos.col = matchidx$col))
        }
    }
}

# begin plotting the data as a whole first a scatter
# plot
scatter4 <- ggplot(prob4data, aes(x = depth, y = phosphate,
    colour = factor(block))) + geom_point() + xlab("Depth") +
    ylab("Phosphate") + ggtitle("Phosphate by Depth") +
    guides(col = guide_legend(title = "Block"))

# to better understand the distributions, a histogram
# for both variables of interest
dhist4 <- ggplot(prob4data, aes(x = depth, fill = factor(block))) +
    geom_histogram(colour = "black", binwidth = 5) + xlab("Depth") +
    ylab("Frequency") + ggtitle("Depth for Blocks") + guides(fill = guide_legend(title = "Block"))
phist4 <- ggplot(prob4data, aes(x = phosphate, fill = factor(block))) +
    geom_histogram(colour = "black", binwidth = 5) + xlab("Phosphate") +
    ylab("Frequency") + ggtitle("Phosphate for Blocks") +
    guides(fill = guide_legend(title = "Block"))

# similar in purpose, a boxplot for both variables of
# interest
dbox4 <- ggplot(prob4data, aes(x = factor(block), y = depth,
    group = block, fill = factor(block))) + geom_boxplot() +
    xlab("Block") + ylab("Depth") + ggtitle("Depth by Block") +
    guides(fill = guide_legend(title = "Block"))
pbox4 <- ggplot(prob4data, aes(x = factor(block), y = phosphate,
    group = block, fill = factor(block))) + geom_boxplot() +
    xlab("Block") + ylab("Phosphate") + ggtitle("Phosphate by Block") +
    guides(fill = guide_legend(title = "Block"))

# plot collective scatter plot separately
scatter4
# combine into a multipanel plot using borrowed
# multipanel function
multiplot(dhist4, dbox4, phist4, pbox4, cols = 2)

# now looking at each block individually create a vector
# of colors in R
colvect <- c("aquamarine4", "coral4", "darkolivegreen",
    "hotpink4", "lavenderblush4", "lightpink4", "mediumorchid4",
    "mediumpurple4", "magenta4", "violetred4", "darkslategrey",
    "burlywood4", "deepskyblue4")

# set figure parameters
```

```r
par(mar = c(1, 1, 1, 1))
par(mfcol = c(4, 4))

# create a scatter plot for each block
for (i in 1:13) {
    subs <- subset(prob4data, block == i)
    p <- plot(x = subs$depth, y = subs$phosphate, xlab = "Depth",
        ylab = "Phosphate", col = colvect[i], main = paste("Block",
            i, sep = " "))
    assign(paste("p", i), p)
}

# recreate the combined scatter plot with altered colors
# for reference
plot(x = prob4data$depth, y = prob4data$phosphate, group.by = prob4data$block,
    col = colvect, xlab = "Depth", ylab = "Phosphate", main = "All Blocks")

par(mfcol = c(1, 1))
# create a matrix of correlation plots
pairs(prob4data, group.by = prob4data$block, col = colvect,
    labels = c("Block", "Depth", "Phosphate"))
```