# Statistics 5014: Homework 4

Due Wednesday September 19 in GitHub, 9 am

*2018-09-16*

## Problem 3: Exploratory Data Analysis with R (Roger D. Peng)

Exploratory data analyis (EDA) focuses on finding variables of key interest, identifying their relationships with one another, finding evidence to accept or reject a stated hypothesis, finding issues in the dataset (e.d., missing data, measurement errors) and determining areas where more data may be needed.

## Problem 4

1. Summary statistics

```
library(readxl)
library(knitr)
# Read in both spreadsheets from HW4_data.xlsx
prob4.data1 <- read_excel("HW4_data.xlsx", sheet = 1)
prob4.data2 <- read_excel("HW4_data.xlsx", sheet = 2)
# Add variable to indicate days
prob4.data1$day <- as.ordered(1)
prob4.data2$day <- as.ordered(2)
# Combine two sheets into one dataset
prob4.data <- rbind(prob4.data1, prob4.data2)
# Function to capitalize first letter of a string
FirstUp <- function(x) {
    substr(x, 1, 1) <- toupper(substr(x, 1, 1))
    x
}
colnames(prob4.data) <- FirstUp(colnames(prob4.data))  # Capitalize column names
# Convert Block variable to factor
prob4.data$Block <- as.ordered(prob4.data$Block)
options(knitr.kable.NA = " ")  # Don't print NA values in kable
# Print tabel with title
kable(summary(prob4.data), digits = 3, format = "pandoc",
    caption = "Table 1. Problem 4 summary statistics")
```

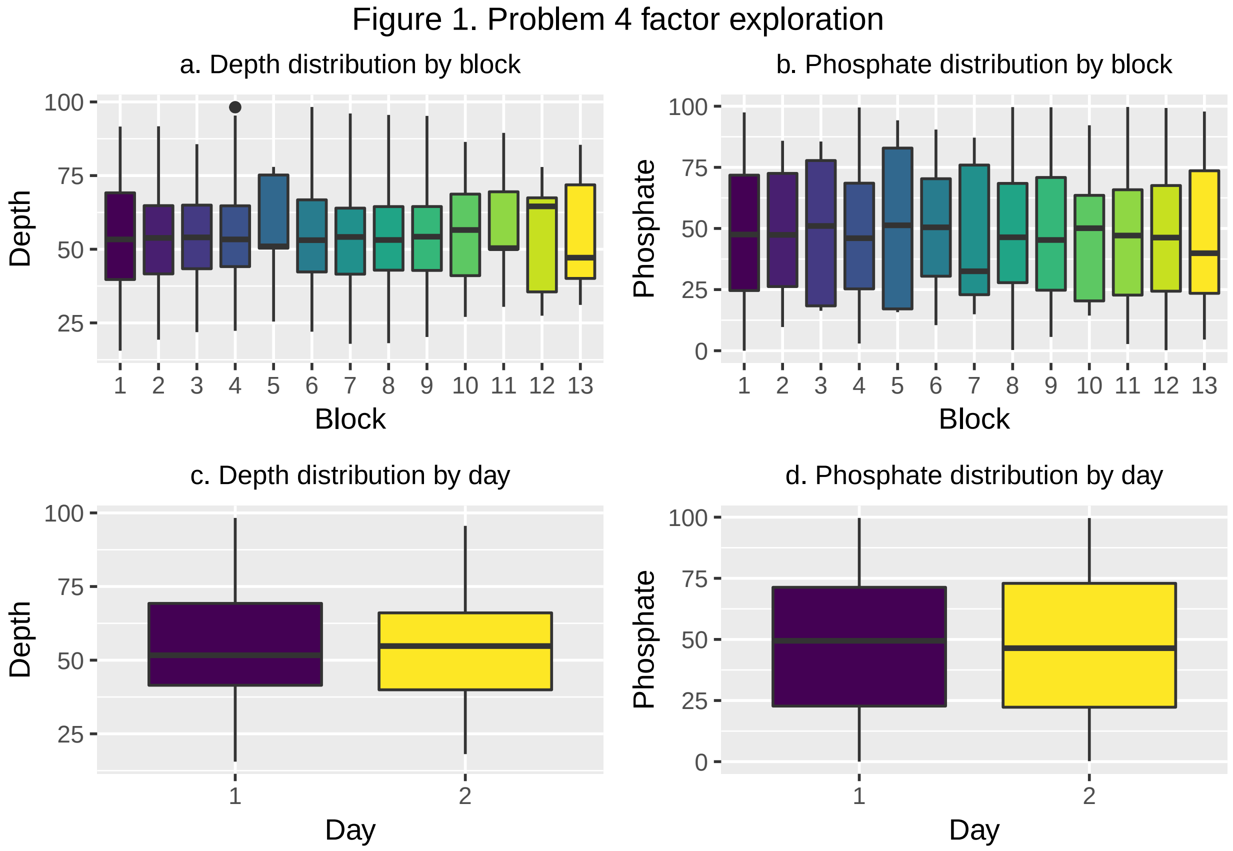
Table 1: Table 1. Problem 4 summary statistics

| Block | Depth | Phosphate | Day |
|---|---|---|---|
| 1 :142 | Min. :15.56 | Min. : 0.01512 | 1:1136 |
| 2 :142 | 1st Qu.:41.07 | 1st Qu.:22.56107 | 2: 710 |
| 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 | | | |

2. Factor exploration

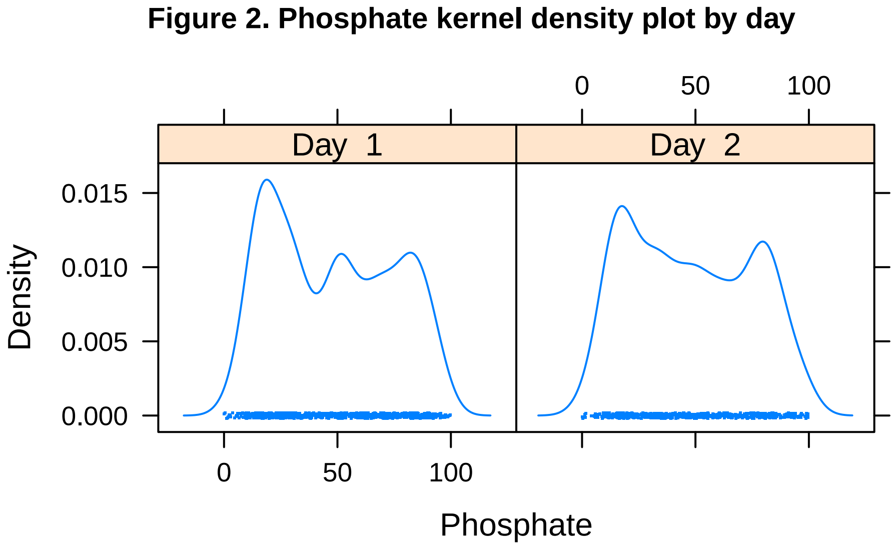
The variables “Block” and “Day” are assumed to be ordinal and hence, factors. Figure 1 visualizes the distributions of the aforementioned factors.

```
library(ggplot2)
library(gridExtra)
# Create 4 individual qplots with titles
prob4.plot1 <- qplot(Block, Depth, data = prob4.data, geom = "boxplot",
    fill = Block) + theme(legend.position = "none") + ggtitle("a. Depth distribution by block") +
    theme(plot.title = element_text(size = 10, hjust = 0.5))
prob4.plot2 <- qplot(Block, Phosphate, data = prob4.data,
    geom = "boxplot", fill = Block) + theme(legend.position = "none") +
    ggtitle("b. Phosphate distribution by block") + theme(plot.title = element_text(size = 10,
    hjust = 0.5))
prob4.plot3 <- qplot(Day, Depth, data = prob4.data, geom = "boxplot",
    fill = Day) + theme(legend.position = "none") + ggtitle("c. Depth distribution by day") +
    theme(plot.title = element_text(size = 10, hjust = 0.5))
prob4.plot4 <- qplot(Day, Phosphate, data = prob4.data,
    geom = "boxplot", fill = Day) + theme(legend.position = "none") +
    ggtitle("d. Phosphate distribution by day") + theme(plot.title = element_text(size = 10,
    hjust = 0.5))
# Combine 4 plots into a 2 x 2 grid and add overall
# title
grid.arrange(prob4.plot1, prob4.plot2, prob4.plot3, prob4.plot4,
    nrow = 2, top = "Figure 1. Problem 4 factor exploration")
```



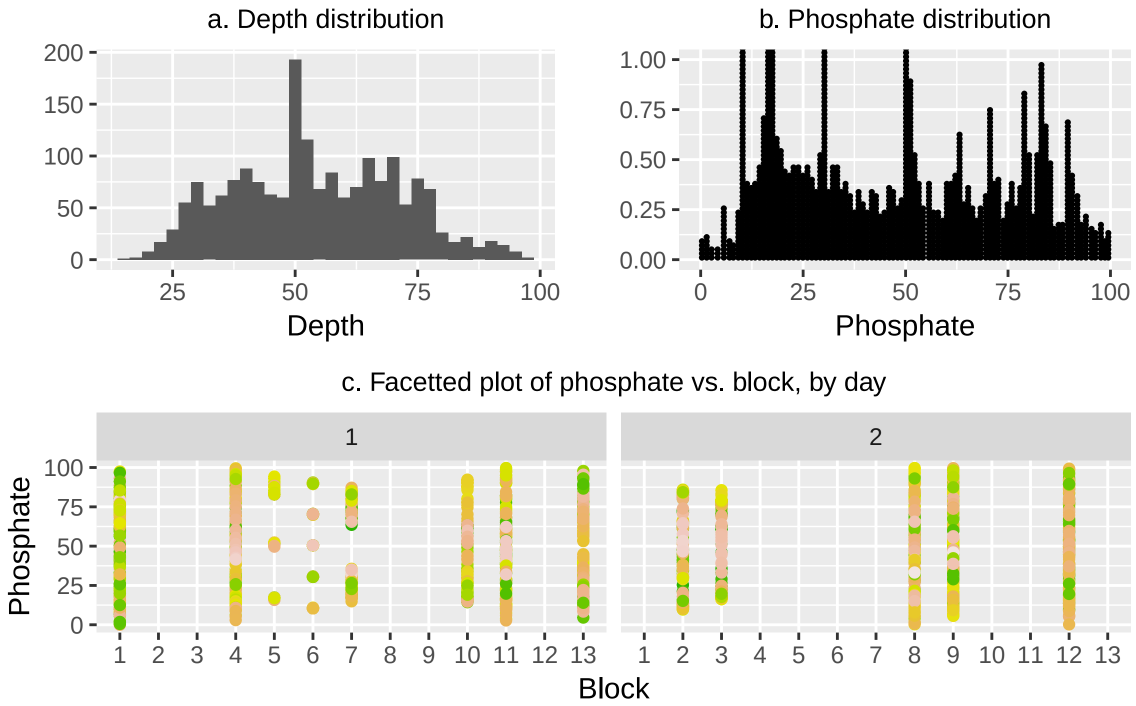
3. Multipanel plots

```
# Create multipanel density plots with lattice
library(lattice)
attach(prob4.data)
# Create factor labels
day.f <- factor(Day, levels = 1:2, labels = paste("Day ",
    levels(prob4.data$Day)))
# Print density plot
densityplot(~Phosphate | day.f, main = list("Figure 2. Phosphate kernel density plot by day",
    fontsize = 9), xlab = "Phosphate", pch = ".")
```



```
# Histogram of Depth
prob4.plot5 <- qplot(Depth, data = prob4.data, binwidth = 2.5) +
    ggtitle("a. Depth distribution") + theme(plot.title = element_text(size = 10,
    hjust = 0.5))
# Dotplot of Phosphate
prob4.plot6 <- qplot(Phosphate, data = prob4.data, geom = "dotplot",
    binwidth = 1) + ggtitle("b. Phosphate distribution") +
    theme(plot.title = element_text(size = 10, hjust = 0.5))
# Faceted scatterplot with a color gradient for
# continuous variable Depth
prob4.plot7 <- qplot(Block, Phosphate, data = prob4.data,
    colour = Depth) + scale_colour_gradientn(colours = terrain.colors(10)) +
    facet_wrap(~Day, nrow = 1) + theme(legend.position = "none") +
    ggtitle("c. Facetted plot of phosphate vs. block, by day") +
    theme(plot.title = element_text(size = 10, hjust = 0.5))
grid.arrange(prob4.plot5, prob4.plot6, prob4.plot7, layout_matrix = rbind(c(1,
    2), c(4, 4)), top = "Figure 3. Multipanel plot example")
```

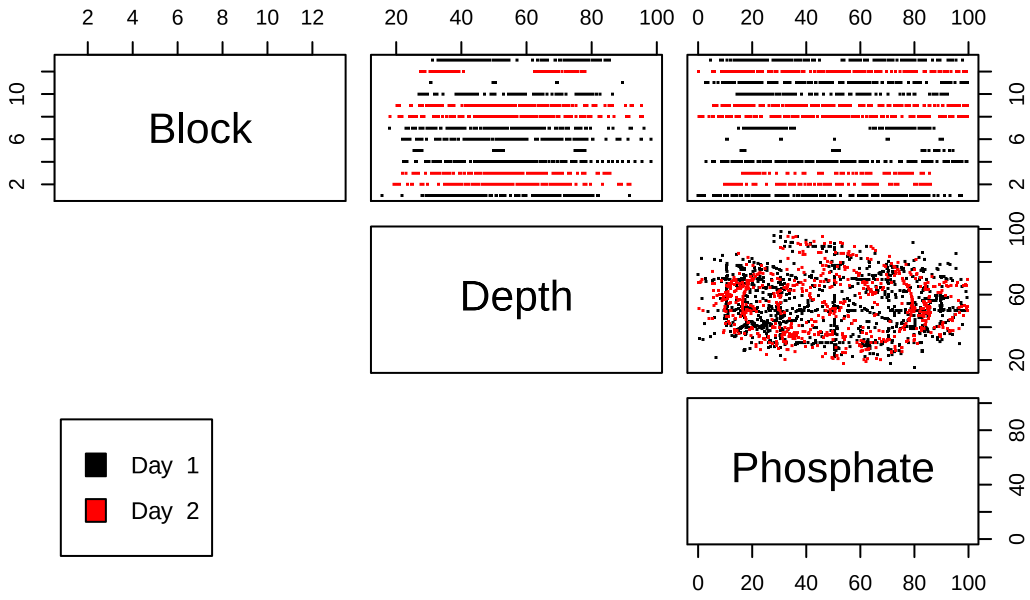
Figure 3. Multipanel plot example



4. Correlation plot

```r
pairs(prob4.data[1:3], main = "Figure 4. Problem 4 scatterplot matrix",
    cex.main = 1, pch = ".", col = prob4.data$Day, lower.panel = NULL)
par(xpd = T)
legend("bottomleft", inset = 0.07, fill = unique(prob4.data$Day),
    legend = paste("Day ", levels(prob4.data$Day)), cex = 0.75)
```

**Figure 4. Problem 4 scatterplot matrix**



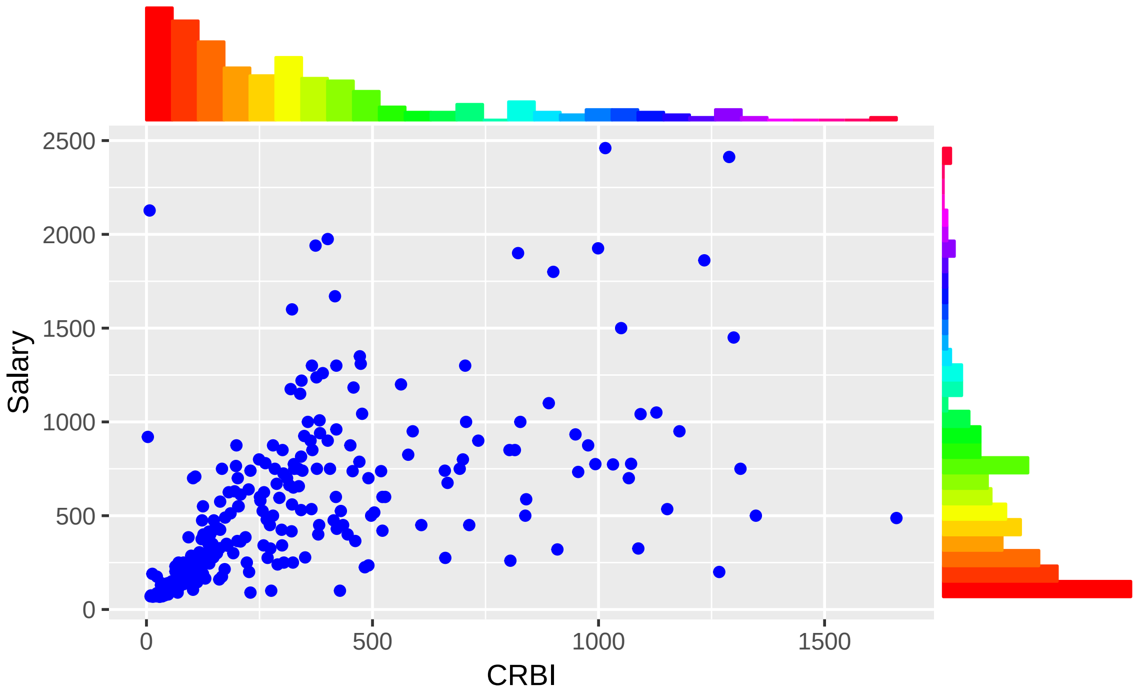
5. A lesson from this dataset

When the data in the current problem was imported as is, the ordinal factors were classified as numeric by default. Not taking this into consideration may have led to misleading statistical summaries or plots. Therefore, upon gathering and tidying datasets, the type of each variable needs to be carefully taken under consideration.

## Problem 5: Create a Scatterplot with Marginal Histograms

```r
# Function requires a 2-variable numerical dataset
library(ggExtra)
MultiHist <- function(input.dataset, bin.num = round(nrow(input.dataset)/10,
    digits = -1)) {
    # Check if the input is a 2-variable dataset
    if (ncol(input.dataset) != 2) {
        stop("Please input a 2-variable dataset.")
    }
    # Check if the input variables are numeric
    if (!all(sapply(input.dataset, is.numeric))) {
        stop("Please input numeric variables only.")
    }
    # Create scatter plot, axis labels, and title
    prob5.plot <- ggplot(input.dataset, aes(x = input.dataset[[colnames(input.dataset)[1]]],
        y = input.dataset[[colnames(input.dataset)[2]]])) +
        geom_point(color = "blue") + theme(legend.position = "none") +
        theme(plot.title = element_text(size = 12, hjust = 0.5)) +
        labs(title = paste("Figure 5. ", colnames(input.dataset)[2],
            "vs. ", colnames(input.dataset)[1]), y = colnames(input.dataset)[2],
            x = colnames(input.dataset)[1])
    # Create marginal histograms
    ggMarginal(prob5.plot, type = "histogram", size = 4,
        bins = bin.num, fill = rainbow(bin.num - 1), color = rainbow(bin.num -
            1))
}
# Major League Baseball Data from the 1986 and 1987
# seasons CRBI: Number of runs batted in during his
# career Salary: 1987 annual salary on opening day in
# thousands of dollars
prob5.data <- fread("https://vincentarelbundock.github.io/Rdatasets/csv/ISLR/Hitters.csv",
    select = c("CRBI", "Salary"), verbose = T)
# Call function using Major League Baseball Data
MultiHist(prob5.data)
```

Figure 5. Salary vs. CRBI


Salary
CRBI