STAT 5014 - Homework 3

Ryan Kingery

9/20/2017

Problem 3

According to the Peng's book, EDA allows one to identify interesting relationships between variables, to check to see if there is or isn't evidence to support or question a hypothesis, or to check for problems with the data set. EDA allows one to decide which questions about the data set are worth pursuing further given time and budgetary constraints on a project.

Problem 4

For this problem I couldn't get the xlsx package to load correctly due to some issue with the rJava package not loading correctly. Therefore, instead of loading the sheets in as xlsx files, I loaded them in separately as csv files and then joined them together. I also went ahead and loaded the separate blocks into indivdual dataframes for later analysis. We can see that the dataframe contains three variables: block, depth, and phoshate.

```
raw_1 <- read.csv("~/Desktop/HW4_data_1.csv")
raw_2 <- read.csv("~/Desktop/HW4_data_2.csv")
df <- bind_rows(raw_1, raw_2)
b1 = subset(df, df$block == 1)
b2 = subset(df, df$block == 2)
b3 = subset(df, df$block == 3)
b4 = subset(df, df$block == 4)
head(df)
```

```
##   block   depth phosphate
## 1     4 55.3846   97.1795
## 2     4 51.5385   96.0256
## 3     4 46.1538   94.4872
## 4     4 42.8205   91.4103
## 5     4 40.7692   88.3333
## 6     4 38.7179   84.8718
```

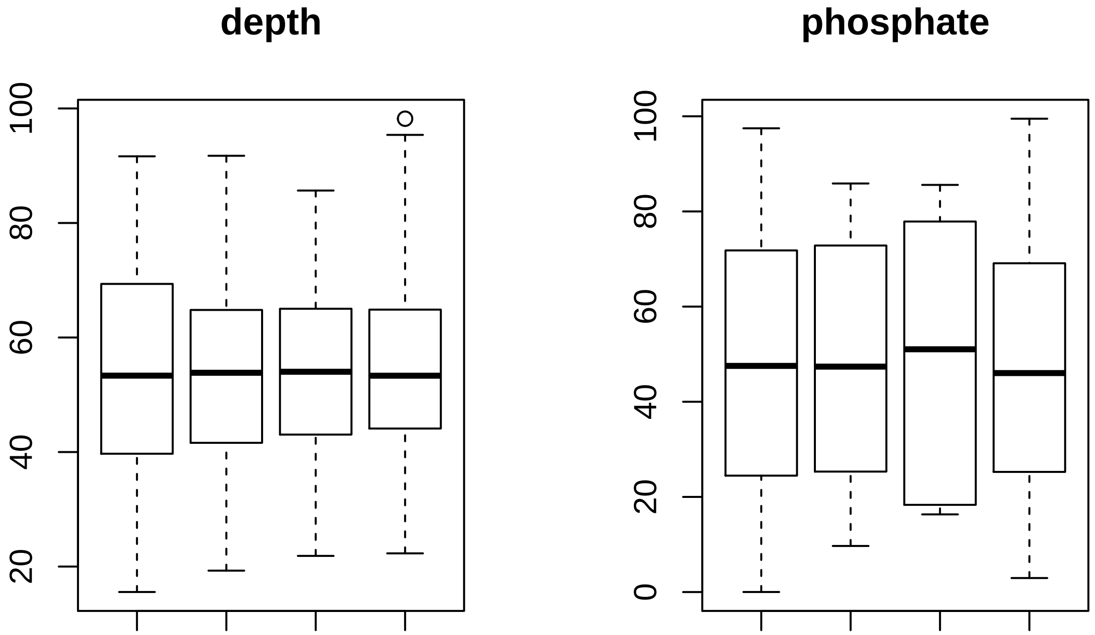
Summary statistics on each column are given below.

```
summary(df)
```

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

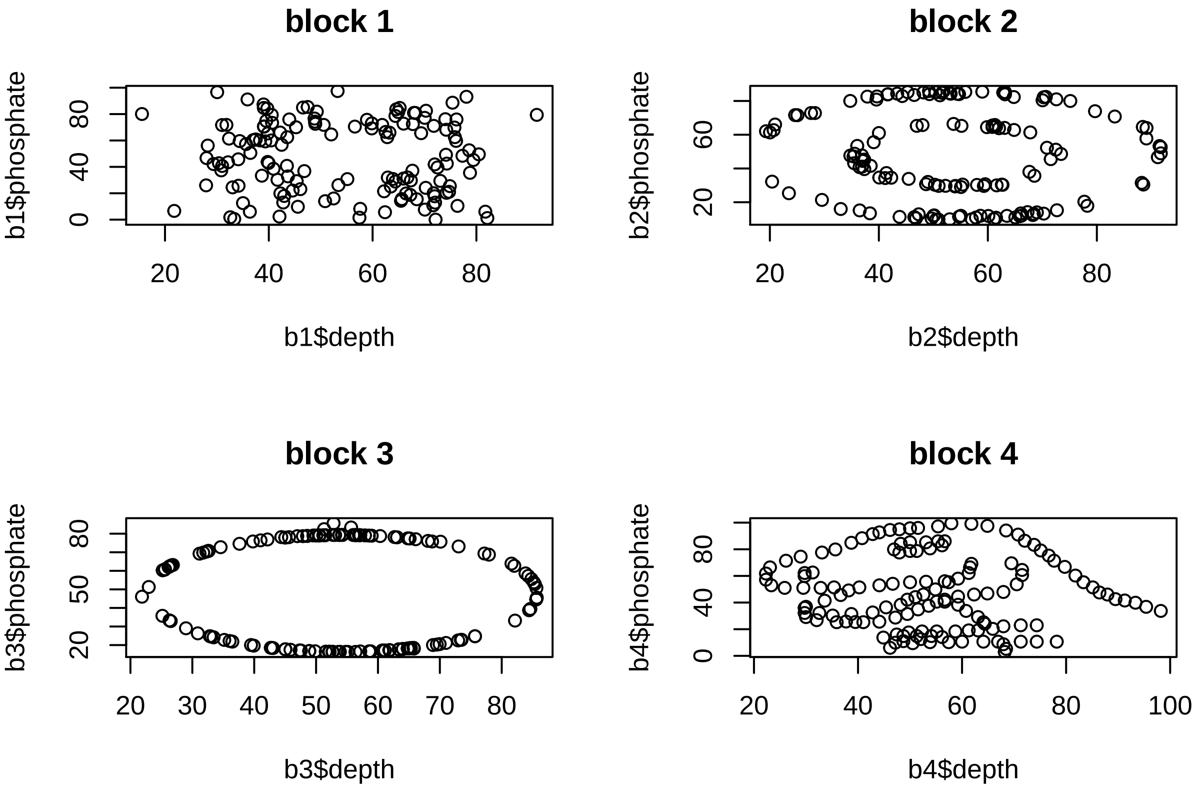
One may perhaps be interested in whether one's choice of block effects the other two variables. Below are boxplots of the other two factors. For depth it doesn't appear that block choice has a significant effect. For phosphate, however, it appears that there is at least some effect present in block 3.

```
par(mfrow = c(1, 2))
boxplot(b1$depth, b2$depth, b3$depth, b4$depth, main = "depth")
boxplot(b1$phosphate, b2$phosphate, b3$phosphate, b4$phosphate, main = "phosphate")
```



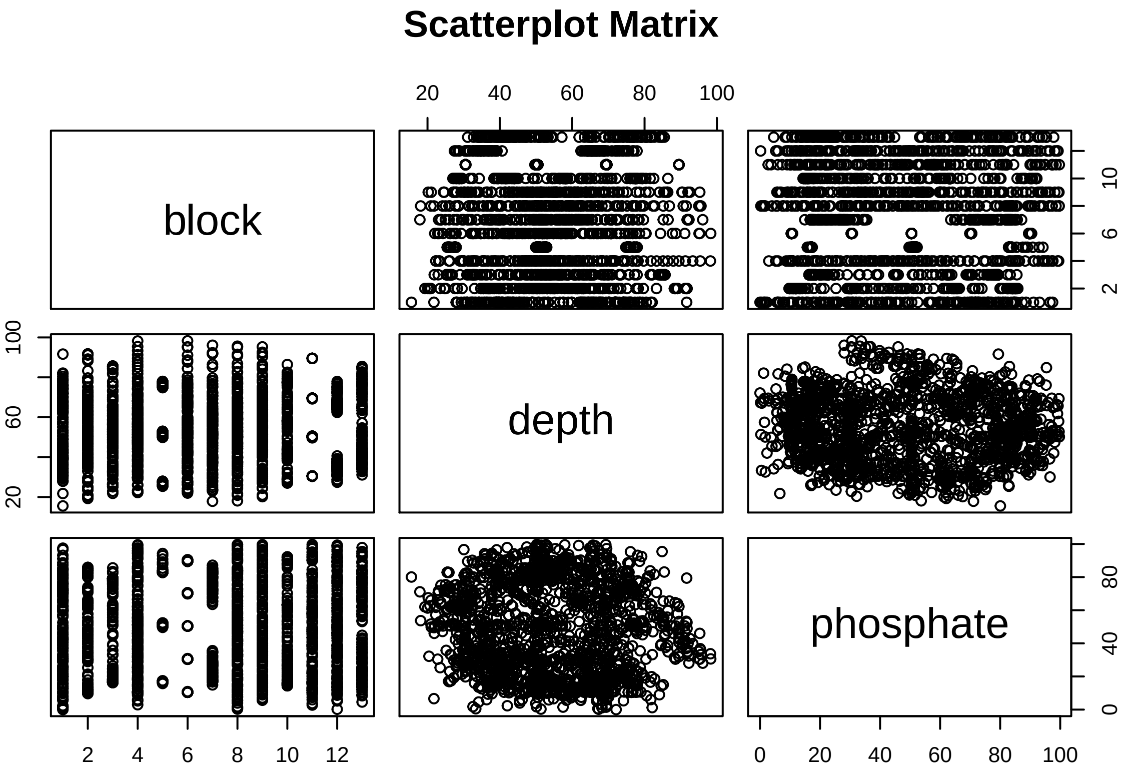
One may also be interested in how depth and phosphate relate across the different blocks. Plots of these relations are given below. One can see that choice of block significantly affects the relationship between depth and phosphate in quite unusual ways.

```
par(mfrow = c(2, 2))
plot(b1$depth, b1$phosphate, main = "block 1")
plot(b2$depth, b2$phosphate, main = "block 2")
plot(b3$depth, b3$phosphate, main = "block 3")
plot(b4$depth, b4$phosphate, main = "block 4")
```



We conclude this section with a scatterplot matrix. From the matrix we can see that the variables present do not appear to show any significant correlation across all the blocks, and for practical purposes may hence be regarded as independent (though evidently not conditionally independent given a fixed block).

```
pairs(~block + depth + phosphate, data = df, main = "Scatterplot Matrix")
```

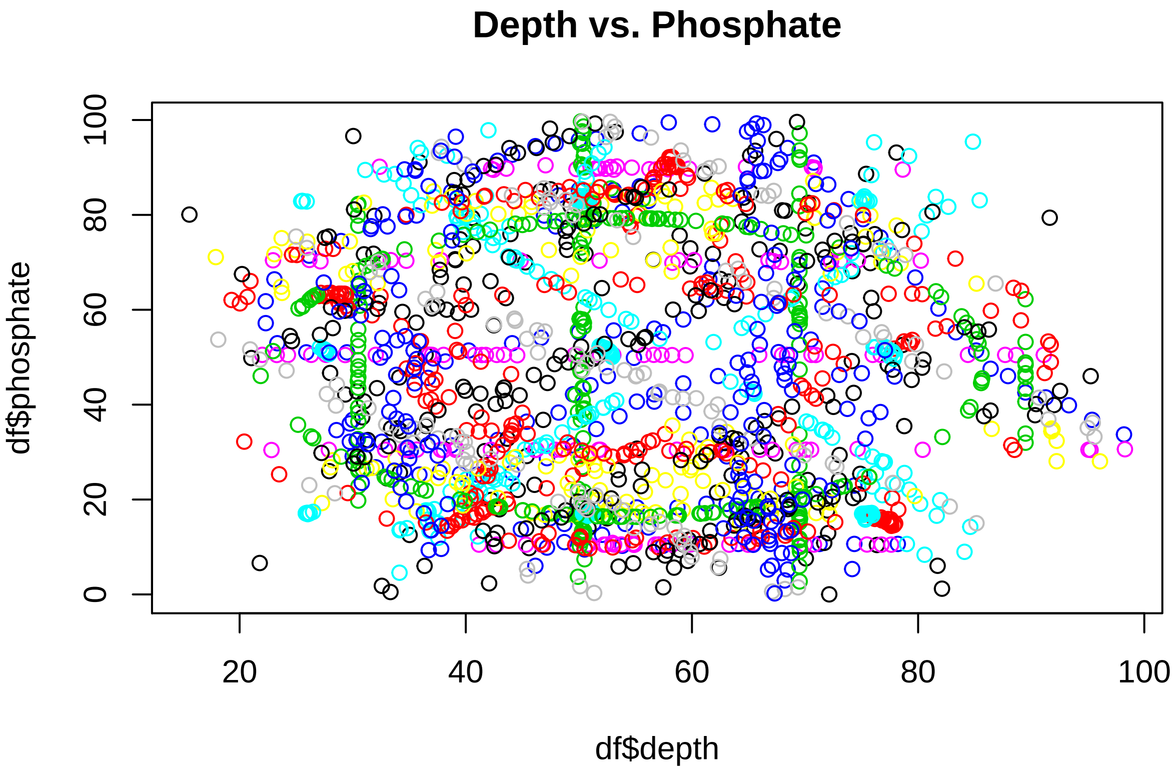


Problem 5

To me the most illuminating figure was the one that showed how the relationship between depth and phosphate varied quite substantially across the different blocks. I assume this is due to physical reasons, but I don't really know. It is weird though compared to the other plots, which seemed to be more straight forward.

Just for weirdness, here is a plot of depth vs. phosphate across all blocks, with each block color-coded.

```
plot(df$depth, df$phosphate, col = df$block, main = "Depth vs. Phosphate")
```


Depth vs. Phosphate
df$phosphate
df$depth