Homework 4

Shreya Chandrasekharan

```
## Loading required package: rJava

## Loading required package: xlsxjars

## Loading required package: boot

##
## Attaching package: 'psych'

## The following object is masked from 'package:boot':
##
##     logit

##
## Attaching package: 'ggplot2'

## The following objects are masked from 'package:psych':
##
##     %+%, alpha

##
## Attaching package: 'DescTools'

## The following objects are masked from 'package:psych':
##
##     ICC, SD
```

Problem 3

Focus of EDA stage of analysis

Exploratory Data Analysis (EDA) helps the researcher get a "quick and dirty" idea of the data. Roger Peng likens the concept of doing EDA to editing a movie. According to him, EDA helps the researcher guage the data and get a better idea of how to move forward with the data before starting any other form of analysis. This way, the researcher would know if it is wise to pursue the analysis at all if the data don't provide any worthwhile evience. EDA is not about the finer details of presentation, or even the final product. It marks the beginning of any type of statistical analysis that the researcher ultimately wants to do.

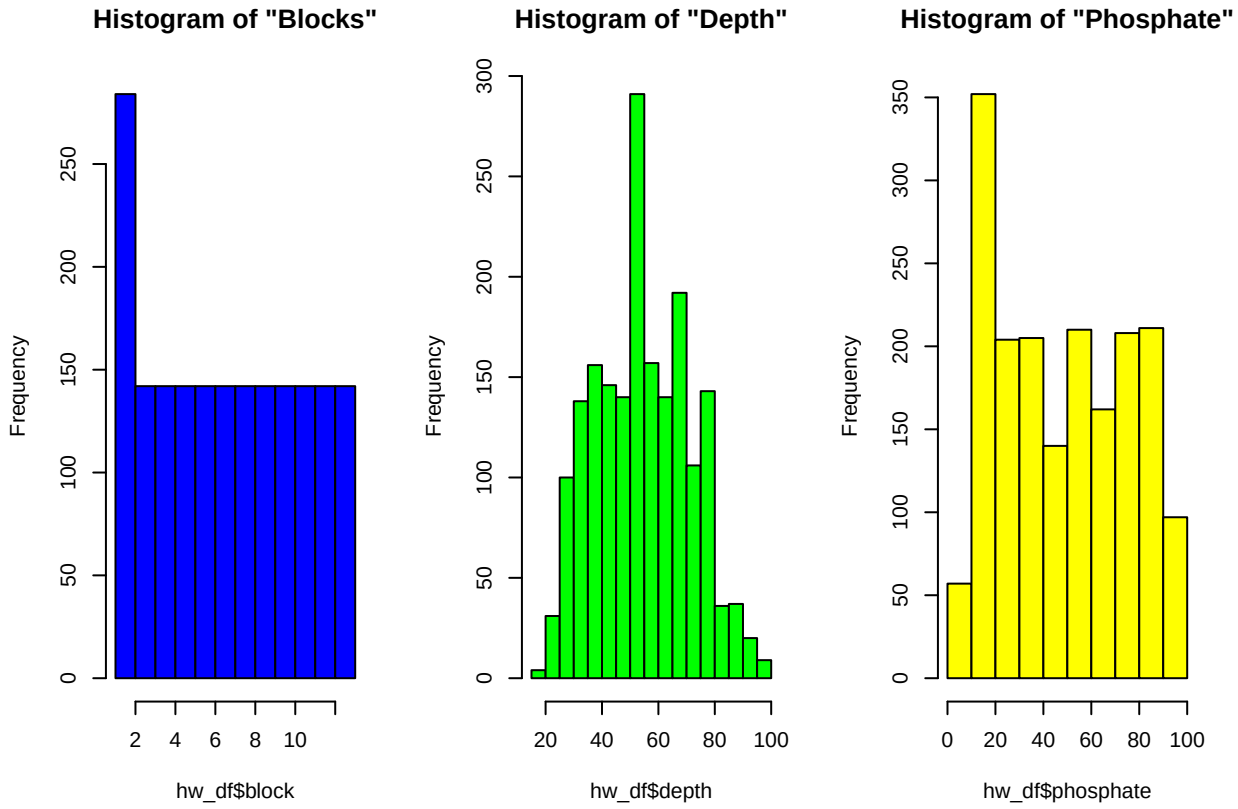
**Problem 4##

```
#This step isn't needed, but just showing that we can read the file this way as well

prob4_data1 <- read.xlsx("C:/Users/Shreya/Downloads/HW4_data.xlsx", sheetIndex = 1)
prob4_data2 <- read.xlsx("C:/Users/Shreya/Downloads/HW4_data.xlsx", sheetIndex = 2)

#Combining both sheets

hw_data <- "C:/Users/Shreya/Downloads/HW4_data.xlsx"
sheets_data <- excel_sheets(hw_data)
hw_df <- map_df(sheets_data, ~ read_excel(hw_data, sheet = .x))
```

```r
#Getting Summary
summary_hw_df <- summary(hw_df)
summary_hw_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
```

```r
#Exploratory Analysis

#First, we look at histograms of each factor type
par(mfrow=c(1,3))
hist(hw_df$block, breaks = 12, main = 'Histogram of "Blocks"', col = "blue")
hist(hw_df$depth, breaks = 12, main = 'Histogram of "Depth"', col = "green")
hist(hw_df$phosphate, breaks = 12, main = 'Histogram of "Phosphate"', col = "yellow")
```




```r
#Correlations

cor(hw_df$depth, hw_df$phosphate, method = "pearson")
```

```
## [1] -0.06601891
```

```r
cor(hw_df$block, hw_df$phosphate, method = "pearson")
```

```
## [1] 3.202565e-05
```

```
cor(hw_df$depth, hw_df$block, method = "pearson")
```

```
## [1] -5.620472e-06
```

```
#Some random analysis (Copied off the internet)

hw_fact <- factanal(covmat = cor(hw_df, use = "complete.obs"), factors = 1, rotation = "varimax")
corMatrix <- cor( hw_df, use="complete.obs" )
hw_fact <- fa( r=corMatrix, factors=1 )
print( hw_fact$loadings)
```

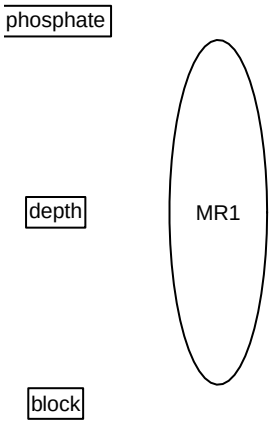
```
##
## Loadings:
##           MR1
## block
## depth     -0.257
## phosphate  0.257
##
##                  MR1
## SS loadings    0.132
## Proportion Var 0.044
```

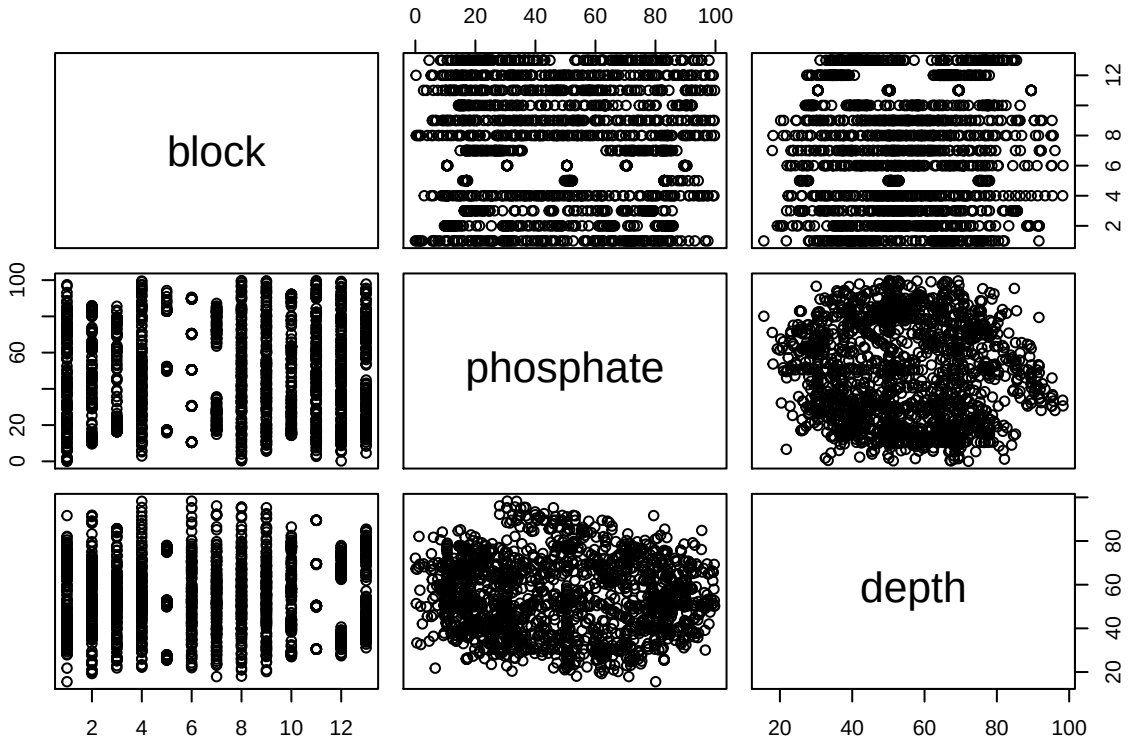
```
print( hw_fact$residual)
```

```
##                  block         depth     phosphate
## block     1.000000e+00  1.320259e-05  1.320258e-05
## depth     1.320259e-05  9.339811e-01 -1.915135e-15
## phosphate 1.320258e-05 -1.915135e-15  9.339811e-01
```

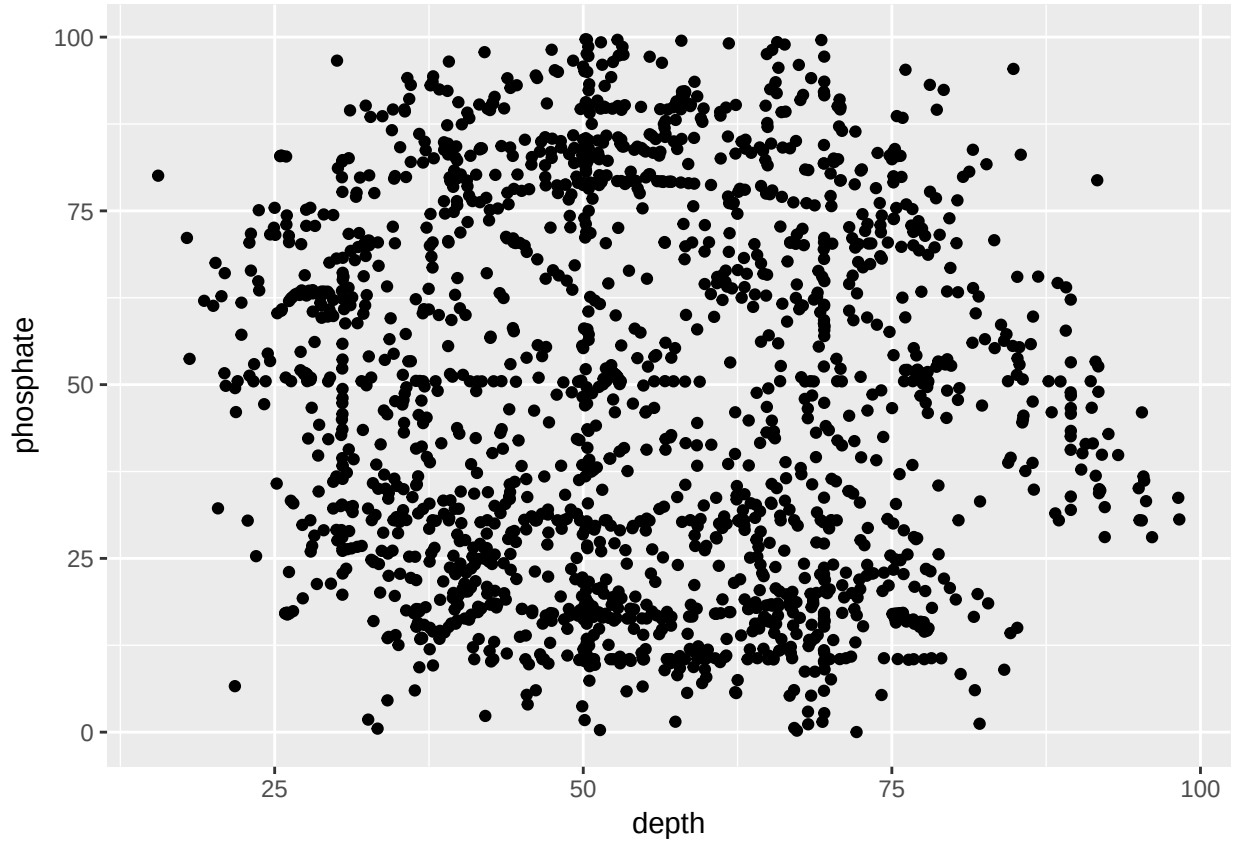
```
fa.diagram(hw_fact)
#We can visualize the factors by calling the function fa.diagram(hw_fact). The square boxes are the obs

#Correlation Plots

pairs(~ block + phosphate + depth, data=hw_df)
```

Factor Analysis

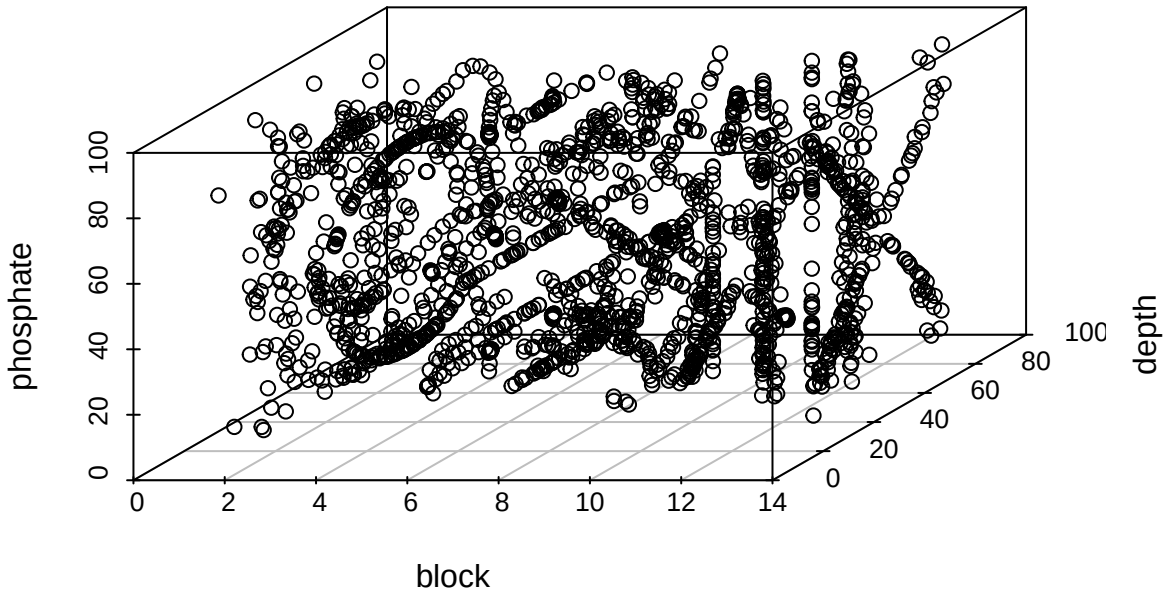





```
PlotCorr(corMatrix)


#ANOVA ('cause why not?)

fit <- aov(phosphate ~ depth + block, data=hw_df)

#Not the best use of ggplot2

plot1 <- qplot(block, depth, data=hw_df)
plot1
plot2 <- qplot(block, phosphate, data=hw_df)
plot2
plot3 <- qplot(depth, phosphate, data=hw_df)
plot3

# Oooh! 3D!

par(mfrow=c(1,1))
```




```
with(data = hw_df,
     scatterplot3d(x = block,
                   y = depth,
                   z = phosphate
     )
)
```



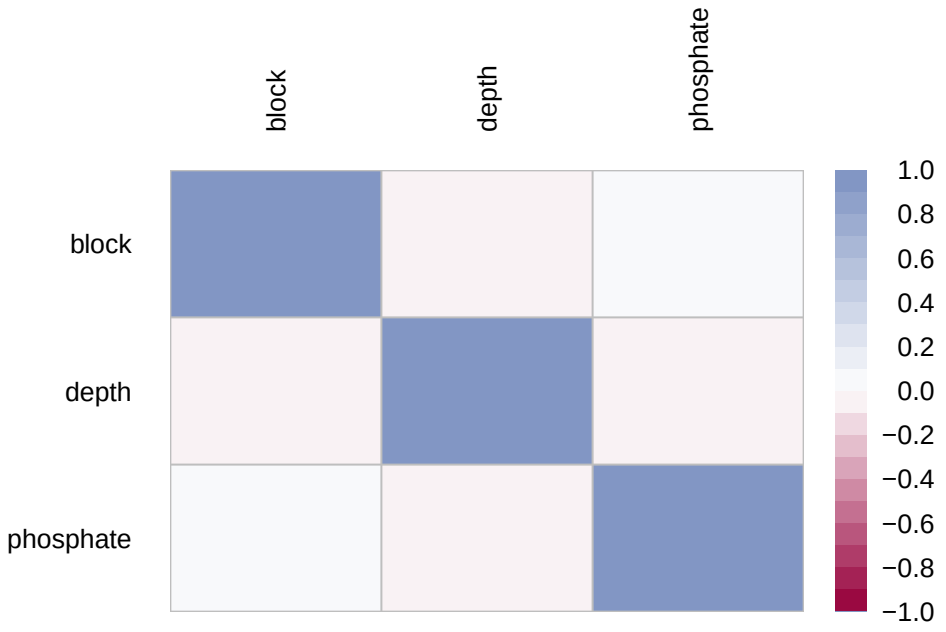
***Lesson*:**

When we know what the variables/factors are, our way of interpreting the data completely changes. For example, here, I focused on the (possible) cause and effect relationships between the three given variables, whereas, I didn't even consider checking for correlation in Problem 6 from HW3.

Problem 5

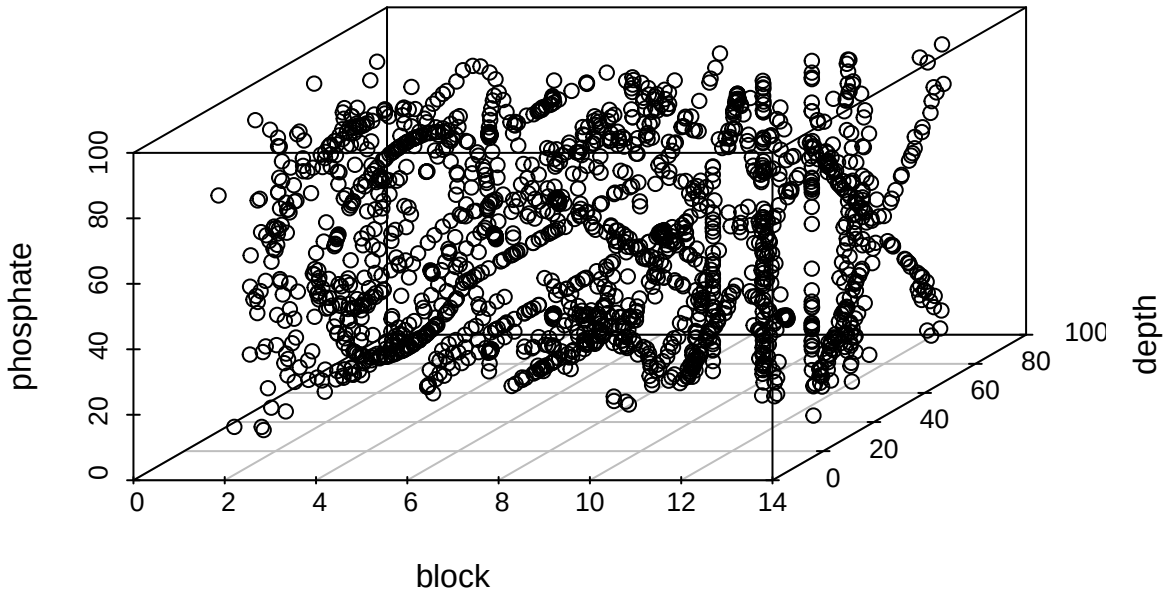
Single, most illuminating figure:

```
par(mfrow=c(1,1))
PlotCorr(corMatrix)
```



Shreya/2017−09−27

```
#I know that the question says "single", still...

with(data = hw_df,
     scatterplot3d(x = block,
                   y = depth,
                   z = phosphate
     )
)
```



My Learning

Exploratory Data Analysis completely changes perspectives with which I view data. Also, knowing the names of variables/factors involved (when using secondary data) has a major impact on how we interpret the data.