STAT 5014 Homework 4

Chaoran Wang

2017-09-26

Fineshed Problem 1 to 2

Problem 3

According to Roger Peng,the EDA stage focuses on indentifying the relationships between variables, checking hypothesis, checking if there is any problems of data, and identifying if more data is needed.

Problem 4

All codes used in this problem are in Appendix. I first read the two sheets in R and combine them.

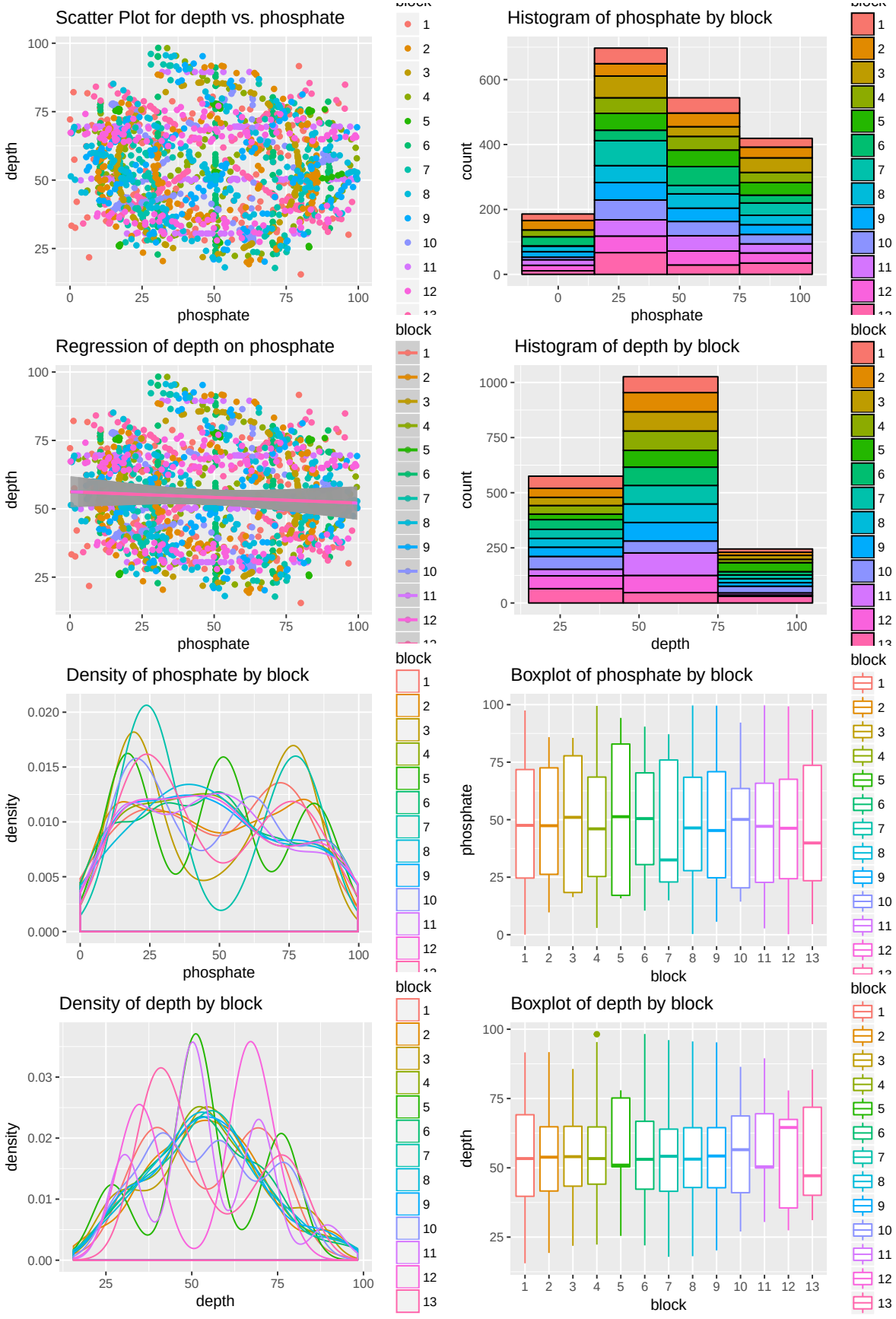
I figure that the "block" column should be factors in this dataset, so I change it to factor from numeric. I create two summary tables below. The second one summarize the data based on each block (factor) which also answer the second question.

Table 1: Problem 4 Data Summary not by block

block	depth	phosphate
1 :142	Min. :15.56	Min. : 0.01512
2 :142	1st Qu.:41.07	1st Qu.:22.56107
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	NA	NA

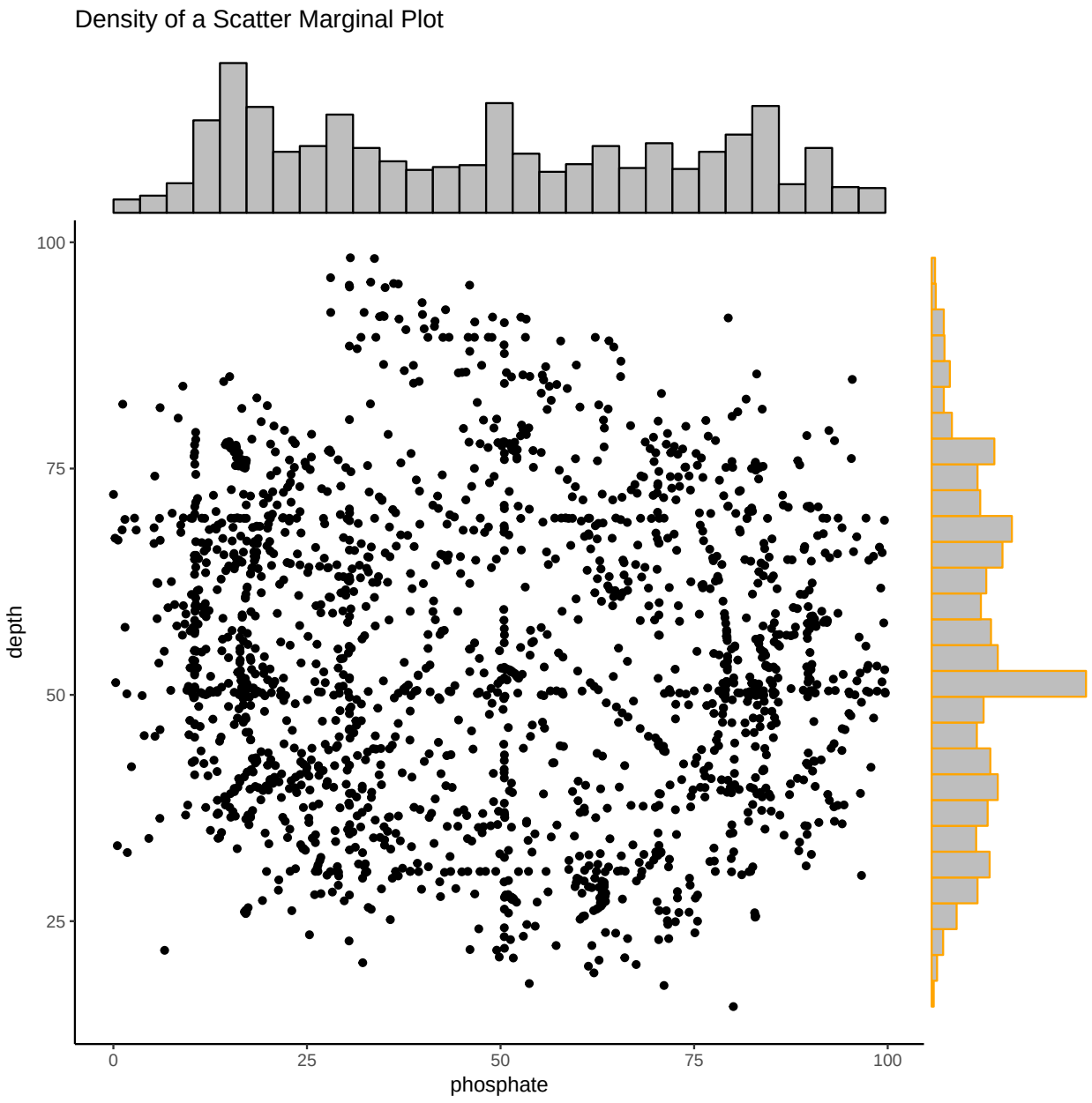
Table 2: Problem 4 Data Summary by block

block	depth.mean	phosphate.mean	depth.sd	phosphate.sd
1	54.26610	47.83472	16.76983	26.93974
2	54.26873	47.83082	16.76924	26.93573
3	54.26732	47.83772	16.76001	26.93004
4	54.26327	47.83225	16.76514	26.93540
5	54.26030	47.83983	16.76774	26.93019
6	54.26144	47.83025	16.76590	26.93988
7	54.26881	47.83545	16.76670	26.94000
8	54.26785	47.83590	16.76676	26.93610
9	54.26588	47.83150	16.76885	26.93861
10	54.26734	47.83955	16.76896	26.93027
11	54.26993	47.83699	16.76996	26.93768
12	54.26692	47.83160	16.77000	26.93790
13	54.26015	47.83972	16.76996	26.93000


Scatter Plot for depth vs. phosphate
depth
phosphate
block
Histogram of phosphate by block
count
phosphate
Regression of depth on phosphate
Histogram of depth by block
depth
Density of phosphate by block
density
Boxplot of phosphate by block
phosphate
Density of depth by block
Boxplot of depth by block

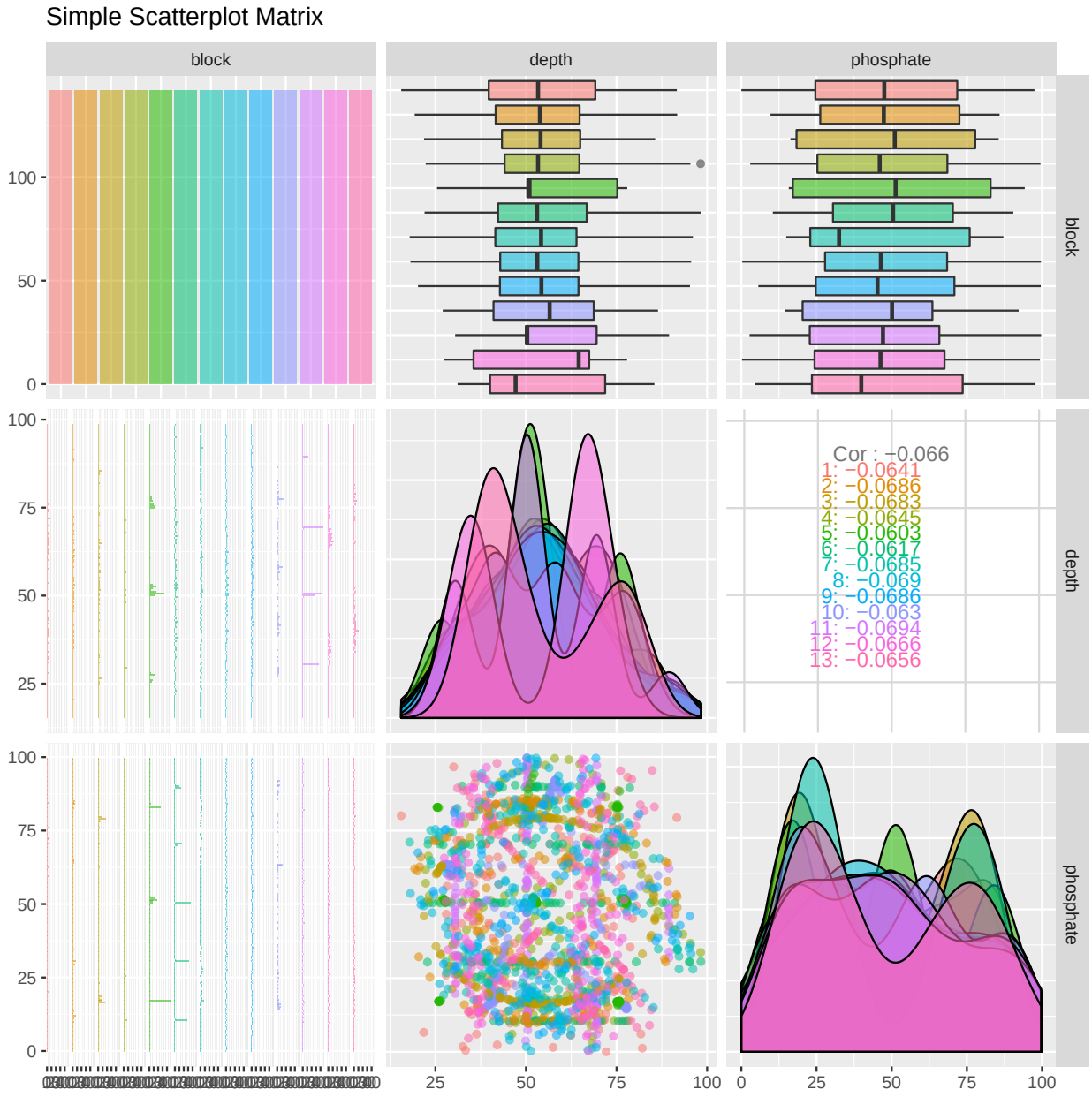
According to Cookbook for R, I creat the multiplot function and use the function to create a multipanel plot for the factor exploration. Based on the plots shown above, we could see there is definitely no obvious linear relationship among variables. 'block' looks not like a significant factor because of the first messy scatter plot. For all blocks, most phosphates lie on (15, 42) while most depth lie on (42, 75). While the 7th block has the highest density for phosphate, the 5th block has the highest density for depth. Besides, the 4th block seems have a mild outlier.

The Density of a Scatter Marginal Plot which comes from Dr. Settlage codes is shown below as well.

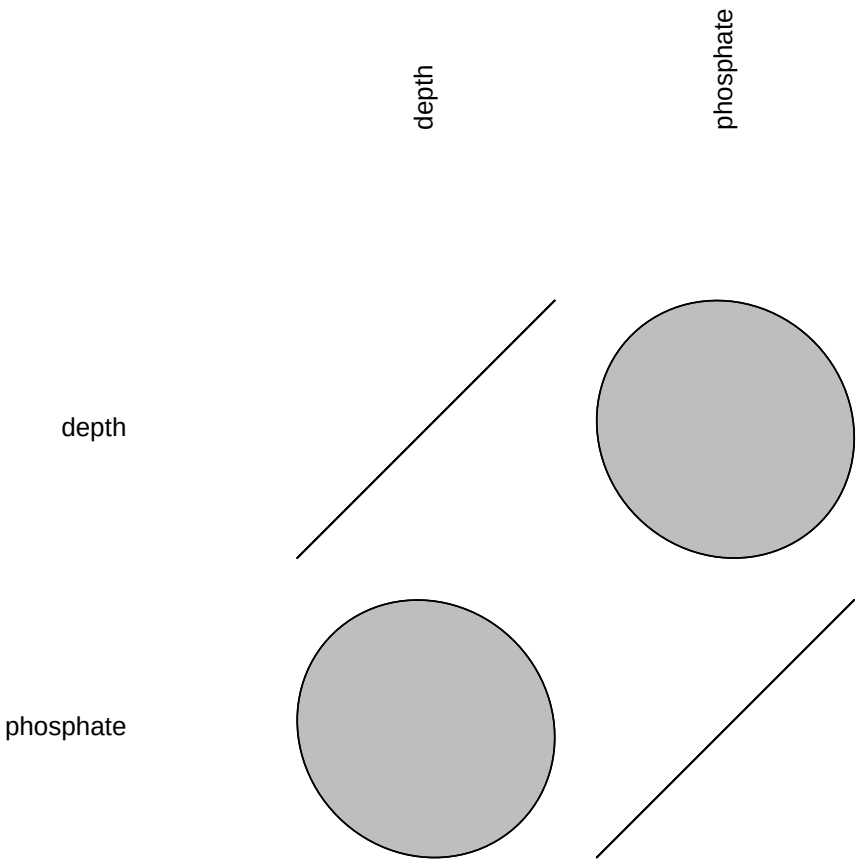


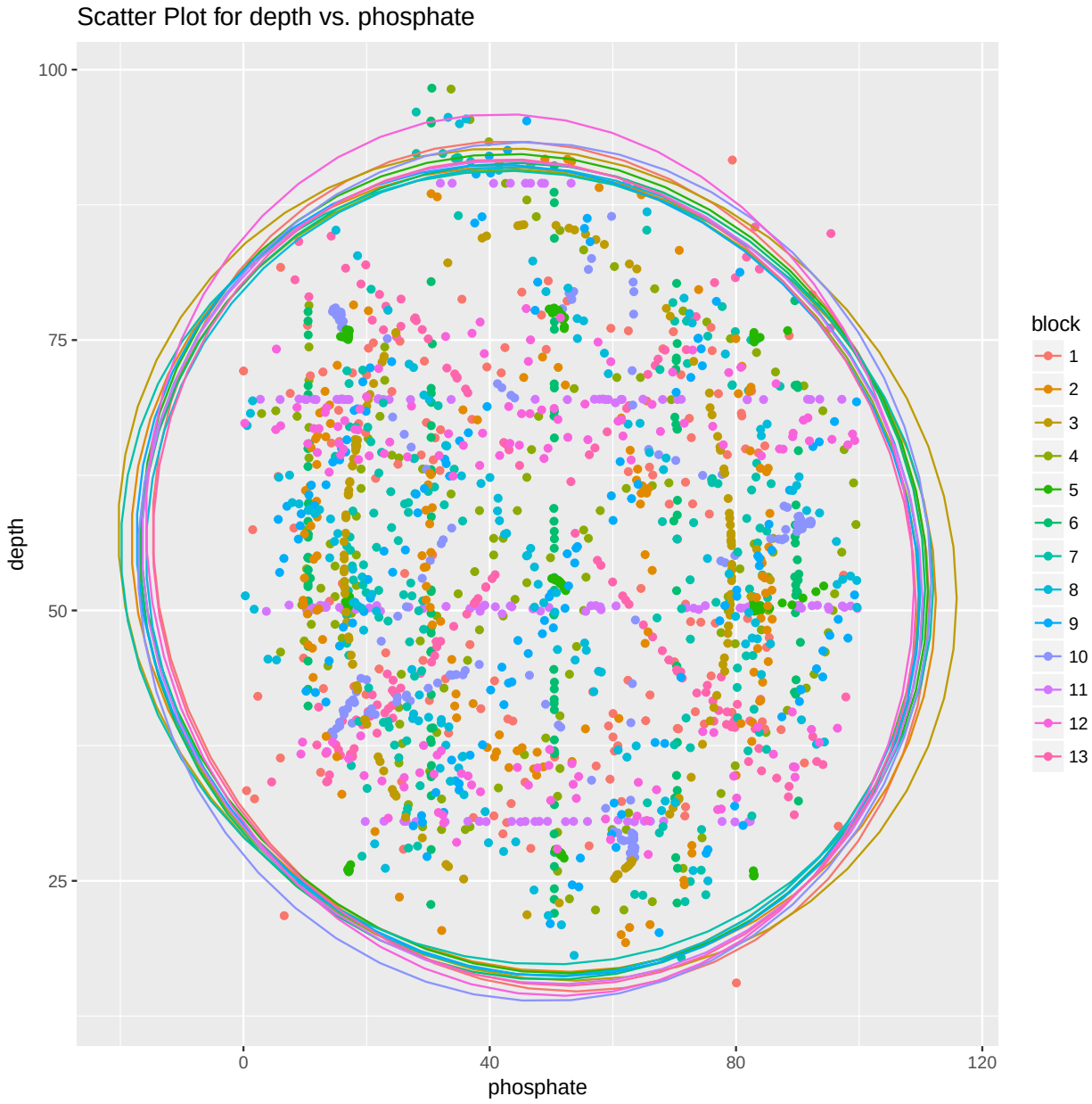
Next, I create correlation plots. Instead of using 'pairs' of basic R, I am using 'ggpairs' in GGally package.

For correlation plots, I am using both 'plotcorr' and 'stat_ellipse'. They look similar to each other.


Simple Scatterplot Matrix
block
depth
phosphate
Cor : −0.066
1: −0.0641
2: −0.0686
3: −0.0683
4: −0.0645
5: −0.0603
6: −0.0617
7: −0.0685
8: −0.069
9: −0.0686
10: −0.063
11: −0.0694
12: −0.0666
13: −0.0656

Correlation Plot





Although the dataset is same from the one in last homework, we are focusing more on factor analysis this time. Considering 'block' as a categprical variable seems to be more reasonable than consider it to be numeric. Hence, the summary statistics which based on 'block' and the plots which treat 'block' as a factor seem to be reasonable.

Problem 6

This exercise shows useful tools, statistics and plots, for exploratory data analysis. And ggplot seems create more beautiful plots than basic R.

Appendix 1: R code

```
############################
#Problem4_data
#get data
############################
prob4_data1 <- read.xlsx("HW4_data.xlsx", sheetIndex = 1)
prob4_data2 <- read.xlsx("HW4_data.xlsx", sheetIndex = 2)
prob4_data <- rbind(prob4_data1, prob4_data2)
############################
############################
#Problem4_analysis
############################
prob4_data[,'block'] <- as.factor(prob4_data[,'block'])
knitr::kable(summary(prob4_data), format = "latex", caption="Problem 4 Data Summary
             not by block", booktabs = T) %>%
  kable_styling(latex_options = c("striped", "hold_position"))
knitr::kable(summaryBy(depth+phosphate ~ block, prob4_data, FUN=c(mean,sd)),
             format = "latex", caption="Problem 4 Data Summary by block",
             booktabs = T) %>%
  kable_styling(latex_options = c("striped", "hold_position"))
############################
# Multiple plot function
# Credit to Cookbook for R
#
# ggplot objects can be passed in ..., or to plotlist (as a list of ggplot objects)
# - cols:   Number of columns in layout
# - layout: A matrix specifying the layout. If present, 'cols' is ignored.
#
# If the layout is something like matrix(c(1,2,3,3), nrow=2, byrow=TRUE),
# then plot 1 will go in the upper left, 2 will go in the upper right, and
# 3 will go all the way across the bottom.
#
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))
    }
  }
}
############################
#Problem4_plot
############################
p1 <- ggplot(prob4_data, aes(x=phosphate, y=depth, colour=block, group=block)) +
   geom_point() +
   ggtitle("Scatter Plot for depth vs. phosphate")
p2 <- ggplot(prob4_data,  aes(x=phosphate, y=depth, colour=block, group=block)) +
   geom_point() + stat_smooth(method = "lm") +
   ggtitle("Regression of depth on phosphate")
p3 <- ggplot(prob4_data, aes(x=phosphate, colour=block)) +
   geom_density() +
   ggtitle("Density of phosphate by block")
p4 <- ggplot(prob4_data, aes(x=depth, colour=block)) +
   geom_density() +
   ggtitle("Density of depth by block")
p5 <- ggplot(prob4_data, aes(x=phosphate, fill=block)) +
   geom_histogram(colour="black", binwidth=30) +
   ggtitle("Histogram of phosphate by block")
p6 <- ggplot(prob4_data, aes(x=depth, fill=block)) +
   geom_histogram(colour="black", binwidth=30) +
   ggtitle("Histogram of depth by block")
p7 <- ggplot(prob4_data, aes(x=block, y=phosphate)) +
   geom_boxplot(stat="boxplot", aes(colour = block)) +
   ggtitle("Boxplot of phosphate by block")
p8 <- ggplot(prob4_data, aes(x=block, y=depth)) +
   geom_boxplot(stat="boxplot", aes(colour = block)) +
   ggtitle("Boxplot of depth by block")
multiplot(p1, p2, p3, p4, p5, p6, p7, p8, cols=2)
############################
# Basic Scatterplot Matrix
ggpairs(prob4_data, aes(colour = block, alpha = 0.4), title = "Simple Scatterplot Matrix",
        lower = list(combo = wrap("facethist", binwidth = 0.5)))
correlation <- round(cor(prob4_data[, 2:3]),3)
plotcorr(correlation, main = "Correlation Plot")
ggplot(prob4_data, aes(x=phosphate, y=depth, colour=block)) + geom_point() +
   stat_ellipse() + ggtitle("Scatter Plot for depth vs. phosphate")
```