HW4_House_Matt

Matthew House

9/26/2017

Problem 3:

In the lecture, there were a few links to Exploratory Data Analysis (EDA) materials. According to Roger Peng, what is the focus of the EDA stage of an analysis? Hint: this is summarized in the free sample portion of his online book.

- According to Roger Peng, the focus of the EDA stage of an analysis is to identify what is important and what can be cut out. He relates this to a film editing room, where the director and others can pair different scenes to see which group tells the most complete story relative to the script. He notes that this stage is important to keep the the project moving forward.

Problem 4:

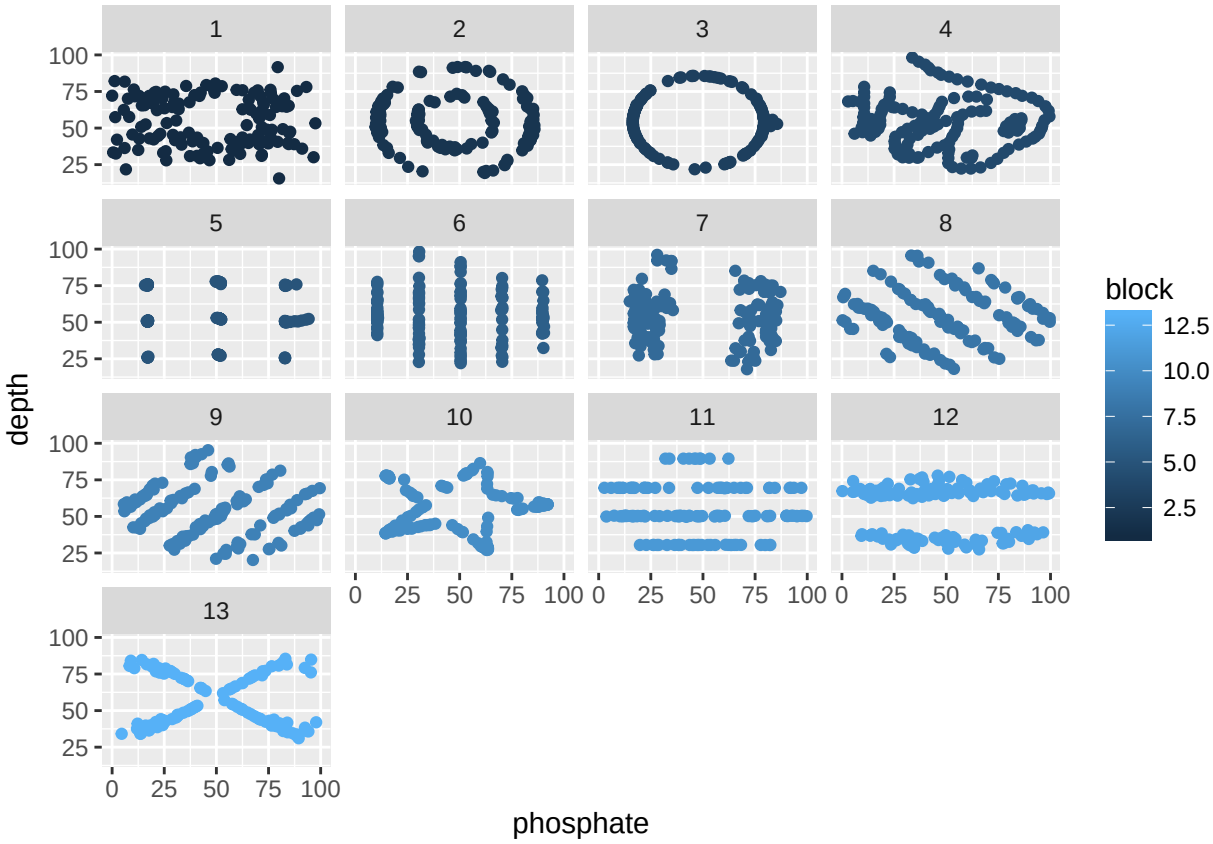
```
    prob4_data1 <- read_excel("HW4_data.xlsx", sheet = 1, col_names = TRUE)
    prob4_data2 <- read_excel("HW4_data.xlsx", sheet = 2, col_names = TRUE)

mydata <- rbind(prob4_data1, prob4_data2)

summydata <- mydata %>%
        group_by(block) %>%
        summarise_all(funs(mean,sd)) %>%

        arrange()

knitr::kable(summydata, caption="Summary Statistics")
```

Table 1: Summary Statistics

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

```
    prob4_data1 <- read_excel("HW4_data.xlsx", sheet = 1, col_names = TRUE)
    prob4_data2 <- read_excel("HW4_data.xlsx", sheet = 2, col_names = TRUE)
```

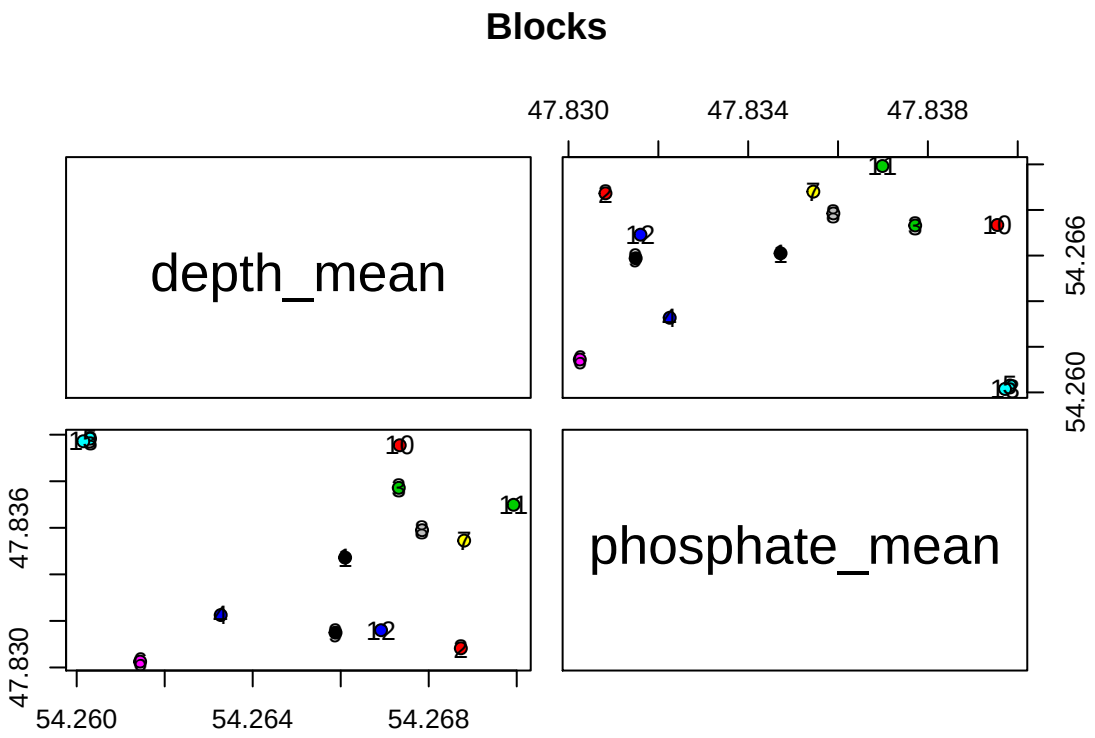
```
mydata <- rbind(prob4_data1, prob4_data2)

grouped <-  group_by(mydata,block)

ggplot(grouped, aes(x = phosphate, y = depth, colour = block)) +
  geom_point() +
  facet_wrap( ~ block)
```



Factors that appear to be present are within groups and plotting depth against phosphate produces shapes for some of the groups and groups of lines for others.

```
   prob4_data1 <- read_excel("HW4_data.xlsx", sheet = 1, col_names = TRUE)
    prob4_data2 <- read_excel("HW4_data.xlsx", sheet = 2, col_names = TRUE)

mydata <- rbind(prob4_data1, prob4_data2)

summydata <- mydata %>%
        group_by(block) %>%
        summarise_all(funs(mean,sd, median, cor(phosphate,depth,
                  use = "pairwise.complete.obs", method = "pearson"))) %>%
        transmute(block, depth_mean, phosphate_mean,
                  depth_sd, phosphate_sd,
                  depth_median, phosphate_median,
                  Correlation = depth_cor) %>%
        arrange()

pairs(summydata[2:3], main = "Blocks",
```

```
      pch = 21, bg = summydata$block[unclass(summydata$block)],
      panel=function(x, y, ...) { points(x, y, ...);
      text(x, y,) })
```




```
knitr::kable(summydata, caption="Summary Statistics")
```

Table 2: Summary Statistics

block	depth_mean	phosphate_mean	depth_sd	phosphate_sd	depth_median	phosphate_median	Correlation
1	54.26610	47.83472	16.76983	26.93974	53.34030	47.53527	-0.0641284
2	54.26873	47.83082	16.76924	26.93573	53.84209	47.38294	-0.0685864
3	54.26732	47.83772	16.76001	26.93004	54.02321	51.02502	-0.0683434
4	54.26327	47.83225	16.76514	26.93540	53.33330	46.02560	-0.0644719
5	54.26030	47.83983	16.76774	26.93019	50.97677	51.29929	-0.0603414
6	54.26144	47.83025	16.76590	26.93988	53.06968	50.47353	-0.0617148
7	54.26881	47.83545	16.76670	26.94000	54.16869	32.49920	-0.0685042
8	54.26785	47.83590	16.76676	26.93610	53.13516	46.40131	-0.0689797
9	54.26588	47.83150	16.76885	26.93861	54.26135	45.29224	-0.0686092
10	54.26734	47.83955	16.76896	26.93027	56.53473	50.11055	-0.0629611
11	54.26993	47.83699	16.76996	26.93768	50.36289	47.11362	-0.0694456
12	54.26692	47.83160	16.77000	26.93790	64.55023	46.27933	-0.0665752
13	54.26015	47.83972	16.76996	26.93000	47.13646	39.87621	-0.0655833

Using groups allowed me to see shapes. I don't know what I am missing for the pairs plots. I didn't see anything that jumped out at me. Perhaps the way that I am doing it is wrong.

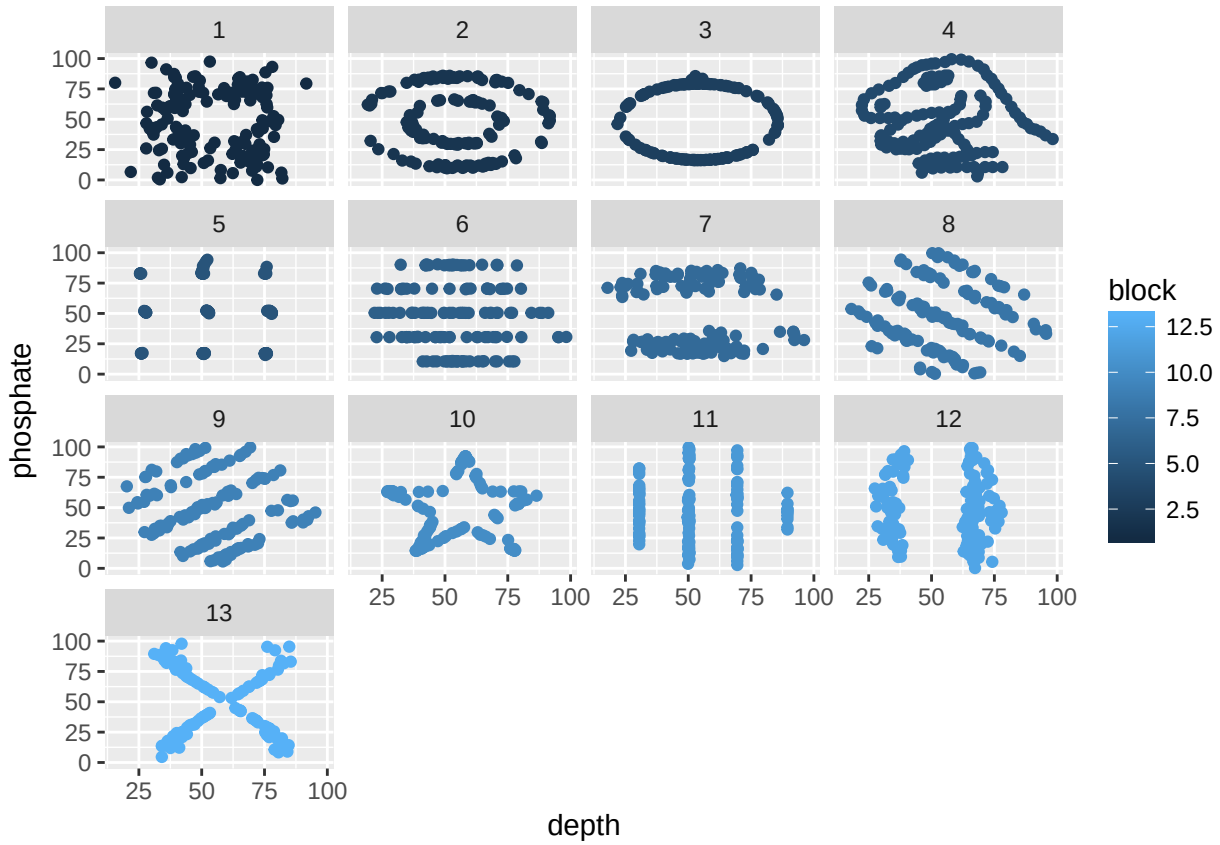
Problem 5

For problem 4, is there a single most illuminating figure that shows a key component of the data?? This is

the figure you should use as your first submission in next weeks contest. Save it as a pdf and make sure it is ready for pushing at the start of class. (Torg 1100, try to be a little early) What did this exercise show you?

Just seaparating the groups showed a distinct difference in what the plots looked like.

```
#Sys.setenv("plotly_username"="wynne")
Sys.setenv("plotly_api_key"="X8YHpLYCaQrYJ1IvTqKk")

#install.packages("webshot")
#webshot::install_phantomjs()

    prob4_data1 <- read_excel("HW4_data.xlsx", sheet = 1, col_names = TRUE)
    prob4_data2 <- read_excel("HW4_data.xlsx", sheet = 2, col_names = TRUE)

mydata <- rbind(prob4_data1, prob4_data2)

grouped <-  group_by(mydata,block)

ggplot(grouped, aes(x = depth, y = phosphate, colour = block)) +
  geom_point() +
  facet_wrap( ~ block)
```



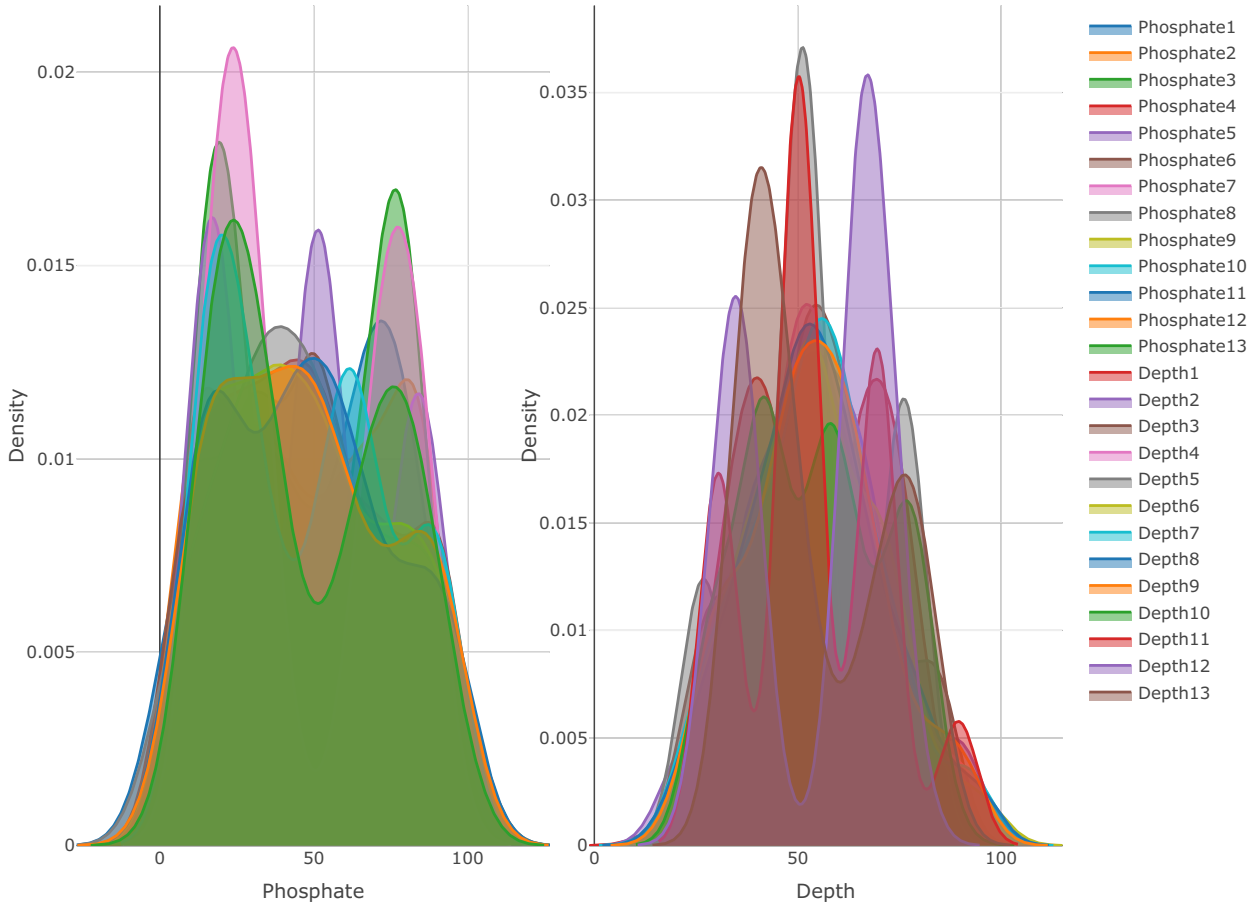

```
# Bob can you look at this? Line 126-255 produces the area fill plots and outputs them
# as two pdfs. One shares a y axis and one does not. Also you were going to look at this
# to see if we couldn't make it 3d and maybe turn 129 lines of code into a function or
# two to drastically shorten this monster.

block1 <- grouped[which(grouped$block == "1"),]
```

```r
density1 <- density(block1$phosphate)

block2 <- grouped[which(grouped$block == "2"),]
density2 <- density(block2$phosphate)

block3 <- grouped[which(grouped$block == "3"),]
density3 <- density(block3$phosphate)

block4 <- grouped[which(grouped$block == "4"),]
density4 <- density(block4$phosphate)

block5 <- grouped[which(grouped$block == "5"),]
density5 <- density(block5$phosphate)

block6 <- grouped[which(grouped$block == "6"),]
density6 <- density(block6$phosphate)

block7 <- grouped[which(grouped$block == "7"),]
density7 <- density(block7$phosphate)

block8 <- grouped[which(grouped$block == "8"),]
density8 <- density(block8$phosphate)

block9 <- grouped[which(grouped$block == "9"),]
density9 <- density(block9$phosphate)

block10 <- grouped[which(grouped$block == "10"),]
density10 <- density(block10$phosphate)

block11 <- grouped[which(grouped$block == "11"),]
density11 <- density(block11$phosphate)

block12 <- grouped[which(grouped$block == "12"),]
density12 <- density(block12$phosphate)

block13 <- grouped[which(grouped$block == "13"),]
density13 <- density(block13$phosphate)



p <- plot_ly(x = ~density1$x, y = ~density1$y, type = 'scatter', mode = 'lines', name = 'Phosphate1', fi

  add_trace(x = ~density2$x, y = ~density2$y, name = 'Phosphate2', fill = 'tozeroy') %>%

  add_trace(x = ~density3$x, y = ~density3$y, name = 'Phosphate3', fill = 'tozeroy') %>%

  add_trace(x = ~density4$x, y = ~density4$y, name = 'Phosphate4', fill = 'tozeroy') %>%

  add_trace(x = ~density5$x, y = ~density5$y, name = 'Phosphate5', fill = 'tozeroy') %>%

  add_trace(x = ~density6$x, y = ~density6$y, name = 'Phosphate6', fill = 'tozeroy') %>%

  add_trace(x = ~density7$x, y = ~density7$y, name = 'Phosphate7', fill = 'tozeroy') %>%
```

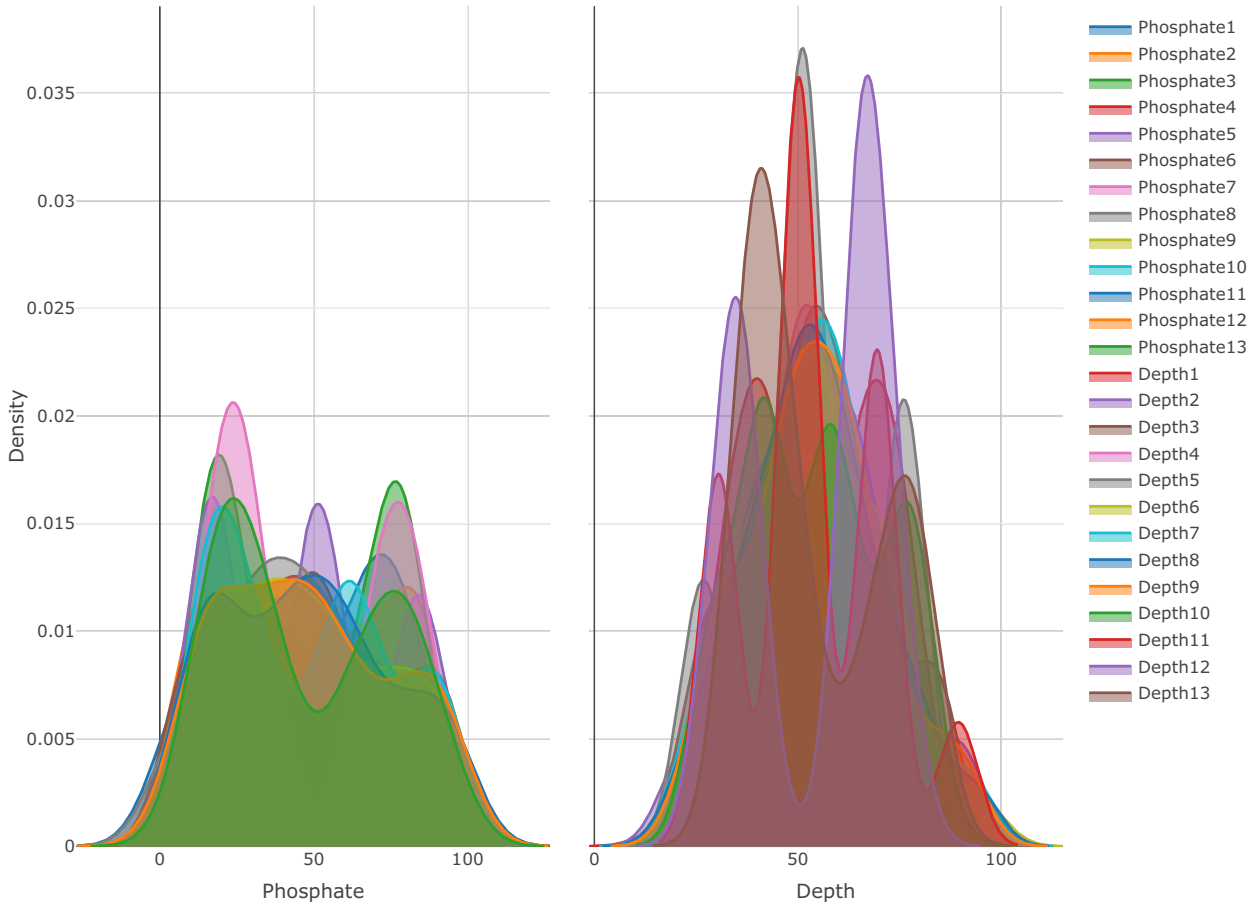
```
  add_trace(x = ~density8$x, y = ~density8$y, name = 'Phosphate8', fill = 'tozeroy') %>%

  add_trace(x = ~density9$x, y = ~density9$y, name = 'Phosphate9', fill = 'tozeroy') %>%

  add_trace(x = ~density10$x, y = ~density10$y, name = 'Phosphate10', fill = 'tozeroy') %>%

  add_trace(x = ~density11$x, y = ~density11$y, name = 'Phosphate11', fill = 'tozeroy') %>%

  add_trace(x = ~density12$x, y = ~density12$y, name = 'Phosphate12', fill = 'tozeroy') %>%

  add_trace(x = ~density13$x, y = ~density13$y, name = 'Phosphate13', fill = 'tozeroy') %>%

  layout(xaxis = list(title = 'Phosphate'),
         yaxis = list(title = 'Density'))


block1a <- grouped[which(grouped$block == "1"),]
density1a <- density(block1a$depth)
block2a <- grouped[which(grouped$block == "2"),]
density2a <- density(block2a$depth)
block3a <- grouped[which(grouped$block == "3"),]
density3a <- density(block3a$depth)
block4a <- grouped[which(grouped$block == "4"),]
density4a <- density(block4a$depth)
block5a <- grouped[which(grouped$block == "5"),]
density5a <- density(block5a$depth)
block6a <- grouped[which(grouped$block == "6"),]
density6a <- density(block6a$depth)
block7a <- grouped[which(grouped$block == "7"),]
density7a <- density(block7a$depth)
block8a <- grouped[which(grouped$block == "8"),]
density8a <- density(block8a$depth)
block9a <- grouped[which(grouped$block == "9"),]
density9a <- density(block9a$depth)
block10a <- grouped[which(grouped$block == "10"),]
density10a <- density(block10a$depth)
block11a <- grouped[which(grouped$block == "11"),]
density11a <- density(block11a$depth)
block12a <- grouped[which(grouped$block == "12"),]
density12a <- density(block12a$depth)
block13a <- grouped[which(grouped$block == "13"),]
density13a <- density(block13a$depth)

p2 <- plot_ly(x = ~density1a$x, y = ~density1a$y, type = 'scatter', mode = 'lines', name = 'Depth1', fi

  add_trace(x = ~density2a$x, y = ~density2a$y, name = 'Depth2', fill = 'tozeroy') %>%

  add_trace(x = ~density3a$x, y = ~density3a$y, name = 'Depth3', fill = 'tozeroy') %>%

  add_trace(x = ~density4a$x, y = ~density4a$y, name = 'Depth4', fill = 'tozeroy') %>%

  add_trace(x = ~density5a$x, y = ~density5a$y,  name = 'Depth5', fill = 'tozeroy') %>%
```

```
  add_trace(x = ~density6a$x, y = ~density6a$y,  name = 'Depth6', fill = 'tozeroy') %>%

  add_trace(x = ~density7a$x, y = ~density7a$y, name = 'Depth7', fill = 'tozeroy') %>%

  add_trace(x = ~density8a$x, y = ~density8a$y, name = 'Depth8', fill = 'tozeroy') %>%

  add_trace(x = ~density9a$x, y = ~density9a$y, name = 'Depth9', fill = 'tozeroy') %>%

  add_trace(x = ~density10a$x, y = ~density10a$y, name = 'Depth10', fill = 'tozeroy') %>%

  add_trace(x = ~density11a$x, y = ~density11a$y, name = 'Depth11', fill = 'tozeroy') %>%

  add_trace(x = ~density12a$x, y = ~density12a$y, name = 'Depth12', fill = 'tozeroy') %>%

  add_trace(x = ~density13a$x, y = ~density13a$y, name = 'Depth13', fill = 'tozeroy') %>%

  layout(xaxis = list(title = 'Depth'),
         yaxis = list(title = 'Density'))
```

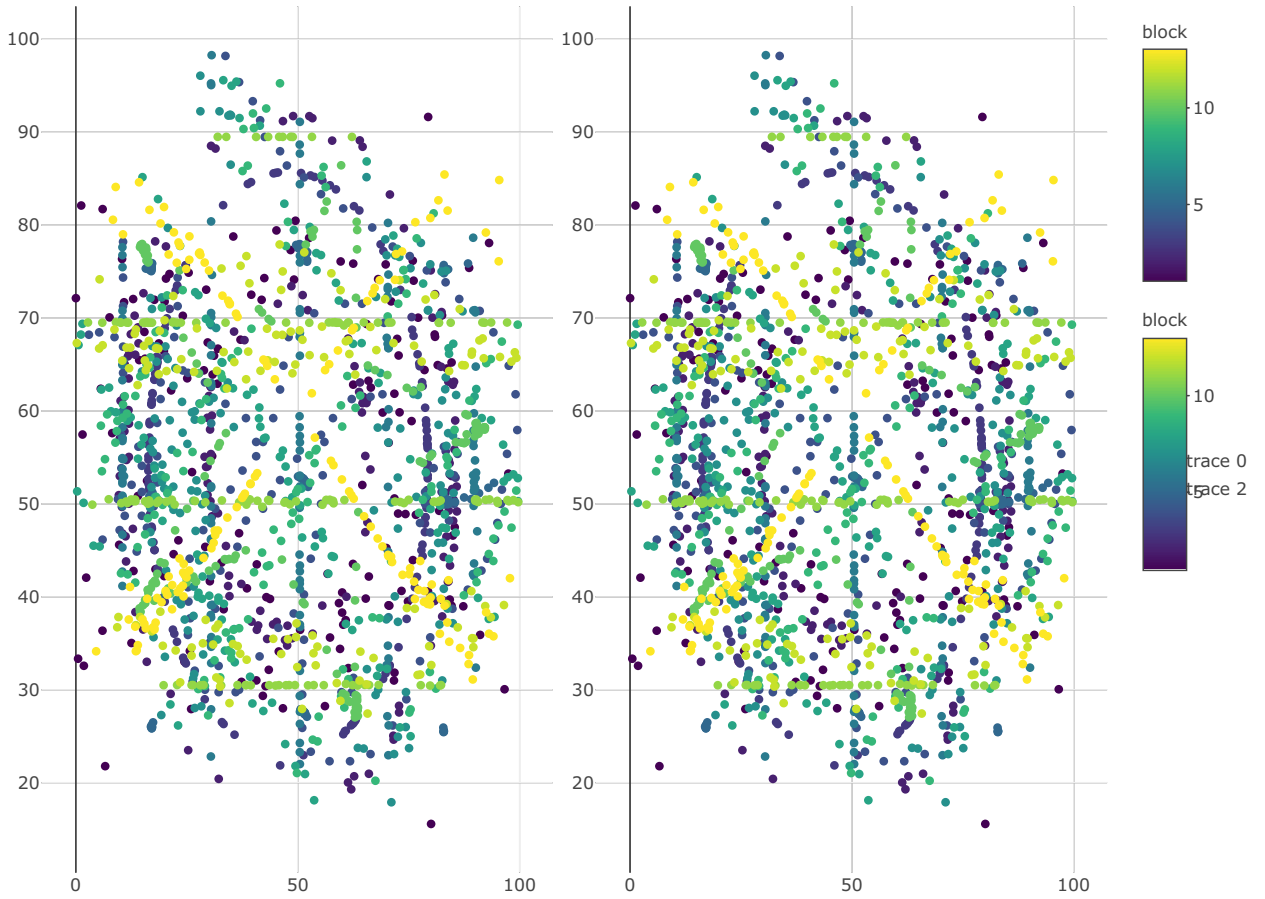
```
export(subplot(p, p2, titleX = TRUE) %>% layout(xaxis = list(title = "Phosphate"), xaxis2 = list(title =
```



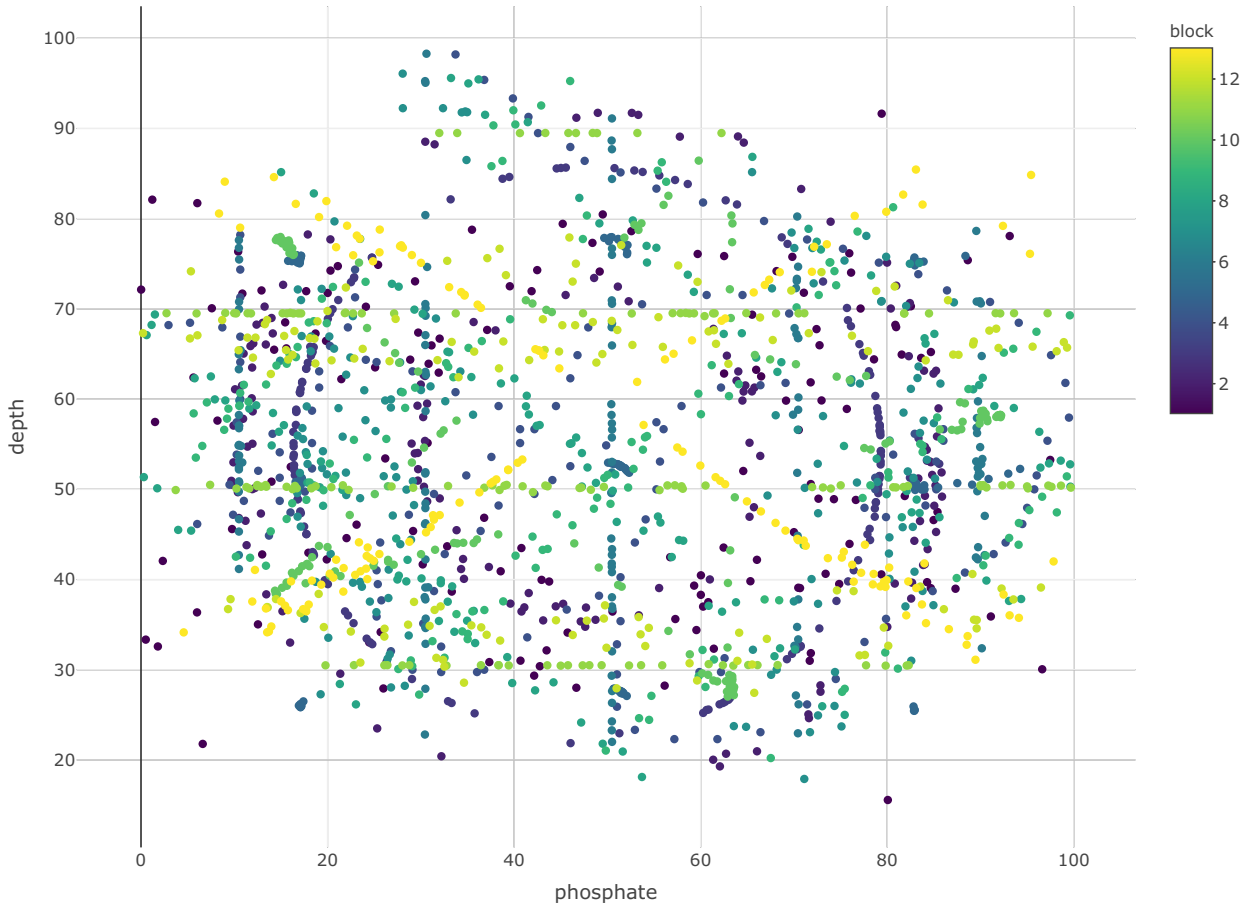

```
export(subplot(p, p2, shareY = TRUE, titleX = TRUE) %>% layout(xaxis = list(title = "Phosphate"), xaxis
```



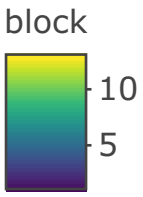

```
p1 <- plot_ly(grouped, x = ~phosphate, y = ~depth, color = ~block, type = "scatter", mode = "markers")
p2 <- plot_ly(grouped, x = ~phosphate, y = ~depth, color = ~block, type = "scatter", mode = "markers")
#p3 <- plot_ly(grouped, x = ~phosphate, y = ~depth, color = ~block, type = "scatter", mode = "markers")
#p4 <- plot_ly(grouped, x = ~phosphate, y = ~depth, color = ~block, type = "scatter", mode = "markers")
export(subplot(p1, p2), "subplots.pdf")
```




```
export(plot_ly(grouped, x = ~phosphate, y = ~depth, color = ~block, type = "scatter", mode = "markers")
```

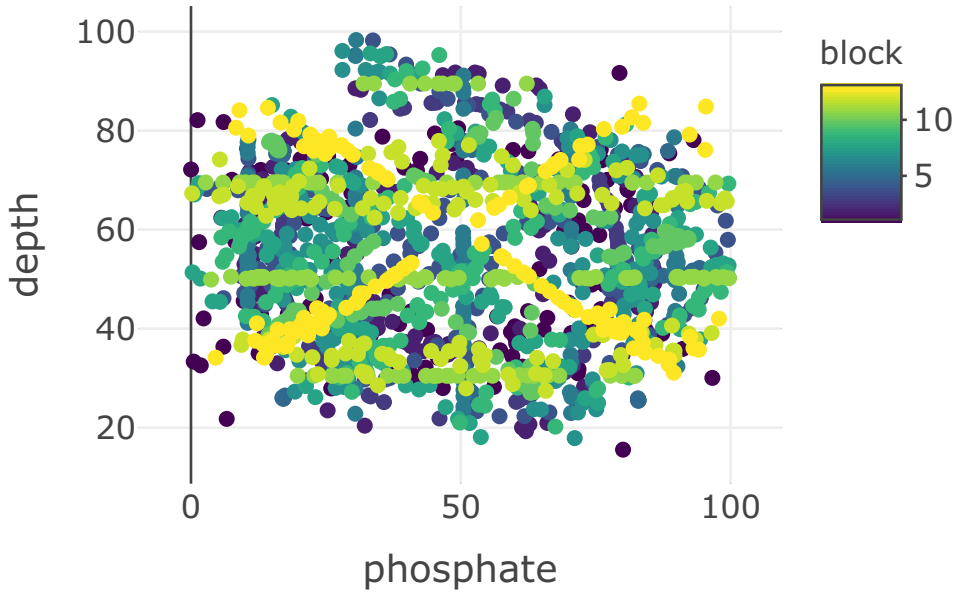



```
# below is the code for the 3d plot in plotly.
# Need to have the plotly package installed
# Don't know how to change the color ramp to be more distinct yet other than the c(color1,
# color2, color3, ...)

plot_ly(grouped, type = "scatter3d", x = ~block , y = ~depth , z = ~phosphate, color = ~block, mode = "
```

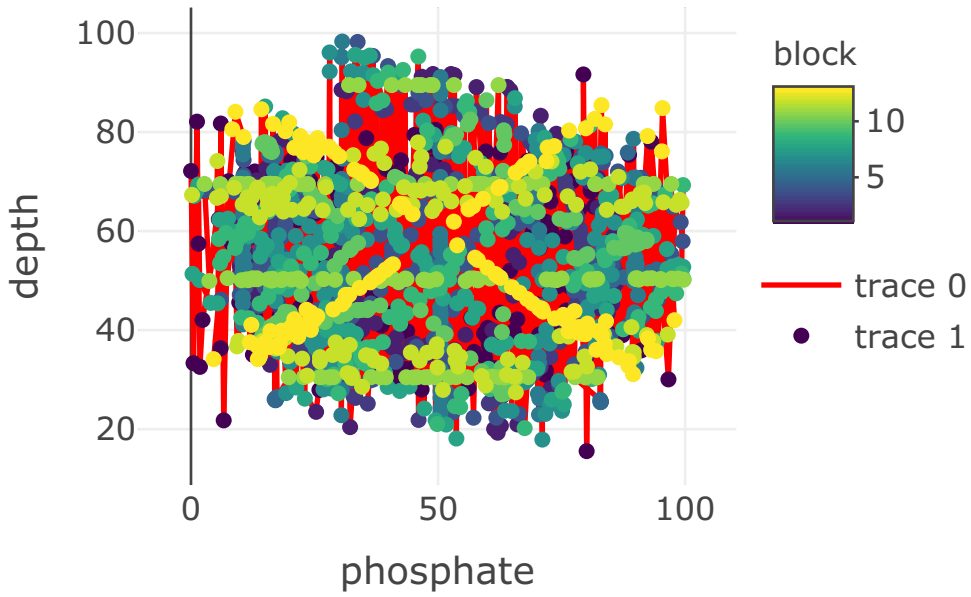


Webgl is not supported by your browser - visit http://get.webgl.org for more info

```
plot_ly(grouped, x = ~phosphate, y = ~depth, type = "scatter", mode = "markers", color = ~block)
```




```
# This one is wild looking with the lines.  It doesn't make much sense.

plot_ly(grouped, x = ~phosphate, y = ~depth, color = I("red")) %>%
add_lines() %>%
add_markers(color = ~block) %>%
layout(showlegend = TRUE)
```




```
############################################################
############################################################
############ END OF HOMEWORK ###############################
############################################################
############################################################

### We will have a more detailed plotting discussion
### and do the following 2 data sets next week
```