HW4_Wei_Yanran

Yanran Wei

September 26, 2017

Problem 3

According to Roger Peng, the focus of the EDA stage of an analysis are listed as below:

1. Show comparisons.
2. Show causality, mechanism, explanation and systematic structure.
3. Show multivariate data.
4. Integrate evidence.
5. Descirbe and document the evidence.
6. Content. Analytical presentations ultimately stand or fall depending on the quality, relevance and integrity of their content.

Problem 4

Question 1

1846 observations are included in the dataset. The summary of the dataset is as below. Three variables are included in the dataset, block, depth and phosphate.

Table 1: Summary of Data

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

Question 2

Looking at the first several lines of the dataset which is shown as below.

Table 2: Sample Data

block	depth	phosphate
4	55.38	97.18
4	51.54	96.03
4	46.15	94.49
4	42.82	91.41
4	40.77	88.33
4	38.72	84.87

Three variables contained in the data, block, depth and phosphate.

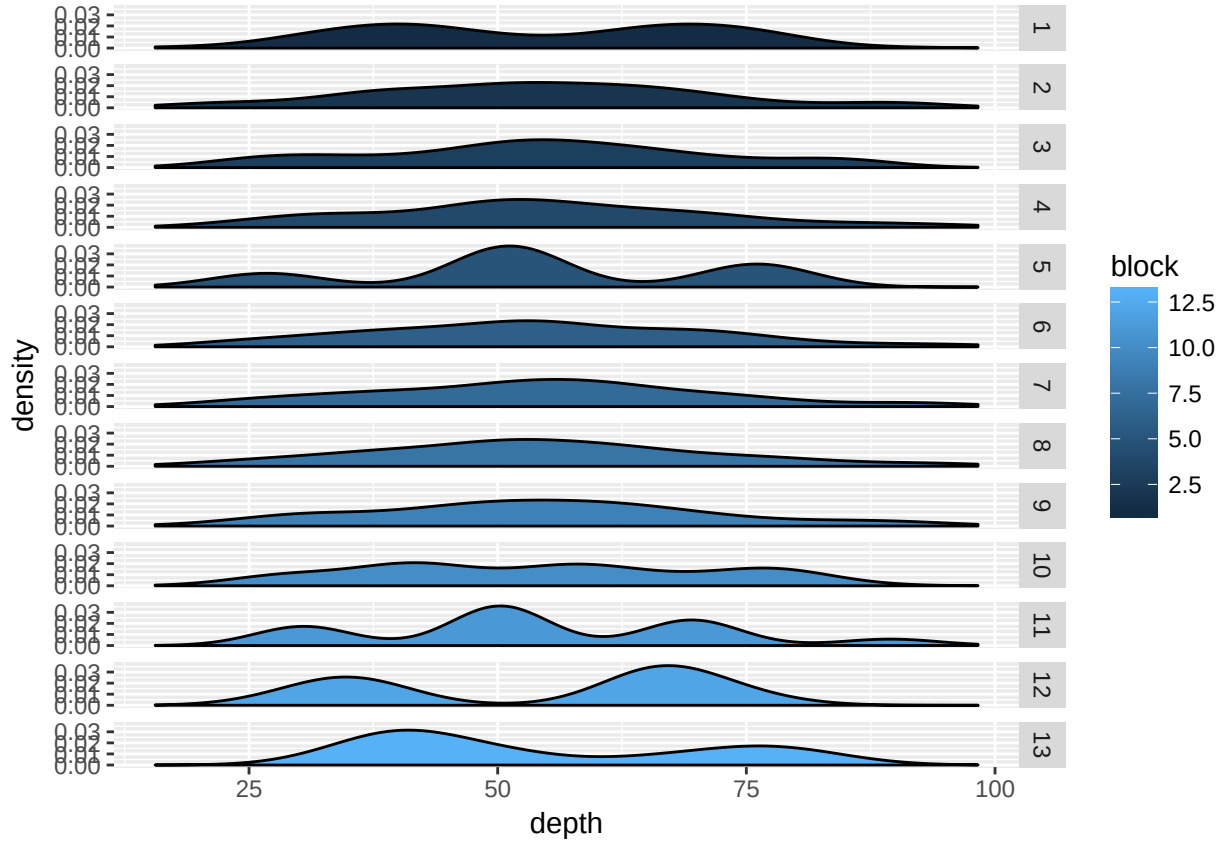
Table 3: Value and Frequency of Block

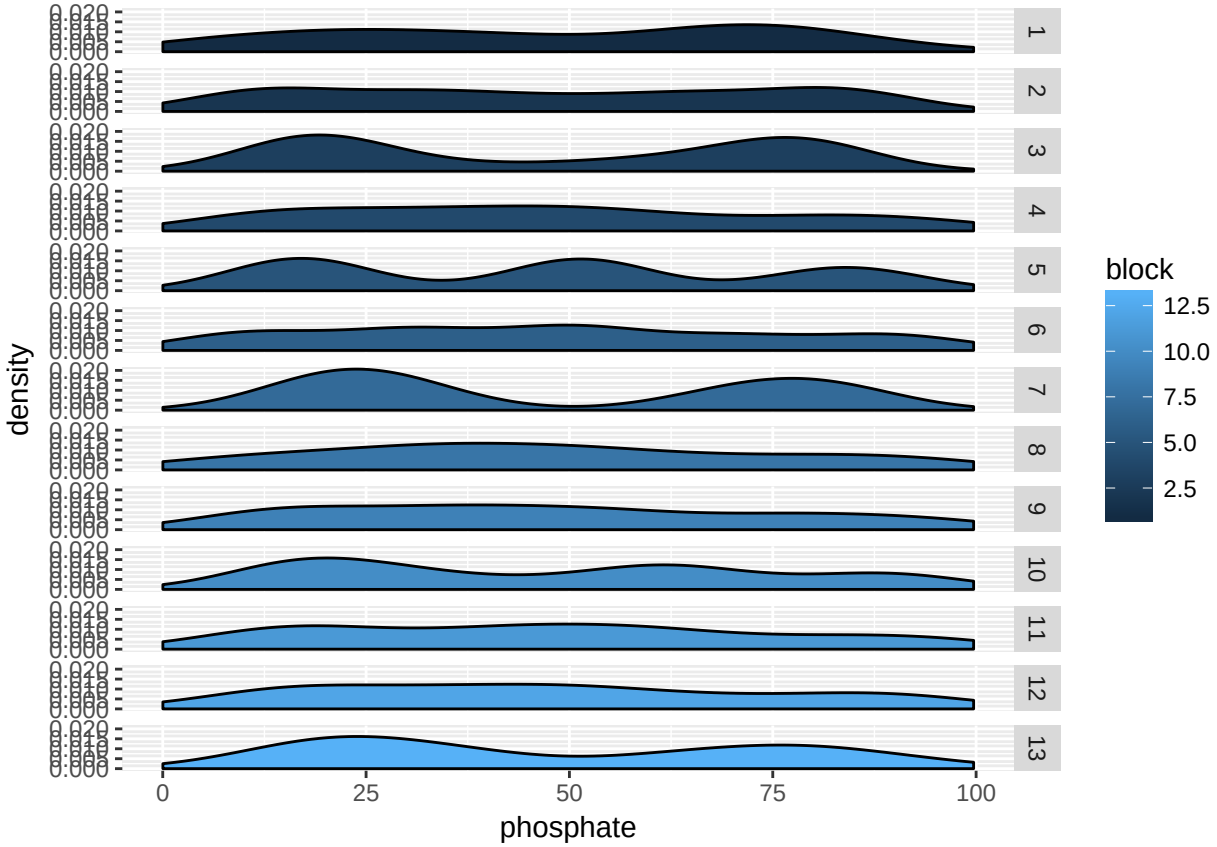
1	2	3	4	5	6	7	8	9	10	11	12	13
142	142	142	142	142	142	142	142	142	142	142	142	142

Block has 13 different values. Each value has 142 corresponding observations. Block is discrete variable whicle depth and phosphate are continuous variables.

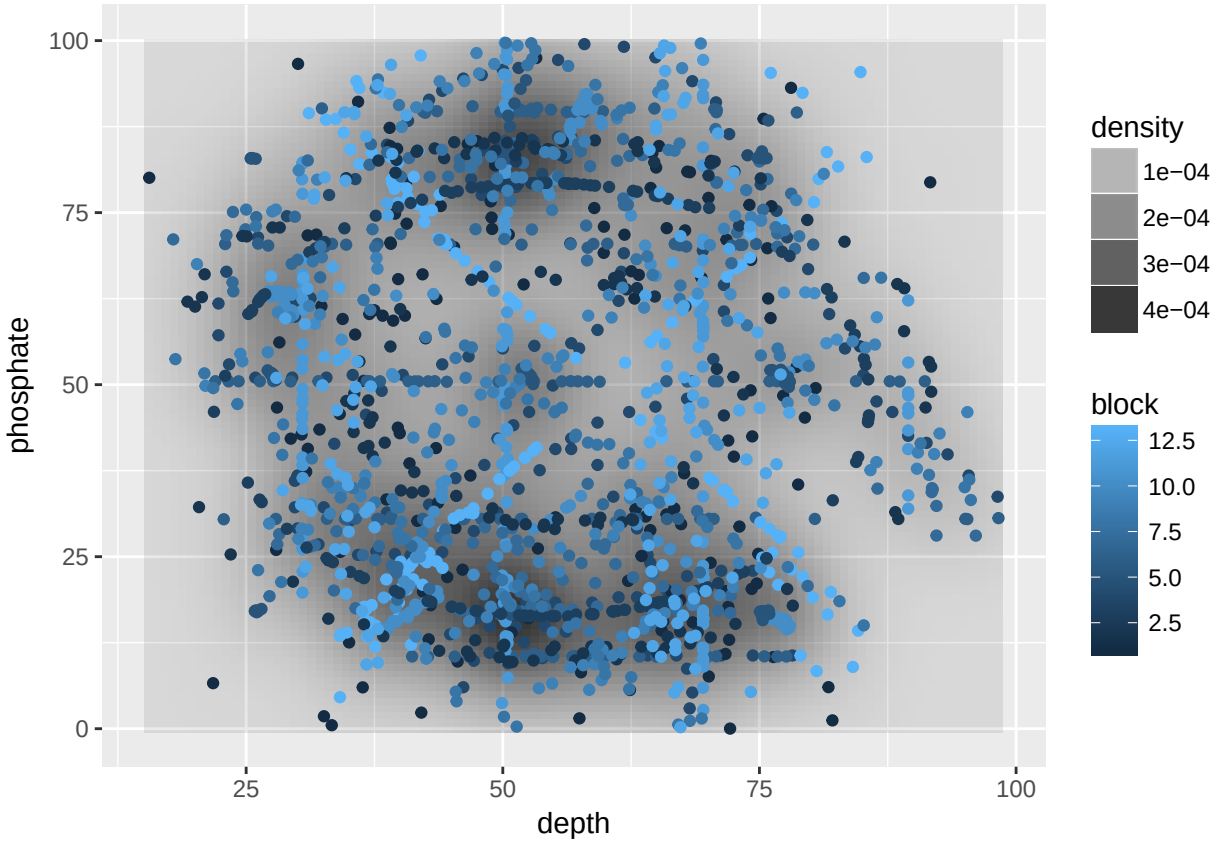
Question 3

As block is discrete vairable, we want to get overview of the distribution of depth and phosphate corresponding to different blocks.

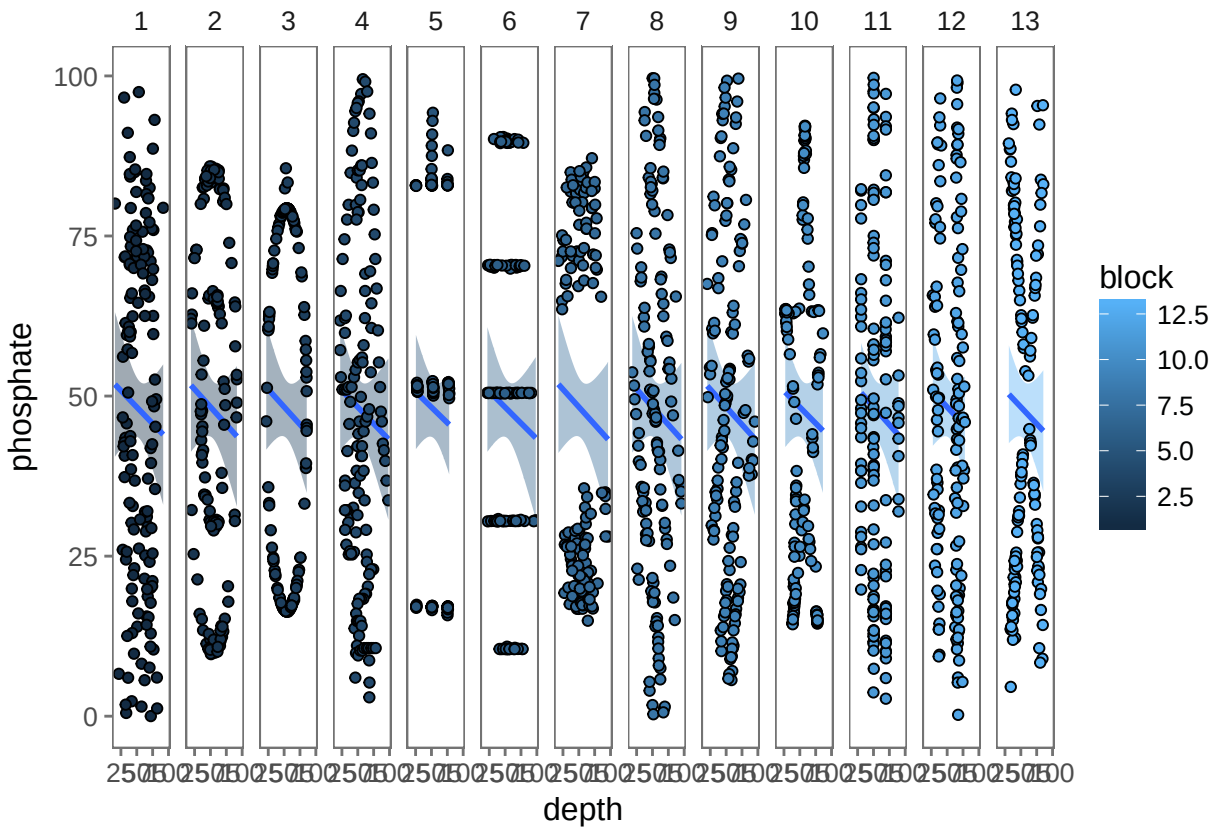




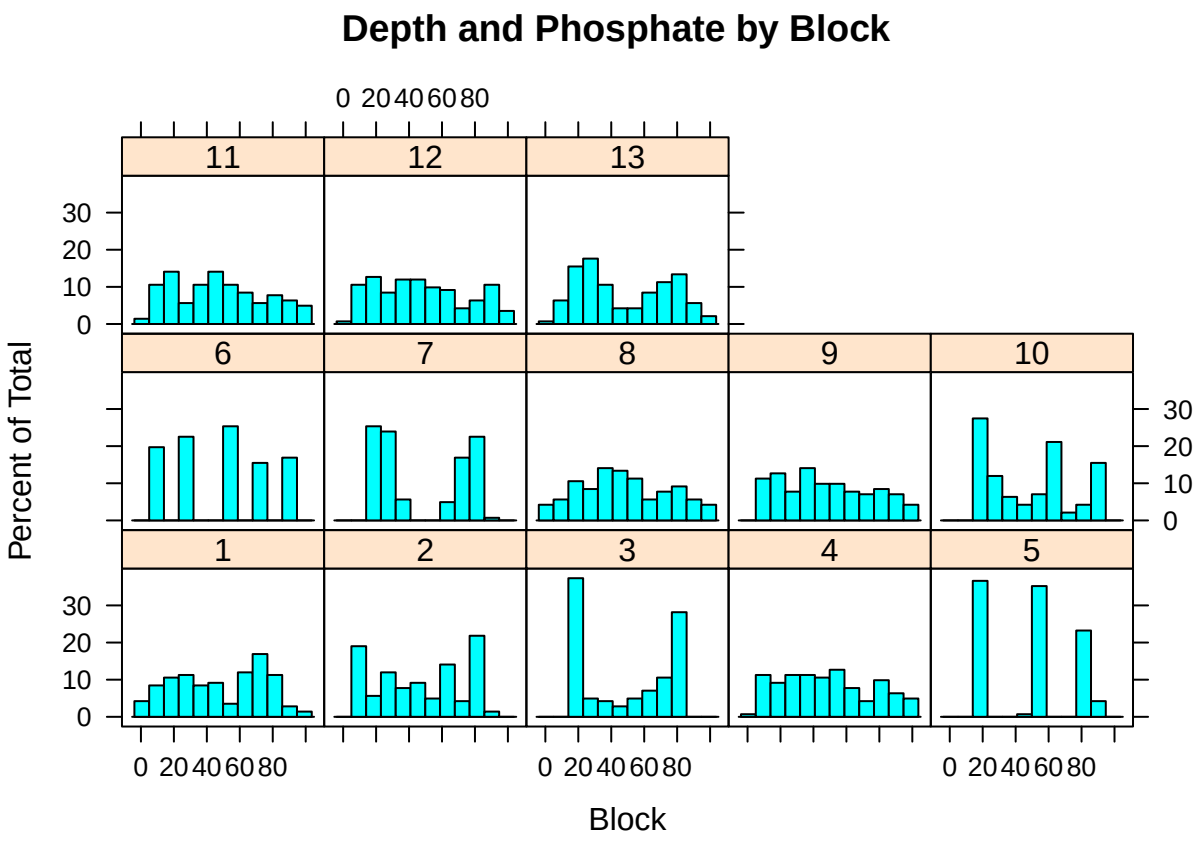
The density of depth seems lower than that of phosphate.



The data points of depth and phosphate are dispersed in the shape of circle according to the graph below. To ignore the influence of block, we made scatterplot and histogram of depth and phosphate for each block.

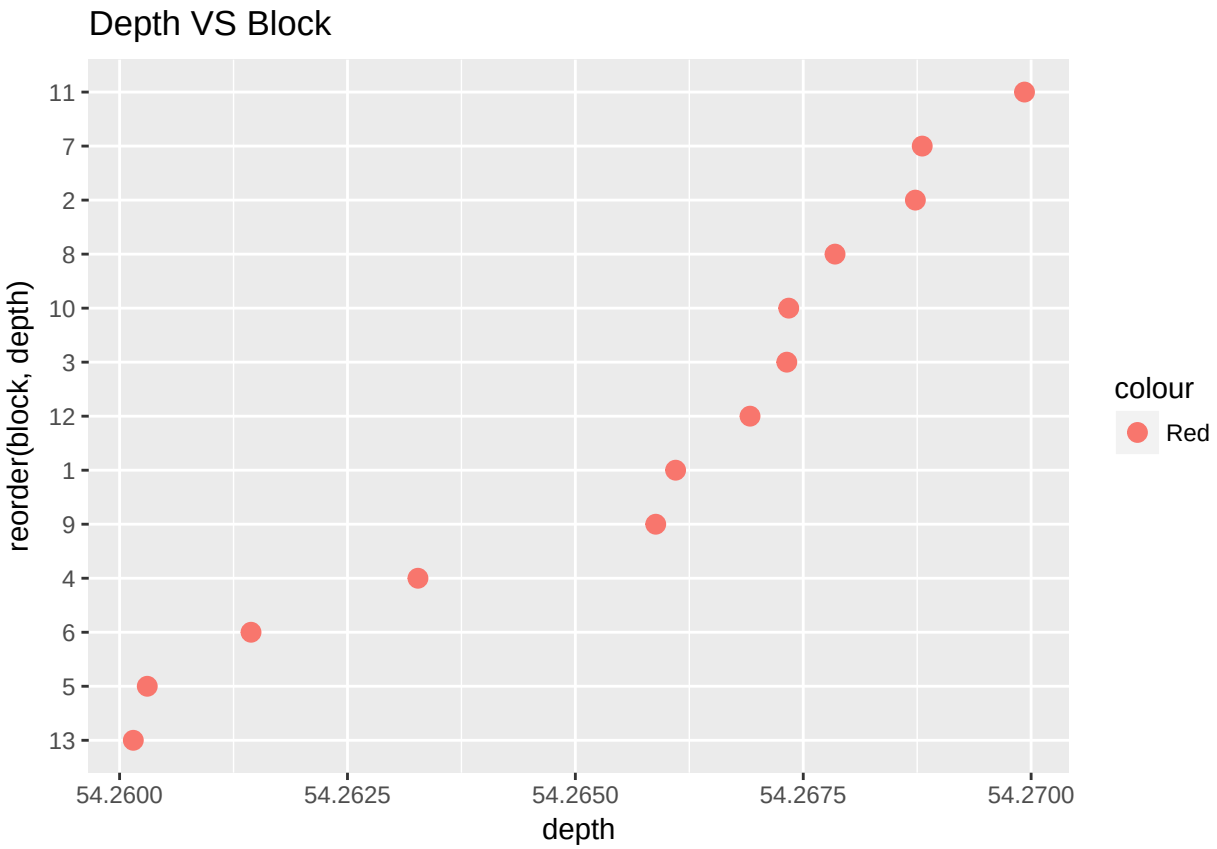


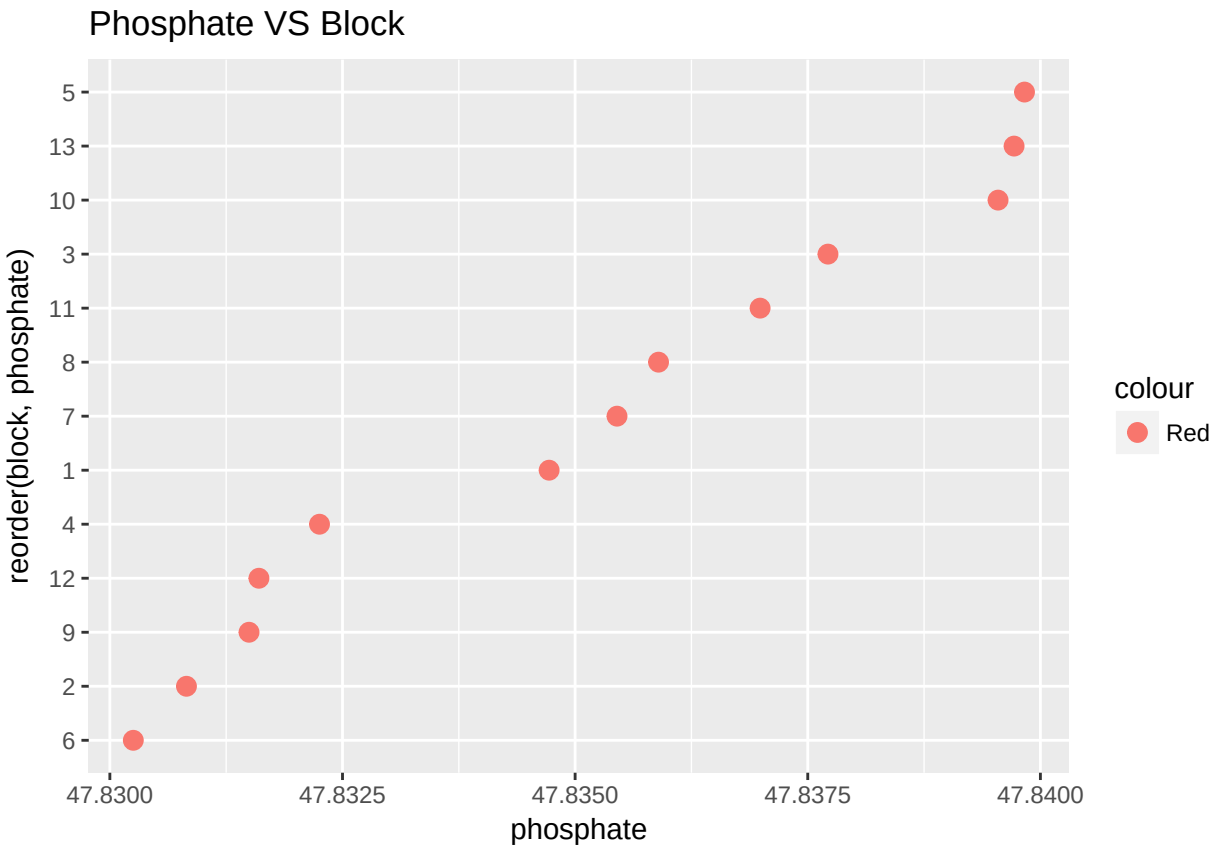
For depth equal to 5, the phosphate value has around five different levels. For other blocks, no obvious linear relationship.



From the histogram, block 5 only has around 5 bars which is identical to what we have observed in the scatterplot. For blocks 1, there are only 5 bars too. No other information I can get from the graph.

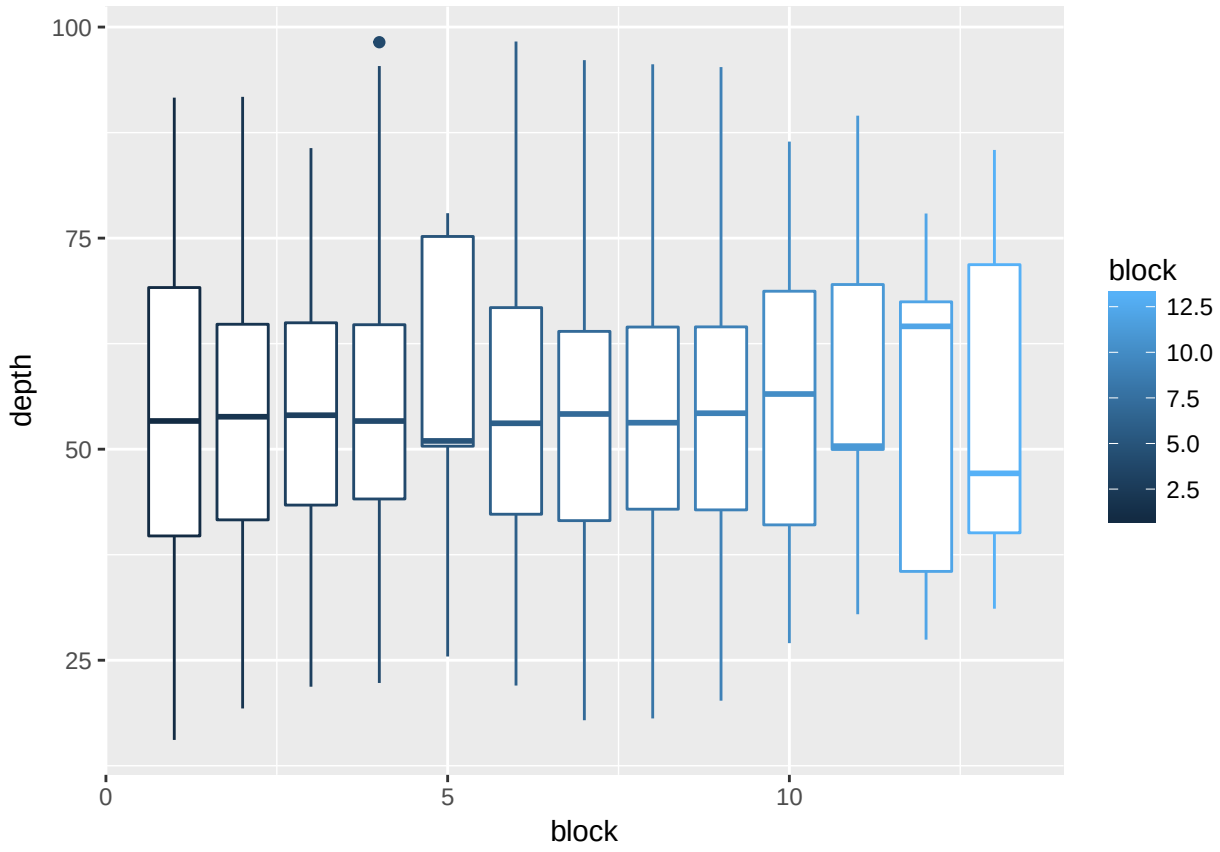
So, what's relationship between block and depth, block and phosphate?

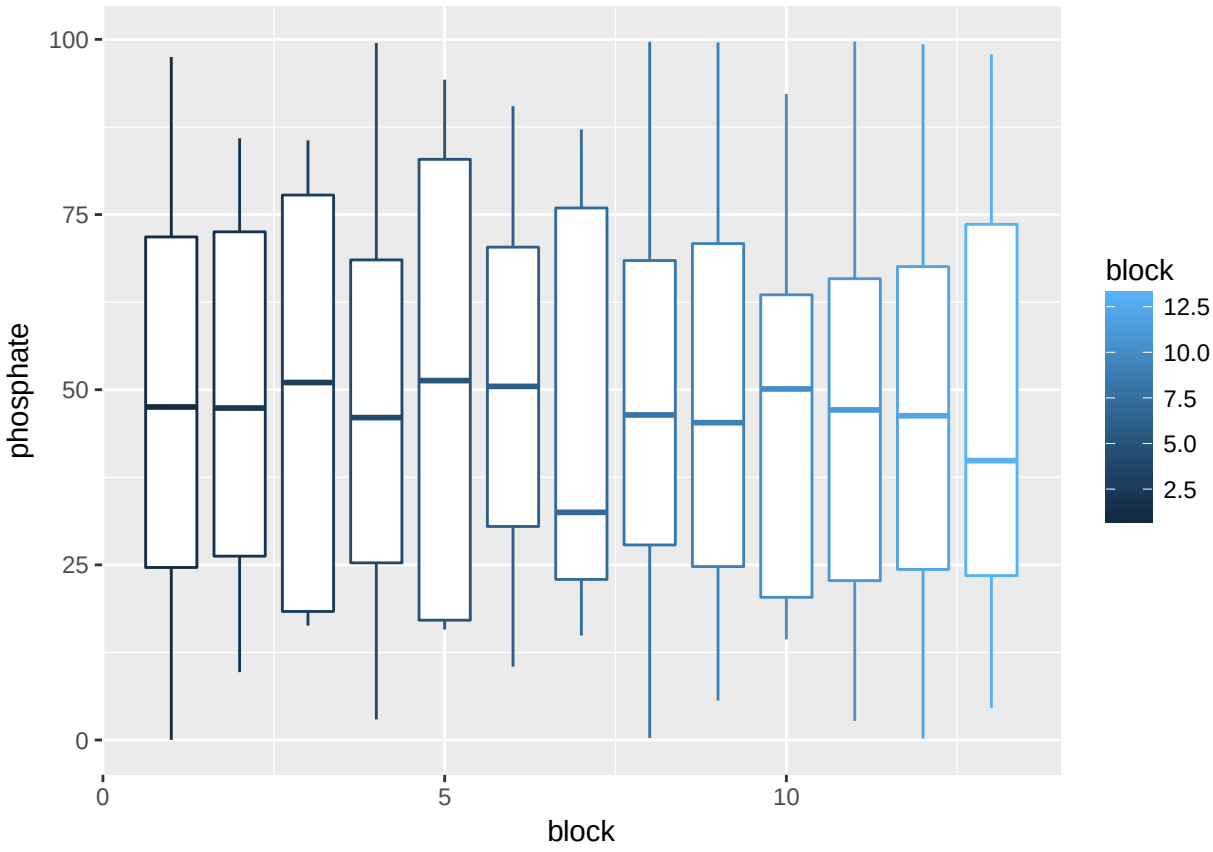

Depth VS Block
reorder(block, depth)
11
7
2
8
10
3
12
1
9
4
6
5
13
54.2600
54.2625
54.2650
54.2675
54.2700
depth
colour
Red

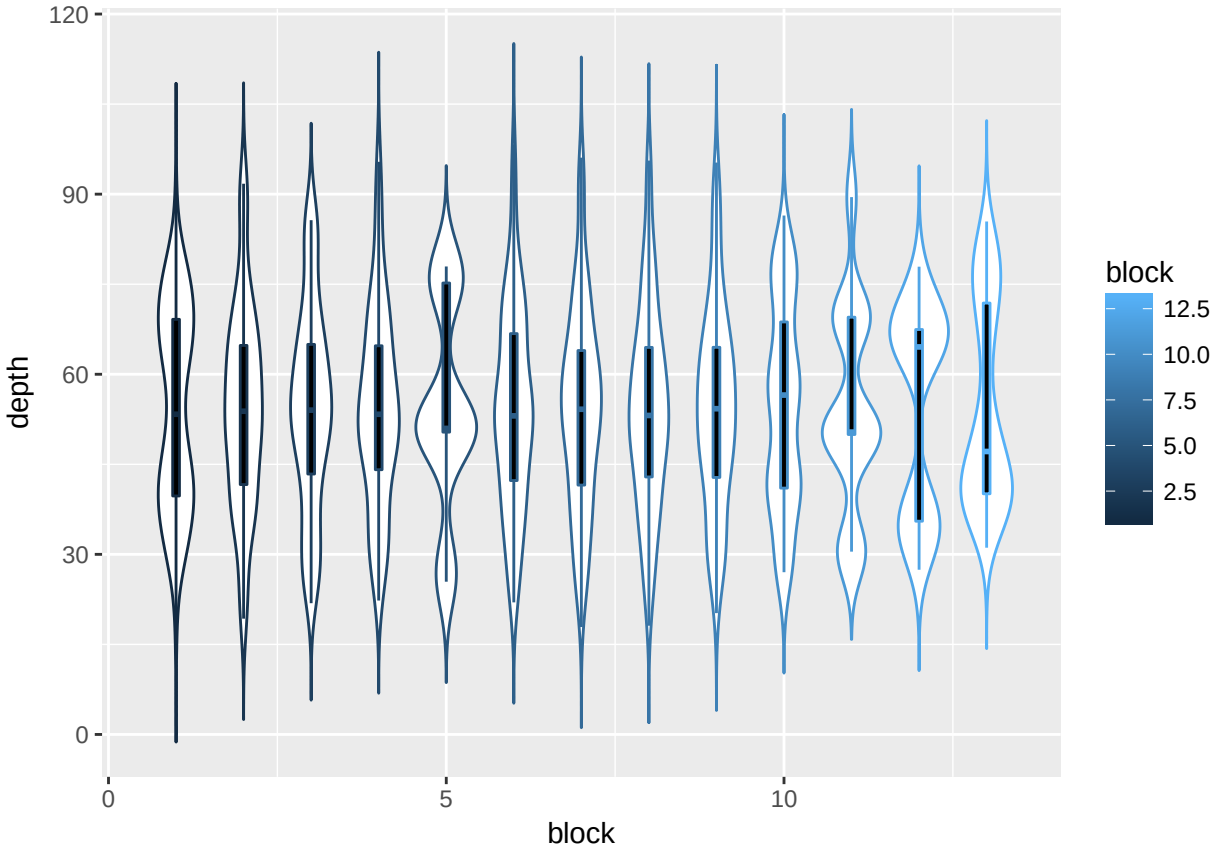


From the first graph, we can see that block 11 has the highest average depth value while block 13 has the lowest depth value. No obvious pattern between depth and block.

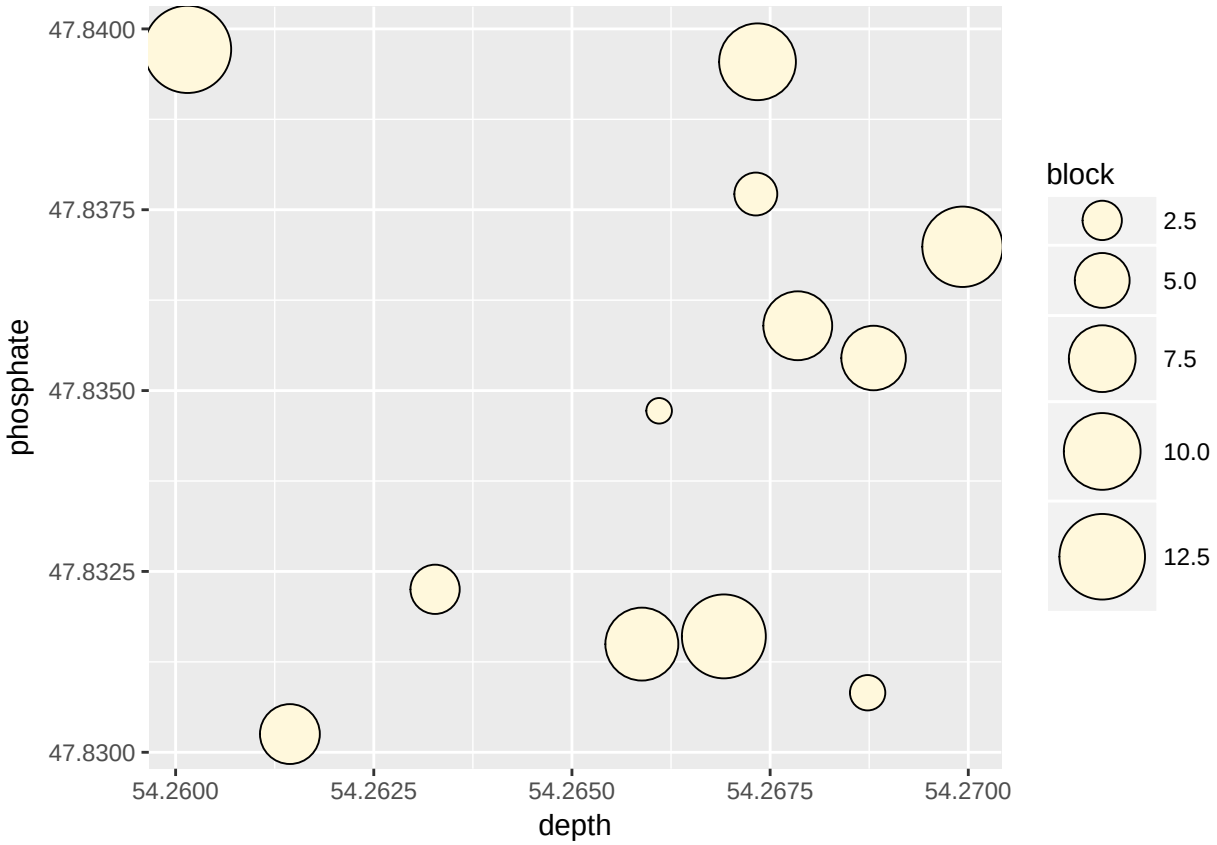
From the second graph, we can see that block 5 has the higest phosphate while block 6 has the lowest phosphate value. No obvious pattern between depth and block, either.


100
75
50
25
depth
0
5
10
block
block
12.5
10.0
7.5
5.0
2.5


phosphate
100
75
50
25
0
0
5
10
block
block
12.5
10.0
7.5
5.0
2.5

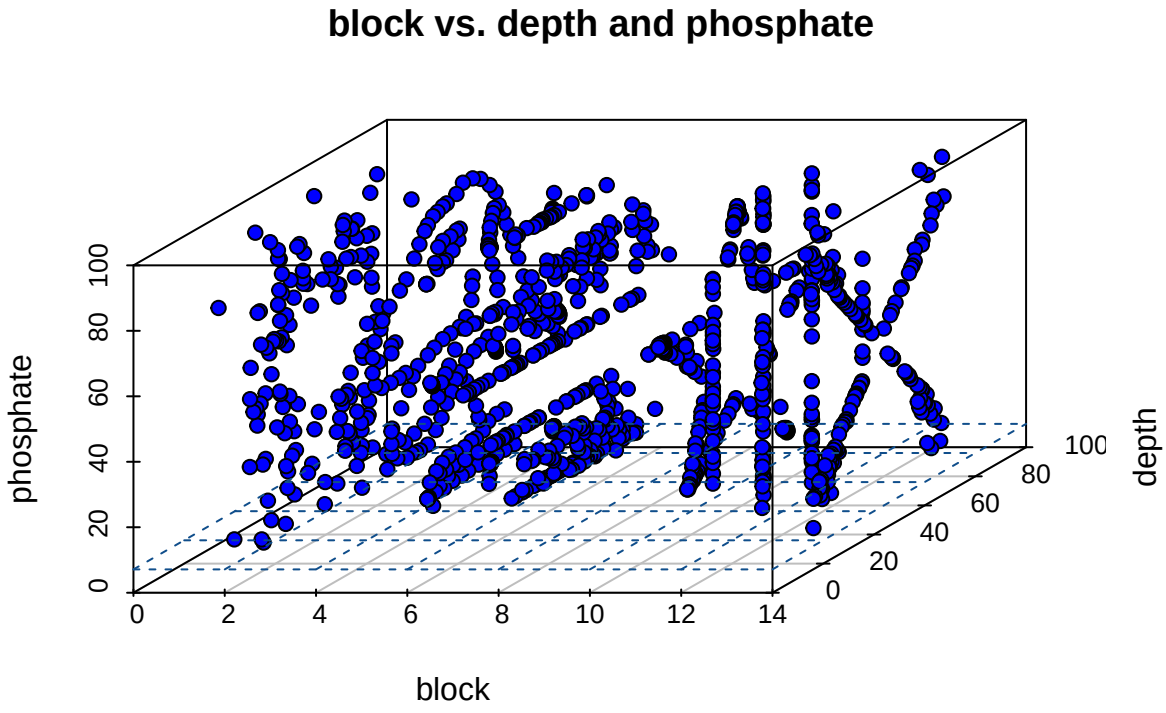


In the box plot, the median depth value for block 1 to 10 does not have big difference while that for block 11 to 13 differs a lot. The median phosphate value for block 7 is the smallest one compared to that of other blocks.

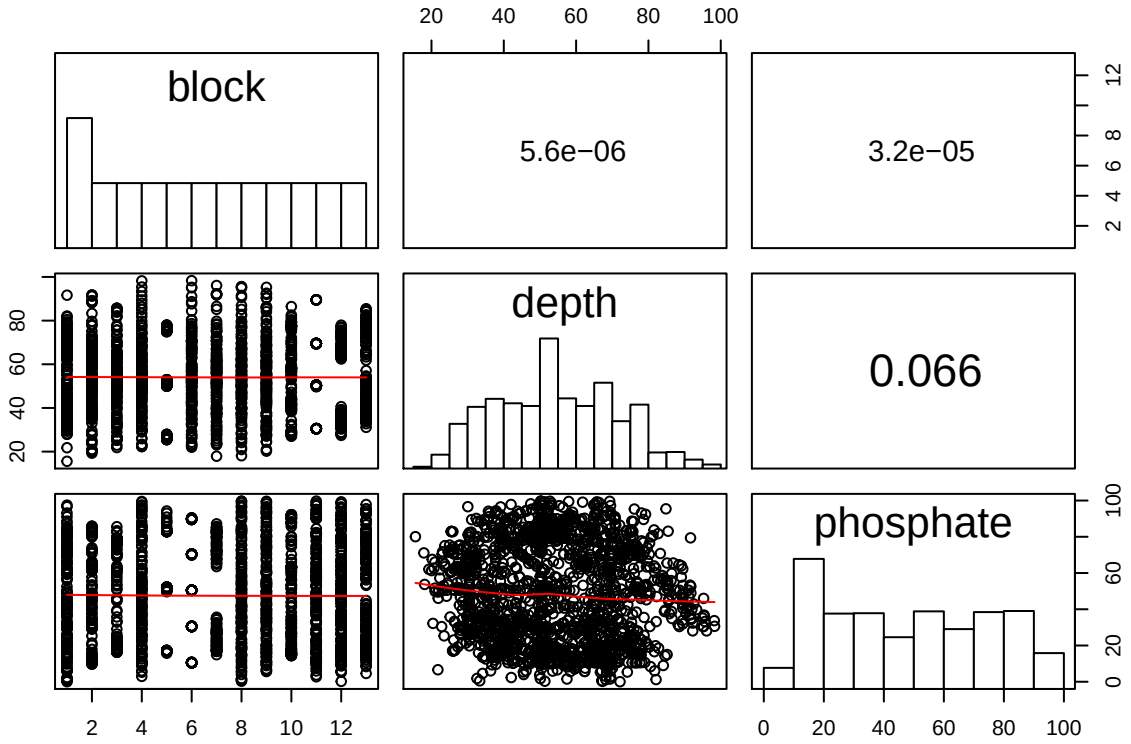


From the bubble plot in which bubble's area represent the size of block, one big bubble is located in left side the graph where the depth value is rather small. The other relatively big bubbles are maily located in the right side of graph when depth is rather large.

Question 4



In the 3D graph, no obvious relationship can be observed between block, depth and phosphate.



The upper right triangle represents correlation between variables. The lower left triangle are scatterplot with smooth lines. The graph on diagnol are bar plots for each variables.

So the correlation between block and depth is almost 0. The correaltion between block and phosphate is also almost 0. The correlation between depth and phosphate is 0.066. So block has no influence on depth and phosphate. However, from the scatterplot between depth and phosphate, we can not make a conclusion that there are realtion ship between these two variables althougth thier corrlation value is 0.066.

Problem 5

In my opinion, the graph showing combination of histogram, scatterplot and correlation value plays an important role. It provide both statistical and visual analysis of variables.