Homework4

Yueyao Wang

September 25, 2017

Problem3

By Roger Peng, the EDA should focus on identifying relationships between variables that are particularly interesting or unexpected, checking to see if there is any evidence for or against a stated hypothesis, checking for problems with the collected data, such as missing data or measurement error), or identifying certain areas where more data need to be collected.

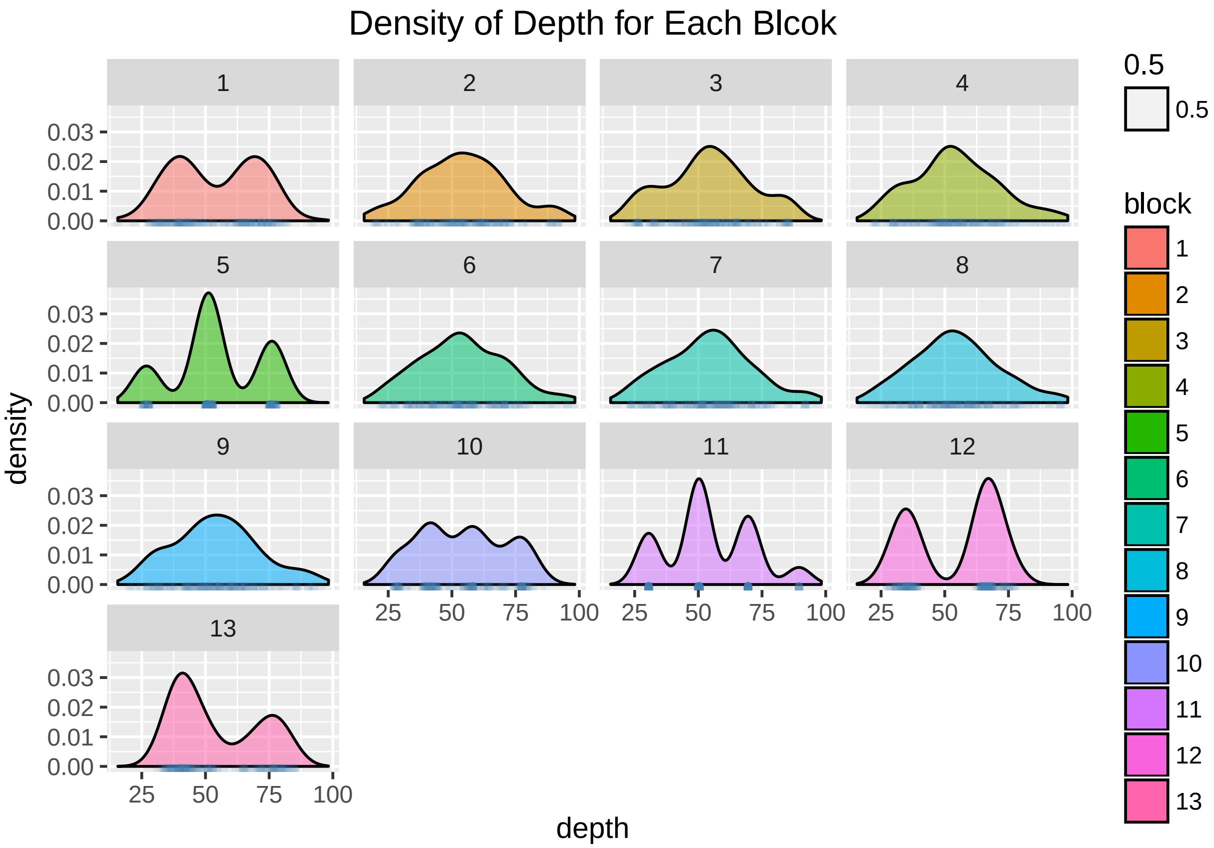
Problem4

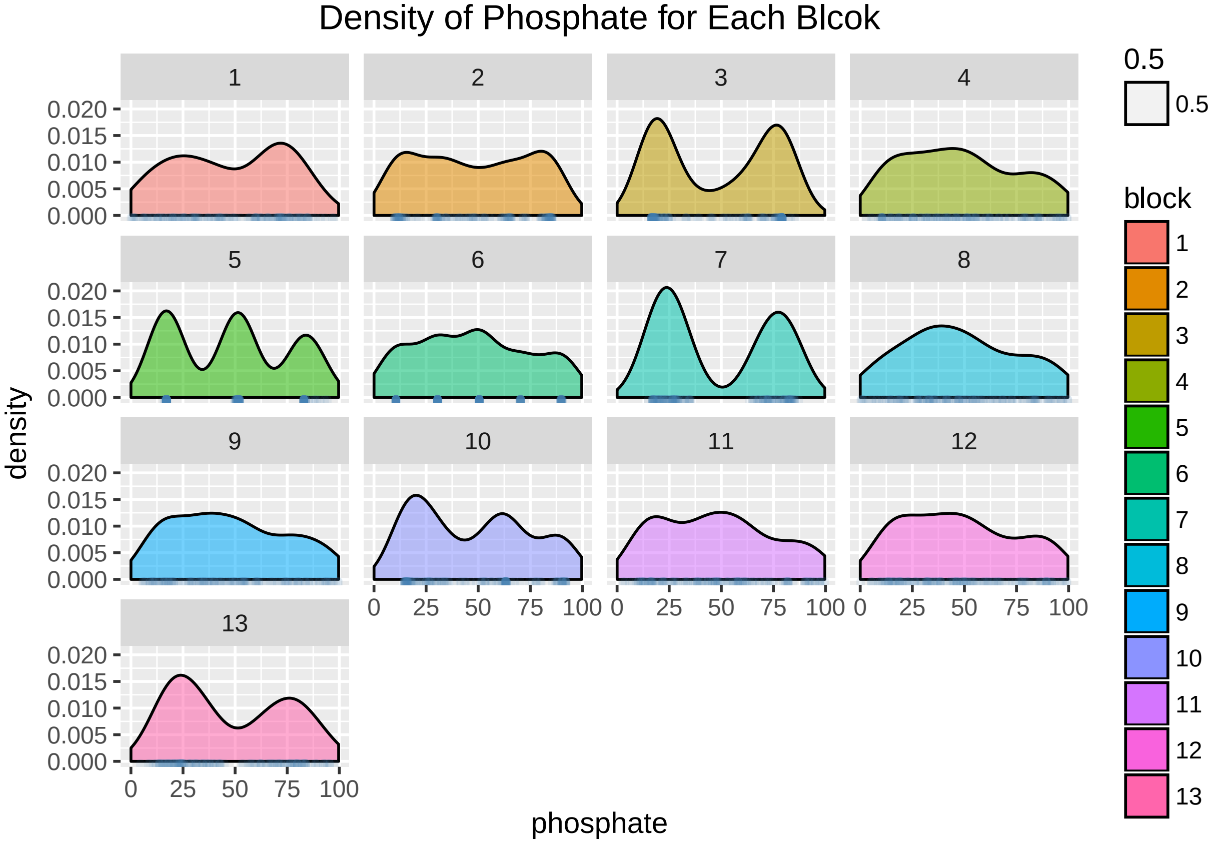
1. Summary Statistics

Table 1: Summary of data by Obersvers

	mean_depth	mean_phosphate	sd_depth	sd_phosphate	correlation
Observer1	54.2661	47.8347	16.7698	26.9397	-0.0641
Observer2	54.2687	47.8308	16.7692	26.9357	-0.0686
Observer3	54.2673	47.8377	16.7600	26.9300	-0.0683
Observer4	54.2633	47.8323	16.7651	26.9354	-0.0645
Observer5	54.2603	47.8398	16.7677	26.9302	-0.0603
Observer6	54.2614	47.8303	16.7659	26.9399	-0.0617
Observer7	54.2688	47.8355	16.7667	26.9400	-0.0685
Observer8	54.2678	47.8359	16.7668	26.9361	-0.0690
Observer9	54.2659	47.8315	16.7689	26.9386	-0.0686
Observer10	54.2673	47.8395	16.7690	26.9303	-0.0630
Observer11	54.2699	47.8370	16.7700	26.9377	-0.0694
Observer12	54.2669	47.8316	16.7700	26.9379	-0.0666
Observer13	54.2602	47.8397	16.7700	26.9300	-0.0656

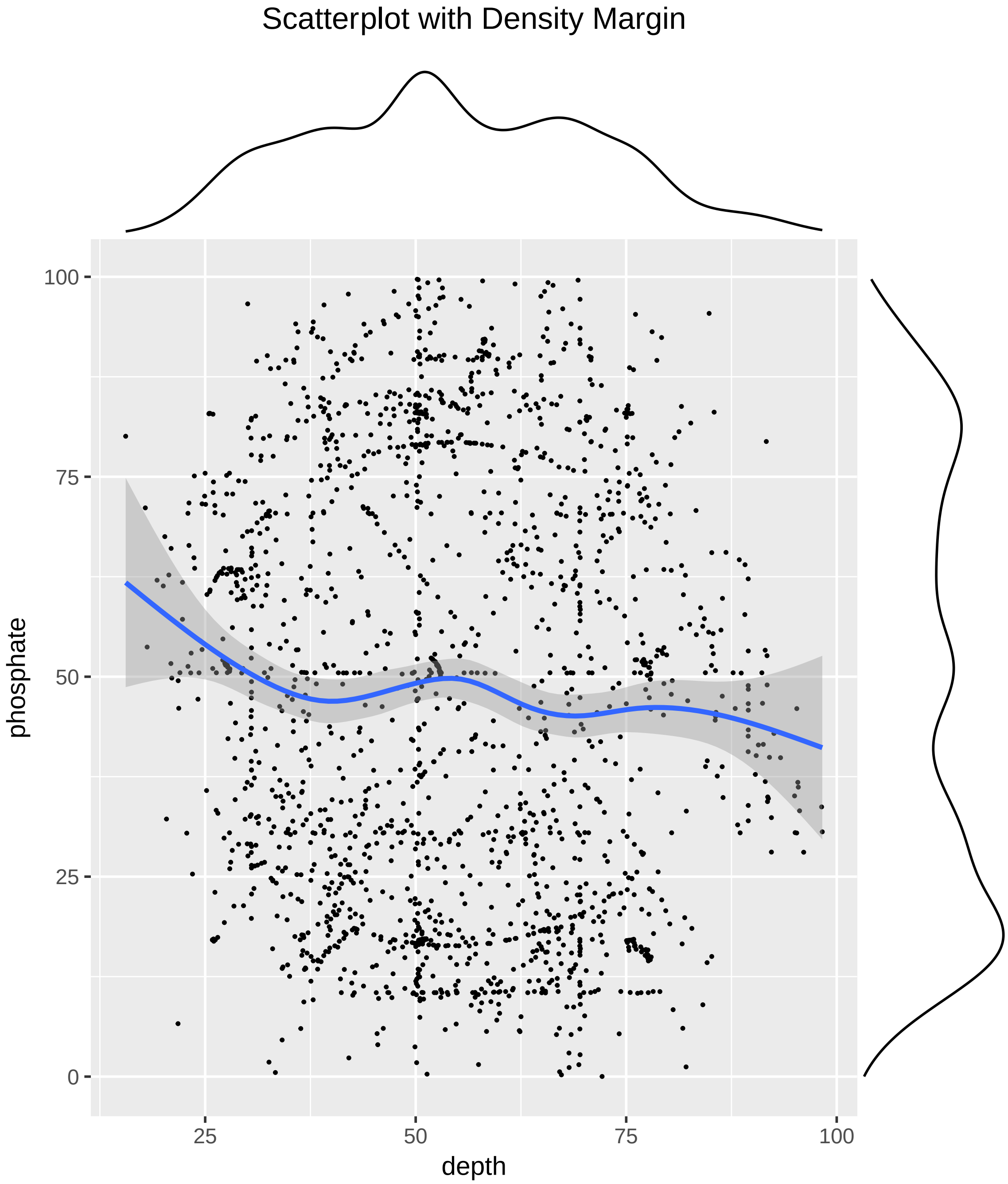
2.

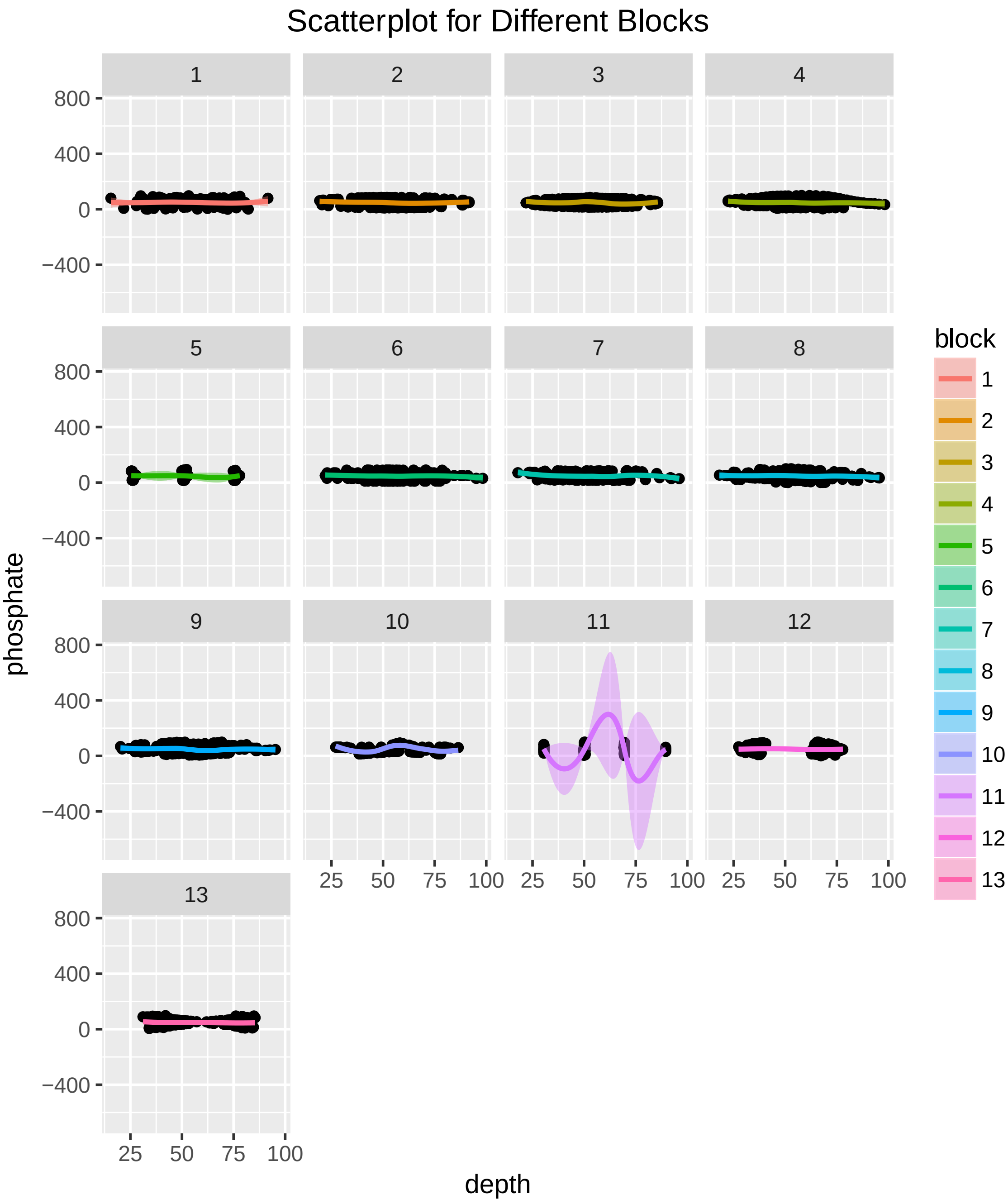




From the above plots we could see that the block is an important factor because the density of phosphate and depth for different are highly different.

3.





4

If we fit the model

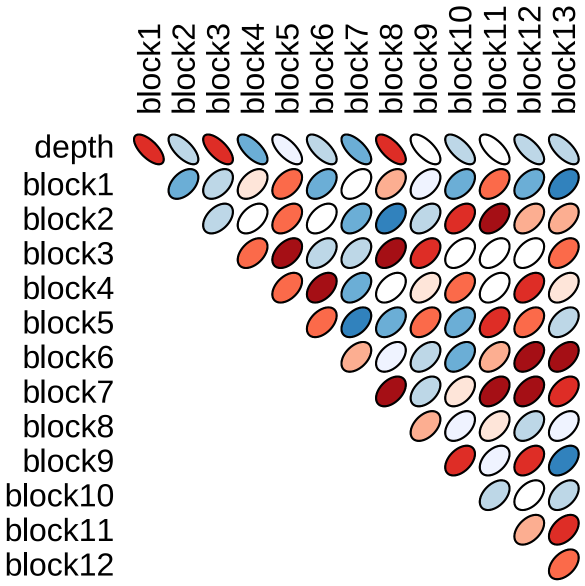
$$phosphate_{ij} = block_i + depth_{ij} + \epsilon$$

, we could get the relationship of phosphate on depth and block. Model coefficients and correlation plots are following:

Table 2: Coefficient of Linear Model

	Estimate	Std. Error	t value	Pr(>|t|)
depth	-0.1060528	0.0374492	-2.831911	0.0046776

	Estimate	Std. Error	t value	Pr(>|t|)
block1	53.5897914	3.0364022	17.649108	0.0000000
block2	53.5861728	3.0364681	17.647533	0.0000000
block3	53.5929174	3.0364328	17.649960	0.0000000
block4	53.5870238	3.0363314	17.648609	0.0000000
block5	53.5942852	3.0362569	17.651433	0.0000000
block6	53.5848287	3.0362855	17.648152	0.0000000
block7	53.5908079	3.0364700	17.649049	0.0000000
block8	53.5911526	3.0364461	17.649302	0.0000000
block9	53.5865433	3.0363968	17.648070	0.0000000
block10	53.5947476	3.0364333	17.650560	0.0000000
block11	53.5924646	3.0364982	17.649431	0.0000000
block12	53.5867593	3.0364227	17.647991	0.0000000
block13	53.5941571	3.0362531	17.651413	0.0000000



5

When we put all blocks data together, the regression line on the scatterplot of depth and phosphate shows us that on average, as depth increasing, the phosphate decreases. However, the dots on the scatterplot spread around and we can't acutally capture that trend. So I did the scatterplot of depth and phosphate again for each block to see their relationship. Except block 11, most blocks phosphate fluctuates around a horizontal line as the depth increasing. This indicates that there isn't much relationship between depth and phosphate for most blocks.

For block 11, from the density plot of depth we can see that the depth for block 11 concentrates on 3 levels and on each level of depth there are different phospahte values. So again, there isn't much relationship between depth and phosphate for block 11.

Appendix: Code

```
#--------load and combine data-------------
prob4_data1 <- read.xlsx("HW4_data.xlsx", sheetIndex = 1)
prob4_data2 <- read.xlsx("HW4_data.xlsx", sheetIndex = 2)
```

```r
prob4_data <- rbind(prob4_data1, prob4_data2)

#------create summary statistics table-----------
descrip_stats <- function(x) {
    # input: x is a dataframe for all samples from 1
    # Observer return:a dataframe of descriptive statstics
    mean_ <- apply(x[, 2:3], MARGIN = 2, FUN = mean)
    sd_ <- apply(x[, 2:3], MARGIN = 2, FUN = sd)
    correlation <- cor(x[, 2], x[, 3])
    d <- data.frame(mean_depth = mean_[1], mean_phosphate = mean_[2],
        sd_depth = sd_[1], sd_phosphate = sd_[2], correlation = correlation)
    return(d)
}
obs_1 <- subset(prob4_data, block == 1)
com_df <- descrip_stats(obs_1)
for (i in 2:13) {
    obs_i <- subset(prob4_data, block == i)
    des_df <- descrip_stats(obs_i)
    com_df <- rbind(com_df, des_df)
}
rownames(com_df) <- paste("Observer", 1:13, sep = "")
kable(com_df, caption = "Summary of data by Obersvers",
    digits = 4) %>% kable_styling(full_width = T)
#----store the dataset used to create ggplots-----
prob4_data_gg <- prob4_data
prob4_data_gg$block <- as.factor(prob4_data_gg$block)
#------density plot-----------
ggplot(data = prob4_data_gg, aes(x = depth, fill = block)) +
    geom_density(aes(alpha = 0.5)) + geom_rug(col = "steelblue",
    alpha = 0.1, size = 1.5) + ggtitle("Density of Depth for Each Blcok") +
    facet_wrap(~block) + theme(plot.title = element_text(hjust = 0.5))

ggplot(data = prob4_data_gg, aes(x = phosphate, fill = block)) +
    geom_density(aes(alpha = 0.5)) + geom_rug(col = "steelblue",
    alpha = 0.1, size = 1.5) + ggtitle("Density of Phosphate for Each Blcok") +
    facet_wrap(~block) + theme(plot.title = element_text(hjust = 0.5))

#----scatterplot with density margin-----
p <- ggplot(data = prob4_data_gg, aes(x = depth, y = phosphate)) +
    geom_point(size = 0.3) + geom_smooth() + ggtitle("Scatterplot with Density Margin") +
    theme(plot.title = element_text(hjust = 0.5))
ggMarginal(p, type = "density")
#---scatterplot for different blocks----
ggplot(data = prob4_data_gg, aes(x = depth, y = phosphate)) +
    geom_point() + geom_smooth(aes(colour = block, fill = block)) +
    facet_wrap(~block) + ggtitle("Scatterplot for Different Blocks") +
    theme(plot.title = element_text(hjust = 0.5))
fit_total <- lm(phosphate ~ depth + block + 0, data = prob4_data_gg)
kable(summary(fit_total)$coefficient, caption = "Coefficient of Linear Model")
cormatrix <- summary(fit_total, correlation = T)$correlation
colors <- c("#A50F15", "#DE2D26", "#FB6A4A", "#FCAE91",
    "#FEE5D9", "white", "#EFF3FF", "#BDD7E7", "#6BAED6",
    "#3182BD", "#08519C")
```

```r
plotcorr(cormatrix, col = colors[(cormatrix * 1e+06)%%11],
    type = "upper")
```