HW4_Porter_Erica

Erica Porter

9/23/2017

Problem 3

According to Roger Peng, some goals of EDA are to examine the data, identify/summarize data characteristics, and identify initial patterns. EDA helps to eliminate initial analyses and efforts that likely won't work, in order to narrow down the collection of possible approaches; EDA certainly helps investogators to gauge which approaches are worthwhile to pursue. Topics of interest during EDA can be: finding issues with the data, understanding its structure, determining variables of interest, and recognizing any evident relationships to investigate throughout the analysis. However, inference and presentation are generally not the focus of Exploratory Data Analysis, as investigators will make progress on these during later stages of analysis. EDA simply sets the stage for more in-depth hypothesis tests, presentation-worthy graphics, and complete explanations.

A few key components of EDA include:

- Display comparisons; this makes hypotheses and graphics more valuable
- Show causality and structure
- Collect evidence and content

Problem 4 & 5

Table 1: First 5 Observations

block	depth	phosphate
4	55.3846	97.1795
4	51.5385	96.0256
4	46.1538	94.4872
4	42.8205	91.4103
4	40.7692	88.3333
4	38.7179	84.8718

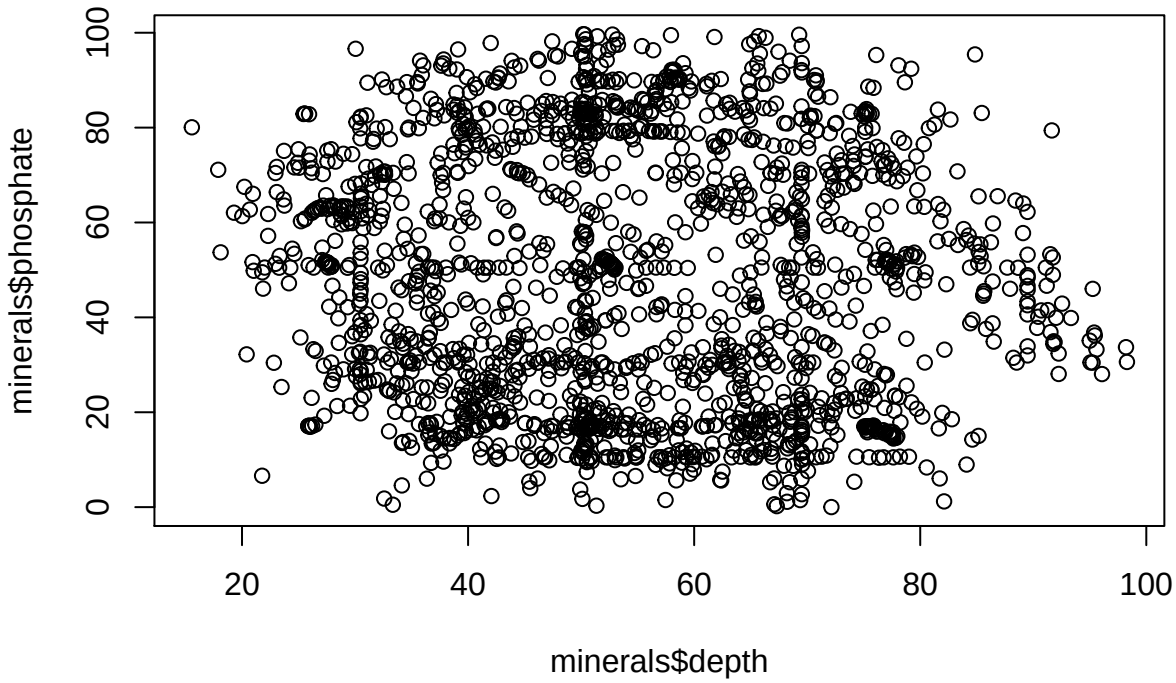
Table 2: Last 5 Observations

block	depth	phosphate
12	34.73183	19.60180
12	33.67444	26.09049
12	75.62726	37.12875
12	40.61013	89.13624
12	39.11437	96.48175
12	34.58383	89.58890

Table 3: Summary of HW4 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

block	depth	phosphate
Max. :13	Max. :98.29	Max. :99.69468

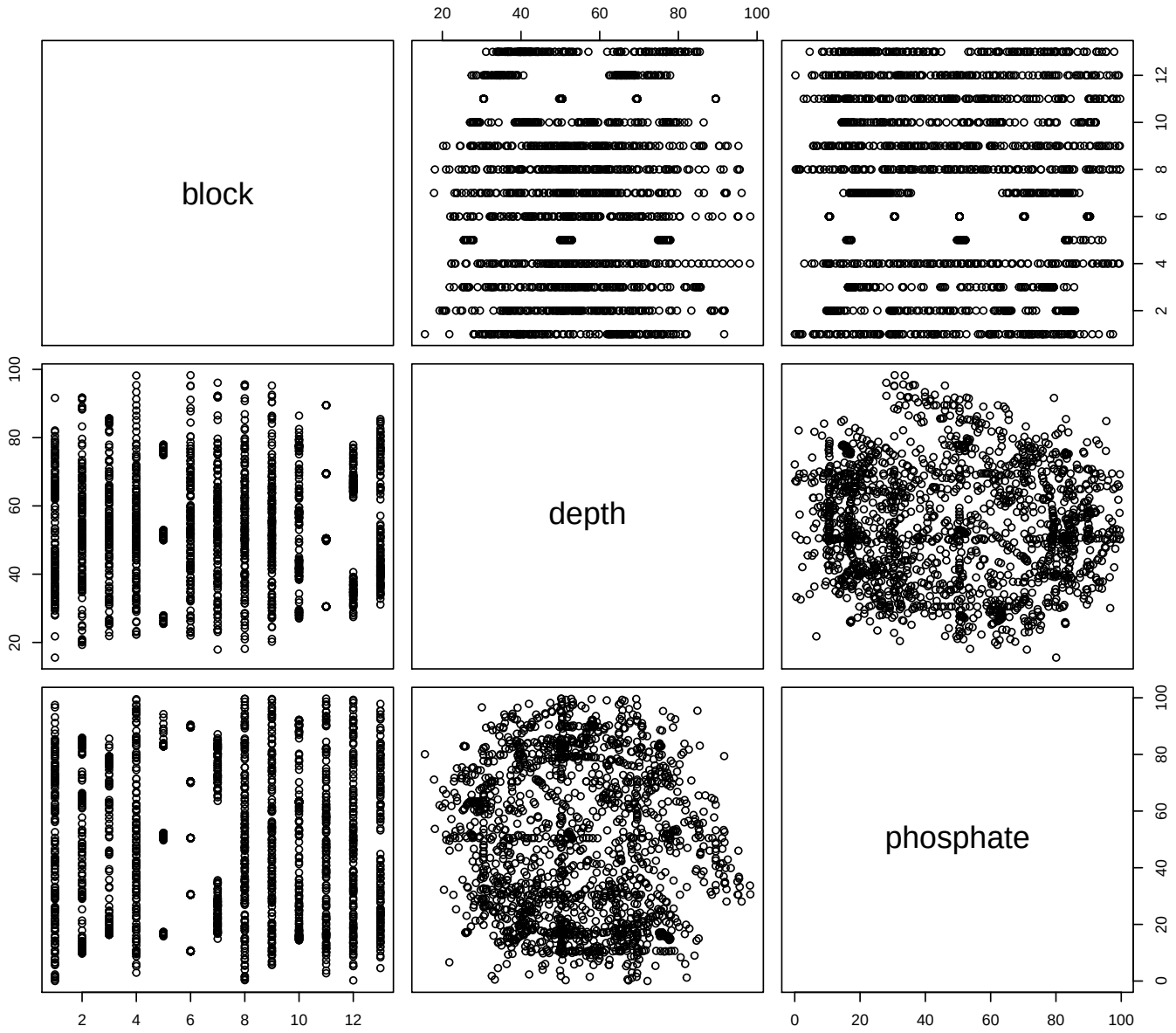



```
dev_data <- readRDS("./HW3_data.rds")

sum_stats <- as.data.frame(matrix(NA,nrow=13,ncol=6))

block_summary <- function(data,observers) {
  temp <- c(mean(subset(data, Observer == observers)$dev1), mean(subset(data, Observer == observers)$dev2), sd
  return(c(observers,temp))
}

for(i in 1:13) {
    sum_stats[i,] <- block_summary(data=dev_data,observers=i)
}

colnames(sum_stats) <- c("Block", "Mean1", "Mean2", "SD1", "SD2", "Correlation")

knitr::kable(sum_stats, caption = "Summary of data by block")
```

Table 4: Summary of data by block

Block	Mean1	Mean2	SD1	SD2	Correlation
1	54.26610	47.83472	16.76983	26.93974	-0.0641284
2	54.26873	47.83082	16.76924	26.93573	-0.0685864
3	54.26732	47.83772	16.76001	26.93004	-0.0683434
4	54.26327	47.83225	16.76514	26.93540	-0.0644719
5	54.26030	47.83983	16.76774	26.93019	-0.0603414
6	54.26144	47.83025	16.76590	26.93988	-0.0617148
7	54.26881	47.83545	16.76670	26.94000	-0.0685042
8	54.26785	47.83590	16.76676	26.93610	-0.0689797

Block	Mean1	Mean2	SD1	SD2	Correlation
9	54.26588	47.83150	16.76885	26.93861	-0.0686092
10	54.26734	47.83955	16.76896	26.93027	-0.0629611
11	54.26993	47.83699	16.76996	26.93768	-0.0694456
12	54.26692	47.83160	16.77000	26.93790	-0.0665752
13	54.26015	47.83972	16.76996	26.93000	-0.0655833



The two sheets of data have the same column names and types, so I combined their rows into one data frame called "minerals". The data set has three variables: "block", "depth", and "phosphate." The "block" variable is initially classified as numeric since it contains integers from 1 to 13, but it appears to be some sort of factor variable, perhaps specifying the areas in which depth and phosphate measurements were taken.

The scatterplot of all data points with "depth" versus "phosphate" is not very clear, but there is no evident overall correlation between the two variables from this plot; there is not an overall trend. The individual scatterplots of "depth" versus "phosphate" appear to demonstrate very different trends/shapes. The data from block 4 plots in the shape of a dinosaur, and block 12 data plots in the shape of a star. This grid of scatterplots by block is the most useful visual to me because it demonstrates that the data are not correlated and the data for the blocks is not distributed similaryl despite their similar summary statistics. Since the data is not correlated, other plots would not improve analysis; individual boxplots would also demostrate the (misleading) similar correlations, and histograms may demonstrate the distributions, but since the data and realtionships are

not real/significant, further models and analysis after that would be unproductive.