STAT 5014: Homework 4

Bobby Soule

9/26/2017

For each assignment, turn in by the due date/time. Late assignments must be arranged prior to submission. In every case, assignments are to be typed neatly using proper English in Markdown.

This week, we spoke about Exploratory Data Analysis and plotting. To begin the homework, we will as usual, start by loading, munging and creating tidy data sets. In this homework, our goal is to create informative (and perhaps pretty) plots showing features or perhaps deficiencies in the data.

Problem 1

Work through the Swirl "Exploratory_Data_Analysis" lesson parts 1 - 10.

Problem 2

As in the last homework, create a new R Markdown file within the project folder within the "04_projecting_knowledge_plots" subfolder (file–>new–>R Markdown–>save as.

The filename should be: HW4_lastname_firstname, i.e. for me it would be HW4_Settlage_Bob

You will use this new R Markdown file to solve problems 4-7.

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.

Problem 4

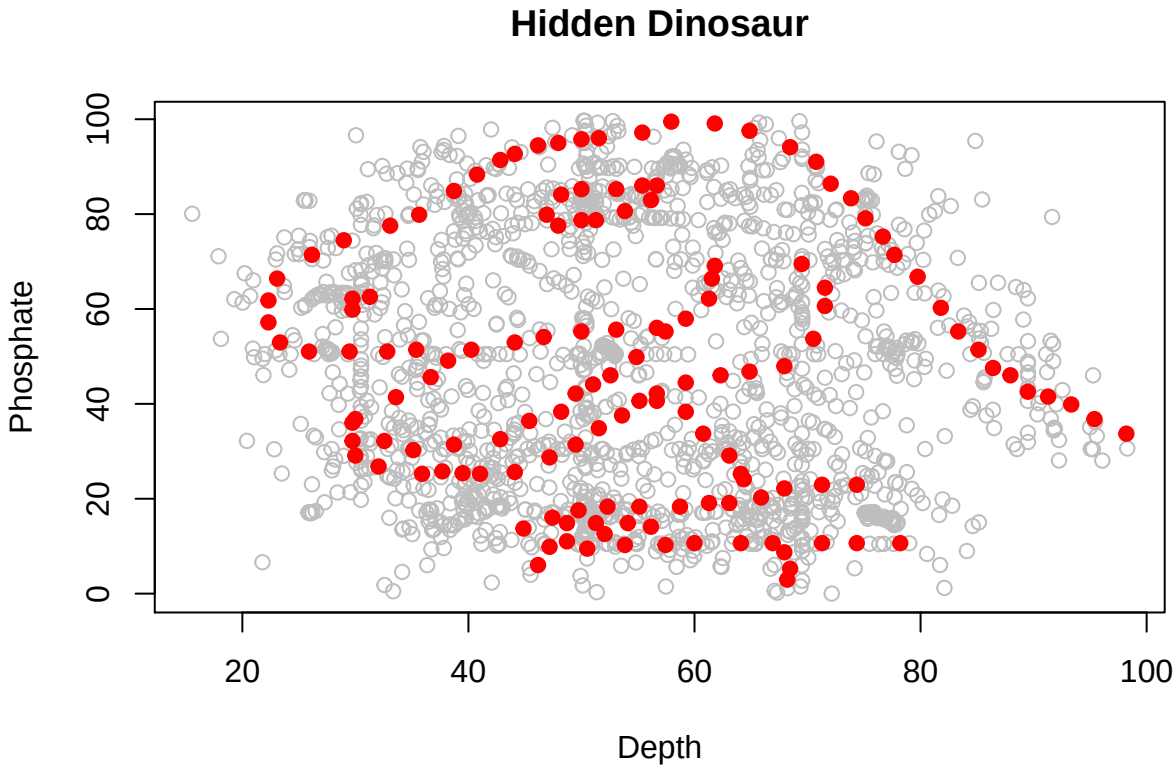
In this weeks folder, there is an Excel file containing the dataset for this problem: HW4_data.xlsx. Read this into R (see below). Make sure you get (and combine) BOTH sheets of data. Work up a well annotated reproducible exploration of this dataset using the principles discussed in class or in the supplementary material.

```
library(readxl)
Day1 <- read_excel("HW4_data.xlsx", sheet = 1) %>%
        mutate(day = as.factor("1"), block = as.factor(block))
Day2 <- read_excel("HW4_data.xlsx", sheet = 2) %>%
        mutate(day = as.factor("2"), block = as.factor(block))
hw4_data <- rbind(Day1, Day2) %>%
        select(day, block:phosphate) %>%
        arrange(day, block, depth)

summary(hw4_data)
```

```
##  day           block          depth          phosphate
##  1:1136   1       :142   Min.   :15.56   Min.   : 0.01512
##  2: 710   4       :142   1st Qu.:41.07   1st Qu.:22.56107
##           5       :142   Median :52.59   Median :47.59445
##           6       :142   Mean   :54.27   Mean   :47.83510
##           7       :142   3rd Qu.:67.28   3rd Qu.:71.81078
##           10      :142   Max.   :98.29   Max.   :99.69468
##           (Other):994
```

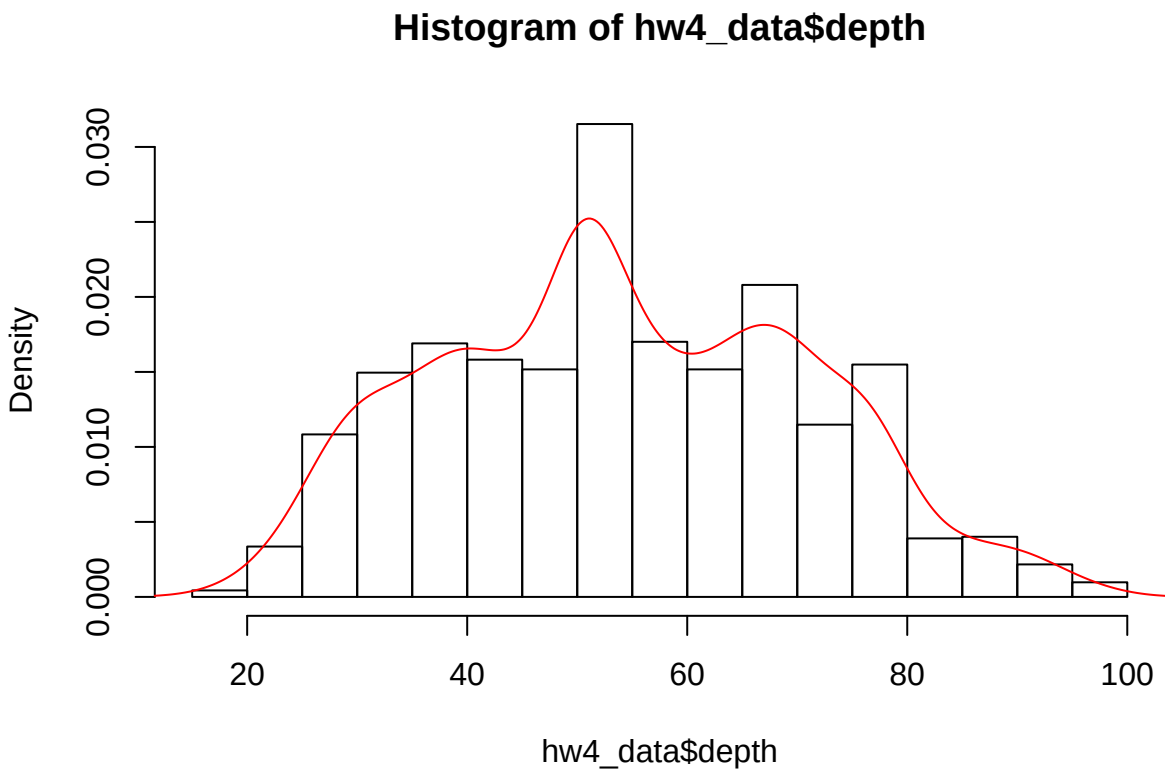
```
plot(hw4_data$depth[hw4_data$block!=4], hw4_data$phosphate[hw4_data$block!=4], col = "grey",
     main = "Hidden Dinosaur", xlab = "Depth", ylab = "Phosphate")
points(hw4_data$depth[hw4_data$block==4], hw4_data$phosphate[hw4_data$block==4], col = "red", pch = 19)
```



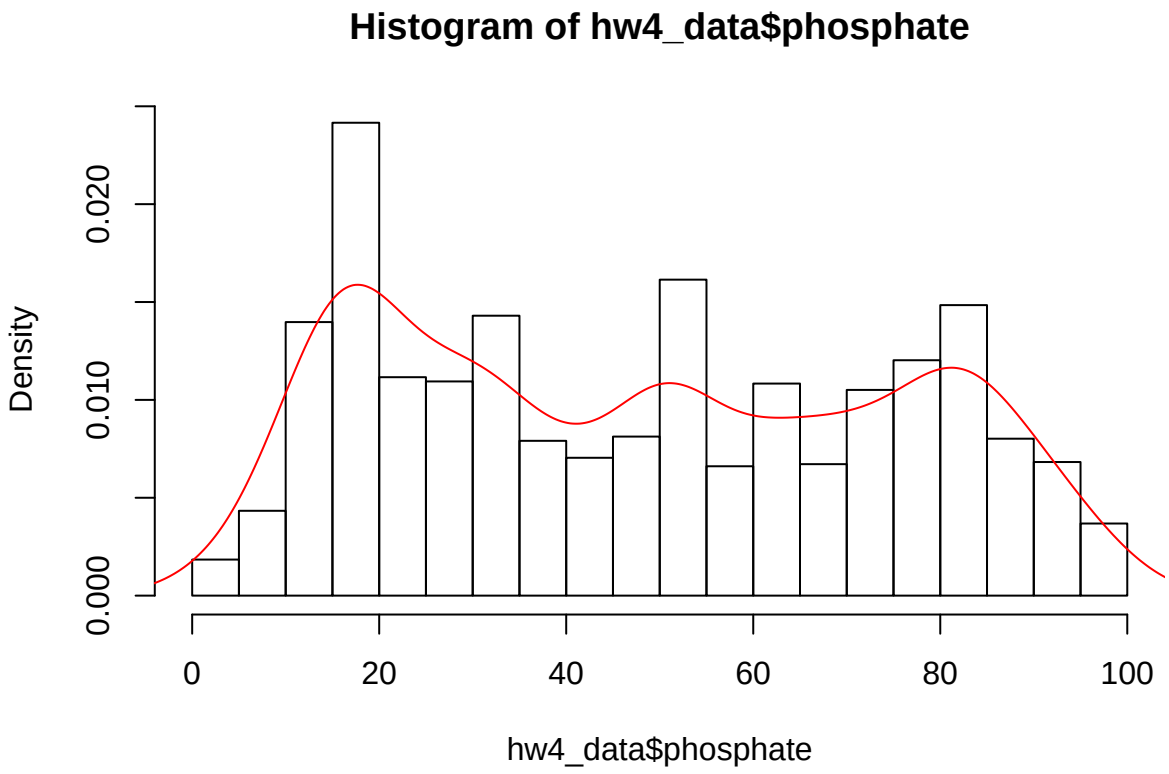

```
cor(hw4_data$depth, hw4_data$phosphate)
```

```
## [1] -0.06601891
```

```
hist(hw4_data$depth, breaks = 20, freq = F)
lines(density(hw4_data$depth), col = "red")
```

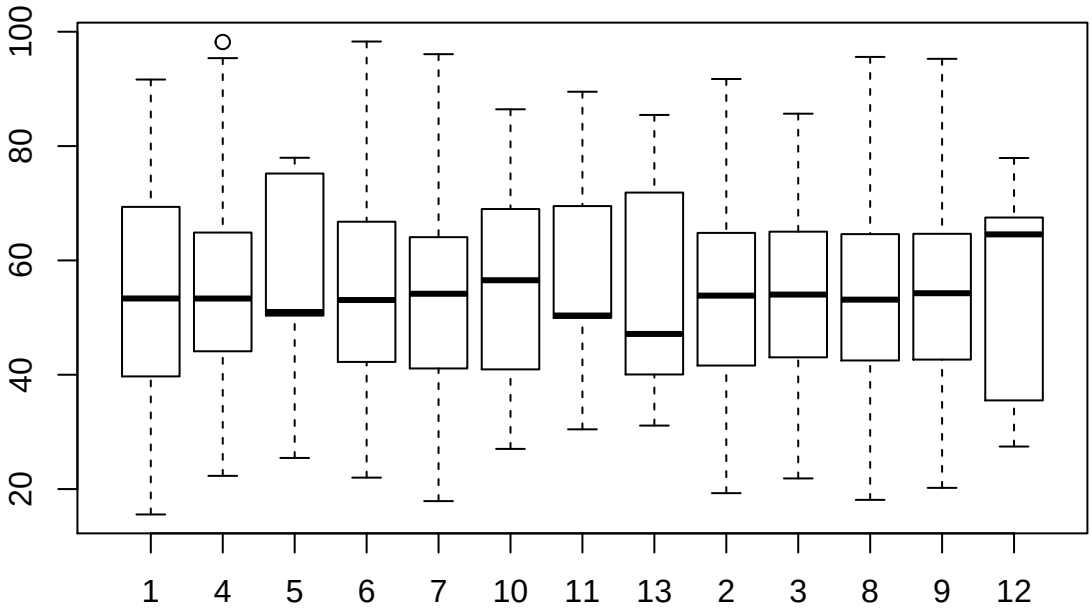



```
hist(hw4_data$phosphate, breaks = 20, freq = F)
lines(density(hw4_data$phosphate), col = "red")
```

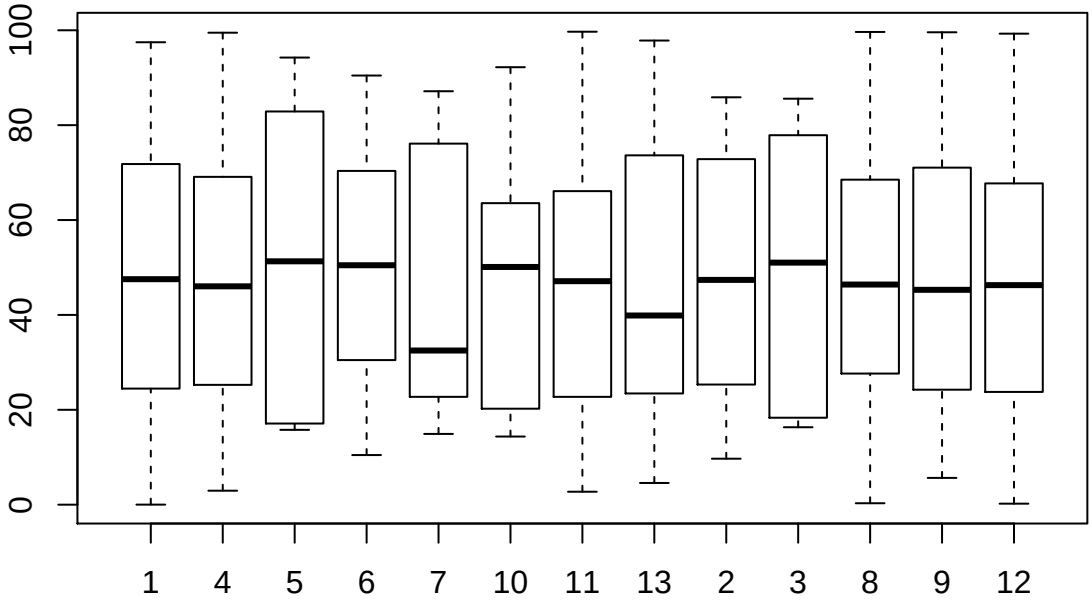



```
summby_block <- summaryBy(depth + phosphate ~ block,
                          data = as.data.frame(hw4_data), FUN = list(mean, sd))
```

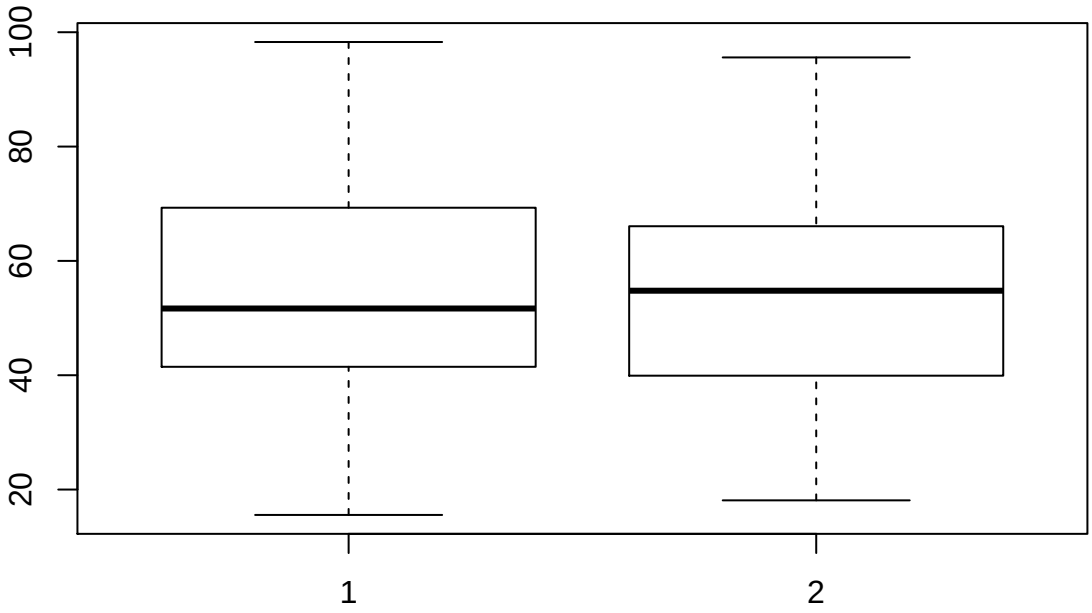
```
boxplot(depth ~ block, data = hw4_data)
```



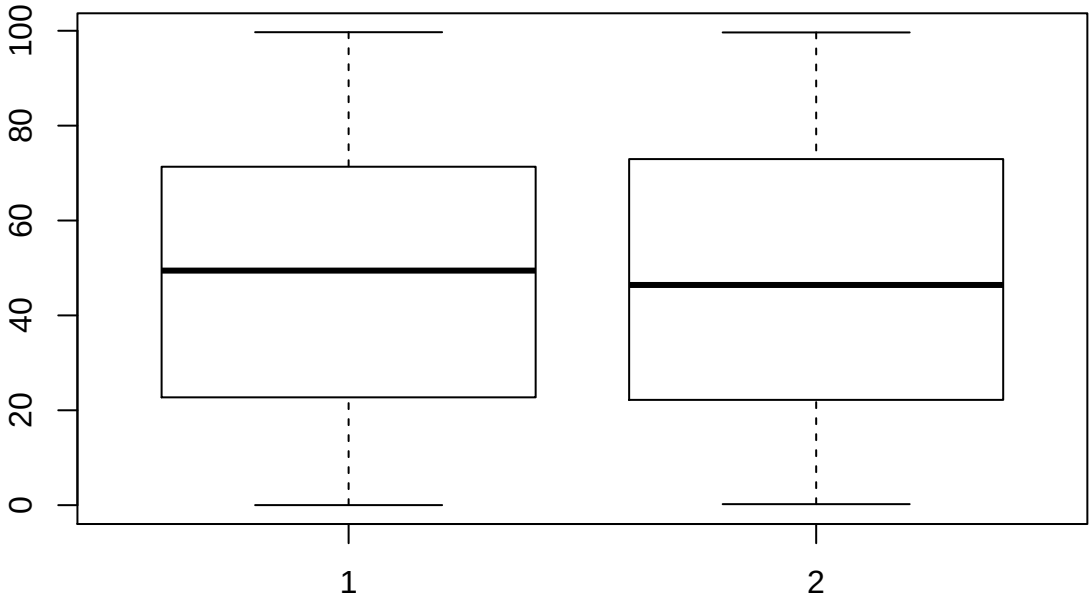

```
boxplot(phosphate ~ block, data = hw4_data)
```




```
summby_day <- summaryBy(depth + phosphate ~ day,
                         data = as.data.frame(hw4_data), FUN = list(mean, sd))
boxplot(depth ~ day, data = hw4_data)
```




```r
boxplot(phosphate ~ day, data = hw4_data)
```




```r
# par(mfrow = c(7,2))
# plot(hw4_data$depth[hw4_data$block==1], hw4_data$phosphate[hw4_data$block==1])
# plot(hw4_data$depth[hw4_data$block==2], hw4_data$phosphate[hw4_data$block==2])
# plot(hw4_data$depth[hw4_data$block==3], hw4_data$phosphate[hw4_data$block==3])
# plot(hw4_data$depth[hw4_data$block==4], hw4_data$phosphate[hw4_data$block==4])
# plot(hw4_data$depth[hw4_data$block==5], hw4_data$phosphate[hw4_data$block==5])
# plot(hw4_data$depth[hw4_data$block==6], hw4_data$phosphate[hw4_data$block==6])
# plot(hw4_data$depth[hw4_data$block==7], hw4_data$phosphate[hw4_data$block==7])
# plot(hw4_data$depth[hw4_data$block==8], hw4_data$phosphate[hw4_data$block==8])
# plot(hw4_data$depth[hw4_data$block==9], hw4_data$phosphate[hw4_data$block==9])
# plot(hw4_data$depth[hw4_data$block==10], hw4_data$phosphate[hw4_data$block==10])
# plot(hw4_data$depth[hw4_data$block==11], hw4_data$phosphate[hw4_data$block==11])
# plot(hw4_data$depth[hw4_data$block==12], hw4_data$phosphate[hw4_data$block==12])
# plot(hw4_data$depth[hw4_data$block==13], hw4_data$phosphate[hw4_data$block==13])
```

Things you need to answer/show:

1. summary statistics (combined in a table?)

2. factor exploration, what factors are present?

3. create a (couple?) multipanel plot (lattice, ggplot2, base R)
   - Base R help: http://www.statmethods.net/advgraphs/layout.html
   - GGplot2: http://www.cookbook-r.com/Graphs/Multiple_graphs_on_one_page_(ggplot2)/
4. correlation plots (?pairs, ?plotcorr)

5. This is the SAME dataset as in Problem 6 of the last homework with the factors relabeled. What is the lesson in this dataset when considering the summary statistics and the plots??

NOTE: DO use this version and NOT the previous version, i.e. load using the xlsx package as above so the factors are correctly labeled.

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?

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

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

Problem 6

Get a dataset, choose ONE:

A. Current Employment Statistics

```
##http://bradleyboehmke.github.io/2015/12/scraping-html-tables.html
library(rvest)
webpage <- read_html("http://www.bls.gov/web/empsit/cesbmart.htm")
tbls <- html_nodes(webpage, "table")
tbls_ls <- webpage %>%
        html_nodes("table") %>%
        .[c(2:5,16:18)] %>%
        html_table(fill = TRUE)

# scrape Table 2. Nonfarm employment...
tbls2_ls$Table1 <- webpage %>%
        html_nodes("#Table2") %>%
        html_table(fill = TRUE) %>%
        .[[1]]
```

```
# Table 3. Net birth/death...
tbls2_ls$Table2 <- webpage %>%
        html_nodes("#Table3") %>%
        html_table() %>%
        .[[1]]
str(tbls2_ls)
```

B. H1b filings

```
library(jsonlite)
library(rvest)
library(pbapply)
library(data.table)

json.cities<-paste0('http://h1bdata.info/cities.php?term=', letters)
all.cities<-unlist(pblapply(json.cities,fromJSON))
city.year<-expand.grid(city=all.cities,yr=seq(2014,2017))
city.year$city<-urltools::url_encode(as.character(city.year$city))
all.urls<-paste0('http://h1bdata.info/index.php?em=&job=&city=', city.year[,1],'&year=', city.year[
getData<-function(url.x){
    x<-read_html(url.x)
    x<-html_table(x)
    x<-data.table(x[[1]])
    return(x)
    Sys.sleep(5)
}
all.h1b<-pblapply(all.urls, getData)
all.h1b<-rbindlist(all.h1b)
saveRDS(all.h1b,'h1b_data.rds', row.names=F)
```

From here, this one is all on you. Do a full on EDA. Note that in MANY situations, interesting features suggesting bringing in additional data, if you feel it, do it. On ANY aspect of that dataset you think is interesting or would make a good plot. Annotate and describe your process and thinking. End with a figure you think best describes some aspect/anomaly/feature of the dataset. This will be your second submission in next weeks contest. Again, save it as pdf and be ready to push it next week.

There could be an opportunity to create a map. Quick and not so dirty:

```
#https://cran.r-project.org/web/packages/fiftystater/vignettes/fiftystater.html
library(ggplot2)
library(fiftystater)

data("fifty_states") # this line is optional due to lazy data loading
crimes <- data.frame(state = tolower(rownames(USArrests)), USArrests)
# map_id creates the aesthetic mapping to the state name column in your data
p <- ggplot(crimes, aes(map_id = state)) +
  # map points to the fifty_states shape data
  geom_map(aes(fill = Assault), map = fifty_states) +
  expand_limits(x = fifty_states$long, y = fifty_states$lat) +
  coord_map() +
  scale_x_continuous(breaks = NULL) +
  scale_y_continuous(breaks = NULL) +
  labs(x = "", y = "") +
  theme(legend.position = "bottom",
        panel.background = element_blank())
```

```
p
ggsave(plot = p, file = "HW4_Problem6_Plot_Settlage.pdf")
```

Problem 7

DO NOT push this homework until the next class! Don't give your classmates any hints on your cool graphics.

When it is time to submit, –ONLY– submit the .Rmd and .pdf solution files. Names should be formatted HW4_lastname_firstname.Rmd

Optional preperation for next class:

Next week we will talk about the apply family of functions. Check out Swirl "R_programming_E" Swirl lessons 10 and 11.