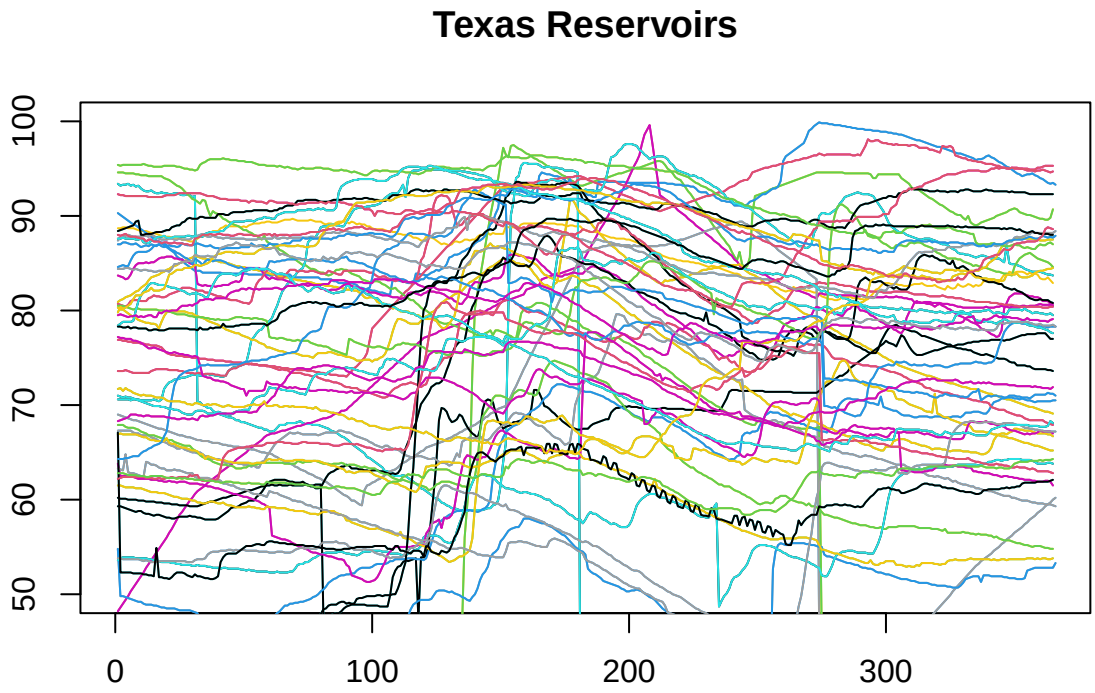
CSI 2300: Intro to Data Science

In-Class Exercise 12: Basic Programming

1. In the Texas reservoirs example from the lecture, the loop is indexed directly over the years from 1971 to 2020. Reprogram the loop to index it over the values 1:50. To recall, the data is read in as follows:

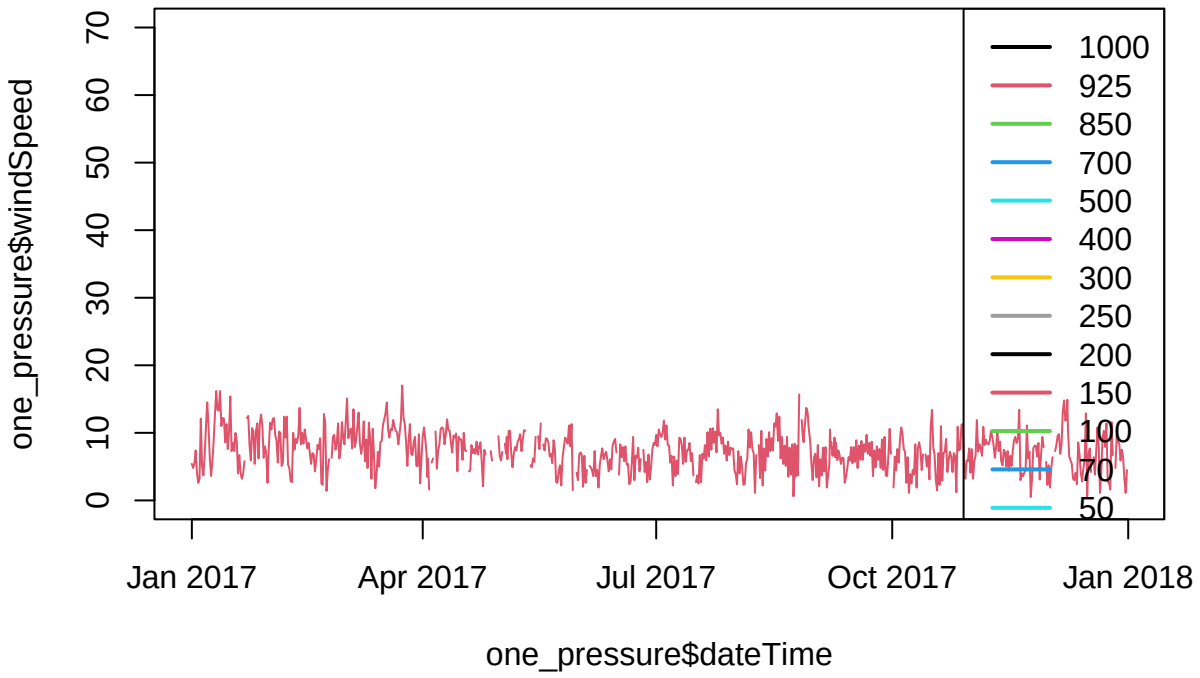
```
suppressMessages(library(lubridate))
www <- "https://www.waterdatafortexas.org/reservoirs/statewide.csv"
water <- read.csv(file=www, header=T, skip=29)
water_year <- year(water$date)

plot(1:365, seq(50, 100, len=365), type="n", xlab='', ylab='', main="Texas Reservoirs")


for(i in 1933:1983){
    one_year <- water[water_year == i, ]
    lines(1:nrow(one_year), one_year$percent_full, col=i)
}

years <- 1933:1983
for (i in 1:50){
    one_year <- water[water_year == years[i], ]
    lines(1:nrow(one_year), one_year$percent_full, col=i)
}
```



2. Load the Corpus Christi radiosonde record, `CorpusCristi.rda`. It will appear in your global environment as `CorpusChristi`. In the in-class exercises for lecture 11, you plotted the wind speed over time at pressure level 925. Create a plot where you overlay the wind speed over the year with each pressure level plotted in a different color. The range on the y-axis should be set to be between 0 and 70.

```
load('dat/CorpusCristi.rda')

pressures <- unique(CorpusCristi$pressure)

colors_by_press <- color.scale(pressures, col = viridis(length(press_unique)))
# Error in color.scale(pressures, col = viridis(length(press_unique))): could not find

one_pressure <- CorpusCristi[which(CorpusCristi$pressure == "1000"),]

plot(one_pressure$dateTime, one_pressure$windSpeed, type="l", col=i, ylim=c(0,70))

for (i in 2:length(pressures)){
    one_pressure <- CorpusCristi[which(CorpusCristi$pressure == pressures[i]),]
    lines(one_pressure$dateTime, one_pressure$windSpeed, type="l", col=colors_by_press)
}
```

```r
# Error: object 'colors_by_press' not found

for (i in 2:10){
    one_pressure <- CorpusCristi[which(CorpusCristi$pressure == pressures[i]),]
    lines(one_pressure$dateTime, one_pressure$windSpeed, type="l", col=colors_by_press)
}
# Error: object 'colors_by_press' not found

for (i in 10:12){
    one_pressure <- CorpusCristi[which(CorpusCristi$pressure == pressures[i]),]
    lines(one_pressure$dateTime, one_pressure$windSpeed, type="l", col=colors_by_press)
}
# Error: object 'colors_by_press' not found

legend('topright', legend=pressures, col=1:17, lwd=2)
```



3. Answer the following questions about the plot that you created in the prior question.

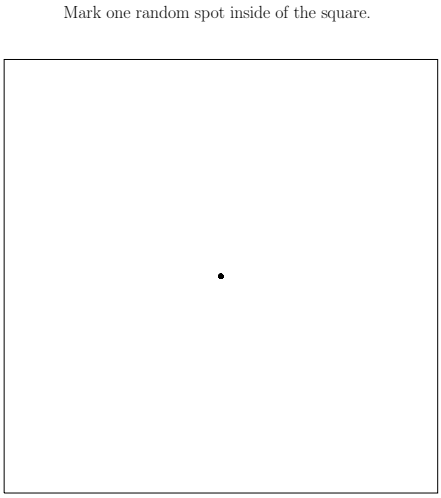
- Do the strongest winds occur in the upper or lower atmosphere? Note, 1000 mb is closest to the earth's surface, and 7 mb is farthest away.

- What time of year are the winds the weakest?

The lower atmosphere. The winds are the weakest in the summer months. October 2017 is particularly low.

- Does the spike due to the hurricane on August 26, 2017 affect all of the pressure levels?
- Would the spike in wind speeds due to the hurricane be considered a global or a local outlier?
- Use the `which.max` command to find the observation whose wind speed is the highest. What is this wind speed, and at what pressure does it occur, and on what date does it occur?

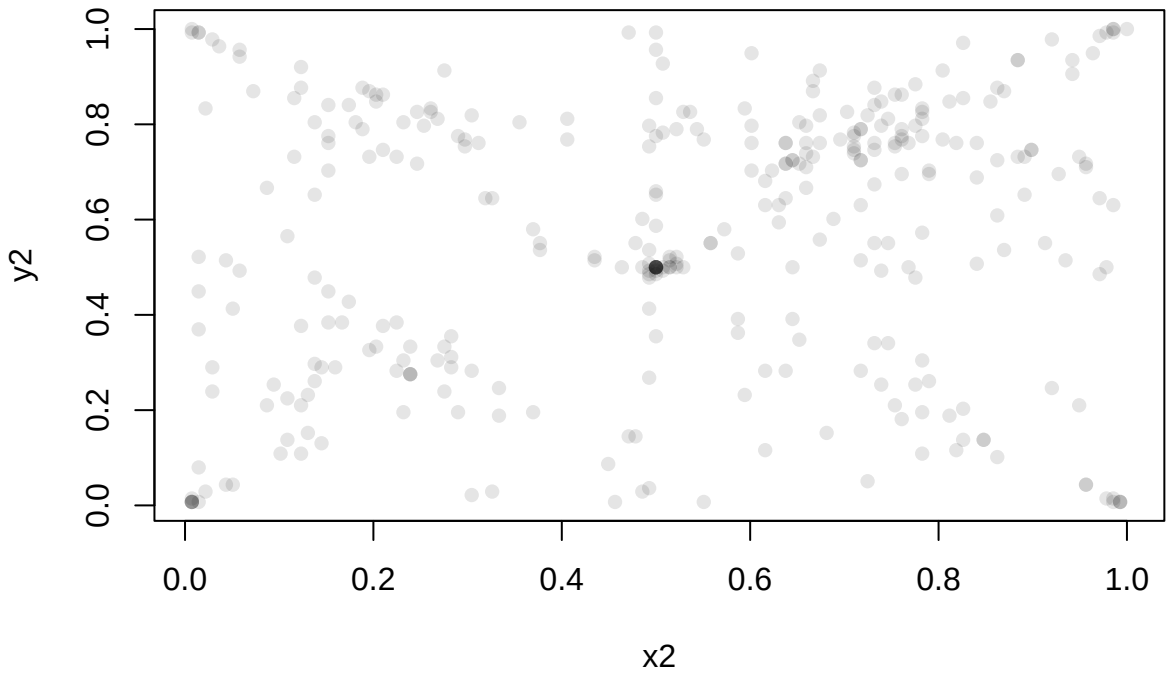
4. In the first data collection asking students to "Mark one random spot inside of the square," the square was completely empty. For a second experiment, a landmark dot was placed in the center of the square, and students were asked again to "Mark one random spot inside of the square," as seen in the figure below.



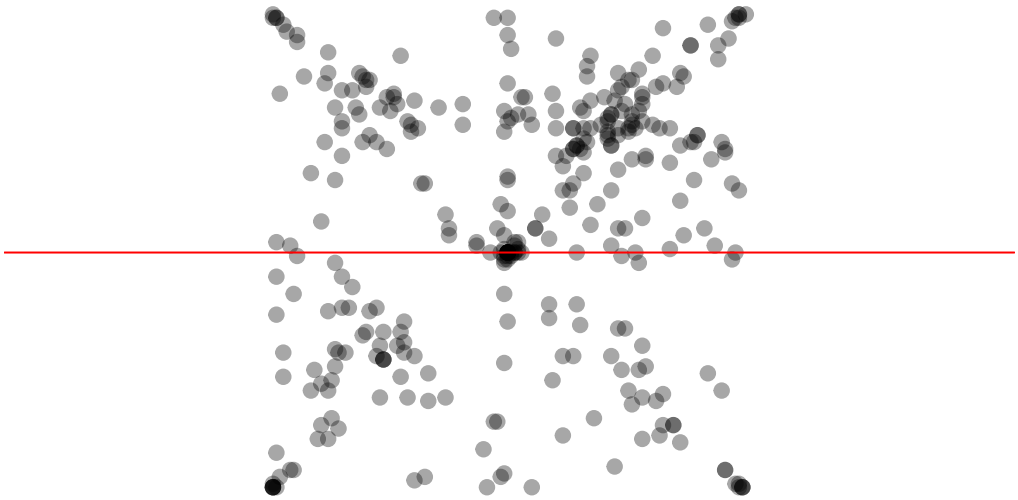
Data collection survey with landmark dot in the center.

The data can be loaded using the code below. Plot the data, and comment on whether or not it appears that the presence of a dot in the center of the square caused students to choose points elsewhere. Alpha-blending for plotting the points will be essential here.

```r
suppressWarnings(suppressMessages(library(spatstat)))
data <- read.csv(file="dat/dot_experiment_data.csv", header=T)
center  <- which(data$type=="centerDot")
#Rescaling data to be on the unit square
x2 <- data$x[center]/13.8
y2 <- data$y[center]/13.8

plot(x2,y2, pch=16 ,col=rgb(0,0,0,.1))
```




```r
plot(x2, y2, col = rgb(0,0,0,.35), pch=19, bty="n", xaxt="n", yaxt="n", asp='1',xlab="",
abline(h=0.5, col='red')
```



5. Create an object of type ppp using the code below.

```
win <- owin(c(0, 1), c(0, 1))
data_dot <- ppp(x2, y2, window = win)
# Warning: data contain duplicated points
```

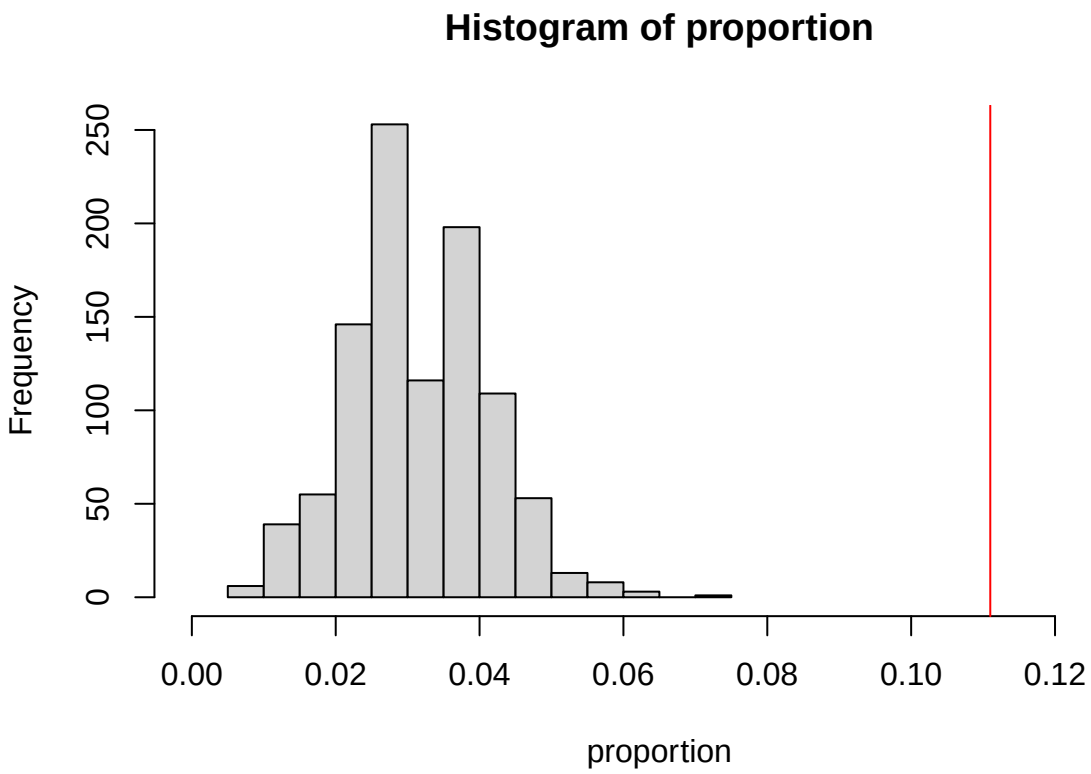
What proportion of the points are within 0.1 units of the center of the square? The center of the square is (0.5, 0.5). You will need to use the distance formula.

```
nn <- length(x2)

how_far <- sqrt((x2 - 0.5)^2 + (y2 - 0.5)^2)

sum(how_far < 0.1)/nn
# [1] 0.1117647
```

6. In this question, you will do a Monte Carlo study in which you compute the proportion of points within 0.1 units of the center of a square. You will repeat this many times and compare the results with the observed data. Complete the following steps:

- Create an empty variable to hold a value.
- Create a `for()` loop to iterate through 1000 times.
- Inside the `for()` loop, simulate a dataset with 340 observations distributed randomly in the unit square using the command `sim_dat <- runifpoint(nn, nsim = 1)`. You can extract the x and y coordinates using `sim_dat$x$` and `sim_dat$y`.
- Compute and save the proportion of points that are within 0.1 units of the center.
- Plot a histogram of these proportions.
- Overlay the proportion of points that were observed in the data with a red vertical line.
- Is the proportion in the observed data similar to or different than points that are randomly distributed in the square?

```
proportion <- NULL
for (i in 1:1000){
    sim_dat <- runifpoint(nn, nsim = 1)
    x <- sim_dat$x
    y <- sim_dat$y
    how_far <- sqrt((x - 0.5)^2 + (y - 0.5)^2)
    proportion[i] <- length(which(how_far < 0.1)) / 340
}

hist(proportion, xlim = c(0,.13))
abline(v=0.111, col='red')
```



The proportion in the observed data is higher than the proportion in the randomly distributed data.