Creating Visualizations in R

Paul Brooks

February 24, 2017

Don't forget to set your working directory (Session -> Set Working Directory)!

Introduction

Read in the SENIC data. Remove the last (extra) column.

```
senicData <- read.table("C:/Users/Larry/OneDrive/Documents/GitHub/Test02/SENIC.csv", header=TRUE, row.na
                        colClasses=c(rep("numeric",7),rep("factor",2),
                        rep("numeric",9), rep("factor",2)))
senicData <- senicData[,-ncol(senicData)]
```

Read in the Lending Club data. The data will be in a data frame called lendingData.

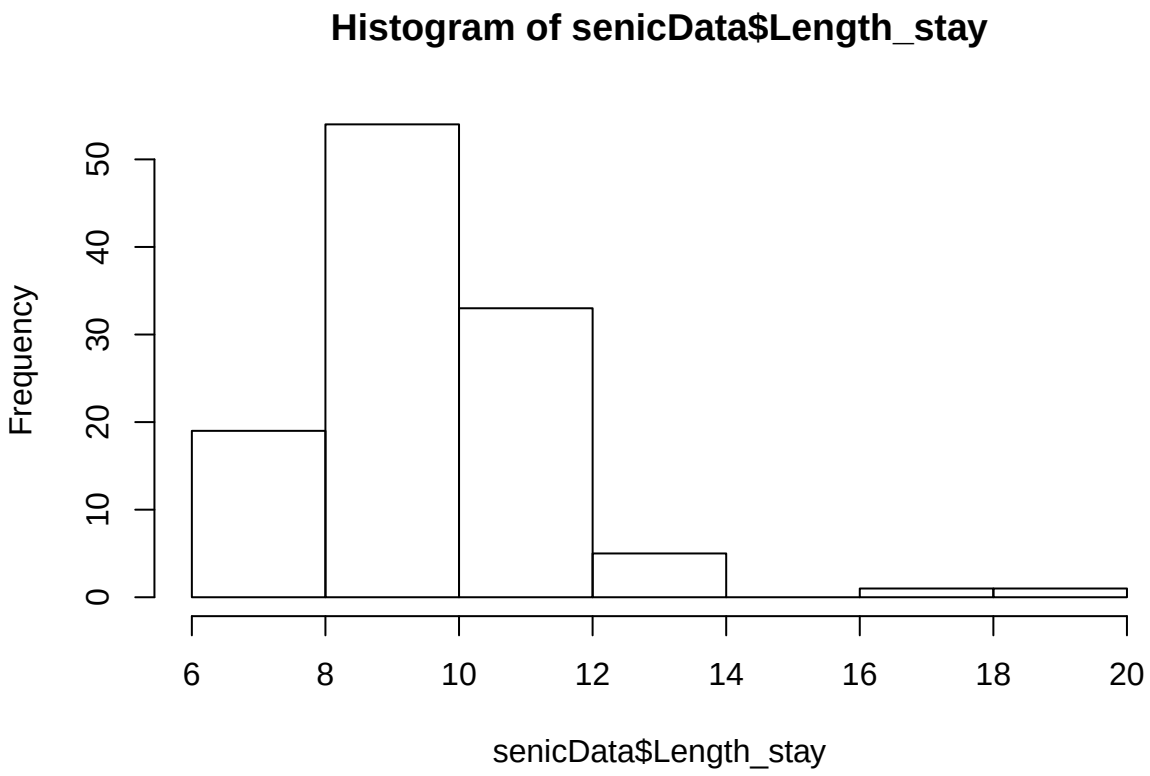
```
load("C:/Users/Larry/OneDrive/Documents/GitHub/Test02/lendingData.rda")
```

Univariate Visualizations

Histogram

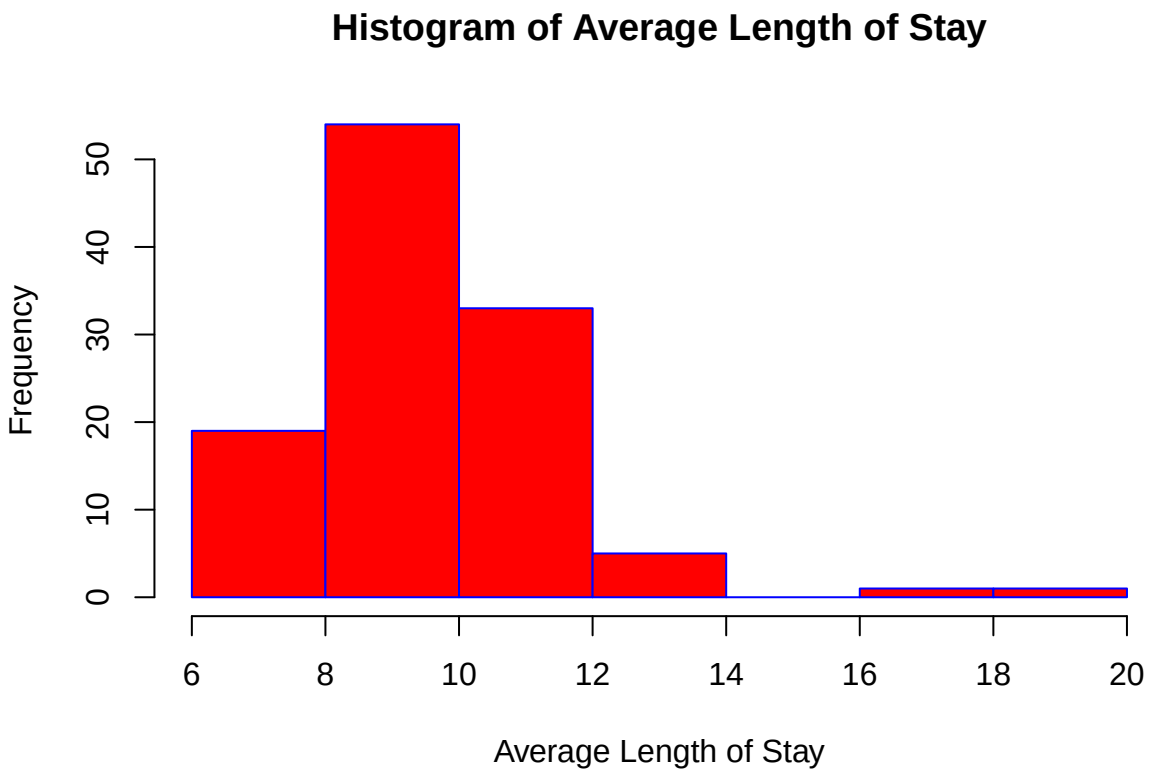
Create a histogram of the average length of stay. The distribution of average lengths is right-skewed.

```
hist(senicData$Length_stay)
```



Add some color, labels, and a title.

```
hist(senicData$Length_stay, col="red",border="blue",
     xlab="Average Length of Stay", main="Histogram of Average Length of Stay")
```



Turn off the graphics device.

```
dev.off()
```

Save the figure as a pdf.

```
pdf("lengthHist.pdf")
hist(senicData$Length_stay, col="red",border="blue",
     xlab="Average Length of Stay", main="Histogram of Average Length of Stay")
dev.off()
```

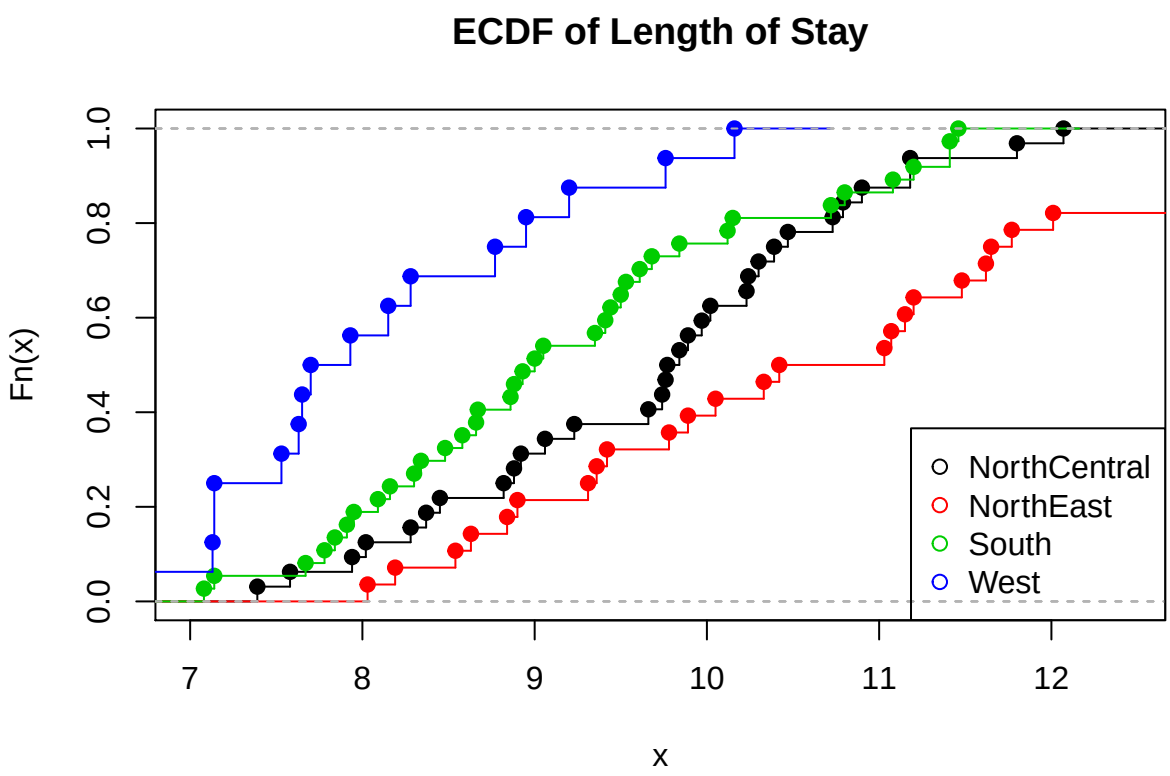
Save the figure as a png file. Note that there are differences in the options for setting the figure size for *pdf()* (inches) and *png()* (pixels).

```
png("lengthHist.png")
hist(senicData$Length_stay, col="red",border="blue",
     xlab="Average Length of Stay", main="Histogram of Average Length of Stay")
dev.off()
```

Empirical Cumulative Distribution Functions

```
plot(ecdf(senicData$Length_stay[senicData$Region_Name == "NorthCentral"]),
     verticals=TRUE, main="ECDF of Length of Stay")
lines(ecdf(senicData$Length_stay[senicData$Region_Name == "NorthEast"]), col=2,
      verticals=TRUE)
lines(ecdf(senicData$Length_stay[senicData$Region_Name == "South"]), col=3,
      verticals=TRUE)
```

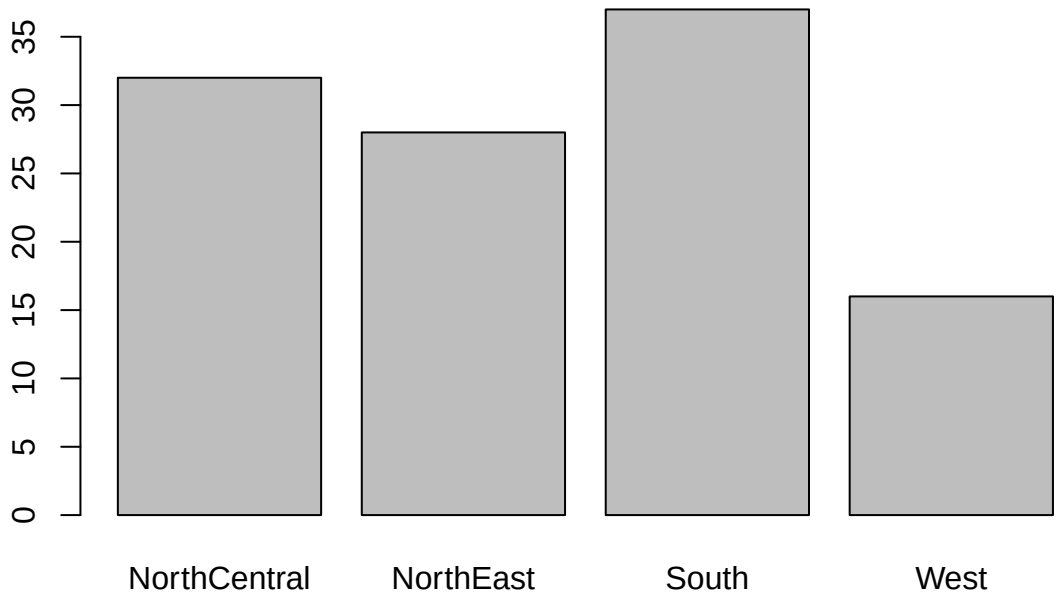
```
lines(ecdf(senicData$Length_stay[senicData$Region_Name == "West"]), col=4,
      verticals=TRUE)
legend("bottomright", c("NorthCentral", "NorthEast", "South", "West"),
       col=c(1,2,3,4), pch=1)
```



Bar Plot

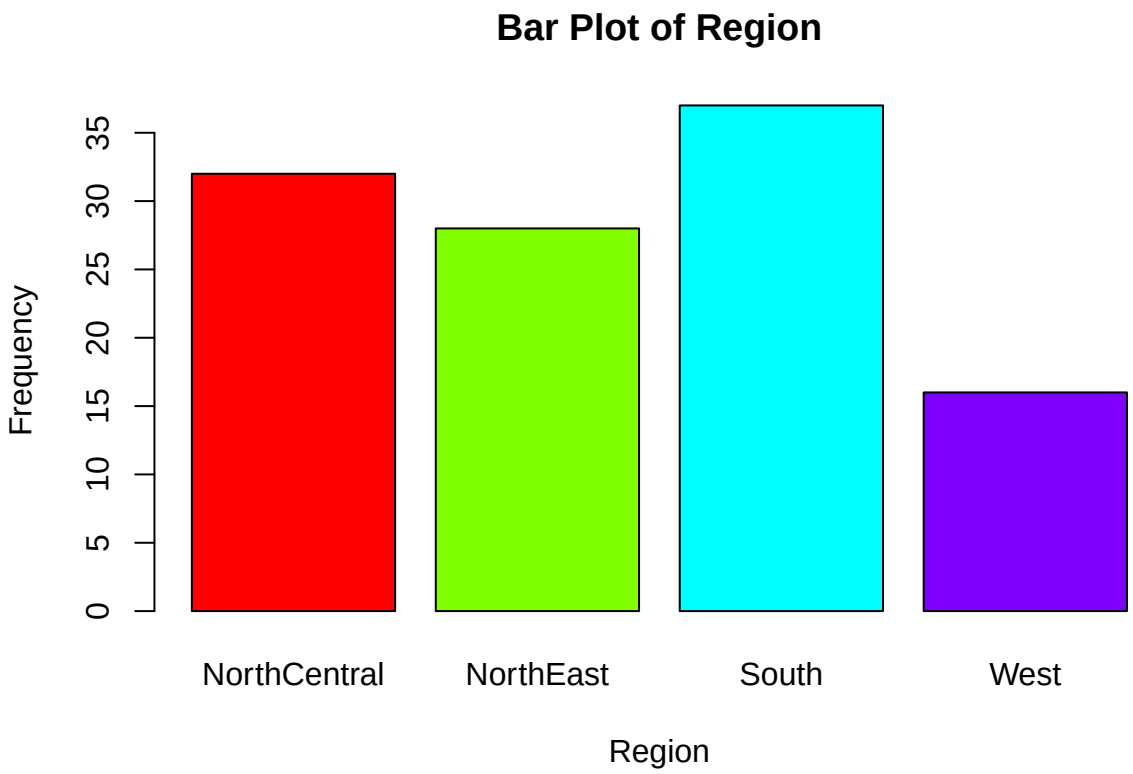
Create a frequency table and bar plot of observations by region.

```
regionCounts <- table(senicData$Region_Name)
barplot(regionCounts)
```



Add axis labels, color, and a title

```
barplot(regionCounts, xlab="Region", ylab="Frequency",
        col=rainbow(unique(senicData$Region_Name)), main="Bar Plot of Region")
```

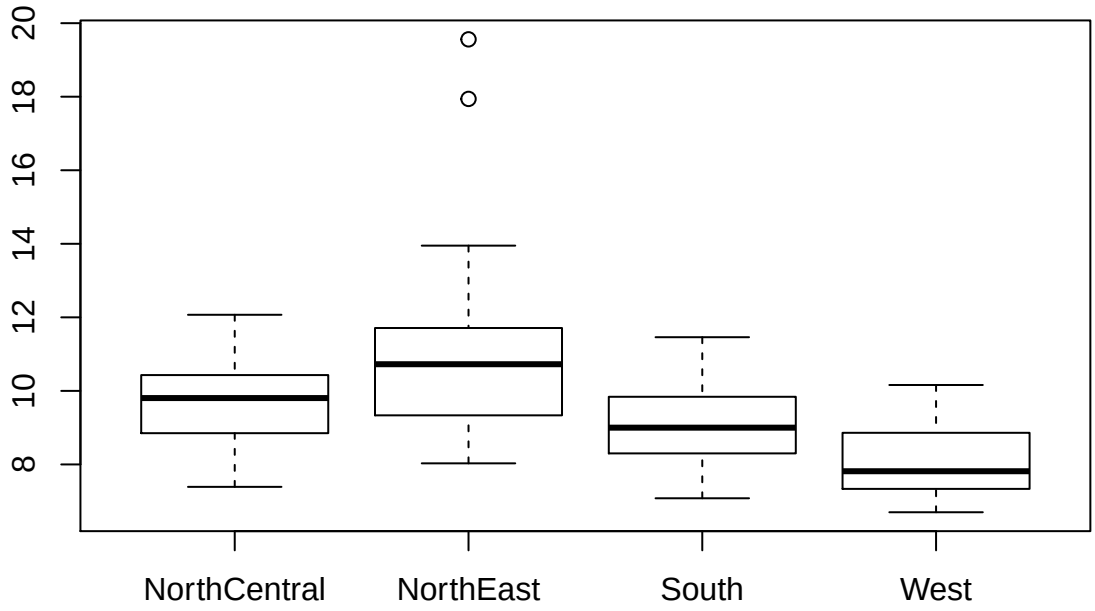


Vizualizing Relationships among Multiple Variables

Box and Whisker Plot

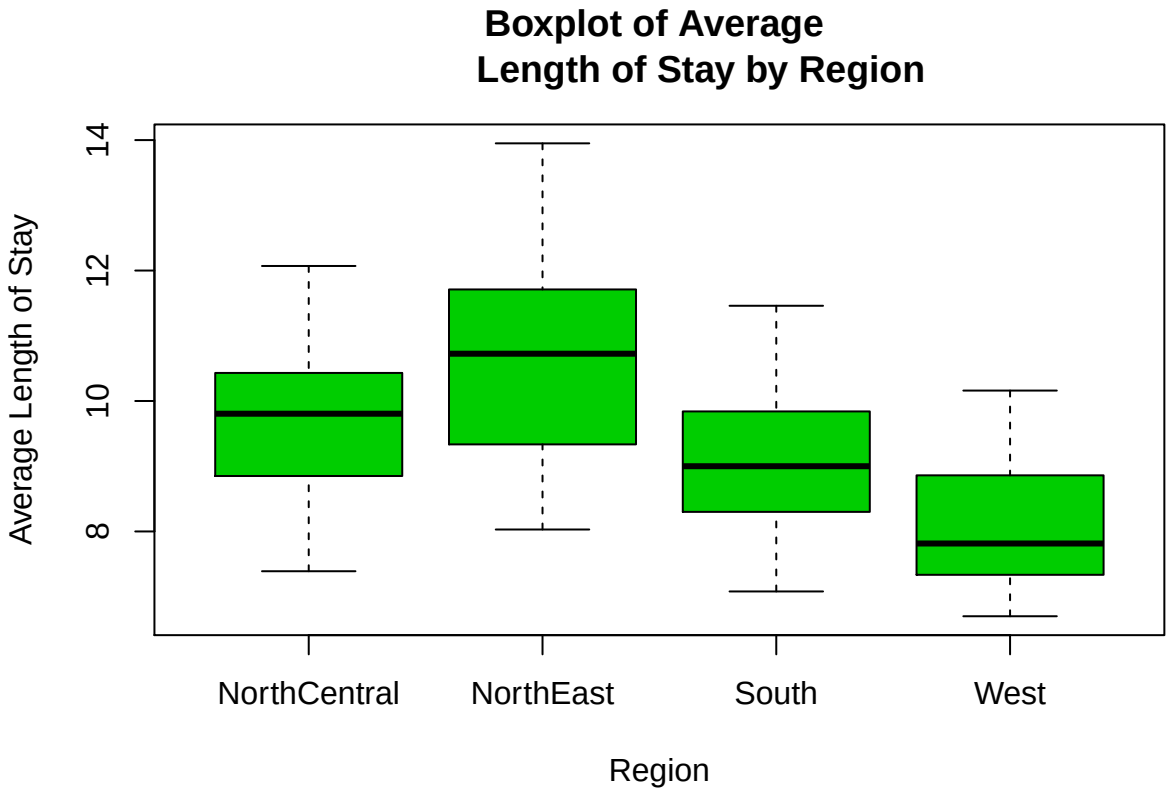
Create a box plot showing the median, interquartile range, and outliers for average length of stay by region.

```
boxplot(senicData$Length_stay ~ senicData$Region_Name)
```



Suppress the plotting of outliers, add labels, color, and a title.

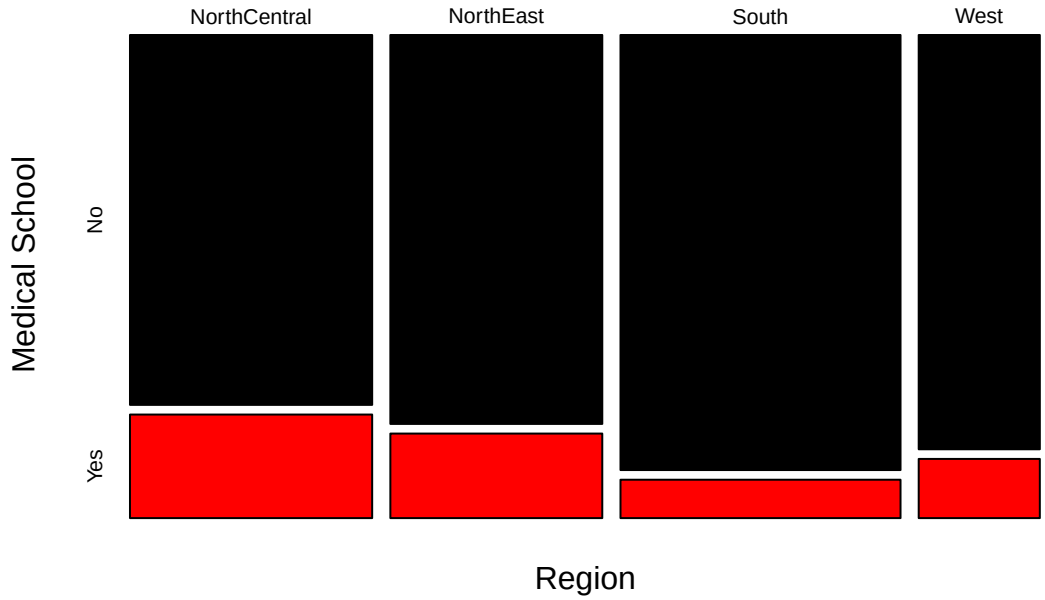
```
boxplot(senicData$Length_stay ~ senicData$Region_Name, xlab="Region",
        ylab="Average Length of Stay", main = "Boxplot of Average
        Length of Stay by Region", col=3, outline=FALSE)
```



Mosaic Plot

Create a table of region and medical school affiliation (Yes/No), and a mosaic plot.

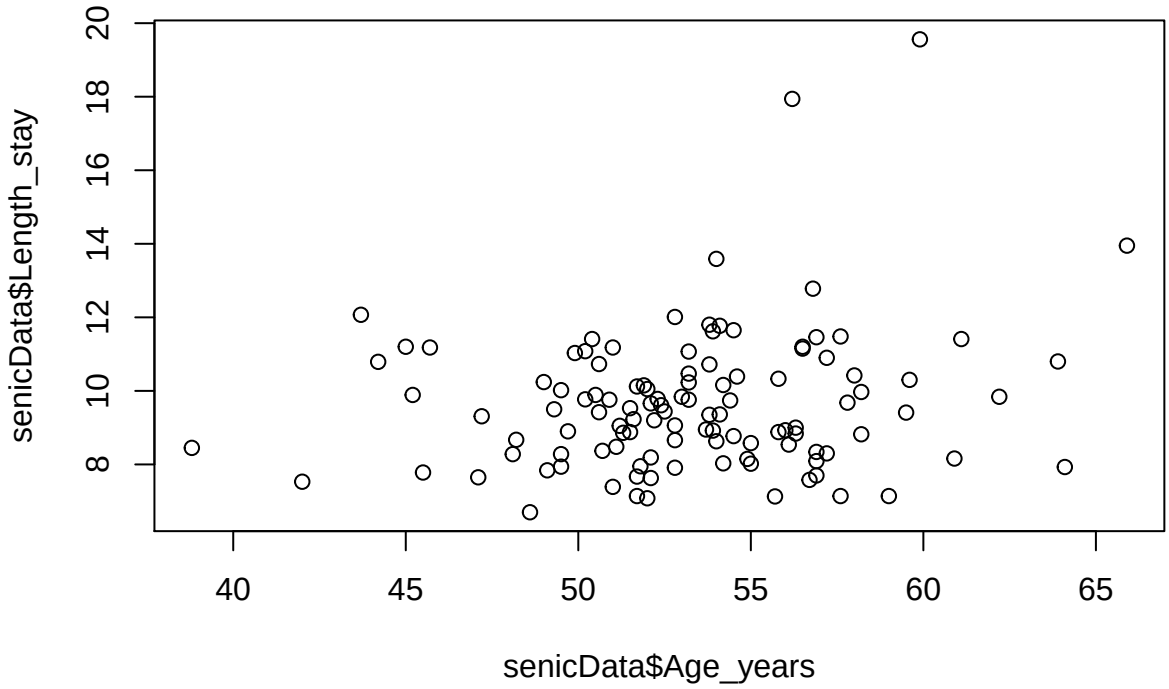
```
medschoolRegionTable <- table(senicData$Region_Name, senicData$Medical_School)
mosaicplot(medschoolRegionTable, color=c(1:2), xlab="Region",
           ylab="Medical School",main="")
```



Scatter Plot

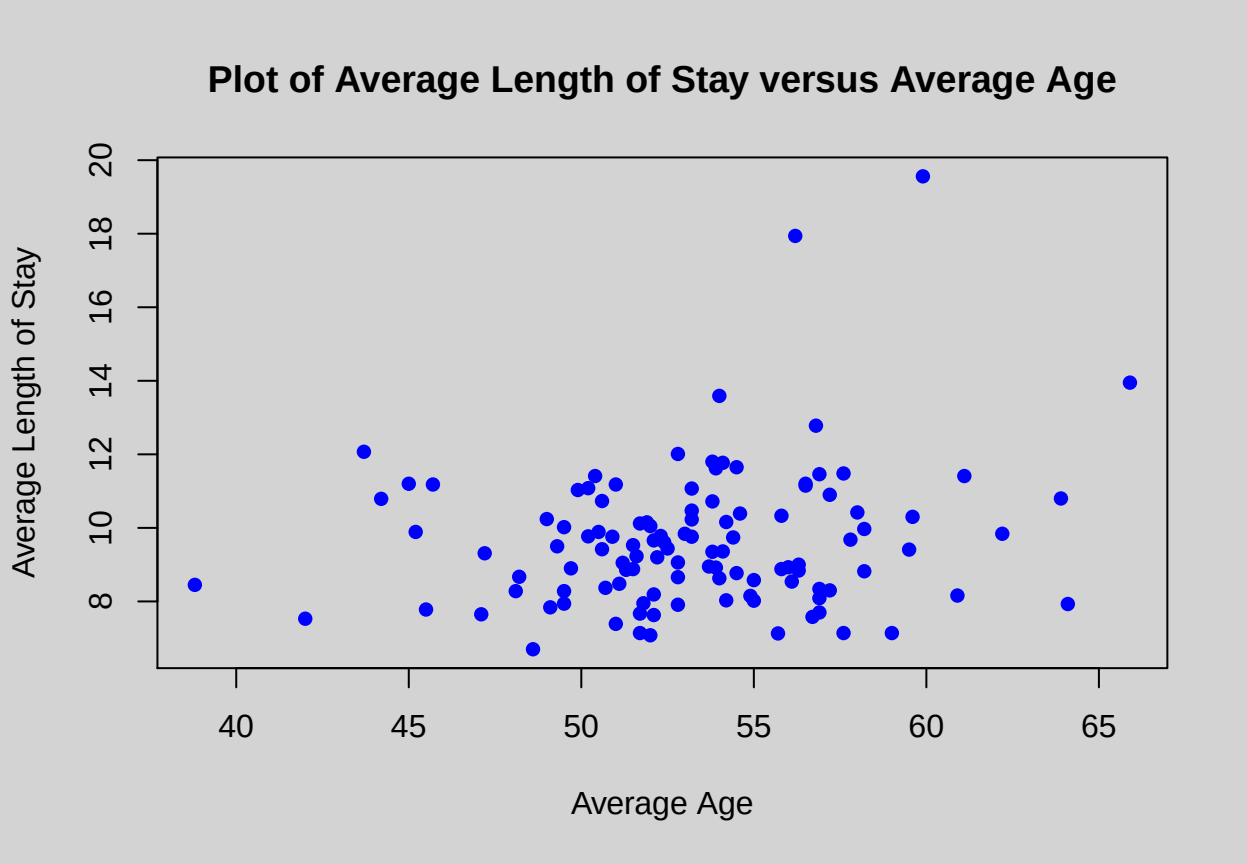
Plot average length of stay versus average age. Note that you specify the x-axis variable first.

```
plot(senicData$Age_years, senicData$Length_stay)
```



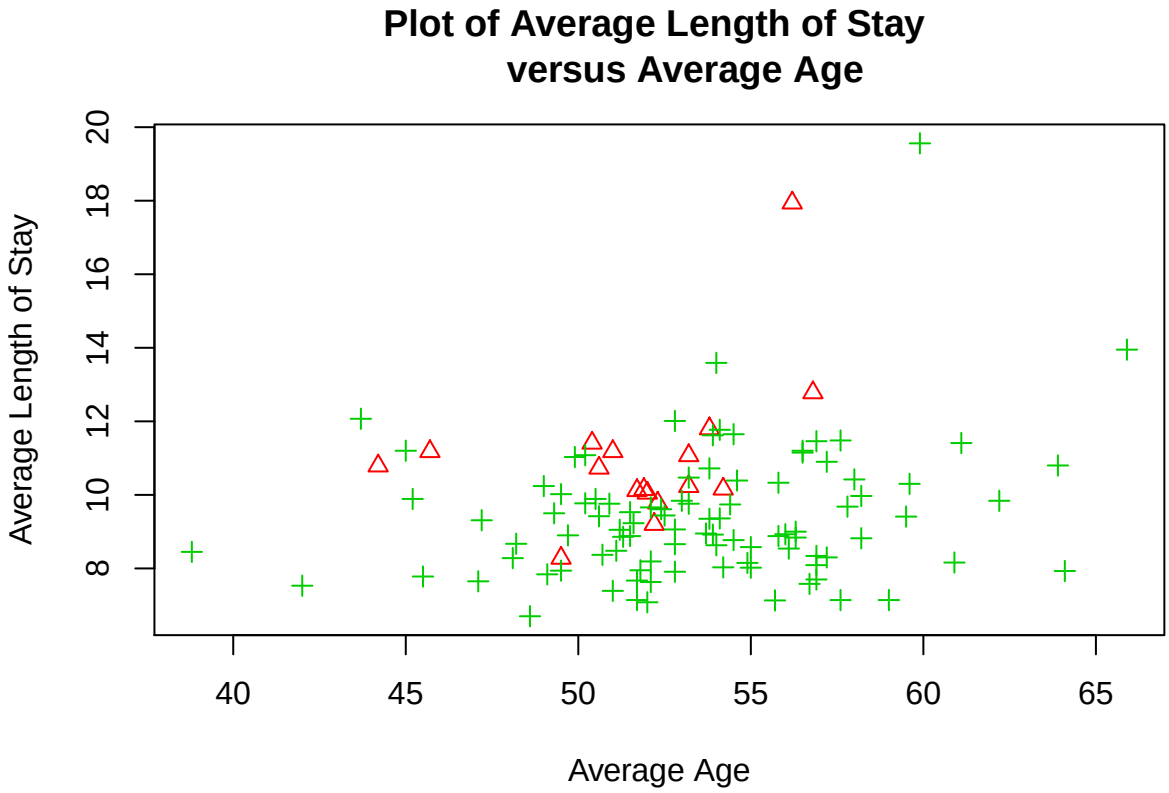
Add labels, color, and change the plotting symbol. *par()* is a function that controls many plotting options. Options specified here hold for all plots in an R window/graphic device until *dev.off()* is called. Change the background color to light gray.

```
par(bg="lightgray")
plot(senicData$Age_years, senicData$Length_stay, xlab="Average Age",
     ylab="Average Length of Stay", col="blue", pch=16,
     main="Plot of Average Length of Stay versus Average Age")
```



Create a scatter plot of average length of stay versus average age with colors indicating if the hospital has an affiliated medical school. The type="n" option creates a blank plot, but sets the axis scales so that all of the data will appear on the plot. The *points()* function adds points to the plot.

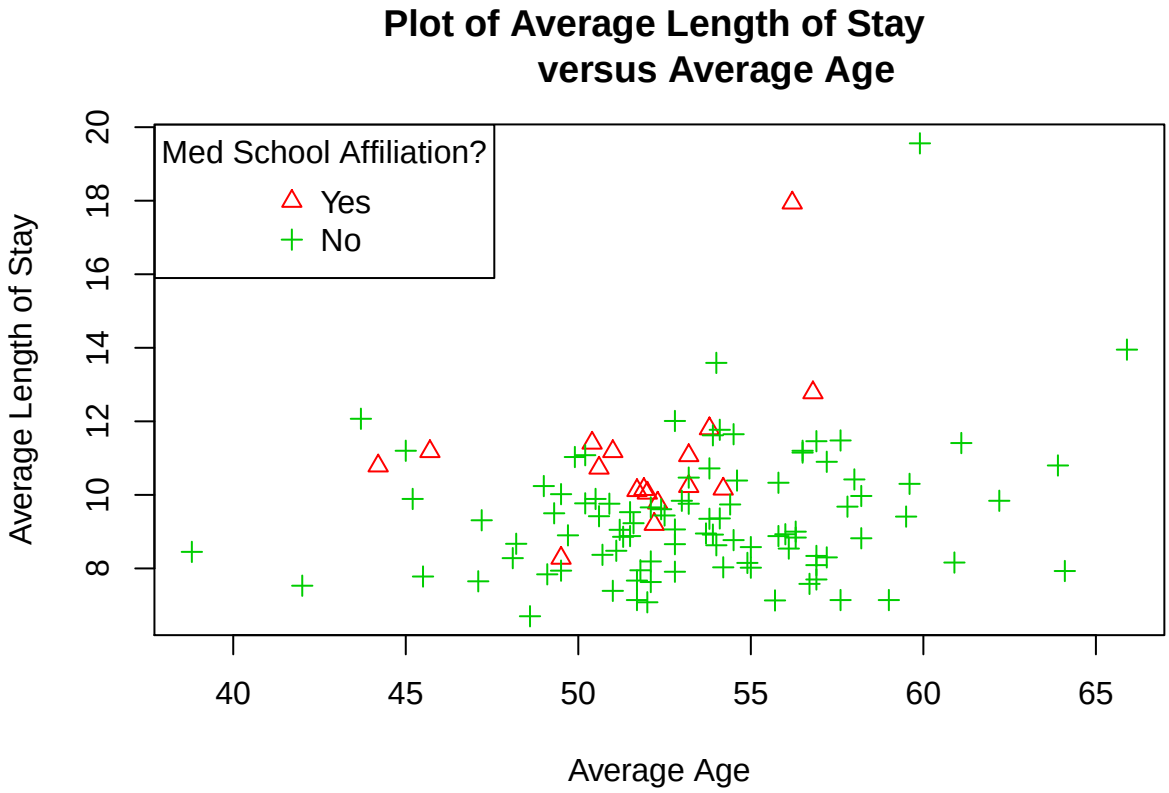
```
plot(senicData$Age_years, senicData$Length_stay, type="n", xlab="Average Age",
     ylab="Average Length of Stay", main="Plot of Average Length of Stay
     versus Average Age")
points(senicData$Age_years[senicData$Medical_School=="Yes"],
       senicData$Length_stay[senicData$Medical_School == "Yes"], col=2, pch=2)
points(senicData$Age_years[senicData$Medical_School=="No"],
       senicData$Length_stay[senicData$Medical_School == "No"], col=3, pch=3)
```



Add a legend. RStudio does not work well with sizing legends. One solution is to call *windows()* on Windows machines or *x11()* on Mac/Linux to open a graphics window. You can specify the size of the window in each function or adjust it interactively, and the plotting commands will execute based on that size.

```
windows()
```

```
plot(senicData$Age_years, senicData$Length_stay, type="n", xlab="Average Age",
     ylab="Average Length of Stay", main="Plot of Average Length of Stay
          versus Average Age")
points(senicData$Age_years[senicData$Medical_School=="Yes"],
       senicData$Length_stay[senicData$Medical_School == "Yes"], col=2, pch=2)
points(senicData$Age_years[senicData$Medical_School=="No"],
       senicData$Length_stay[senicData$Medical_School == "No"], col=3, pch=3)
legend("topleft", title="Med School Affiliation?", legend=c("Yes", "No"),
       col=c(2,3), pch=c(2,3))
```



For 3D plots, install the *rgl* package.

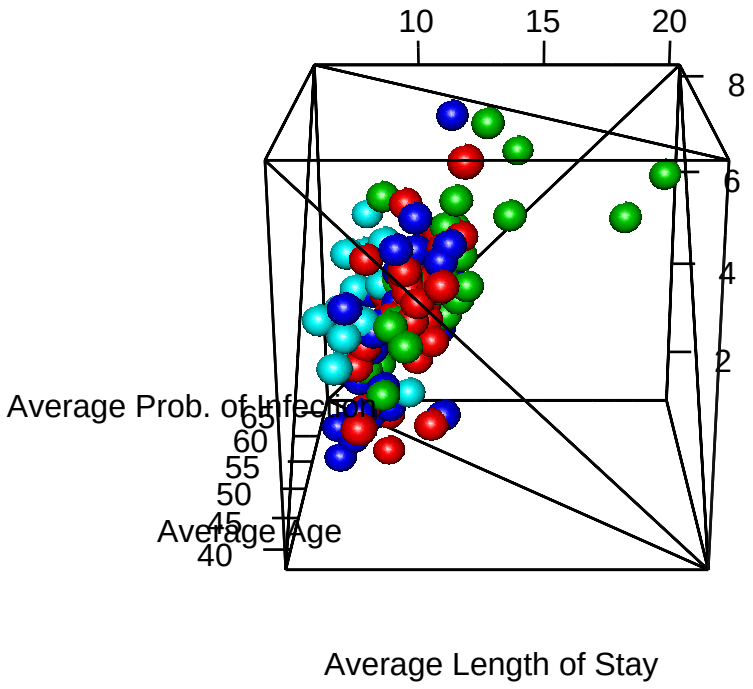
```
install.packages("rgl")
```

Load the library.

```
library(rgl)
```

Plot infection percentage versus average length of stay and average age. Use color to indicate the region. *rgl* is a different graphics API than the base graphics. Plot the points, resize the window, and then add a legend.

```
plot3d(senicData$Length_stay, senicData$Age_years, senicData$Infection_pct,
       col=as.numeric(senicData$Region_Name)+1, type="s",
       xlab="Average Length of Stay",ylab="Average Age",
       zlab="Average Prob. of Infection")
legend3d("topright", levels(senicData$Region_Name),
         col=1:length(levels(senicData$Region_Name))+1, pch=16)
```

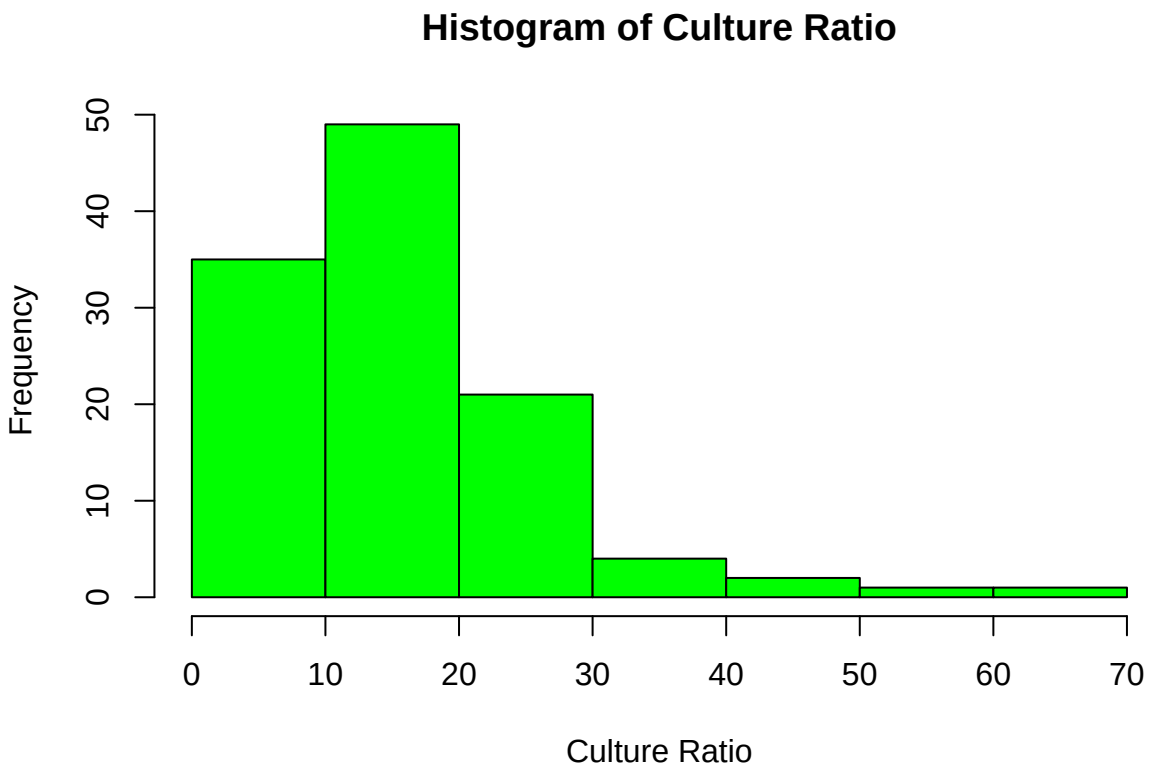


Solutions to Exercises

SENIC Data

1. Create a histogram of the culture ratio. The distribution is right-skewed. Most hospitals perform 5-25 cultures for every infection, while a few hospitals perform 40-70 cultures per infection.

```
hist(senicData$Culture_ratio, col="green", xlab="Culture Ratio",
     ylab="Frequency", main="Histogram of Culture Ratio")
```

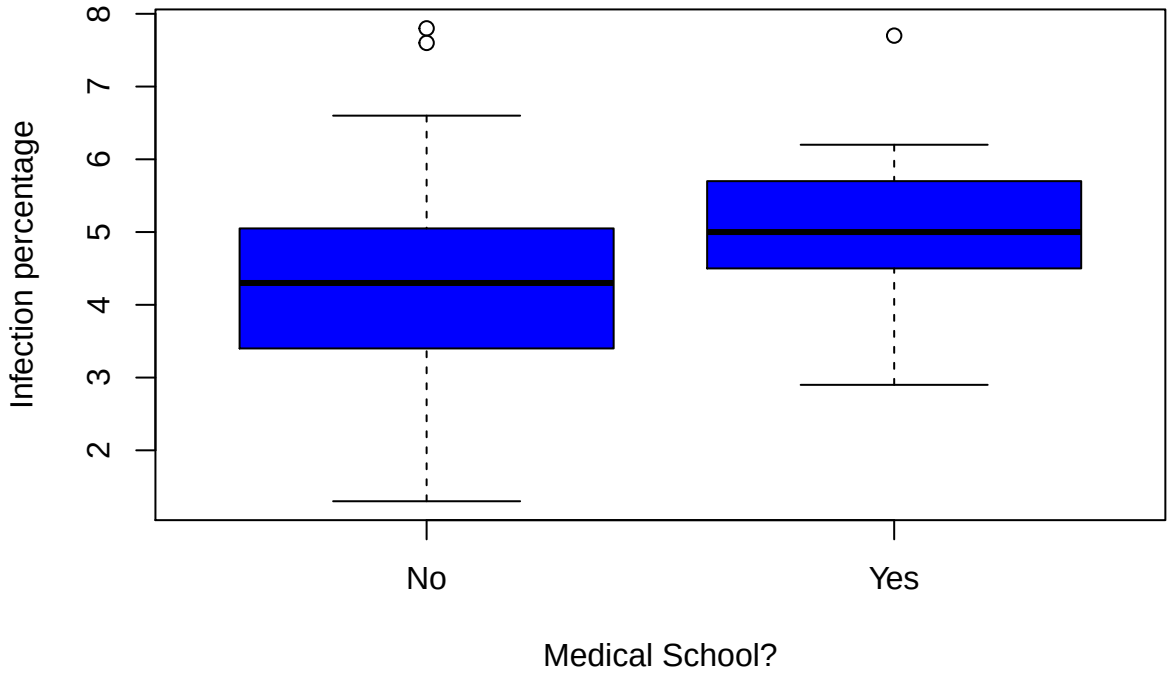


2. The default length of whiskers is in the help topic for boxplot. The "range" option determines the length of the whiskers. The default is that a whisker extends to a point in the data that is up to 1.5 times the interquartile range.

```
?boxplot
```

3. Create a boxplot of the infection percentage. Hospitals with medical schools tend to have higher infection percentages. The median infection percentage is higher, and the first quartile of infection percentage for the hospitals with medical schools is larger than the median for those without.

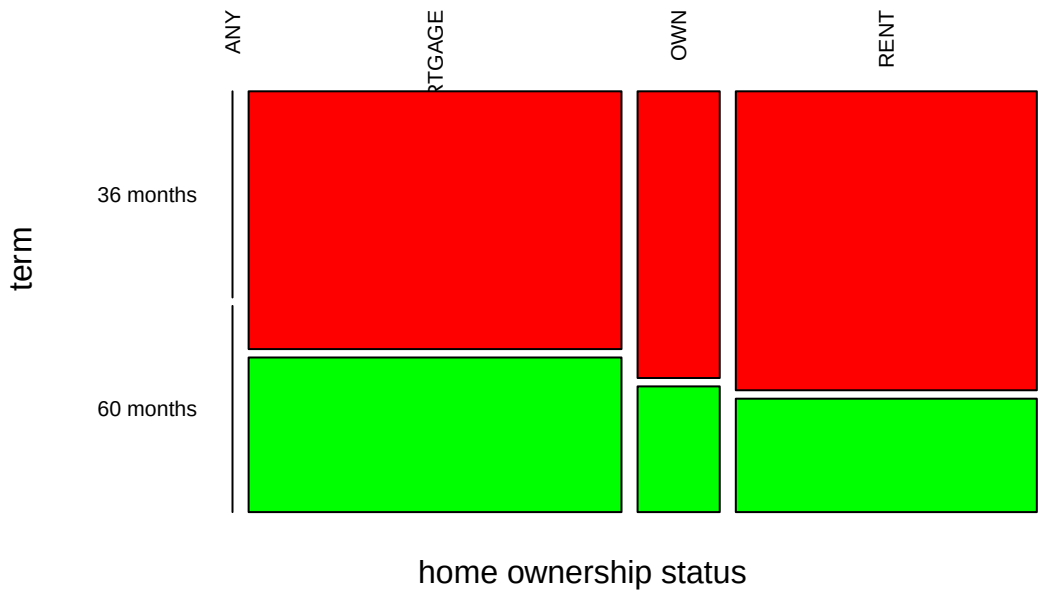
```
boxplot(senicData$Infection_pct ~ senicData$Medical_School,
        xlab="Medical School?", ylab="Infection percentage", main="", col=4)
```



Lending Club Data

1. Create a mosaic plot of home ownership status versus term.

```
homeownerTermTable <- table(lendingData[,c("home_ownership", "term")])
mosaicplot(homeownerTermTable, color=rainbow(3), xlab="home ownership status",
           ylab="term", main="", las=2)
```



2. Create a time series plot to investigate whether the loan amounts are seasonal.

The issue dates are given by month, so first aggregate loan amounts by month. Then create a line plot.

```
library(dplyr)
```

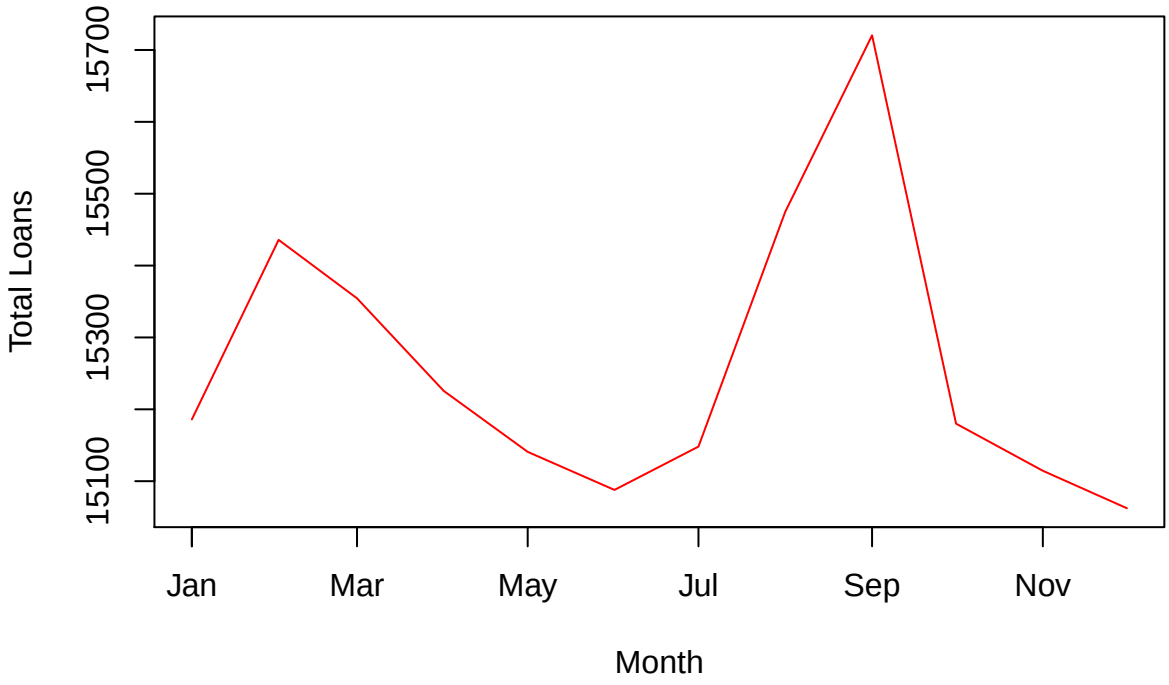
```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

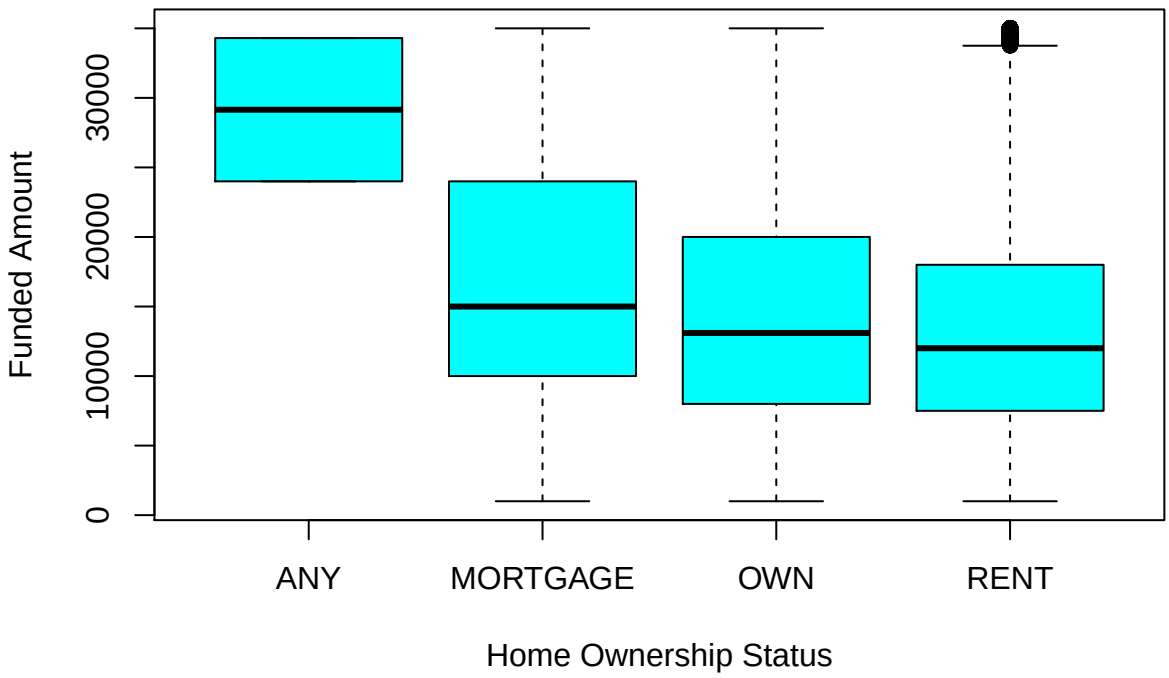
```
lendingData$issue_d <- as.POSIXct(lendingData$issue_d)
loansByMonth <- lendingData %>%
                select(loan_amnt, issue_d) %>%
                group_by(issue_d) %>%
                arrange(issue_d) %>%
                summarise(mean(loan_amnt))

plot(loansByMonth, type="l", col="red", xlab="Month", ylab="Total Loans")
```



3. Create a boxplot of funded amount by home ownership status.

```
boxplot(lendingData$funded_amnt ~
        lendingData$home_ownership, xlab="Home Ownership Status",
        ylab="Funded Amount", main="", col=5)
```



The median funded is smallest for renters.

4. Create a box plot of funded amount by home ownership and term. *ggplot2* is a powerful but unintuitive (to me, at least!) plotting package. Install *ggplot2*.

```
install.packages("ggplot2")
```

Load *ggplot2* and *reshape2*.

```
library(ggplot2)
library(reshape2)
fundedData <- lendingData[, c("funded_amnt", "home_ownership", "term")]
```

Melt the data for the three variables. The result has four columns: home_ownership, term, variable (funded_amnt), and value.

```
fundedMeltedData <- melt(fundedData)
```

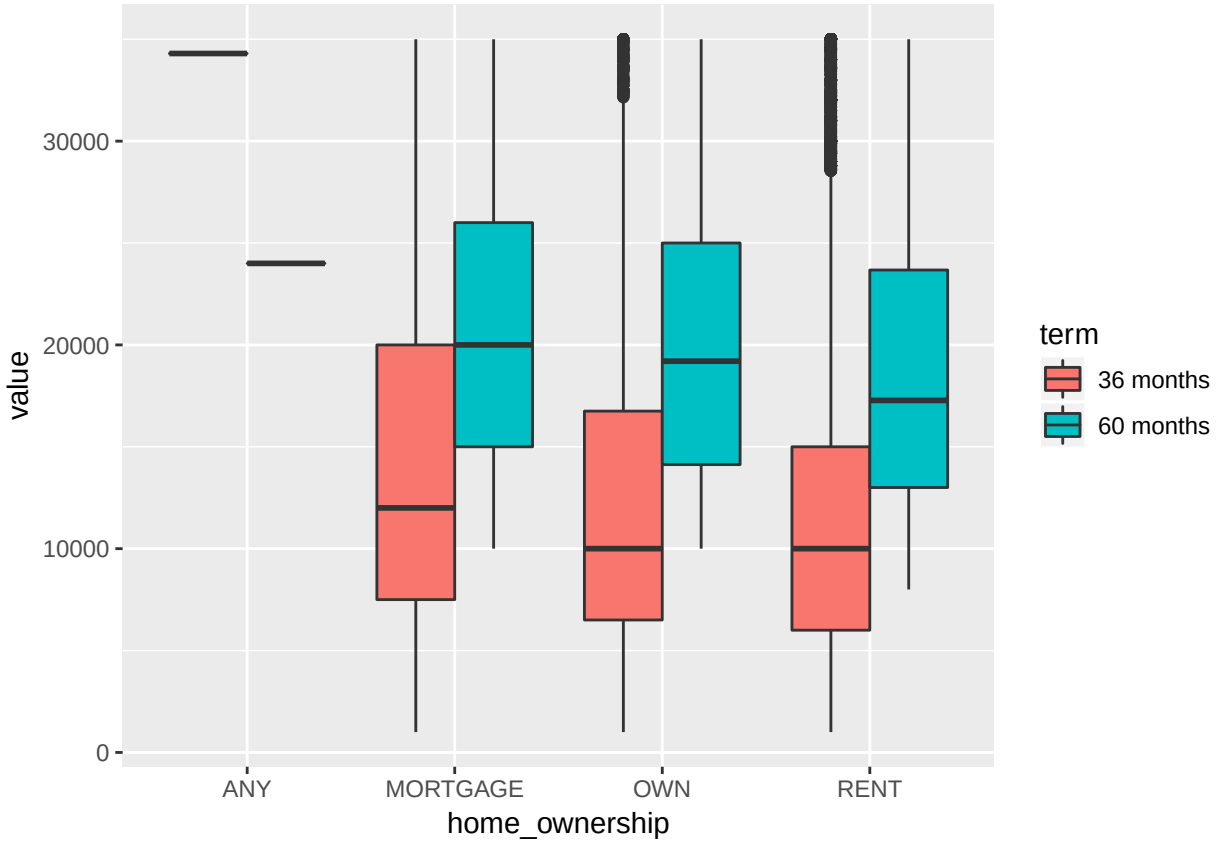
```
## Using home_ownership, term as id variables
```

Turn off any active graphics devices.

```
dev.off()
```

Create a box plot with one box for each home ownership status and term grouped together by home ownership status.

```
myBoxPlot <- ggplot(fundedMeltedData, aes(x=home_ownership, y=value), group=term) +
             geom_boxplot(aes(fill=term))
print(myBoxPlot)
```



Save the plot to a file.

```
ggsave(myBoxPlot, file="fundedHomeTerm.jpg", width=10, height=5)
```