Intro to R for Epidemiologists

Lab 3 (1/29/15)

Many of these questions go beyond the information provided in the lecture. Therefore, you may need to use R help files and the internet to search for answers. Feel free to ask questions of the instructor, the TAs, or your classmates, but try to work through as much as you can independently.

For the lab, you are expected to create an R script (.R file in the R editor) with your code corresponding to each question. Begin each question with a commented line of code indicating the question. As an example:

```r
# Jenna Krall

# Question 1.
head(iris)
```

Part 1. Ozone vs. temperature

The `airquality` dataset in the `datasets` package includes daily air quality measurements taken at Roosevelt Island in New York City from May to September 1973.

Data Source: New York State Department of Conservation (ozone data) and the National Weather Service (meteorological data)

Look at the help file for the `airquality` dataset and use the following instructions to reproduce the plot below.

1. Remove missing values from the `airquality` dataset.
   a. Use `complete.cases` to determine which rows have complete data (i.e. no missing data)
   b. Subset your data using the information in (a.) and bracket notation
2. Create a scatterplot of ozone versus temperature (see below)
   a. Label the axes (include units)
   b. Add a title to the plot
   c. Change the plotting symbol
   d. Change the point color so that the color corresponds to the month of the measurement. For the colors, use `c("darkslategray1", "darkslategray2", "darkslategray3", "darkslategray4", "darkslategray")`.
3. Add a loess line
4. Add a legend for the colors

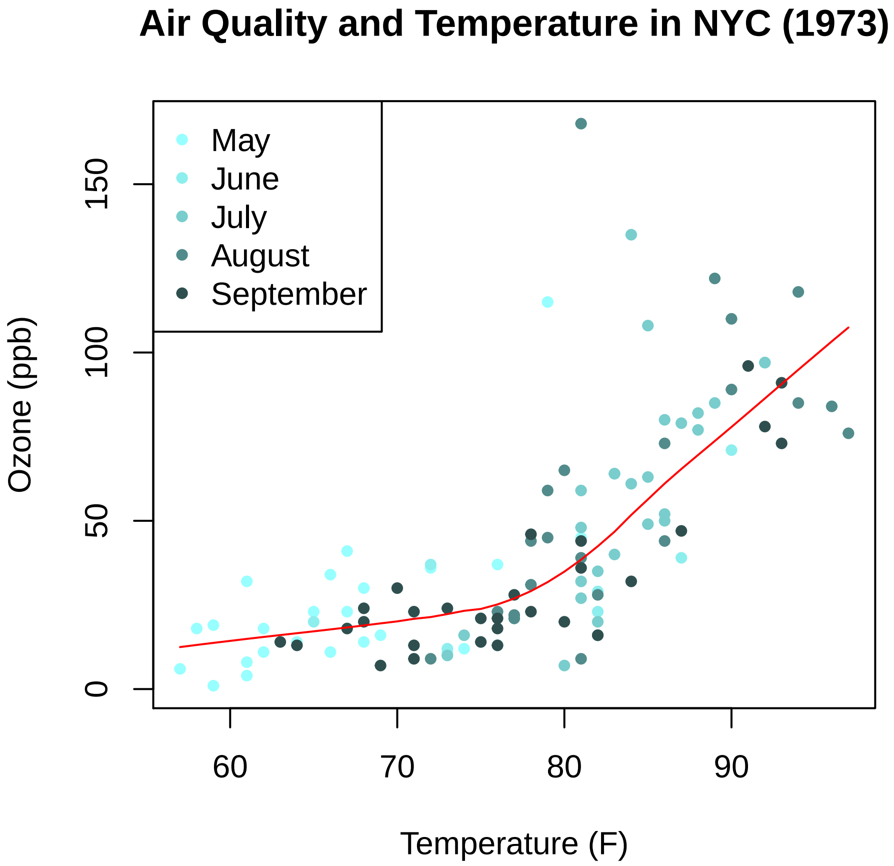
```r
# Part 1
# Remove missing values
aq_cc <- airquality[complete.cases(airquality), ]

# Initialize your vector
col.Month <- vector(length = length(aq_cc$Month))
# See possible values of month
# table(aq_cc$Month)
# Fill in col.Month with colors
colsall <- c("darkslategray1", "darkslategray2", "darkslategray3",
  "darkslategray4", "darkslategray")
col.Month[aq_cc$Month == 5] <- colsall[1]
col.Month[aq_cc$Month == 6] <- colsall[2]
col.Month[aq_cc$Month == 7] <- colsall[3]
col.Month[aq_cc$Month == 8] <- colsall[4]
col.Month[aq_cc$Month == 9] <- colsall[5]

# Create the scatterplot with title and proper axes labels
# Also change plotting symbol and color
plot(Ozone ~ Temp, data = aq_cc, xlab = "Temperature (F)",
  ylab = "Ozone (ppb)", main = "Air Quality and Temperature in NYC (1973)",
  pch = 20, col = col.Month)

# Add loess line
lowess_aq <- lowess(aq_cc$Temp, aq_cc$Ozone)
lines(lowess_aq$x, lowess_aq$y, col = "red")

# Add legend
months <- c("May", "June", "July", "August", "September")
legend("topleft", legend = months, col = colsall, pch = 20)
```



Part 2. Bar plots

Download the dataset `haireyecolor_matrix.csv` from the course website. This dataset includes hair color, eye color, and sex for 592 statistics students.

1. Read in the dataset
2. Limit the dataset to only females
3. Separately for each hair color, create a vector of proportions representing eye color (e.g. what proportion of brunettes have brown eyes?)
4. Create a bar plot for each hair color
   a. Add a title to each plot
   b. Label the axes in each plot
   c. Ensure the y-axis limits are correct. Hint: `?plot.default`
   d. Change the direction of the axis labels.
   e. Change the colors to reflect the eye color represented
5. Consolidate all plots so that you have two rows of figures: one with black, blond, and brown hair and one with only red hair (as shown). Hint: `?layout`
6. Add x-axis text to only the last plot for red hair. You will need to increase the margins on the bottom for the last plot (Hint: `?par`) and add the text after the plot (Hint: `?mtext`).

```
#Part 2
# Read in the data
hec <- read.csv("haireyecolor_matrix.csv", stringsAsFactors = FALSE)

#Select just female observations
hec_female <- hec[hec$Sex == "Female", ]


#set up
layout(matrix(c(1, 2, 3, 4, 4, 4), ncol = 3, byrow = T))
cols <- c("burlywood4", "blue", "turquoise", "green")

# Plot a.
hec_use <- hec_female[hec_female$Hair == "Black", ]
# Find proportions
prop_hec <- hec_use$value/ sum(hec_use$value)
names(prop_hec) <-  hec_use$Eye

#Create barplot
barplot(prop_hec, ylab = "Proportion", xlab = "", ylim = c(0, 1),
  main = "Eye color for Black hair", col = cols, las = 2)

# Plot b.
hec_use <- hec_female[hec_female$Hair == "Blond", ]
# Find proportions
prop_hec <- hec_use$value/ sum(hec_use$value)
names(prop_hec) <-  hec_use$Eye
```

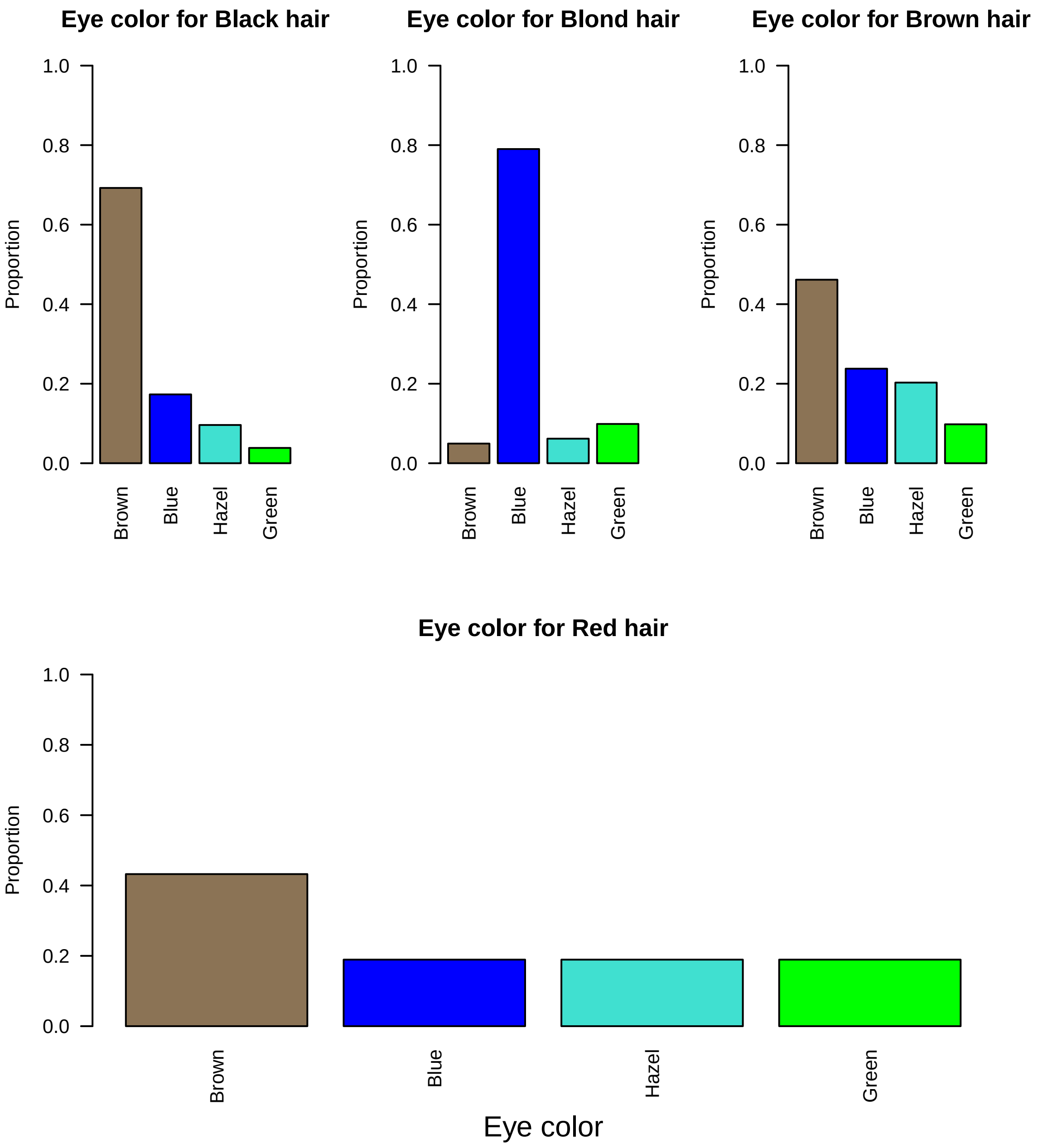
```r
#Create barplot
barplot(prop_hec, ylab = "Proportion", xlab = "", ylim = c(0, 1),
  main = "Eye color for Blond hair", col = cols, las = 2)

# Plot c.
hec_use <- hec_female[hec_female$Hair == "Brown", ]
# Find proportions
prop_hec <- hec_use$value/sum(hec_use$value)
names(prop_hec) <-  hec_use$Eye

#Create barplot
barplot(prop_hec, ylab = "Proportion", xlab = "", ylim = c(0, 1),
  main = "Eye color for Brown hair", col = cols, las = 2)

# Plot d.
hec_use <- hec_female[hec_female$Hair == "Red", ]
# Find proportions
prop_hec <- hec_use$value/sum(hec_use$value)
names(prop_hec) <-  hec_use$Eye

#Create barplot
mar.def <- par()$mar
par(mar = mar.def + c(2, 0, 0, 0))
barplot(prop_hec, ylab = "Proportion", xlab = "", ylim = c(0, 1),
  main = "Eye color for Red hair", col = cols, las = 2)
mtext("Eye color", side = 1, line = 4)
```



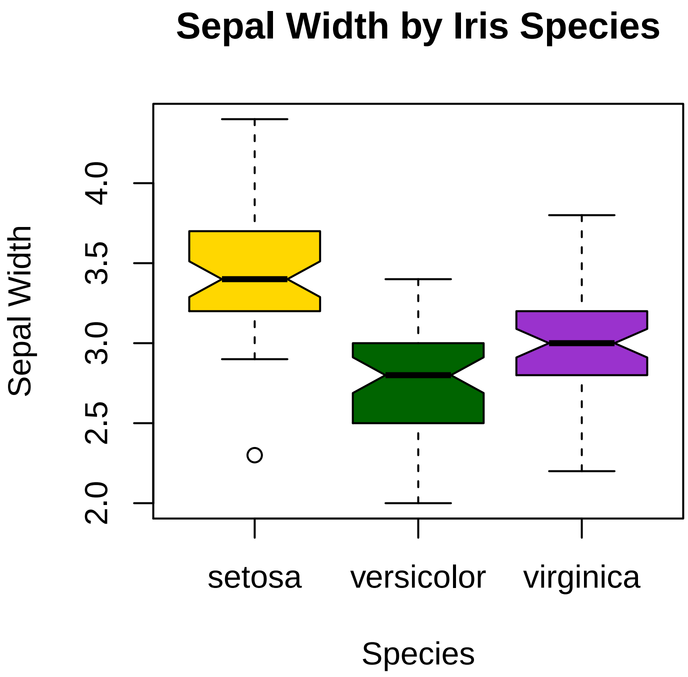
Part 3. Boxplot of sepal width by species

Use the `iris` dataset in R to create notched boxplots of sepal width by species

1. Make each box a different color
2. Add axis labels
3. Add a title

```
#Part 3
# Create notched box plots with
# 2. different colors
# 3. axis labels
```

```r
# 4. Title
boxplot(Sepal.Width ~ Species, data = iris, notch = TRUE, col = c("gold",
  "darkgreen", "darkorchid3"), ylab = "Sepal Width", xlab = "Species",
  main = "Sepal Width by Iris Species")
```


Sepal Width by Iris Species
Sepal Width
setosa
versicolor
virginica
Species