ISTA 116: Lab Assigment #3 (50 pts)

Due Friday, Sept. 30, 5 P.M.

Problem 1:

(25 pts total + 3 pts possible Extra Credit)

The `Berkeley.csv` dataset (on d2l) contains data about graduate admissions to six departments at UC Berkeley in 1973. Each observation represents one applicant. Applicants are classified by their sex (`Sex`), the department they applied to (`Department`), and whether they were admitted (`Admission`).

As part of a lawsuit, UC Berkeley was accused of having admissions practices that discriminated against women. Let's investigate the evidence for this allegation using the 1973 data.

a. (2 pts) Create a data frame in `R` from this data, and then create a contingency table displaying joint frequencies for sex and admission status.

b. (4 pts) Create a new table containing the conditional proportions of admissions status given sex (use `prop.table()`).

c. (3 pts) Create a bar plot with a group or stack for each sex, with bars showing the conditional proportions calculated in part (b) (Hint: The `barplot()` function will group/stack within columns of a contingency table. If you need to group based on rows, you can use the `t()` function (for "transpose") on you contingency table first, to switch rows and columns).

d. (3 pts) Is there a suggestion of bias against women, based on parts (b) and (c)?

e. (2 pts) Let's take department into account as well. Create a three-way contingency table showing joint frequencies by `Sex`, `Admission` and `Department`.

f. (4 pts) Compute proportions of admissions, conditioning on *both* `Sex` *and* `Department`. (Use `prop.table()` again, this time on the table you created in (e). To condition on more than one variable, give a vector for the `margin=` argument).

g. (4 pts) Comment on the differences in acceptance rates by gender for the six departments (Optional: produce a bar plot or a series of bar plots depicting these differences).

h. (3 pts) Offer a possible explanation for any differences between the conclusions you made in parts (d) and (g).

i. (EC: 3 pts) Investigate your hypothesis from (h) using the data by calculating any other conditional proportions that are relevant to the question.

Problem 2:

(25 pts)

The `MPG.csv` dataset contains information about the number of miles-per-gallon (the `MPG` variable) achieved by cars with various physical characteristics (Engine displacement (`Eng.Displ`), number of cylinders (`Cyl`) and transmission type (`Transm`). Let's compare the miles-per-gallon achieved by cars with manual vs. automatic transmissions.

a. (5 pts) Plot the `MPG` density curves for the automatic (`Transm == Autom`) and manual (`Transm == Manual`) transmissions on top of each other. You might need to play around with the `xlim=` and `ylim=` arguments to avoid cutting anything off. Distinguish the curves by color and/or line type.

b. (4 pts) Compute the means and five-number summaries for the two distributions.

c. (4 pts) Produce side-by-side box plots of the two distributions.

d. (4 pts) Which type of transmission tends to get better gas mileage in general? What information did you use to make this determination?

e. (4 pts) Which type of transmission does the car with the greatest gas mileage have? What information did you use to answer this question?

f. (4 pts) If these answers aren't the same, give a possible reason for the discrepancy. If a new data set were collected, what do you think would be more likely to change: your answer to part (d), or your answer to part (e)? Why?