ISTA 116: Lab Assigment #2 (50 pts)

Your Name Here

Due Sept. 13-14 in Lab

The d2l site contains a dataset called `NJSpeeding3.csv`, containing data collected about speeding tickets issued in New Jersey. The `Speed` variable is the speed that ticketed drivers were reported to have been going. The `Overlimit` variable indicates how far over the posted limit this was. Finally, the `License` variable is a factor indicating what state, district or territory issued the license plate on the car. Canadian plates are listed as `CN`; in cases where the plate information is unknown, the value is `U` . (Note: the actual string has a space after the `U`)

Problem 1: Categorical Data (25 pts)

a. (3 pts) Create a table showing how often tickets were issued according to the driver's state of origin (`License`). Display your code and the resulting table.

b. (3 pts) Convert the frequency table you created in part (b) to a table showing percentages. The function `prop.table()` will give you decimal proportions; convert these to percentages with an arithmetic calculation, sort, and display the results, rounded to 2 decimal places using `round()` (Note: It's a good idea *not* to round the values in the table itself – only do the rounding for display purposes). As always, show your code, and the resulting table.

c. (3 pts) Now, sort the results in decreasing order (use `sort()` — **remember:** `sort()` by itself doesn't change anything; you need to assign the result somewhere!).

d. (4 pts) Create a pie chart, showing the number of tickets issued by driver's home state (you can use one of the tables you created in parts a-c. Which one looks best? How informative is it?). Include an informative title using the `main=` argument.

e. (4 pts) Describe a strategy you might use to improve the information value of your pie chart.

Figure 1: Your pie chart can go here!

f. (4 pts) An alternative to a pie chart is a bar plot, which makes it even easier to see what's bigger than what, and by how much; plus you can show the actual counts. Use the table you created in part (e) to create a barplot, with sensible axis labels (`xlab=` and `ylab=`) as well as a main title (`main=`). Use color (`col=`) if you want.

g. (4 pts) What do you notice about the distribution of tickets issued? What are some factors that account for the differences in numbers? Is there anything surprising about it?

Problem 2: Numeric Data (25 pts)

a. (4 pts) Using the `Speed` variable, create a histogram, with suitable title and axis labels, showing **proportions** on the y-axis (use `prob = TRUE`). Experiment with different values of `breaks=`. On the same plot, overlay a density curve.

b. (4 pts) The distribution should look mostly like a "bell curve", with one or two notable differences. What difference(s) do you notice? Speculate about what aspects of the real-world situation might contribute to the shape.

c. (2 pts) Convert the values in `Speed` from miles-per-hour to kilometers-per-hour (there are about 1.61 km in 1 mile). Store the results in a new variable.

d. (5 pts) Compute $\bar{x}$ and s for both the original Speed variable (in mph) and the new variable (in km/h). How did the conversion affect these values?

e. (5 pts) Convert the miles-per-hour variable into a new variable containing z-scores. Do the same for the km/h variable. Produce side-by-side histograms of both. What do you notice?

f. (5 pts) Compute the five-number summary and produce a box plot of the `Overlimit` variable. Comment on the shape of the distribution. What factors might explain the shape?