EDAV Fall 2020 PSet 1

Read *Graphical Data Analysis with R*, Ch. 3

Grading is based both on your graphs and verbal explanations. Follow all best practices as discussed in class. *Do not expect the assignment questions to spell out precisely how the graphs should be drawn. Sometimes guidance will be provided, but the absense of guidance does not mean that all choices are ok.*

You must use R for this assignment, either base R or tidyverse, which includes **ggplot2**. If you're new to R, I recommend learning **ggplot2** rather than base R graphics. Instructions for getting started in R are available in this chapter of edav.info.

The datasets in this assignment are from the **alr4** package which can be installed from CRAN:

```
install.packages("alr4")
```

1. lakes

Choose one of the numeric variables in the `lakes` dataset.

a) Plot a histogram of the variable.

b) Create a new factor variable called `Species_size` from the `Species` column that divides the `Species` variable into three groups of equal range with levels `small`, `medium` and `large`. (Hint: use the `cut()` function.) Display the first 10 values of `Species_size`.

c) Plot histograms, faceted by `Species_size`, for the same variable.

d) Plot multiple boxplots, grouped by `Species_size` for the same variable. The boxplots should be ordered by decreasing median from left to right.

e) Plot overlapping density curves of the same variable, one curve per factor level of `Species_size`, on a single set of axes. Each curve should be a different color.

f) Summarize the results of c), d) and e): what unique information, *specific to this variable*, is provided by each of the three graphical forms? In other words, describe one by one what you learn from the histograms, boxplots, and density curves that the others don't show as well or at all.

g) Briefly research the lake with the highest number of `Species`. What factors do you think contribute to this? Is this also the largest lake?

2. Challeng

a) Draw two histograms of the `temp` variable in the `Challeng` dataset in the **alr4** package, with binwidths of 5 years and `boundary = 0`, one right open and one right closed. How do they compare?

b) Redraw the histogram using the parameters that you consider most appropriate for the data. Explain why you chose the parameters that you chose.

3. **drugcost**

a) Use **tidyr** functions to convert the numeric columns in the `drugcost` dataset in the **alr4** package to two columns: `variable` and `value`. The first few rows should be:

```
   variable      value
1       COST       1.34
2       COST       1.34
3       COST       1.38
4       COST       1.22
```

Use this form to plot histograms of all of the variables in one plot by faceting on `variable`. What patterns do you observe?

For the remaining parts we will consider different methods to test for normality.

b) Choose one of the variables with a unimodal shape histogram and draw a true normal curve on top on the histogram. How do the two compare?

c) Perform the Shapiro-Wilk test for normality of the variable using the `shapiro.test()` function. What do you conclude?

d) Draw a quantile-quantile (QQ) plot of the variable. Does it appear to be normally distributed?

e) Use the **nullabor** package to create a lineup of histograms in which one panel is the real data and the others are fake data generated from a null hypothesis of normality. Can you pick out the real data? If so, how does the shape of its histogram differ from the others?

f) Show the lineup to someone else, not in our class (anyone, no background knowledge required). Ask them which plot looks the most different from the others. Did they choose the real data?

g) Briefly summarize your investigations. Did all of the methods produce the same result?