EvalFest How-Tos for Graphical Exploratory Data Analysis in R

Graphical Exploratory Data Analysis in R

Let's explore the data graphically using R. I use RStudio as my editor. We are building on videos in our Data Management series. In the Database Organization and Password Protection and Data Cleaning and Recoding videos, we discuss how to prepare your data for analysis.

> **Now you might be thinking, why bother?** Looking at the data in various graphical forms is another way of ensuring that the data were collected and coded as expected. If you produce a graph of the data that is very unexpected, you will want to go back to the raw data to try to determine what is wrong.
>
> This step is really about looking at the data in a graphical format to ensure that nothing is amiss. If you want more information about formatting graphs or creating other graphs, we suggest Stephanie Evergreen's blog, 'How to Build Data Visualizations in Excel' (http://stephanieevergreen.com/how-to/) and Ann K Emery's online tutorials (http://annkemery.com/tag/tutorials/) as good resources.

Exploring the data graphically

For now, we're still getting to know our data, and so let's get started.

1. Single Bar graph
    a. Let's look at a single bar graph for the variable Gender.
        i. Bar graphs are typically used to display categorical data.
    b. You should recognize lines 2, 3, 5, and 6 as the code to set up the data for R.
        i. Note the extra piece of code in line 3 that begins with na.strings. This code tells R to ignore any instances of missing data. This is a critical piece of code to ensure that our tables and graphs do not include a category for missing data.
    c. In order to produce a bar graph in R, you have to first create a frequency table. Line 9 creates the frequency table.
        i. Notice the na.omit instruction – this again tells R to ignore missing cases.
    d. And line 10 produces the bar graph.
    e. There are many aspects of the graph that can be formatted, such as color, data labels, etc. However, because we are simply exploring our data, we won't format the graph any further at this point.

2. Side-by-side bar graph
    a. Let's look at Overall Rating by Gender. For this, we will produce a side-by-side bar graph.
    b. In order to produce a side-by-side bar graph, we once again have to start with a table. But this time, we will create a contingency, or two-way table, that will include the counts for each bivariate combination of Overall Rating and Gender.
    c. Line 13 tells RStudio to make the contingency table, and Line 14 produces the side-by-side bar graph.
        i. Note that, unless instructed otherwise, R does everything in alphabetical order with working with string or text data (as opposed to numeric data). So the first bar in each overall rating group indicates the count of females, and the second bar indicates the count of males.
    d. Again, look for patterns in the graph that you may not have expected. Does everything look okay? Should you proceed? These are questions you want to ask at this step.
    e. As you can see, Very Good (4) and Excellent (5) are the most frequent ratings, regardless of Gender. If we think back to the mean, median, and standard deviation we computed for OverallCode in the Numeric Exploratory Data Analysis video, these values align with what we see in this graph. The mean of 4.39 is less than the median of 5, the standard deviation is relatively small (.759), and the data are left skewed, showing most of our data at the upper end of the rating scale.