Manual Data Exploration and Analysis

Tim Vigers

2/5/2020

Questions

1. Which cereal has the highest and lowest caloric density?
    - This is a query of extremes (one attribute).
2. Which cereal brand has the highest average caloric density?
    - This requires computing a derived value (analysis) and querying for an extreme.
3. Is there an association between shelf and manufacturer (i.e. does a particular brand tend to be on the top shelf)?
    - This requires computing a derived value (analysis) for two attributes.
4. What is the distribution of sodium content?
    - This is a query across all values of an attribute.
5. Which cereals are the most similar across all nutrients?
    - This would be a derived attribute (some sort of distance metric) across many attributes.

Insights

1. There appears to be a slight correlation between sugars and calories. This is sort of what you'd expect, but it's nice to see it borne out in the data when visualized in a scatter plot (and would not have been obvious just looking at the data). One interesting thing I noticed, though, is that the correlation appears to be much stronger when you plot calories on the x axis and sugars on the y. I also noticed that there are negative sugar values after plotting these data.
2. There appear to be slightly different distributions of calorie content by manufacturer. For example, most General Mills cereals contain 110 calories, whereas the other manufacturers seem to be somewhat more variable. However, I'm not really sure how to test this in Excel.
3. General Mills and Kellog have approximately 60% of the "market share" in this dataset (which I noticed by making a pie chart). I may not have seen this without visualizing it, and am curious whether or not these samples are representative. But I'm not sure how to do that analysis without more data.
4. There are quite a few 0 values for sodium, but the rest of the data appears to be fairly normally distributed. If I were doing this in R, I would test to
5. One thing I'm really curious about is which nutrients are different between hot and cold cereals. Unfortunately I'm not great at this sort of analysis in Excel (or by hand), and there are only 3 data points for hot cereals so it would be nice to gather more data.

The Process

For this analysis I opened the data in Excel and did my best to plot it. I find some plots are difficult in Excel, like histograms for example, so I did those by hand. Eventually I figured out how to make a histogram in Excel though (see the second sheet of the uploaded Excel document), which was helpful in confirming my hand drawing. Also, I made X-Y scatterplots for variables that I thought might be correlated or was curious about.

Limitations

I found using Excel for this assignment incredibly frustrating, although it was a good exercise in some ways. I have several go-to plots when I'm working in R, so using a program I'm less familiar with forced me to think a little more carefully about exactly what I wanted to do. However, because it took me forever to figure out how to make even the most basic plots, I felt pretty limited in terms of what I could actually visualize. For continuous variables, all I could think to do was very simple X-Y scatterplots to visually assess correlation, and I never figured out how to make a box and whisker plot, which is really useful for finding outliers. Finally, I realized two limitations about myself: 1) that I don't know how to do many statistical tests by hand, and almost completely rely on R's modeling functions for analysis; and 2) that I'm not particularly creative in my exploratory data analysis.