IMC463: Machine Learning for IMC
Assignment 5: Transformations
Due: Thursday, May 26 by 1pm

Submit homework at the before class on Canvas. Work in your homework groups and submit one copy of the homework per group with the names of all your group members.

1. Breakfast cereal case Data frame `UScereal` distributed with the MASS library (install MASS then type `library(mass)`) describes 65 commonly available breakfast cereals in the US, based on information available on the mandatory food label on the package. The measurements are normalized to a serving size of one American cup (see help page `?UScereal`). I will use the term *numerical nutrition variables* to mean all variables except for `mfr`, `shelf` and `vitamins` (variables 2:8, 10). This data set is an exercise in the classic V&R3, where the authors asked, (Q1) Is there any way to discriminate among the major manufacturers by cereal characteristics, or do they each have a balanced portfolio of cereals? (Q2) Are there interpretable clusters of cereals? (Q3) Can you describe why cereals are displayed on high, low or middle shelves? I add: (Q4) How to visualize the data and show clusters or manufacturers? Keep the following in mind as you search for clusters:

    - This problem addresses a core topic in marketing management, **market segmentation** (as opposed to **customer segmentation** discussed earlier, where the customers in some market segment are further partitioned into smaller groups). Customers have heterogeneous wants and needs. Large manufacturers like General Mills identify clusters (segments) based on these wants and needs and then create ("target") a brand for each segment.[1] Smaller manufacturers may not be able to compete across segments, and instead adopt a **niche strategy**, where they focus on small segment with unique needs that large manufactures may ignore. Your clusters should identify the mainstream segments as well as niches. Visualizations should show the strategies of different manufactures.
    - Should you standardize and/or transform the data prior to clustering and/or using PCA/FA? Should you cluster on the raw numerical variables, PCs or factors? Are there interpretable clusters of cereals?
    - I suggest examining both PCs and varimax rotated PCs. Which do you prefer? Plotting components in a two-dimensional scatterplot is a form of a **perceptual map**. Empty regions may reveal opportunities to launch new brands (or cereals that nobody wants to eat!).
    - Which clustering methods do you suggest? Why?
    - I think that unsupervised (PCA/cluster) methods are the way to go with Q1, Q2 and Q4, but one could make a case to use supervised for Q3 (or parts of Q1).

[1] A fundamental approach is "STP:" segment, target and position.