Unsupervised Learning I: Association Rules

Introduction

In this lesson and the next, you'll use un-supervised learning techniques to mine for patterns in the colleges data set. In this lesson, you'll use the arules package to mine for associations, and next time clustering.

Data

The College Scorecard is a data set provided by the White House that is designed to increase transparency around college decision making. The data set includes information on geographic location, standardized test performance, and the demographic make-up of the student body. Use the str (structure) and summary functions to familiarize yourself with the data, and read more about it on the official government website.

Unlike the previous two problem areas, we do not have an outcome variable.

Instructions

Follow the steps below to construct association rules with the arules package and the colleges data set.

1. Read in the data using the `read.delim` function. Then use `ggplot2`, `ggpairs`, and `dplyr` to identify interesting relationships in the data. Write a short description of one interesting pattern you identified.
2. Prepare your data for association rule mining by transforming it into a set of transactions. Use the inspect and summary functions to view the transactions.
3. Generate rules with the apriori function with a support of 0.01 and a confidence of 0.60.
4. Try the following combinations of support and confidence: [0.10, 0.60], [0.01, 0.10]. What happens to the number of rules as the support increases? (**Hint:** use the summary function to see the number of rules).
5. In the text we constructed earnings quartiles and explored the associations in top earners by filtering the rules for the top quartile of earners. Now, re-filter the rules to explore the bottom 25% of earners (Q1). Report at least 1 interesting finding. **Hint:** Use the subset and inspect functions to filter the right-hand side (rhs) for 'earnings_quartiles=Q1'.