# Problem Set 8

QMBE 3740: Data Mining

Module: Association Rules

## Part 1

**Question 1**

You work in a hospital and have access to patient medical records. You decide to use association rules on a variety of datasets available to you. In this context, what are examples of association rules that you might discover that fit into each of the following cateogries?

a Actionable

b Trivial

c Inexplicable

**Question 2**

Think of an organization where you currently work, have worked in the past, or an organization you are familiar with (like a school, community group, etc.). What is an application of association rules that might be useful in that environment?

**Question 3**

Continue to explore the `groceries.csv` dataset that we used in class and that was presented in the Chapter 11 case study. Answer the following questions.

a What are the 10 least frequently purchased items?

b If you change the minimum rule length to 3, how many rules to you generate? What if you change it to 4? (Use the same support / confidence thresholds used in the case study)

c Change the minimum rule length back to 2 and produce a list of rules involving either soda or whipped/sour cream (you'll need to study the `subset()` function)

## Part 2

Use the `Market_Basket_Optimisation.csv` dataset provided on Canvas and perform association rule mining as we did in class with both the groceries and lastfm datasets. Perform the following tasks and answer the related questions.

1. Read the transactions into R.
2. Use the `summary()` function to answer the questions:
    - How many transactions are in the data?

- How many distinct items are in the data? Using the formula used in class (and in the book), calculate the number of possible itemsets we would theoretically have to evaluate in a brute force approach.

3. Using the `summary()` function output, create a graph showing the distribution of transaction sizes in the data.
4. Using the `itemFrequency()` function, create a dataset of items and their frequencies and determine the ten most frequent items, and the ten least frequent items.
5. Use descriptives statistics on the item frequencies to determine a reasonable support threshold (use `confidence=0.25` and `minlen = 2`) and generate the association rules using the apriori algorithm.
6. Evaluate the rules and answer:
   - How many association rules were generated?
   - How many different rule lengths are there and how many rules are in each length?
   - Printout the top 12 association rules by confidence.
   - Printout the top 12 association rules by lift.
7. Using the `subset()` function, printout the top 10 association rules by lift, that do not include the 6 most frequent items.
8. Discuss a couple of the rules you find most interesting and explain how you think they might be used in a retail context.