Association mining can be applied to many data problems beyond the well-known example of finding relationships between different products in customer shopping data. In this homework assignment, we will explore real data from the banking sector and look for patterns associated with the likelihood of responding positively to a direct marketing campaign and signing up for a term deposit with the bank (stored in the variable “y”). You can find out more about the variables in this dataset here: <https://archive.ics.uci.edu/ml/datasets/bank+marketing>

Please make sure you have included an attribution statement (see syllabus if you have questions).

**Part 1: Explore Data Set**

1. Copy the contents of the following URL to a dataframe called *bank*:

<https://ist387.s3.us-east-2.amazonaws.com/data/bank-full.csv>

Hint: Even though this is a .csv file, chances are R won’t be able to read it in correctly using the read\_csv() function. If you take a closer look at the contents of the URL file, you may notice each field is separated by a semicolon (;) rather than a comma. In situations like this, consider using something like this:

*bank <- read.table(url, sep=";", header = TRUE)*

Make sure there are 41,188 rows and 21 columns in your *bank* df.

1. Next, we will focus on some key factor variables from the dataset, and convert a few numeric ones to factor variables. Execute the following commands and write a comment describing how the conversion for each numeric variable works and what the variables in the resulting dataframe are.

*bank\_new <- data.frame(job=bank$job,*

*marital=bank$marital,*

*housing\_loan=bank$housing,*

*young=as.factor((bank$age<median(bank$age))),*

*contacted\_more\_than\_once=as.factor((bank$campaign>1)),*

*contacted\_before\_this\_campaign=as.factor((bank$previous<0)),*

*success=(bank$y))*

1. Count the number of successful term deposit sign-ups, using the *table( )* command on the *success* variable.
2. Express the results of problem C as percentages by sending the results of the *table( )* command into the *prop.table( )* command
3. Using the same techniques, show the percentages for the *marital* and *housing\_loan* variables as well.

**Part 2: Coerce the data frame into transactions**

1. Install and library two packages: arules and arulesViz.
2. Coerce the *bank\_new* data frame into a sparse transactions matrix called *bankX*.
3. Use the *itemFrequency( )* and *itemFrequencyPlot( )* commands to explore the contents of *bankX*. What do you see?
4. This is a fairly large dataset, so we will explore only the first 10 observations in the *bankX* transaction matrix:

*inspect(bankX[1:10])*

Explain the difference between *bank\_new* and *bankX* in a block comment.

**Part 3: Use arules to discover patterns**

**Support** is the proportion of times that a particular set of items occurs relative to the whole dataset. **Confidence** is proportion of times that the consequent occurs when the antecedent is present..

1. Use apriori to generate a set of rules with **support** over 0.005 and **confidence** over 0.3, and trying to predict who **successfully** signed up for a term deposit. Hint: You need to define the right-hand side rule (rhs).
2. Use inspect()to review of the ruleset.
3. Use the output of inspect( ) or inspectDT( ) and describe any 2 rules the algorithm found.