# Association Rules Code

**Figure 1**  
*Load Libraries and Import .csv File*

library(arules)

## Loading required package: Matrix

##   
## Attaching package: 'arules'

## The following objects are masked from 'package:base':  
##   
## abbreviate, write

library(arulesViz)  
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.2 v dplyr 1.0.7  
## v tidyr 1.1.3 v stringr 1.4.0  
## v readr 1.4.0 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x tidyr::expand() masks Matrix::expand()  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()  
## x tidyr::pack() masks Matrix::pack()  
## x dplyr::recode() masks arules::recode()  
## x tidyr::unpack() masks Matrix::unpack()

library(Matrix)  
mod6course <- read.csv("Coursetopics.csv", header = TRUE)

**Figure 2**  
*Explore Dataset*

head(mod6course)

## Intro DataMining Survey Cat.Data Regression Forecast DOE SW  
## 1 1 1 0 0 0 0 0 0  
## 2 0 0 1 0 0 0 0 0  
## 3 0 1 0 1 1 0 0 1  
## 4 1 0 0 0 0 0 0 0  
## 5 1 1 0 0 0 0 0 0  
## 6 0 1 0 0 0 0 0 0

str(mod6course)

## 'data.frame': 365 obs. of 8 variables:  
## $ Intro : int 1 0 0 1 1 0 1 0 1 0 ...  
## $ DataMining: int 1 0 1 0 1 1 0 0 0 0 ...  
## $ Survey : int 0 1 0 0 0 0 0 0 0 0 ...  
## $ Cat.Data : int 0 0 1 0 0 0 0 1 0 1 ...  
## $ Regression: int 0 0 1 0 0 0 0 0 0 0 ...  
## $ Forecast : int 0 0 0 0 0 0 0 1 0 0 ...  
## $ DOE : int 0 0 0 0 0 0 0 1 0 0 ...  
## $ SW : int 0 0 1 0 0 0 0 1 0 0 ...

**Figure 3**  
*Convert to Matrix and Transaction Database*

mod6course <- as.matrix(mod6course)  
mod6course.trans <- as(mod6course,"transactions")  
summary(mod6course.trans)

## transactions as itemMatrix in sparse format with  
## 365 rows (elements/itemsets/transactions) and  
## 8 columns (items) and a density of 0.2136986   
##   
## most frequent items:  
## Intro SW Cat.Data Regression Survey (Other)   
## 144 81 76 76 68 179   
##   
## element (itemset/transaction) length distribution:  
## sizes  
## 0 1 2 3 4 5 6 7   
## 7 234 39 46 26 9 3 1   
##   
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 1.00 1.00 1.71 2.00 7.00   
##   
## includes extended item information - examples:  
## labels  
## 1 Intro  
## 2 DataMining  
## 3 Survey

str(mod6course.trans)

## Formal class 'transactions' [package "arules"] with 3 slots  
## ..@ data :Formal class 'ngCMatrix' [package "Matrix"] with 5 slots  
## .. .. ..@ i : int [1:624] 0 1 2 1 3 4 7 0 0 1 ...  
## .. .. ..@ p : int [1:366] 0 2 3 7 8 10 11 12 16 17 ...  
## .. .. ..@ Dim : int [1:2] 8 365  
## .. .. ..@ Dimnames:List of 2  
## .. .. .. ..$ : NULL  
## .. .. .. ..$ : NULL  
## .. .. ..@ factors : list()  
## ..@ itemInfo :'data.frame': 8 obs. of 1 variable:  
## .. ..$ labels: chr [1:8] "Intro" "DataMining" "Survey" "Cat.Data" ...  
## ..@ itemsetInfo:'data.frame': 0 obs. of 0 variables

inspect(mod6course.trans[1:20])

## items   
## [1] {Intro,DataMining}   
## [2] {Survey}   
## [3] {DataMining,Cat.Data,Regression,SW}   
## [4] {Intro}   
## [5] {Intro,DataMining}   
## [6] {DataMining}   
## [7] {Intro}   
## [8] {Cat.Data,Forecast,DOE,SW}   
## [9] {Intro}   
## [10] {Cat.Data}   
## [11] {Intro}   
## [12] {DataMining}   
## [13] {DataMining}   
## [14] {DataMining,Survey,Forecast}   
## [15] {DataMining,Survey}   
## [16] {Intro,Survey,Cat.Data,Forecast,DOE,SW}  
## [17] {Intro,Regression}   
## [18] {Intro,Forecast}   
## [19] {Intro}   
## [20] {Regression}

**Figure 4**  
*Association Rules with 0.1 Confidence*

rules <- apriori(mod6course.trans,parameter = list(supp = 0.01, conf = 0.1,target = "rules"))

## Apriori  
##   
## Parameter specification:  
## confidence minval smax arem aval originalSupport maxtime support minlen  
## 0.1 0.1 1 none FALSE TRUE 5 0.01 1  
## maxlen target ext  
## 10 rules TRUE  
##   
## Algorithmic control:  
## filter tree heap memopt load sort verbose  
## 0.1 TRUE TRUE FALSE TRUE 2 TRUE  
##   
## Absolute minimum support count: 3   
##   
## set item appearances ...[0 item(s)] done [0.00s].  
## set transactions ...[8 item(s), 365 transaction(s)] done [0.00s].  
## sorting and recoding items ... [8 item(s)] done [0.00s].  
## creating transaction tree ... done [0.00s].  
## checking subsets of size 1 2 3 4 done [0.00s].  
## writing ... [232 rule(s)] done [0.00s].  
## creating S4 object ... done [0.00s].

options(digits = 3)  
inspect(head(sort(rules, by = "lift"), n = 5))

## lhs rhs support confidence coverage  
## [1] {Intro,Regression,Forecast} => {DataMining} 0.0137 0.714 0.0192   
## [2] {Intro,Survey,DOE} => {Cat.Data} 0.0110 0.800 0.0137   
## [3] {Intro,DataMining,Cat.Data} => {Regression} 0.0164 0.750 0.0219   
## [4] {Intro,DataMining,Regression} => {Forecast} 0.0137 0.500 0.0274   
## [5] {Intro,Survey,Cat.Data} => {Forecast} 0.0137 0.500 0.0274   
## lift count  
## [1] 4.01 5   
## [2] 3.84 4   
## [3] 3.60 6   
## [4] 3.58 5   
## [5] 3.58 5

**Figure 5**  
*Association Rules with 0.5 Confidence*

rules2 <- apriori(mod6course.trans,parameter = list(supp = 0.01, conf = 0.5,target = "rules"))

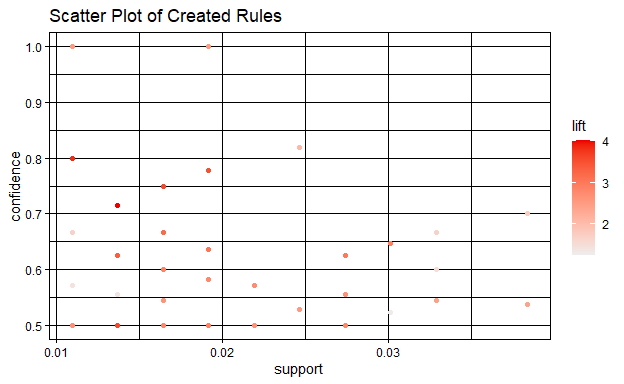
## Apriori  
##   
## Parameter specification:  
## confidence minval smax arem aval originalSupport maxtime support minlen  
## 0.5 0.1 1 none FALSE TRUE 5 0.01 1  
## maxlen target ext  
## 10 rules TRUE  
##   
## Algorithmic control:  
## filter tree heap memopt load sort verbose  
## 0.1 TRUE TRUE FALSE TRUE 2 TRUE  
##   
## Absolute minimum support count: 3   
##   
## set item appearances ...[0 item(s)] done [0.00s].  
## set transactions ...[8 item(s), 365 transaction(s)] done [0.00s].  
## sorting and recoding items ... [8 item(s)] done [0.00s].  
## creating transaction tree ... done [0.00s].  
## checking subsets of size 1 2 3 4 done [0.00s].  
## writing ... [54 rule(s)] done [0.00s].  
## creating S4 object ... done [0.00s].

inspect(head(sort(rules2, by = "lift"), n = 5))

## lhs rhs support confidence coverage  
## [1] {Intro,Regression,Forecast} => {DataMining} 0.0137 0.714 0.0192   
## [2] {Intro,Survey,DOE} => {Cat.Data} 0.0110 0.800 0.0137   
## [3] {Intro,DataMining,Cat.Data} => {Regression} 0.0164 0.750 0.0219   
## [4] {Intro,DataMining,Regression} => {Forecast} 0.0137 0.500 0.0274   
## [5] {Intro,Survey,Cat.Data} => {Forecast} 0.0137 0.500 0.0274   
## lift count  
## [1] 4.01 5   
## [2] 3.84 4   
## [3] 3.60 6   
## [4] 3.58 5   
## [5] 3.58 5

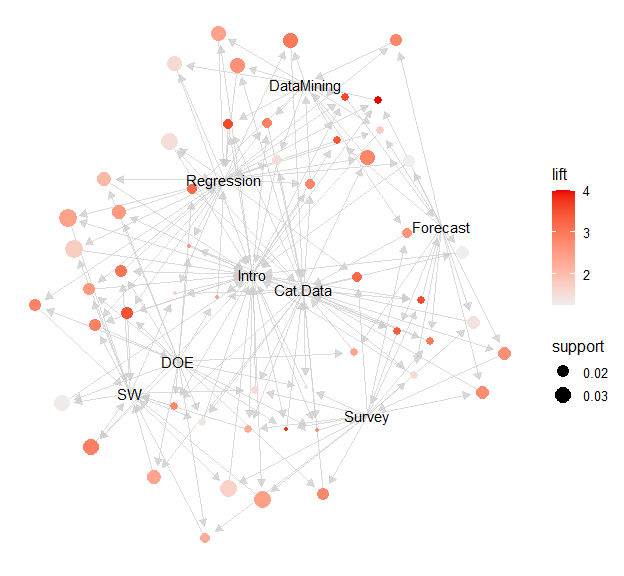
**Figure 6**  
*Scatter Plot of Rules*

plot(rules2, jitter = 0, main = "Scatter Plot of Created Rules")



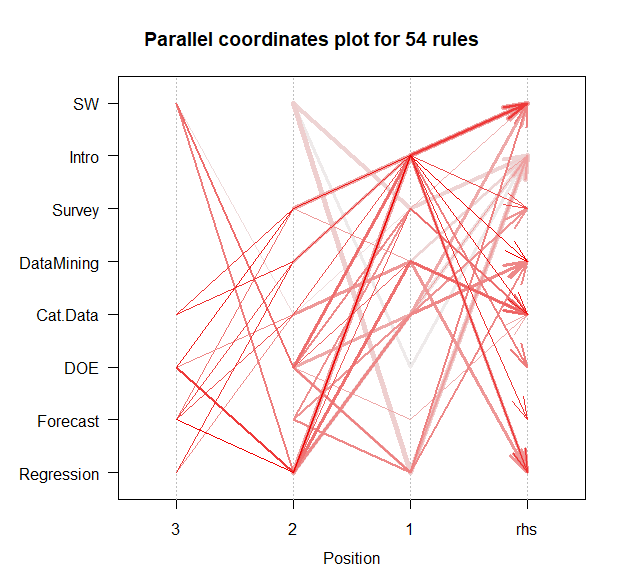
**Figure 7** *Graph of Rules*

plot(rules2, method = "graph")



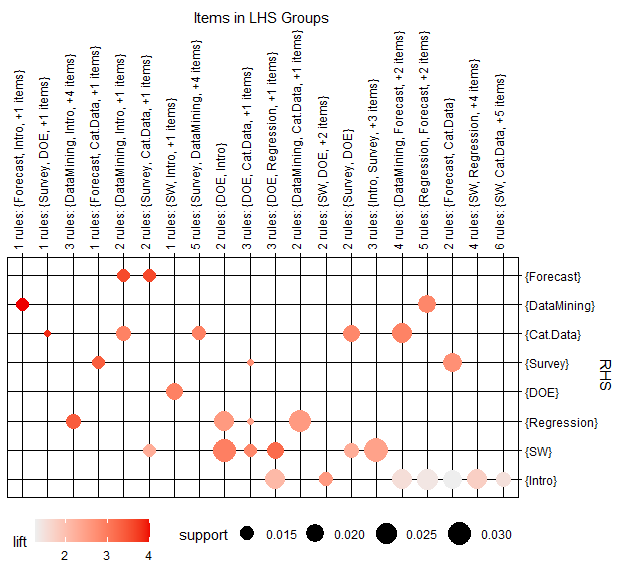
**Figure 8** *Parallel Coordinates Plot of Rules*

plot(rules2, method = "paracoord", control = list(reorder = TRUE))



**Figure 9**  
*Item Groups Plot of Rules*

plot(rules2, method = "grouped")



# Description and Interpretation

After loading appropriate libraries and importing the *Coursetopics.csv* to R (Figure 1), the data set is explored using the head and str commands (Figure 2). Head displays the first 6 rows of data and reveals eight columns of courses data. The rows represent each student observation. If students are enrolled in the course, the column value is one, and if not enrolled, the value is zero. The data does not initially have any columns that need to be removed. Str or the structure command lists the data set dimensions and the attribute labels along with attribute type. The data frame has eight columns and 365 rows of student observations. All attributes are integer. Next in Figure 3, the data frame is converted to a matrix with the as.matrix command, and columns are transformed to transaction data creating the *mod6course.trans* matrix. Running a summary of this matrix verifies we still have the same columns and rows as before. Summary also provides the most frequent courses selected in the data. The structure str function is run again which verifies the matrix is transaction data. Lastly, the inspect command is run on the first 20 rows to verify the data within each row is still intact.

With some exploration complete, the data can now be passed to the technology to find itemsets and possible associations. Several algorithms can be used to compile association rules. The Apriori algorithm is one of the most common and will be used here. This algorithm starts by identifying all single items, then moves on to find all two-item itemsets, then three-item itemsets, etc. until all possible itemsets are found. In Figure 4, the Apriori algorithm generates a list of association rules using the apriori function given the parameters defined in the assignment. For the first example, the minimum support is 0.01 and the confidence parameter is 0.1. From this, R Studio creates 232 rules shown in the results of the algorithm. Next, the rules are sorted by lift, and the first five rows are displayed using the `inspect’ function. These rules follow an if/then template of *if antecedent(s), then consequent* (Schmueli et al, 2018). The five rules can be interpreted into the following five statements.

**1. If a student enrolls in Intro, Regression, and Forecast courses, then with confidence 71.4% the student will also enroll in the DataMining course. This rule has a lift ratio of 4.01.**  
**2. If a student enrolls in Intro, Survey, and DOE, then with confidence 80% the student will also enroll in Cat.Data. This rule has a lift ratio of 3.84**  
**3. If a student enrolls in Intro, DataMining, and Cat.Data, then with confidence 75% the student will also enroll in Regression. This rule has a lift ratio of 3.60**  
**4. If a student enrolls in Intro, DataMining, and Regression, then with confidence 50% the student will also enroll in Forecast. This rule has a lift ratio of 3.58.**  
**5. If a student enrolls in Intro, Survey, and Cat.Data, then with confidence 50% the student will also enroll in Forecast. This rule has a lift ratio of 3.58.**

For these statements, confidence indicates the rate at which the consequent and antecedent(s) co-occur, and the lift ratio is a measure of rule strength against a benchmark value. The assignment also instructs to re-run the Apriori model using the same minimum support with a confidence of 0.5. These results are shown in Figure 5. The top five rules results by lift did not change, however, the number of rules generated did decrease to 54. This second model will be used for the final exploration using visualizations.

The first visualization (Figure 6) is a scatter plot of the rules by support (x-axis) and confidence (y-axis). Support indicates how many transactions are affected, and confidence shows what rate a consequent will be found. The lift ratio for each rule is indicated by the color intensity at each point. Next, a graph of the rules is created using the plot method type *graph* (Figure 7). This graph is complex, however, it is clear to see the course *Intro* is involved in many of the rule relationships. Support is represented by the size of the circle, and lift is represented by color intensity. Using the plot method *paracoord* produces a parallel coordinates plot of the 54 rules (see Figure 8). Again, this plot has many rule associations pointing to the *Intro* course. Many institutions require incoming students to attend an introductory course. If true, then this column of data would only contain the single result of one to indicate all students were assigned to the course, and subsequently, the column should be removed from the model. However, after checking in Excel, the data in the column is mixed and therefore will remain.

The final graph (Figure 9) is a balloon matrix plot that lists the consequent items on the x-axis labeled to the right as RHS (right-hand side), and the groups of antecedents on the y-axis labeled across the top as LHS (left-hand side). Again, the size of the balloon indicates the support and the darker color intensity indicates higher lift values. For example, the rule with antecedents *Forecast, Intro, plus one item* has a high lift ratio with the *DataMining* consequent. By contrast, the rule that contains *Regression, Forecast plus two items* have higher support but a lower lift ratio with *Intro*.

# References:

Colorado State University Global (CSUG). (2021). *Module 6: Relationships in unsupervised learning* [Interactive lecture]. Canvas. <https://portal.csuglobal.edu>.

RStudio Team. (2021). *RStudio: Integrated development environment for R. RStudio*. PBC, Boston, MA. <http://www.rstudio.com/>.

Shmueli, G., Bruce, P.C., Yahav, I., Patel, N.R., and Lichtendahl, K.C. (2018). *Data mining for business analytics: Concepts, techniques, and application sin R.* Wiley Publishing. ISBN: 9781118879337.