Joyce Woznica

jlwoznic@syr.edu

October 27, 2019

Joyce Woznica  
Homework #3

Contents

[Introduction 3](#_Toc22835536)

[Analysis 3](#_Toc22835537)

[The Data 3](#_Toc22835538)

[Data Load, Cleanse, Munge and Preparation 3](#_Toc22835539)

[Exploratory Data Analysis 5](#_Toc22835540)

[Descriptive Statistics 5](#_Toc22835541)

[Visualization 7](#_Toc22835542)

[Results 10](#_Toc22835543)

[Rules of Interest 10](#_Toc22835544)

[Unmarried, No Children, No Mortgage or Car 10](#_Toc22835545)

[Unmarried, No Children, No Mortgage 10](#_Toc22835546)

[Forties with a Single Child 10](#_Toc22835547)

[Female, Middle Income, One Child 11](#_Toc22835548)

[Middle Income with a Single Child 11](#_Toc22835549)

[Conclusions 12](#_Toc22835550)

List of Figures

[Figure 2: Data set Description 3](#_Toc22835551)

[Figure 3: Bank Data Column Names 4](#_Toc22835552)

[Figure 4: Verification of No Missing Data 4](#_Toc22835553)

[Figure 4: Structure of the Bank data 4](#_Toc22835554)

[Figure 5: Data set after Age Grouping 5](#_Toc22835555)

[Figure 6: R Code to transform data set 5](#_Toc22835556)

[Figure 7: Data set after Cleaning and Manipulation 5](#_Toc22835557)

[Figure 8: Function to run Association Rules 6](#_Toc22835558)

[Figure 8: Calling myARM Function 6](#_Toc22835559)

[Figure 10: Ruling down the Rules 6](#_Toc22835560)

[Figure 11: First Twenty Rules obtained with Support, Confidence and Life 7](#_Toc22835561)

[Figure 12: Interactive Plot of Top Ten Rules by Decreasing Support 7](#_Toc22835562)

[Figure 13: Interactive Plot of Top Ten Rules Sorted by Decreasing Confidence 8](#_Toc22835563)

[Figure 14: Scatterplot of Top Twenty Rules by Decreasing Confidence 9](#_Toc22835564)

[Figure 15: Rule 13 with Support=0.075, Confidence=0.938 and Lift=2.05 9](#_Toc22835565)

# 

# Introduction

The marketing department maintains information for a bank on customers which includes data on their demographics and the accounts that they currently hold with the bank. This is very important when launching new products and determine which customers to target for these products. When launching their “Personal Equity Plan” (PEP)[[1]](#footnote-1), the marketing department gathered information about the customers that purchased this PEP and hope to use that information to build better customer profiles.

This document outlines the data collected, the loading process, any modifications required to the data and why, initial visualizations and information about the data, the results of association rule mining on the data.

# Analysis

## The Data

For each customer the following data was collected by the bank for analysis.

| **Variable** | **Definition** |
| --- | --- |
| ***id*** | Customer unique identification number |
| ***age*** | Customer age in years |
| ***sex*** | Gender of customer – MALE or FEMALE |
| ***region*** | Represents if the customer lives in the INNER\_CITY, RURAL, SUBURBAN or TOWN |
| ***income*** | Customer income |
| ***married*** | Marital status of customer – YES or NO |
| ***children*** | Number of children for customer |
| ***car*** | Customer owns a car – YES or NO |
| ***save\_act*** | Customer has a savings account – YES or NO |
| ***current\_act*** | Customer has a current account at the bank – YES or NO |
| ***mortgage*** | Customer has a mortgage – YES or NO |
| ***pep*** | Did the customer purchase a PEP after the last mailing from the bank – YES or NO |

Figure 2: Data set Description

Each record, there are 600 customer records with 12 characteristics, makes up a customer description where the final field, ***pep***, indicating if that particular customer purchased a Personal Equity Plan.

### Data Load, Cleanse, Munge and Preparation

#### Data Load

To load the School Data, the *read.csv* function was used with any empty data replaced with spaces. The column names were left as is for the analysis. An example of the first few rows of the “BankDF” data frame are shown in the following figure.

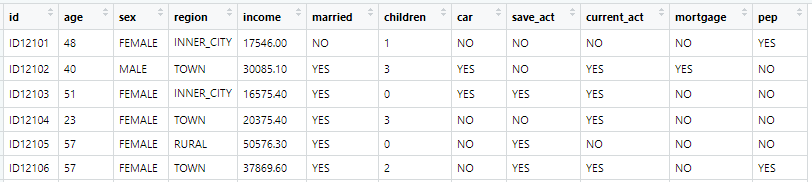


Figure 3: Bank Data Column Names

#### Data Cleanse

A check was run on this resulting loaded data using *is.na* to determine if any fields were missing data. It was determined by this verification that there was no missing data.



Figure 4: Verification of No Missing Data

#### Data Munge and Preparation

A verification of the structure of the bank data frame was run to determine what types of data manipulation may be required to prepare the data set for association rule mining. This is shown in the following figure.

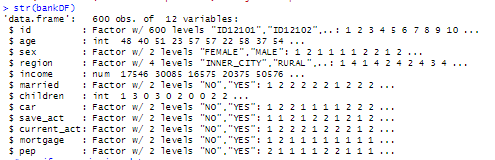


Figure 4: Structure of the Bank data

To reconfigure the data for the best possible usage for association rule discovery, several manipulations were executed. The first was taken from examples provided by Jeremy Bolton’s code to group the ages into “bins.” This was done using the code in the following figure and the results of the first few rows of the data frame are also presented.

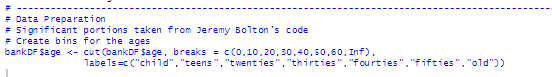
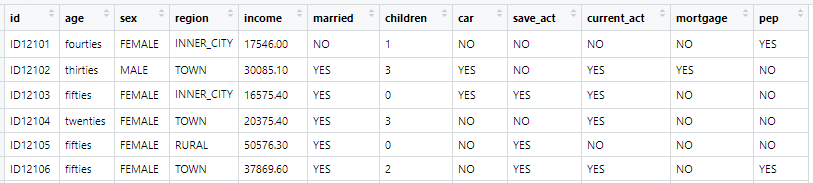
 

Figure 5: Data set after Age Grouping

Additional manipulation to the data frame was required using the commands presented in the following figure.

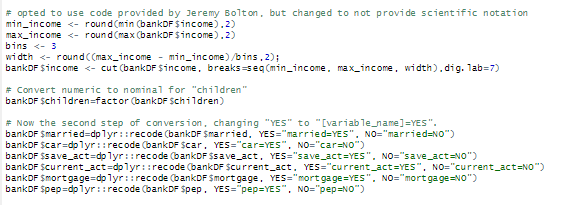


Figure 6: R Code to transform data set

The resulting data set is reflected in the first few rows shown in the following figure.

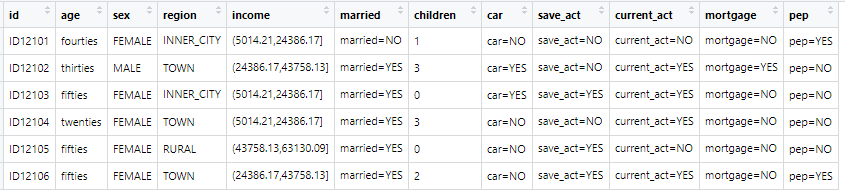


Figure 7: Data set after Cleaning and Manipulation

## Exploratory Data Analysis

### Descriptive Statistics

The next step was to run association rule discovery. In performing this analysis, different parameters were used to determine the strongest rules for this particular data set. Approaching the data set in a way to provide rules that will result in purchasing the Personal Equity Plan, it is important to set the rules in in such a way that the “right hand side” of all rules is set to a result of “pep=YES”. As a result, the following function was created to return the rules of interest.

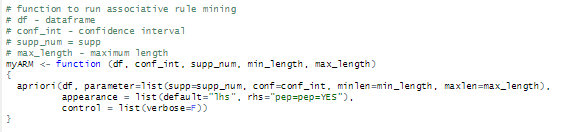


Figure 8: Function to run Association Rules

Using this method, a simple call could be made to manipulate the parameters for the *apriori* function to bring a different rule set. These results could then be inspected and/or sorted by confidences or support so that a better understanding of the rule could be obtained.

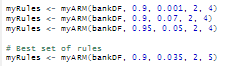


Figure 8: Calling myARM Function

Many different combinations of confidence, support, minimum length and maximum length were evaluated. In some cases, selecting a higher minimum with a higher confidence provided less than twenty rules for evaluation. In other cases, raising the minimum confidence and support numbers provided limited to no results.

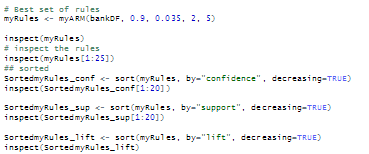


Figure 10: Ruling down the Rules

For a solid set of rules, it was determined that using the following parameters gave strong rules for the combination of characteristics that would result in a customer purchasing a Personal Equity Plan.:

* Confidence minimum of 0.9
* Support minimum of 0.035
* Minimum number for ‘left hand side’ = 2
* Maximum number of ‘left hand side’ = 5

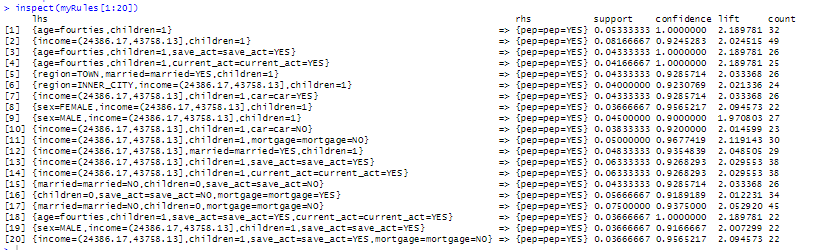


Figure 11: First Twenty Rules obtained with Support, Confidence and Life

### Visualization

Some interesting plots were run on the rules to show some visualization of what was taking place. The first was plotting the top ten sorted rules based on support in an interactive map.

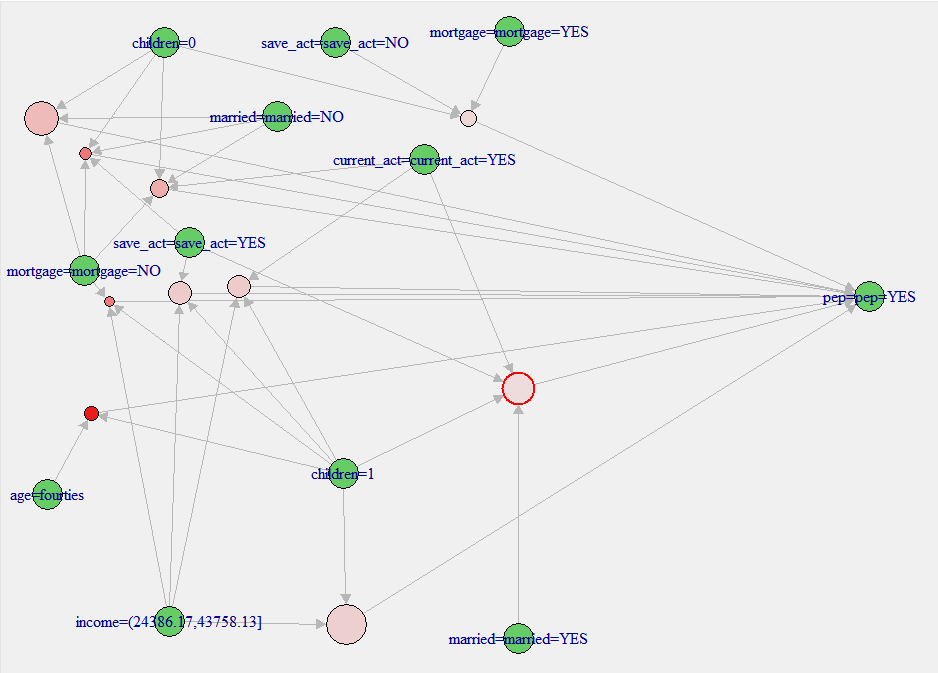


Figure 12: Interactive Plot of Top Ten Rules by Decreasing Support

This was followed by a similar visualization of the top ten rules sorted by decreasing confidence which clearly showed some very different results.

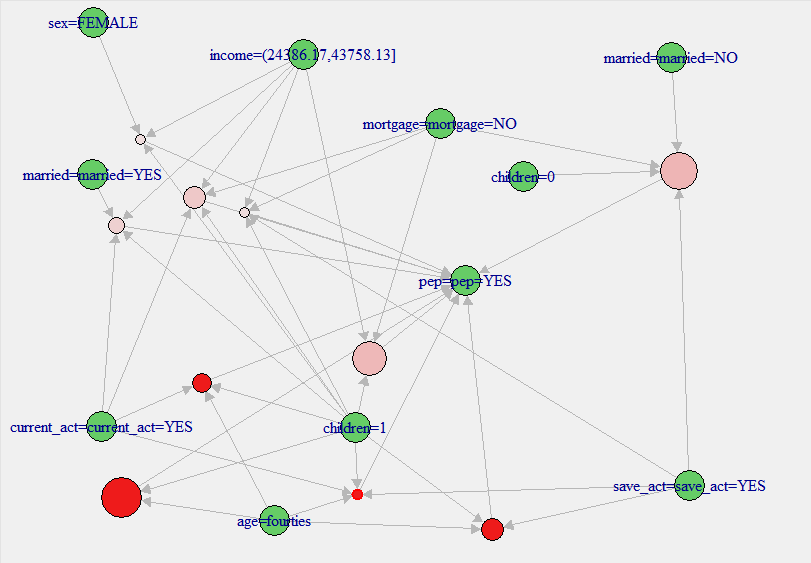


Figure 13: Interactive Plot of Top Ten Rules Sorted by Decreasing Confidence

A visualization was then done using a deprecated function that still provided the plot, *plotly\_arules* which shows a very interesting scatterplot of the rules. The following figure shows the top twenty rules sorted by decreasing confidence displayed with support on the *x*-axis and Confidence on the *y­*-axis and color coded by lift.

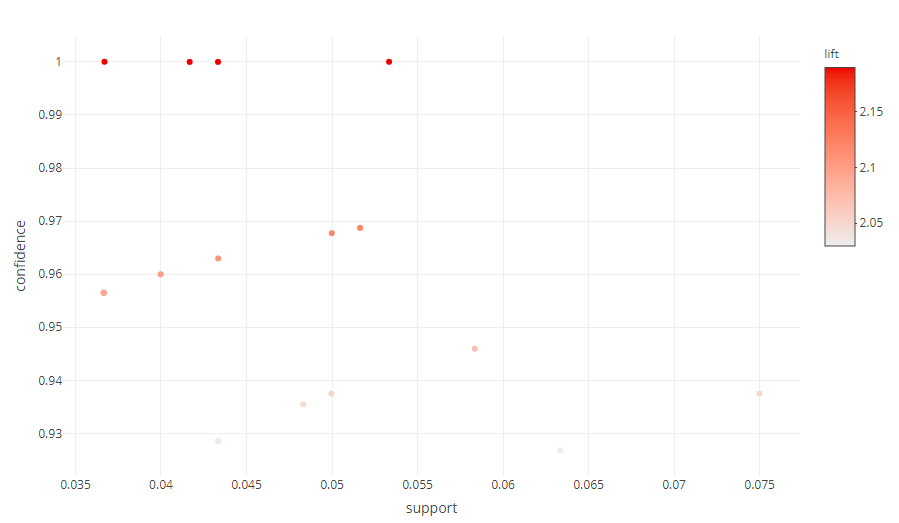


Figure 14: Scatterplot of Top Twenty Rules by Decreasing Confidence

Not only does this show an interesting grouping of the rules, but it is interactive as well. For example, by hovering over a point on the plot, the rules are displayed as shown in the following figure.

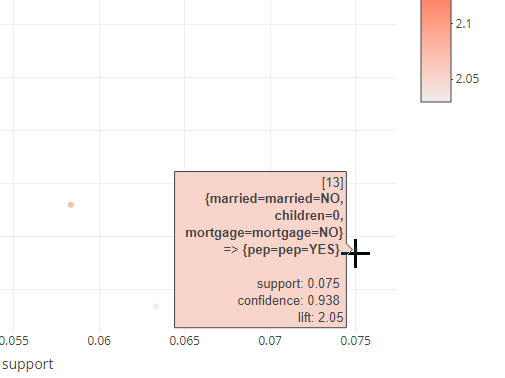


Figure 15: Rule 13 with Support=0.075, Confidence=0.938 and Lift=2.05

This final plotting method served useful in selecting rules of interest that are discussed in the Results section.

# Results

As shown in the previous section, many interesting rules were generated by the analysis and based on this information, several interesting rules have been noted as well as explanation and recommendations based on these rules.

## Rules of Interest

### Unmarried, No Children, No Mortgage or Car

The first rule of interest, rule 34, is the following:



This rule is interesting for several reasons, but primarily because this customer is not married, has no children, no car and no mortgage. At first it may seem odd that an individual with this history would be interested in an equity plan. However, this could be because he/she has no solid credit history or needs assistance in overal money management so that he/she can begin to accumulate some overall wealth. This is an opportunity for the bank to target unmarried customers without children meeting these other criteria to help them invest and grown their equity with the firm.

### Unmarried, No Children, No Mortgage

The second rule of interest, rule 17, is the following:



This rule is interesting for similar reasons of the previous rule. For a customer to be married with no children and no mortgage, one would have to assume that this individual is again attempting to build wealth or a credit history to enable moving forward with things like a mortgage in the future. This again poses a very good opportunity for the bank to target this demographic to build customer loyalty.

### Forties with a Single Child

The next rule of interest, rule 1, is the following:



This rule had the highest confidence in comparison to our other rules and can be considered a bit predictable. However, it has a very high confidence and a moderatly high support and lift. This rule in interesting because of the lack of other characteristics showing up as part of the left hand side of the rule – it is simple with only age=forties and one child. This customer looks to be planning for the future and his/her child early enough to make a difference with his/her investment approach. Again, this demographic, married or not, may make a good target for new products offered by the bank.

### Female, Middle Income, One Child

The fourth rule of interest, rule 8, is the following:



In reviewing this rule, it can be seen that the confidence is quite high and the lift is solid. Females with a single child could represent single parents, since married is not indicated as part of the rule. This is an excellent target for the bank to find women earning a decent wage that are looking to build equity. Other products related to investing for their children and buidling their retirement may be good programs for this demographic.

### Middle Income with a Single Child

The final rule of interest, rule 2, is the following:



This rule has a very high support compared to the other rules and a high confidence as well.

In this case, two out of the actual ten characteristics of interest (income and children) are being used to generate a fairly strong support number of ~0.08167. Support provides information about how frequent items like this income bracket and one child appear in a set in all the transactions. The more items, usually the lower the support.

The confidence is very close to one in this rule. Confidence is essentially the conditional probability of the occurrence of the consequent given the antecdent. However, this can be misleading even if the confidence is quite high because pep=YES appears often in the data set (274 times).

By adding lift into the mix, a better picture can be obtained of the strength of the rule. Lift considers both the support of the rule and the overall data set. If the lift is exactly 1, then the left hand side and right hand side are independent of each other. If the lift is greater than 1, this points to a stronger dependence on each other.

The combination of confidence, lift and support of this rule make it strong overall.

Similar to the previous rule, it can indicate that this middle income might be interested in saving for the future and this could be prompted by the child in the home. Using this information, again targeting mutual funds, UTMA or 529 plans for their child might prove successful in this market.

# Conclusions

As a result of the analysis done on this data, general realizations can be made from the rules generated:

* Having at least one child implies purchasing a PEP
* Having the second to the highest income bracket (24386.17,43758.13] but not the highest income bracket (43758.13,63130.09] implies purchasing a PEP
* Having a mortgage was not a major contributing factor
* Highest confidence was from customers in the forties
* Being married did not have significant affect
* Having a savings account played in a role in purchase a PEP

Based on this information and the rules from the previous section, it is recommended that customers with a single child in the second to highest income bracket be targeted for new products around investment opportunities and growing wealth. In addition, products around investments for their child’s future may be good for the target demographics discussed may prove profitable for the firm. A similar campaign could be run for customers in their forties to grow equity and increase retirement earnings would also be a good opportunity for the bank.

1. Note: A “Personal Equity Plan” was introduced in the United Kingdom in 1986 and were replaced by Individual Savings Accounts in 1999 and are no longer a viable investment option. [↑](#footnote-ref-1)