Data 630 – Fall 2018

Assignment 1 – Association Rules in R

Kenneth Lulie

Dr. Gates

University of Maryland University College

October 1, 2018

**Introduction**

The study of compensation for labor is important to current and prospective employees, their employers and the government as well. Understanding the factors that lead to different wage outcomes allows students to make choices to improve their outcomes. A major issue of today is the student loan debt crisis, with the average student graduating in 2016 owing over $37,000 in student loans (Friedman, 2018). Access to accurate information regarding wage outcomes of different majors and universities allows college students to compare their options.

Additionally, this field of study allows the government to understand potential causes of income inequality of women and minorities. Understanding these causes allows the government to evaluate what the most important factors are, and then intervene to improve those outcomes. For example, the study of compensation showed that women are underrepresented in STEM majors which are generally have higher compensation than other majors (Hill, 2018). Today, there is an awareness of this issue and a result there are many programs and scholarships targeted at increasing the rate of women entering STEM majors.

The objective of this paper is knowledge discovery of wage data using association rules. Association rules can assist in understanding which different factors combine to lead to different outcomes (Han, 2011). The rules generated by this method are easy to understand which eases the journey from discovery to action and can often provide insights that would be impossible or difficult to find without some form of machine learning.

**Analysis and Model Demonstration**

**Data Information and Cleaning**

The data being used in this analysis is a file named “wages.csv” provided as a class resource. This dataset has 534 observations of 11 variables. There were no unique ID variables. These variables describe the characteristics of wage earners and the amount of money they earned (See Figure 1). The South variable is the only one initially provided as a factor with the levels of “yes” and “no” presumably indicating if the wage earner is located in the south. Education, Age, and Experience are provided as integer variables, presumably in the unit of years. Sex, Union, and Marital Status are provided as integer variables of 1 or 0. Race, Occupation, and Sectors are provided as integers, presumably translated to represent different groups. Wage is provided as a number.

Figure 1 Wage Data



The data was reviewed to identify any cases of missing data. The first step taken was to review for any cases of a “NA” value in the data. However, no values of NA were discovered in the data. Next, the data was reviewed using the summary command to look for outliers that could indicate problems with the data, such as value of 0 in a field such as age or wage. Again, the data was of high quality with no observations being of suspect nature. The summary data showed observations that fit the profile of working class adults. The age ranged from 18 to 64, the traditional working ages. The education ranged from 2 to 18 showing the variety of education in the working public. The wage variable also ranged from 1 to 44.5, in unknown units.

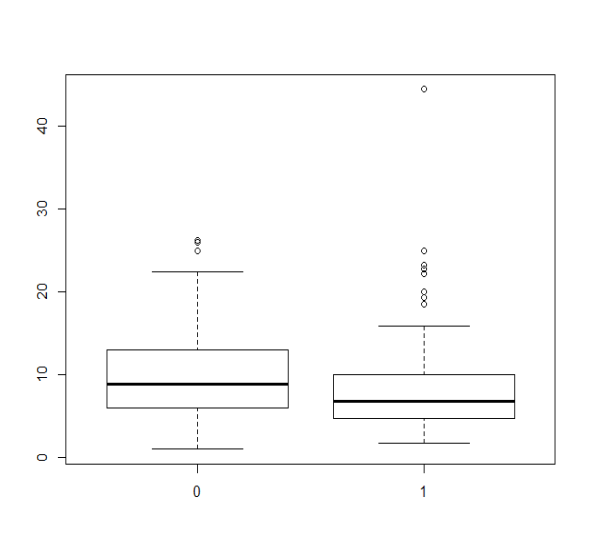
 As the objective of this paper is to discover knowledge about the factors that impact wage outcomes, different variables were reviewed in their relationship to wages using the aggregate command. The average value of wage per different grouping of sex and union was calculated (See Figure 2). This data showed that workers with a union flag of 1 earned 25% more wages than those with a flag of 0. This indicates that being a union member could increase wages. When the average value of wage was reviewed by sex, we saw that one gender earned nearly 26% more on average than the other gender. This indicates an inequality e in wage outcomes become genders.

Figure 3 Boxplot of Wage by Sex

Figure 2 Average value of Wage by Sex and Union Status shows imbalance.



The imbalance in wages by sex could be driven by outliers. To explore further, a boxplot was generated (See Figure 3). The boxplot shows us that the median values for each gender are also imbalanced as the mean is, indicating that this outcome is not driven by outliers in the data.

**Data Preprocessing**

To perform association rule mining, the variables must be in factor form. However, the data provided only has one variable initially in factor form. The rest must be factorized. Sex, Union, Marital Status, Race, Occupation, and Sectors were factorized using their existing values as the summary data showed they were in integer form with 6 or less values each.

However, despite being integers Experience and Age each had a range of around 40. This is too many values to simply factorize, as the association mining rules would have too many values to run efficiently. Therefore, these variables were factorized using the discretize command with the frequency method and a break parameter of 6. This divided the values into 6 ranges, each with an equal number of observations but not equal intervals. The education variable represented a special case, as it is one of the most important factors in the dataset that can be affected by the decisions of the worker. Education is widely understood in preexisting categories of years, ranging from no high school diploma, high school diploma, some college, bachelors degree, and post bachelor education. Therefore, this variable was factorized using the discretization command with fixed intervals equaling the previous mentioned categories. This will allow for easier understanding of the impact of different educational choices.

Finally, the wage variable was also factorized using the frequency method. Initially, the number of breaks was set to 6. However, this was found to cause difficulties in the model stage. Therefore, a new value of 3 breaks was selected to represent low earners, average earners, and above average earners. These ranges were 1 to less than 6, 6 to less than 10, and 10 to 44.5.

**Subsection: Analysis and Model Methods**

Association rules is a knowledge discovery from data technique to discover relationships between itemsets (Han, 2011). The most common used algorithm for this purpose is the aprioi algorithm explained by R. Agrawal and R. Skirant in 1994. Since then, this algorithm has been used for different purposes, although it is most famous for identifying frequently bought together items at retailers.

Association rules represent patterns for items or values that are frequently associated or found together (Hill, 2011). A hypothetical example of an association rule would be “Windshield Wiper ≥ Windshield Wiper Fluid [support = 5%, confidence = 80%, lift= 2.36]. This rule says that if a windshield wiper is found in an observation, for example a purchase at a retailer, that purchase will also include Windshield Wiper Fluid. The support indicates that this combination is found in 5% of the total data, and the confidence indicates that in the item sets where this combination is found, this rule is true 80% of the time. The lift indicates the correlation of the association. A lift of 1 indicates that they are independent, and a lift above 1 indicates positive correlation. The lift is one of the most important values to measure an association rule by.

The confidence and the support values are important parameters to be selected when running the algorithm. If the confidence value is too low, the rule might be mere noise or chance. If the confidence value is too low, the rule may not be accurate enough for your purposes. Another important parameter to select is the minimum and maximum length of the rule to ensure your rules are the appropriate complexity.

While there are different methods of generating association rules, the most used is the Apriori algorithm. This algorithm uses the Apriori property, which states that all non-empty subsets of a frequent item set must also be frequent. This is an antimonotonicity property, which means that a set cannot pass a test, all of its subsets will fail the same test as well (Hill, 2011). This allows the algorithm to be more efficient when running. The Apriori algorithm uses a “bottom up” method where candidate generation is run testing the parameters and frequent subsets are extended one item at a time. A pruning stage is also typically used to eliminate duplicative rules.

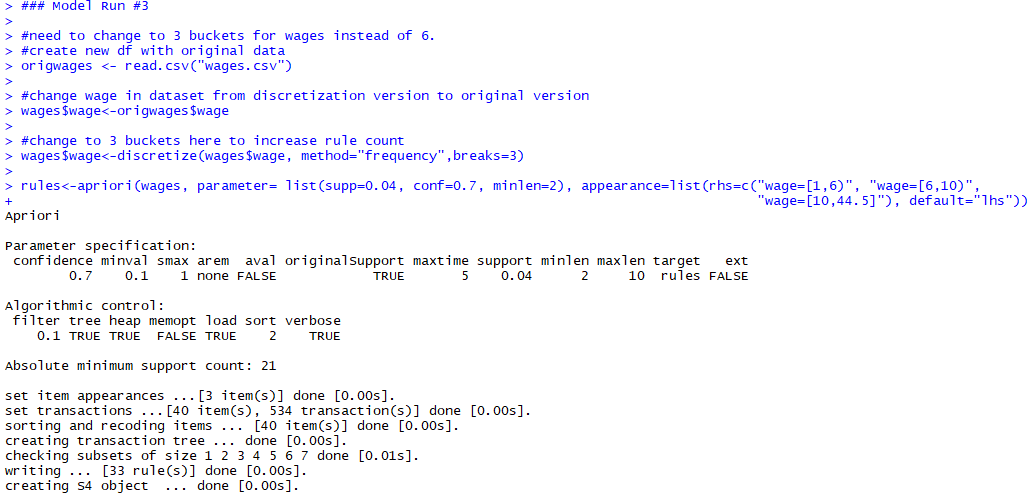
**Association Rules Model**

The first model was generated using R Studio with the “apriori” command from the “arules” library (See Figure 13). The assumptions for this model is that the data is accurate and that the support and confidence values chosen are statistically significant enough for the purpose of understanding the data. The parameters used to build the first model required a support of 5%, a confidence of 70%, a minimum length of 2 items, and a maximum length of 10 items. The model also requires that the right hand side of the rule include the wage variable. The first run of this model showed 0 results (See Figure 7).

The second model was generated with the same parameters except lowering the support to .2% and the confidence to 60%. This resulted in a set of 21 rules (See Figure 8). These rules were pruned by creating a matrix using the subset command, and then comparing this subset to find redundant rules. These redundant rules were removed from the set of rules, resulting in a final count of only 17 rules (See Figure 9).

After lowering the parameters for the second model and still not achieving powerful results, the target variable was suspected to have too many values at 6 which made it difficult for the algorithm to find rules that satisfied the support levels defined. The wage variable was refactored into 3 ranges down from 6, and the algorithm was rerun with support raised to 4% and the confidence was raised to 70% and otherwise default parameters (See Figure 4). For this model, there were 23 rules remaining after pruning (See Figure 10).

Figure 4 - Refactoring Wages Resulted in Higher Quality Rules

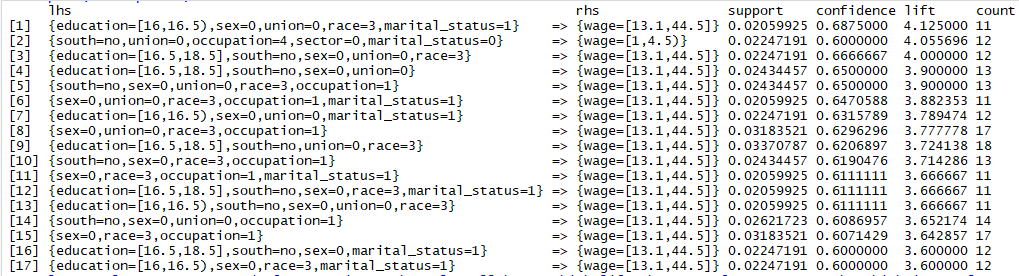


**Results and Model Evaluation**

**First & Second Models**

The output of the first model which required support of 5% or about 27 minimum observations, confidence of 70% and a minimum length of two items with a maximum of 10 items resulted in no rules being generated. The hypothesis for this result was that the parameters were too restrictive. This lead to the second model where the support was lowered to 2% or 11 minimum observations, and the confidence was lowered to 60% with the rest of the parameters remaining the same. After pruning, this resulted in 17 rules (See Figure 5).

Figure 5 - Model Two Rules Suffer from Poor Observation Counts and Low Accuracy

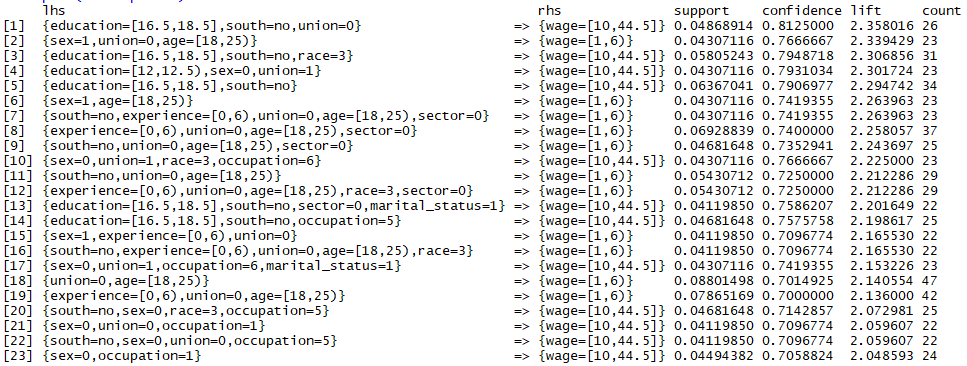


The rules generated by the second model had high lift scores, between 3.6 to 4.125. This generally indicates rules that would be of interest as that indicates they are highly positively correlated. However, as the parameters were lowered to only require 2% support, this combined with the relatively small sample size resulted in rules with very low counts of observations. The rule with the highest support had only 18 observations supporting it. Additionally, the confidence was low with every single rule between 60 and 70% confidence indicating these specific rules were barely right more often than they were wrong. This led to a conclusion that these rules were likely not reliable and could simply be over-fitting to random noise. Further, all the rules except for one predicted only the highest category of wages. Ideally, the model would generate rules that predict more than one outcome.

The analysis of the rules did reveal interesting trends that fit in with what we know about the factors that lead to wage outcomes. Many of the rules which predicted the highest wage outcome required an education level of 16 years or above which seems to confirm the common knowledge that higher education generally leads to higher wage earning outcomes. Additionally, a common item in the rules predicting high wages was that the employee was not located in the south, which in the United States has lower average wages.

**Third Model**

After the second model appeared to over-fit, the wage variable was refactored into only three ranges instead of the original six ranges to test the theory that less target ranges would allow for more robust rules of higher statistical quality. The parameters were raised accordingly to require support of 4% or 22 minimum observations to ensure that the rules were generally more robust. Additionally, the confidence level was raised to a minimum of 70% to make sure these rules had a higher standard of accuracy. After pruning, this resulted in 23 rules (See Figure 6).

Figure 6 - Model 3 Association Rules Show higher Quality

The rules generated by model 3 have lifts ranging from 2.05 to 2.36. This indicates that the rules are relatively strongly positively correlated which is a sign of health for the model. Additionally, the confidence of the rules ranges from a low of 70.5% to a higher of 81% again indicating much more accurate rules than were generated for the second model. Finally, the support ranged from 4.01% to 8.8% showing that these rules are more likely to be found true if the dataset was to be expanded and less likely to be the result of random noise. When reviewing the rules, it is important to note that the rules are split roughly predicting the highest or the lowest wage outcomes (See Figure 11). While ideally, rules including the middle category would also be generated, this isn’t a failure of the model to meet the objective because it the below average and high average outcomes that we are attempting to discover knowledge of.

The rule with the strongest confidence of 81.25% was that workers with an education level of over 16 years to 18 years, that aren’t in the south, and aren’t in a union will have the highest wage level of 10 to 44.5. An interesting inclusion in this rule is that the absence of a union predicts a higher wage, while unions are generally known for raising wages. An explanation could be that this criteria serves to exclude workers in the education sector where workers are generally more educated and wages are generally lower but unions are more prevalent. This explanation would also fit well with another rule that predicts if you are between 18-25 years of age, and not in a union you will be a below average workers. Additionally, many higher wage predicting rules require a sex of 0 showing a gender imbalance in the higher wage earner category.

**Conclusions**

The result of the models show that there were interesting factors to be found in the data to help understand what leads to different wage outcomes. From the initial data set we can see that wages are lower in the south, that there is a difference in wage outcome by sex, that higher education can be important to higher wages, and that unions are important for higher wages for younger people.

This exercise shows that knowledge discovery through data is capable of helping understand complex issues with many factors and variables. These insights can help us not only make decisions to improve outcomes at the individual level, but also at a societal level. As more and more data becomes available, discovering knowledge will be more and more important to maximizing different outcomes ranging from understanding wage data to medical outcomes.

**Limitations and Improvements**

The major limitation of this model was that the data set of 534 observations was rather low for this type of knowledge discovery. Additionally, drawing actionable conclusions is difficult as this is a test data set of unknown origin. For future improvements, using a dataset with a much greater and known population would allow for more improvements.

**References**

Why So Few? Women in Science, Technology, Engineering, and Mathematics. (n.d.). Retrieved from <https://www.aauw.org/research/why-so-few/>

Friedman, Z. (2018, June 13). Student Loan Debt Statistics In 2018: A $1.5 Trillion Crisis. Retrieved from <https://www.forbes.com/sites/zackfriedman/2018/06/13/student-loan-debt-statistics-2018/#73b368497310>

Han, Kamber, and Pei (2011). Data Mining: Concepts and Techniques, Third Edition Retrieved September 14, 2018 from http://hanj.cs.illinois.edu/cs412/bk3/01.pdf

Appendix A

Figure 7 - Model 1 Generated no Results

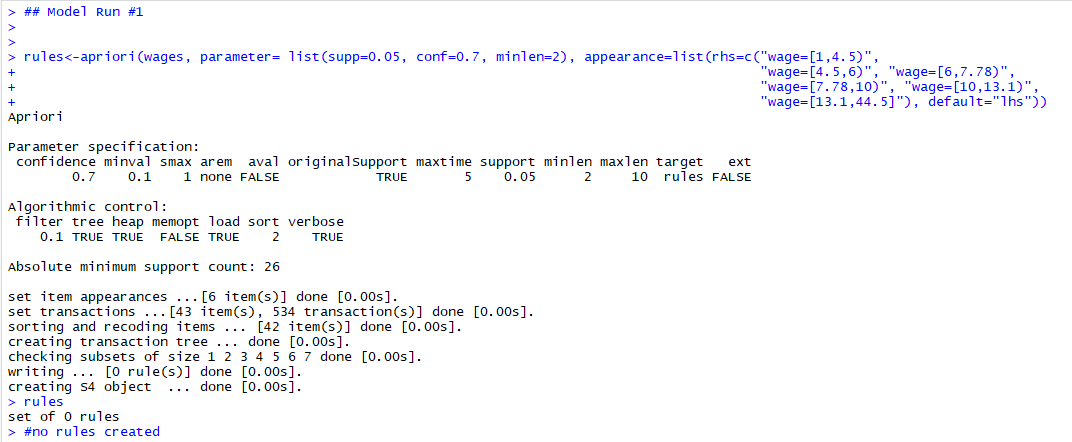
****

Figure 8 - Model Two Generated 21 Initial Rules

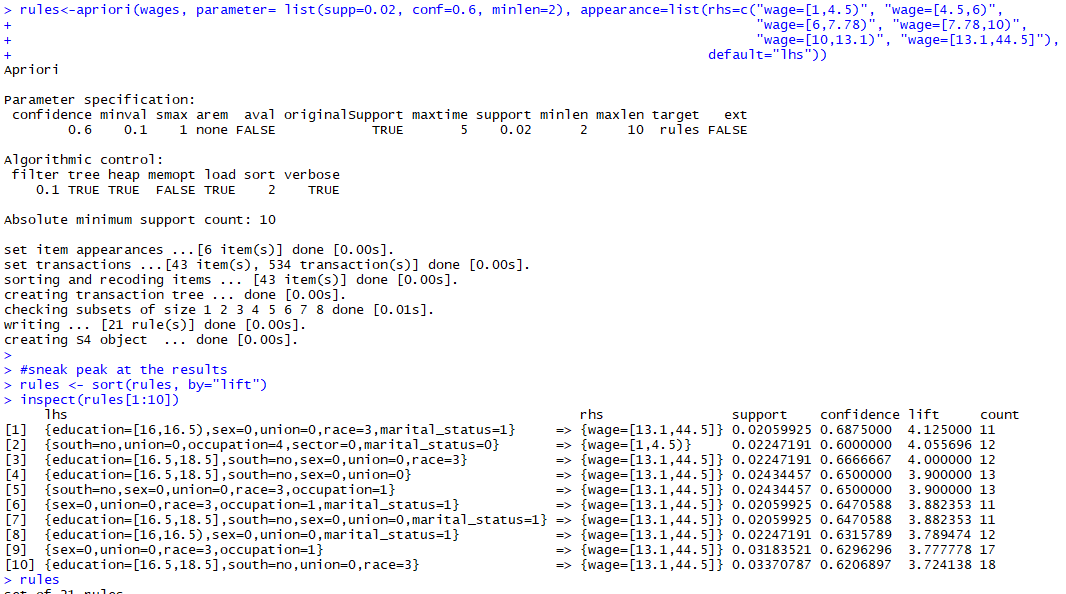
****

Figure 9 - Model 2 - 17 Rules remaining after Pruning Duplicative Rules

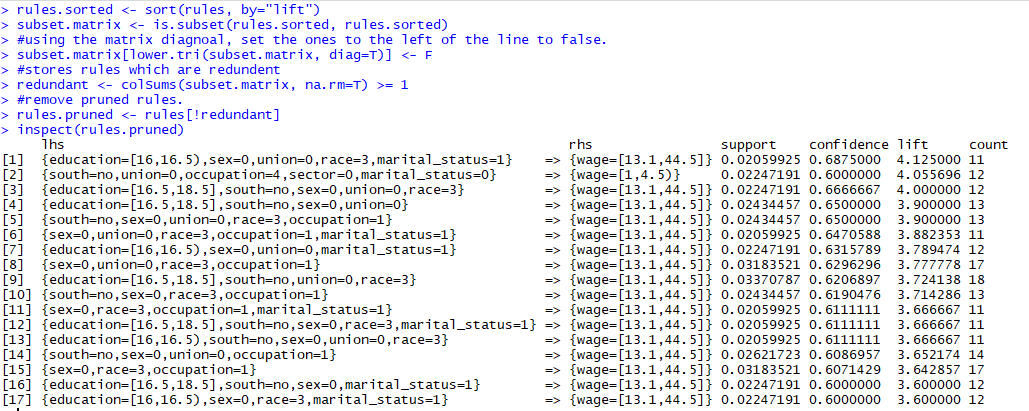
****

Figure 10 - Model 3 Resulted in 23 Rules after Pruning

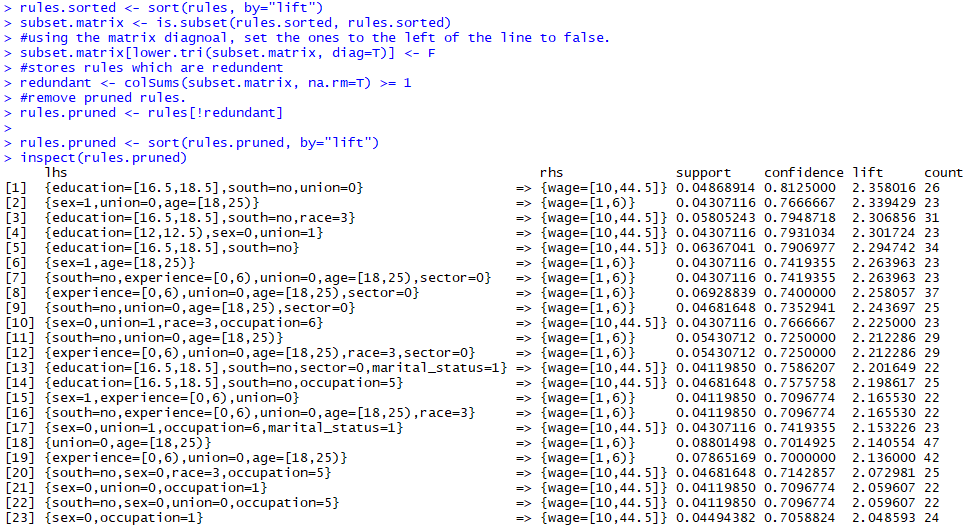
****

Figure 11 - Model 3 Rules Matrix Shows Mix in Lift

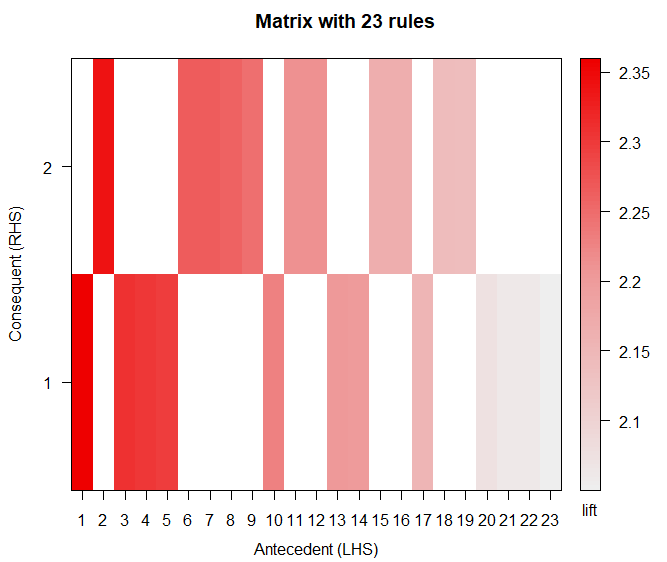
****

Figure 12 – R Code used to generate Project

# Assignment 1

# by Kenneth Lulie, Data 630 - Ami Gates

# Created 9/30/2018

# Worked on 10/1/2018

#Using the "wages.csv" Data obtained from the course resources.

#objective will be to mine for rules to predict wage.

#Will be reusing code as possible from week 2 and week 3 exercises

#preloading libraries

#assumes packages "arules, arulesviz, TSP, data.table" are already installed

library("arules")

library("arulesViz")

#load in file

setwd("D:/UMUC/630/Week 3/Assignment 1")

wages <- read.csv("wages.csv")

head(wages)

summary(wages)

#data looks really good, looking at the mins and maxs i don't see any obvious areas to correct

#only questionable data is a "1" for wage which appears to be under minimum wage but without knowing

#the actual location and employment status impossible to know for certain.

str(wages)

#need to discretive education, experience, wage, age

#need to factorize sex, union, race, occupation, sector, and marital\_status

#534 observations of 11 variables.

#no ID variable to remove.

#Review, no NAs

apply(wages, 2, function (wages) sum(is.na(wages)))

#no Nas returned, very high quality data.

#Review mean of sex

(aggregate(wages[, 6], list(wages$sex), mean))

#review difference by south

(aggregate(wages[, 6], list(wages$south), mean))

(aggregate(wages[, 6], list(wages$union), mean))

boxplot(wages$wage ~ wages$south, data = wages)

boxplot(wages$wage ~ wages$sex, data = wages)

#lets start converting to factors using list above

wages$sex <- as.factor(wages$sex)

wages$union <- as.factor(wages$union)

wages$race <- as.factor(wages$race)

wages$occupation<- as.factor(wages$occupation)

wages$sector <- as.factor(wages$sector)

wages$marital\_status <- as.factor(wages$marital\_status)

str(wages)

#all done no errors

#now to discretize the other variables.

#i think this dataset would be more of general knowledge and understanding than finding hidden connections.

#As such, frequency i think would be better

#for education, i'll just define myself, but all the others seem like frequency will be fine.

#6 buckets will be fine, seems to be standard and should be enough.

wages$experience<-discretize(wages$experience, method="frequency",breaks=6)

wages$wage<-discretize(wages$wage, method="frequency",breaks=6)

wages$age<-discretize(wages$age, method="frequency",breaks=6)

#Education is split into less than high school, some college, 4 year degree, post grad.

wages$education<-discretize(wages$education, method="fixed",breaks=c(0, 11.5, 12, 12.5, 15.5, 16, 16.5, 18.5))

str(wages)

summary(wages$sex)

summary(wages$union)

summary(wages$race)

summary(wages$occupation)

summary(wages$sector)

summary(wages$marital\_status)

#interesting distributions

## Model Run #1

rules<-apriori(wages, parameter= list(supp=0.05, conf=0.7, minlen=2), appearance=list(rhs=c("wage=[1,4.5)",

"wage=[4.5,6)", "wage=[6,7.78)",

"wage=[7.78,10)", "wage=[10,13.1)",

"wage=[13.1,44.5]"), default="lhs"))

rules

#no rules created

## Model Run #2

rules<-apriori(wages, parameter= list(supp=0.02, conf=0.6, minlen=2), appearance=list(rhs=c("wage=[1,4.5)", "wage=[4.5,6)",

"wage=[6,7.78)", "wage=[7.78,10)",

"wage=[10,13.1)", "wage=[13.1,44.5]"),

default="lhs"))

#sneak peak at the results

rules <- sort(rules, by="lift")

inspect(rules[1:10])

rules

#now we have to prune it. Using code from the exercise 3

#pruning the returned rules.

#creating a matrix of the subsets.

rules.sorted <- sort(rules, by="lift")

subset.matrix <- is.subset(rules.sorted, rules.sorted)

#using the matrix diagnoal, set the ones to the left of the line to false.

subset.matrix[lower.tri(subset.matrix, diag=T)] <- F

#stores rules which are redundent

redundant <- colSums(subset.matrix, na.rm=T) >= 1

#remove pruned rules.

rules.pruned <- rules[!redundant]

inspect(rules.pruned)

#only 17 rules were generated after pruning. The rules all have a high life, by range from 11 to 18 obs which is too low.

#also confidence is between .6 and .7 which is low.

#plot model 2 rules on a matrix

plot(rules.pruned, method="matrix", measure=c("lift", "confidence"))

### Model Run #3

#need to change to 3 buckets for wages instead of 6.

#create new df with original data

origwages <- read.csv("wages.csv")

#change wage in dataset from discretization version to original version

wages$wage<-origwages$wage

#change to 3 buckets here to increase rule count

wages$wage<-discretize(wages$wage, method="frequency",breaks=3)

rules<-apriori(wages, parameter= list(supp=0.04, conf=0.7, minlen=2), appearance=list(rhs=c("wage=[1,6)", "wage=[6,10)",

"wage=[10,44.5]"), default="lhs"))

#now we have to prune it. Using code from the exercise 3

#pruning the returned rules.

#creating a matrix of the subsets.

rules.sorted <- sort(rules, by="lift")

subset.matrix <- is.subset(rules.sorted, rules.sorted)

#using the matrix diagnoal, set the ones to the left of the line to false.

subset.matrix[lower.tri(subset.matrix, diag=T)] <- F

#stores rules which are redundent

redundant <- colSums(subset.matrix, na.rm=T) >= 1

#remove pruned rules.

rules.pruned <- rules[!redundant]

rules.pruned <- sort(rules.pruned, by="lift")

inspect(rules.pruned)

# 23 rules were generated after pruning. Lift went down, but support and confidence are much better quality

#plot model 3 rules on a matrix

plot(rules.pruned, method="matrix", measure=c("lift", "confidence"))