M4L4 Homework Assignment

Joshua Conte

October 15, 2017

1 M4L4 Homework Assignment

R studio was configured with the following parameters before beginning the project:

```
# clears the console in RStudio
cat("\014")
```

```
# clears environment
rm(list = ls())

# Load required packages
require(arules)
require(arulesViz)
```

1.1 Load Data.

I used the "Income" database that comes with the arules package.

This data set originates from an example in the book 'The Elements of Statistical Learning' (see Section source). The data set is an extract from this survey. It consists of 8993 instances (obtained from the original data set with 9409 instances, by removing those observations with the annual income missing) with 14 demographic attributes. The data set is a good mixture of categorical and continuous variables with a lot of missing data. This is characteristic of data mining applications. The Income data set contains the data already prepared and coerced to transactions.

Below is my R code:

```
data("Income")
summary(Income)
```

```
## transactions as itemMatrix in sparse format with
##  6876 rows (elements/itemsets/transactions) and
##  50 columns (items) and a density of 0.28
##
## most frequent items:
##      language in home=english education=no college graduate
##                          6277                           4849
##         number in household=1   ethnic classification=white
##                          4757                           4605
##         years in bay area=10+                        (Other)
##                          4446                          71330
##
## element (itemset/transaction) length distribution:
## sizes
##   14
## 6876
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      14      14      14      14      14      14
##
## includes extended item information - examples:
##              labels variables     levels
## 1 income=$0-$40,000    income $0-$40,000
## 2   income=$40,000+    income   $40,000+
## 3          sex=male       sex       male
```

```
##
## includes extended transaction information - examples:
##   transactionID
## 1             2
## 2             3
## 3             4
```

1.1.1 Format

Income data set contains 8993 observations on the following 14 variables:

- **income** an ordered factor with levels [0,10) < [10,15) < [15,20) < [20,25) < [25,30) < [30,40) < [40,50) < [50,75) < 75+
- **sex** a factor with levels male female
- **marital status** a factor with levels married cohabitation divorced widowed single
- **age** an ordered factor with levels 14-17 < 18-24 < 25-34 < 35-44 < 45-54 < 55-64 < 65+
- **education** an ordered factor with levels grade <9 < grades 9-11 < high school graduate < college (1-3 years) < college graduate < graduate study
- **occupation** a factor with levels professional/managerial sales laborer clerical/service homemaker student military retired unemployed
- **years in bay area** an ordered factor with levels <1 < 1-3 < 4-6 < 7-10 < >10
- **dual incomes** a factor with levels not married yes no
- **number in household** an ordered factor with levels 1 < 2 < 3 < 4 < 5 < 6 < 7 < 8 < 9+
- **number of children** an ordered factor with levels 0 < 1 < 2 < 3 < 4 < 5 < 6 < 7 < 8 < 9+
- **householder status** a factor with levels own rent live with parents/family
- **type of home** a factor with levels house condominium apartment mobile Home other
- **ethnic classification** a factor with levels American Indian Asian Black East Indian Hispanic pacific islander white other
- **language in home** a factor with levels English Spanish other

1.2 Analyzing the Data

Look at the first five transactions.

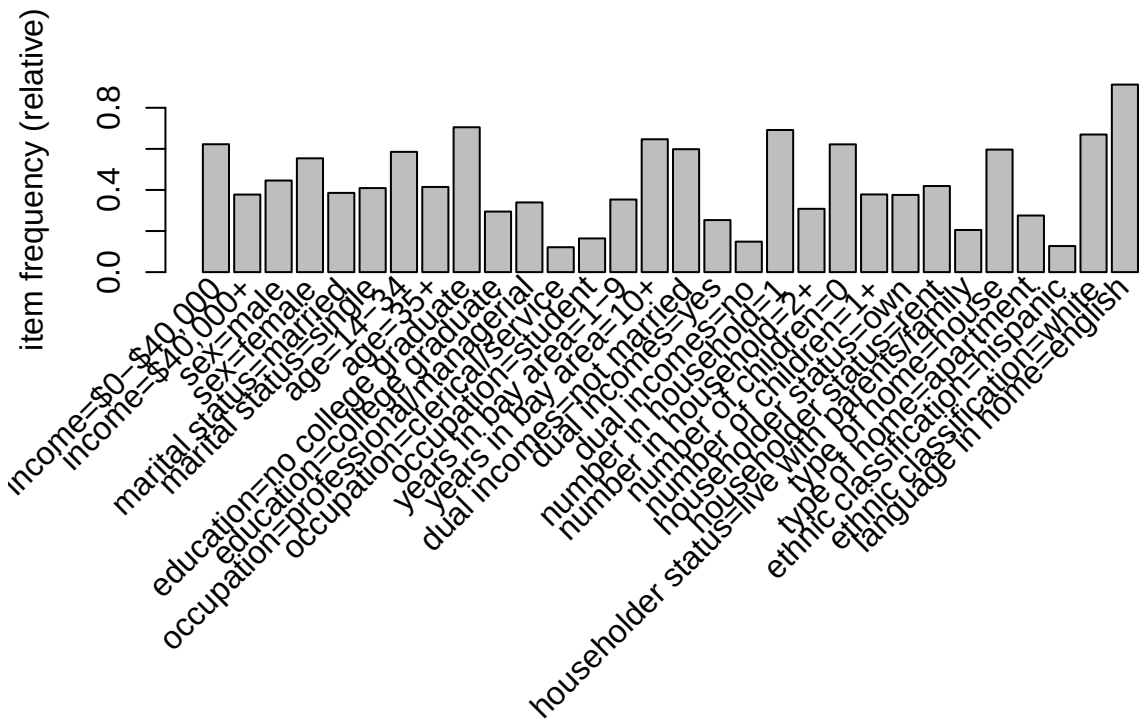
```
# look at the first five transactions
inspect(Income[1:5])
```

```
##     items                                          transactionID
## [1] {income=$40,000+,
##      sex=male,
##      marital status=married,
##      age=35+,
##      education=college graduate,
##      occupation=homemaker,
##      years in bay area=10+,
##      dual incomes=no,
##      number in household=2+,
##      number of children=1+,
##      householder status=own,
##      type of home=house,
##      ethnic classification=white,
##      language in home=english}                             2
## [2] {income=$40,000+,
##      sex=female,
```

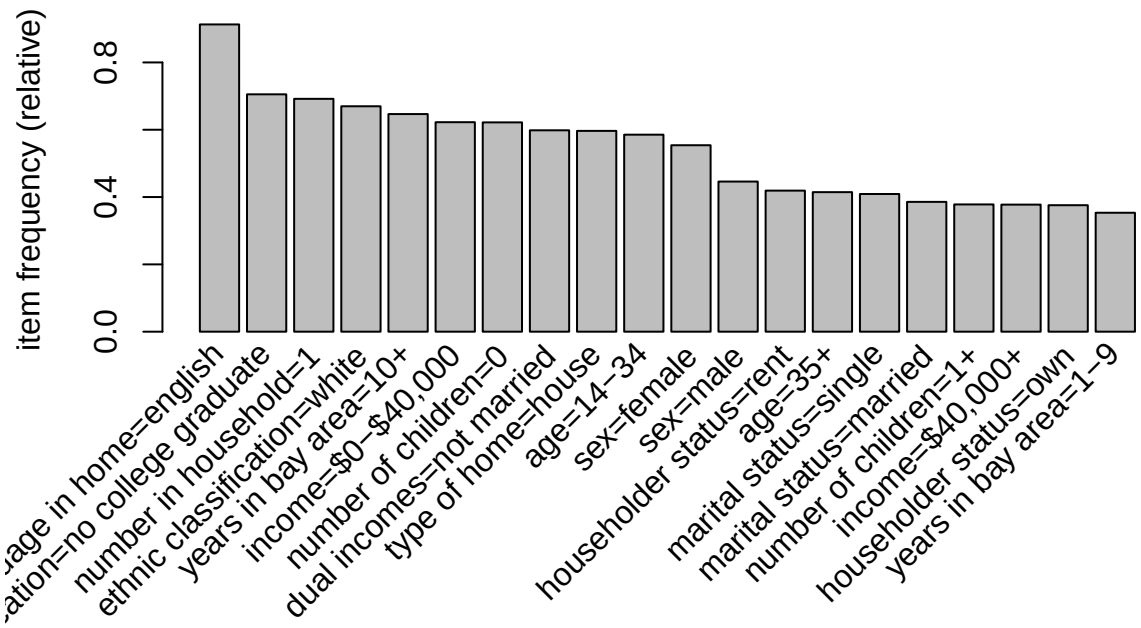
```
##      marital status=married,
##      age=14-34,
##      education=college graduate,
##      occupation=professional/managerial,
##      years in bay area=10+,
##      dual incomes=yes,
##      number in household=1,
##      number of children=1+,
##      householder status=rent,
##      type of home=apartment,
##      ethnic classification=white,
##      language in home=english}                                  3
## [3] {income=$0-$40,000,
##      sex=female,
##      marital status=single,
##      age=14-34,
##      education=no college graduate,
##      occupation=student,
##      years in bay area=10+,
##      dual incomes=not married,
##      number in household=2+,
##      number of children=1+,
##      householder status=live with parents/family,
##      type of home=house,
##      ethnic classification=white,
##      language in home=english}                                  4
## [4] {income=$0-$40,000,
##      sex=female,
##      marital status=single,
##      age=14-34,
##      education=no college graduate,
##      occupation=student,
##      years in bay area=1-9,
##      dual incomes=not married,
##      number in household=2+,
##      number of children=1+,
##      householder status=live with parents/family,
##      type of home=house,
##      ethnic classification=white,
##      language in home=english}                                  5
## [5] {income=$40,000+,
##      sex=male,
##      marital status=married,
##      age=35+,
##      education=no college graduate,
##      occupation=retired,
##      years in bay area=10+,
##      dual incomes=no,
##      number in household=1,
##      number of children=0,
##      householder status=own,
##      type of home=house,
##      ethnic classification=white,
##      language in home=english}                                  6
```

Plot the frequency.

```
# plot the frequency
# if getting the error
# Error in plot.new() : figure margins too large in RStudio
# use dev.off() to Rrsetting your graphics device
# dev.off() will remove any leftover options or settings
#
itemFrequencyPlot(Income, support = 0.1)
```

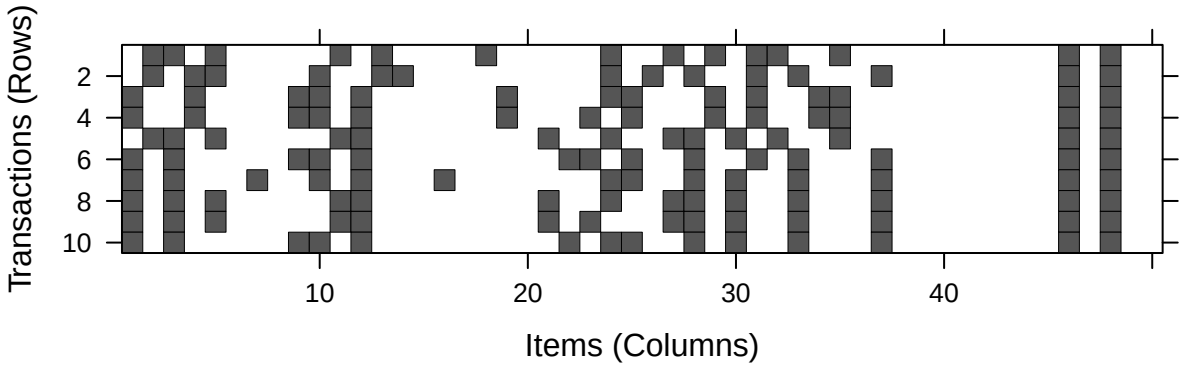



```
itemFrequencyPlot(Income, topN = 20)
```

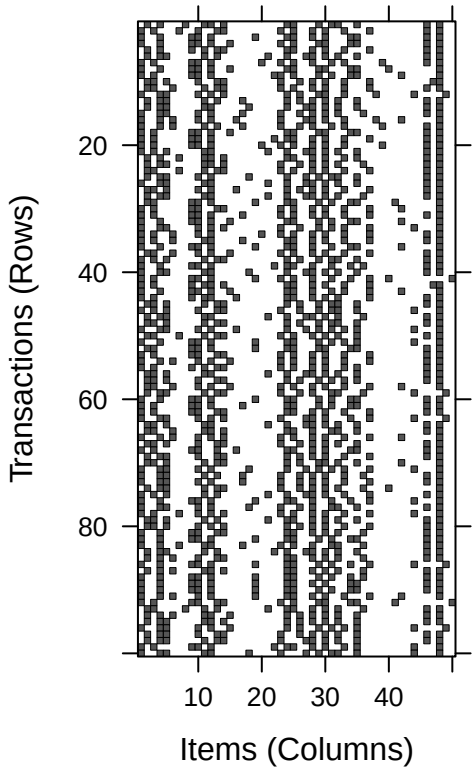


Visualization of some of the transactions.

```
# a visualization of first ten transactions
image(Income[1:10])
```




```
# visualization of a random sample of 100 transactions
image(sample(Income, 100))
```



1.3 Apriori algorithm

I began with the values: support = 0.01, confidence = 0.99, and mionlen = 4. Then I started adjusting the support value so that when the redundancies were removed the final set of rules would be around 50. I eventually set the support value to 0.17.

To find the redundancies, I found using *is.redundant()* is much more efficient and faster than using the example in the module. Before removing the redundancies, I sorted the rules by lift.

Below is my R code:

```
# set better support and confidence levels to get around 50 rules
income <-
  apriori(Income, parameter = list(
    support = 0.17,
    confidence = 0.99,
    minlen = 4
  ))
```

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##        0.99    0.1    1 none FALSE            TRUE       5    0.17      4
##  maxlen target   ext
##      10  rules FALSE
##
```

```
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 1168
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[50 item(s), 6876 transaction(s)] done [0.00s].
## sorting and recoding items ... [26 item(s)] done [0.00s].
## creating transaction tree ... done [0.02s].
## checking subsets of size 1 2 3 4 5 6 done [0.02s].
## writing ... [90 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

```
summary(income)
```

```
## set of 90 rules
##
## rule length distribution (lhs + rhs):sizes
##  4  5  6
## 51 36  3
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.000   4.000   4.000   4.467   5.000   6.000
##
## summary of quality measures:
##     support         confidence          lift           count
##  Min.   :0.1706   Min.   :0.9901   Min.   :1.085   Min.   :1173
##  1st Qu.:0.1804   1st Qu.:0.9916   1st Qu.:1.086   1st Qu.:1240
##  Median :0.1943   Median :1.0000   Median :1.671   Median :1336
##  Mean   :0.2054   Mean   :0.9972   Mean   :1.489   Mean   :1412
##  3rd Qu.:0.2166   3rd Qu.:1.0000   3rd Qu.:1.671   3rd Qu.:1490
##  Max.   :0.3240   Max.   :1.0000   Max.   :1.671   Max.   :2228
##
## mining info:
##    data ntransactions support confidence
##  Income          6876    0.17       0.99
```

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

# A more efficient way to find redundant rules
redundant<- is.redundant(rules.sorted)
summary(redundant)
```

```
##    Mode   FALSE    TRUE
## logical      53      37
```

```
# Remove redundant rules
rules.pruned <- rules.sorted[!redundant]

# Verify rules
rules.pruned
```

```
## set of 53 rules
```

```
# Print rules
inspect(rules.pruned)
```

```
##      lhs                                                rhs                          support confidenc
## [1]  {marital status=single,
##       age=14-34,
##       householder status=live with parents/family} => {dual incomes=not married} 0.1858639  1.000000
## [2]  {marital status=single,
##       education=no college graduate,
##       householder status=live with parents/family} => {dual incomes=not married} 0.1801920  1.000000
## [3]  {sex=male,
##       marital status=single,
##       age=14-34}                                   => {dual incomes=not married} 0.1828098  1.000000
## [4]  {sex=male,
##       marital status=single,
##       language in home=english}                    => {dual incomes=not married} 0.1847004  1.000000
## [5]  {sex=female,
##       marital status=single,
##       age=14-34}                                   => {dual incomes=not married} 0.1826643  1.000000
## [6]  {sex=female,
##       marital status=single,
##       language in home=english}                    => {dual incomes=not married} 0.1800465  1.000000
## [7]  {marital status=single,
##       age=14-34,
##       type of home=house}                          => {dual incomes=not married} 0.2111693  1.000000
## [8]  {marital status=single,
##       age=14-34,
##       number of children=0}                        => {dual incomes=not married} 0.2079697  1.000000
## [9]  {income=$0-$40,000,
##       marital status=single,
##       age=14-34}                                   => {dual incomes=not married} 0.3000291  1.000000
## [10] {marital status=single,
##       age=14-34,
##       years in bay area=10+}                       => {dual incomes=not married} 0.2168412  1.000000
## [11] {marital status=single,
##       age=14-34,
##       ethnic classification=white}                 => {dual incomes=not married} 0.2187318  1.000000
## [12] {marital status=single,
##       age=14-34,
##       number in household=1}                       => {dual incomes=not married} 0.2126236  1.000000
## [13] {marital status=single,
##       age=14-34,
##       education=no college graduate}               => {dual incomes=not married} 0.2994474  1.000000
## [14] {marital status=single,
##       age=14-34,
##       language in home=english}                    => {dual incomes=not married} 0.3240256  1.000000
## [15] {income=$0-$40,000,
##       marital status=single,
##       type of home=house}                          => {dual incomes=not married} 0.1740838  1.000000
## [16] {marital status=single,
##       education=no college graduate,
##       type of home=house}                          => {dual incomes=not married} 0.1983711  1.000000
## [17] {marital status=single,
##       type of home=house,
```

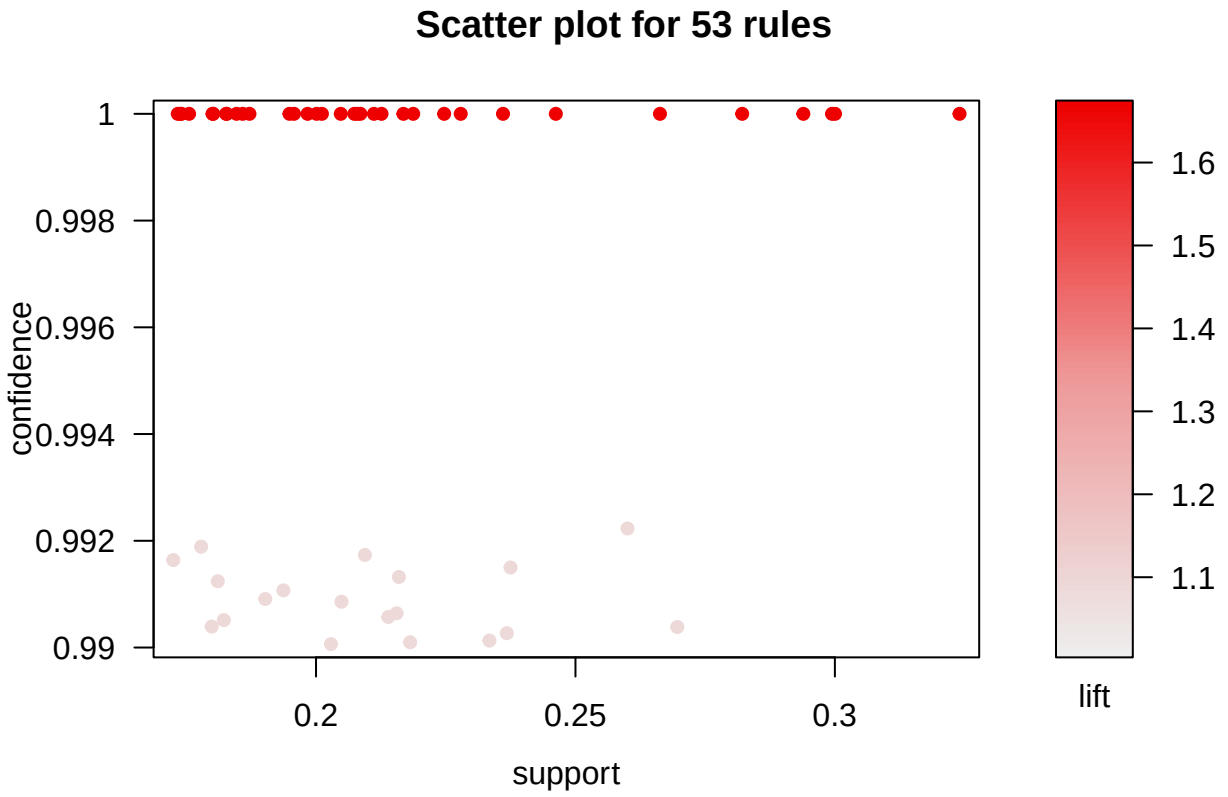
```
##       language in home=english}                    => {dual incomes=not married} 0.2001163  1.0000000
## [18] {income=$0-$40,000,
##       marital status=single,
##       number of children=0}                        => {dual incomes=not married} 0.1948807  1.0000000
## [19] {marital status=single,
##       number of children=0,
##       ethnic classification=white}                 => {dual incomes=not married} 0.1733566  1.0000000
## [20] {marital status=single,
##       number in household=1,
##       number of children=0}                        => {dual incomes=not married} 0.2047702  1.0000000
## [21] {marital status=single,
##       number of children=0,
##       language in home=english}                    => {dual incomes=not married} 0.2278941  1.0000000
## [22] {income=$0-$40,000,
##       marital status=single,
##       years in bay area=10+}                       => {dual incomes=not married} 0.1957533  1.0000000
## [23] {income=$0-$40,000,
##       marital status=single,
##       ethnic classification=white}                 => {dual incomes=not married} 0.2011344  1.0000000
## [24] {income=$0-$40,000,
##       marital status=single,
##       number in household=1}                       => {dual incomes=not married} 0.2085515  1.0000000
## [25] {income=$0-$40,000,
##       marital status=single,
##       education=no college graduate}               => {dual incomes=not married} 0.2662885  1.0000000
## [26] {income=$0-$40,000,
##       marital status=single,
##       language in home=english}                    => {dual incomes=not married} 0.2939209  1.0000000
## [27] {marital status=single,
##       education=no college graduate,
##       years in bay area=10+}                       => {dual incomes=not married} 0.2073880  1.0000000
## [28] {marital status=single,
##       years in bay area=10+,
##       language in home=english}                    => {dual incomes=not married} 0.2246946  1.0000000
## [29] {marital status=single,
##       number in household=1,
##       ethnic classification=white}                 => {dual incomes=not married} 0.1737929  1.0000000
## [30] {marital status=single,
##       education=no college graduate,
##       ethnic classification=white}                 => {dual incomes=not married} 0.1871728  1.0000000
## [31] {marital status=single,
##       ethnic classification=white,
##       language in home=english}                    => {dual incomes=not married} 0.2462187  1.0000000
## [32] {marital status=single,
##       education=no college graduate,
##       number in household=1}                       => {dual incomes=not married} 0.1755381  1.0000000
## [33] {marital status=single,
##       number in household=1,
##       language in home=english}                    => {dual incomes=not married} 0.2360384  1.0000000
## [34] {marital status=single,
##       education=no college graduate,
##       language in home=english}                    => {dual incomes=not married} 0.2821408  1.0000000
## [35] {number in household=1,
##       type of home=house,
```

```
##       ethnic classification=white}                => {language in home=english} 0.2600349  0.992230
## [36] {sex=female,
##       years in bay area=10+,
##       number in household=1,
##       ethnic classification=white}                => {language in home=english} 0.1778650  0.991889
## [37] {marital status=married,
##       type of home=house,
##       ethnic classification=white}                => {language in home=english} 0.2094241  0.991735
## [38] {sex=female,
##       householder status=own,
##       ethnic classification=white}                => {language in home=english} 0.1724840  0.991638
## [39] {householder status=own,
##       type of home=house,
##       ethnic classification=white}                => {language in home=english} 0.2374927  0.991499
## [40] {years in bay area=10+,
##       householder status=own,
##       ethnic classification=white}                => {language in home=english} 0.2159686  0.991321
## [41] {occupation=professional/managerial,
##       number of children=0,
##       ethnic classification=white}                => {language in home=english} 0.1810646  0.991242
## [42] {marital status=married,
##       years in bay area=10+,
##       ethnic classification=white}                => {language in home=english} 0.1937173  0.991071
## [43] {income=$40,000+,
##       householder status=own,
##       ethnic classification=white}                => {language in home=english} 0.1902269  0.990909
## [44] {occupation=professional/managerial,
##       number in household=1,
##       ethnic classification=white}                => {language in home=english} 0.2049156  0.990857
## [45] {number in household=1,
##       householder status=own,
##       ethnic classification=white}                => {language in home=english} 0.2155323  0.990641
## [46] {sex=female,
##       number in household=1,
##       number of children=0,
##       ethnic classification=white}                => {language in home=english} 0.2139325  0.990572
## [47] {income=$40,000+,
##       marital status=married,
##       ethnic classification=white}                => {language in home=english} 0.1822280  0.990513
## [48] {income=$40,000+,
##       age=35+,
##       ethnic classification=white}                => {language in home=english} 0.1799011  0.990392
## [49] {sex=female,
##       number in household=1,
##       ethnic classification=white}                => {language in home=english} 0.2696335  0.990384
## [50] {age=35+,
##       years in bay area=10+,
##       ethnic classification=white}                => {language in home=english} 0.2367656  0.990267
## [51] {number of children=0,
##       type of home=house,
##       ethnic classification=white}                => {language in home=english} 0.2334206  0.990129
## [52] {age=35+,
##       type of home=house,
##       ethnic classification=white}                => {language in home=english} 0.2181501  0.990099
```

```
## [53] {marital status=married,
##       householder status=own,
##       ethnic classification=white}                => {language in home=english} 0.2028796  0.990063
```

Visualizing Association Rules:

```
plot(rules.pruned)
```

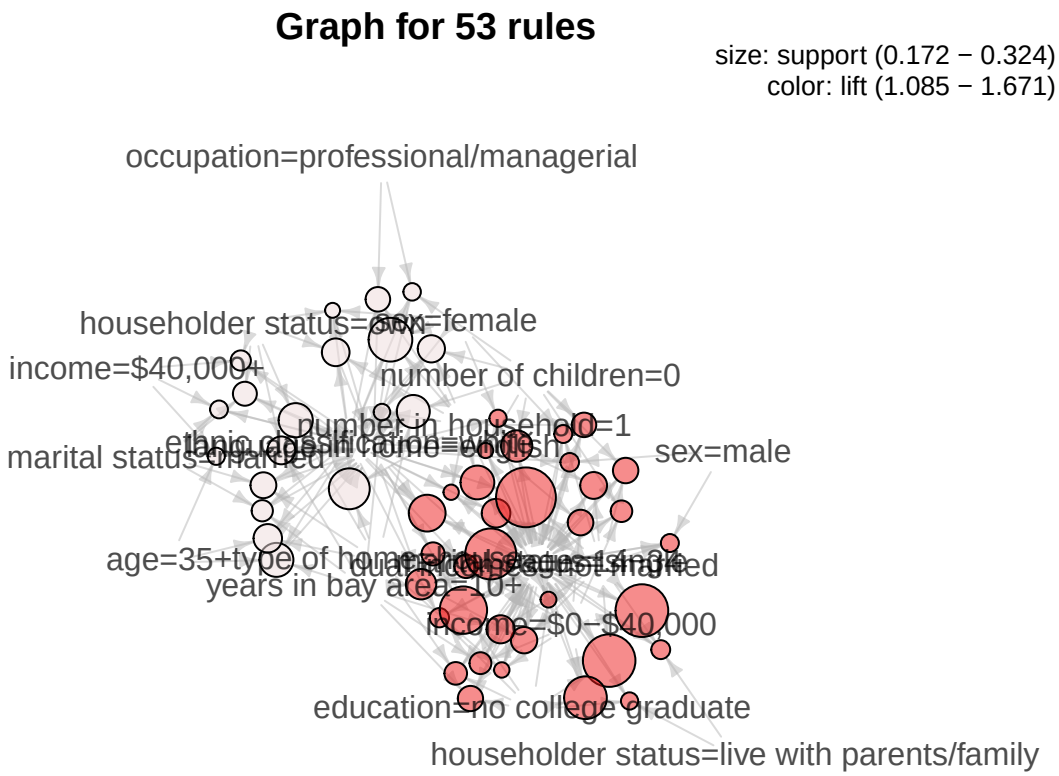



```
plot(rules.pruned, method="graph", control=list(type="items"))
```

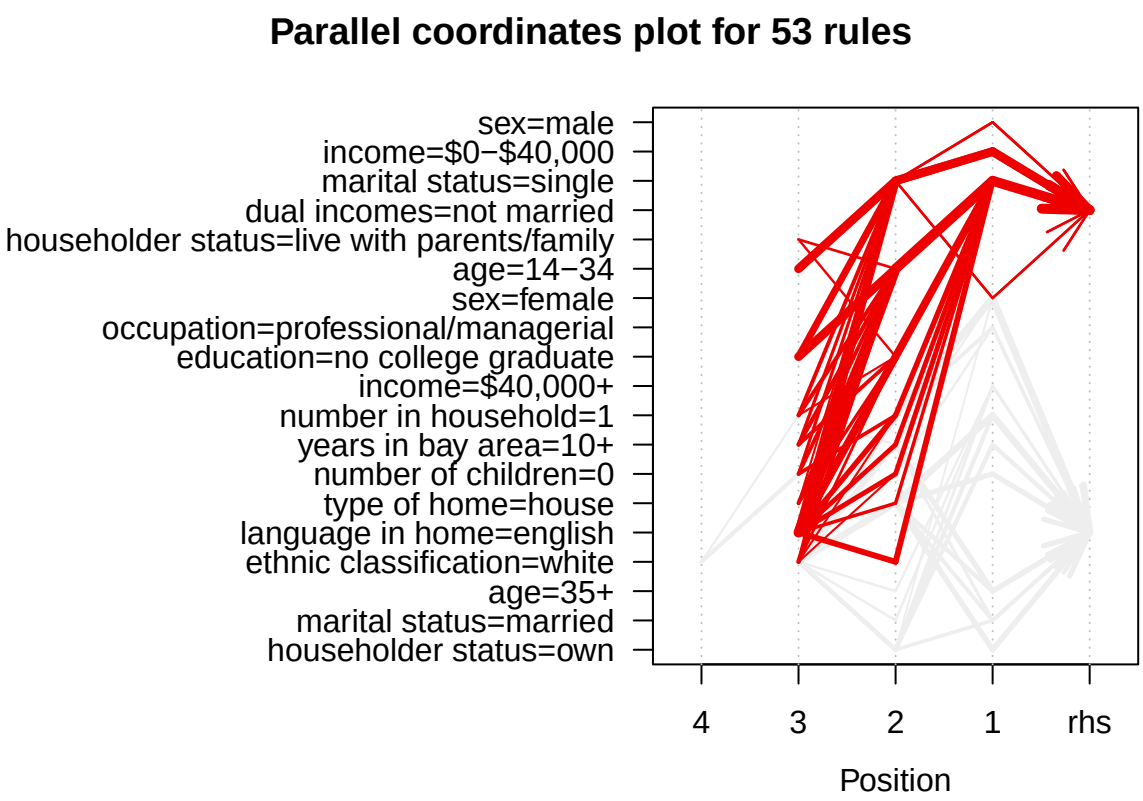
```
## Warning: Unknown control parameters: type
```

```
## Available control parameters (with default values):
## main  =  Graph for 53 rules
## nodeColors    =  c("#66CC6680", "#9999CC80")
## nodeCol   =  c("#EE0000FF", "#EE0303FF", "#EE0606FF", "#EE0909FF", "#EE0C0CFF", "#EE0F0FFF", "#EE121
## edgeCol   =  c("#474747FF", "#494949FF", "#4B4B4BFF", "#4D4D4DFF", "#4F4F4FFF", "#515151FF", "#535353
## alpha     =  0.5
## cex   =  1
## itemLabels    =  TRUE
## labelCol  =  #000000B3
## measureLabels     =  FALSE
## precision     =  3
## layout    =  NULL
## layoutParams  =  list()
## arrowSize     =  0.5
## engine    =  igraph
## plot  =  TRUE
```

```
## plot_options  =  list()
## max   =  100
## verbose   =  FALSE
```




```
plot(rules.pruned, method="paracoord", control=list(reorder=TRUE))
```



Calculating lift and conviction:

```
# Lift and conviction for the first five rules. Note, the number convention is worng.
cbind(
  as(rules.pruned[1:5], "data.frame"),
  conviction = interestMeasure(rules.pruned[1:5], "conviction", rules.pruned[1:5])
)
```

```
##
## 1                     {marital status=single,age=14-34,householder status=live with parents/family}
## 2  {marital status=single,education=no college graduate,householder status=live with parents/family}
## 15                                                  {sex=male,marital status=single,age=14-34}
## 16                                  {sex=male,marital status=single,language in home=english}
## 17                                                {sex=female,marital status=single,age=14-34}
##      support confidence     lift count conviction
## 1  0.1858639          1 1.671366  1278         NA
## 2  0.1801920          1 1.671366  1239         NA
## 15 0.1828098          1 1.671366  1257         NA
## 16 0.1847004          1 1.671366  1270         NA
## 17 0.1826643          1 1.671366  1256         NA
```

```
# Lift and conviction for the last five rules. Note, the number convention is worng.
cbind(
  as(rules.pruned[49:53], "data.frame"),
  conviction = interestMeasure(rules.pruned[49:53], "conviction", rules.pruned[49:53])
)
```

```
##
```

```
## 49              {sex=female,number in household=1,ethnic classification=white} => {language in home=
## 48                 {age=35+,years in bay area=10+,ethnic classification=white} => {language in home=
## 50       {number of children=0,type of home=house,ethnic classification=white} => {language in home=
## 47                    {age=35+,type of home=house,ethnic classification=white} => {language in home=
## 6  {marital status=married,householder status=own,ethnic classification=white} => {language in home=
##      support confidence     lift count conviction
## 49 0.2696335  0.9903846 1.084895  1854   9.059919
## 48 0.2367656  0.9902676 1.084767  1628   8.951025
## 50 0.2334206  0.9901295 1.084615  1605   8.825798
## 47 0.2181501  0.9900990 1.084582  1500   8.798575
## 6  0.2028796  0.9900639 1.084543  1395   8.767462
```

1.4 Questions

1. **Which rules make sense to you? Highlight the five best and five worst of your rule set.**
   - Reviewing the rules, they all make sense to me. I think that they are all ok, so for this question I am going to choose 1 - 5 as the best and 49 - 53 as the worst.
   - **The Best:**
     1. {marital status=single, age=14-34, householder status=live with parents/family} => {dual incomes=not married}
     2. {marital status=single, education=no college graduate, householder status=live with parents/family} => {dual incomes=not married}
     3. {sex=male, marital status=single, age=14-34} => {dual incomes=not married}
     4. {sex=male, marital status=single, language in home=english} => {dual incomes=not married}
     5. {sex=female, marital status=single, age=14-34} => {dual incomes=not married}
   - **The Worst**
     49. {sex=female, number in household=1, ethnic classification=white} => {language in home=english}
     50. {age=35+, years in bay area=10+, ethnic classification=white} > {language in home=english}
     51. {number of children=0, type of home=house, ethnic classification=white} => {language in home=english}
     52. {age=35+, type of home=house, ethnic classification=white} => {language in home=english}
     53. {marital status=married, householder status=own, ethnic classification=white} => {language in home=english}
2. **How did you choose the level of support and confidence?**
   - I began with the values: support = 0.01, confidence = 0.99, and mionlen = 4. Then I started adjustingthe support value so that when the redundancies were removed the final set of rules would be around 50. . I settled on 0.17 for support and 0.99 for confidence.
3. **What is the lift and conviction of your best and worst rules?**
   - **The Best:**
     1. lift 1.671366 / conviction NA
     2. lift 1.671366 / conviction NA
     3. lift 1.671366 / conviction NA
     4. lift 1.671366 / conviction NA
     5. lift 1.671366 / conviction NA
   - **The Worst**
     49. lift 0.2696335 / conviction 9.059919
     50. lift 0.2367656 / conviction 8.951025
     51. lift 0.2334206 / conviction 8.825798
     52. lift 0.2181501 / conviction 8.798575
     53. lift 0.2028796 / conviction 8.767462
4. **Visualize your 50 association rules. Where do the best and worst end up in your plot?**
   - The first five are high confidence with low support, so they are all in the upper left corner of the

scatter plot. The last five are higher support with lower confidence, which are the ones on the lower middle/right of the scatter plot.

5. **Does the model make sense?**
   - Yes the model makes sense, the first five rules are all single with a single income, that makes sense, usually a single person does not have duel incomes. The last five rules also make sense, they all live in California and are white and the household language is English, I think that makes sense. After reviewing all of the rules, nothing jumped out at me as being strange or misclassified.