Project 6

Association Rules

Brooke O'Donnell

December 7th, 2020

```
data(package="arules")

data("Adult")
```

1. Please use function or functions you see fit to answer the following questions:

```
set.seed(123)
summary(Adult)
```

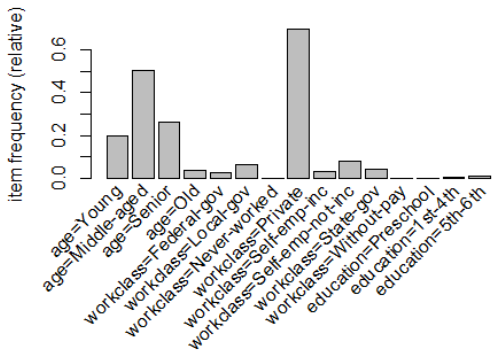
a. what is the density of the sparse matrix?
   - density=.10899, about 11% of cells in the sparse matrix have "1's"

b. What is the most frequent item (or feature)?
   - workclass=private

c. How many transactions (respondents) contain more that 11 (not including 11) features (items)?
   - 30162 transactions contain more than 11 feature items

2. Plot the top 15 most frequent items (features) in the data.

```
itemFrequency(Adult[1:15])

itemFrequencyPlot(Adult[,1:15])
```



a. What are the top 2 most frequent items?
   - The top 2 most frequent items are workclass=Private and age=Middle-aged

3. Use the "apriori" function to create 2-item association rules. Please set the "support" and "confidence" to levels at which no more than 50 rules are generated.

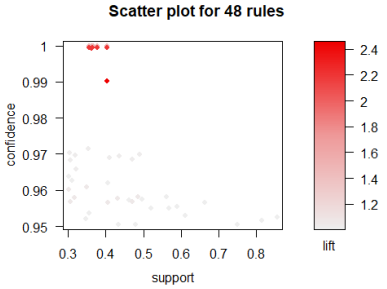
```
adult.rules.2<-apriori(Adult, parameter=list(support=0.30, confidence=0.95, minlen=2, maxlen=3))

adult.rules.2

inspect(adult.rules.2)

summary(adult.rules.2)

## set of 48 rules

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.000   2.000   3.000   2.729   3.000   3.000
##
## summary of quality measures:
##     support         confidence        coverage          lift
##  Min.   :0.3010   Min.   :0.9504   Min.   :0.3123   Min.   :0.997
##  1st Qu.:0.3544   1st Qu.:0.9566   1st Qu.:0.3631   1st Qu.:1.004
##  Median :0.4034   Median :0.9623   Median :0.4037   Median :1.017
##  Mean   :0.4362   Mean   :0.9702   Mean   :0.4508   Mean   :1.251
##  3rd Qu.:0.4813   3rd Qu.:0.9924   3rd Qu.:0.5054   3rd Qu.:1.496
##  Max.   :0.8548   Max.   :0.9999   Max.   :0.8974   Max.   :2.453

count
##  Min.   :14700
##  1st Qu.:17311
##  Median :19704
##  Mean   :21305
##  3rd Qu.:23507
##  Max.   :41752
```

a. How many association rule are generated?
   - 48 association rules are generated

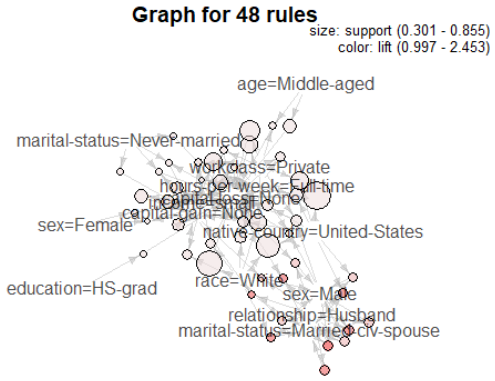
b. What are the mean and the range for the "lift" for your rules?
   - lift mean=1.251
   - lift range=.997-2.5

4. Plot the association rules generated from the last step using the "graph" method.

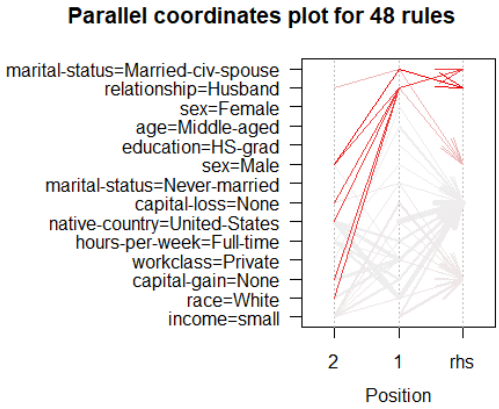
```
plot(adult.rules.2)
```




```
plot(adult.rules.2, method="graph",
control=list(type="items"))
```




```
plot(adult.rules.2, method="paracoord", control=list(reorder=TRUE))
```



a. Do you see any "item" or "items" that function as the central attractions of many association rules?

- Yes

b. What items are these centers?

- It looks like workclass=Private, hours=per-week, capital-loss=None, and capital-gain=None are items centered as central attractions of many association rules.

5. Please select the five rules with the highest lifts.

```
inspect(adult.rules.2)

inspect(sort(adult.rules.2, by="lift")[1:5])

##     lhs                                rhs                                support confidence  coverage     lift count
## [1] {marital-status=Married-civ-spouse,
##      sex=Male}                      => {relationship=Husband}           0.4034028  0.9901503
0.4074157 2.452877 19703
## [2] {relationship=Husband,
##      race=White}                    => {marital-status=Married-civ-spouse} 0.3654232  0.9994400
```

```
0.3656279 2.181270 17848
## [3] {relationship=Husband}           => {marital-status=Married-civ-spouse} 0.4034233
0.9993914 0.4036690 2.181164 19704
## [4] {relationship=Husband,
##      sex=Male}                       => {marital-status=Married-civ-spouse} 0.4034028  0.9993913
0.4036485 2.181164 19703
## [5] {relationship=Husband,
##      capital-gain=None}              => {marital-status=Married-civ-spouse} 0.3550018
0.9993660 0.3552271 2.181109 17339
```

a. Explain what the rule(s) means if it (or they) involves "capital-gain=none."

- When a respondent answers [marital-status=Never-married] they will likely answer [capital-gain=None]. Confidence tells you that these will be paired almost 96% of the time. The highest lift produced is 2.18 this is when [relationship=Husband] and [marital-status=Married] are together which is obvious those answers would be associated.

6. Please create a subset of rules that only involve women.

```
capital.gain.rules<-subset(adult.rules.2, items %in% "sex=Female")
inspect(capital.gain.rules)

##     lhs                          rhs              support   confidence
## [1] {sex=Female}                 => {capital-loss=None} 0.3201753 0.9657856
## [2] {sex=Female,capital-gain=None} => {capital-loss=None} 0.3009705 0.9636817
##     coverage  lift     count
## [1] 0.3315180 1.013121 15638
## [2] 0.3123132 1.010914 14700

capital.gain.rules

## set of 2 rules
```

a. How many rules are there?

- 2 rules

b. Please choose one rule and interpret the meanings of its support, confidence, and lift?

- When considering all transactions sex=Female and capital-loss=None will be paired 32% of the time. I can be 97% confident that when a respondent answers sex=Female they will answer capital-loss=None. The lift produced a 1.0 which prove thi to be a high associationl.

c. What social or sociological implications can you infer from the all the rules?

- Females might not be taught about how to properly invest their money because there seems to be such a high assoiation between [sex=Female] and [capital-loss=None].