UTD 2016 Spring

CS6301.5U1 Advanced Computational Methods for Data Science

Assignment 1 – Association Rule Mining

Name: Yu Zhang, Mingyue Sun

NetID: yxz141631, mxs151730

1. Data Description and Analysis Methods

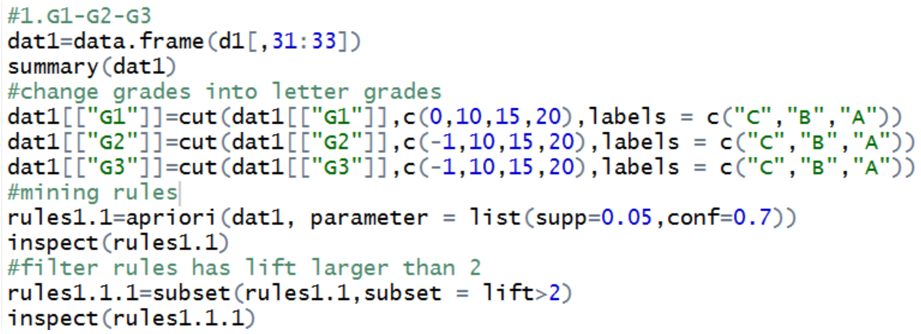
Two datasets are provided which provides the information of student family background, living habits and academic behaviors in Math and Portuguese courses. Both of those two datasets contains 33 columns. The data size of Math course is 395. For Portuguese course it’s 649.

In this section, we will use Association Rule Mining algorithm to analysis potential valuable information that may hides behind different attributes. To start with, we divided those potential relations(rules) into three parts: Academic (2.1-2.2), Personal Life and Connections between personal life (2.3-2.4) and academic behaviors (2.5).

2. Analysis

We will use the “apriori” function in “arules” package of R language to do the ARM. This algorithm accepts categorical data only. Therefore, after grouping attributes we are interested in, we may transfer numeric data into factor or categorize them into different levels. We will use the same attribute grouping to analyze two datasets.

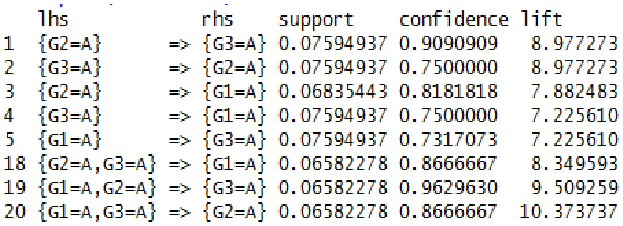
2.1 Associations between Grades in 3 Exams



```
#1.G1-G2-G3
dat1=data.frame(d1[,31:33])
summary(dat1)
#change grades into letter grades
dat1[["G1"]]=cut(dat1[["G1"]],c(0,10,15,20),labels = c("C","B","A"))
dat1[["G2"]]=cut(dat1[["G2"]],c(-1,10,15,20),labels = c("C","B","A"))
dat1[["G3"]]=cut(dat1[["G3"]],c(-1,10,15,20),labels = c("C","B","A"))
#mining rules
rules1.1=apriori(dat1, parameter = list(supp=0.05,conf=0.7))
inspect(rules1.1)
#filter rules has lift larger than 2
rules1.1.1=subset(rules1.1,subset = lift>2)
inspect(rules1.1.1)
```

Data Pre-Processing: We categorize the grades into 3 levels: C: 0-10, B: 11-15, A:16-20. And we chose the rules that has lift larger than 2. The following are results:

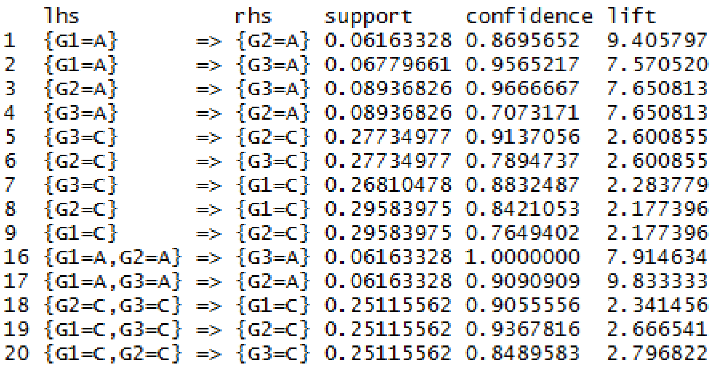
- Math:



```
   lhs              rhs    support    confidence lift
1  {G2=A}        => {G3=A} 0.07594937 0.9090909   8.977273
2  {G3=A}        => {G2=A} 0.07594937 0.7500000   8.977273
3  {G2=A}        => {G1=A} 0.06835443 0.8181818   7.882483
4  {G3=A}        => {G1=A} 0.07594937 0.7500000   7.225610
5  {G1=A}        => {G3=A} 0.07594937 0.7317073   7.225610
18 {G2=A,G3=A} => {G1=A} 0.06582278 0.8666667   8.349593
19 {G1=A,G2=A} => {G3=A} 0.06582278 0.9629630   9.509259
20 {G1=A,G3=A} => {G2=A} 0.06582278 0.8666667  10.373737
```

The lifts found in rules between grades from 3 exams are high (all of them are higher than 7). This indicates than the rules found are relatively strong. When looking into the details, we find that all the rules are associated with grades value of A. The most valuable rules may be in line 18, 19 and 20. Because they have highest lift and most of them contains some of the rules above. From this result, we can clearly find out if a student gets high score in one of the exam, it’s likely that he will get high score again. Also if a student gets high score in two exams, the remaining exam score is probably high too.

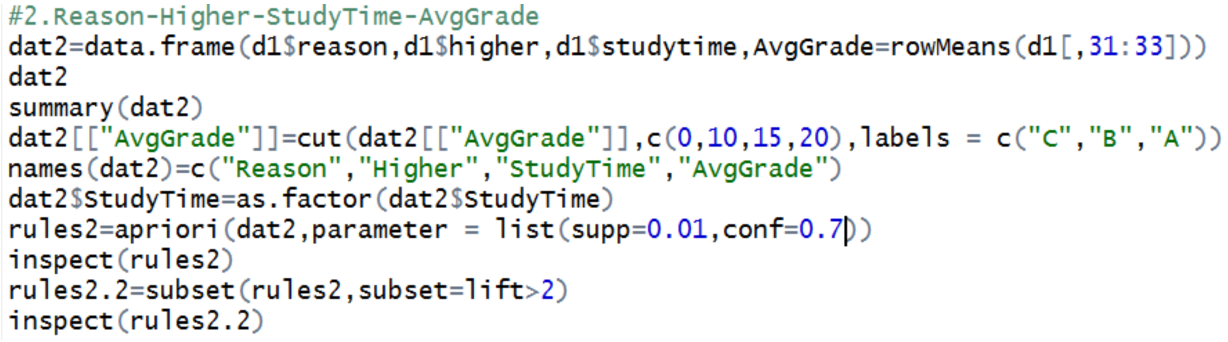
- Portuguese:



```
   lhs             rhs     support    confidence lift
1  {G1=A}       => {G2=A} 0.06163328 0.8695652  9.405797
2  {G1=A}       => {G3=A} 0.06779661 0.9565217  7.570520
3  {G2=A}       => {G3=A} 0.08936826 0.9666667  7.650813
4  {G3=A}       => {G2=A} 0.08936826 0.7073171  7.650813
5  {G3=C}       => {G2=C} 0.27734977 0.9137056  2.600855
6  {G2=C}       => {G3=C} 0.27734977 0.7894737  2.600855
7  {G3=C}       => {G1=C} 0.26810478 0.8832487  2.283779
8  {G2=C}       => {G1=C} 0.29583975 0.8421053  2.177396
9  {G1=C}       => {G2=C} 0.29583975 0.7649402  2.177396
16 {G1=A,G2=A} => {G3=A} 0.06163328 1.0000000  7.914634
17 {G1=A,G3=A} => {G2=A} 0.06163328 0.9090909  9.833333
18 {G2=C,G3=C} => {G1=C} 0.25115562 0.9055556  2.341456
19 {G1=C,G3=C} => {G2=C} 0.25115562 0.9367816  2.666541
20 {G1=C,G2=C} => {G3=C} 0.25115562 0.8489583  2.796822
```

The result of Portuguese course dataset is also good. Though the lifts may not as high as above, fortunately it shows rules in values other than A. Similar with result above, student getting high score is likely to get high score again. But line 18-20 also tell us that students that get low score, it's also likely that they will get another low score.

2.2 Reason to Join Class-Willing to Take Higher Course-Study Time-Average Grade

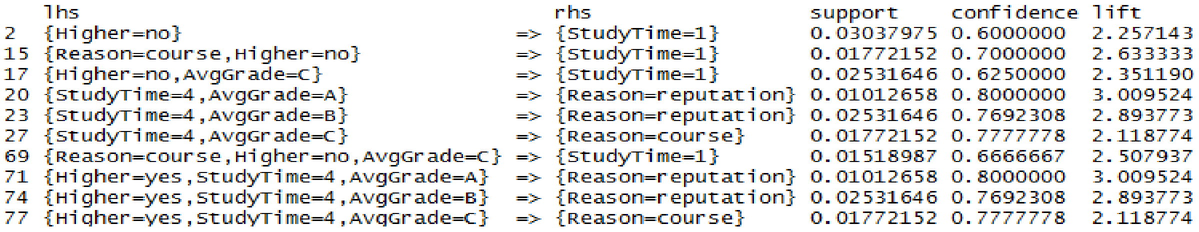


```
#2.Reason-Higher-StudyTime-AvgGrade
dat2=data.frame(d1$reason,d1$higher,d1$studytime,AvgGrade=rowMeans(d1[,31:33]))
dat2
summary(dat2)
dat2[["AvgGrade"]]=cut(dat2[["AvgGrade"]],c(0,10,15,20),labels = c("C","B","A"))
names(dat2)=c("Reason","Higher","StudyTime","AvgGrade")
dat2$StudyTime=as.factor(dat2$StudyTime)
rules2=apriori(dat2,parameter = list(supp=0.01,conf=0.7))
inspect(rules2)
rules2.2=subset(rules2,subset=lift>2)
inspect(rules2.2)
```

Data Pre-Processing: 1. take average grade as new attribute and turn them into letter grades. 2. Factoring the study time into categories. 3. Lowering the support threshold into 0.01 because 3 attributes are multi valued (more than 3), high sup value may prevent us seeing the good results.

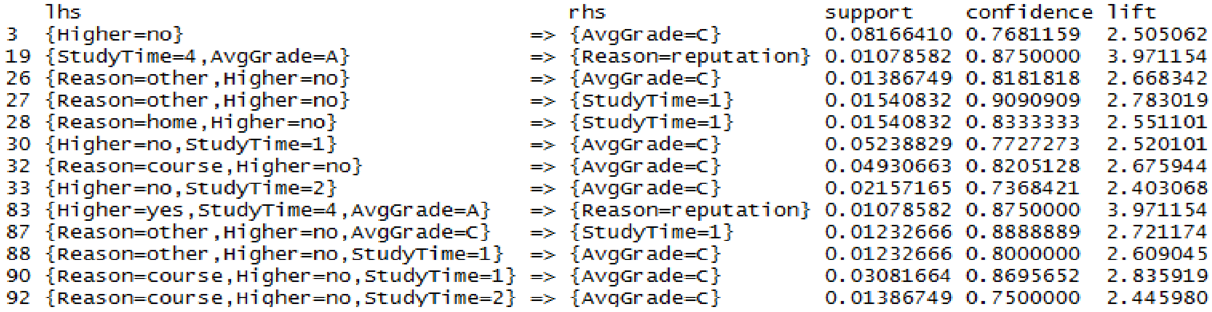
Results:

Math:



```
   lhs                                  rhs                  support    confidence lift
2  {Higher=no}                       => {StudyTime=1}        0.03037975 0.6000000  2.257143
15 {Reason=course,Higher=no}         => {StudyTime=1}        0.01772152 0.7000000  2.633333
17 {Higher=no,AvgGrade=C}            => {StudyTime=1}        0.02531646 0.6250000  2.351190
20 {StudyTime=4,AvgGrade=A}          => {Reason=reputation} 0.01012658 0.8000000  3.009524
23 {StudyTime=4,AvgGrade=B}          => {Reason=reputation} 0.02531646 0.7692308  2.893773
27 {StudyTime=4,AvgGrade=C}          => {Reason=course}     0.01772152 0.7777778  2.118774
69 {Reason=course,Higher=no,AvgGrade=C} => {StudyTime=1}     0.01518987 0.6666667  2.507937
71 {Higher=yes,StudyTime=4,AvgGrade=A}  => {Reason=reputation} 0.01012658 0.8000000  3.009524
74 {Higher=yes,StudyTime=4,AvgGrade=B}  => {Reason=reputation} 0.02531646 0.7692308  2.893773
77 {Higher=yes,StudyTime=4,AvgGrade=C}  => {Reason=course}     0.01772152 0.7777778  2.118774
```

Portuguese:



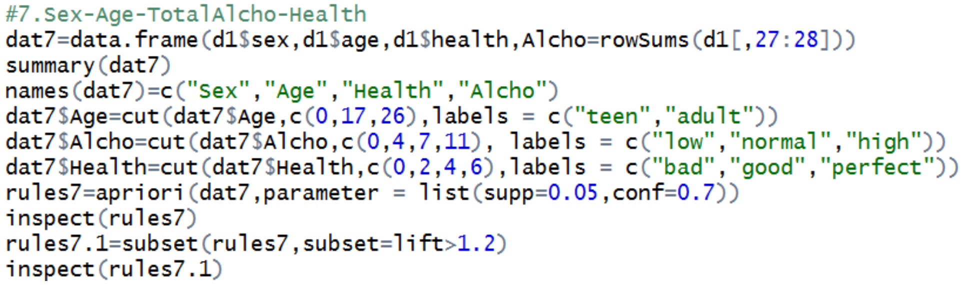
```
    lhs                                          rhs                   support    confidence lift
3   {Higher=no}                               => {AvgGrade=C}          0.08166410 0.7681159  2.505062
19  {StudyTime=4,AvgGrade=A}                  => {Reason=reputation}   0.01078582 0.8750000  3.971154
26  {Reason=other,Higher=no}                  => {AvgGrade=C}          0.01386749 0.8181818  2.668342
27  {Reason=other,Higher=no}                  => {StudyTime=1}         0.01540832 0.9090909  2.783019
28  {Reason=home,Higher=no}                   => {StudyTime=1}         0.01540832 0.8333333  2.551101
30  {Higher=no,StudyTime=1}                   => {AvgGrade=C}          0.05238829 0.7727273  2.520101
32  {Reason=course,Higher=no}                 => {AvgGrade=C}          0.04930663 0.8205128  2.675944
33  {Higher=no,StudyTime=2}                   => {AvgGrade=C}          0.02157165 0.7368421  2.403068
83  {Higher=yes,StudyTime=4,AvgGrade=A}       => {Reason=reputation}   0.01078582 0.8750000  3.971154
87  {Reason=other,Higher=no,AvgGrade=C}       => {StudyTime=1}         0.01232666 0.8888889  2.721174
88  {Reason=other,Higher=no,StudyTime=1}      => {AvgGrade=C}          0.01232666 0.8000000  2.609045
90  {Reason=course,Higher=no,StudyTime=1}     => {AvgGrade=C}          0.03081664 0.8695652  2.835919
92  {Reason=course,Higher=no,StudyTime=2}     => {AvgGrade=C}          0.01386749 0.7500000  2.445980
```

For Math course: 1. Student with no interest of higher course usually use very little time in studying. This rule is stronger when the student's average grade is low or their motivation to attend is "course". 2. Students spending much time in studying and getting higher grades is probably because they want to get reputation through the course study. 3.From line 69, we can assume that if a student spent much time in studying but doesn't get good grades, the reason he chose the course may be the course itself, i.e. the student loves the course no matter he is talented in it or not.

For Portuguese course: 1. Rules found in this dataset is similar to those above. But with several additional points. 2. Students with no specific motivation to learn the course and don't want to take higher courses usually spend very little time in studying. 3. If a student is not willing to take higher courses and spend little time in studying, his grades may be very low (indicated from line 88,90,92).

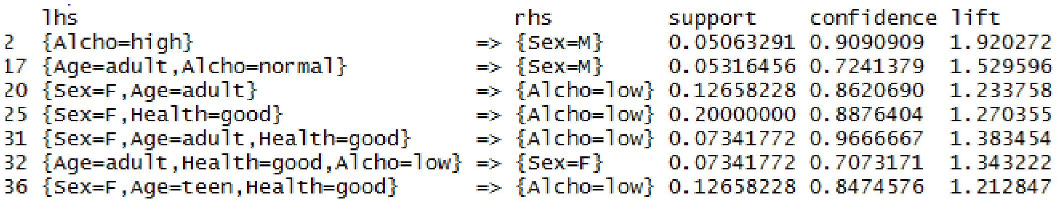
2.3 Sex-Age-Total Alcohol Per Week- Health Condition

Data Pre-Processing: 1. Categorize age into "teen"(0-17) and "adult"(18+) 2. Sum up workday and weekend alcohol consumption and categorize it into "low", "normal", and "high" 3. Categorize health condition from 1-5 to "bad" (0-2), "good" (3-4), and "perfect" (5).



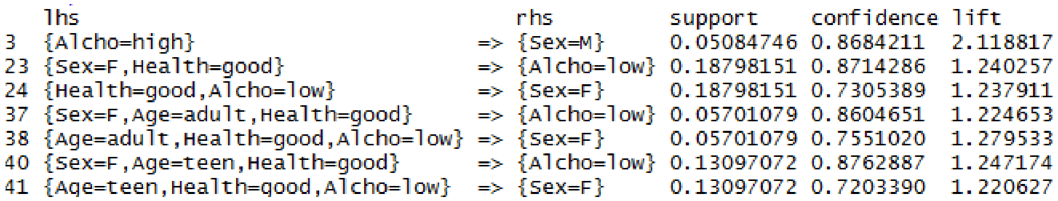
```
#7.Sex-Age-TotalAlcho-Health
dat7=data.frame(d1$sex,d1$age,d1$health,Alcho=rowSums(d1[,27:28]))
summary(dat7)
names(dat7)=c("Sex","Age","Health","Alcho")
dat7$Age=cut(dat7$Age,c(0,17,26),labels = c("teen","adult"))
dat7$Alcho=cut(dat7$Alcho,c(0,4,7,11), labels = c("low","normal","high"))
dat7$Health=cut(dat7$Health,c(0,2,4,6),labels = c("bad","good","perfect"))
rules7=apriori(dat7,parameter = list(supp=0.05,conf=0.7))
inspect(rules7)
rules7.1=subset(rules7,subset=lift>1.2)
inspect(rules7.1)
```

Math:



```
   lhs                                  rhs           support    confidence lift
2  {Alcho=high}                      => {Sex=M}       0.05063291 0.9090909  1.920272
17 {Age=adult,Alcho=normal}          => {Sex=M}       0.05316456 0.7241379  1.529596
20 {Sex=F,Age=adult}                 => {Alcho=low}   0.12658228 0.8620690  1.233758
25 {Sex=F,Health=good}               => {Alcho=low}   0.20000000 0.8876404  1.270355
31 {Sex=F,Age=adult,Health=good}     => {Alcho=low}   0.07341772 0.9666667  1.383454
32 {Age=adult,Health=good,Alcho=low} => {Sex=F}       0.07341772 0.7073171  1.343222
36 {Sex=F,Age=teen,Health=good}      => {Alcho=low}   0.12658228 0.8474576  1.212847
```

Portuguese:

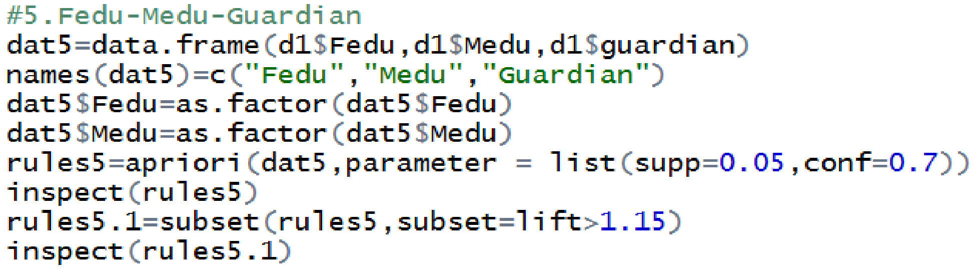


```
   lhs                                       rhs         support    confidence lift
3  {Alcho=high}                           => {Sex=M}     0.05084746 0.8684211  2.118817
23 {Sex=F,Health=good}                    => {Alcho=low} 0.18798151 0.8714286  1.240257
24 {Health=good,Alcho=low}                => {Sex=F}     0.18798151 0.7305389  1.237911
37 {Sex=F,Age=adult,Health=good}          => {Alcho=low} 0.05701079 0.8604651  1.224653
38 {Age=adult,Health=good,Alcho=low}      => {Sex=F}     0.05701079 0.7551020  1.279533
40 {Sex=F,Age=teen,Health=good}           => {Alcho=low} 0.13097072 0.8762887  1.247174
41 {Age=teen,Health=good,Alcho=low}       => {Sex=F}     0.13097072 0.7203390  1.220627
```

From two datasets, we will get similar (same) results:

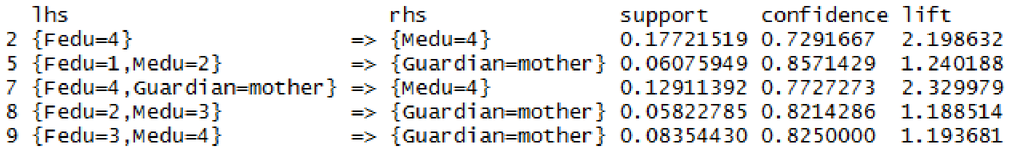
1.From line 2 and 17 we can tell that if a student's weekly alcohol consumption is high, or with normal consumption volume in age of 18 or above, the student's gender usually is male. 2. Adult and teenager female student with good health usually have low alcohol consumption.

2.4 Father Education Level-Mother Education Level-Guardian



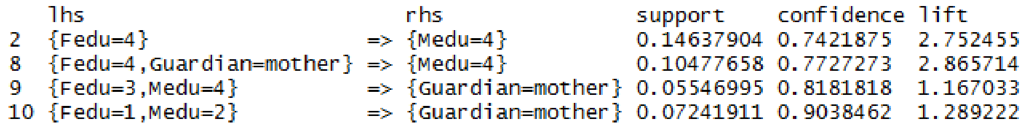
```
#5.Fedu-Medu-Guardian
dat5=data.frame(d1$Fedu,d1$Medu,d1$guardian)
names(dat5)=c("Fedu","Medu","Guardian")
dat5$Fedu=as.factor(dat5$Fedu)
dat5$Medu=as.factor(dat5$Medu)
rules5=apriori(dat5,parameter = list(supp=0.05,conf=0.7))
inspect(rules5)
rules5.1=subset(rules5,subset=lift>1.15)
inspect(rules5.1)
```

Math:



```
  lhs                              rhs               support    confidence lift
2 {Fedu=4}                      => {Medu=4}          0.17721519 0.7291667  2.198632
5 {Fedu=1,Medu=2}               => {Guardian=mother} 0.06075949 0.8571429  1.240188
7 {Fedu=4,Guardian=mother}      => {Medu=4}          0.12911392 0.7727273  2.329979
8 {Fedu=2,Medu=3}               => {Guardian=mother} 0.05822785 0.8214286  1.188514
9 {Fedu=3,Medu=4}               => {Guardian=mother} 0.08354430 0.8250000  1.193681
```

Portuguese:

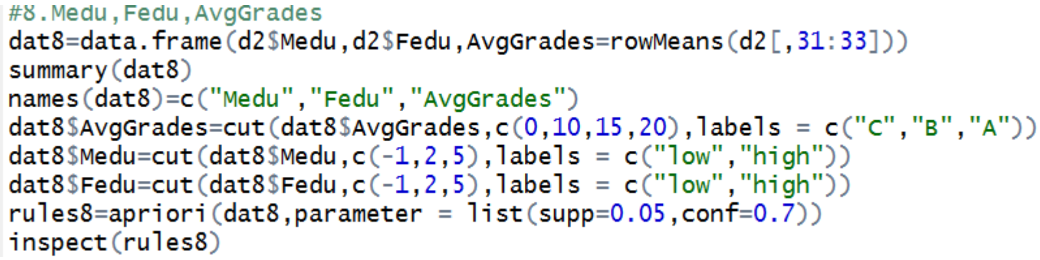


```
   lhs                              rhs               support    confidence lift
2  {Fedu=4}                      => {Medu=4}          0.14637904 0.7421875  2.752455
8  {Fedu=4,Guardian=mother}      => {Medu=4}          0.10477658 0.7727273  2.865714
9  {Fedu=3,Medu=4}               => {Guardian=mother} 0.05546995 0.8181818  1.167033
10 {Fedu=1,Medu=2}               => {Guardian=mother} 0.07241911 0.9038462  1.289222
```

The results we find are similar in two datasets:

1. From Math (2) and Portuguese(2,8), if education level of father side is higher education and above, it's likely that mother education level is also higher education. 2. Math (8,9) and Portuguese (9,10) indicates that if mother education level is higher than father education level, mothers usually plays a role as guardian of student.

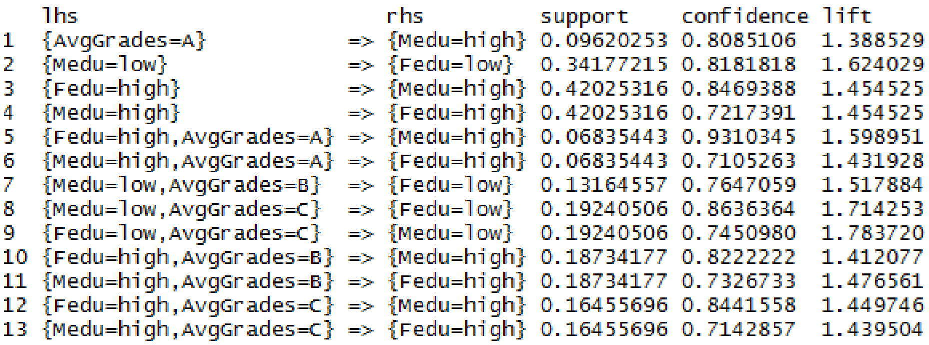
2.5 Father Education Level-Mother Education Level-Average Grade



```
#8.Medu,Fedu,AvgGrades
dat8=data.frame(d2$Medu,d2$Fedu,AvgGrades=rowMeans(d2[,31:33]))
summary(dat8)
names(dat8)=c("Medu","Fedu","AvgGrades")
dat8$AvgGrades=cut(dat8$AvgGrades,c(0,10,15,20),labels = c("C","B","A"))
dat8$Medu=cut(dat8$Medu,c(-1,2,5),labels = c("low","high"))
dat8$Fedu=cut(dat8$Fedu,c(-1,2,5),labels = c("low","high"))
rules8=apriori(dat8,parameter = list(supp=0.05,conf=0.7))
inspect(rules8)
```

Data Pre-Processing: 1. Categorize total grades into 3 levels. 2. Categorize parent education level into 2 levels.

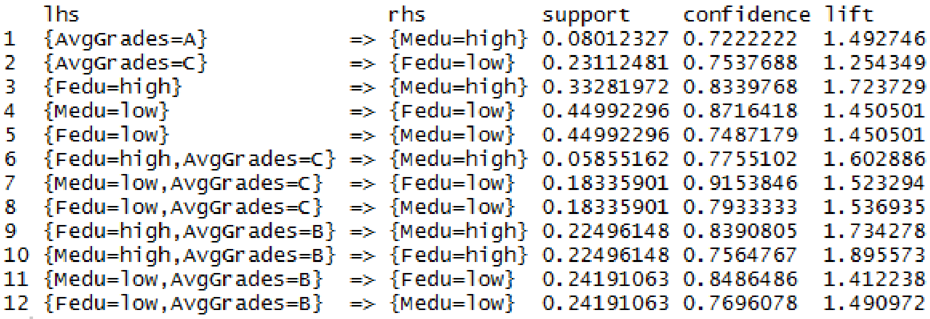
Math:



```
   lhs                           rhs           support    confidence lift
1  {AvgGrades=A}              => {Medu=high} 0.09620253 0.8085106  1.388529
2  {Medu=low}                 => {Fedu=low}  0.34177215 0.8181818  1.624029
3  {Fedu=high}                => {Medu=high} 0.42025316 0.8469388  1.454525
4  {Medu=high}                => {Fedu=high} 0.42025316 0.7217391  1.454525
5  {Fedu=high,AvgGrades=A} => {Medu=high} 0.06835443 0.9310345  1.598951
6  {Medu=high,AvgGrades=A} => {Fedu=high} 0.06835443 0.7105263  1.431928
7  {Medu=low,AvgGrades=B}  => {Fedu=low}  0.13164557 0.7647059  1.517884
8  {Medu=low,AvgGrades=C}  => {Fedu=low}  0.19240506 0.8636364  1.714253
9  {Fedu=low,AvgGrades=C}  => {Medu=low}  0.19240506 0.7450980  1.783720
10 {Fedu=high,AvgGrades=B} => {Medu=high} 0.18734177 0.8222222  1.412077
11 {Medu=high,AvgGrades=B} => {Fedu=high} 0.18734177 0.7326733  1.476561
12 {Fedu=high,AvgGrades=C} => {Medu=high} 0.16455696 0.8441558  1.449746
13 {Medu=high,AvgGrades=C} => {Fedu=high} 0.16455696 0.7142857  1.439504
```

1. From line 1, if the student's grade is good, his or her mother probably have accepted higher education. 2. If a student's mother has low education level, usually his or her father also has low education level (line 2,7,8), the chance may be higher if the student has low average grade.

Portuguese:



```
   lhs                           rhs           support    confidence lift
1  {AvgGrades=A}              => {Medu=high} 0.08012327 0.7222222  1.492746
2  {AvgGrades=C}              => {Fedu=low}  0.23112481 0.7537688  1.254349
3  {Fedu=high}                => {Medu=high} 0.33281972 0.8339768  1.723729
4  {Medu=low}                 => {Fedu=low}  0.44992296 0.8716418  1.450501
5  {Fedu=low}                 => {Medu=low}  0.44992296 0.7487179  1.450501
6  {Fedu=high,AvgGrades=C} => {Medu=high} 0.05855162 0.7755102  1.602886
7  {Medu=low,AvgGrades=C}  => {Fedu=low}  0.18335901 0.9153846  1.523294
8  {Fedu=low,AvgGrades=C}  => {Medu=low}  0.18335901 0.7933333  1.536935
9  {Fedu=high,AvgGrades=B} => {Medu=high} 0.22496148 0.8390805  1.734278
10 {Medu=high,AvgGrades=B} => {Fedu=high} 0.22496148 0.7564767  1.895573
11 {Medu=low,AvgGrades=B}  => {Fedu=low}  0.24191063 0.8486486  1.412238
12 {Fedu=low,AvgGrades=B}  => {Medu=low}  0.24191063 0.7696078  1.490972
```

1. Similar with first result above 2. Parents' education levels are usually in the same level (line 2-12). 3. If a student has low average grade, his father is likely to have low education level.