Week 10 Lab in R: Association Rule Mining

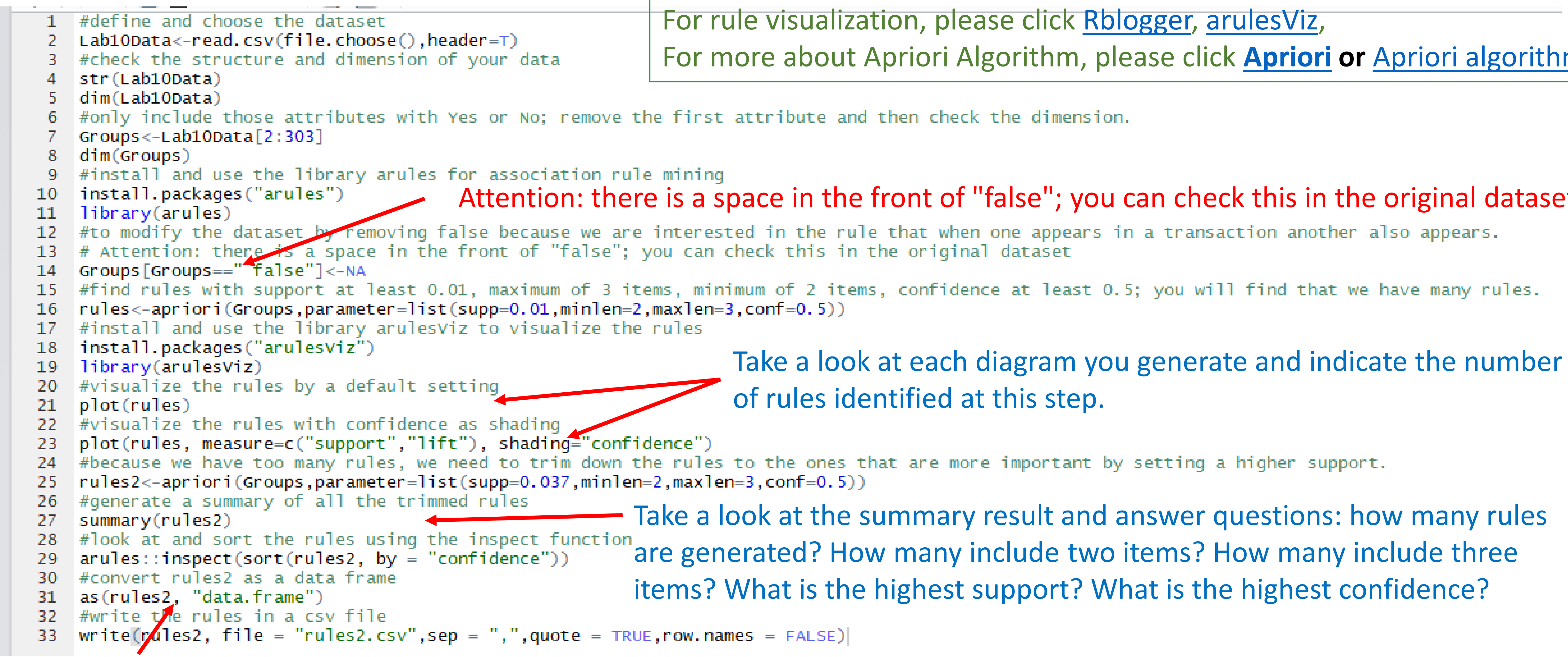
Dr. Liang (Leon) Chen

Instructions

- R script for Association Rule Mining is provided in the next slide, please follow it and complete the lab in R.
- You do not need to type notes (starting at #), but it’s a good manner to have them in you script.
- In order to see codes and notes clearly, I show the script in RStudio.
- After completing this lab, please submit your answers to the questions in LA5 Submission on WTClass.
- **The apriori function will be used in this lab. For more details about it, please visit this site.**

R Script for Association Rule Mining

For more examples, please click arules
For rule visualization, please click Rblogger, arulesViz,
For more about Apriori Algorithm, please click **Apriori** **or** Apriori algorithm



```
#define and choose the dataset
Lab10Data<-read.csv(file.choose(),header=T)
#check the structure and dimension of your data
str(Lab10Data)
dim(Lab10Data)
#only include those attributes with Yes or No; remove the first attribute and then check the dimension.
Groups<-Lab10Data[2:303]
dim(Groups)
#install and use the library arules for association rule mining
install.packages("arules")
library(arules)
#to modify the dataset by removing false because we are interested in the rule that when one appears in a transaction another also appears.
# Attention: there is a space in the front of "false"; you can check this in the original dataset
Groups[Groups==" false"]<-NA
#find rules with support at least 0.01, maximum of 3 items, minimum of 2 items, confidence at least 0.5; you will find that we have many rules.
rules<-apriori(Groups,parameter=list(supp=0.01,minlen=2,maxlen=3,conf=0.5))
#install and use the library arulesViz to visualize the rules
install.packages("arulesViz")
library(arulesViz)
#visualize the rules by a default setting
plot(rules)
#visualize the rules with confidence as shading
plot(rules, measure=c("support","lift"), shading="confidence")
#because we have too many rules, we need to trim down the rules to the ones that are more important by setting a higher support.
rules2<-apriori(Groups,parameter=list(supp=0.037,minlen=2,maxlen=3,conf=0.5))
#generate a summary of all the trimmed rules
summary(rules2)
#look at and sort the rules using the inspect function
arules::inspect(sort(rules2, by = "confidence"))
#convert rules2 as a data frame
as(rules2, "data.frame")
#write the rules in a csv file
write(rules2, file = "rules2.csv",sep = ",",quote = TRUE,row.names = FALSE)
```

Attention: there is a space in the front of "false"; you can check this in the original dataset

Take a look at each diagram you generate and indicate the number of rules identified at this step.

Take a look at the summary result and answer questions: how many rules are generated? How many include two items? How many include three items? What is the highest support? What is the highest confidence?

Take a look at the result after running this command. Compare these rules with those generated at Step 4.2 in the RapidMiner Lab. Are they the same? (Sort the column Confidence in the RapidMiner)