Lab7

Joel Bassett - Jebwizard

April 21, 2017

1 and 2 - Reading

3 - Implement association rules

admissions <- read.csv("C:/Users/bassej4/Downloads/binary.csv")  
  
head(admissions)

## admit gre gpa rank  
## 1 0 380 3.61 3  
## 2 1 660 3.67 3  
## 3 1 800 4.00 1  
## 4 1 640 3.19 4  
## 5 0 520 2.93 4  
## 6 1 760 3.00 2

admissions[,names(admissions)] <- lapply(admissions[,names(admissions)] , factor)  
  
  
library(arules)

## Loading required package: Matrix

##   
## Attaching package: 'arules'

## The following objects are masked from 'package:base':  
##   
## abbreviate, write

print("apriori with defaults")

## [1] "apriori with defaults"

rules.all <- apriori(admissions)

## Apriori  
##   
## Parameter specification:  
## confidence minval smax arem aval originalSupport maxtime support minlen  
## 0.8 0.1 1 none FALSE TRUE 5 0.1 1  
## 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: 40   
##   
## set item appearances ...[0 item(s)] done [0.00s].  
## set transactions ...[164 item(s), 400 transaction(s)] done [0.00s].  
## sorting and recoding items ... [6 item(s)] done [0.00s].  
## creating transaction tree ... done [0.00s].  
## checking subsets of size 1 2 done [0.00s].  
## writing ... [1 rule(s)] done [0.00s].  
## creating S4 object ... done [0.00s].

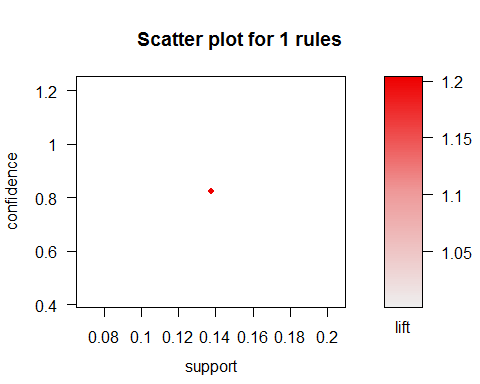
rules.all

## set of 1 rules

library(arulesViz)

## Loading required package: grid

plot(rules.all)



#plot(rules.all, method = "grouped")

4 -

elect <- read.csv("C:/Users/bassej4/Documents/RPI Classes/OpenSource/datalab7/elect.txt")  
  
str(elect)

## 'data.frame': 50 obs. of 4 variables:  
## $ State : Factor w/ 50 levels "Alabama","Alaska",..: 1 2 3 4 5 6 7 8 9 10 ...  
## $ TotalVote: int 1883449 312598 2012585 1054945 12419857 2130330 1578769 375190 7609810 3301875 ...  
## $ BushPct : num 62.5 61.1 54.9 54.3 44.4 ...  
## $ KerryPct : num 36.8 35.5 44.4 44.5 54.3 ...

head(elect)

## State TotalVote BushPct KerryPct  
## 1 Alabama 1883449 62.46 36.84  
## 2 Alaska 312598 61.07 35.52  
## 3 Arizona 2012585 54.87 44.40  
## 4 Arkansas 1054945 54.31 44.55  
## 5 California 12419857 44.36 54.31  
## 6 Colorado 2130330 51.69 47.02

tail(elect)

## State TotalVote BushPct KerryPct  
## 45 Vermont 312309 38.80 58.94  
## 46 Virginia 3198367 53.68 45.48  
## 47 Washington 2859084 45.64 52.82  
## 48 West Virginia 755887 56.06 43.20  
## 49 Wisconsin 2997007 49.32 49.70  
## 50 Wyoming 243428 68.86 29.07

attributes(elect)

## $names  
## [1] "State" "TotalVote" "BushPct" "KerryPct"   
##   
## $class  
## [1] "data.frame"  
##   
## $row.names  
## [1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23  
## [24] 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46  
## [47] 47 48 49 50

summary(elect)

## State TotalVote BushPct KerryPct   
## Alabama : 1 Min. : 243428 Min. :36.78 Min. :26.00   
## Alaska : 1 1st Qu.: 755991 1st Qu.:46.48 1st Qu.:39.37   
## Arizona : 1 Median : 1816321 Median :52.70 Median :46.56   
## Arkansas : 1 Mean : 2441319 Mean :53.14 Mean :45.67   
## California: 1 3rd Qu.: 2975852 3rd Qu.:59.52 3rd Qu.:52.45   
## Colorado : 1 Max. :12419857 Max. :71.54 Max. :61.94   
## (Other) :44

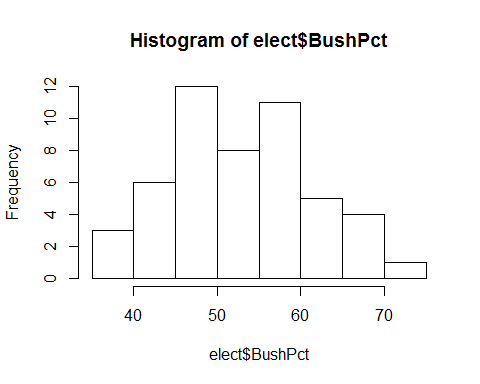
var(elect$BushPct)

## [1] 71.39077

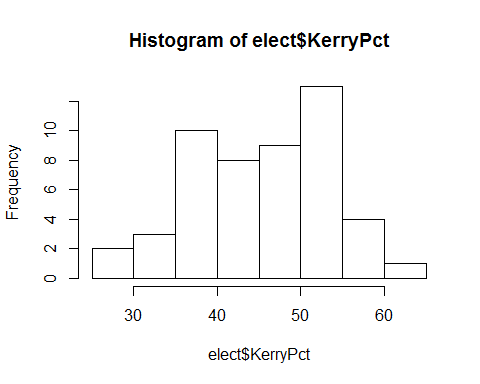
var(elect$KerryPct)

## [1] 72.06318

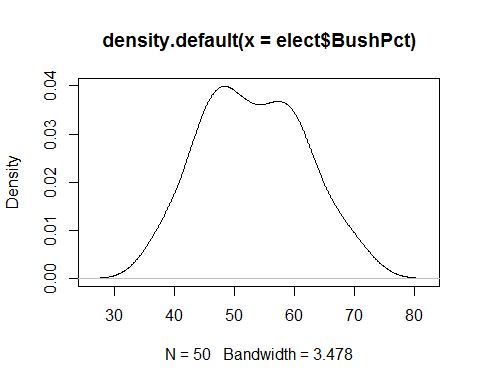
hist(elect$BushPct)



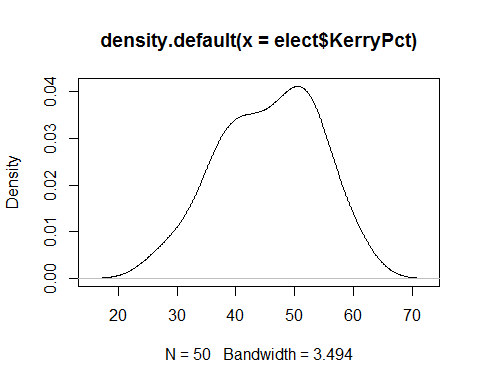
hist(elect$KerryPct)



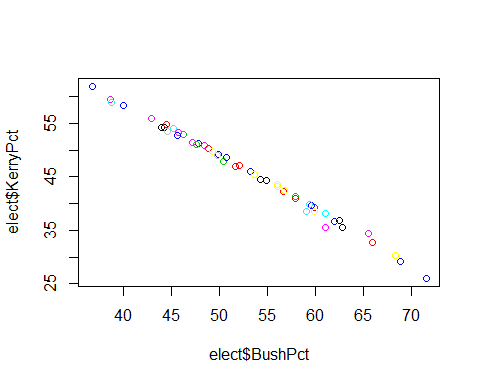
plot(density(elect$BushPct))



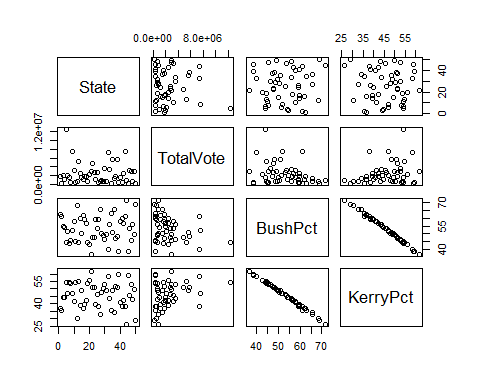
plot(density(elect$KerryPct))



with(elect, plot(elect$BushPct, elect$KerryPct, col = elect$TotalVote))



pairs(elect)



topmovies <- read.csv("C:/Users/bassej4/Documents/RPI Classes/OpenSource/datalab7/topmovies.txt")  
  
str(topmovies)

## 'data.frame': 1000 obs. of 5 variables:  
## $ rank: int 1 2 3 4 5 6 7 8 9 10 ...  
## $ name: Factor w/ 996 levels "10","10000 B.C.",..: 73 912 754 683 632 246 680 543 921 660 ...  
## $ box : Factor w/ 992 levels "$ 52.581","$ 52.823",..: 992 991 990 989 988 987 986 985 984 983 ...  
## $ date: Factor w/ 349 levels "April 01","April 02",..: 75 76 153 260 254 181 254 168 188 239 ...  
## $ year: num 2009 1997 2008 1977 2004 ...

head(topmovies)

## rank name box date year  
## 1 1 Avatar $759.563 December 18 2009  
## 2 2 Titanic $600.788 December 19 1997  
## 3 3 The Dark Knight $533.184 July 18 2008  
## 4 4 Star Wars: Episode IV - A New Hope $460.998 May 25 1977  
## 5 5 Shrek 2 $437.212 May 19 2004  
## 6 6 E.T. the Extra-Terrestrial $434.975 June 11 1982

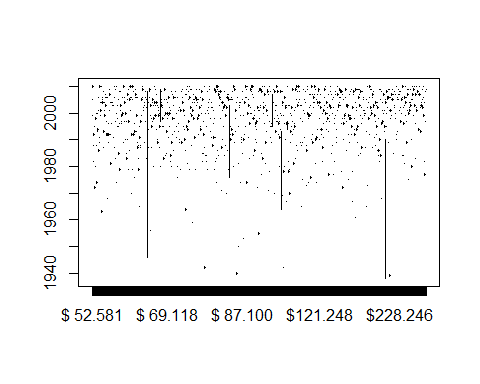
tail(topmovies)

## rank name box date year  
## 995 995 Beethoven $ 57.114 April 03 1992  
## 996 996 Annie $ 57.059 May 21 1982  
## 997 997 Beaches $ 57.042 December 21 1988  
## 998 998 Message in a Bottle $ 52.823 February 12 1999  
## 999 999 Resident Evil: Afterlife $ 56.615 September 10 2010  
## 1000 1000 Kicking and Screaming $ 52.581 May 13 2005

summary(topmovies)

## rank name box   
## Min. : 1.0 101 Dalmatians : 2 $ 65.000: 2   
## 1st Qu.: 250.8 The Amityville Horror: 2 $ 67.128: 2   
## Median : 500.5 The Karate Kid : 2 $ 72.779: 2   
## Mean : 500.5 The Longest Yard : 2 $ 82.500: 2   
## 3rd Qu.: 750.2 10 : 1 $100.192: 2   
## Max. :1000.0 10000 B.C. : 1 $102.300: 2   
## (Other) :990 (Other) :988   
## date year   
## December 25: 21 Min. :1938   
## December 15: 14 1st Qu.:1992   
## December 17: 11 Median :2000   
## December 21: 11 Mean :1998   
## December 22: 11 3rd Qu.:2006   
## November 21: 11 Max. :2010   
## (Other) :921

with(topmovies, plot(topmovies$box, topmovies$year, col = topmovies$rank))



pairs(topmovies)

