title: "HW1" author: "Anindita Deb" date: "3/2/2022" output: pdf_document

################################ First Task : Exploring College Data of ISLR2 Pakage ##

###########Visulaization of College Dataset of ISLR2 Package without scaling###
College_modified<-College College_modified$Apps<-log(College_modified$Apps) College_modified$Accept<-log(College_modified$Accept) College_modified$Enroll<-log(College_modified$Enroll) College_modified$Top10perc<-log(College_modified$Top10perc) College_modified$Top25perc<-log(College_modified$Top25perc) College_modified$F.Undergrad<-log(College_modified$F.Undergrad) College_modified$P.Undergrad<-log(College_modified$P.Undergrad) College_modified$Outstate<-log(College_modified$Outstate) College_modified$Room.Board<-log(College_modified$Room.Board) College_modified$Books<-log(College_modified$Books) College_modified$Personal<-log(College_modified$Personal) College_modified$PhD<-log(College_modified$PhD) College_modified$Terminal<-log(College_modified$Terminal) College_modified$S.F.Ratio<-log(College_modified$S.F.Ratio) College_modified$perc.alumni<-log(College_modified$perc.alumni) College_modified$Expend<-log(College_modified$Expend) College_modified$Grad.Rate<-log(College_modified$Grad.Rate) M = cor(College[,2:18]) corrplot(M) M = cor(College_Modified[,2:18]) corrplot(College_Modified) ################Partitioning the dataset into private and public## Private_College_Data<-subset(College,Private=="Yes") Public_College_Data<-subset(College,Private=="No") setwd("C:/Users/anind/OneDrive/Desktop/Spring2022/EAS/") save(Private_College_Data, file = "Private_College_Data.Rdata") save(Public_College_Data, file = "Public_College_Data.Rdata") Private_College_Data<-Private_College_Data[order(Private_College_Data$Apps, decreasing = TRUE),] Public_College_Data<-Public_College_Data[order(Public_College_Data$Apps, decreasing = TRUE),] Private_College_Data=Private_College_Data[Private_College_Data$Top25perc>median(Private_College_Data$Top25perc),] Public_College_Data=Public_College_Data[Public_College_Data$Top25perc>median(Public_College_Data$Top25perc),] x11() par(mfrow=c(1,2)) hist(Public_College_Data$Grad.Rate) hist(Private_College_Data$Grad.Rate) Public_College_Data[["GradRateMod"]]=ordered(cut(Public_College_Data[["Grad.Rate"]],c(0,30,50,80,100), labels=c("Low","Medium","High","Low")))

Private_College_Data[["GradRateMod"]]=ordered(cut(Private_College_Data[["Grad.Rate"]],c(0,60,85,100),labels=c("Low","High","Medium"))) save(Private_College_Data, file = "Private_College_Data_mod.Rdata") save(Public_College_Data, file = "Public_College_Data_mod.Rdata") ##################### End of Task 1###

######################Second Task : Learning to implement Generalized Association Rules of Market Data ### summary(marketing) h<-hist(marketing$Income, main="Histogram for Income", xlab="Income", border="blue", col="green", las=1, breaks=5)

> colnames(marketing) [1] "Income" "Sex" "Marital" "Age"
> [5] "Edu" "Occupation" "Lived" "Dual_Income" [9] "Household" "Householdu18" "Status" "Home_Type"
> [13] "Ethnic" "Language"

h<-hist(marketing$Income, main="Histogram for Income", xlab="Income", border="blue", col="green", las=1, breaks=5) marketing$Income[marketing$Income == 1]<-"Less than $10,000"
marketing$Income[marketing$Income == 2]<-"$10,000 to $14,999"
marketing$Income[marketing$Income == 3]<-"$15,000 to $19,999"
marketing$Income[marketing$Income == 4]<-"$20,000 to $24,999"
marketing$Income[marketing$Income == 5]<-"$25,000 to $29,999"
marketing$Income[marketing$Income == 6]<-"$30,000 to $39,999"
marketing$Income[marketing$Income == 7]<-"$40,000 to $49,999"
marketing$Income[marketing$Income == 8]<-"$50,000 to $74,999"
marketing$Income[marketing$Income == 9]<-"$75,000 or more"

marketing$Sex[marketing$Sex == 1] <-"Male" marketing$Sex[marketing$Sex == 2] <-"Female"

marketing$Sex<-as.factor(marketing$Sex) h<-hist(marketing$Marital, main="Histogram for Marital Status", xlab="Marital Status", border="blue", col="green", las=1, breaks=5) marketing$Marital[marketing$Marital == 1] <-"Married" marketing$Marital[marketing$Marital == 2] <-"Living together, not married" marketing$Marital[marketing$Marital == 3] <-"Divorced or separated" marketing$Marital[marketing$Marital == 4] <-"Widowed" marketing$Marital[marketing$Marital == 5] <-"Single, never married"

h<-hist(marketing$Age, main="Histogram for Age", xlab="Age", border="blue", col="green", las=1, breaks=3) marketing$Age[marketing$Age == 1] <-"14 thru 17" marketing$Age[marketing$Age == 2] <-"18 thru 24" marketing$Age[marketing$Age == 3] <-"25 thru 34" marketing$Age[marketing$Age == 4] <-"35 thru 44" marketing$Age[marketing$Age == 5] <-"45 thru 54" marketing$Age[marketing$Age == 6] <-"55 thru 64" marketing$Age[marketing$Age == 7] <-"65 and Over"

```
h<-hist(marketing$Edu, main="Histogram for Education", xlab="Education", border="blue", col="green", las=1, breaks=5) marketing$Edu[marketing$Edu == 1] <-"Grade 8 or less"
marketing$Edu[marketing$Edu == 2] <-"Grades 9 to 11" marketing$Edu[marketing$Edu == 3] <-
"Graduated high school" marketing$Edu[marketing$Edu == 4] <-"1 to 3 years of college"
marketing$Edu[marketing$Edu == 5] <-"College graduate" marketing$Edu[marketing$Edu == 6] <-
"Grad Study" h<-hist(marketing$Occupation, main="Histogram for Occupation", xlab="Occupation",
border="blue", col="green", las=1, breaks=5) marketing$Occupation[marketing$Occupation == 1] <-
"Professional/Managerial" marketing$Occupation[marketing$Occupation == 2] <-"Sales Worker"
marketing$Occupation[marketing$Occupation == 3] <-"Factory Worker/Laborer/Driver"
marketing$Occupation[marketing$Occupation == 4] <-"Clerical/Service Worker"
marketing$Occupation[marketing$Occupation == 5] <-"Homemaker"
marketing$Occupation[marketing$Occupation == 6] <-"Student, HS or College"
marketing$Occupation[marketing$Occupation == 7] <-"Military"
marketing$Occupation[marketing$Occupation == 8] <-"Retired"
marketing$Occupation[marketing$Occupation == 9] <-"Unemployed"

marketing$Lived[marketing$Lived == 1] <-"Less than one year" marketing$Lived[marketing$Lived == 2]
<-"One to three years" marketing$Lived[marketing$Lived == 3] <-"Four to six years"
marketing$Lived[marketing$Lived == 4] <-"Seven to ten years" marketing$Lived[marketing$Lived == 5]
<-"More than ten years"

marketing$Dual_Income[marketing$Dual_Income == 1] <-"Not Married"
marketing$Dual_Income[marketing$Dual_Income == 2] <-"Yes"
marketing$Dual_Income[marketing$Dual_Income == 3] <-"No"

marketing$Household[marketing$Household == 1] <-"One"
marketing$Household[marketing$Household == 2] <-"Two"
marketing$Household[marketing$Household == 3] <-"Three"
marketing$Household[marketing$Household == 4] <-"Four"
marketing$Household[marketing$Household == 5] <-"Five"
marketing$Household[marketing$Household == 6] <-"Six" marketing$Household[marketing$Household
== 7] <-"Seven" marketing$Household[marketing$Household == 8] <-"Seven"
marketing$Household[marketing$Household == 9] <-"Nine or more"

marketing$Householdu18[marketing$Householdu18 == 0] <-"None"
marketing$Householdu18[marketing$Householdu18 == 1] <-"One"
marketing$Householdu18[marketing$Householdu18 == 2] <-"Two"
marketing$Householdu18[marketing$Householdu18 == 3] <-"Three"
marketing$Householdu18[marketing$Householdu18 == 4] <-"Four"
marketing$Householdu18[marketing$Householdu18 == 5] <-"Five"
marketing$Householdu18[marketing$Householdu18 == 6] <-"Six"
marketing$Householdu18[marketing$Householdu18 == 7] <-"Seven"
```

```
marketing$Householdu18[marketing$Householdu18 == 8] <-"Eight"
marketing$Householdu18[marketing$Householdu18 == 9] <-"Nine or more"

marketing$Status[marketing$Status == 1]<-"Own" marketing$Status[marketing$Status == 2]<-"Rent"
marketing$Status[marketing$Status == 3]<-"Live with Parents/Family"

marketing$Home_Type[marketing$Home_Type == 1]<-"House"
marketing$Home_Type[marketing$Home_Type == 2]<-"Condominium"
marketing$Home_Type[marketing$Home_Type == 3]<-"Apartment"
marketing$Home_Type[marketing$Home_Type == 4]<-"Mobile Home"
marketing$Home_Type[marketing$Home_Type == 5]<-"Other"

marketing$Ethnic[marketing$Ethnic == 1]<-"American Indian" marketing$Ethnic[marketing$Ethnic == 2]
<-"Asian" marketing$Ethnic[marketing$Ethnic == 3]<-"Black" marketing$Ethnic[marketing$Ethnic == 4]
<-"East Indian" marketing$Ethnic[marketing$Ethnic == 5]<-"Hispanic"
marketing$Ethnic[marketing$Ethnic == 6]<-"Pacific Islander" marketing$Ethnic[marketing$Ethnic == 7]
<-"White"
marketing$Ethnic[marketing$Ethnic == 8]<-"Other"

marketing$Language[marketing$Language == 1]<-"English" marketing$Language[marketing$Language
== 2]<-"Spanish" marketing$Language[marketing$Language == 3]<-"Other"

marketing$Income<-as.factor(marketing$Income) marketing$Sex<-as.factor(marketing$Sex)
marketing$Marital<-as.factor(marketing$Marital) marketing$Age<-as.factor(marketing$Age)
marketing$Edu<-as.factor(marketing$Edu) marketing$Occupation<-as.factor(marketing$Occupation)
marketing$Lived<-as.factor(marketing$Lived) marketing$Dual_Income<-
as.factor(marketing$Dual_Income) marketing$Household<-as.factor(marketing$Household)
marketing$Householdu18<-as.factor(marketing$Householdu18) marketing$Status<-
as.factor(marketing$Status) marketing$Home_Type<-as.factor(marketing$Home_Type)
marketing$Ethnic<-as.factor(marketing$Ethnic) marketing$Language<-as.factor(marketing$Language)

size<-nrow(marketing) set.seed(1) marketing_mod<-marketing marketing_mod$Income<-
sample(marketing_mod$Income,size=size,replace=TRUE) marketing_mod$Sex<-
sample(marketing_mod$Sex,size=size,replace=TRUE) marketing_mod$Marital<-
sample(marketing_mod$Marital,size=size,replace=TRUE) marketing_mod$Age<-
sample(marketing_mod$Age,size=size,replace=TRUE) marketing_mod$Edu<-
sample(marketing_mod$Edu,size=size,replace=TRUE) marketing_mod$Occupation<-
sample(marketing_mod$Occupation,size=size,replace=TRUE) marketing_mod$Lived<-
sample(marketing_mod$Lived,size=size,replace=TRUE) marketing_mod$Dual_Income<-
sample(marketing_mod$Dual_Income,size=size,replace=TRUE) marketing_mod$Household<-
sample(marketing_mod$Household,size=size,replace=TRUE) marketing_mod$Householdu18<-
sample(marketing_mod$Householdu18,size=size,replace=TRUE) marketing_mod$Status<-
```

sample(marketing_mod$Status,size=size,replace=TRUE) marketing_mod$Home_Type<-
sample(marketing_mod$Home_Type,size=size,replace=TRUE) marketing_mod$Ethnic<-
sample(marketing_mod$Ethnic,size=size,replace=TRUE) marketing_mod$Language<-
sample(marketing_mod$Language,size=size,replace=TRUE)

library("rpart") library("rpart.plot") Y<-rep(1,size=size) marketing<-cbind(marketing,Y) Y<-
rep(0,size=size) marketing_mod<-cbind(marketing_mod,Y) overall_mar<-
rbind(marketing,marketing_mod) overall_mar$Y<-as.factor(overall_mar$Y) model.controls <-
rpart.control(minbucket = 2, minsplit = 4, xval = 10) fit.overall_mar<-rpart(Y~., data =
overall_mar,method="class",control = model.controls)
min_cp=which(fit.overall_mar$cptable[,4]==min(fit.overall_mar$cptable[,4])) pruned_fit_overall_mar<-
prune(fit.overall_mar,cp=fit.overall_mar$cptable[min_cp,1]) rpart.plot(fit.overall_mar,main="Classification
Tree for unsupervised learning - Marketing Data Before Pruning")
rpart.plot(pruned_fit_overall_mar,main="Classification Tree for unsupervised learning - Marketing Data
After Pruning") rpart.rules(pruned_fit_overall_mar,cover=TRUE) rpart_summary<-
pruned_fit_overall_mar[1] rpart.rules(Y,cover=TRUE) fit.overall_mar<-rpart(Y~., data =
overall_mar,method="class",control = model.controls)
plot(fit.overall_mar$cptable[,4],type="o",lty=1,col='blue',main="Cp for model selection",ylab="cv error")
x<-which(fit.overall_mar$frame[,'var']=="" & fit.overall_mar$frame[,'yval']==2) n<-
as.vector(fit.overall_mar$frame[x,'n']) probability_mat<-fit.overall_mar$frame[x,'yval2'] probability<-
as.vector(probability_mat[,5]) support<-(probability*n)/size support_percent*<-support*100
fit.overall_mar_mod<-cbind(fit.overall_mar$frame[x,],support_percent)
####################Including only those fit.overall_mar_mod <- fit.overall_mar$frame[x,]
rpart.plot(pruned_fit_overall_mar,main="Classification Tree for unsupervised learning - Marketing Data")
x=rpart.rules(pruned_fit_overall_mar, cover = TRUE) z=x$Y z<-as.numeric(z) support<-(0.20*z)/(2*size)

actual <- rpart(Y="1" ~ ., data = overall_mar) x11() rpart.plot(actual, type = 3, clip.right.labs = FALSE,
branch = .3, under = TRUE) rpart.rules(overall_mar, cover = TRUE)

####################Task 3 - Exploring Boston Housing Data and Generating Association rules
using Apriori Algorithm#########################

###############Visualizing the data and generating histogram plots
###
sum(is.na(Boston)) is.null(Boston)

crim zn indus chas nox rm age dis rad tax ptratio lstat medv

h<-hist(Boston$age ,Boston$crim, main="Histogram for Crime Rate vs Age", xlab="Age", ylab="Crime
Rate", border="blue", col="green", las=1, breaks=5) h<-hist(Boston$indus ,Boston$crim,
main="Histogram for Crime Rate vs non-retailArea", xlab="Propertion of non-retail business",
ylab="Crime Rate", border="blue", col="green", las=1, breaks=5) h<-hist(Boston$lstat ,Boston$crim,

```
main="Histogram for Crime Rate vs Status", xlab="Status", ylab="Crime Rate", border="blue", col="green", las=1, breaks=5) h<-hist(Boston$zn,Boston$nox, main="Histogram Nitrogen Level vs ResArea", xlab="ResArea", ylab="Nitrogen", border="blue", col="green", las=1, breaks=5) h<-hist(Boston$indus,Boston$nox, main="Histogram Nitrogen Level vs NonRetailArea", xlab="NonRetailArea", ylab="Nitrogen", border="blue", col="green", las=1, breaks=5)

h<-hist(Boston$zn,Boston$tax, main="Histogram Tax vs Res", xlab="ResidentialArea", ylab="Tax", border="blue", col="green", las=1, breaks=5) h<-hist(Boston$indus,Boston$tax, main="Histogram Tax vs NonRetailReg", xlab="NonRetailReg", ylab="Tax", border="blue", col="green", las=1, breaks=5) h<-hist(Boston$dis,Boston$nox, main="Histogram Nitrogen Level vs Weighted Mean Distance", xlab="Weighted Mean Distance", ylab="Nitrogen Level", border="blue", col="green", las=1, breaks=3) h<-hist(Boston$chas,Boston$nox, main="Histogram Nitrogen Level vs Charles River", xlab="Track Bounds Charles River", ylab="Nitrogen Level", border="blue", col="green", las=1, breaks=5) h<-hist(Boston$chas,Boston$crim, main="Histogram Crime Rate vs Charles River", xlab="Track Bounds Charles River", ylab="Crime Rate", border="blue", col="green", las=1, breaks=5) h<-hist(Boston$dis,Boston$ptratio, main="Histogram of Student/Teacher ratio vs dis", xlab="Weighted mean of distances to Boston Employment Cent", ylab="Student/Teacher ratio", border="blue", col="green", las=1, breaks=3) h<-hist(Boston$dis,Boston$crim, main="Histogram of Crime Rate vs dis", xlab="Weighted mean of distances to Boston Employment Cent", ylab="Crime Rate", border="blue", col="green", las=1, breaks=5) h<-hist(Boston$dis,Boston$zn, main="Histogram of Residential Area vs dis", xlab="Weighted mean of distances to Boston Employment Cent", ylab="Residential Area", border="blue", col="green", las=1, breaks=3)

h<-hist(Boston$age, main="Histogram for Age", xlab="Age", border="blue", col="green", las=1) h<-hist(Boston$chas, main="Histogram for Chas", xlab="chas", border="blue", col="green") Boston[["age"]] <- ordered(cut(Boston[["age"]], c(0, 35, 60, 80, 100)), labels = c("Young", "Middle-aged", "Senior", "Elderly")) Boston$chas[Boston$chas == 0] <-"No" Boston$chas[Boston$chas == 1] <-"Yes"

Boston$chas<-as.factor(Boston$chas)

h<-hist(Boston$nox, main="Histogram for Nitrogen Level", xlab="Nitrogen Conc.", border="blue", col="green", las=1, breaks=3) Boston[["nox"]] <- ordered(cut(Boston[["nox"]], c(0.2, 0.5, 0.7, 1)), labels = c("Low", "Medium", "High"))

h<-hist(Boston$rm, main="Histogram for Avg no of rooms", xlab="Avg no of rooms", border="blue", col="green", las=1, breaks=3) Boston[["rm"]] <- ordered(cut(Boston[["rm"]], c(2,4,7,10)), labels = c("Small", "Medium", "Large")) h<-hist(Boston$zn, main="Histogram for Residential Area", xlab="Residential Area", border="blue", col="green", las=1, breaks=3) Boston[["zn"]] <- ordered(cut(Boston[["zn"]], c(0,40,70,100)), labels = c("Small", "Average", "Large")) h<-hist(Boston$indus, main="Histogram Non Retail Business Acre", xlab="Non Retail Area", border="blue", col="green", las=1, breaks=3) Boston[["indus"]] <- ordered(cut(Boston[["indus"]], c(0,10,20,30)), labels = c("Small", "Medium", "Large"))
```

h<-hist(Boston$dis, main="Weighted Mean distance to 5 Boston Employment Centers", xlab="Weighted Mean distance", border="blue", col="green", las=1, breaks=3) Boston[["dis"]] <- ordered(cut(Boston[["dis"]], c(0,5,10,15)), labels = c("Close", "Average", "Distant")) h<-hist(Boston$rad, main="Index of accessibility to radial highways", xlab="Index of accessibility", border="blue", col="green") Boston$rad[Boston$rad ==1] <-"Close" Boston$rad[Boston$rad ==2] <-"Close" Boston$rad[Boston$rad ==3] <-"Close" Boston$rad[Boston$rad ==4] <-"Close" Boston$rad[Boston$rad ==5] <-"Close" Boston$rad[Boston$rad ==6] <-"Average" Boston$rad[Boston$rad ==7] <-"Average" Boston$rad[Boston$rad ==8] <-"Average" Boston$rad[Boston$rad ==24] <-"Distant" Boston$rad<-as.factor(Boston$rad)

h<-hist(Boston$tax, main="Histogram of tax", xlab="Tax", border="blue", col="green", las=1, breaks=3) Boston[["tax"]] <- ordered(cut(Boston[["tax"]], c(0,250,500,800)), labels = c("Low", "Medium", "High"))

h<-hist(Boston$ptratio, main="Histogram of Student/Teacher ratio", xlab="Student/Teacher ratio", border="blue", col="green", las=1, breaks=3) Boston[["ptratio"]] <- ordered(cut(Boston[["ptratio"]], c(10,15,20,25)), labels = c("Small", "Medium", "Large"))

h<-hist(Boston$lstat, main="Histogram of lower status", xlab="lower status", border="blue", col="green", las=1, breaks=3) Boston[["lstat"]] <- ordered(cut(Boston[["lstat"]], c(0,15,30,40)), labels = c("VeryLow", "Middle", "Class")) h<-hist(Boston$medv, main="Histogram of median value of owner occupied houses", xlab="medv", border="blue", col="green", las=1, breaks=3) Boston[["medv"]] <- ordered(cut(Boston[["medv"]], c(0,20,40,60)), labels = c("Cheap", "Medium", "Expensive")) h<-hist(Boston$crim, main="Histogram of Crime Rate", xlab="Crime", border="blue", col="green", las=1, breaks=3) Boston[["crim"]] <- ordered(cut(Boston[["crim"]], c(0,20,60,100)), labels = c("Average", "High", "VeryHigh")) library("arules") Boston <- as(Boston, "transactions") summary(Boston) itemFrequencyPlot(Boston, support = 0.01, cex.names = 0.8)

rules <- apriori(Boston, parameter = list(support = 0.01, confidence = 0.6)) rulesLowCrim <- subset(rules, subset = rhs %in% "crim=Average" & lhs %in% "dis=Close" & lift>1) inspect(head(sort(rulesLowCrim,by="confidence"))) rulesptratio <- subset(rules, subset = rhs %in% "ptratio=Small" & lift>5) inspect(head(sort(rulesptratio,by="confidence")))

```
knitr::opts_chunk$set(echo = TRUE)
```

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

Including Plots

You can also embed plots, for example:

```
plot(pressure)
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.