hw1.R

Brian Kang

Wed Apr 10 09:49:26 2019

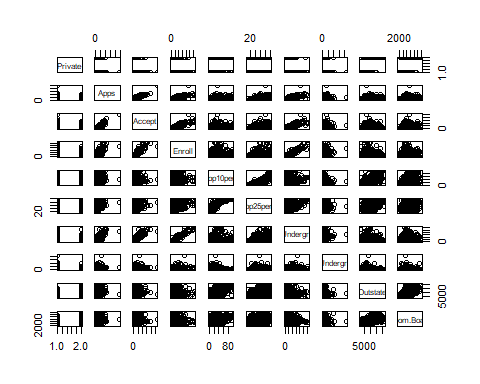
# Hw1  
# 4  
# a)  
setwd(dir = "C:/Users/slexi/Documents/ECON484")  
#install.packages("ISLR", lib = "C:/Users/slexi/Documents/ECON484")  
library(ISLR)

## Warning: package 'ISLR' was built under R version 3.5.3

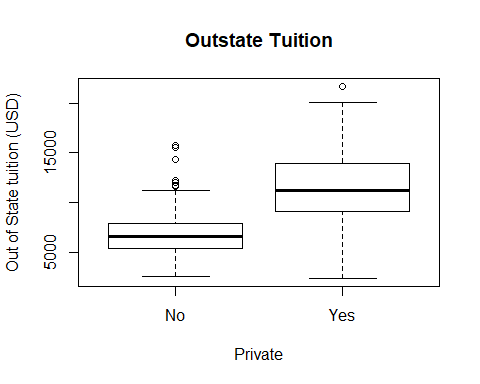
data("College")  
college <- read.csv("College.csv")  
  
# b)  
# The names of colleges are considered as a data still,  
# so make a new "private" column with the names of the   
# colleges.  
rownames(college) = college[,1]  
fix(college)  
  
# Now get rid of the old "public" column with the names  
# that were still considered data  
college = college[,-1]  
fix(college)  
  
# c)  
# i)  
summary(college)

## Private Apps Accept Enroll Top10perc   
## No :212 Min. : 81 Min. : 72 Min. : 35 Min. : 1.00   
## Yes:565 1st Qu.: 776 1st Qu.: 604 1st Qu.: 242 1st Qu.:15.00   
## Median : 1558 Median : 1110 Median : 434 Median :23.00   
## Mean : 3002 Mean : 2019 Mean : 780 Mean :27.56   
## 3rd Qu.: 3624 3rd Qu.: 2424 3rd Qu.: 902 3rd Qu.:35.00   
## Max. :48094 Max. :26330 Max. :6392 Max. :96.00   
## Top25perc F.Undergrad P.Undergrad Outstate   
## Min. : 9.0 Min. : 139 Min. : 1.0 Min. : 2340   
## 1st Qu.: 41.0 1st Qu.: 992 1st Qu.: 95.0 1st Qu.: 7320   
## Median : 54.0 Median : 1707 Median : 353.0 Median : 9990   
## Mean : 55.8 Mean : 3700 Mean : 855.3 Mean :10441   
## 3rd Qu.: 69.0 3rd Qu.: 4005 3rd Qu.: 967.0 3rd Qu.:12925   
## Max. :100.0 Max. :31643 Max. :21836.0 Max. :21700   
## Room.Board Books Personal PhD   
## Min. :1780 Min. : 96.0 Min. : 250 Min. : 8.00   
## 1st Qu.:3597 1st Qu.: 470.0 1st Qu.: 850 1st Qu.: 62.00   
## Median :4200 Median : 500.0 Median :1200 Median : 75.00   
## Mean :4358 Mean : 549.4 Mean :1341 Mean : 72.66   
## 3rd Qu.:5050 3rd Qu.: 600.0 3rd Qu.:1700 3rd Qu.: 85.00   
## Max. :8124 Max. :2340.0 Max. :6800 Max. :103.00   
## Terminal S.F.Ratio perc.alumni Expend   
## Min. : 24.0 Min. : 2.50 Min. : 0.00 Min. : 3186   
## 1st Qu.: 71.0 1st Qu.:11.50 1st Qu.:13.00 1st Qu.: 6751   
## Median : 82.0 Median :13.60 Median :21.00 Median : 8377   
## Mean : 79.7 Mean :14.09 Mean :22.74 Mean : 9660   
## 3rd Qu.: 92.0 3rd Qu.:16.50 3rd Qu.:31.00 3rd Qu.:10830   
## Max. :100.0 Max. :39.80 Max. :64.00 Max. :56233   
## Grad.Rate   
## Min. : 10.00   
## 1st Qu.: 53.00   
## Median : 65.00   
## Mean : 65.46   
## 3rd Qu.: 78.00   
## Max. :118.00

# ii)  
pairs(college[,1:10])



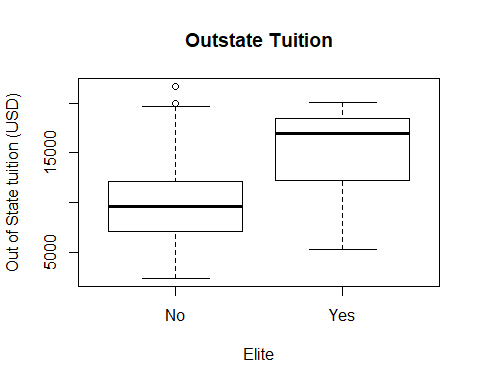
# iii)  
plot(college$Private, college$Outstate,   
 xlab = "Private",   
 ylab ="Out of State tuition (USD)",   
 main = "Outstate Tuition")



# iv)  
Elite = rep("No", nrow(college))  
Elite[college$Top10perc > 50] = "Yes"  
Elite = as.factor(Elite)  
college = data.frame(college, Elite)  
summary(Elite)

## No Yes   
## 699 78

plot(college$Elite, college$Outstate,   
 xlab = "Elite",   
 ylab ="Out of State tuition (USD)",   
 main = "Outstate Tuition")



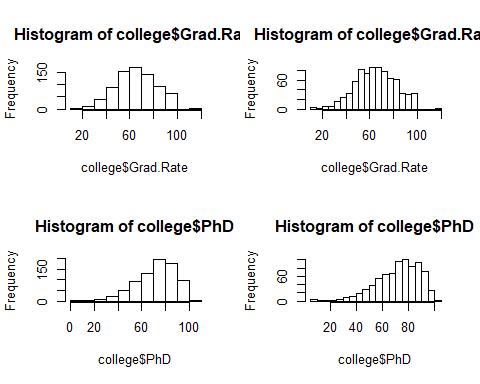
# v)  
par(mfrow=c(2,2))  
hist(college$Grad.Rate)  
hist(college$Grad.Rate, breaks = 20)  
summary(college$Grad.Rate) # There is a college with a 118% graduation rate...

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 10.00 53.00 65.00 65.46 78.00 118.00

row.names(college[nrow(college[college$Grad.Rate==118,]),])

## [1] "Abilene Christian University"

hist(college$PhD)  
hist(college$PhD, breaks = 30)



summary(college$PhD) # There is a college with a 103% of workers with a PhD...

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 8.00 62.00 75.00 72.66 85.00 103.00

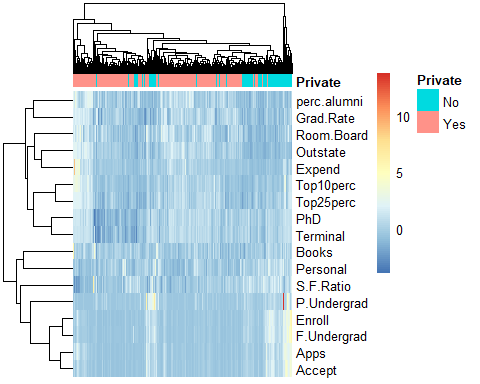
row.names(college[nrow(college[college$PhD==103,]),])

## [1] "Abilene Christian University"

# We can see that same data for this college is incorrect.  
  
# vi)  
par(mfrow=c(4,5))  
name <- colnames(college)  
#for (i in c(1:ncol(college))) {  
# hist(as.numeric(college[,i]), xlab = name[i], main = NULL)  
#}  
  
# found online, heatmap of overall variables  
#install.packages("pheatmap")  
library(pheatmap)

## Warning: package 'pheatmap' was built under R version 3.5.3

pheatmap(t(as.matrix(scale(College[, 2:18]))),  
 annotation=College[1],  
 show\_colnames=FALSE)



# As epected from the histograms we can see that the application counts,  
# acceptances, enrollments all show exponential trends. We can see that  
# the top 25% absorbs the top 10% trend. Out of state tuition, room board,  
# student-faculty ratio, graduation rate seem normal. Books, personal,  
# perc. alumni, expend are skewed right. PhD and terminal are skewed left.  
# F&P undergrad's heavy skew is expected.  
# The heatmap is standardized and transposed for convenience, meaning  
# the vertical bars correspond to every one college and the horizontals are  
# the 19 variables each college information is stored. The general behavior  
# I see is the general spread out of numbers (the blue color) but there  
# seems to be a couple exceptions. There is one college with a lot of books;  
# one with particularly high expenditure; one with a lot of part time  
# undergrads and applications and acceptances.