Chapter 08 데이터 시각화

신봉균 20191624

2023-04-19

## 코드 8-1

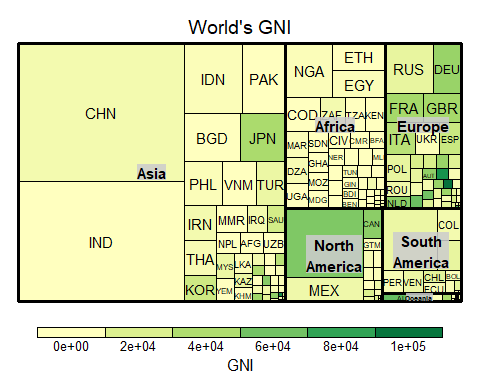
library(treemap) #treemap 패키지 불러오기

## Warning: 패키지 'treemap'는 R 버전 4.2.3에서 작성되었습니다

data(GNI2014) #데이터 불러오기기  
head(GNI2014)

## iso3 country continent population GNI  
## 3 BMU Bermuda North America 67837 106140  
## 4 NOR Norway Europe 4676305 103630  
## 5 QAT Qatar Asia 833285 92200  
## 6 CHE Switzerland Europe 7604467 88120  
## 7 MAC Macao SAR, China Asia 559846 76270  
## 8 LUX Luxembourg Europe 491775 75990

treemap(GNI2014,  
 index=c('continent','iso3'), #계층구조 설정(대륙->국가)  
 vSize='population', #타일의 크기  
 type = 'value', # 타일 컬러링 방법법  
 vColor= 'GNI', #타일의 컬러  
 title= 'World\'s GNI') #트리맵 제목

 **library(treemap)**으로 트리맵을 그리는 데 필요한 **treemap** 패키지를 불러온 후 **data(GNI2014)**로 실습에 필요한 **GNI2014** 데이터를 불러온다.

타일의 면적은 인구수와 비례한다.

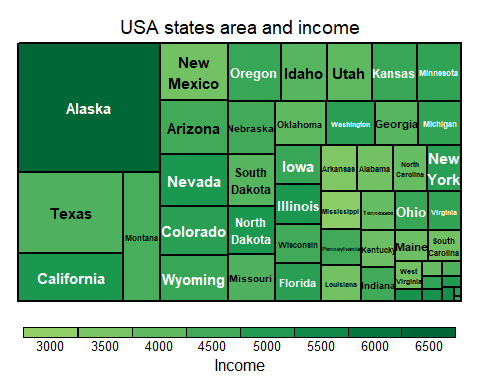
타일의 색은 GNI를 의미한다. 소득이 높을 수록 진한 초록색에 가깝고 낮을 수록 노란색에 가깝다.

타일들의 굵은 테두리선에 의해 대륙으로 묶여져 있다.

타일을 보면 Asiz의 인구가 타 대륙에 비해 압도적으로 많은 것을 알 수 있다.

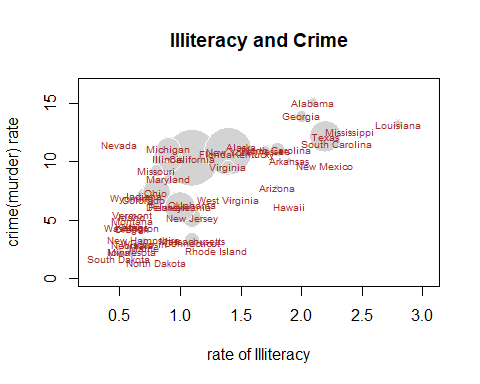
## 코드 8-2

library(treemap) #treemap 패키지 불러오기  
st= data.frame(state.x77) #매트릭스를 데이터프레임으로 변환  
st= data.frame(st, stname= rownames(st)) #주 이름 열 stname 추가  
  
treemap(st,  
 index = c('stname'), #타일에 주 이름 표기  
 vSize= 'Area', #타일의 크기  
 vColor = 'Income', #타일의 컬러  
 type= 'value', #타일 컬러링 방법  
 title = 'USA states area and income') # 트리맵의 제목



## 코드 8-3

st= data.frame(state.x77) #매트릭스를 데이터프레임으로 변환  
symbols(st$Illiteracy, st$Murder, #원의 x,y 좌표의 열  
 circles= st$Population, #원의 반지름의 열  
 inches = 0.3, #원의 크기 조절값  
 fg= 'white', #원의 테두리 색  
 bg= 'lightgray', #원의 바탕색  
 lwd= 1.5, #원의 테두리선 두께  
 xlab='rate of Illiteracy',  
 ylab='crime(murder) rate',  
 main= 'Illiteracy and Crime')  
  
text(st$Illiteracy, st$Murder, #텍스트가 출력될 x,y좌표  
 rownames(st), #출력할 텍스트  
 cex=0.6, #폰트 크기  
 col='brown') #폰트 컬러

 버블차트는 **symbols()** 와 **text()**의 조합으로 생성된다.

**symbols()** 는 2차원 좌표 상에 자료값을 원으로 표시하는 기능을 한다.

**text()**는 원 위에 텍스트, 여기서는 주의 이름을 표시하는 기능을 한다.

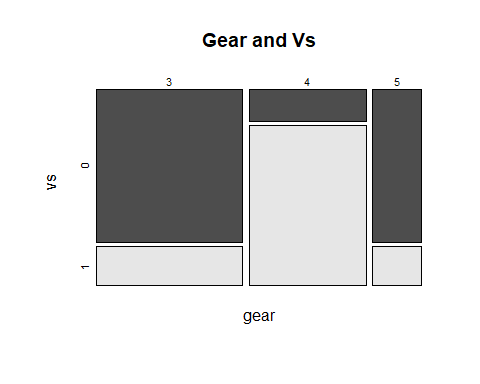
분석: 원의 위치를 관찰해보면 전반적으로 문맹률이 높아질수록 범죄율이 증가하는 추세를 확인할 수 있다.또한, 원의 크기로 보아 인구수가 많을 수록 범죄율이 증가하는 것 또한 확인할 수 있다.

## 코드 8-4

head(mtcars)

## mpg cyl disp hp drat wt qsec vs am gear carb  
## Mazda RX4 21.0 6 160 110 3.90 2.620 16.46 0 1 4 4  
## Mazda RX4 Wag 21.0 6 160 110 3.90 2.875 17.02 0 1 4 4  
## Datsun 710 22.8 4 108 93 3.85 2.320 18.61 1 1 4 1  
## Hornet 4 Drive 21.4 6 258 110 3.08 3.215 19.44 1 0 3 1  
## Hornet Sportabout 18.7 8 360 175 3.15 3.440 17.02 0 0 3 2  
## Valiant 18.1 6 225 105 2.76 3.460 20.22 1 0 3 1

mosaicplot(~gear+vs, data= mtcars, color=TRUE,  
 main='Gear and Vs')



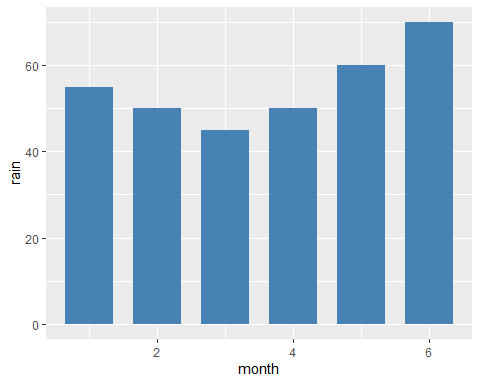
**mtcars** 데이터셋은 **32 by 11** 임으로 11개의 **feature**을 가지고 있다. 모자이프 플롯을 작성하는 함수 **mosaicplot()**을 이용하여 그림을 작성하였다.

## 코드 8-5

library(ggplot2)  
month = c(1,2,3,4,5,6)  
rain = c(55,50,45,50,60,70)  
df= data.frame(month, rain)  
df

## month rain  
## 1 1 55  
## 2 2 50  
## 3 3 45  
## 4 4 50  
## 5 5 60  
## 6 6 70

ggplot(df, aes(x= month, y=rain)) + geom\_bar(stat='identity', #막대의 높이는 y축의 열의 값  
 width = 0.7, #막대의 폭 지정  
 fill='steelblue') # 막대의 색 지정정



**library(ggplot2)** 을 불러온다.

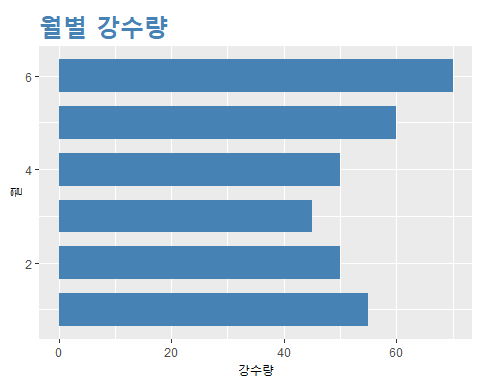
**aes(x=month, y=rain)** : aes()는 그래프를 그리기 위한 x축, y축의 열을 지정한다.

**geom\_bar()**함수에 **stat=‘identity’**는 막대의 높이는 y축에 해당하는 열의 값으로 지정

**width = 0.7**은 막대의 폭 지정 **fill=‘steelblue’**은 막대의 색갈 지정

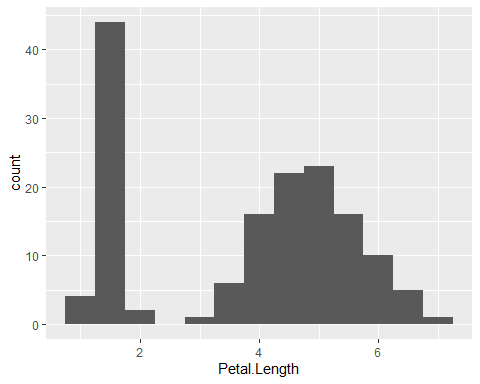
## 코드 8-6

ggplot(df, aes(x = month, y = rain)) +  
 geom\_bar(stat = "identity", width = 0.7, fill = "steelblue") +  
 ggtitle("월별 강수량") +  
 theme(plot.title = element\_text(size = 25, face = "bold", colour = "steelblue")) +  
 xlab("월") + ylab("강수량") +  
 coord\_flip()

 **theme(plot.title = element\_text(size = 25, face = “bold”, colour = “steelblue”))** 지정된 그래프에 대한 제목의 폰트 크기, 색 등을 지정한다. 이 경우 폰트 크기는 25, 볼드 처리, 폰트 컬러는 강청색으로 지정했다.

## 코드 8-7

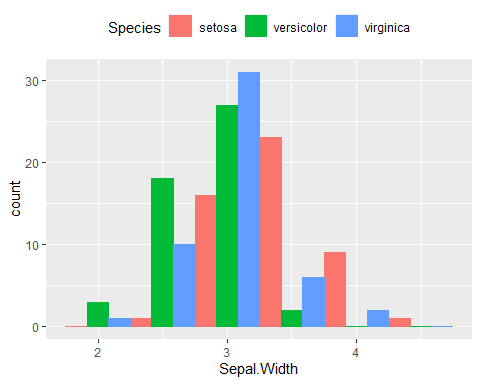
library(ggplot2)  
  
ggplot(iris, aes(x=Petal.Length)) + #그래프를 그릴 데이터 가정  
 geom\_histogram(binwidth = 0.5) # 히스토그램 작성



**ggplot()**에 그래프 작성대상 데이터는 iris로 설정 **Petal.Length**열을 **aes()**의 x매개변수에 값으로 준다.

## 코드 8-8

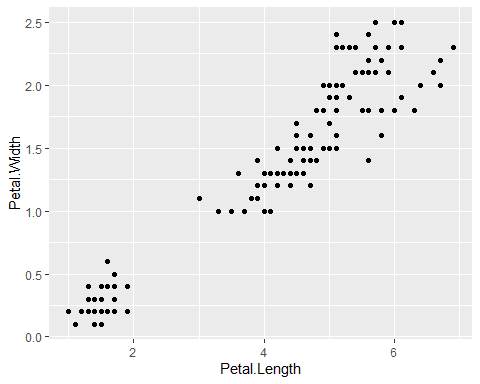
library(ggplot2)  
  
ggplot(iris, aes(x= Sepal.Width, fill= Species, color=Species)) +  
 geom\_histogram(binwidth = 0.5, position='dodge') +  
 theme(legend.position = 'top')



**position= ‘dodge’** 이 히스토그램은 3개의 품종의 히스토그램이 하나의 그래프에 작성된다. 동일 구간에 대해 3개의 막대가 그려진다. position은 동일 구간의 막대들을 어떻게 그럴지를 지정하는데, ’dodge’는 막대들을 겹치지 않고 병렬로 그리도록 지정하는 것이다.

## 코드 8-9

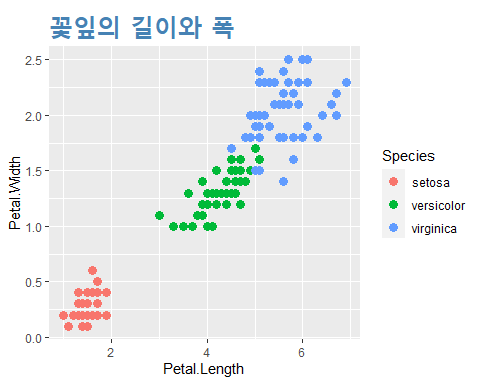
library(ggplot2)  
  
ggplot(data= iris, aes(x=Petal.Length, y=Petal.Width)) +  
 geom\_point()



**geom\_point()** 함수는 산점도를 작성하는 함수이다.

## 코드 8-10

library(ggplot2)  
  
ggplot(data= iris, aes(x=Petal.Length, y=Petal.Width,  
 color= Species)) +  
 geom\_point(size= 3) +  
 ggtitle('꽃잎의 길이와 폭')+ #그래프의 제목 지정  
 theme(plot.title = element\_text(size=25, face = 'bold', colour = 'steelblue'))

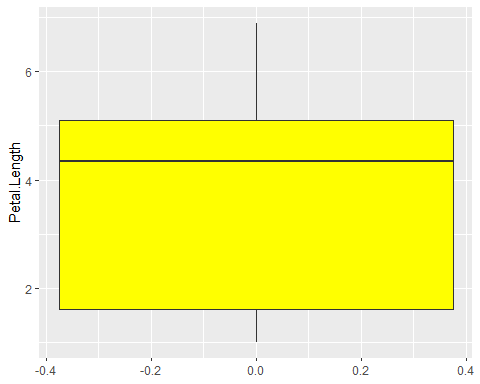


aes 안에 colore= Species로 점의 색을 품종에 따라 다르게 설정

size=3는 점의 크기를 설정하는 매개변수이다.

## 코드 8-11

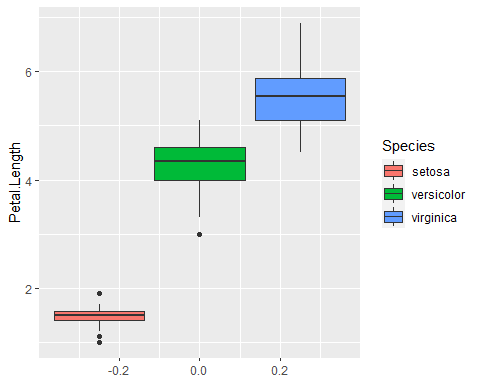
library(ggplot2)  
  
ggplot(data= iris, aes(y=Petal.Length)) +  
 geom\_boxplot(fill='yellow')



상자그림을 그리는 함수의 이름은 **geom\_boxplot()**이다.

## 코드 8-12

library(ggplot2)  
  
ggplot(data= iris, aes(y=Petal.Length, fill=Species)) +  
 geom\_boxplot()



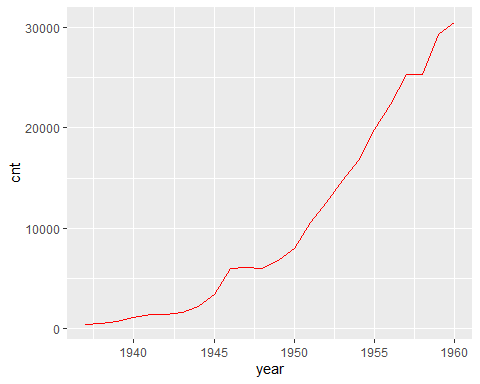
종별 꽃잎의 길이로 boxplot을 작성하였다.

## 코드 8-13

library(ggplot2)  
  
year= 1937:1960  
cnt= as.vector(airmiles)  
df= data.frame(year, cnt) #데이터 준비  
head(df) #데이터 확인

## year cnt  
## 1 1937 412  
## 2 1938 480  
## 3 1939 683  
## 4 1940 1052  
## 5 1941 1385  
## 6 1942 1418

ggplot(data= df, aes(x= year, y=cnt)) + #선그래프 작성  
 geom\_line(col='red')



선그래프를 작성하는 함수는 **geom\_line()**이다.

## 코드 8-14

library(Rtsne)

## Warning: 패키지 'Rtsne'는 R 버전 4.2.3에서 작성되었습니다

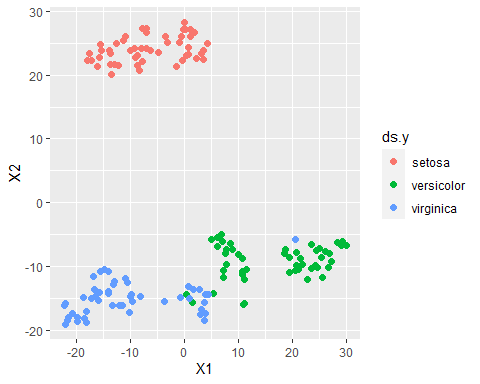
library(ggplot2)  
ds= iris[,-5] #품종 정보 제외  
  
##중복 데이터 제거  
  
dup= which(duplicated(ds))  
dup #143번째 행 중복

## [1] 143

ds= ds[-dup,]  
ds.y = iris$Species[-dup] #중복을 제외한 품종 정보  
  
## t-SNE 실행  
tsne = Rtsne(ds, dim=2, perplexity=10)  
  
##축소 결과 시각화  
  
df.tsne = data.frame(tsne$Y)  
head(df.tsne)

## X1 X2  
## 1 -8.6108195 23.11012  
## 2 1.1292051 27.18890  
## 3 0.4508762 23.04645  
## 4 0.7283821 24.27282  
## 5 -8.7641922 21.46322  
## 6 -15.2630139 23.95376

ggplot(df.tsne, aes(x=X1, y=X2, color=ds.y)) +  
 geom\_point(size=2)



ds는 차원 축소 대상 데이터셋이다.

dim=2는 ds를 몇 차원으로 축소할지 지정하는데 2또는 3이 일반적이다.

perplexity=10은 차원 축소 과정에서 데이터를 샘플링하는데 샘플의 개수를 몇 개로 할지 지정한다.(대상 데이터의 행의 수)/3 보다 작게 지정한다.

## 코드 8-15

library('car')

## Warning: 패키지 'car'는 R 버전 4.2.3에서 작성되었습니다

## 필요한 패키지를 로딩중입니다: carData

## Warning: 패키지 'carData'는 R 버전 4.2.3에서 작성되었습니다

library('rgl')

## Warning: 패키지 'rgl'는 R 버전 4.2.3에서 작성되었습니다

library('mgcv')

## 필요한 패키지를 로딩중입니다: nlme

## This is mgcv 1.8-41. For overview type 'help("mgcv-package")'.

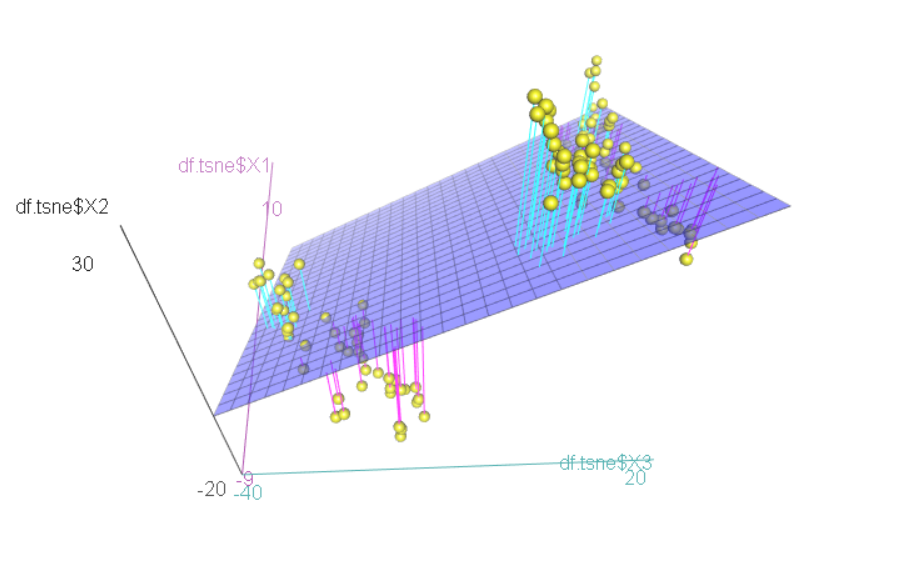
library('Rtsne')  
  
tsne = Rtsne(ds, dims=3, perplexity=10)  
df.tsne = data.frame(tsne$Y)  
head(df.tsne)

## X1 X2 X3  
## 1 -17.54991 -2.320246 21.08685  
## 2 -27.83605 -2.342905 19.22657  
## 3 -28.03360 -2.456148 23.30532  
## 4 -28.48170 -1.033006 22.38860  
## 5 -17.35360 -2.732398 22.68071  
## 6 -10.15114 -1.530328 20.19814

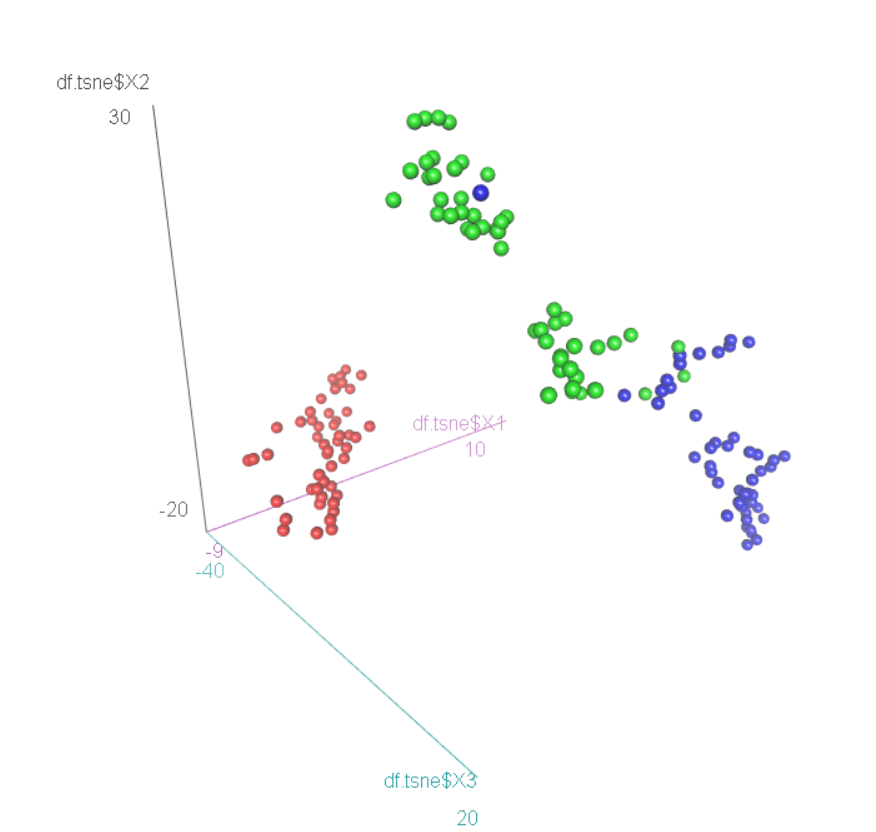
#회귀면이 포함된 3차원 산점도  
  
scatter3d(x=df.tsne$X1, y=df.tsne$X2, z=df.tsne$X3)

## 필요한 네임스페이스를 로딩합니다: MASS

#회귀면이 없는 3차원 산점도  
  
points= as.integer(ds.y)  
color= c('red','green','blue')  
scatter3d(x=df.tsne$X1, y=df.tsne$X2, z=df.tsne$X3,  
 point.col = color[points], #점의 색을 품종별로 다르게  
 surface = FALSE) #회귀면을 표시하지 않음



회귀면이 있는 3차원 산점도 이다.



회귀면이 없는 3차원 산점도 이다.