제8강
데이터 시각화

Section 01

데이터 시각화 기법

1. 데이터 시각화 기법

1. 데이터 시각화의 중요성

- 데이터 시각화(**data visualization**) : 숫자 형태의 데이터를 그래프나 그림 등의 형태로 표현하는 과정
- 데이터 분석 과정에서 중요한 기술 중의 하나
- 데이터를 시각화 하면 데이터가 담고 있는 정보나 의미를 보다 쉽게 파악
- 경우에 따라서는 시각화 결과로부터 중요한 영감을 얻기도 함
- 독립적인 교육과정이 따로 존재할 만큼 중요

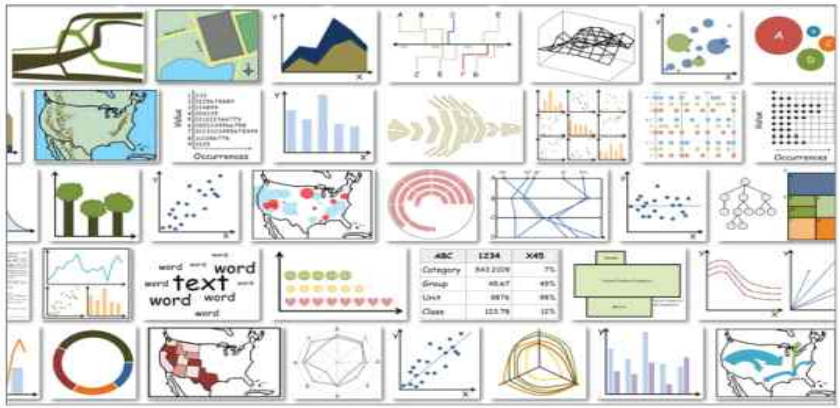


그림 8–1 다양한 데이터 시각화 사례들

1. 데이터 시각화 기법

2. 트리맵

2.1 GNI2014 데이터셋으로 트리맵 작성하기

- 사각타일의 형태로 구성되어 있으며, 각 타일의 크기와 색깔로 데이터의 크기를 나타냄
- 각각의 타일은 계층 구조가 있기 때문에 데이터에 존재하는 계층 구조도 표현
- **treemap** 패키지 설치 필요
- 예제 데이터셋 : **treemap** 패키지 안에 포함된 **GNI2014. 2014**년도의 전 세계 국가별 인구, 국민총소득**(GNI)**, 소속 대륙의 정보를 담고 있음

코드 8-1

```
library(treemap)                          # treemap 패키지 불러오기
data(GNI2014)                             # 데이터 불러오기
head(GNI2014)                             # 데이터 내용보기
treemap(GNI2014,
        index=c("continent","iso3"),      # 계층구조 설정(대륙-국가)
        vSize="population",               # 타일의 크기
        vColor="GNI",                     # 타일의 컬러
        type="value",                     # 타일 컬러링 방법
        bg.labels="yellow",               # 레이블의 배경색
        title="World's GNI")              # 트리맵 제목
```

1. 데이터 시각화 기법

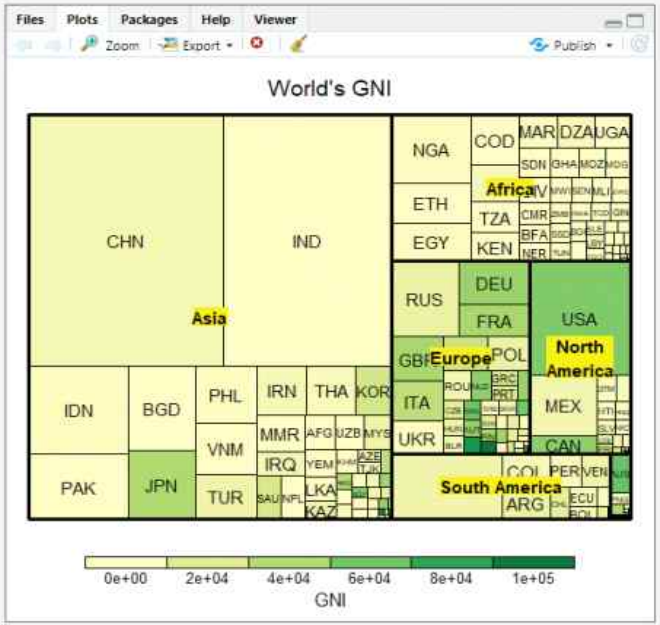
```
> library(treemap)                    # treemap 패키지 불러오기
> data(GNI2014)                       # 데이터 불러오기
> head(GNI2014)                       # 데이터 내용보기
  iso3         country     continent population    GNI
3  BMU         Bermuda North America      67837 106140
4  NOR          Norway        Europe    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
```

열의 이름	의미
iso3	국가를 식별하는 표준코드
country	국가명
continent	국가가 속한 대륙명
population	국가의 인구
GNI	국가의 국민총소득

표 8-1 GNI2014의 각 열의 의미

1. 데이터 시각화 기법

```
> treemap(GNI2014,
+         index=c("continent","iso3"),   # 계층구조 설정(대륙-국가)
+         vSize="population",            # 타일의 크기
+         vColor="GNI",                  # 타일의 컬러
+         type="value",                  # 타일 컬러링 방법
+         bg.labels="yellow",            # 레이블의 배경색
+         title="World's GNI")           # 트리맵 제목
>
```



- 타일의 면적은 인구수와 비례
- 타일의 색은 **GNI**를 의미함
- 소득이 높을수록 진한 초록색에 가까움
- 소득이 낮을수록 노랑색에 가깝다.
- 그림으로 보면 아시아의 인구가 매우 많다.(중국과 인도)
- 소득은 역시 북아메리카, 유럽 등이다.

1. 데이터 시각화 기법

- `GNI2014`
  트리맵을 그릴 대상의 데이터셋이다. 데이터프레임 형태여야 한다.
- `index=c("continent","iso3")`
  트리맵 상에서 타일들이 대륙(continent) 안에 국가(iso3)의 형태로 배치되는 것을 지정한다.
- `vSize="population"`
  타일의 크기를 결정하는 열을 지정하며, 여기서는 인구수(population)로 지정하였다.
- `vColor="GNI"`
  타일의 색을 결정하는 열을 지정하는데, 여기서는 소득(GNI)으로 지정하였다.
- `type="value"`
  타일의 컬러링 방법을 지정하는 것으로 "value"는 vColor에서 지정한 열에 저장된 값의 크기에 의해 색이 결정됨을 의미한다. "value" 외에도 "index", "comp", "dens" 등을 지정할 수 있다.
- `bg.labels="yellow"`
  대륙을 나타내는 레이블의 배경색을 지정한다.
- `title="World's GNI"`
  트리맵의 제목을 지정한다.

1. 데이터 시각화 기법

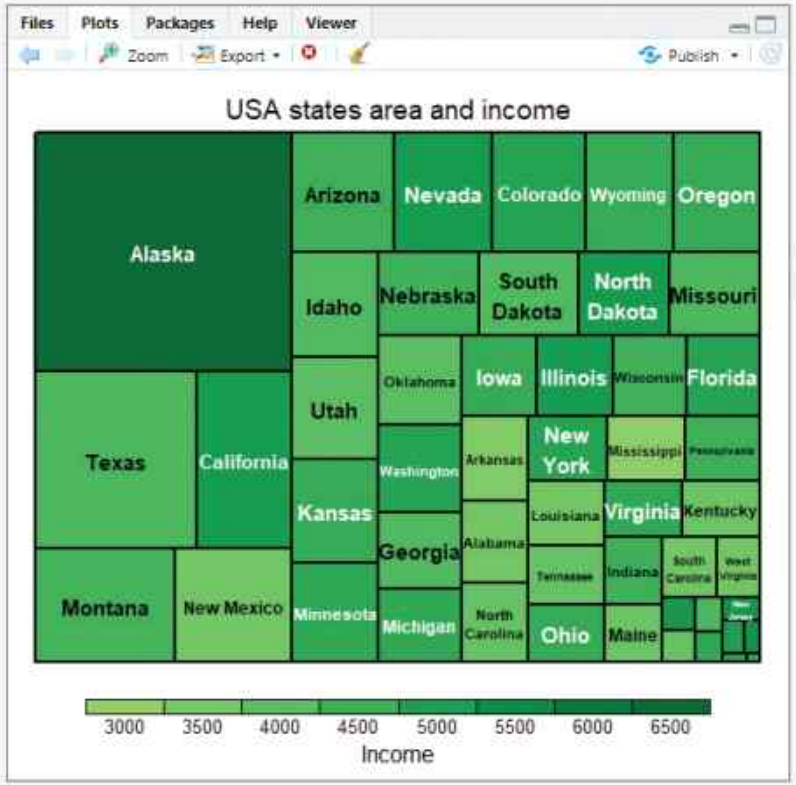
2.2 state.x77 데이터셋으로 트리맵 작성하기

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 stname을 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
```

1. 데이터 시각화 기법



타일의 면적은 주의 면적
타일의 색은 주의 소득
알래스카가 면적도 크고 소득도 높다.

1. 데이터 시각화 기법

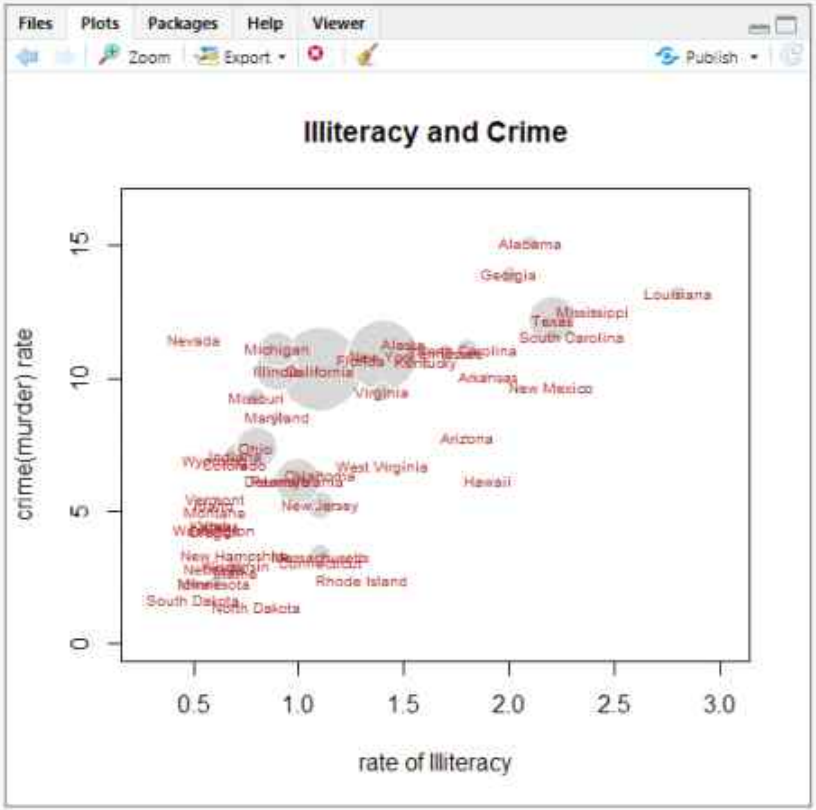
3. 버블차트

- 버블 차트(**bubble chart**): 앞에서 배운 산점도 위에 버블의 크기로 정보를 표시하는 시각화 방법
- 산점도가 **2**개의 변수에 의한 위치 정보를 표시한다면, 버블 차트는 **3**개의 변수 정보를 하나의 그래프에 표시

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, y 좌표
        rownames(st),                      # 출력할 텍스트
        cex=0.6,                           # 폰트 크기
        col="brown")                       # 폰트 컬러
```

1. 데이터 시각화 기법



st$Illiteracy(문맹률), st$Murder(범죄율)

전반적으로 문맹률이 높아질수록 범죄율이 증가하는 추세

인구수가 많은 주가 대체로 범죄율도 높은 것을 확인

범죄율이 가장 낮은 주는 North Dakota

1. 데이터 시각화 기법

- `st$Illiteracy, st$Murder`
  2차원 좌표의 x축과 y축을 나타낼 열을 지정한다(여기서 x축은 문맹률, y축은 범죄율(살인율)). x축의 값과 y축의 값이 만나는 지점에 원이 그려진다.
- `circles=st$Population`
  원의 크기(반지름)를 결정할 열을 지정한다(여기서는 인구수).
- `inches=0.3`
  원의 크기를 조절하는 매개변수로, 매개변수값이 클수록 원이 크게 그려진다.
- `fg="white"`
  원의 테두리선 색을 지정한다.
- `bg="lightgray"`
  원의 바탕색을 지정한다.
- `lwd=1.5`
  원의 테두리선 두께를 지정한다.
- `xlab="rate of Illiteracy"`
  x축의 레이블을 지정한다.
- `ylab="crime(murder) rate"`
  y축의 레이블을 지정한다.
- `main="Illiteracy and Crime"`
  그래프의 제목을 지정한다.

1. 데이터 시각화 기법

- `st$Illiteracy, st$Murder`
  텍스트를 표시할 위치에 대한 x축과 y축 좌표값을 나타내는데, symbols() 함수에 있는 원의 x축과 y축 좌표값과 일치시킨다.
- `rownames(st)`
  표시할 텍스트를 지정한다. st의 행 이름은 미국 각 주의 이름이다.
- `cex=0.6`
  텍스트의 크기를 지정한다.
- `col="brown"`
  텍스트의 색을 지정한다.

1. 데이터 시각화 기법

4. 모자이크 플롯

- 모자이크 플롯(**mosic plot**): 다중변수 범주형 데이터에 대해 각 변수의 그룹별 비율을 면적으로 표시하여 정보를 전달

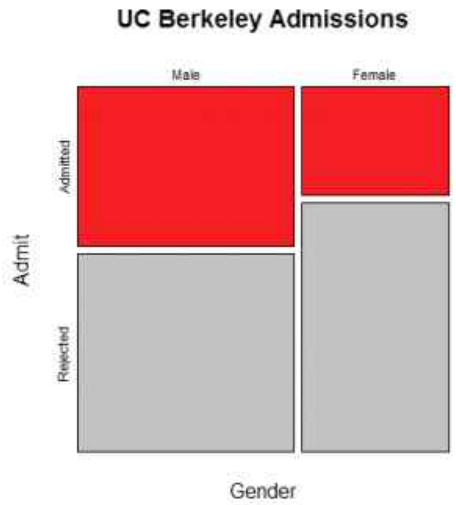
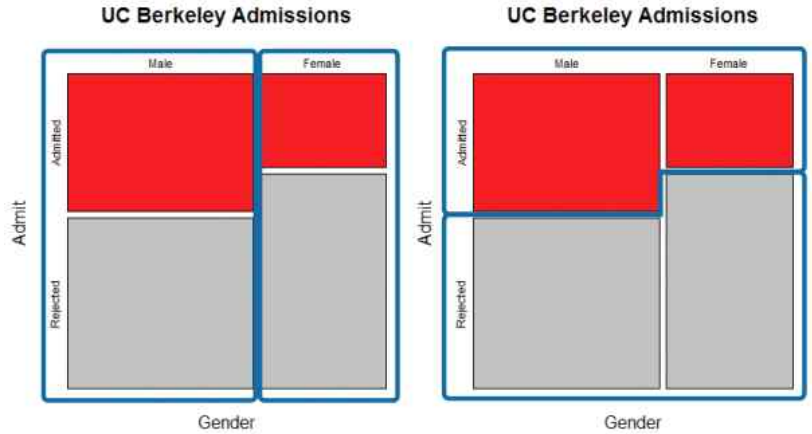


그림 8–2 모자이크 플롯의 예

예제 데이터: UCBAdmissions
미국의 버클리대학교 대학원의 지원자와 합격자 통계를 성별, 학과별로 정리
아래는 지원자와 합격자 통계를 성별로 구분하여 모자이크 플롯으로 나타낸 것

1. 데이터 시각화 기법



(a) 남성 지원자와 여성 지원자의 비율 (b) 합격자와 불합격자의 비율

그림 8-3 모자이크 플롯의 해석 1

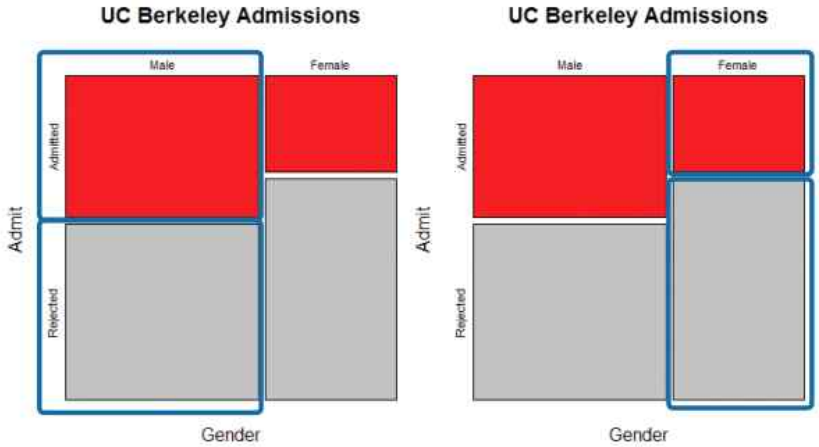
왼쪽의 전체 면적이 남성(male) 지원자의 수를 나타내고, 오른쪽의 전체 면적이 여성(female) 지원자의 수를 나타냄

남성 지원자의 수가 여성 지원자 수에 비해 1.5배 정도 많음

위쪽 빨간색 면적은 합격자의 수를, 아래쪽 회색 면적은 불합격자의 수를 나타냄

전체 지원자에서 합격자의 비율이 50%가 안 되는 것을 확인

1. 데이터 시각화 기법



(a) 남성 지원자의 합격자와 불합격자 비율 (b) 여성 지원자의 합격자와 불합격자 비율

그림 8-4 모자이크 플롯의 해석 2

남성 지원자의 합격자 비율과 불합격자 비율

여성 지원자의 합격자 비율과 불합격자 비율

여성 지원자의 합격률이 남성 지원자의 합격률보다 눈에 띄게 낮음

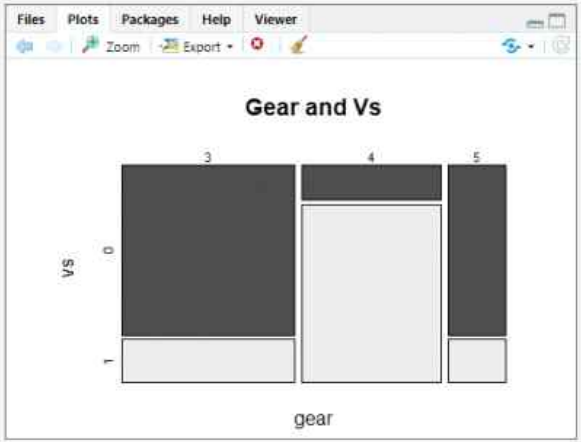
1. 데이터 시각화 기법

코드 8-4

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

```
> 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")
```

1. 데이터 시각화 기법

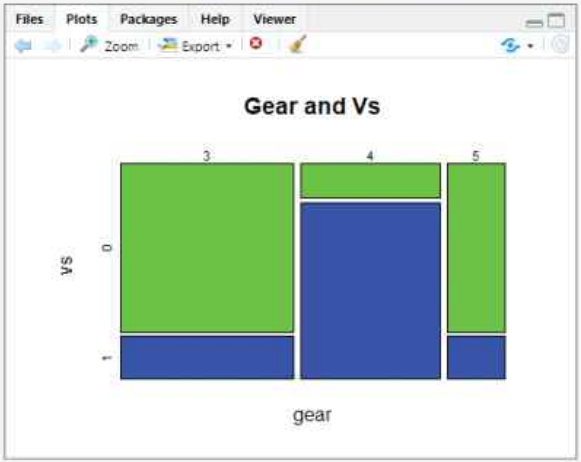


기어의 개수가 가장 많은 경우는 3개이다.
기어의 개수가 5개는 드물다.
엔진의 형태는 0, 1정도가 반반정도 된다.
기어의 개수가 홀수인 경우는 0타입이 많고.
짝수인 경우 1이 훨씬 많다.

- `~gear+vs`
  모자이크 플롯을 작성할 대상 변수를 지정한다. ~ 다음의 변수가 x축 방향으로 표시되고, + 다음의 변수가 y축 방향으로 표시된다.
- `data = mtcars`
  모자이크 플롯을 작성할 대상 데이터셋을 지정한다.
- `color=TRUE`
  y축 변수의 그룹별로 음영을 달리하여 표시한다.
- `main ="Gear and Vs"`
  모자이크 플롯의 제목을 지정한다.

1. 데이터 시각화 기법

```
> mosaicplot(~gear+vs, data = mtcars, color=c("green","blue"),
+            main ="Gear and Vs")
```



감사합니다.