

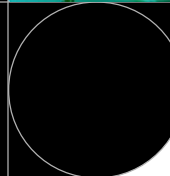
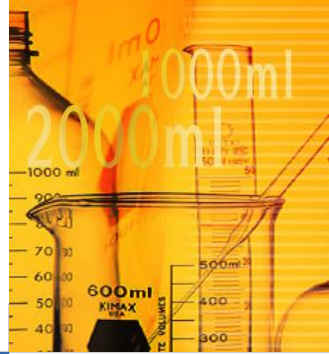
# Machine learning

## Chapter 11

# Data visualization

Sejong Oh

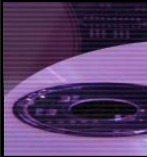
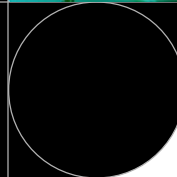
Bio Information Technology Lab.



# Contents

- PCA
- Tree map
- Mosaic plot
- 지도상에 데이터 표현

PCA



- PCA (Principal Component Analysis)
  - 주성분 분석
  - $X = n \times p$  인 배열 데이터에 대해
  - $X$  의 column space 를 차원 축소하여  $n$  개의 개체를 잘 구분하여 보이도록 하는 작업
  - 인간이 인식할수 있는 차원은 3차원까지 이므로 3차원을 넘는 데이터(변수의 개수가 4개 이상)는 시각화해서 보기가 어렵다. 고차원 데이터를 2차원 또는 3차원으로 축소하여 시각화 하는 방법이 PCA

- 차원 축소의 IDEA (3차원  $\rightarrow$  2차원)
  - 3차원상의 개체를 2차원상에 투영 (projection)
  - 이때 정보의 손실이 있을수 밖에 없는데, 정보 손실을 최소화 하도록 투영한다



- R code

```
data("USArrests")
str(USArrests)
head(USArrests)
pca.US <- prcomp(USArrests, scale=TRUE)      # run PCA
pca.US
aa = predict(pca.US)
plot(aa[,1:2])
rownames(pca.US$x) <- 1:50
biplot(pca.US, scale=0, cex=0.8)
```

```
> head(USArrests)
```

	Murder	Assault	UrbanPop	Rape
Alabama	13.2	236	58	21.2
Alaska	10.0	263	48	44.5
Arizona	8.1	294	80	31.0
Arkansas	8.8	190	50	19.5
California	9.0	276	91	40.6
Colorado	7.9	204	78	38.7

```
>
```

살인

폭행

도시화율

성폭행

# PCA

```
> pca.US <- prcomp(USArrests, scale=TRUE)
```

```
> pca.US
```

각변수들이 각기 다른 단위를 갖는 경우(키, 몸무게)

Standard deviations:

```
[1] 1.5748783 0.9948694 0.5971291 0.4164494
```

값(편차)이 클수록 개체 식별에 중요한 역할을 한다는 의미

Rotation:

	PC1	PC2	PC3	PC4
Murder	-0.5358995	0.4181809	-0.3412327	0.64922780
Assault	-0.5831836	0.1879856	-0.2681484	-0.74340748
UrbanPop	-0.2781909	-0.8728062	-0.3780158	0.13387773
Rape	-0.5434321	-0.1673186	0.8177779	0.08902432

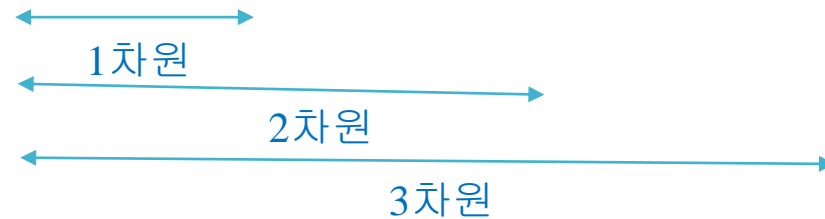
```
>
```

고유 벡터 행렬



> predict(pca.US) 차원 축소를 했을때 각 개체의 좌표값

	PC1	PC2	PC3	PC4
Alabama	-0.97566045	1.12200121	-0.43980366	0.154696581
Alaska	-1.93053788	1.06242692	2.01950027	-0.434175454
Arizona	-1.74544285	-0.73845954	0.05423025	-0.826264240
Arkansas	0.13999894	1.10854226	0.11342217	-0.180973554
California	-2.49861285	-1.52742672	0.59254100	-0.338559240
Colorado	-1.49934074	-0.97762966	1.08400162	0.001450164
Connecticut	1.34499236	-1.07798362	-0.63679250	-0.117278736
Delaware	-0.04722981	-0.32208890	-0.71141032	-0.873113315
Florida	-2.98275967	0.03883425	-0.57103206	-0.095317042
Georgia	-1.62280742	1.26608838	-0.33901818	1.065974459
Hawaii	0.90348448	-1.55467609	0.05027151	0.893733198

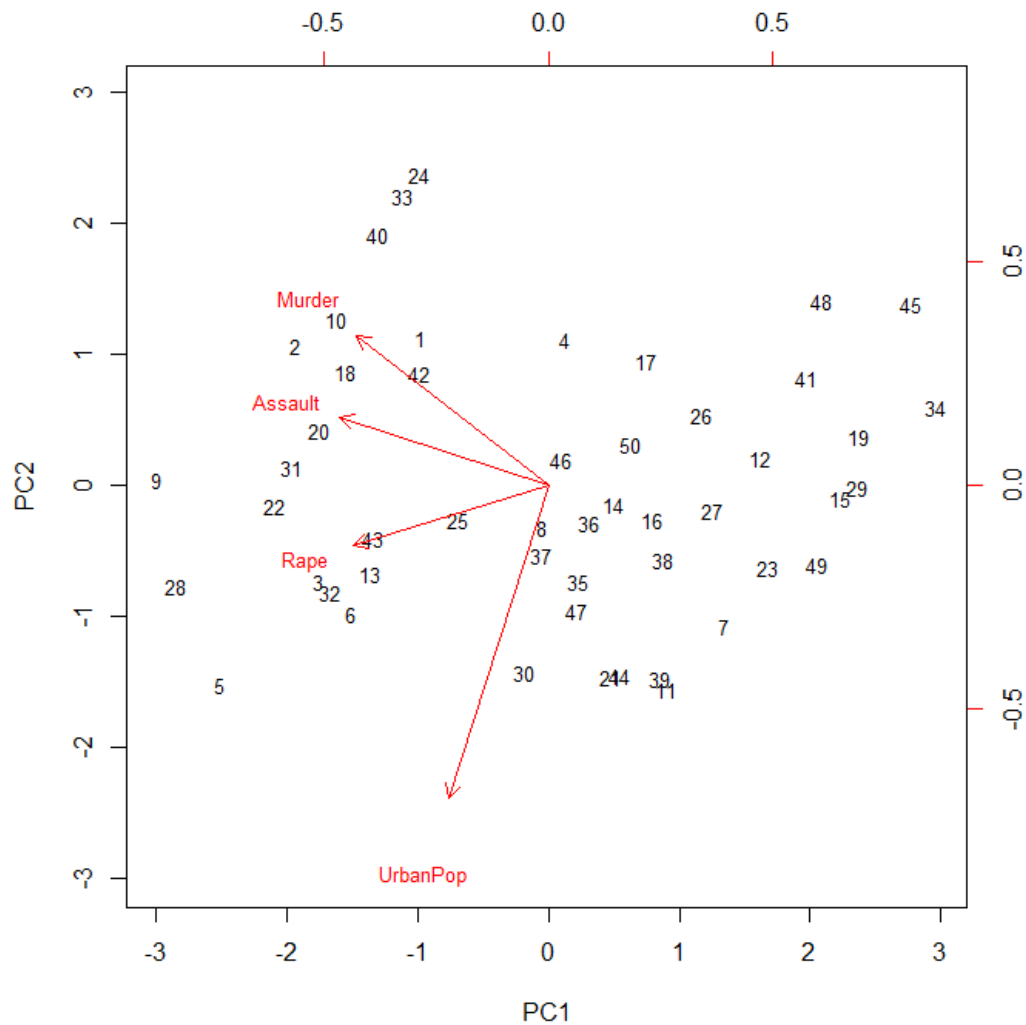


```
plot(predict(pca.US) [,1:2])
```

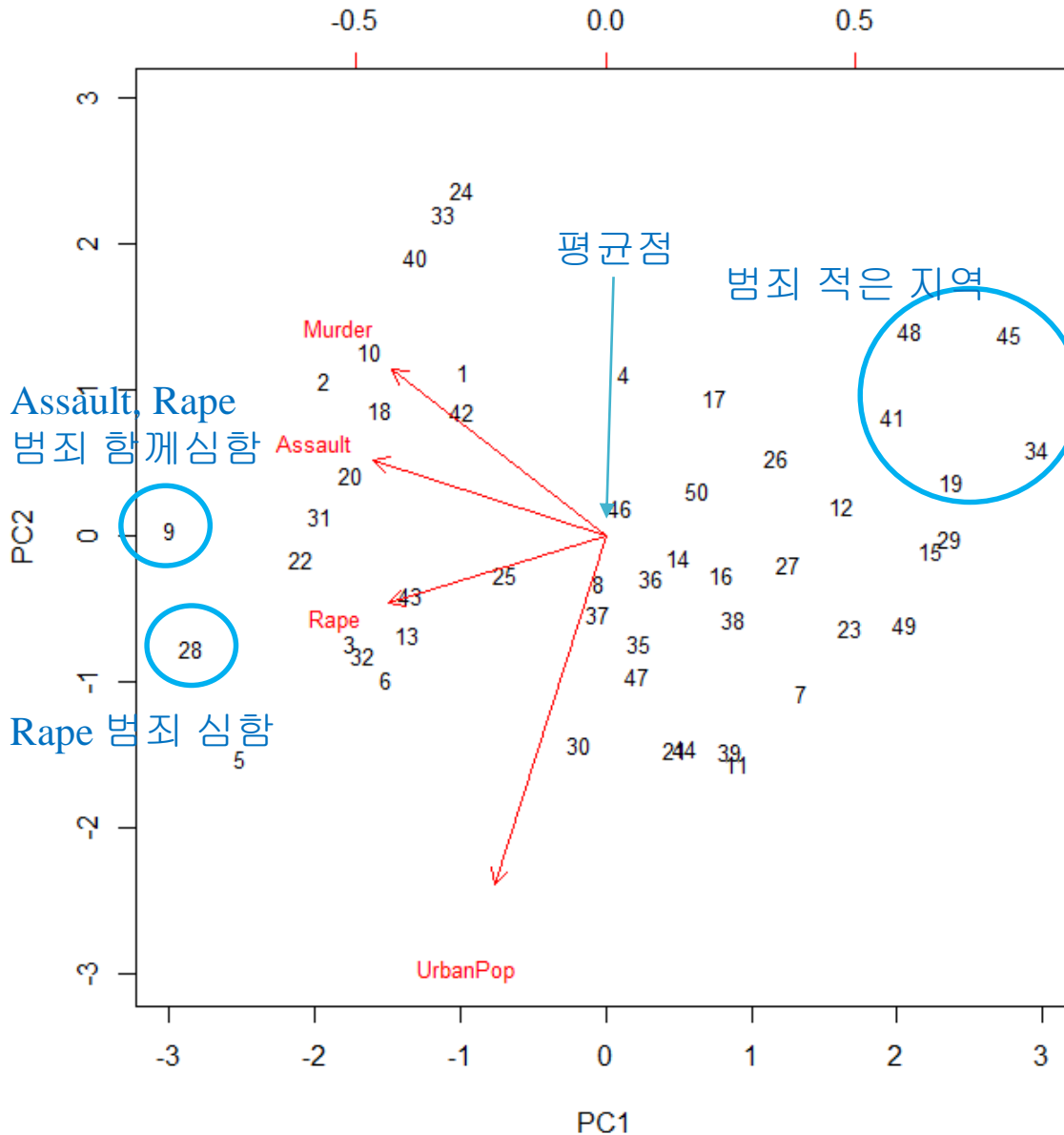
# 2차원 산점도보기

# PCA

```
> rownames(pca.US$x) <- 1:50      #주 이름을 숫자로 변경  
> biplot(pca.US, scale=0, cex=0.8)
```



# PCA

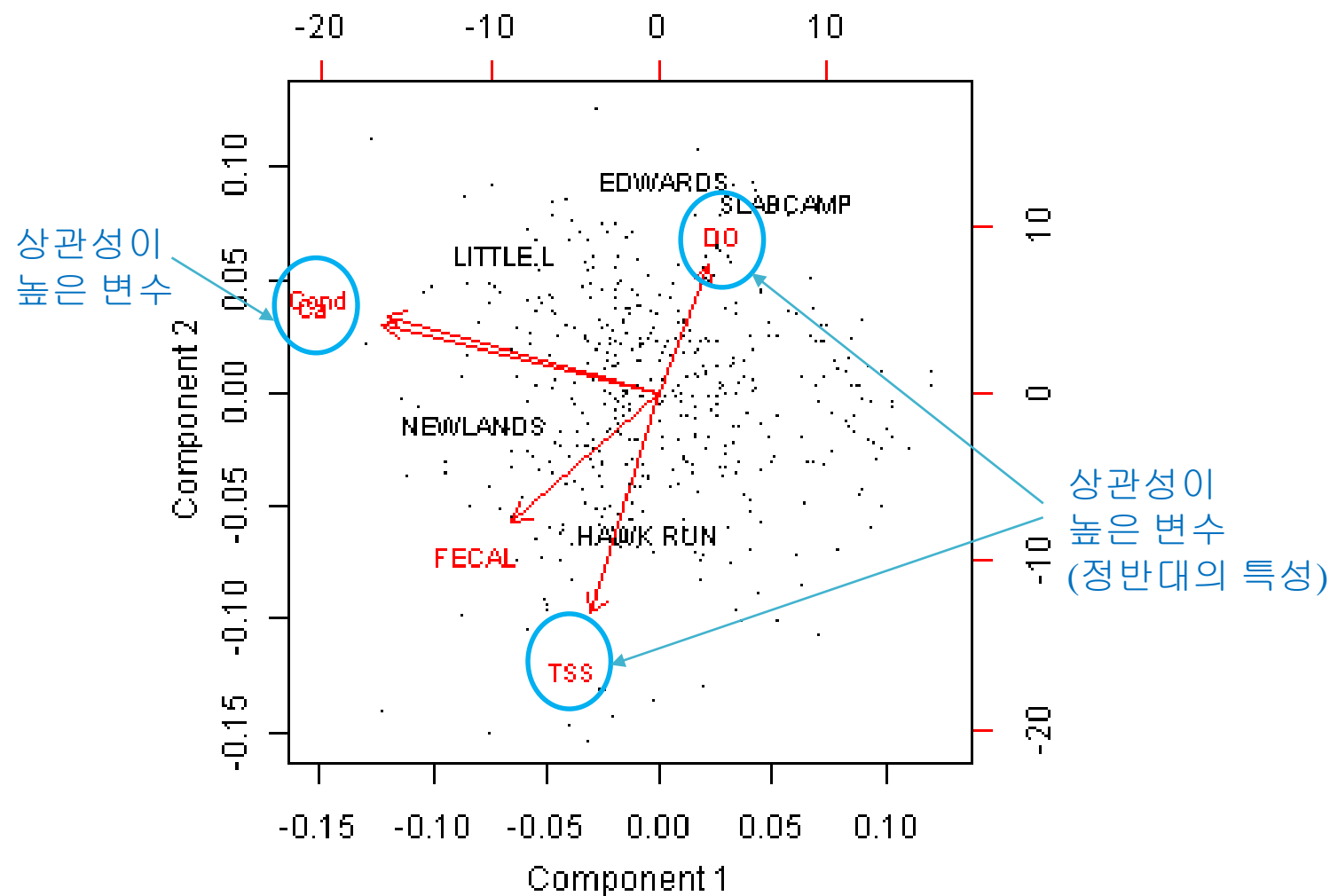


화살표는 각 변수를 나타냄

평균점을 중심으로 화살표 방향은  
평균보다 범죄가 많은 곳,  
반대방향은 평균보다 범죄가 적은 곳

Murder, Assault 화살표 사이가  
가까움 : Murder, Assault 범죄가  
함께 일어나는 경향이 있음을 의미  
(두변수 사이에 상관성 높음)

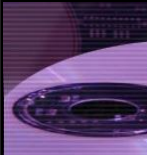
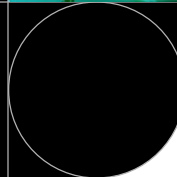
화살표 길이가 길수록 개체  
식별에 중요한 변수임을 의미



- 대상 데이터에 많은 변수가 있는 경우 PCA 로 부터 얻을 수 있는 정보는 제한적
  - 변수들을 클러스링한 후 클러스터를 대표하는 변수들만 뽑아서 PCA 를 진행
- PCA 는 MDS (multidimensional scaling)의 일종
  - PCA minimizes dimensions, preserving covariance of data.
  - MDS minimizes dimensions, preserving distance between data points.
  - They are same, if covariance in data = Euclidean distance between data points in high dimension.
  - They are different, if distance measure is different.

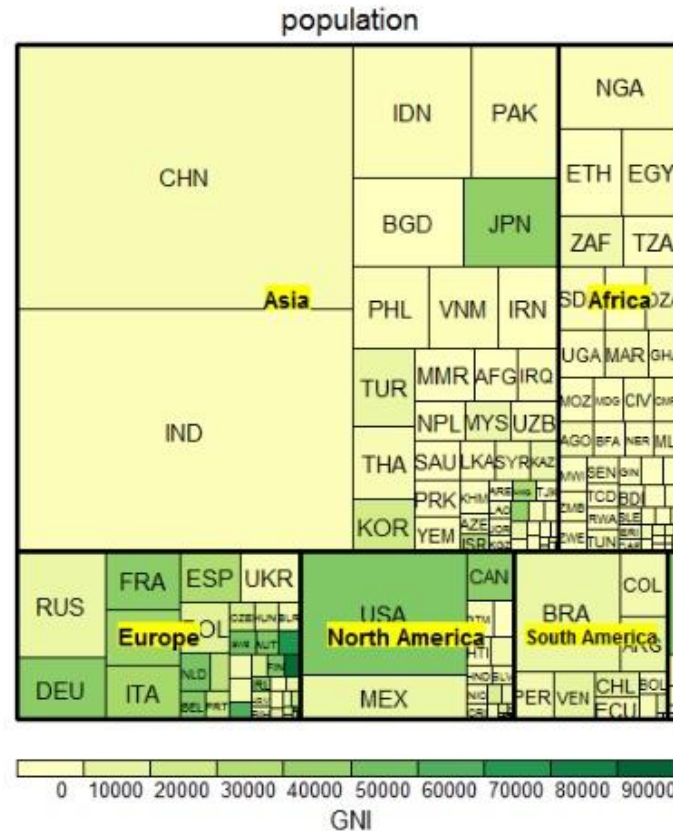
- iris dataset에 대해 PCA 분석을 하시오
  - 5번째 컬럼은 제외한다

# Tree map



# 나무지도(tree map)

- 나무지도는 데이터가 갖는 계층구조를 타일 모양으로 표현한 것
- 타일은 계층적 속성을 가지며, 계층은 컬러로 표현된다





# 나무지도(tree map)

- 설치가 필요한 패키지
  - ◉ treemap
- 실습에 사용할 데이터셋
  - ◉ GNI2014 (treemap)
  - ◉ 208개 국가의 1인당 총소득(gross national income) 데이터
  - ◉ 국가는 대륙(continent)으로 그룹핑되고 국가명은 국제표준(iso3)으로 지칭된다.
  - ◉ 국가정보는 population(인구)과 GNI(1인당 국민소득)이다

```
> 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
```

# 나무지도(tree map)

```
library (treemap)
data (GNI2014)           # 데이터 불러오기
str (GNI2014)            # 데이터 내용보기
treemap (GNI2014 ,
         index=c ("continent", "iso3") ,
         vSize="population", # 타일의 크기
         vColor="GNI",       # 타일의 컬러
         type="value",       # 타일 컬러링 방법
         bg.labels="yellow") # 레이블의 배경색
```

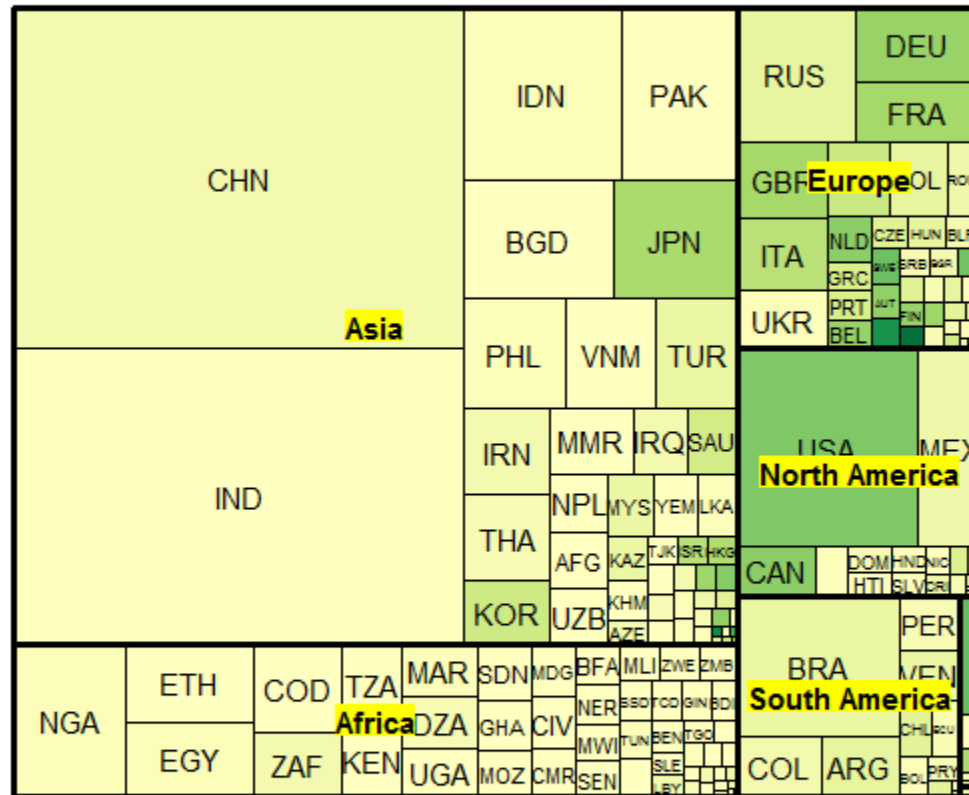
**index=c ("continent", "iso3")**

: 개체의 단위를 지정하는데 계층적 구조를 갖는 경우 상위 층을 먼저 넣는다. 대륙을 먼저 표현하고 그 안에 국가를 넣으라는 의미

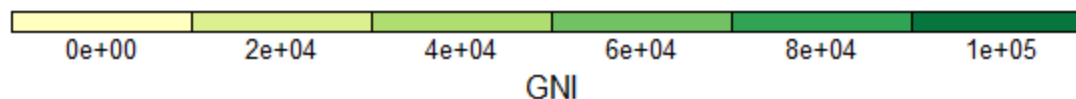
**type="value"**

: **vColor** 에서 지정한 값에 의해서 타일의 컬러가 결정됨

# population



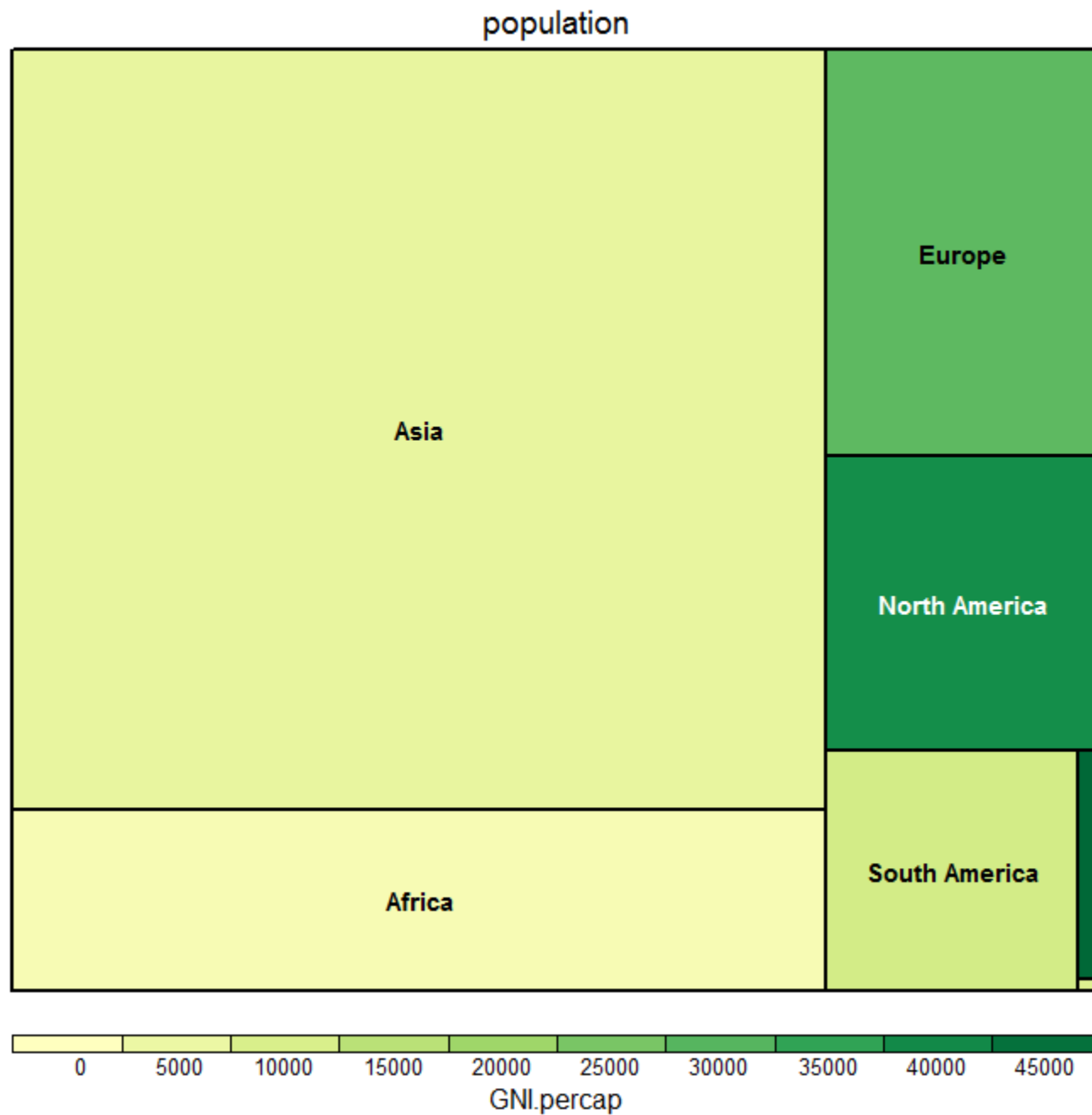
타일의 크기 : 국토면적  
타일의 색 : 국민소득



# 나무지도(tree map)

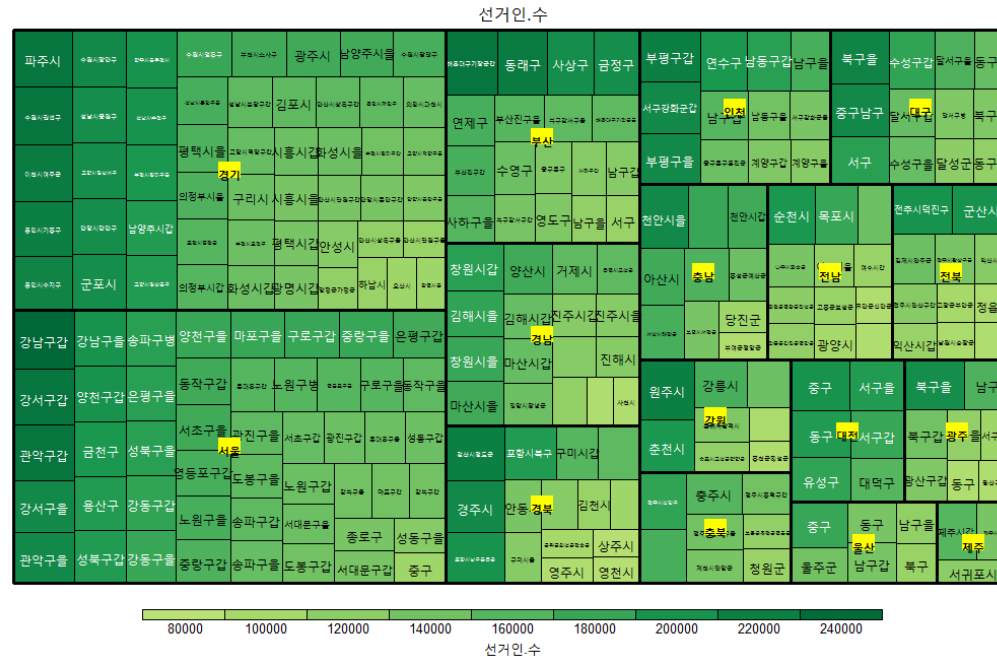
```
# 국가별 국민 총소득을 계산해서 GNI.total 컬럼에 저장
GNI2014$GNI.total <-
  GNI2014$population*GNI2014$GNI
head(GNI2014)
# 국가별 국민 총소득을 대륙별로 합계내서 GNI2014.a 에 저장
GNI2014.a <- aggregate(GNI2014[,4:6],
  by=list(GNI2014$continent),sum)
# 대륙별 합계를 대륙 인구수로 나누어 GNI.percap 컬럼에 저장
GNI2014.a$GNI.percap <-
  GNI2014.a$GNI.total/GNI2014.a$population

treemap(GNI2014.a,
  index=c("Group.1"),
  vSize="population",
  vColor="GNI.percap",
  type="value",
  bg.labels="yellow")
```



# [연습 1]

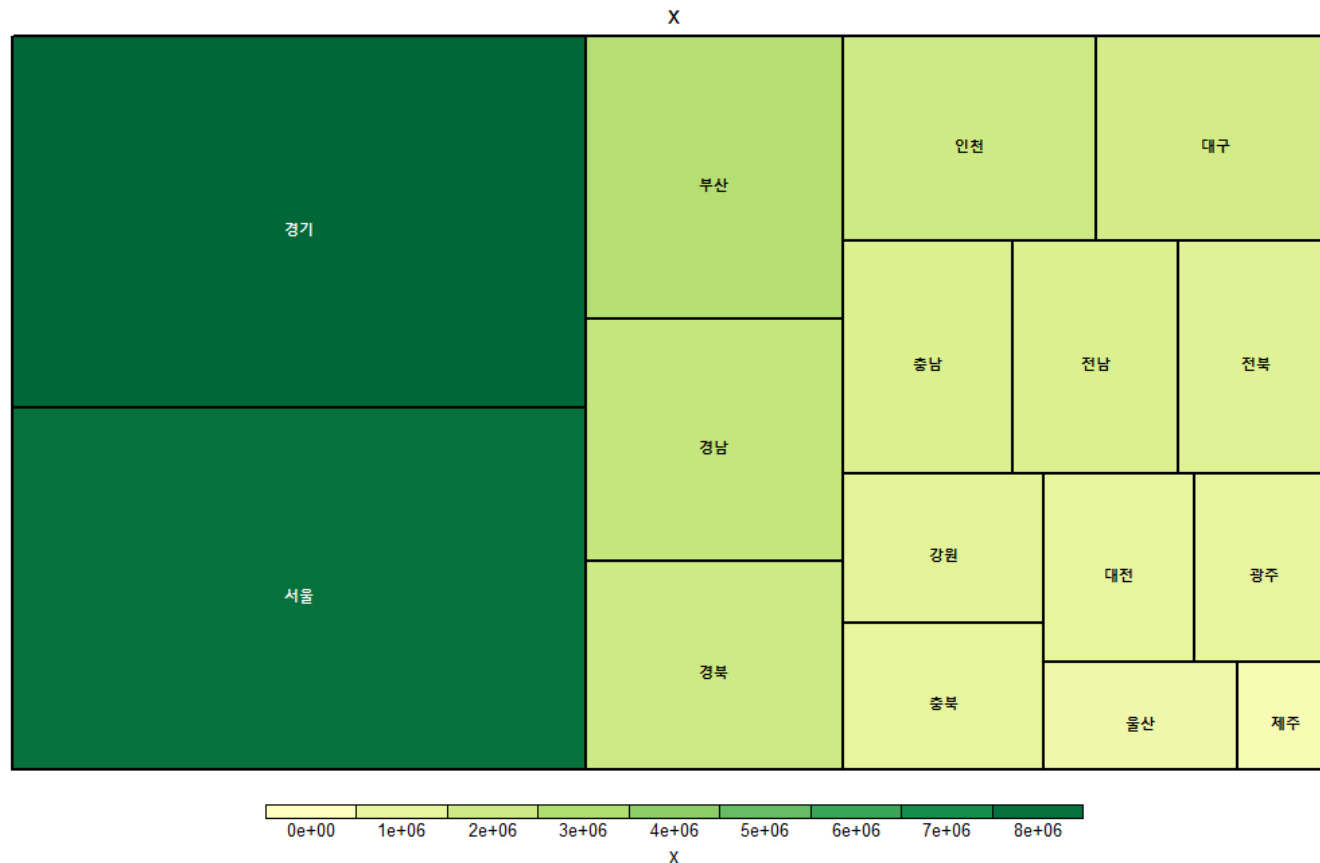
- 1. “국회의원\_선거구\_유권자수.csv” 파일의 내용을 가지고 다음과 같은 treemap 을 작성하시오



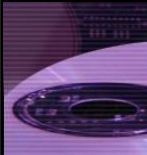
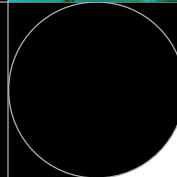
- 타일 하나는 각 선거구를 의미
- 굵은 검은띠 블록은 선거구가 속한 시도를 의미
- 타일의 면적, 색깔은 선거인수를 의미

## [연습 1]

- 2. “국회의원\_선거구\_유권자수.csv” 파일의 내용으로 부터 시도별 선거인 수를 집계(합계계산)하여 다음과 같은 treemap 을 작성하시오



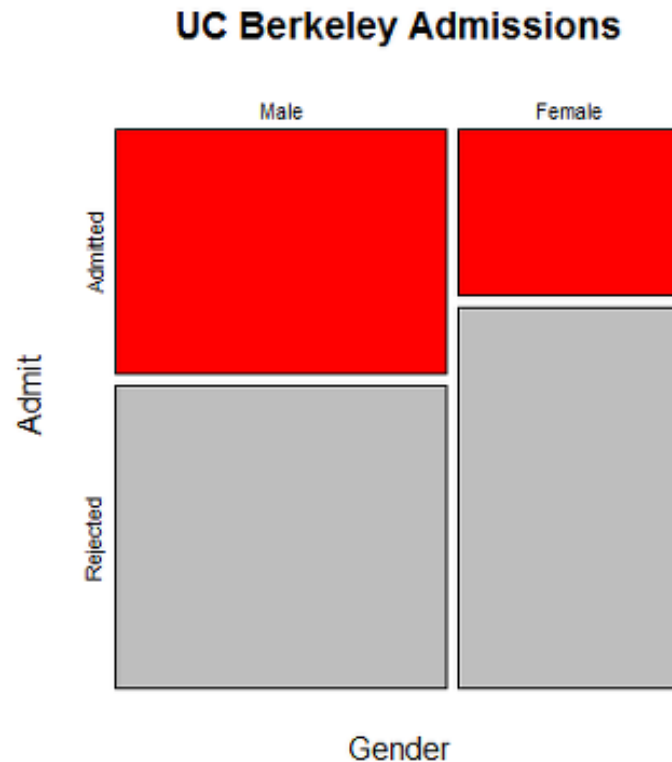
# Mosaic plot





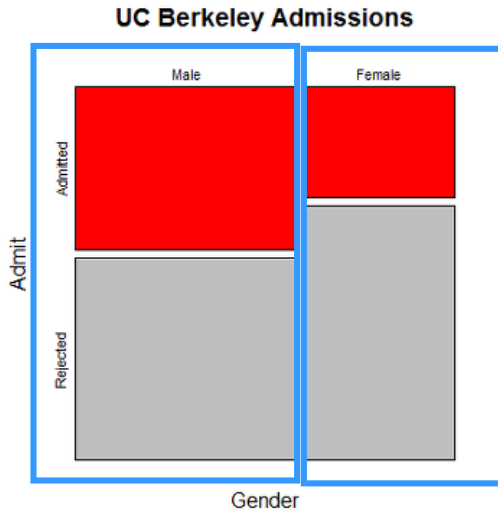
# 모자이크 플롯(mosaic plot)

- 모자이크 플롯(mosaic plot)은 2원 3원 교차표의 시각화이다. 전체 정사각 도형을 교차표의 행 빈도에 비례하는 직사각 도형으로 나누고 다시 각 도형을 행 내 열의 빈도에 해당하는 직사각 도형으로 나눈다.

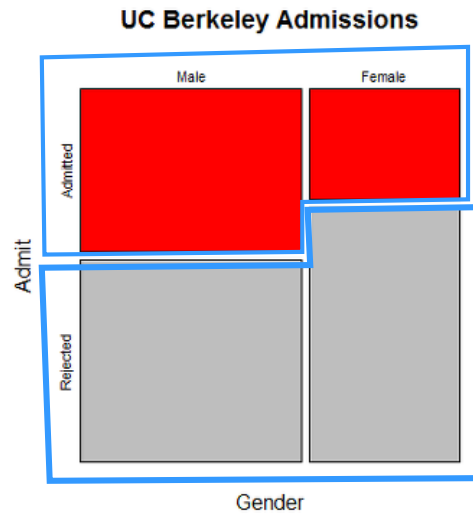


UC Berkeley 대학원  
입시 통계

# 모자이크 플롯 (mosaic plot)

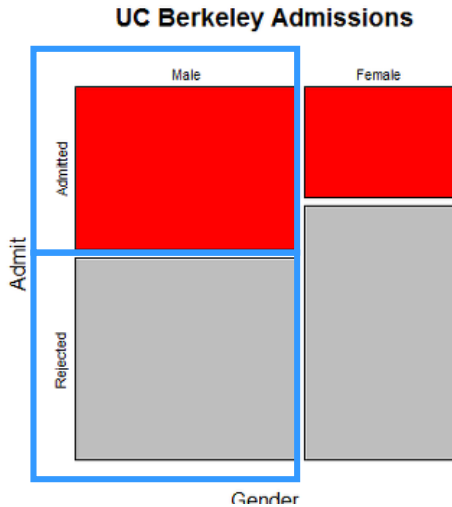


버클리 지원자 중 남성, 여성의 비율  
(면적의 크기가 비율을 나타낸다)

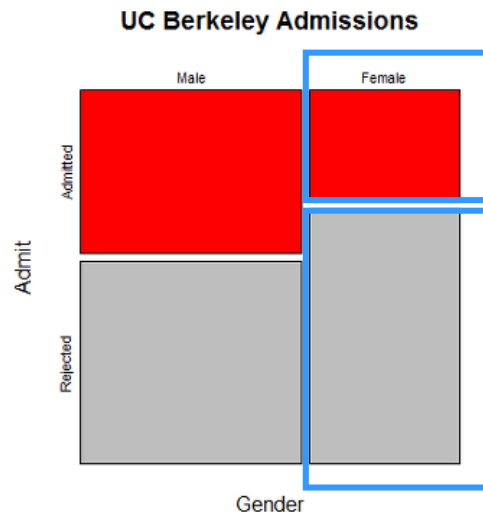


버클리 지원자 중 합격자와, 불합격자의 비율  
(면적의 크기가 비율을 나타낸다)

# 모자이크 플롯 (mosaic plot)

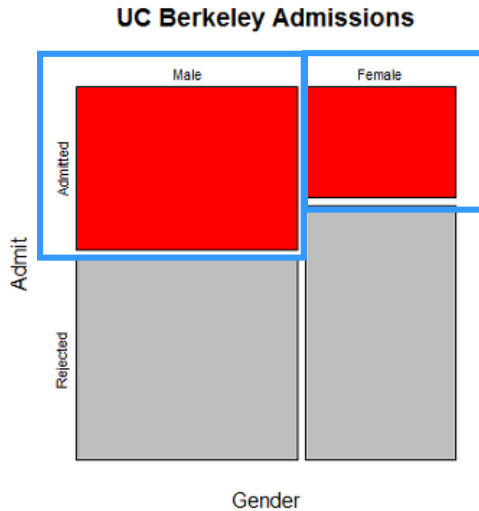


버클리 남성 지원자 중 합격자, 불합격자의 비율  
(면적의 크기가 비율을 나타낸다)



버클리 여성 지원자 중 합격자, 불합격자의 비율  
(면적의 크기가 비율을 나타낸다)

# 모자이크 플롯 (mosaic plot)



버클리 남성 합격자와, 여성 합격자의 비율  
(면적의 크기가 비율을 나타낸다)

(전체적으로는 남성이 합격자 수, 합격률에 있어서 여성보다 앞서는 것을 알 수 있다. -> 남녀차별 문제 제기)

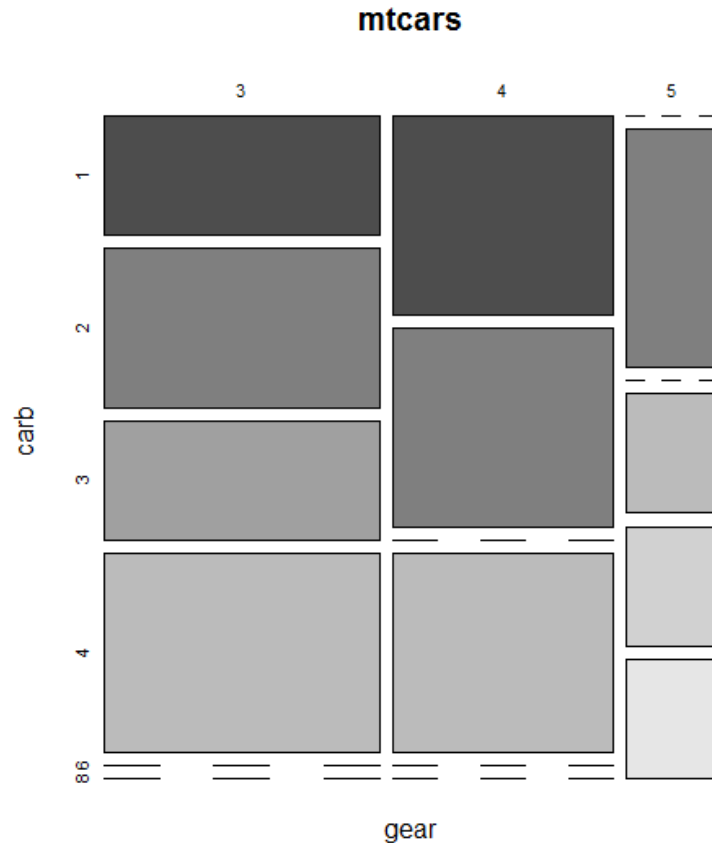
이와 같이 모자이크 플롯은 여러가지 정보를 한눈에 표현할 수 있다

# 모자이크 플롯 (mosaic plot)

- 설치 필요 패키지
  - 없음
- 실습용 데이터셋
  - mtcars
  - Titanic

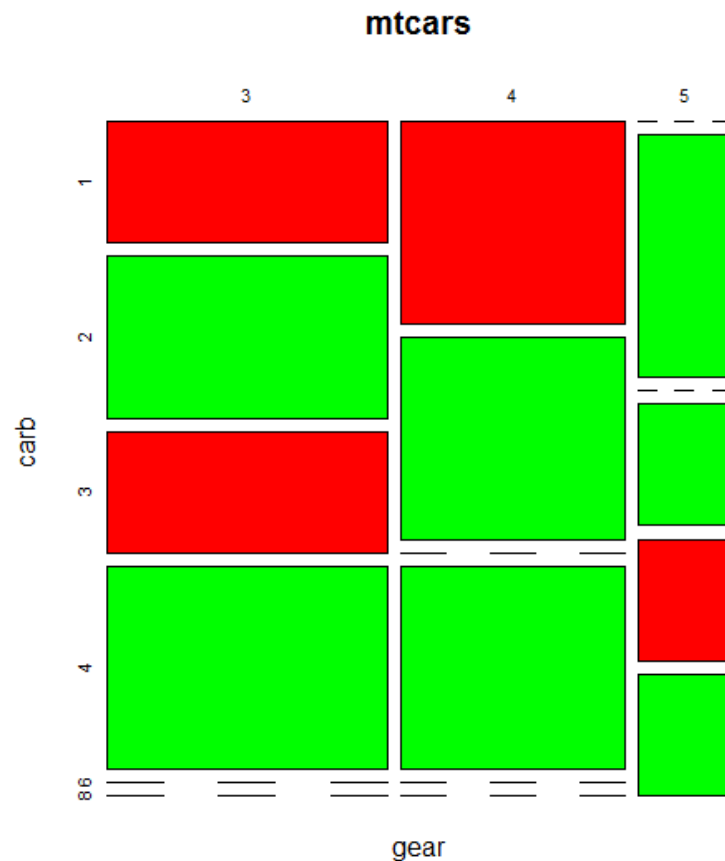
# 모자이크 플롯 (mosaic plot)

```
rm(list=ls())      # 앞의 작업 결과 clear  
# matrix 형태로 데이터가 존재하는 경우  
head(mtcars)  
mosaicplot(~gear+carb, data = mtcars, color = TRUE)
```



# 모자이크 플롯 (mosaic plot)

```
mosaicplot(~gear+carb, data = mtcars,  
           color = c("red", "green"))
```



# 모자이크 플롯 (mosaic plot)

# 교차표 형태로 데이터가 존재하는 경우

**Titanic**

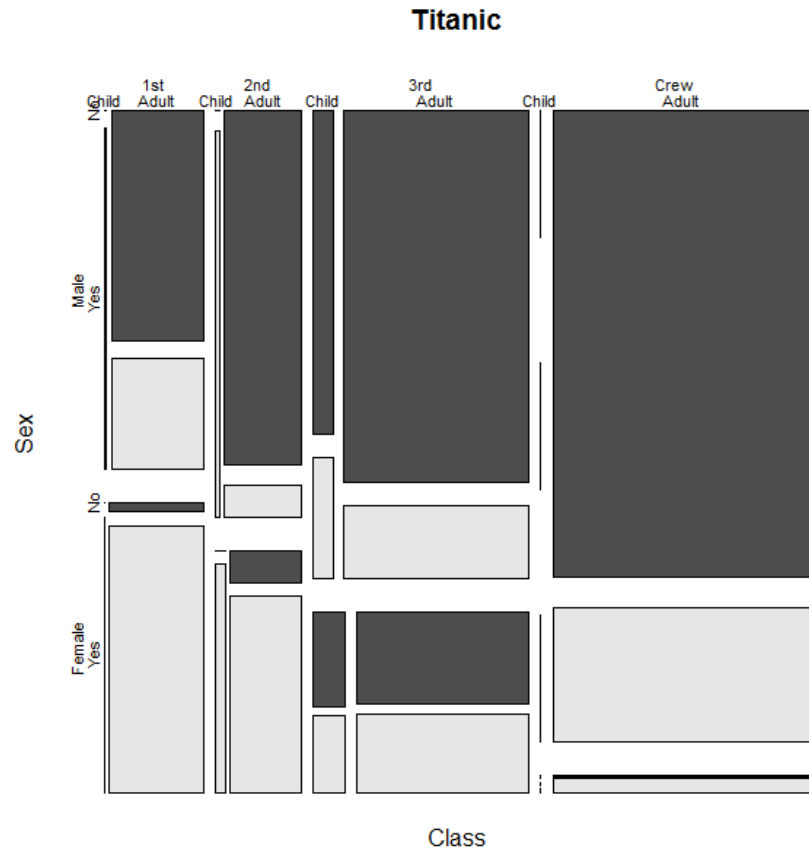
```
mosaicplot(Titanic, color = TRUE, off=5)
```

```
> Titanic  
, , Age = Child, Survived = No
```

	Sex	
Class	Male	Female
1st	0	0
2nd	0	0
3rd	35	17
Crew	0	0

```
, , Age = Adult, Survived = No
```

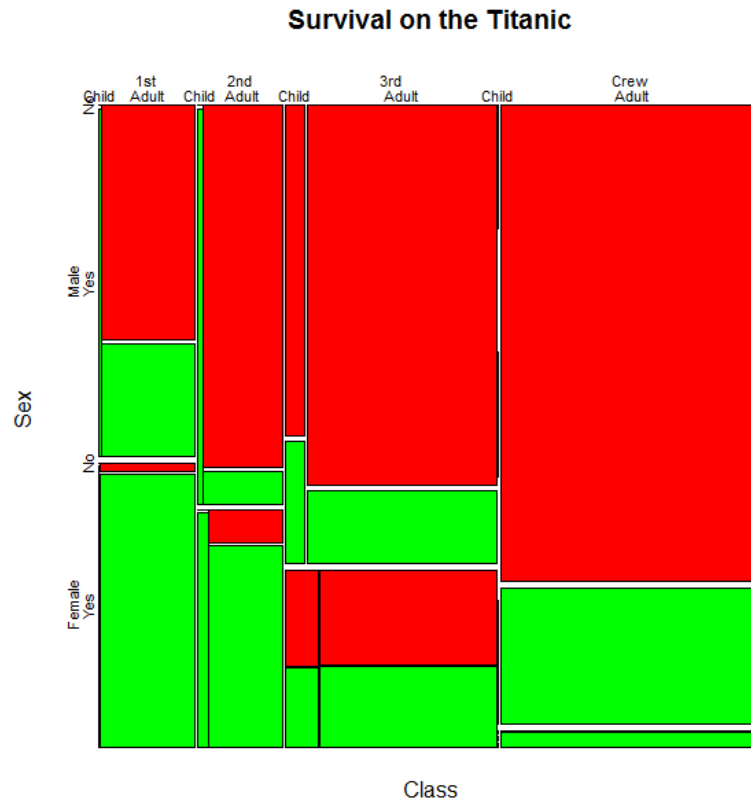
	Sex	
Class	Male	Female
1st	118	4
2nd	154	13





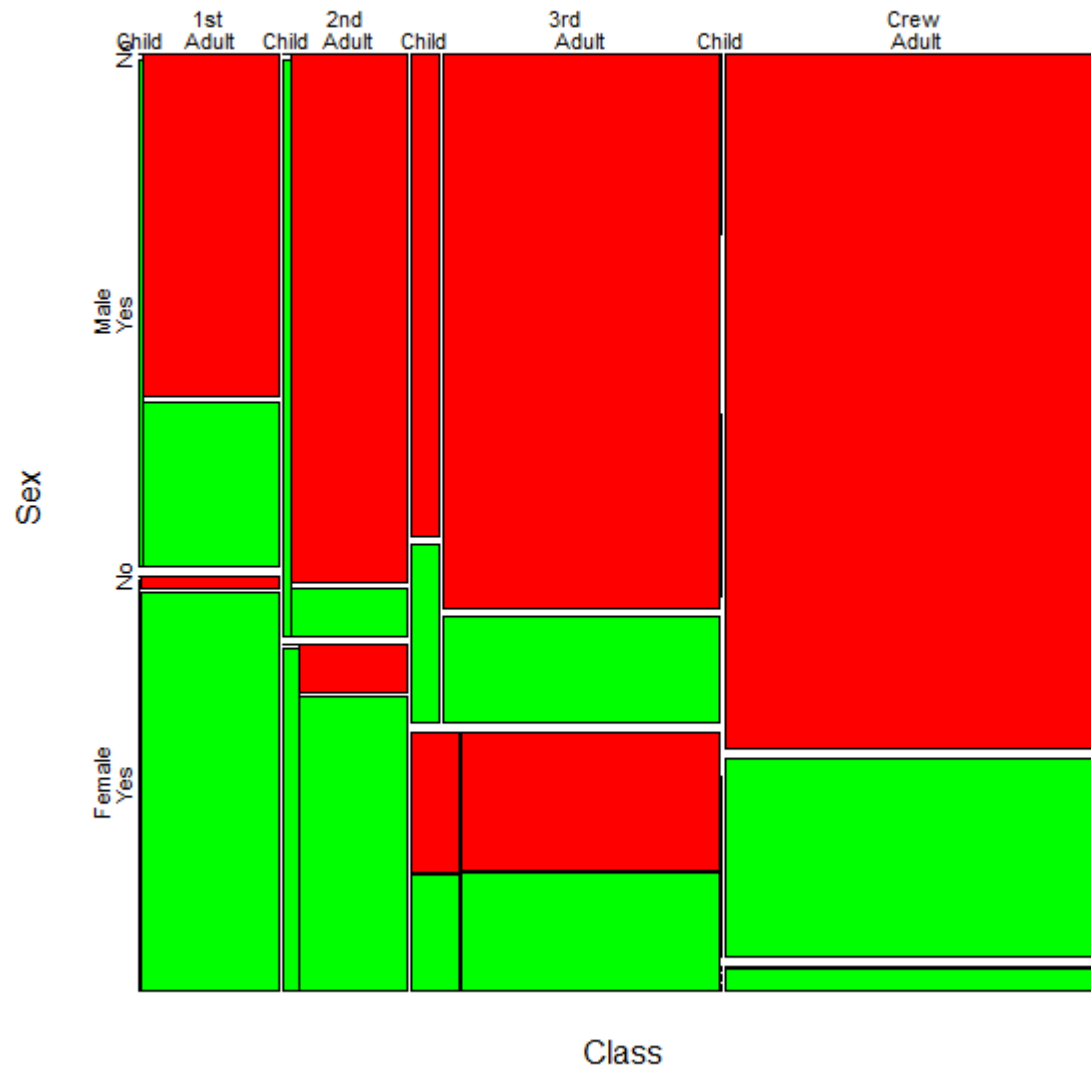
# 모자이크 플롯 (mosaic plot)

```
mosaicplot(Titanic,  
            main = "Survival on the Titanic",  
            color = c("red", "green"),  
            off=1) # 블록들 사이의 간격 지정
```



붉은색 : 사망  
연두색 생존

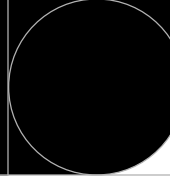
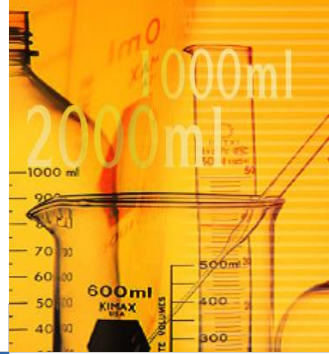
## Survival on the Titanic



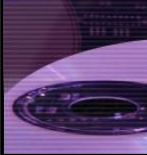
붉은색 : 사망  
연두색 생존

## [연습 4]

1. HairEyeColor 데이터셋에 대해 모자이크 플롯을 작성하시오.  
여기서 관찰할 수 있는 정보는 무엇인가



# 지도상에 데이터 표현하기



- 구글맵 API
- 특정지역 지도보기
- 마커 출력
- 데이터의 크기 지도에 출력

- 구글맵 API 기능을 이용하여 구글지도상에 정보를 표시할 수 있다.



출처

<https://vijaybarve.wordpress.com/tag/ggplot2/>

- 설치가 필요한 패키지
  - ggmap : 구글맵과 연동을 위해 필요
  - ggplot2 : 구글맵 위에 그래프 출력을 위해 필요

# 특정 지역 지도 보기

```
library(ggmap)
gc <- geocode(enc2utf8("용인")) # 지점의 경도위도
cen <- as.numeric(gc)           # 경도위도를 숫자로
map <- get_googlemap(center=cen) # 지도생성
ggmap(map)                      # 지도 화면에 보이기
```

**geocode()**

: 지역명을 경도와 위도로 변환

**as.numeric(gc)**

: 경도와 위도를 숫자로 변환

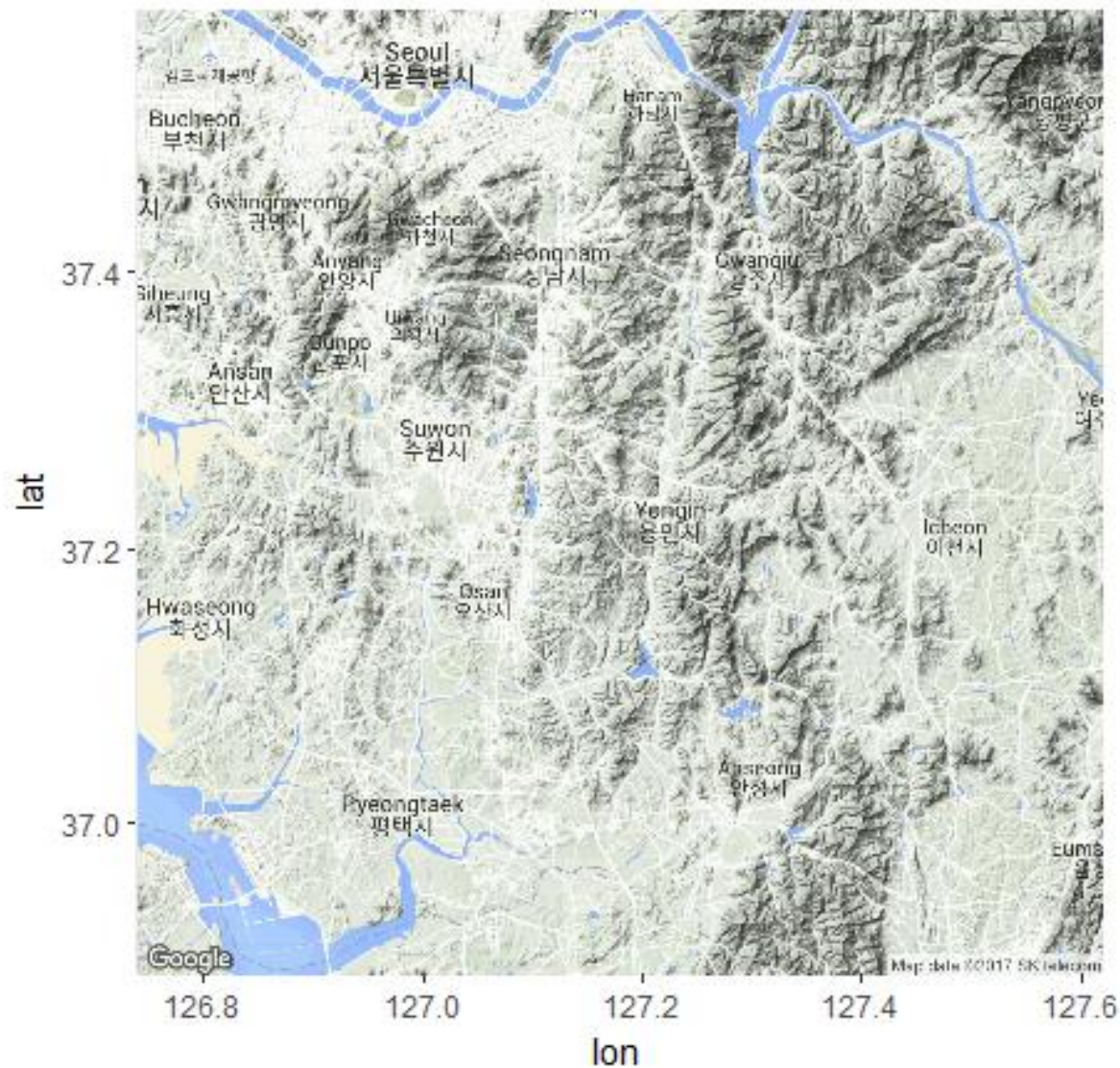
**get\_googlemap()**

: 지정된 지역의 구글 지도를 가져온다

**center=cen** : 지도의 중심점 지정

```
> gc
      lon      lat
1 127.1776 37.24109
> cen
[1] 127.17755 37.24109
```

# 특정 지역 지도 보기





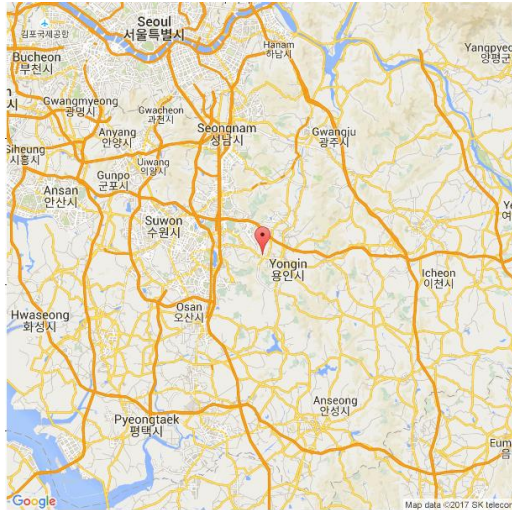
- `get_googlemap`

Parameter	설명
center	지도의 중심좌표
zoom	지도의 확대크기로서 3(대륙)~21(빌딩). 기본값은 10 (도시)
size	지도의 가로와 세로 픽셀 크기. 기본값은 640 x 640 (c(640,640))
maptype	출력될 지도유형 (다음페이지 참조) 기본값은 " <b>terrain</b> "

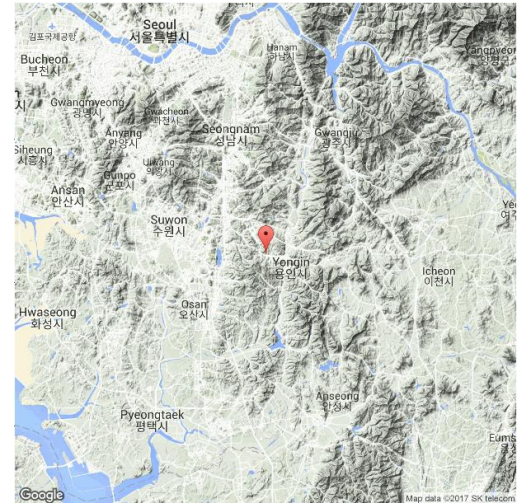
# 마커 출력

- maptype

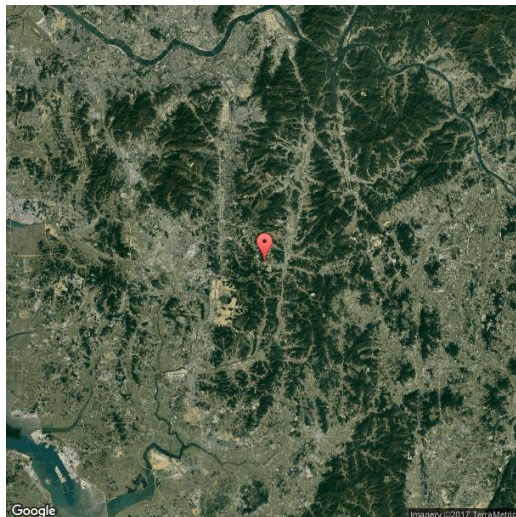
"roadmap"



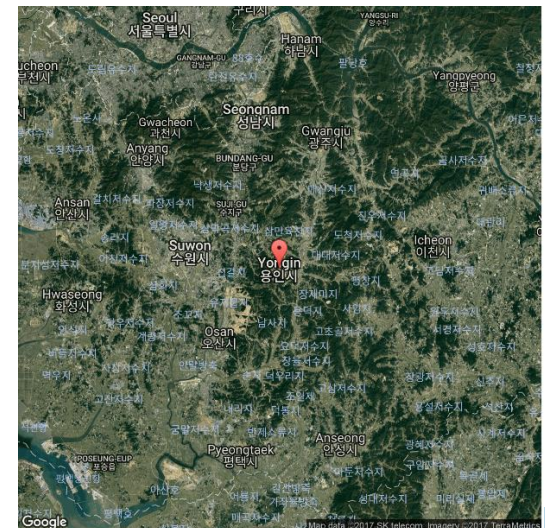
"terrain"




"satellite"



"hybrid"



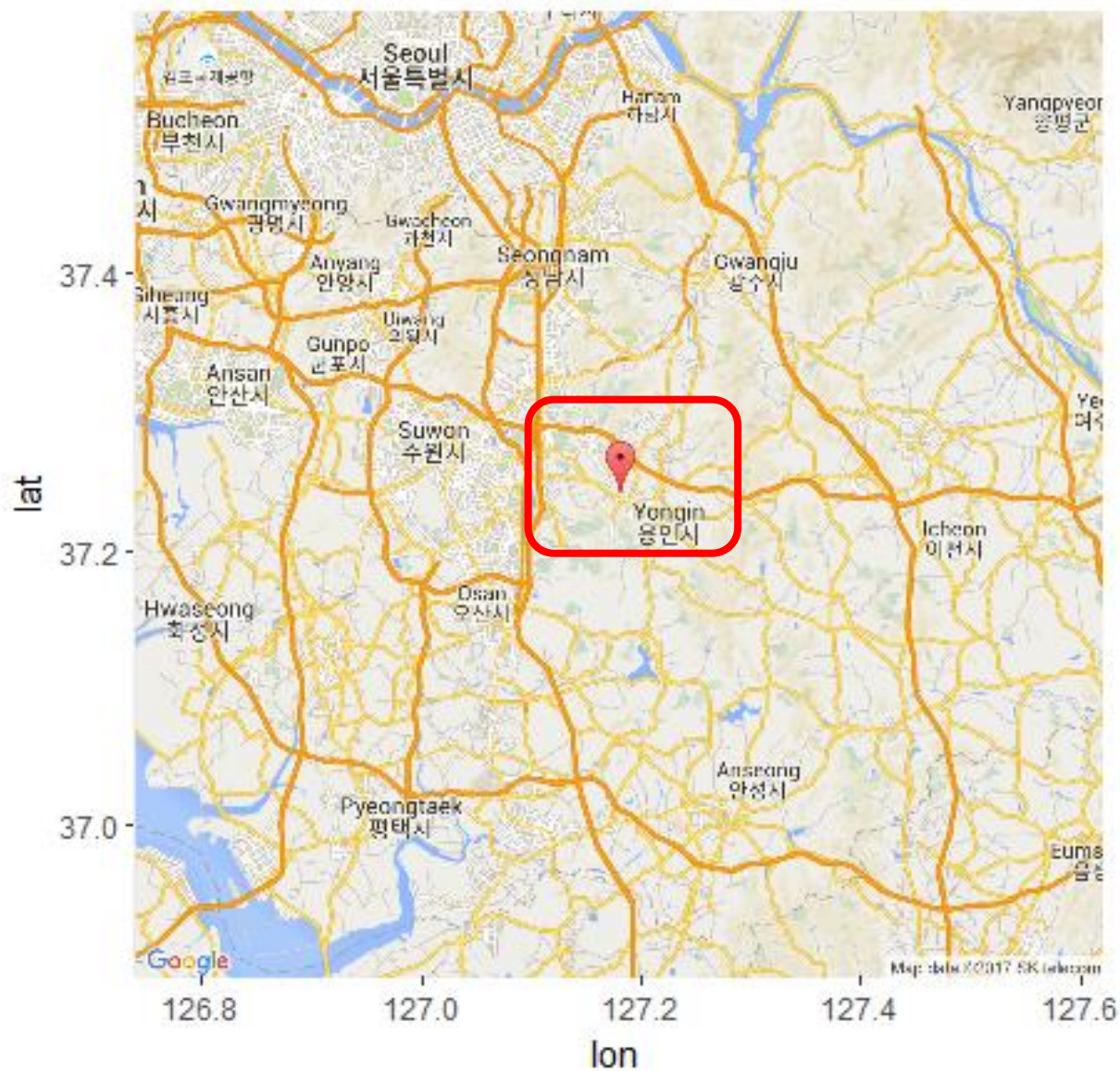
# 마커 출력

- 지도의 위도,경도 위치에 마커(  )를 출력한다

- 용인에 마커를 표시해보기

```
library(ggmap)
gc <- geocode(enc2utf8("용인")) # 지점의 경도위도
cen <- as.numeric(gc)           # 경도위도를 숫자로
map <- get_googlemap(center=cen, # 지도의 중심
                      maptype="roadmap", # 지도의 형태
                      marker=gc)        # 마커의 위치
ggmap(map)                       # 지도 화면에 보이기
```





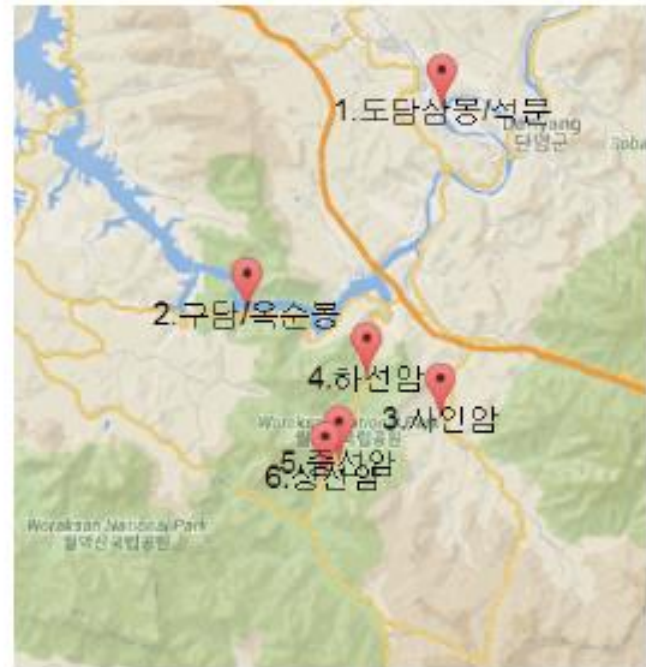
# 마커 출력

- 여러 지역의 마커 표시하기

충청북도 단양군에 있는 명소인 단양 팔경의 위치를 표시하는 관광 안내 지도를 만들어 보자.

지명	주소
도담삼봉/석문	매포읍 삼봉로 644-33
구담/옥순봉	단성면 월악로 3827
사인암	대강면 사인암2길 42
하선암	단성면 선암계곡로 1337
중선암	단성면 선암계곡로 868-2
상선암	단성면 선암계곡로 790

## 미리 보기



```
library(ggmap)
library(ggplot2)

names <- c("1.도담삼봉/석문", "2.구담/옥순봉", "3.사인암",
           "4.하선암", "5.중선암", "6.상선암")
addr <- c("매폍읍 삼봉로 644-33",
          "단성면 월악로 3827",
          "대강면 사인암2길 42",
          "단성면 선암계곡로 1337",
          "단성면 선암계곡로 868-2",
          "단성면 선암계곡로 790")
gc <- geocode(enc2utf8(addr)) #주소를 경도, 위도로 변환
df <- data.frame(name=names,
                  lon=gc$lon,
                  lat=gc$lat)
```

```
> df
```

		name	lon	lat
1	1.	도담삼봉/석문	128.3433	37.00300
2	2.	구담/옥순봉	128.2560	36.93046
3	3.	사인암	128.3404	36.89439
4	4.	하선암	128.3094	36.90788
5	5.	중선암	128.2969	36.87783
6	6.	상선암	128.2907	36.87222

```
>
```

*(continue)*

```
cen <- c(mean(df$lon), mean(df$lat))  
map <- get_googlemap(center=cen,  
  maptype="roadmap",  
  zoom=11,  
  marker=gc)
```

```
ggmap(map)
```

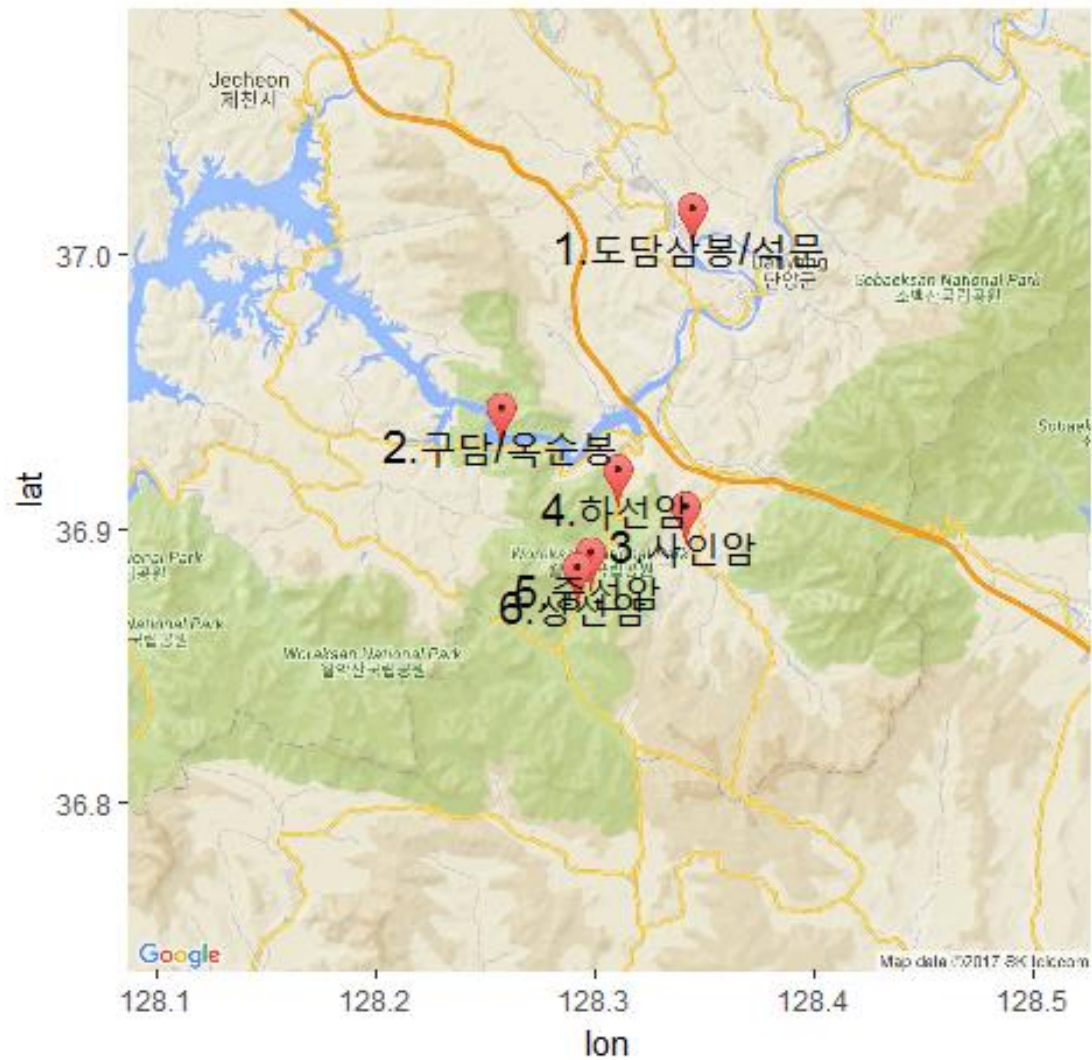
# 지도+마커 화면에 보이기

# 명소이름 지도위에 표시하기

```
gmap <- ggmap(map)  
gmap+geom_text(data=df,  
  aes(x=lon, y=lat),  
  size=5,  
  label=df$name)
```

데이터 프레임 df를 기반으로 경도(lon)와 위도(lat)를 x, y 좌표로 설정하고(aes), 출력할 문자의 크기를 5로 설정하며(size), df의 name 값을 문자로 출력





# 마커 출력

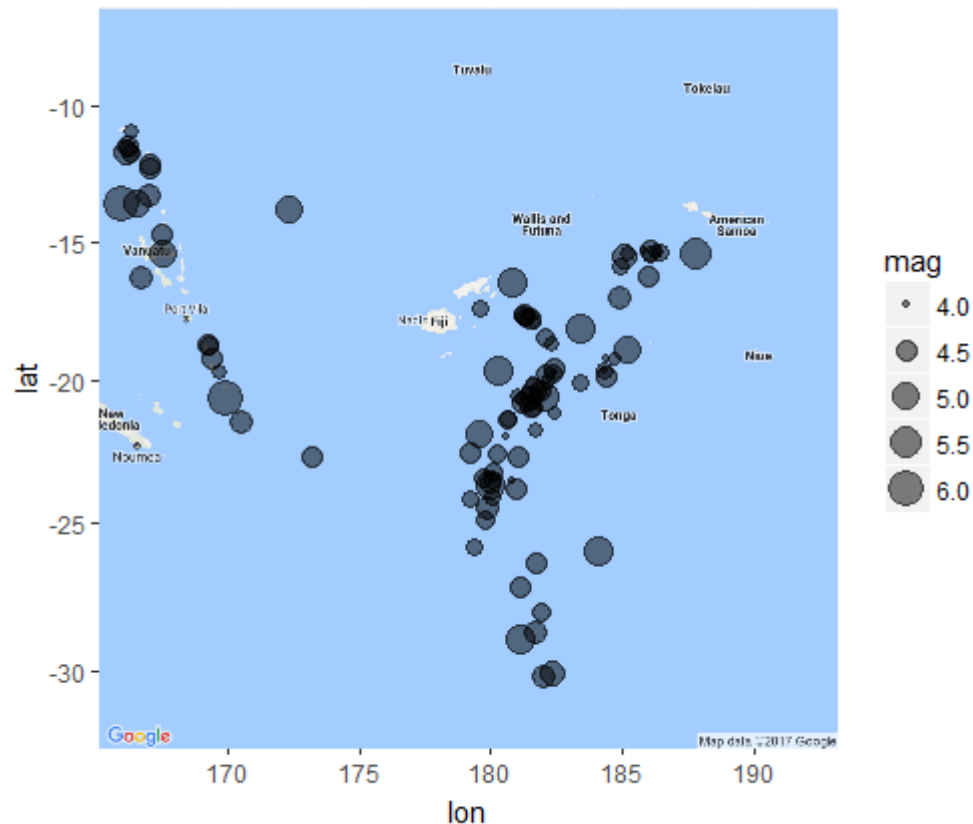
- 구글지도를 이용하여 특정 지역의 경도, 위도 알아내기
  - <https://maps.google.com>



위도      경도

# 데이터의 크기를 지도에 표현하기

- R 에서 제공하는 지진 발생 데이터(quakes)를 이용하여 지진규모를 발생지역에 표시해보자



# 데이터의 크기를 지도에 표현하기

```
> head(quakes)
```

	lat	long	depth	mag	stations
1	-20.42	181.62	562	4.8	41
2	-20.62	181.03	650	4.2	15
3	-26.00	184.10	42	5.4	43
4	-17.97	181.66	626	4.1	19
5	-20.42	181.96	649	4.0	11
6	-19.68	184.31	195	4.0	12

위도

경도

지진규모(진도)

진앙지 깊이

# 데이터의 크기를 지도에 표현하기

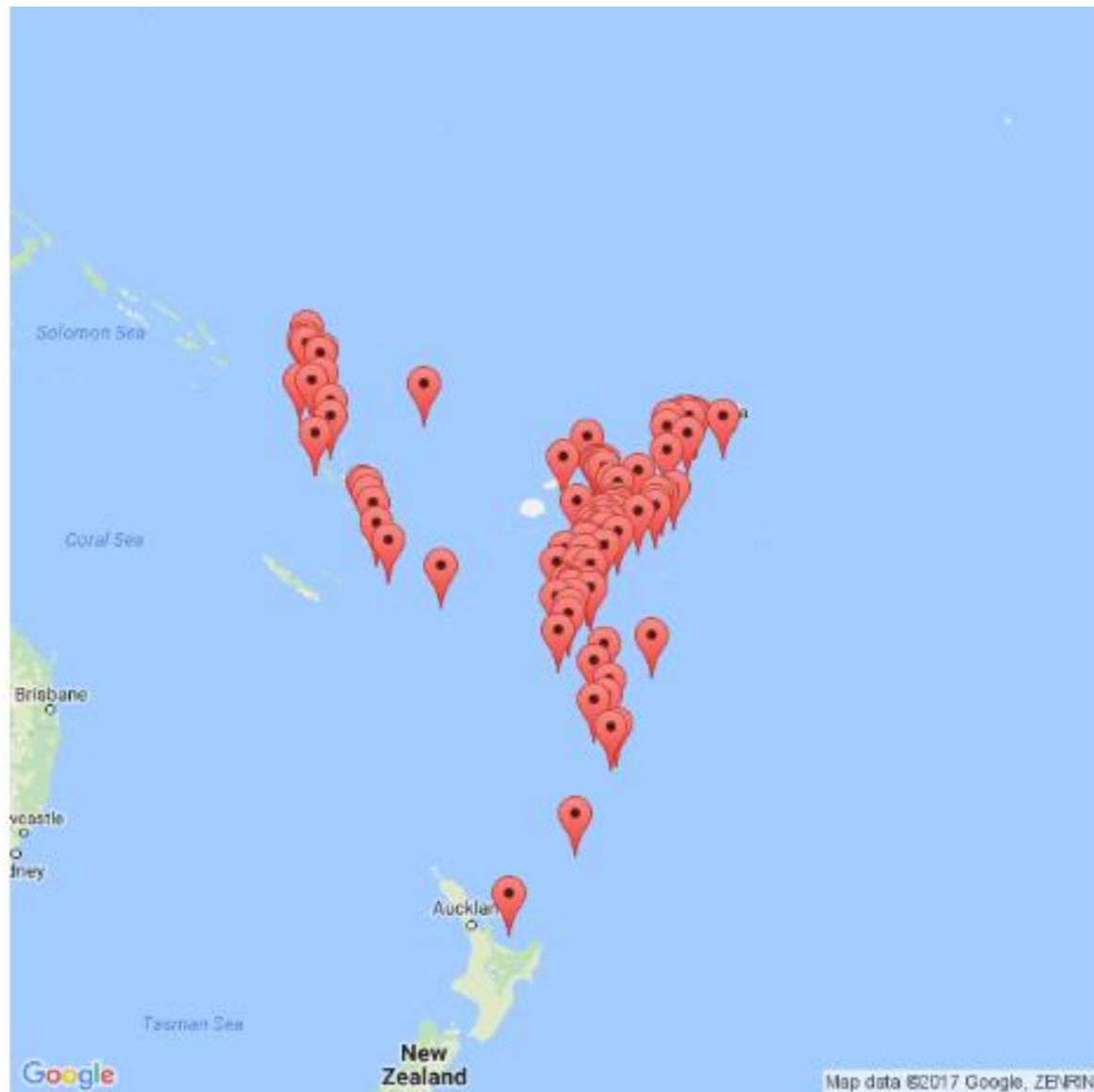
```
library(ggmap)
library(ggplot2)

df <- head(quakes, 100)
cen <- c(mean(df$lon), mean(df$lat))
gc <- data.frame(lon=df$lon, lat=df$lat)
gc$lon <- ifelse(gc$lon>180, -(360-gc$lon), gc$lon)
gc
map <- get_googlemap(center=cen,
  maptype="roadmap",
  zoom=4,
  marker=gc)
ggmap(map, extent="device")
```

④ 경도가 180도를 넘는 경우,  
0~-180도 사이로 변환  
⇨ ifelse(조건, 조건이 참일 경우  
의 값, 조건이 거짓일 경우의 값)

x,y축 레이블을 보이지 않게함.

# 데이터의 크기를 지도에 표현하기



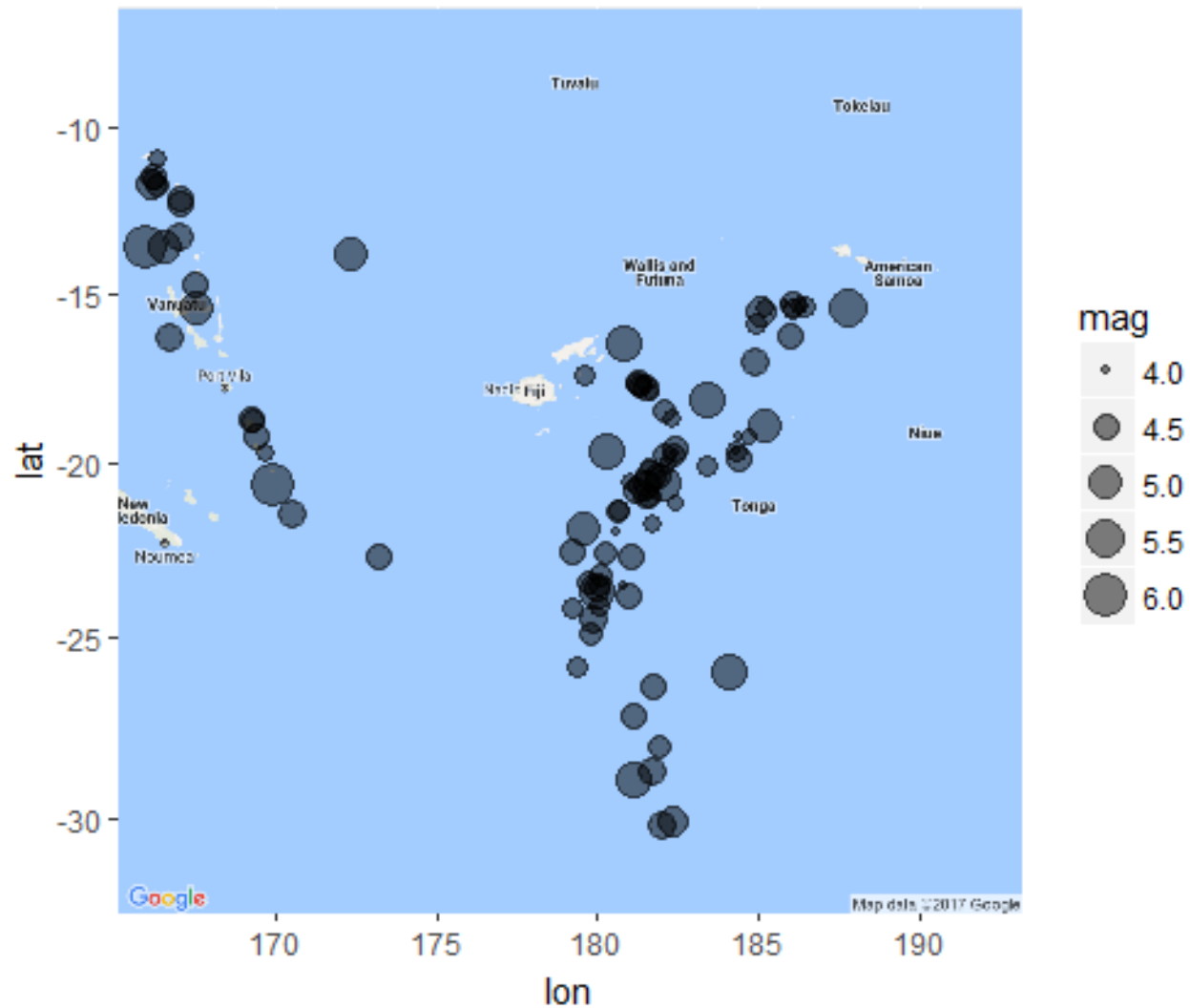
# 데이터의 크기를 지도에 표현하기

*(continue)*

```
map <- get_googlemap(center=cen,  
  maptype="roadmap",  
  zoom=5)  
gmap <- ggmap(map)  
gmap+geom_point(data=df,  
  aes(x=long,y=lat,size=mag),  
  alpha=0.5)
```

지도 위에 df 데이터 내의 경도(long)와 위도(lat)를 x, y 좌표로 하는 위치에 원의 크기(mag)로 표시. 투명도(alpha)는 완전 투명(0)과 완전 불투명(1)의 중간(0.5)

# 데이터의 크기를 지도에 표현하기





## [과제 1]

- 서울지역의 이마트 지점을 지도상에 마커로 표시하시오
  - 지점명도 함께 출력
  - 지점 주소 검색 : <http://emart.ssg.com/>

### 이마트 서비스 안내

---

- 이마트 지점 찾기 (휴점일)
- 이마트 고객만족 센터
- 이마트 이벤트/행사
- 신세계 포인트카드
- SSG PAY

(화면의 하단부에 있음)