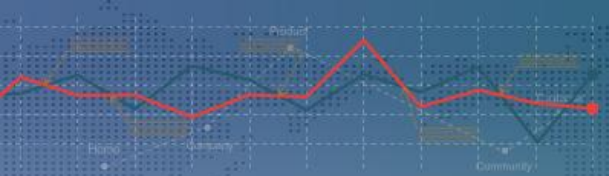
5강 이변량 데이터의 시각화 (2)

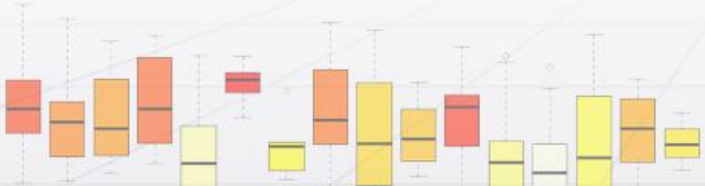
고려대학교 통계학과 허명회 교수

1. 모자이크 플롯
2. 나무 지도



1. 모자이크 플롯 (mosaic plot)

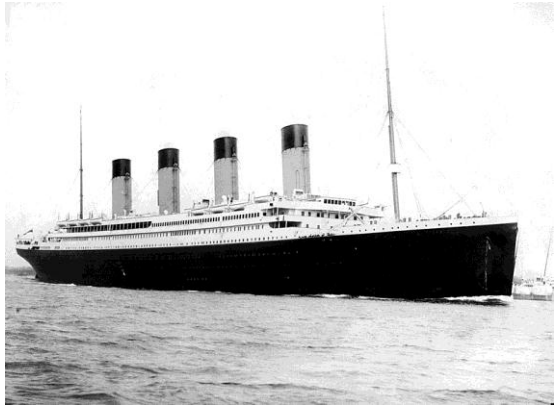
- 범주형 변수와 교차표
- 모자이크 플롯
- 3원 교차표의 모자이크 플롯
- 세분화된 2원 교차표의 활용
- 심프슨의 파라독스



1 모자이크 플롯

범주형 변수 (categorical data)

- 명목 범주형 변수 및 자료
- 보기 1. Titanic



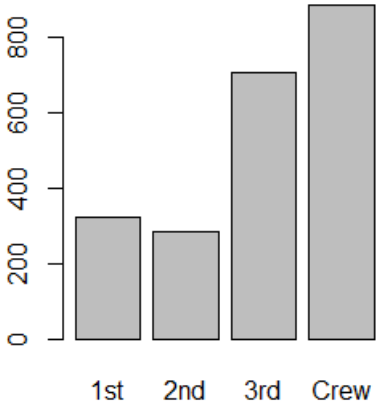
- `Class` (1등실,2등실,3등실,선원실)
- `Sex` (남자,여자)
- `Age` (성인,아동)
- `Survived` (사망,생존)

1 모자이크 플롯

범주형 변수(categorical data)

보기 1. Titanic

```
> data(Titanic)
> str(Titanic)
  table [1:4, 1:2, 1:2, 1:2]
  ...
> apply(Titanic,1,sum)
  1st   2nd   3rd Crew
  325   285   706   885
> barplot(apply(Titanic,1,sum))
```



1 모자이크 플롯

교차표 (cross tabulation)

- 2원 교차표: 행*열

```
> apply(Titanic,c(1,4),sum)
        Survived
Class     No Yes
  1st    122 203
  2nd    167 118
  3rd    528 178
  Crew   673 212
```

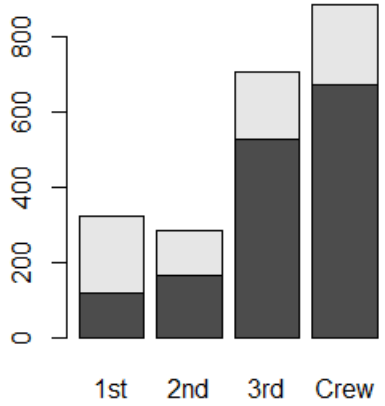
* 행 = Class, 열 = Survived

1 모자이크 플롯

교차표 (cross tabulation)

2원 교차표: 행*열

```
> apply(Titanic,c(4,1),sum)
          Class
Survived 1st 2nd 3rd Crew
     No  122 167 528  673
     Yes 203 118 178  212
```

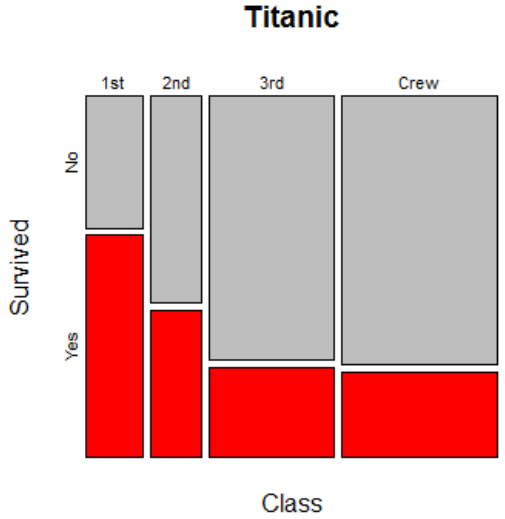



```
> barplot(apply(Titanic,c(4,1),sum))
```

1 모자이크 플롯

모자이크 플롯 (mosaic plot)

- 교차표를 행의 빈도에 비례하는 폭의 수직 막대를 세우고
- 다음엔 각 수직 막대를 행 내 열의 빈도에 비례하게 나눔



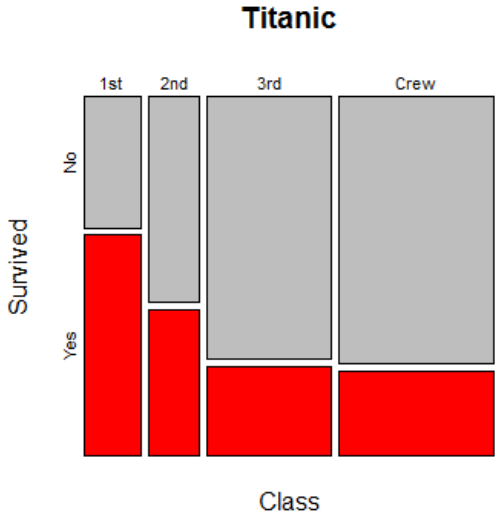
- `Titanic` 적용

`Class` 별 생존율 비교가 쉽다.

모자이크 플롯 (mosaic plot)

- R 스크립트

```
mosaicplot(~ Class+Survived,data=Titanic,
           color=c("grey","red"))
```

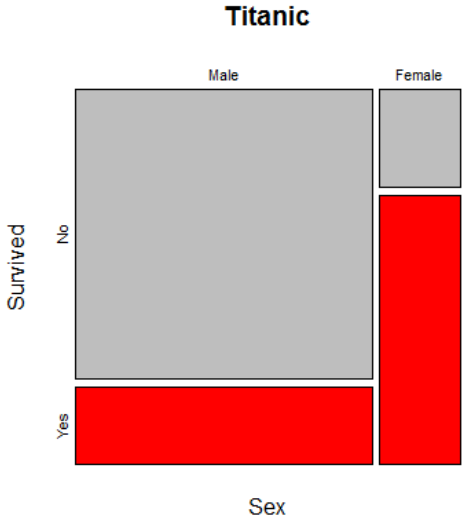


- Class별로 생존율이 다르다.

모자이크 플롯 (mosaic plot)

- R 스크립트

```
mosaicplot(~ Sex+Survived,data=Titanic,
                color=c("grey","red"))
```

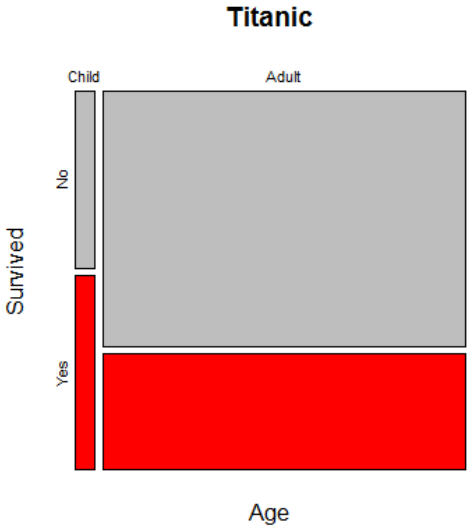


- Sex별로 생존율이 다르다.

모자이크 플롯 (mosaic plot)

- R 스크립트

```
mosaicplot(~ Age+Survived,data=Titanic,
           color=c("grey","red"))
```



- Age별로 생존율이 다르다.

1 모자이크 플롯

3원 교차표의 모자이크 플롯

- Class 별로 성별 생존율이 다른가?

```
apply(Titanic,c(2,4,1),sum)
```

```
, , Class = 1st

        Survived
Sex       No Yes
  Male   118  62
  Female   4 141

, , Class = 2nd

        Survived
Sex       No Yes
  Male   154  25
  Female  13  93

, , Class = 3rd

        Survived
Sex       No Yes
  Male   422  88
  Female 106  90

, , Class = Crew

        Survived
Sex       No Yes
  Male   670 192
  Female   3  20
```

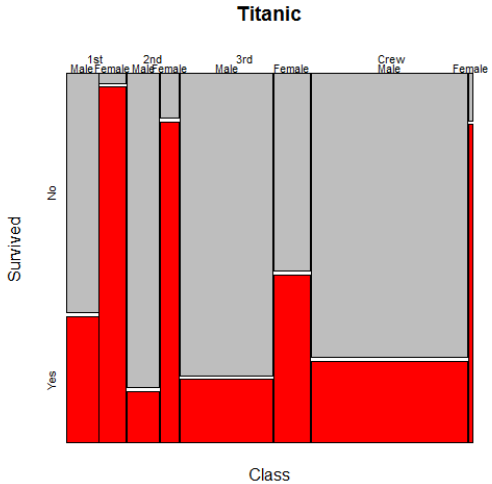
→ 3원 교차표

1 모자이크 플롯

3원 교차표의 모자이크 플롯

- Class 별로 성별 생존율이 다른가?

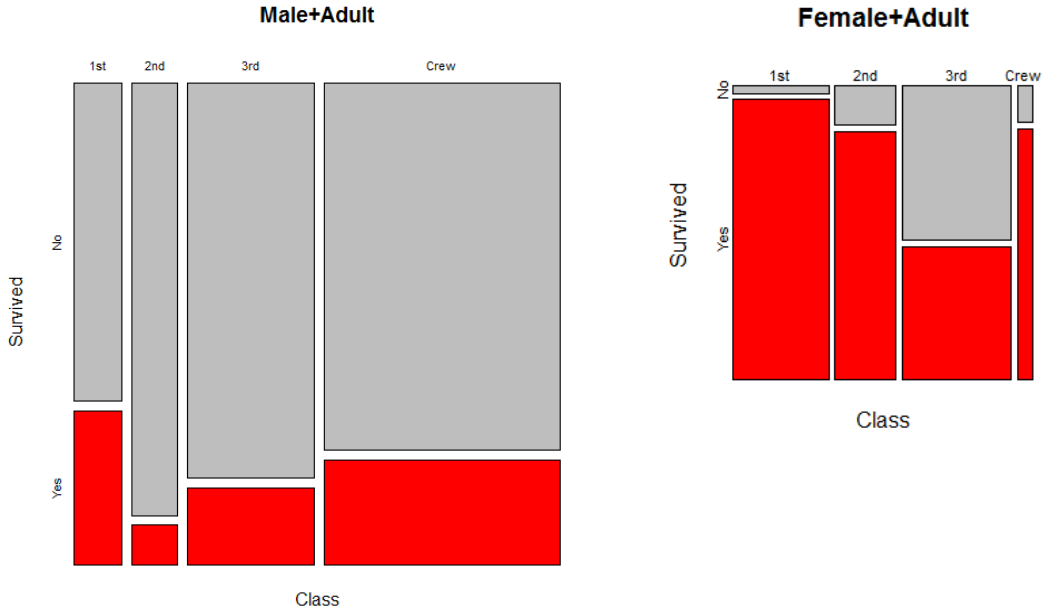
```
mosaicplot(~ Class+Sex+Survived,data=Titanic,
    color=c("grey","red"),dir=c("v","v","h"),off=1)
```



1 모자이크 플롯

세분화된 2원 교차표의 활용 (대안)

- 성인의 경우 Class 별, 성별 생존율은 어떤가?



1 모자이크 플롯

세분화된 2원 교차표의 활용 (대안)

- R 스크립트 : 성인의 경우 Class 별, 성별 생존율은 어떤가?

```
mosaicplot(~ Class+Survived,
  data=as.table(Titanic[,"Male","Adult",]),
  color=c("grey","red"),main="Male+Adult")
mosaicplot(~ Class+Survived,
  data=as.table(Titanic[,"Female","Adult",]),
  color=c("grey","red"),main="Female+Adult")
```

1 모자이크 플롯

심프슨의 파라독스 (Simpson's paradox)

- 사례 : 버클리 대학원 입학자료

```
> data(UCBAdmissions)
> str(UCBAdmissions)
> apply(UCBAdmissions,c(1,2),sum)
                Gender
Admit            Male Female
Admitted         1198    557
Rejected         1493   1278
```

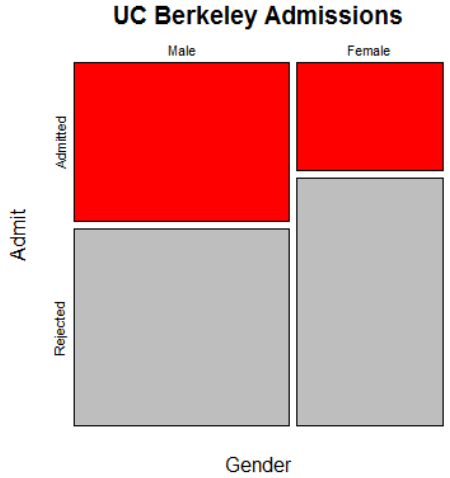
  - 남자 합격률 44.5% > 여자 합격률 30.4%

1 모자이크 플롯

심프슨의 파라독스 (Simpson's paradox)

- 사례: 버클리 대학원 입학자료

```
> mosaicplot(~Gender+Admit, data=UCBAdmissions,
    color=c("red","grey"), main = "UC Berkeley Admissions")
```



1 모자이크 플롯

심프슨의 파라독스 (Simpson's paradox)

- 사례: 버클리 대학원 입학자료

```
> UCBAdmissions
, , Dept = A

          Gender
Admit       Male Female
  Admitted   512     89
  Rejected   313     19

, , Dept = B

          Gender
Admit       Male Female
  Admitted   353     17
  Rejected   207      8
```

A 학과의 경우
남자합격률 62.1% < 여자 합격률 82.4%

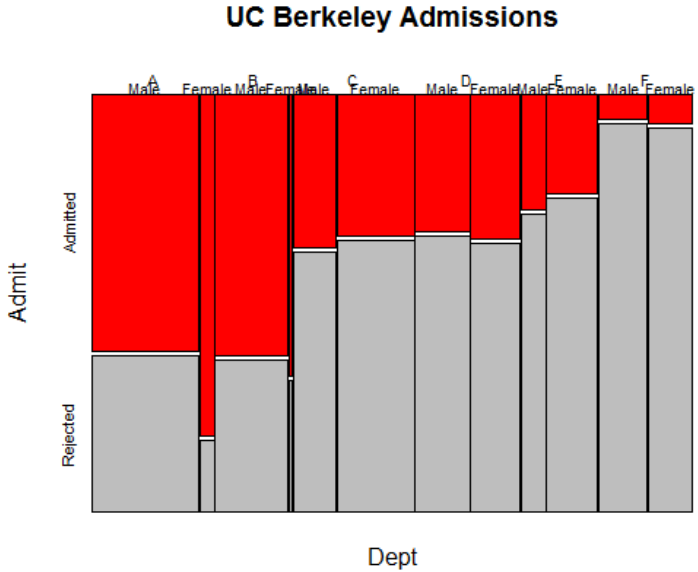
B 학과의 경우
남자 63% < 여자 68%

1 모자이크 플롯

심프슨의 파라독스 (Simpson's paradox)

- 사례: 버클리 대학원 입학자료

```
> UCBAdmissions
```

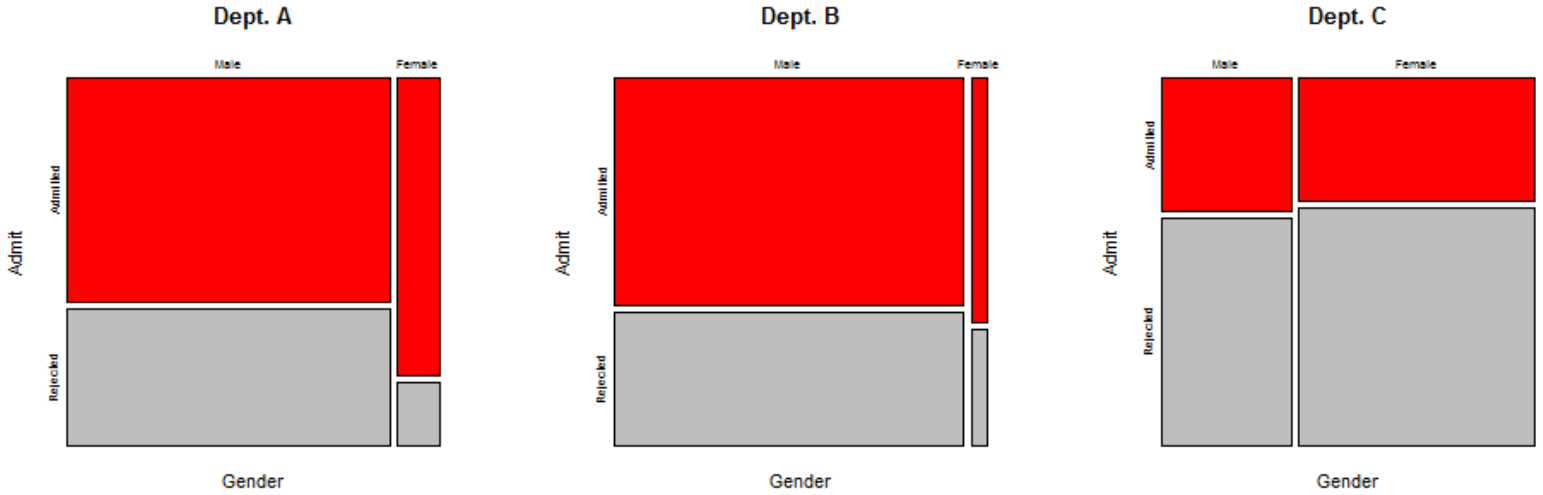



```
windows(height=5, width=6)
mosaicplot(~Dept+Gender+Admit,
   data=UCBAdmissions,
   color=c("red","grey"),
   dir=c("v","v","h"), off=1,
   main = "UC Berkeley Admissions")
```

1 모자이크 플롯

심프슨의 파라독스 (Simpson's paradox)

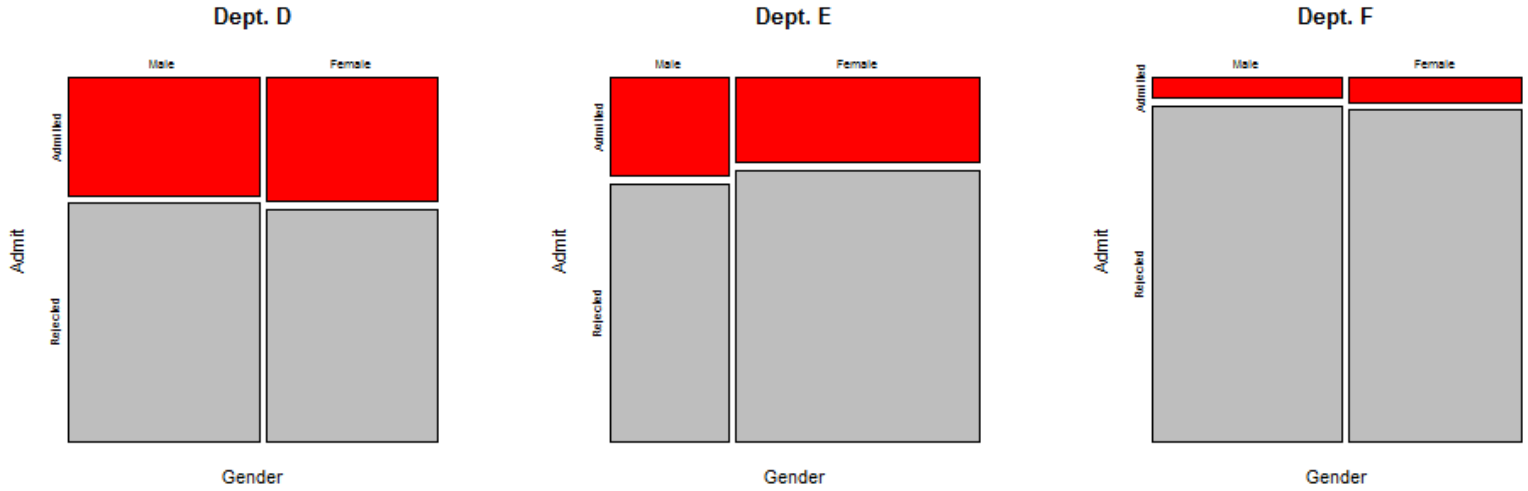
- 사례 : 버클리 대학원 입학자료 (학과 A, B, C)



1 모자이크 플롯

심프슨의 파라독스 (Simpson's paradox)

- 사례 : 버클리 대학원 입학자료 (학과 D, E, F)



1 모자이크 플롯

심프슨의 파라독스 (Simpson's paradox)

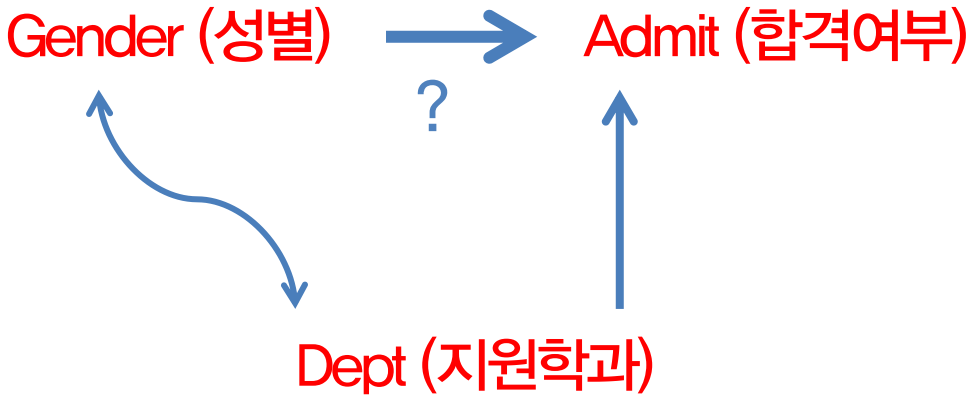
- 사례: 버클리 대학원 입학자료에서 학과별 모자이크 플롯 R 스크립트

```
windows(height=3.6, width=10); par(mfrow=c(1,3))
mosaicplot(~Gender+Admit, data=as.table(UCBAdmissions[,,"A"]),
    color=c("red","grey"), main = "Dept. A")
mosaicplot(~Gender+Admit, data=as.table(UCBAdmissions[,,"B"]),
    color=c("red","grey"), main = "Dept. B")
mosaicplot(~Gender+Admit, data=as.table(UCBAdmissions[,,"C"]),
    color=c("red","grey"), main = "Dept. C")
```

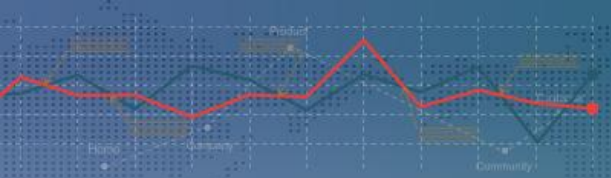
1 모자이크 플롯

심프슨의 파라독스 (Simpson's paradox)

- 왜, 어떤 이유로 이런 논리적 모순이 발생하는가?

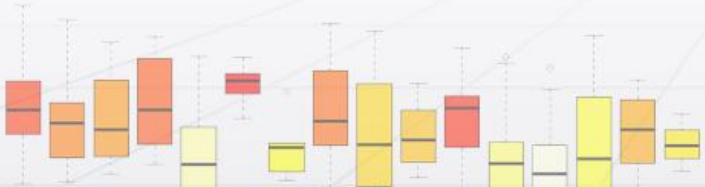


* 교락효과 (confounding effect): 제 3의 변수의 영향



2. 나무 지도 (tree map)

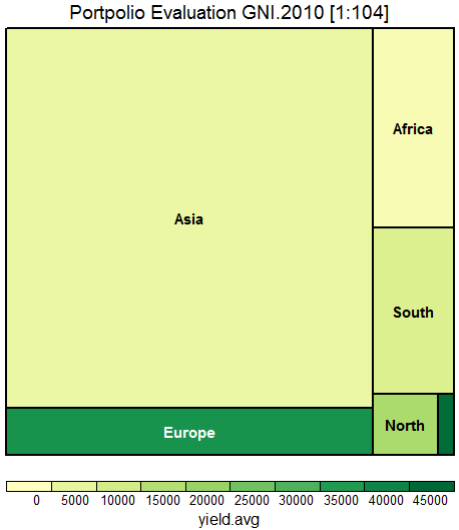
- 나무지도는…
- 보기 1. GNI2010 자료
- R 스크립트
- R 그래프

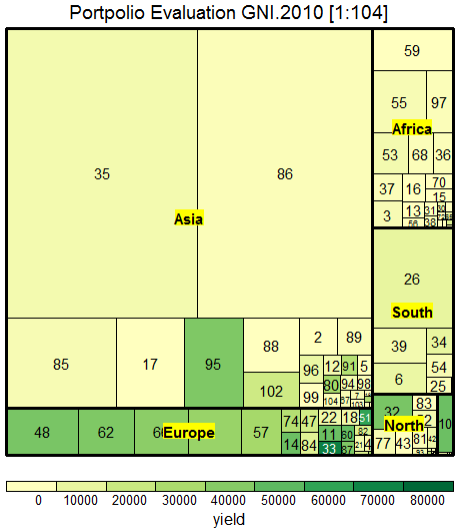


2 나무 지도

나무 지도(tree map)는 ….

- 컬러 타일을 계층적으로 배열하여 붙인 그림





2 나무 지도

나무 지도(tree map)는 ….

- 포트폴리오(portfolio) 데이터에 적용

투자영역	투자액	수익률
Asia		
Africa		
Europe		
South A.		
North A.		
:		

2 나무 지도

나무 지도(tree map)는 ….

- 계층적 구조의 포트폴리오 데이터에 적용

투자영역	항 목	투자액	수익률
Euro	투자영역	항 목	투자액	수익률
	Afric	투자영역	항 목	투자액	수익률
		Asia	1		
			2		
			3		
			:		

2 나무 지도

보기 1. GNI2010 자료

- 6개 대륙 208개 국가의 인구, 1인당 소득(GNI)

2 나무 지도

보기 1. GNI2010 자료

- 6개 대륙 208개 국가의 인구, 1인당 소득(GNI)

```
> library(treemap)
> GNI.2010 <- read.table("GNI-2010.txt",header=T)[1:104,]
> str(GNI.2010)
  'data.frame': 104 obs. of 4 variables:
  $ item      투자항목
  $ sector    투자영역
  $ principal 투자액
  $ yield     수익률
```

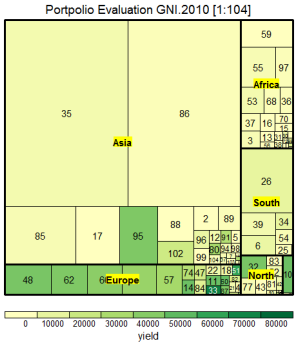
* 이하, 104개 행을 사용

2 나무 지도

R 스크립트 (1/2)

GNI.2010 자료의 나무 지도

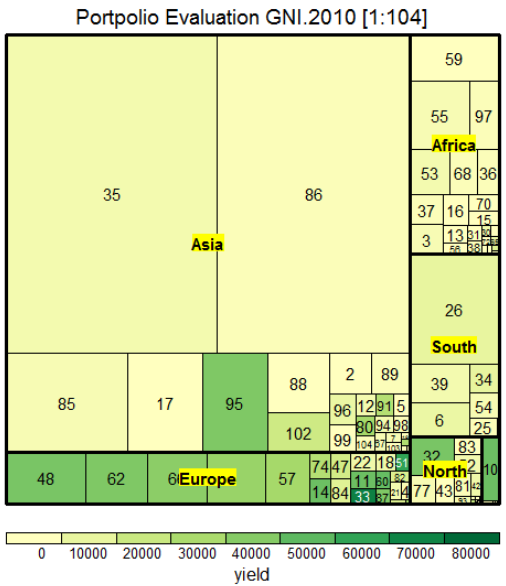
```
windows(height=8, width=7)
treemap(GNI.2010, index=c("sector", "item"),
     vSize="principal", vColor="yield",
     type="value", bg.labels="yellow",
     title="Portpolio Evaluation GNI.2010 [1:104]")
```

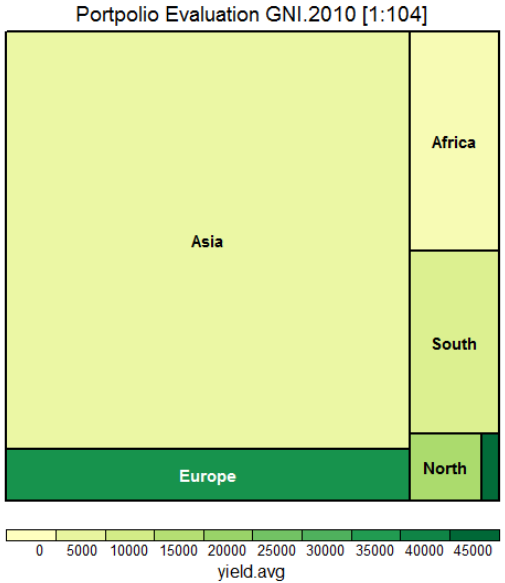


2 나무 지도

R 그래프

GNI.2010 자료의 나무 지도





R 스크립트 (2/2)

GNI.2010 자료의 나무 지도 (앞 쪽의 오른쪽 그림)

```
GNI.2010$yield.total <- GNI.2010$principal*as.numeric(GNI.2010$yield)
GNI.2010.a <- aggregate(GNI.2010[,3:5],by=list(GNI.2010$sector),sum)
GNI.2010.a$yield.avg <- GNI.2010.a$yield.total/GNI.2010.a$principal
windows(height=8, width=7)
treemap(GNI.2010.a,index=c("Group.1"),vSize="principal",
    vColor="yield.avg",type="value",bg.labels="yellow",
    title="Portpolio Evaluation GNI.2010 [1:104]")
```

정리

- 모자이크 플롯 : 교차표를 시각화한다.
  - 결합 칸의 빈도가 모자이크 타일의 크기로 반영된다.

- 나무 지도 : 포트폴리오 자료를 시각화한다.
  - 규모적 특성 (투자액) → 타일의 크기
  - 밀도적 속성 (수익률) → 타일의 컬러
  - 계층적 자료의 표현