# 다변량통계분석 Practice 1

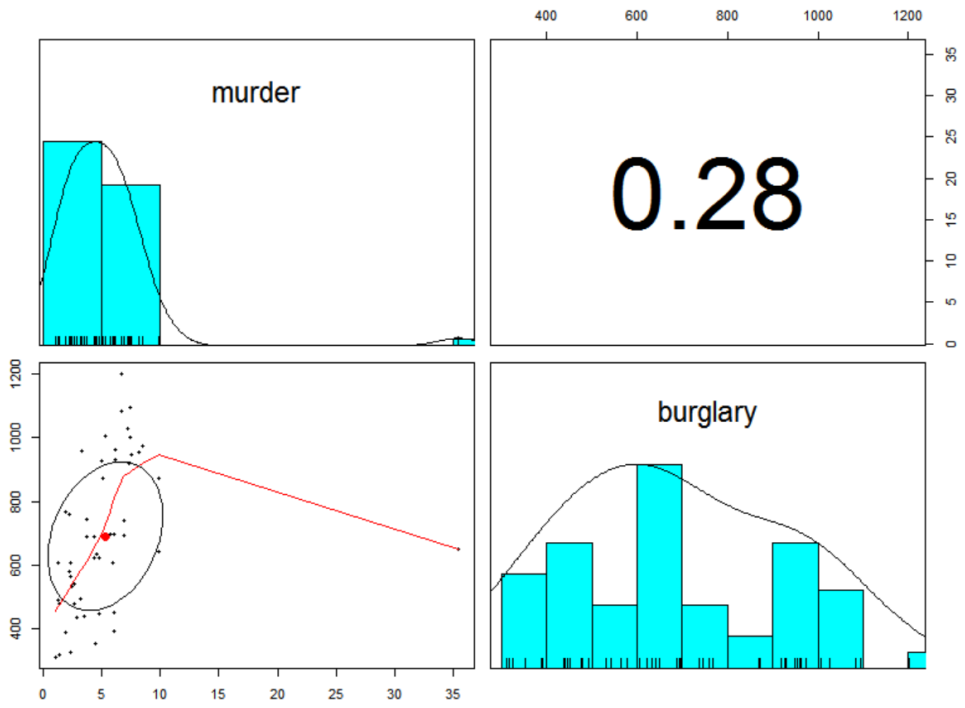
빅데이터경영MBA / U2016040 / 김우현

Crime.csv는 2005년 미국의 범죄율 데이터로 범죄 유형별 발생건을 인구 100,000명 중의 발생 비율로 표시하였다. 살인, 강도, 폭행, 절도 등 총 7가지 범죄를 포함하는데 이 중 살인(murder)와 절도(burglary) 사이의 관계를 살피려고 한다.

```
data <- read.csv("crime.csv", header = T)
head(data, 10)

crime <- data[-1, ]                    # United States (미국 전체) 제거
rownames(crime) <- c(1:nrow(crime))
head(crime, 10)
```

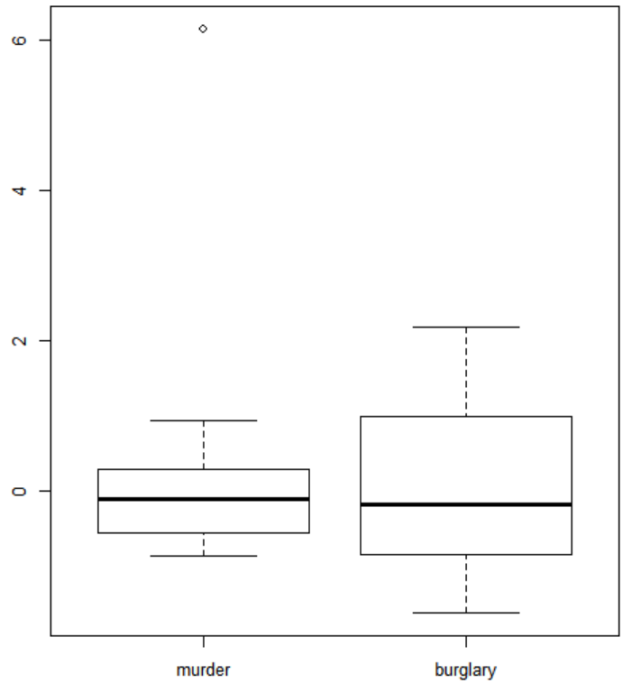
A. 두 변수 사이의 산점도를 단변량 분포와 함께 그리시오. 상관계수도 함께 살피시오.

```
library(psych)
pairs.panels(cdata)
cor.test(cdata$murder, cdata$burglary)   # cor = 0.28
```

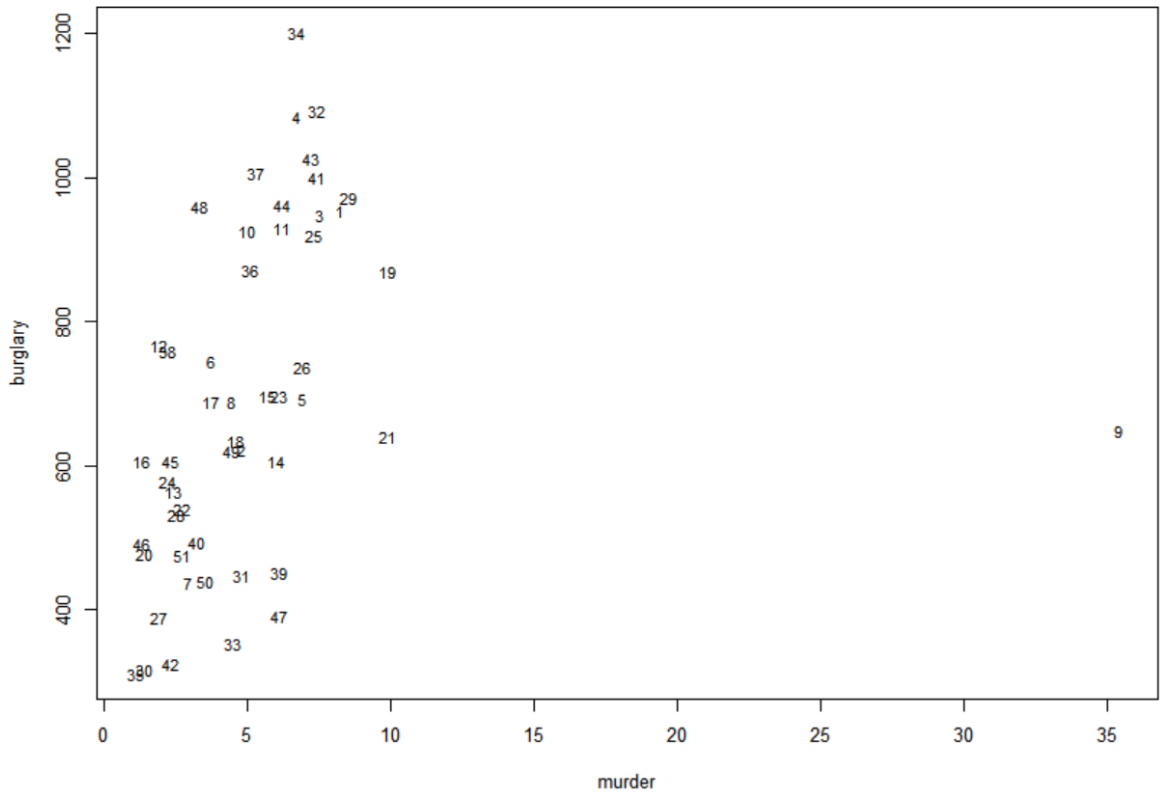


B. 위를 통해 이상점 존재여부를 판단하고 존재한다면 해당 주를 확인하고 제거하시오. 제거 후 변수들 사이의 관계가 어떻게 변화하는지 살피시오.

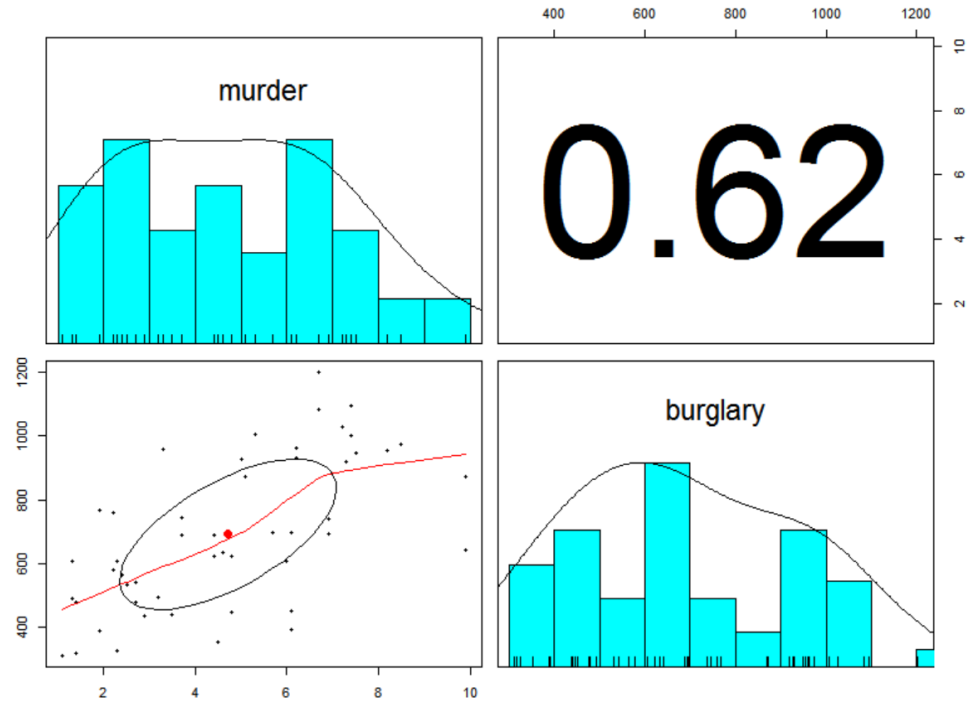
```
cdata <- crime[ , c("murder", "burglary")]
head(cdata)

# 발생비율의 단위가 다르므로 비교를 위해 scaling 한다.
boxplot(scale(cdata))   # murder에 특이하게 높은 이상점 존재
```




```
# 사분위수를 이용한 outlier 확인
# fivenum : minimum, lower-hinge, median, upper-hinge, maximum
a <- cdata$murder
which(a > fivenum(a)[4] + 1.5*IQR(a))   # outlier = 9

# 산점도를 통한 outlier 확인
cdata <- crime[ , c("state", "murder", "burglary")]
head(cdata, 10)

plot(burglary ~ murder, data = cdata, cex = 0.8, type = "n")
text(cdata$murder, cdata$burglary, cex = 0.8, labels = rownames(cdata))   # outlier = 9
```

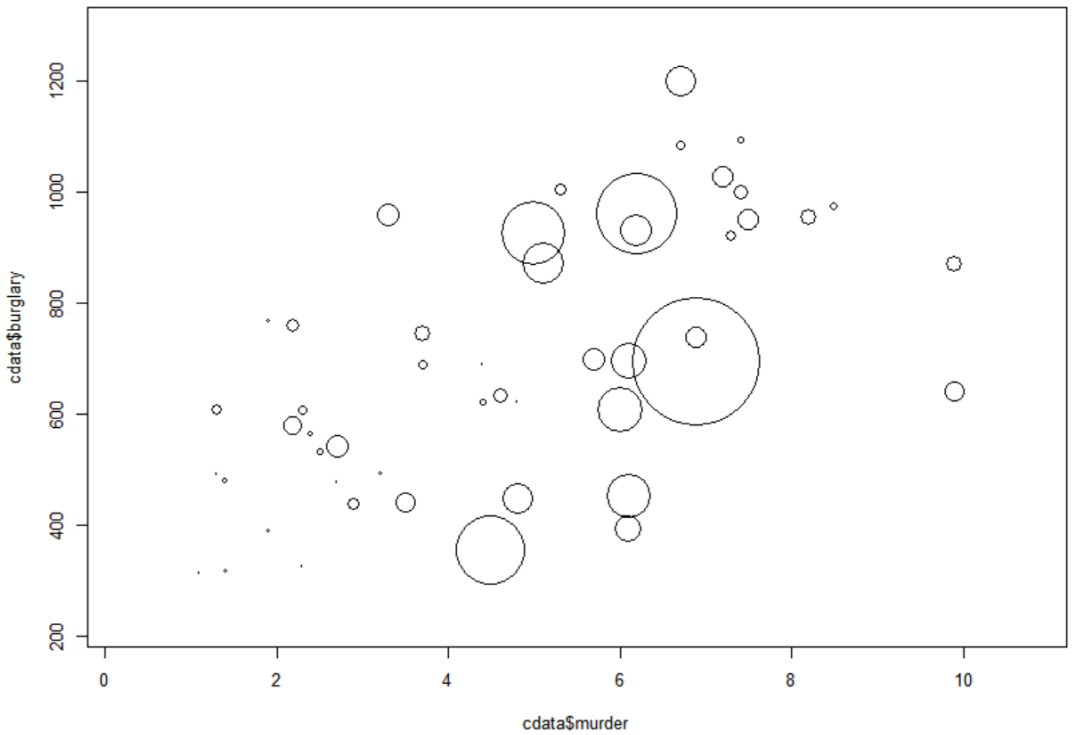



```
outlier = 9
cdata[outlier, c("state", "murder", "burglary")]   # District of Columbia

# 이상점(outlier)
                  state           murder   burglary
 9 District of Columbia           35.4     649.7

# outlier 제거 후 상관계수 확인
cdata <- cdata[-outlier, ]
head(cdata, 10)

cor.test(cdata$murder, cdata$burglary)
pairs.panels(cdata[ , c(2:3)])          # cor = 0.62 상관계수 증가
```




C. 살인, 절도와 인구(population)의 관계를 함께 관찰하기 위해 bubble plot을 그리고 관찰한 사실을 기술하시오.

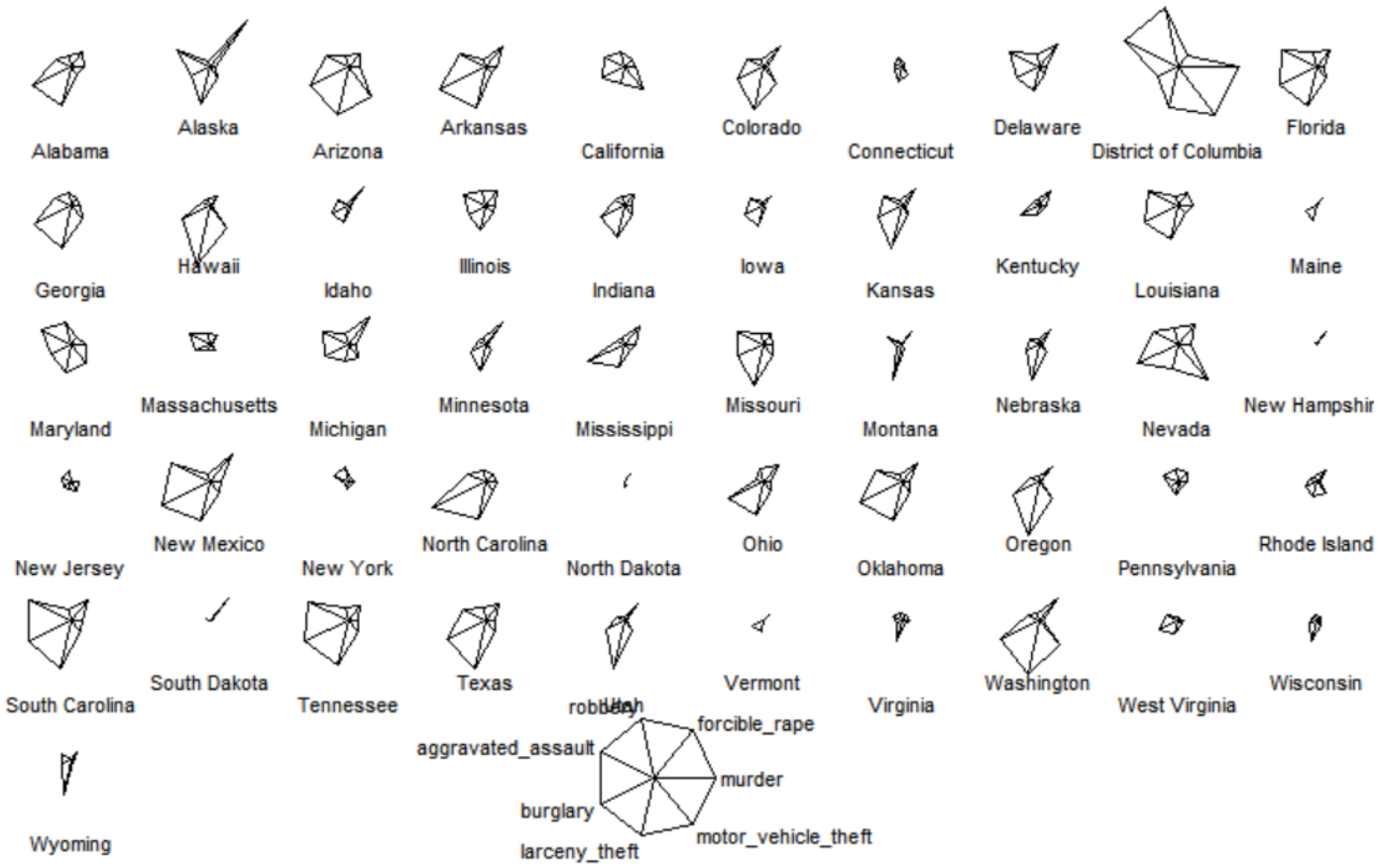
```
# outlier 제거시
cdata <- crime[-outlier, c("state", "murder", "burglary", "population")]
symbols(cdata$murder, cdata$burglary, circles = cdata$population, inches = 0.5)
```



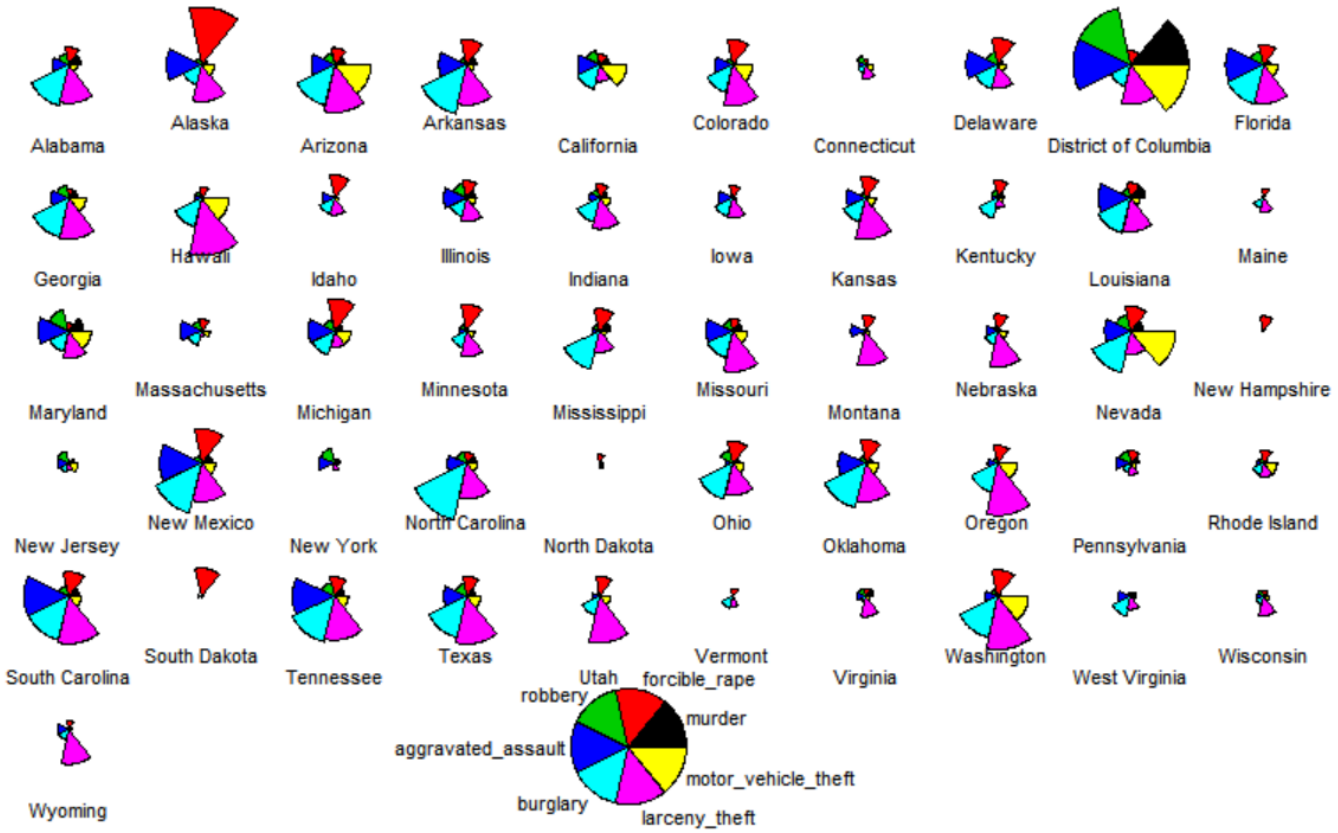

살인과 절도 발생률은 어느 정도 상관관계가 있으나, 인구와는 상관관계가 없다.

D. 7 가지 범죄의 발생 건수를 heatmap, 별그림, 나이팅게일 차트로 표현하고 범죄 발생 특징 간의 패턴이 비슷한 주들이 있는지 살피시오.

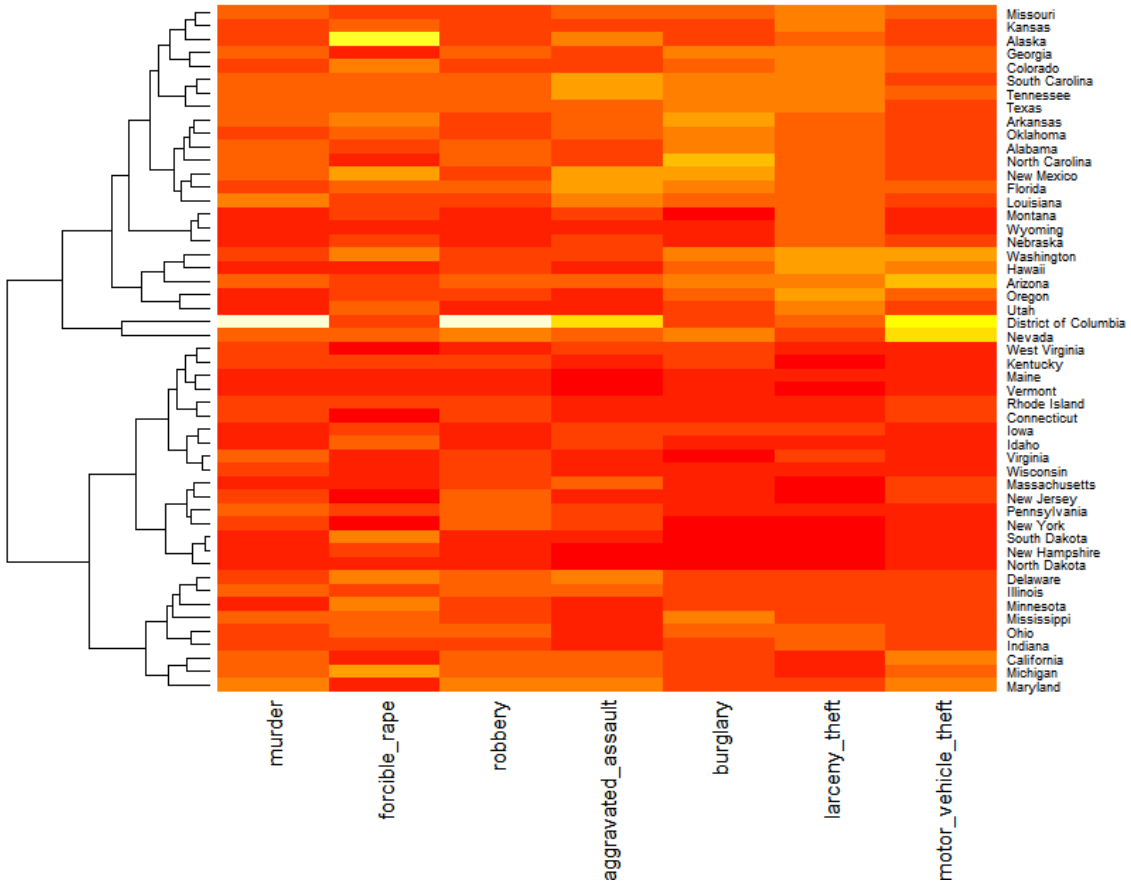
```
cdata2 <- crime
rownames(cdata2) <- cdata2$state
cdata2 <- cdata2[ , -c(1,9)]     # state 를 rowname 으로 변경. 인구 변수 제거.
head(cdata2, 10)

# star
stars(cdata2, key.loc = c(11, 2), cex = 0.8, ncol = 10)
```




```
# nightingale
stars(cdata2, key.loc = c(11, 2), cex = 0.8, draw.segments = T, ncol = 10)
```




```
# Heatmap
cdata2 <- as.matrix(cdata2)
heatmap(cdata2, scale = "column", Colv = NA, cexCol = 0.9, margins = c(8, 5))
```



미주리, 캔사스, 조지아, 콜로라도, 테네시 주 등은 larcency theft 발생이 상대적으로 다른 주보다 높은 경향을 보인다.
뉴욕, 뉴햄프셔, 사우스 다코다, 노스 다코다 주등은 다른 주들에 비해 rape를 제외한 다른 범죄발생률이 낮다.

---

R code

```
# 2005년 미국의 범죄율 데이터
# 인구 100,000명 중의 발생 비율
data <- read.csv("crime.csv", header = T)
head(data, 10)

crime <- data[-1, ]                    # United States (미국 전체) 제거
rownames(crime) <- c(1:nrow(crime))
head(crime, 10)

#---------------------------------------------------------------
# A. 살인(murder)와 절도(burglary) 사이의 산점도를 단변량 분포와 함께 그리시오.
# 상관계수도 함께 살피시오.

cdata <- crime[ , c("murder", "burglary")]
head(cdata)

# 발생비율의 단위가 다르므로 비교를 위해 scaling 한다.
boxplot(scale(cdata))   # murder 가 특이하게 높은 이상점 존재
cor.test(cdata$murder, cdata$burglary)

library(psych)
pairs.panels(cdata)     # cor = 0.28

#---------------------------------------------------------------
# B. 위를 통해 이상점 존재여부를 판단하고 존재한다면 해당 주를 확인하고 제거하시오.
# 제거 후 변수들 사이의 관계가 어떻게 변화하는지 살피시오.

# 사분위수를 이용한 outlier 확인
```

```
# fivenum : minimum, lower-hinge, median, upper-hinge, maximum
a <- cdata$murder
which(a > fivenum(a)[4] + 1.5*IQR(a))   # 9

# 산점도를 통한 outlier 확인
cdata <- crime[ , c("state", "murder", "burglary")]
head(cdata, 10)

plot(burglary ~ murder, data = cdata, cex = 0.8, type = "n")
text(cdata$murder, cdata$burglary, cex = 0.8, labels = rownames(cdata))   # outlier = 9

outlier = 9
cdata[outlier, c("state", "murder", "burglary")]   # District of Columbia

# outlier 제거 후 상관계수 확인
cdata <- cdata[-outlier, ]
head(cdata, 10)

cor.test(cdata$murder, cdata$burglary)
pairs.panels(cdata[ , c(2:3)])          # cor = 0.62 상관계수 증가

#-------------------------------------------------------------
# C. 살인, 절도와 인구(population)의 관계를 함께 관찰하기 위해 bubble plot을 그리고
# 관찰한 사실을 기술하시오.

# 원본
symbols(crime$murder, crime$burglary, circles = crime$population, inches = 0.5)

# outlier 제거시
cdata <- crime[-outlier, c("state", "murder", "burglary", "population")]
symbols(cdata$murder, cdata$burglary, circles = cdata$population, inches = 0.5)

# 살인과 절도 발생률은 어느 정도 상관관계가 있으나, 인구와는 상관관계가 없다.

#-------------------------------------------------------------
# D. 7가지 범죄의 발생 건수를 heatmap, 별그림, 나이팅게일 차트로 표현하고
# 범죄 발생 특징 간의 패턴이 비슷한 주들이 있는지 살피시오.
```

```r
cdata2 <- crime
rownames(cdata2) <- cdata2$state
cdata2 <- cdata2[ , -c(1,9)]     # state를 rowname으로 변경. 인구 변수 제거.
head(cdata2, 10)

# star & nightingale - dataframe
# star
stars(cdata2, key.loc = c(11, 2), cex = 0.8, ncol = 10)
# nightingale
stars(cdata2, key.loc = c(11, 2), cex = 0.8, draw.segments = T, ncol = 10)

# heatmap - matrix
cdata2 <- as.matrix(cdata2)
heatmap(cdata2, scale = "column", Colv = NA, cexCol = 0.9, margins = c(8, 5))
```