[프라임칼리지] 데이터시각화 - 2차 과제물(100점 만점)

정시제출기한: **25.06.02(월) 23:59**
추가연장제출: 25.06.03(화) 23:59
- [과제] 메뉴
이 름: 신승엽

□ 모든 문항에 대해 답만 쓰지 말고 코드 및 결과에 대한 자신의 해석을 반드시 포함해야 합니다.
모든 문항에 대해 **"R** 또는 **R studio** 또는 **Google Colab(R** 언어**)"**를 사용해 계산하여 1)작성한 '코드'와 2)코드를 실행한 '프로그래밍 결과(캡처이미지)'를 각각 첨부하고 3)결과에 대한 자신의 해석을 기술해주세요. - 총 4개 문항
(자신의 해석은 R코드 명령어의 동작에 대한 해석이 아닌 실행한 결과에 대한 해석을 의미합니다.)
- **R** 실습을 다루는 과제이므로 엑셀이나 공학용 계산기, 파이썬 등 다른 프로그램이나 손으로 직접 계산한 답안은 평가가 불가능합니다.
- 답안 파일은 반드시 전체 문항에 대해 하나의 파일**(PDF/HWP/DOCX)**만 제출하시기 바랍니다. (문항별로 따로 파일을 제출하거나, R 파일을 별도로 제출하지 마시기 바랍니다.)

선행 코드:

```
install.packages("ggplot2")
install.packages("maps")
install.packages("showtext")
install.packages("mapproj")
library(ggplot2)
library(maps)
library(showtext)
font_add_google("Nanum Gothic", "nanumgothic")
showtext_auto()
```

문제**1.**

R에 내장된 "USArrests" 데이터셋(별도 패키지 설치 필요 없음 – 대소문자 유의)을 활용하여 다음 작업을 수행하시오. (25점)

1) 미국 지도를 그리고 각 주의 살인율(Murder)을 지도에 표현하라.
- 살인율이 높을수록 진한 색을 사용하고, 색상은 6단계 이상의 연속적인 그라데이션으로 표시하라.
- 주 경계선과 주 이름 레이블을 지도에 추가하라.

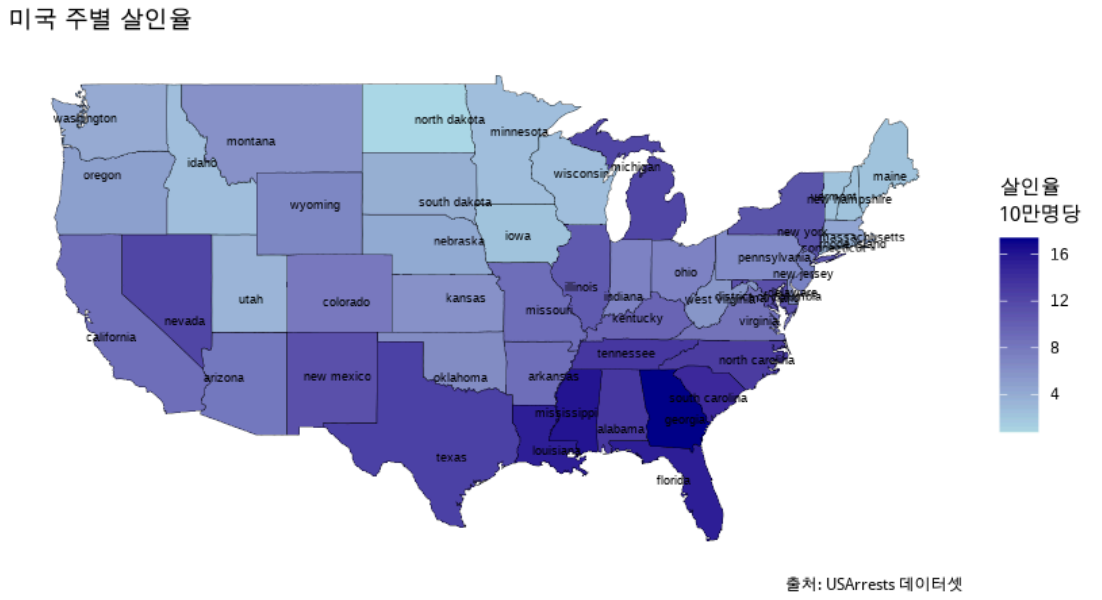
```
(* 힌트:
install.packages("ggplot2")
install.packages("maps")
library(ggplot2)
library(maps)
data("USArrests")
states_map <- map_data("state") # 주별 경계 데이터 불러오기
USArrests$state <- tolower(state.name) # 주 이름을 소문자로 변환하여 USArrests 데이터에 추가
# 주별 경계 데이터와 USArrests 데이터 병합

map_data <- merge(states_map, USArrests, by.x = "region", by.y = "state")
map_data # 병합된 데이터 확인)
```

코드:

```
# USArrests 데이터셋
data("USArrests")
```

```r
# 데이터셋의 행 이름(주 이름)을 소문자로 변환하여 state 열로 추가
USArrests$state <- tolower(rownames(USArrests))

# 미국 주 경계 데이터
states_map <- map_data("state")

# 지도 데이터와 범죄 데이터 병합
map_data <- merge(states_map, USArrests, by.x = "region", by.y =
"state")

# 그룹과 순서로 정렬
map_data <- map_data[order(map_data$group, map_data$order), ]

# 주 이름 표시를 위한 중심점 계산
state_centers <- aggregate(cbind(long, lat) ~ region, data =
states_map, FUN = mean)

# 살인율을 표시한 미국 지도 생성
ggplot() +
 # 주 경계
 geom_polygon(data = map_data,
              aes(x = long, y = lat, group = group, fill = Murder),
              color = "black", linewidth = 0.1) +
 # 주 이름 표시
 geom_text(data = state_centers,
           aes(x = long, y = lat, label = region),
           size = 2.5, inherit.aes = FALSE) +
 # 색상 그라데이션 설정
 scale_fill_gradient(low = "lightblue", high = "darkblue",
                     name = "살인율\n10만명당", n.breaks = 6) +
 coord_map() +
 labs(title = "미국 주별 살인율",
      caption = "출처: USArrests 데이터셋") +
 theme_minimal() +
 theme(axis.title = element_blank(),
       axis.text = element_blank(),
       axis.ticks = element_blank(),
       panel.grid = element_blank())
```

결과:



해석:
미국 남부 지역(조지아, 루이지애나, 앨라배마, 플로리다 등)이 가장 높은 살인율(진한 파란색)을 보이고 있다. 반면, 북부와 중서부 지역(노스다코타, 사우스다코타, 아이오와, 미네소타 등)은 상대적으로 낮은 살인율(연한 파란색)을 나타낸다. 가장 높은 살인율을 보이는 지역은 대략 15-16명/10만 명 수준으로 가장 낮은 지역의 약 4배에 달한다. 미시간, 뉴욕, 캘리포니아와 같이 대도시가 많은 주들이 중간~높은 수준의 살인율을 보이는 경향이 있다. 지도는 단순히 살인율만 보여주지만, 이러한 패턴 뒤에는 사회경제적 불평등, 역사적 요인, 총기 소유 관련 법률 등 복합적인 요인들이 작용했을 가능성이 크다. 이러한 지역적 차이는 당시 미국 사회의 불균등한 발전과 다양한 사회적 요인들을 반영하는 것으로 보인다.

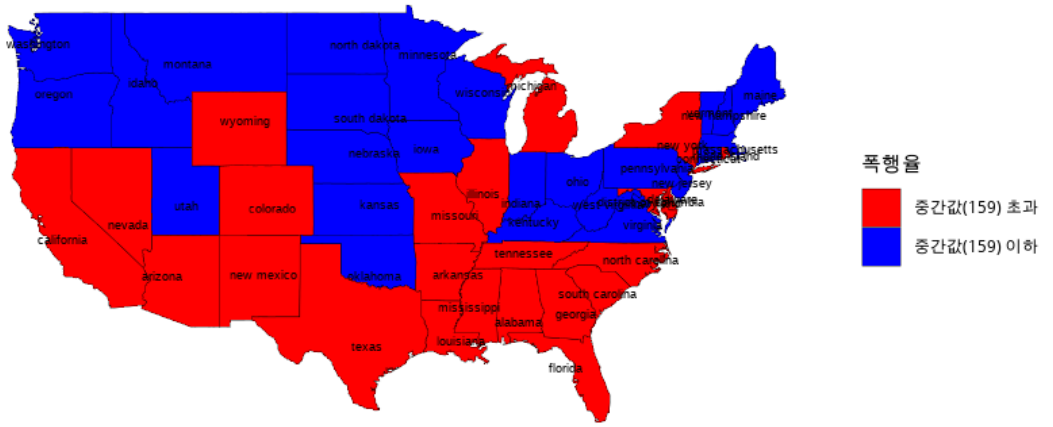
2) 전국 폭행율(Assault)의 중간값을 계산하고, 이를 초과하는 주와 그 이하인 주를 구분하여 지도에 다르게 표시하라.
- 폭행율이 전국 폭행율의 '중간값'을 초과하는 주는 빨간색으로, 그 외 주는 파란색으로 표현하라.

코드:
```
# 폭행율(Assault)의 중간값 계산
assault_median <- median(USArrests$Assault)
```

```r
# 중간값을 초과하는지 여부에 따라 그룹 변수 추가
USArrests$assault_group <- ifelse(USArrests$Assault > assault_median, "high", "low")

# 지도 데이터와 범죄 데이터 병합
map_data <- merge(states_map, USArrests, by.x = "region", by.y = "state")

# 그룹과 순서로 정렬
map_data <- map_data[order(map_data$group, map_data$order), ]

# 폭행율 중간값 기준 미국 지도 생성
ggplot() +
 # 주 경계
 geom_polygon(data = map_data,
                  aes(x = long, y = lat, group = group, fill = assault_group),
              color = "black", linewidth = 0.1) +
 # 주 이름 표시
 geom_text(data = state_centers,
           aes(x = long, y = lat, label = region),
           size = 2.5, inherit.aes = FALSE) +
 # 색상 설정
 scale_fill_manual(name = "폭행율",
                   values = c("high" = "red", "low" = "blue"),
                              labels = c("high" = paste0("중간값(", assault_median, ") 초과"),
                                         "low" = paste0("중간값(", assault_median, ") 이하"))) +
 coord_map() +
 labs(title = "미국 주별 폭행율") +
 theme_minimal() +
 theme(axis.title = element_blank(),
       axis.text = element_blank(),
       axis.ticks = element_blank(),
       panel.grid = element_blank())
```

결과:

미국 주별 폭행율



해석:
빨간색은 폭행율이 중간값을 초과하는 주를, 파란색은 중간값 이하인 주를 나타낸다. 지도를 분석해보면 미국 남부 지역(텍사스, 플로리다, 조지아, 루이지애나 등)과 서부 해안 지역(캘리포니아, 네바다)이 주로 중간값보다 높은 폭행율을 보이고 있다. 반면 북부와 북동부 지역의 여러 주들(워싱턴, 오레곤, 메인, 버몬트 등)은 상대적으로 낮은 폭행율을 나타내고 있다.
특히 주목할 점은 살인율 지도와 유사한 지역적 패턴이 관찰된다는 것이다. 이는 폭력 범죄가 특정 지역에 집중되는 경향이 있음을 시사한다. 미시간, 뉴욕 등 대도시가 많은 일부 북부 주들도 높은 폭행율을 보이고 있어, 도시화와 범죄율 간의 연관성을 추측해볼 수 있다. 이러한 지역별 차이는 사회경제적 요인, 인구 밀도, 법 집행 정책 등 다양한 변수에 의해 영향을 받았을 것이다.

3) USArrests 데이터셋에서 살인(Murder), 폭행(Assault), 도시인구(UrbanPop), 강간(Rape) 각각에 대해 범죄율이 가장 높은 상위 5개 주와 가장 낮은 하위 5개 주를 추출하여 추출된 데이터를 이용해서 각 변수별로 막대 그래프를 그리고, 해당 주의 특징을 분석하라.
[R code]
USArrests

```
> 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
```

코드:

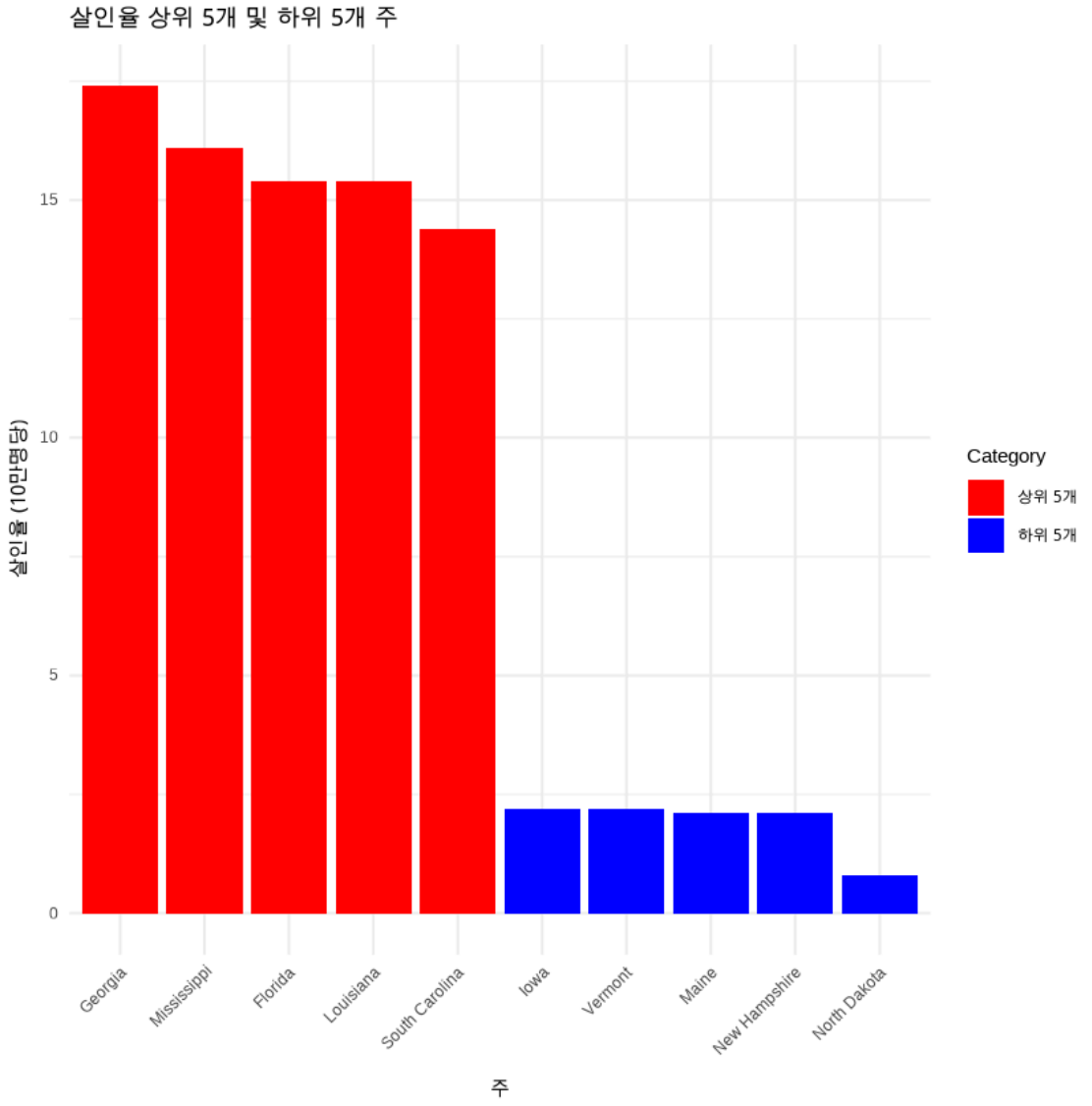
```r
# 필요한 패키지 로드
library(ggplot2)
library(dplyr)

# USArrests 데이터셋 로드
data("USArrests")

# 데이터프레임에 주 이름을 열로 추가
USArrests$State <- rownames(USArrests)

# 1. 각 변수별로 상위 5개, 하위 5개 주 추출

# 살인(Murder) 상위 5개, 하위 5개 주
murder_high <- USArrests %>%
 arrange(desc(Murder)) %>%
 head(5)

murder_low <- USArrests %>%
 arrange(Murder) %>%
 head(5)

murder_data <- rbind(murder_high, murder_low)
murder_data$Category <- c(rep("상위 5개", 5), rep("하위 5개", 5))
murder_data$State    <-    factor(murder_data$State,    levels    =
murder_data$State[order(murder_data$Murder, decreasing = TRUE)])

# 폭행(Assault) 상위 5개, 하위 5개 주
assault_high <- USArrests %>%
 arrange(desc(Assault)) %>%
 head(5)

assault_low <- USArrests %>%
 arrange(Assault) %>%
 head(5)

assault_data <- rbind(assault_high, assault_low)
assault_data$Category <- c(rep("상위 5개", 5), rep("하위 5개", 5))
```

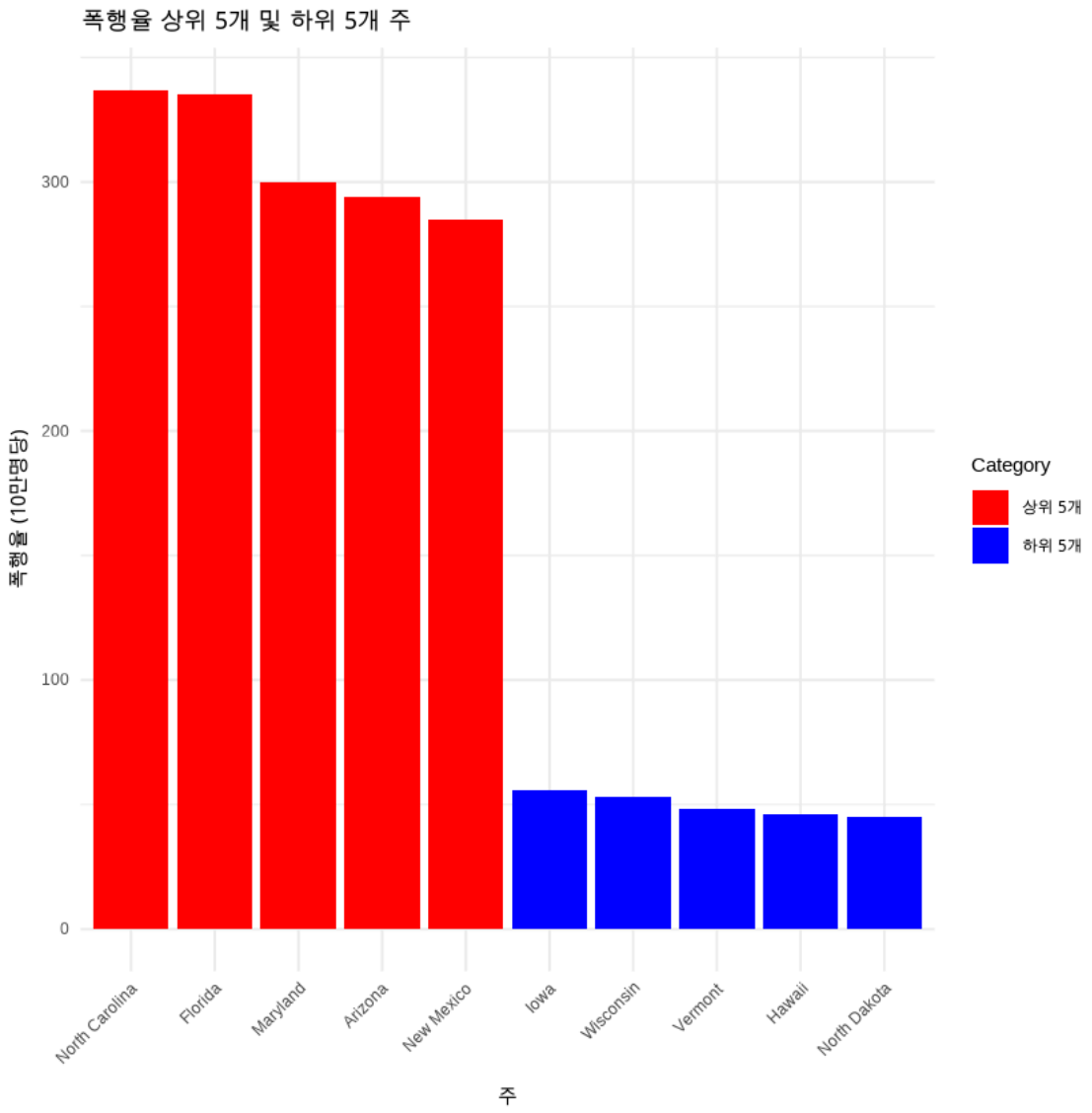
```r
assault_data$State <- factor(assault_data$State, levels =
assault_data$State[order(assault_data$Assault, decreasing = TRUE)])

# 도시인구(UrbanPop) 상위 5개, 하위 5개 주
urbanpop_high <- USArrests %>%
 arrange(desc(UrbanPop)) %>%
 head(5)

urbanpop_low <- USArrests %>%
 arrange(UrbanPop) %>%
 head(5)

urbanpop_data <- rbind(urbanpop_high, urbanpop_low)
urbanpop_data$Category <- c(rep("상위 5개", 5), rep("하위 5개", 5))
urbanpop_data$State <- factor(urbanpop_data$State, levels =
urbanpop_data$State[order(urbanpop_data$UrbanPop, decreasing =
TRUE)])

# 강간(Rape) 상위 5개, 하위 5개 주
rape_high <- USArrests %>%
 arrange(desc(Rape)) %>%
 head(5)

rape_low <- USArrests %>%
 arrange(Rape) %>%
 head(5)

rape_data <- rbind(rape_high, rape_low)
rape_data$Category <- c(rep("상위 5개", 5), rep("하위 5개", 5))
rape_data$State <- factor(rape_data$State, levels =
rape_data$State[order(rape_data$Rape, decreasing = TRUE)])

# 2. 각 변수별 막대 그래프 생성

# 살인(Murder) 막대 그래프
murder_plot <- ggplot(murder_data, aes(x = State, y = Murder, fill
= Category)) +
 geom_bar(stat = "identity") +
  scale_fill_manual(values = c("상위 5개" = "red", "하위 5개" =
"blue")) +
 labs(title = "살인율 상위 5개 및 하위 5개 주",
```

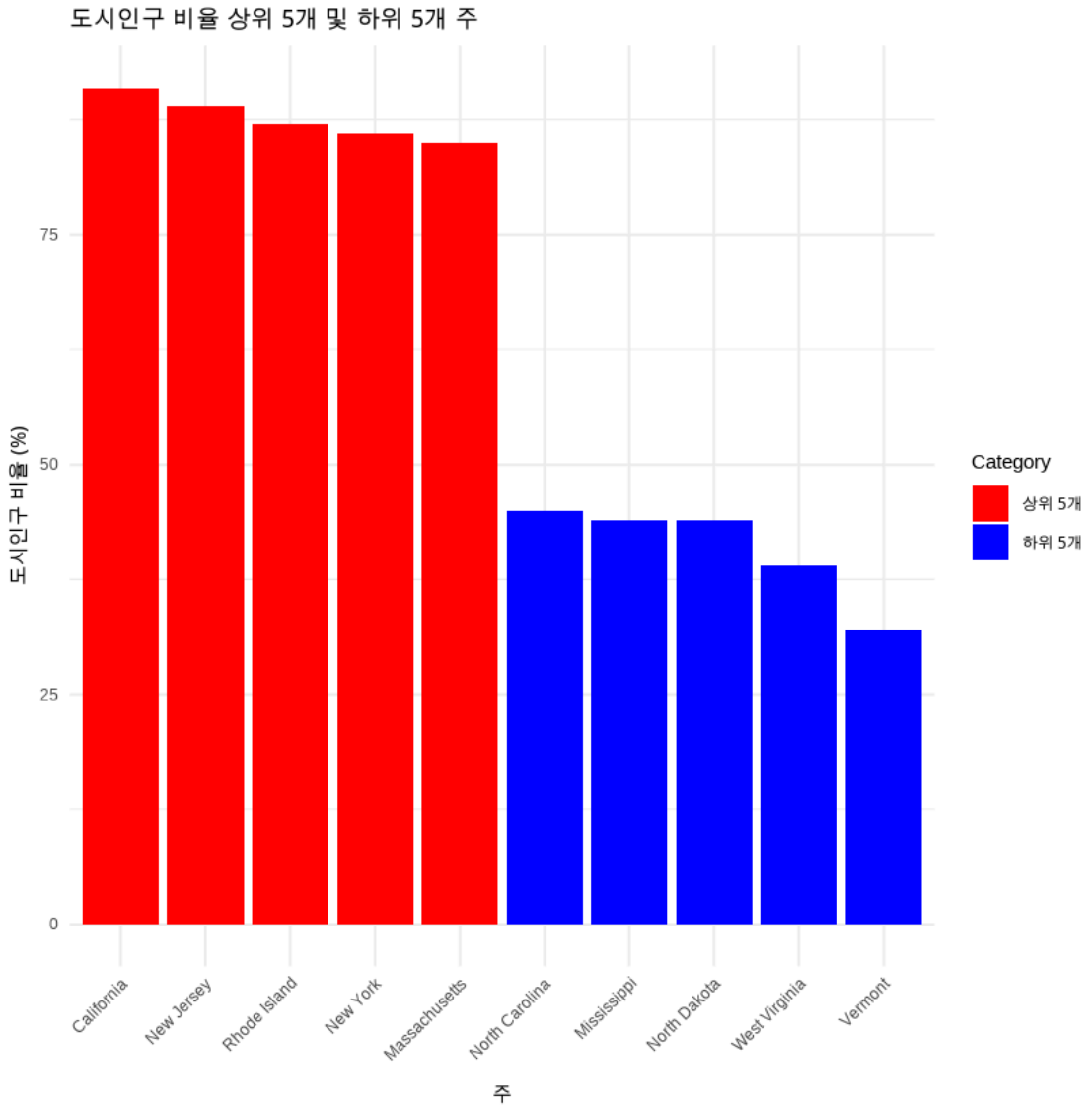
```
      x = "주",
      y = "살인율 (10만명당)") +
 theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))

# 폭행(Assault) 막대 그래프
assault_plot <- ggplot(assault_data, aes(x = State, y = Assault,
fill = Category)) +
 geom_bar(stat = "identity") +
  scale_fill_manual(values = c("상위 5개" = "red", "하위 5개" =
"blue")) +
 labs(title = "폭행율 상위 5개 및 하위 5개 주",
      x = "주",
      y = "폭행율 (10만명당)") +
 theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))

# 도시인구(UrbanPop) 막대 그래프
urbanpop_plot <- ggplot(urbanpop_data, aes(x = State, y = UrbanPop,
fill = Category)) +
 geom_bar(stat = "identity") +
  scale_fill_manual(values = c("상위 5개" = "red", "하위 5개" =
"blue")) +
 labs(title = "도시인구 비율 상위 5개 및 하위 5개 주",
      x = "주",
      y = "도시인구 비율 (%)") +
 theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))

# 강간(Rape) 막대 그래프
rape_plot <- ggplot(rape_data, aes(x = State, y = Rape, fill =
Category)) +
 geom_bar(stat = "identity") +
  scale_fill_manual(values = c("상위 5개" = "red", "하위 5개" =
"blue")) +
 labs(title = "강간율 상위 5개 및 하위 5개 주",
      x = "주",
      y = "강간율 (10만명당)") +
 theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

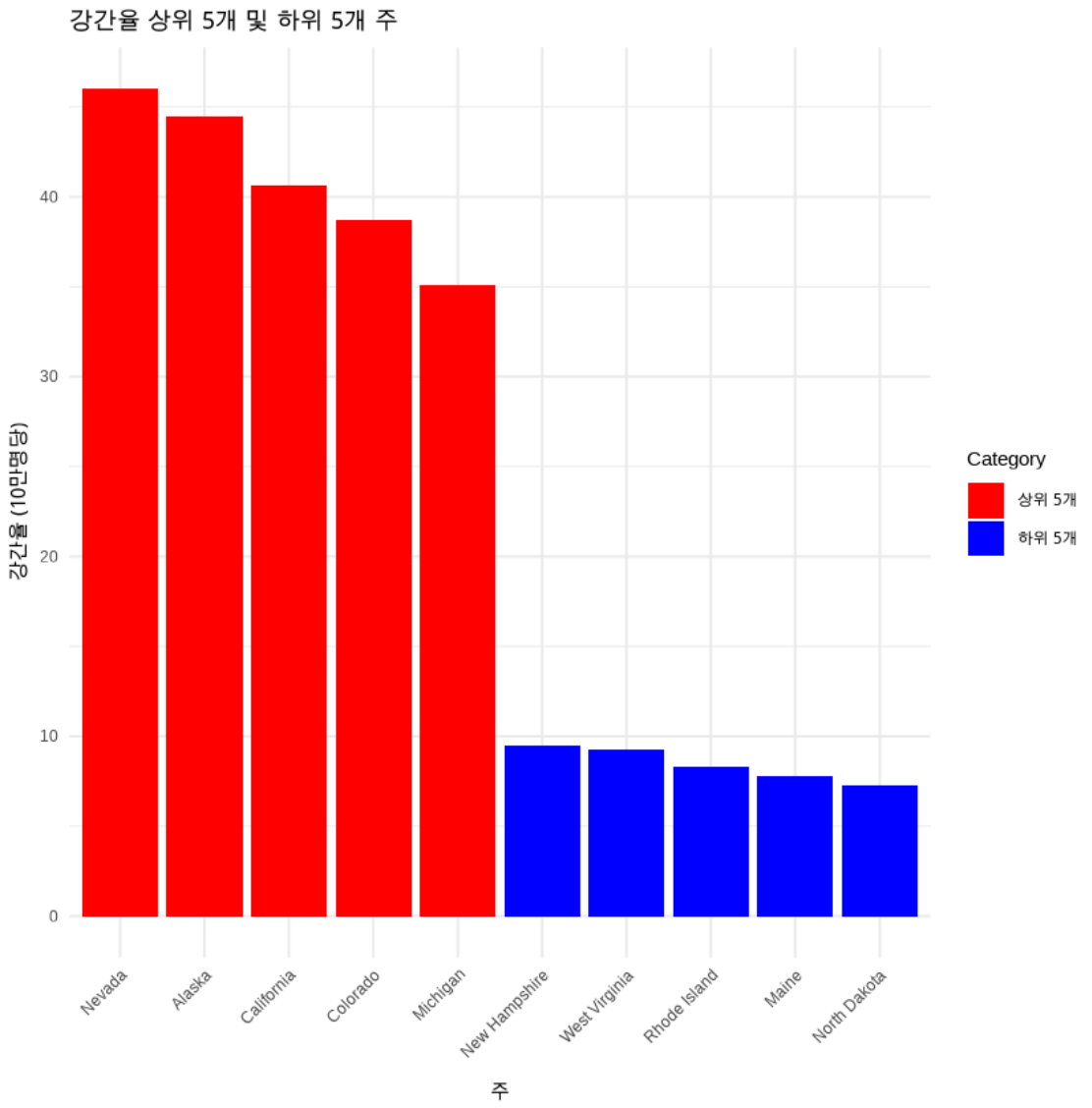
```r
# 3. 그래프 출력
print(murder_plot)
print(assault_plot)
print(urbanpop_plot)
print(rape_plot)
```

결과:




폭행율 상위 5개 및 하위 5개 주
폭행율 (10만명당)
300
200
100
0
North Carolina
Florida
Maryland
Arizona
New Mexico
Iowa
Wisconsin
Vermont
Hawaii
North Dakota
주
Category
상위 5개
하위 5개


도시인구 비율 상위 5개 및 하위 5개 주
도시인구 비율 (%)
75
50
25
0
California
New Jersey
Rhode Island
New York
Massachusetts
North Carolina
Mississippi
North Dakota
West Virginia
Vermont
주
Category
상위 5개
하위 5개



해석:
살인율(Murder) 분석:
살인율이 가장 높은 주는 조지아(17.4), 미시시피, 플로리다, 루이지애나, 사우스캐롤라이나 순으로 모두 남부 지역에 집중되어 있다. 반면 가장 낮은 주는 노스다코타(0.8), 뉴햄프셔, 메인, 버몬트, 아이오와로 대부분 북부 및 북동부에 위치하고 있다. 상위 5개 주의 살인율은 하위 5개 주보다 약 7-20배 높은 수준이다.

폭행율(Assault) 분석:
폭행율은 노스캐롤라이나(337), 플로리다, 메릴랜드, 애리조나, 뉴멕시코가 가장 높으며, 역시 대부분 남부와 서부 지역에 분포한다. 가장 낮은 주는 노스다코타(45), 하와이, 버몬트, 위스콘신, 아이오와로 주로 북부와 일부 중서부 지역이다. 상위 지역의 폭행율은 하위 지역보다 약 6-7배 높다.

도시인구 비율(UrbanPop) 분석:
도시인구 비율이 가장 높은 주는 캘리포니아(91%), 뉴저지, 로드아일랜드, 뉴욕, 매사추세츠로 주로 동부 해안과 서부 해안의 산업화된 지역이다. 가장 낮은 주는 버몬트(32%), 웨스트버지니아, 노스다코타, 미시시피, 노스캐롤라이나로 주로 농촌 지역이 많은 곳들이다. 도시화율의 상위 주와

하위 주 간 격차는 약 2-3배로 다른 범죄 지표들보다는 상대적으로 작다.

강간율(Rape) 분석:
강간율이 가장 높은 주는 네바다(46), 알래스카, 캘리포니아, 콜로라도, 미시간 순으로 서부와 중서부 지역에 분포한다. 가장 낮은 주는 노스다코타(7.3), 메인, 로드아일랜드, 웨스트버지니아, 뉴햄프셔로 주로 북동부 지역이다. 상위 지역의 강간율은 하위 지역보다 약 5-6배 높다.

문제**2.**
R 패키지 “datarium”(패키지 별도 설치 필요! 모두 소문자)에 내장된 marketing(모두 소문자) 데이터셋은 광고 미디어에 사용한 비용과 판매액의 데이터이다. facebook 컬럼은 facebook 광고비로 사용한 금액이고, sales 컬럼은 판매액이다.

1) facebook을 x축, sales를 y축으로 하는 산점도를 그리시오.
2) facebook을 독립변수(설명변수), sales를 종속변수(반응변수, 결과변수)로 하는 회귀직선을 산점도 위에 그리고 산점도의 제목으로 본인의 학번을 출력하시오. (25점)
(힌트: datarium 패키지를 설치, 로드한 후 콘솔에 ‘dat <- marketing’ 을 입력하여 실행하면 marketing 데이터셋이 dat라는 변수에 저장된다.)
[R code]

```
install.packages("datarium")
library("datarium")
dat <- marketing
dat
```

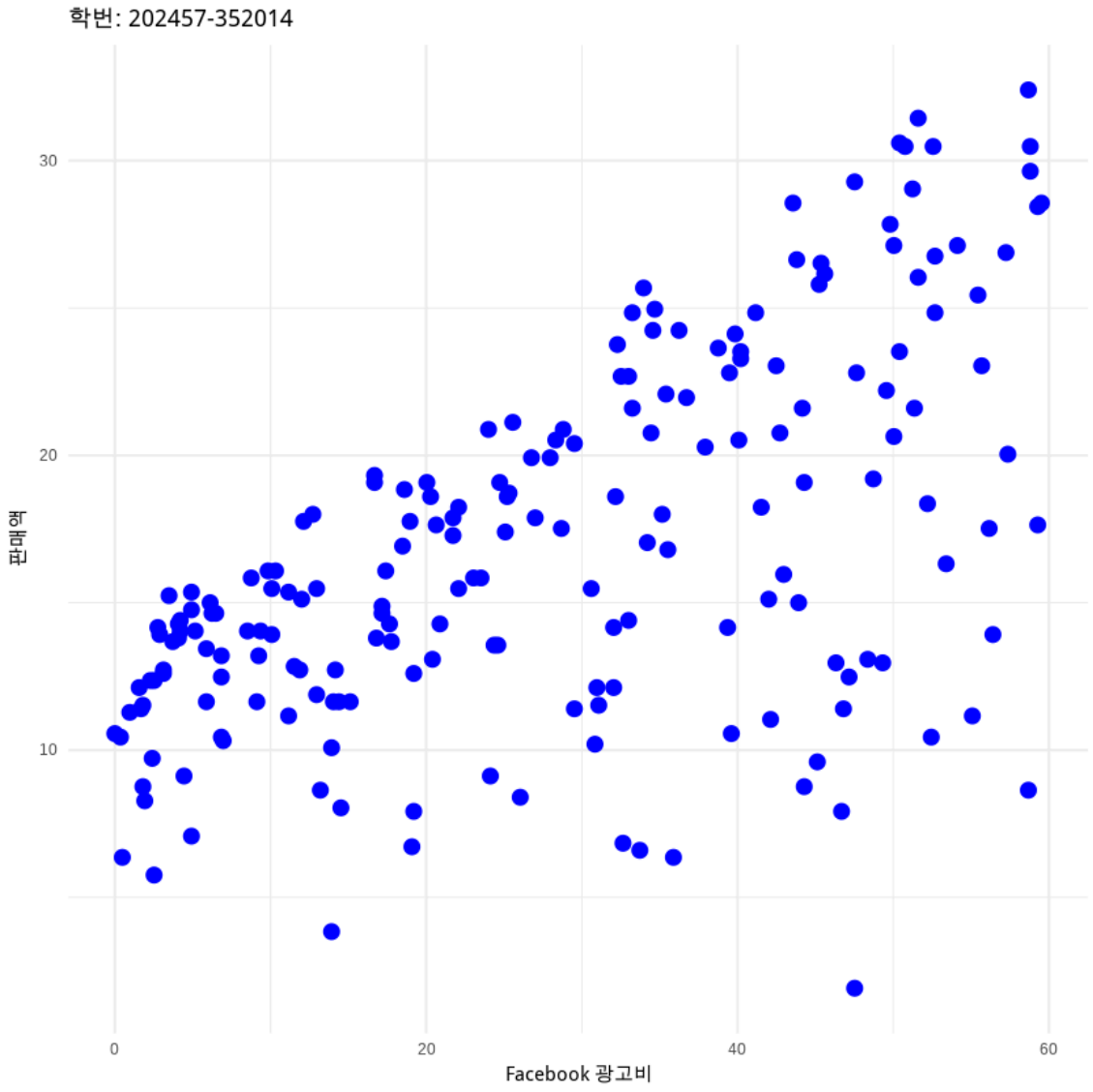
```
> library("datarium")
> dat <- marketing
> dat
    youtube facebook newspaper sales
1    276.12    45.36     83.04 26.52
2     53.40    47.16     54.12 12.48
3     20.64    55.08     83.16 11.16
```

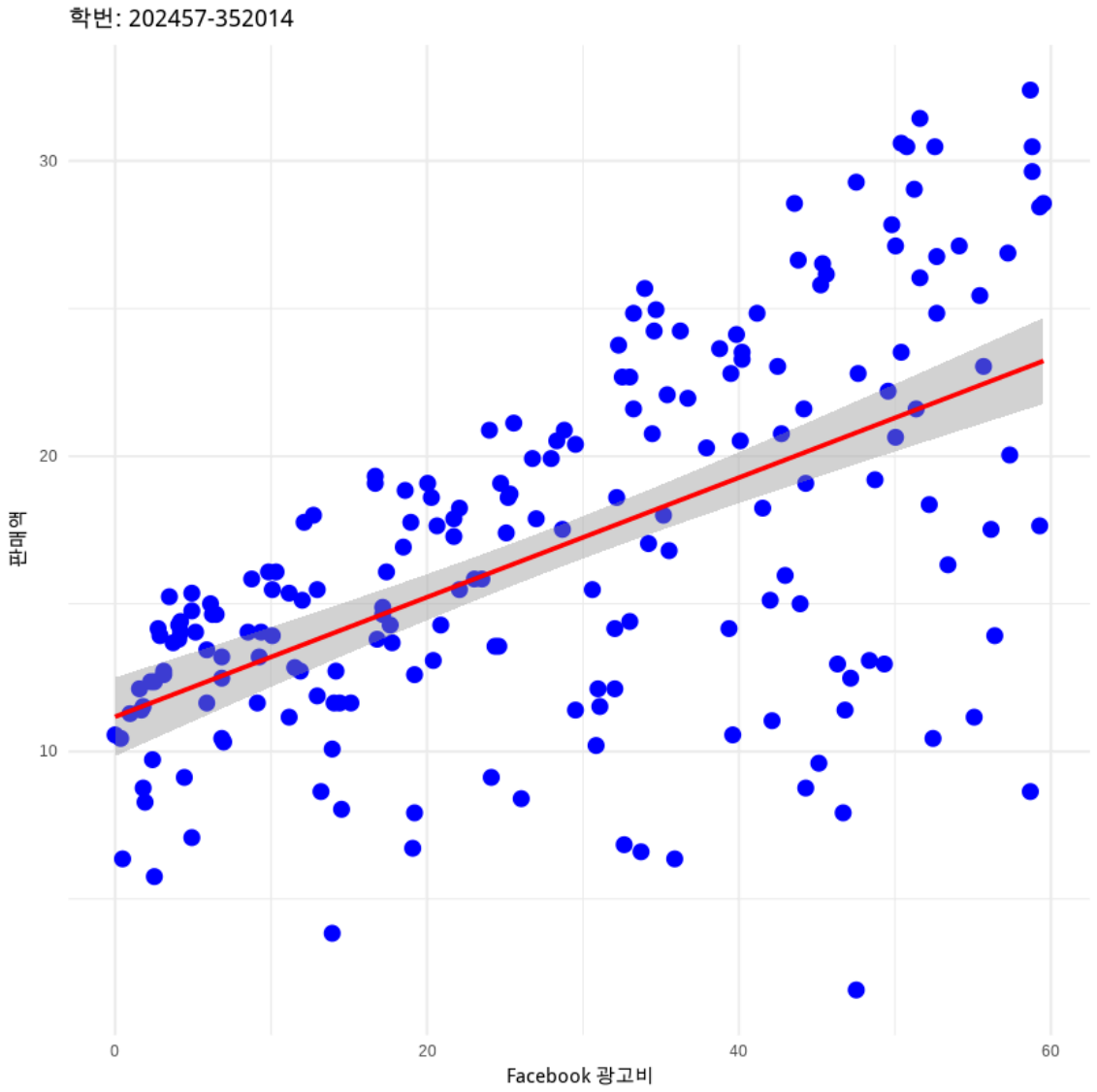
코드:

```
# 필요한 패키지 설치 및 로드
install.packages("datarium")
library(datarium)
library(ggplot2)

# marketing 데이터셋 불러오기
dat <- marketing

# 데이터 확인
head(dat)

# 1) facebook을 x축, sales를 y축으로 하는 산점도
plot1 <- ggplot(dat, aes(x = facebook, y = sales)) +
 geom_point(color = "blue", size = 3) +
 labs(x = "Facebook 광고비", y = "판매액",
      title = "학번: 202457-352014") +
 theme_minimal()
```

```
# 산점도 출력
print(plot1)
# 2) facebook을 독립변수(설명변수), sales를 종속변수(반응변수, 결과변수)로 하는 회귀직선을 산점도 위에 그리고 산점도의 제목으로 본인의 학번을 출력하시오
# 회귀모델 생성
model <- lm(sales ~ facebook, data = dat)

# 회귀직선이 추가된 산점도 생성
plot2 <- ggplot(dat, aes(x = facebook, y = sales)) +
 geom_point(color = "blue", size = 3) +
 geom_smooth(method = "lm", formula = y ~ x, color = "red", se = TRUE) +
 labs(x = "Facebook 광고비", y = "판매액",
      title = "학번: 202457-352014") +
 theme_minimal()

# 최종 산점도 출력
print(plot2)
```

결과:


학번: 202457-352014
30
20
10
판매량
0
20
40
60
Facebook 광고비



해석:

1)
위 산점도는 Facebook 광고비와 판매액 간의 관계를 보여주는 그래프이다. Facebook 광고비가 증가함에 따라 판매액도 전반적으로 증가하는 경향을 보이고 있다. 광고비가 낮은 영역(0-20)에서는 판매액이 대체로 10-15 사이에 분포하는 반면, 광고비가 높은 영역(50-60)에서는 판매액이 20-30 범위로 더 높게 나타나는 경향이 있다. 그러나 같은 광고비 수준에서도 판매액의 편차가 크게 나타나는데, 이는 Facebook 광고비 외에도 판매액에 영향을 미치는 다른 요인들이 존재함을 시사한다.

2)
첫 번째 산점도에 선형 회귀분석 결과를 추가한 것으로, 회귀식은 sales = 11.17 + 0.2 × facebook이다. 빨간색 회귀선은 두 변수 간의 평균적인 선형 관계를 나타내며, Facebook 광고비가 1단위 증가할 때마다 판매액은 평균적으로 0.2단위 증가한다고 해석할 수 있다. 회귀선 주변의 회색 영역은 95% 신뢰구간을 나타내는데, 이 구간이 좁다는 것은 회귀계수 추정의 정확도가 상대적으로 높음을 의미한다. 그러나 많은 데이터 포인트들이 회귀선에서 상당히 벗어나 있어, 이 모델의 설명력은 제한적이며 Facebook 광고비 외에도 판매액에 영향을 미치는 다른 중요한

요인들이 있음을 알 수 있다.

문제**3.**
R 패키지 “vcd”(패키지 별도 설치 필요! 모두 소문자)에 내장된 “Arthritis”(오타 및 대소문자 구분 유의) 데이터셋은 류마티스 관절염 환자를 대상으로 한 임상시험 결과 데이터이다. 각 행은 각 환자를 나타내며, 변수 Treatment는 그룹(Treated = 새로운 치료제를 투약한 그룹, Placebo = 위약을 받은 그룹)을 나타낸다. 변수 Sex는 성별을, Improved는 치료 결과(None = 차도 없음, Some = 약간 좋아짐, Marked = 매우 좋아짐)를 나타낸다. 새로운 치료제 투약 여부가 치료 결과와 연관이 있는지, 성별과 치료 결과 간에 연관이 있는지를 데이터 시각화를 통해서 탐구하시오. (참고: barplot 또는 누적 barplot 사용) (25점)
[R code]
install.packages("vcd")
library("vcd")
Arthritis

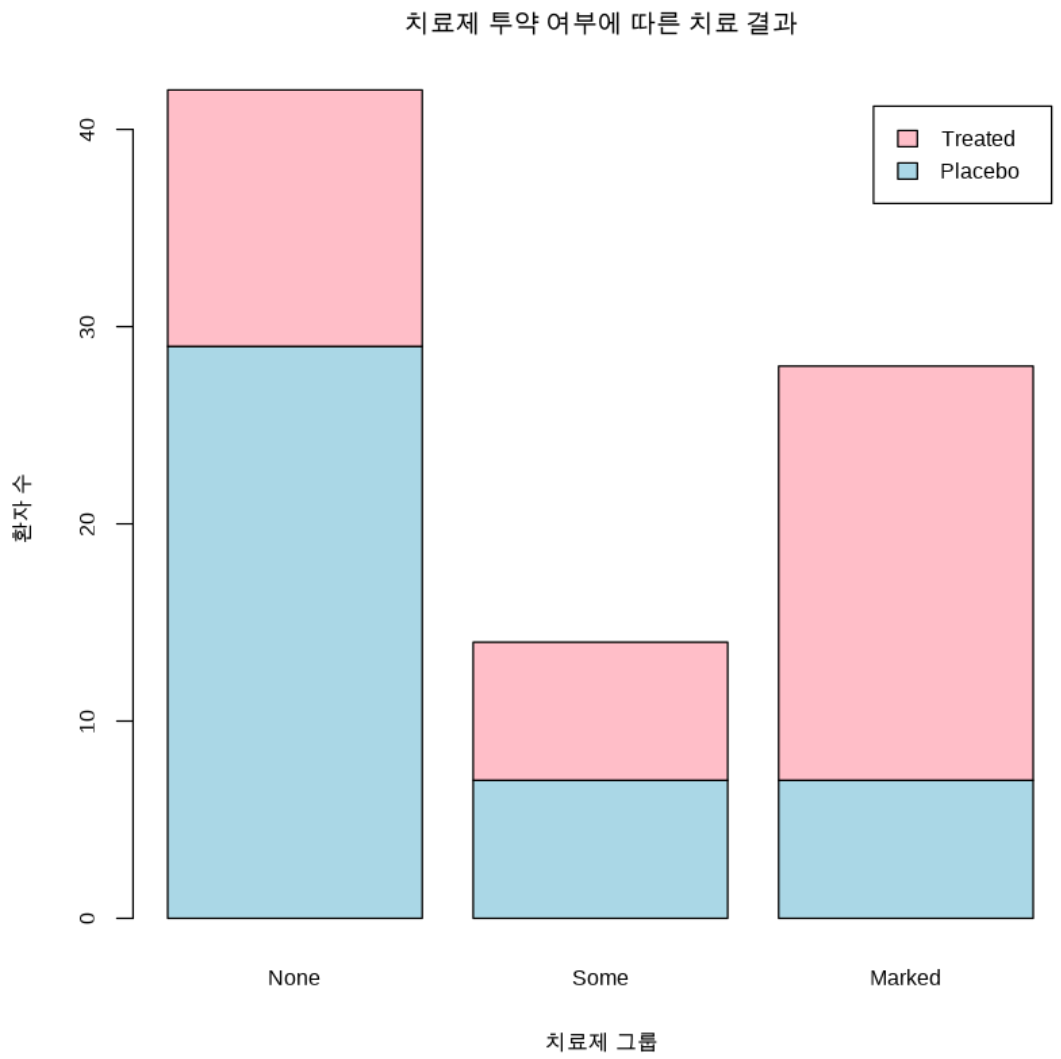
```
> library("vcd")
필요한 패키지를 로딩중입니다: grid
> Arthritis
   ID Treatment    Sex Age Improved
1  57   Treated   Male  27     Some
2  46   Treated   Male  29     None
3  77   Treated   Male  30     None
```

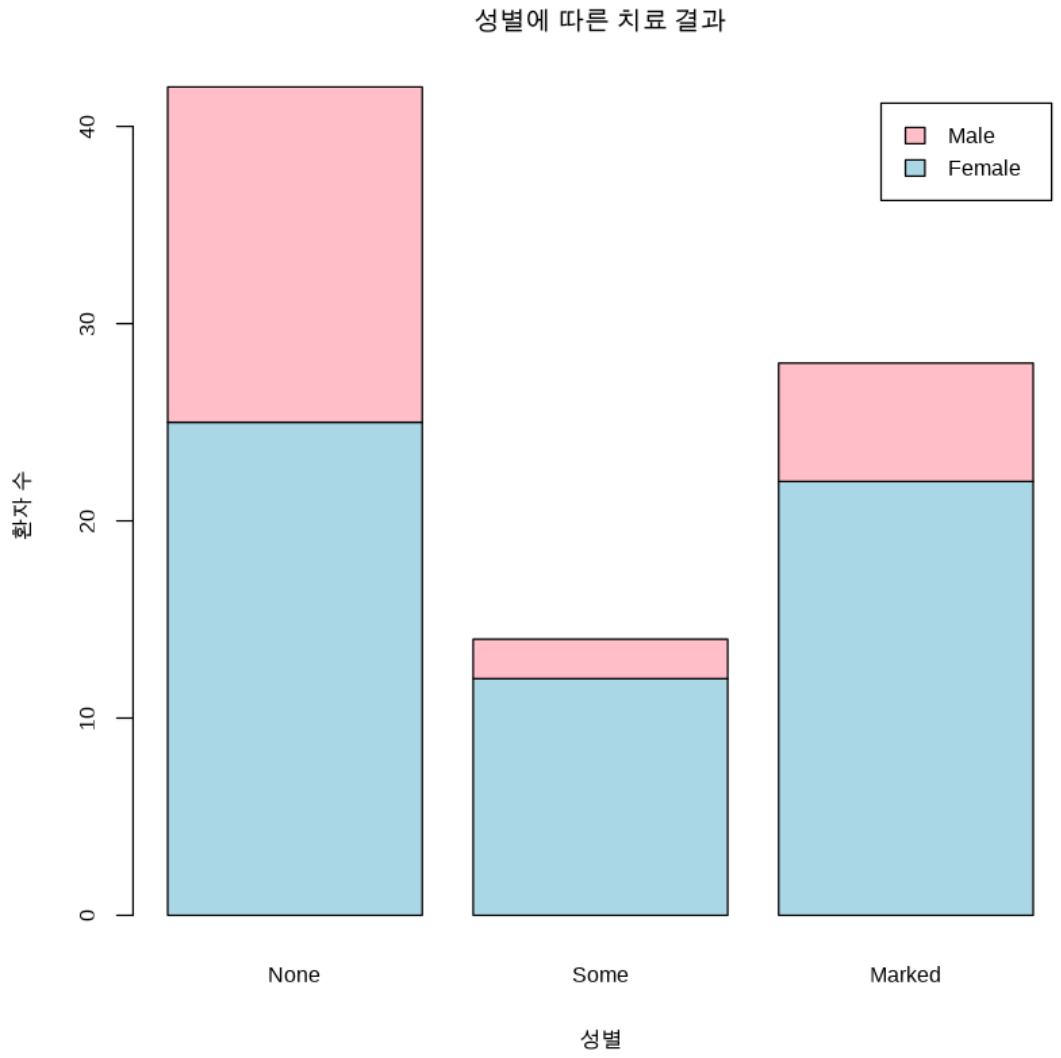
코드:

```
# 필요한 패키지 설치 및 로드
install.packages("vcd")
library(vcd)

# Arthritis 데이터셋 불러오기
data(Arthritis)

# 1. 치료제 투약 여부(Treatment)와 치료 결과(Improved) 간의 연관성 시각화
# 빈도표 생성
treatment_improved <- table(Arthritis$Treatment,
Arthritis$Improved)

# 누적 막대 그래프 생성
barplot(treatment_improved,
        beside = FALSE,
        col = c("lightblue", "pink", "lightgreen"),
        legend.text = TRUE,
        main = "치료제 투약 여부에 따른 치료 결과",
        xlab = "치료제 그룹",
        ylab = "환자 수")

# 2. 성별(Sex)과 치료 결과(Improved) 간의 연관성 시각화
# 빈도표 생성
sex_improved <- table(Arthritis$Sex, Arthritis$Improved)
```

```r
# 누적 막대 그래프 생성
barplot(sex_improved,
        beside = FALSE,
        col = c("lightblue", "pink", "lightgreen"),
        legend.text = TRUE,
        main = "성별에 따른 치료 결과",
        xlab = "성별",
        ylab = "환자 수")
```

결과:





해석:
첫 번째 그래프는 치료제 투약 여부에 따른 치료 결과를 보여주는데, 치료 효과가 없었던(None) 환자들 중에서는 위약(Placebo) 그룹의 비율이 새로운 치료제(Treated) 그룹보다 훨씬 높음을 알 수 있다. 반면 뚜렷한 개선(Marked)을 보인 환자들 중에는 치료제 그룹이 위약 그룹보다 현저히 많아, 새로운 치료제가 위약에 비해 류마티스 관절염 증상 개선에 효과적이라는 것을 시사한다. 약간의 개선(Some)이 있었던 환자들은 두 그룹 간에 큰 차이가 없으나, 여전히 치료제 그룹의 비율이 조금 더 높은 것으로 보인다.
두 번째 그래프는 성별에 따른 치료 결과를 보여주는데, 전체적으로 여성(Female) 환자가 남성(Male) 환자보다 많은 것이 특징이다. 치료 효과가 없었던(None) 환자들과 약간의 개선(Some)이 있었던 환자들, 그리고 뚜렷한 개선(Marked)을 보인 환자들 모두에서 여성의 수가 더 많지만, 이는 전체 여성 환자 수가 더 많기 때문으로 보인다. 남성과 여성 간의 치료 결과 분포를 비율로 고려했을 때는 큰 차이가 없어 보여, 성별과 치료 결과 사이에는 뚜렷한 연관성이 보이지 않는다고 할 수 있다. 결론적으로, 치료제 투약 여부는 치료 결과와 강한 연관성이 있으며 새로운 치료제가 효과적임을 보여주지만, 성별은 치료 결과와 뚜렷한 연관성을 보이지 않는다.

문제**4.**
R에 내장된 “airquality” 데이터셋(별도 패키지 설치 필요 없음)은 1973년 5월부터 9월까지 뉴욕의

대기질에 관한 데이터셋이다. 변수 Ozone은 대기 중 오존의 양, Solar.R은 태양방사선의 양, Wind는 풍속, Temp는 기온을 나타낸다. 이 네 가지 변수(Ozone, Solar.R, Wind, Temp)에 대한 산점도 행렬을 그리고, 이 산점도 행렬에서 알 수 있는 변수들 간의 관계에 대하여 서술하시오. (25점)

[R code]

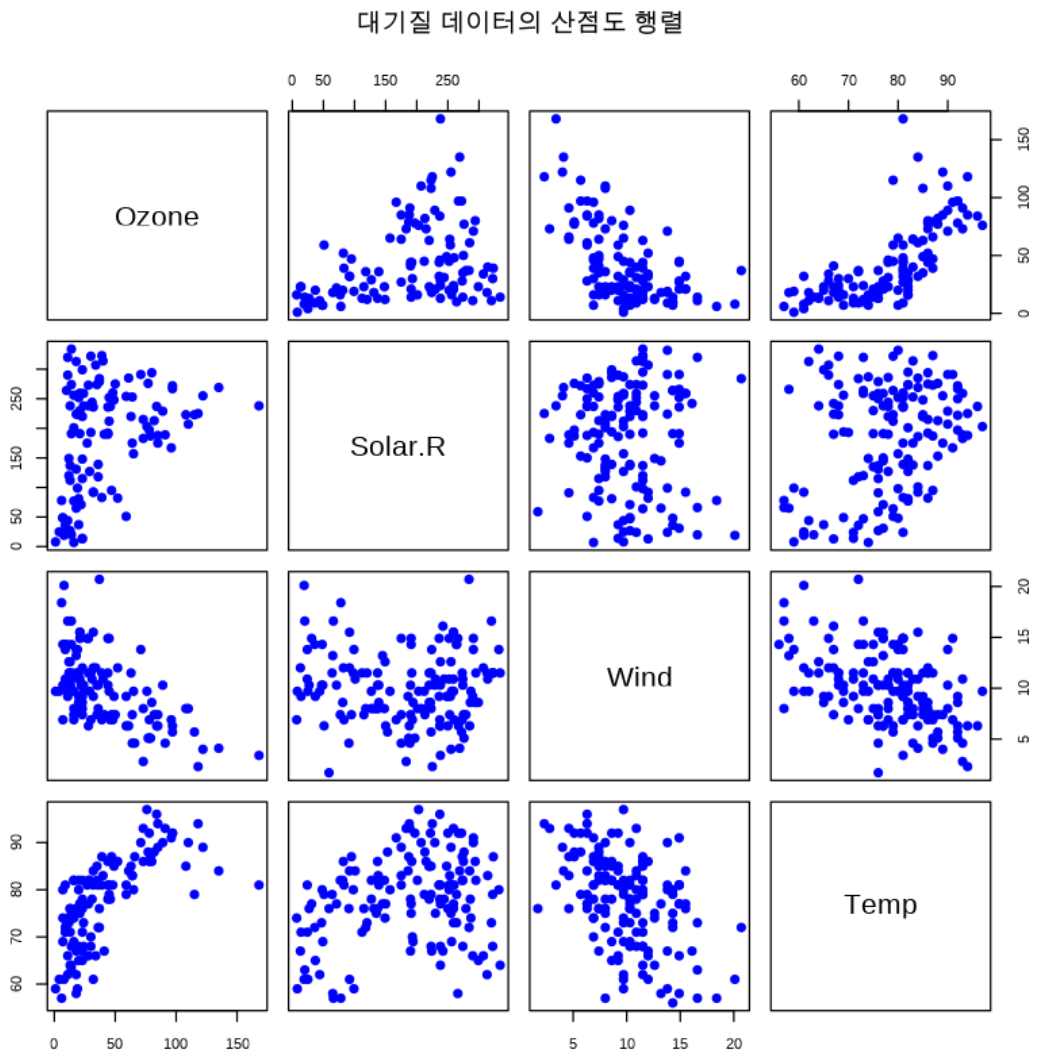
airquality

```
> airquality
    Ozone Solar.R Wind Temp Month Day
1      41     190  7.4   67     5   1
2      36     118  8.0   72     5   2
3      12     149 12.6   74     5   3
```

코드:

```
# airquality 데이터셋 확인
head(airquality)

# 산점도 행렬
pairs(airquality[, c("Ozone", "Solar.R", "Wind", "Temp")],
      main = "대기질 데이터의 산점도 행렬",
      pch = 19,
      col = "blue")

# 상관계수 계산 (NA 제외)
cor(airquality[, c("Ozone", "Solar.R", "Wind", "Temp")], use =
"complete.obs")
```

결과:



해석:
Ozone(오존)과 Temp(기온)의 관계:
오존과 기온 사이에는 뚜렷한 양의 상관관계가 있다. 기온이 상승할수록 대기 중 오존 농도도 함께 증가하는 경향을 보인다. 이는 고온에서 광화학 반응이 활발해져 오존 생성이 증가하기 때문으로 해석할 수 있다.
Ozone(오존)과 Wind(풍속)의 관계:
오존과 풍속 사이에는 명확한 음의 상관관계가 있다. 풍속이 증가할수록 오존 농도는 감소하는 경향을 보인다. 이는 바람이 강할수록 오존이 분산되어 농도가 낮아지기 때문이다.
Ozone(오존)과 Solar.R(태양방사선)의 관계:
약한 양의 상관관계가 관찰된다. 태양방사선이 강할수록 오존 농도가 다소 증가하는 경향이 있으나, 그 관계는 기온만큼 강하지 않다.
Wind(풍속)과 Temp(기온)의 관계:
풍속과 기온 사이에는 음의 상관관계가 있다. 기온이 높을수록 풍속이 낮은 경향을 보인다.
Solar.R(태양방사선)과 Temp(기온)의 관계:

태양방사선과 기온 사이에는 약한 양의 상관관계가 있다. 태양방사선이 강할 때 기온도 높은 경향이 있다.
Solar.R(태양방사선)과 Wind(풍속)의 관계:
뚜렷한 상관관계가 보이지 않으며, 데이터 포인트가 비교적 고르게 분포되어 있다.

대기 오존 농도는 기온과 가장 강한 양의 상관관계를 보이고, 풍속과는 강한 음의 상관관계를 보인다. 즉, 더운 날씨에 바람이 적을 때 오존 오염이 심해지는 경향이 있다. 이러한 정보는 대기질 예측 모델 개발이나 오존 경보 시스템 구축에 중요한 기초 자료가 될 수 있다.

<과제물 작성 관련 유의사항**>**

1. 모든 문항은 과제물 파일에 R 프로그램 코드와 실행한 결과물(캡처 이미지)을 같이 포함하여 작성한다.
2. 모든 문항에 대해 하나의 파일로 작성하며, 실행 결과에 대한 캡처 이미지는 적정 크기로 하여 제출하는 문서 파일의 본문 내에 삽입하여 포함한다.