강원대학교
전자·AI시스템공학과

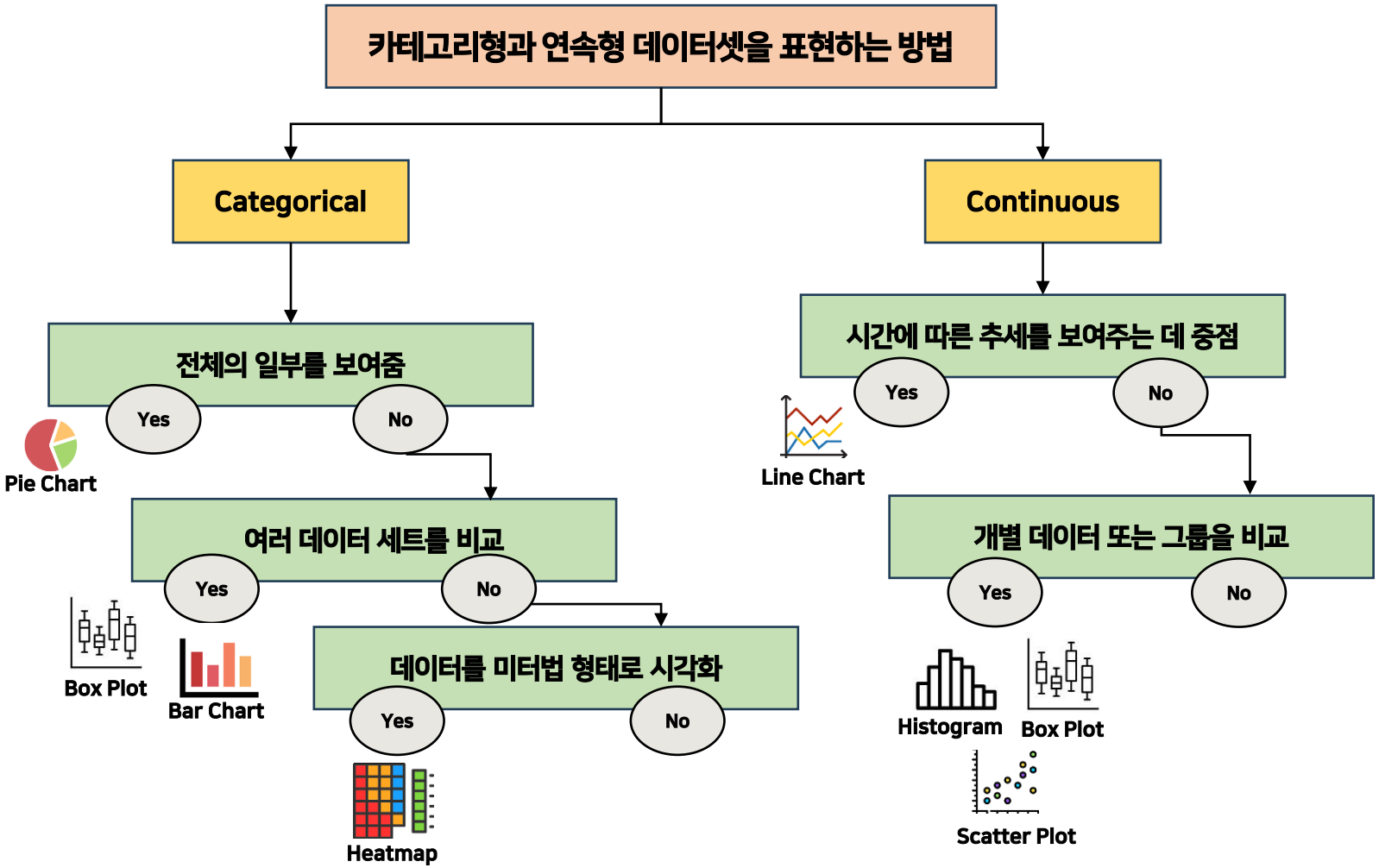
# 머신러닝1
## - Exploratory Data Analysis -

KNU
강원대학교

## 기술통계-탐색적 데이터 분석(EDA)

- EDA는 Exploratory Data Analysis의 약자로, 탐색적 데이터 분석을 의미함
- 데이터 분석을 시작하기 전에 데이터를 다양한 각도에서 관찰하고 이해하는 과정
- 데이터의 기본적인 특성, 구조, 패턴, 이상치, 변수 간의 관계 등을 파악함으로써 분석가가 보다 유익한 인사이트를 얻음
- 데이터에 대한 이해를 바탕으로 더 효율적인 분석 계획을 세울 수 있도록 하는 과정

## 기술통계-탐색적 데이터 분석(EDA)



## 기술통계-탐색적 데이터 분석(EDA)

### R에서 데이터를 그래프로 표현하는 방법

- **ggplot2**

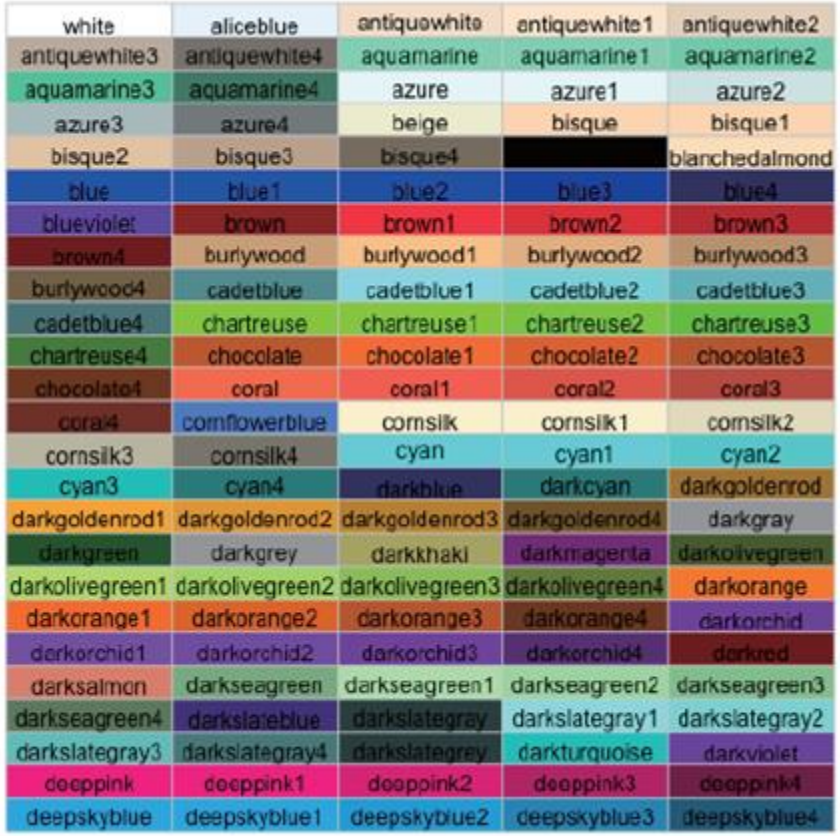
```
ggplot(데이터 프레임, aes(x = 변수, y = 변수, fill = 그래프의 색상을 나타내는 변수))+
        geom_*(stat = "identity", width = 막대의 넓이, ) +
        scale_fill_manual(values = 막대 그래프를 지정된 색으로 수정) +
        labs(title = 그래프의 타이틀 제목) +
        xlab(x축 제목) +
        ylab(y축 제목) +
        theme_void() +
        theme(
                text = element_text(color = "blue"),
                title = element_text(size = 16, face = "bold"),
                axis.text = element_text(size = 12)
         )
```

## 기술통계-탐색적 데이터 분석(EDA)

### R에서 데이터를 그래프로 표현하는 방법

- 그래프 컬러 모음

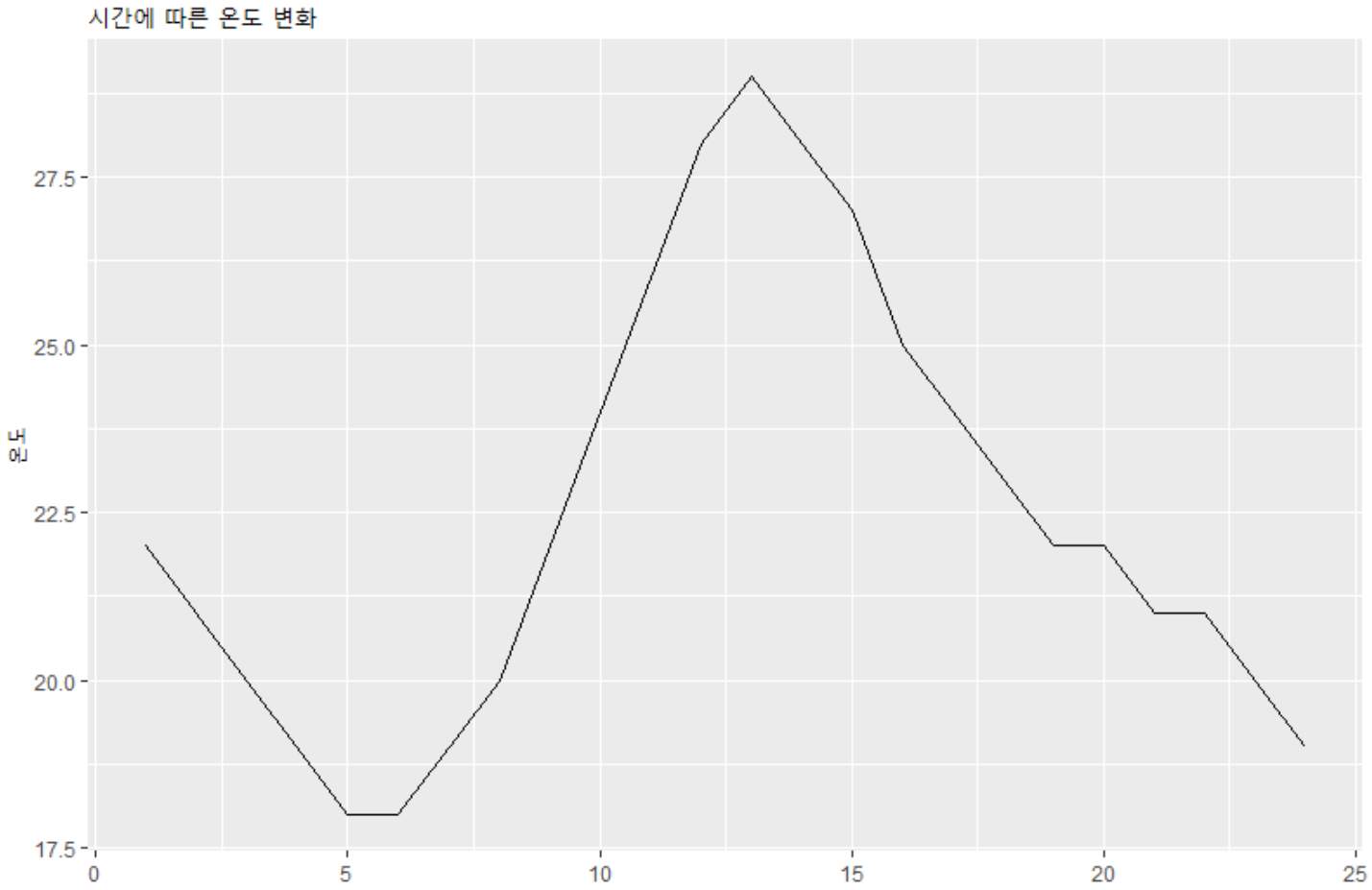
```
> colors()
  [1] "white"           "aliceblue"       "antiquewhite"
  [4] "antiquewhite1"   "antiquewhite2"   "antiquewhite3"
  [7] "antiquewhite4"   "aquamarine"      "aquamarine1"
 [10] "aquamarine2"     "aquamarine3"     "aquamarine4"
 [13] "azure"           "azure1"          "azure2"
 [16] "azure3"          "azure4"          "beige"
 [19] "bisque"          "bisque1"         "bisque2"
 [22] "bisque3"         "bisque4"         "black"
 [25] "blanchedalmond"  "blue"            "blue1"
 ...
[628] "thistle3"        "thistle4"        "tomato"
[631] "tomato1"         "tomato2"         "tomato3"
[634] "tomato4"         "turquoise"       "turquoise1"
[637] "turquoise2"      "turquoise3"      "turquoise4"
[640] "violet"          "violetred"       "violetred1"
[643] "violetred2"      "violetred3"      "violetred4"
[646] "wheat"           "wheat1"          "wheat2"
[649] "wheat3"          "wheat4"          "whitesmoke"
[652] "yellow"          "yellow1"         "yellow2"
[655] "yellow3"         "yellow4"         "yellowgreen"
```



## 기술통계-탐색적 데이터 분석(EDA)

### 탐색적 데이터 분석(EDA) – Continuous

- 라인그래프(Line Chart) : 시간의 흐름에 따른 데이터의 변화 추이를 시각화하는 데 사용함

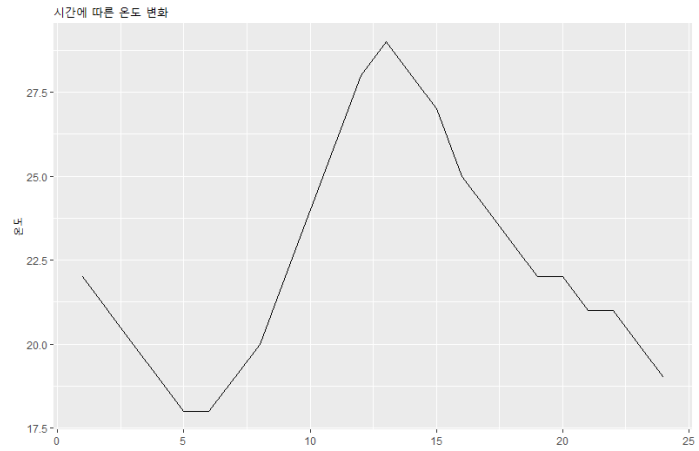


## 기술통계-탐색적 데이터 분석(EDA)

### 탐색적 데이터 분석(EDA) – Continuous

- 라인그래프(Line Chart) : 시간의 흐름에 따른 데이터의 변화 추이를 시각화하는 데 사용함

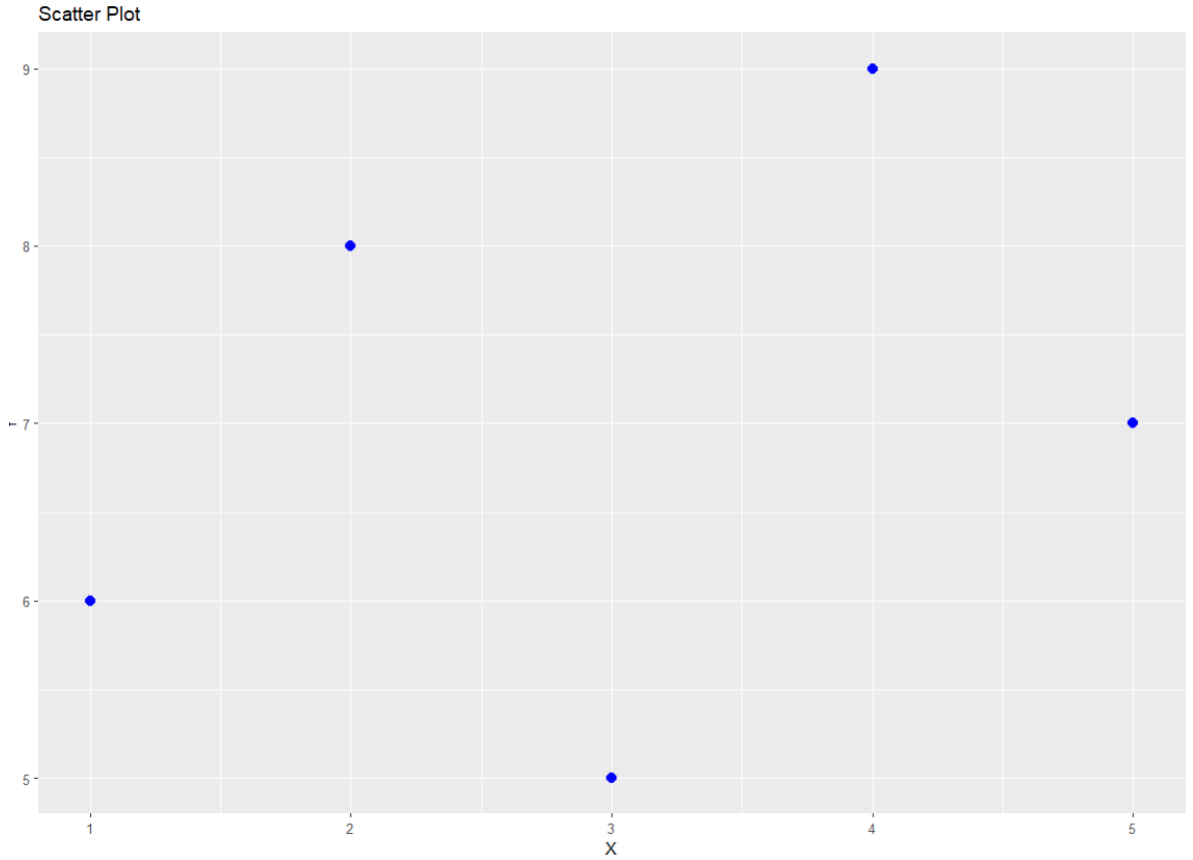
```
time <- seq(1, 24) # 24시간
temperature <- c(22, 21, 20, 19, 18, 18, 19, 20, 22, 24, 26, 28, 29, 28, 27, 25, 24, 23, 22, 22, 21, 21, 20, 19)

data <- data.frame(time=time, temp=temperature)

ggplot(data, aes(x=time, y=temp)) +
  geom_line() + # 선 그래프 추가
  labs(title="시간에 따른 온도 변화") +
  xlab("시간") +
  ylab("온도")
```



## 기술통계-탐색적 데이터 분석(EDA)

### 탐색적 데이터 분석(EDA) – Continuous

- 산점도(Scatter Plot) : 산점도는 직교 좌표계를 이용해 좌표상의 점들을 표시함으로써 두 개 변수 간의 관계를 나타내는 그래프 방법

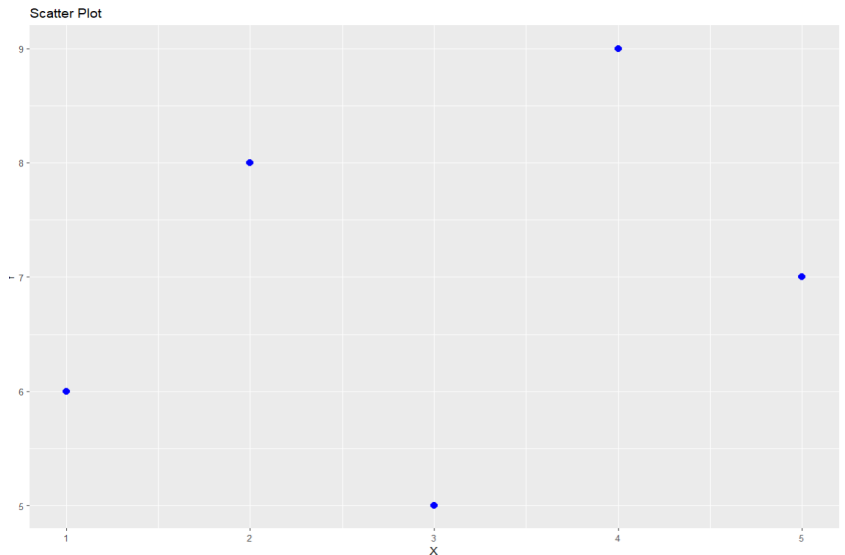


## 기술통계-탐색적 데이터 분석(EDA)

### 탐색적 데이터 분석(EDA) – Continuous

- 산점도(Scatter Plot) : 산점도는 직교 좌표계를 이용해 좌표상의 점들을 표시함으로써 두 개 변수 간의 관계를 나타내는 그래프 방법

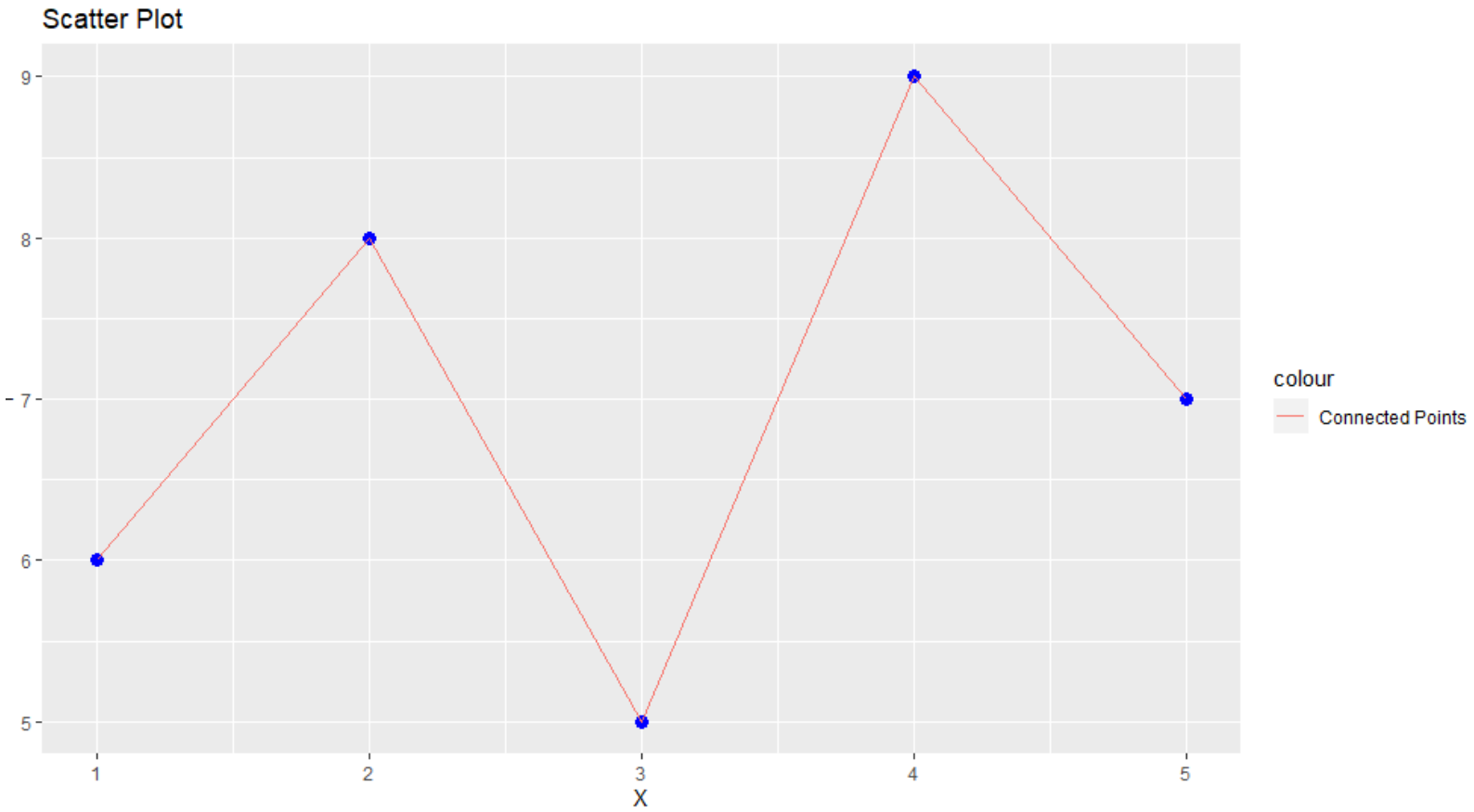
```
df <- data.frame( x = c(1, 2, 3, 4, 5),  y = c(6, 8, 5, 9, 7))

ggplot(df, aes(x = x, y = y)) +
            geom_point(color = "blue", size = 3) +
             labs(title = "Scatter Plot") +
            xlab("X") +
            ylab("Y")
```



## 기술통계-탐색적 데이터 분석(EDA)

### 탐색적 데이터 분석(EDA) – Continuous

- 산점도(Scatter Plot) : 산점도는 직교 좌표계를 이용해 좌표상의 점들을 표시함으로써 두 개 변수 간의 관계를 나타내는 그래프 방법

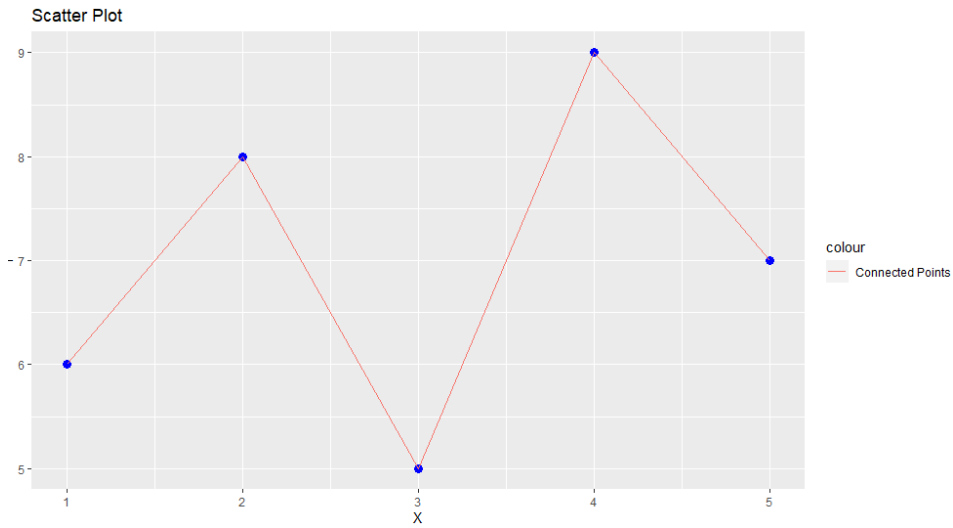


## 기술통계-탐색적 데이터 분석(EDA)

### 탐색적 데이터 분석(EDA) – Continuous

- 산점도(Scatter Plot) : 산점도는 직교 좌표계를 이용해 좌표상의 점들을 표시함으로써 두 개 변수 간의 관계를 나타내는 그래프 방법

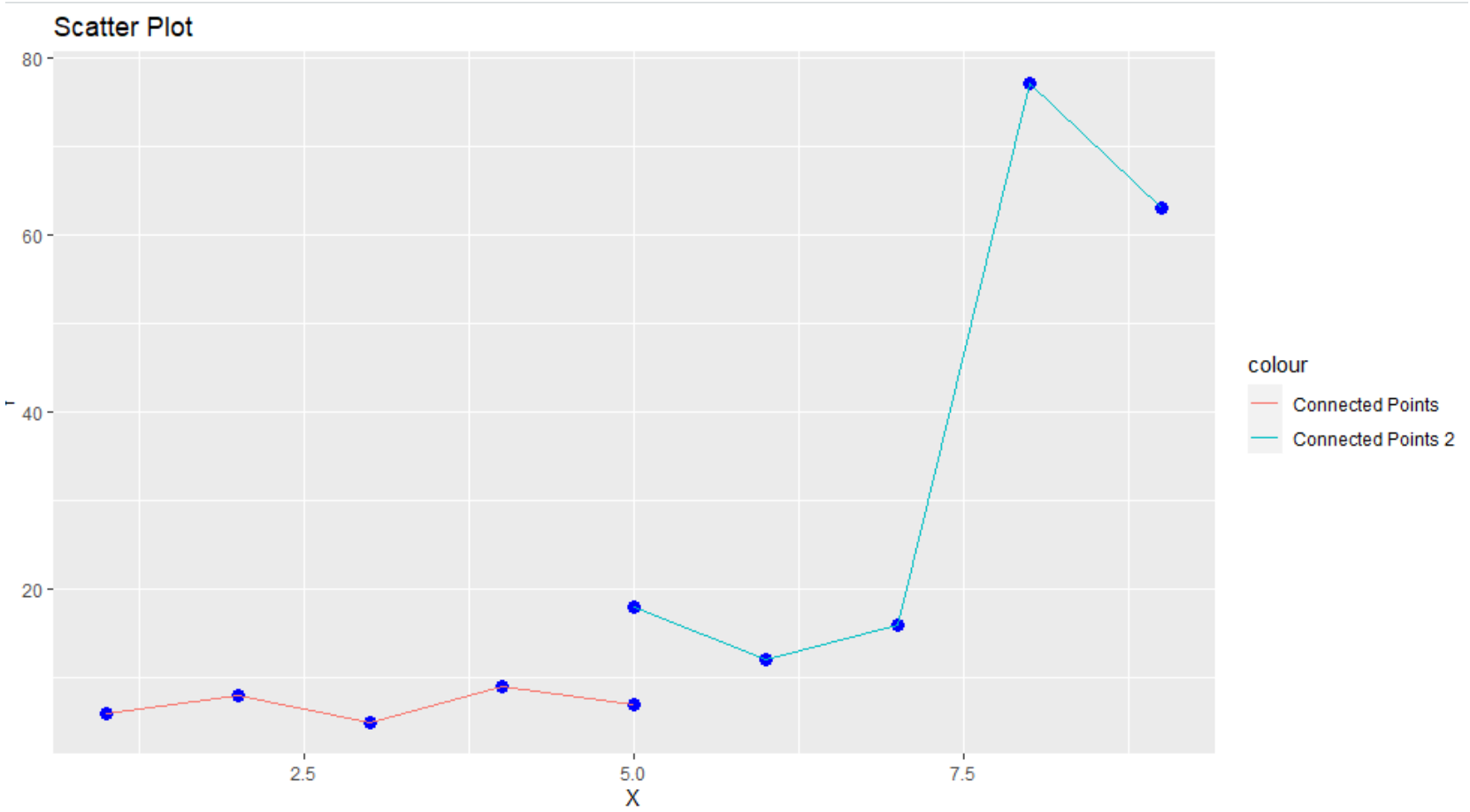
```
df <- data.frame( x = c(1, 2, 3, 4, 5),  y = c(6, 8, 5, 9, 7))

ggplot(df, aes(x = x, y = y)) +
        geom_point(color = "blue", size = 3) +
        geom_line(aes(color = "Connected Points"), size = 0.5) +
        labs(title = "Scatter Plot") +
        xlab("X") +
        ylab("Y")
```



## 기술통계-탐색적 데이터 분석(EDA)

### 탐색적 데이터 분석(EDA) – Continuous

- 산점도(Scatter Plot) : 산점도는 직교 좌표계를 이용해 좌표상의 점들을 표시함으로써 두 개 변수 간의 관계를 나타내는 그래프 방법

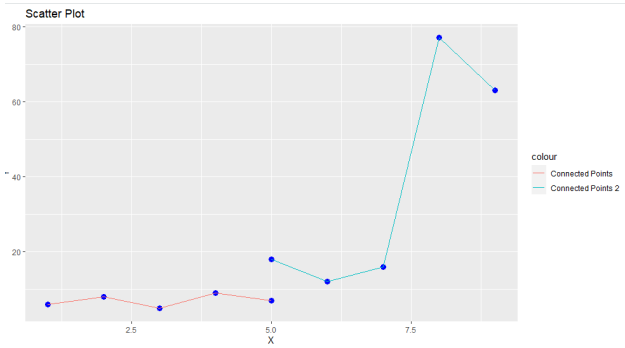


## 기술통계-탐색적 데이터 분석(EDA)

### 탐색적 데이터 분석(EDA) – Continuous

- **산점도(Scatter Plot) : 산점도는 직교 좌표계를 이용해 좌표상의 점들을 표시함으로써 두 개 변수 간의 관계를 나타내는 그래프 방법**

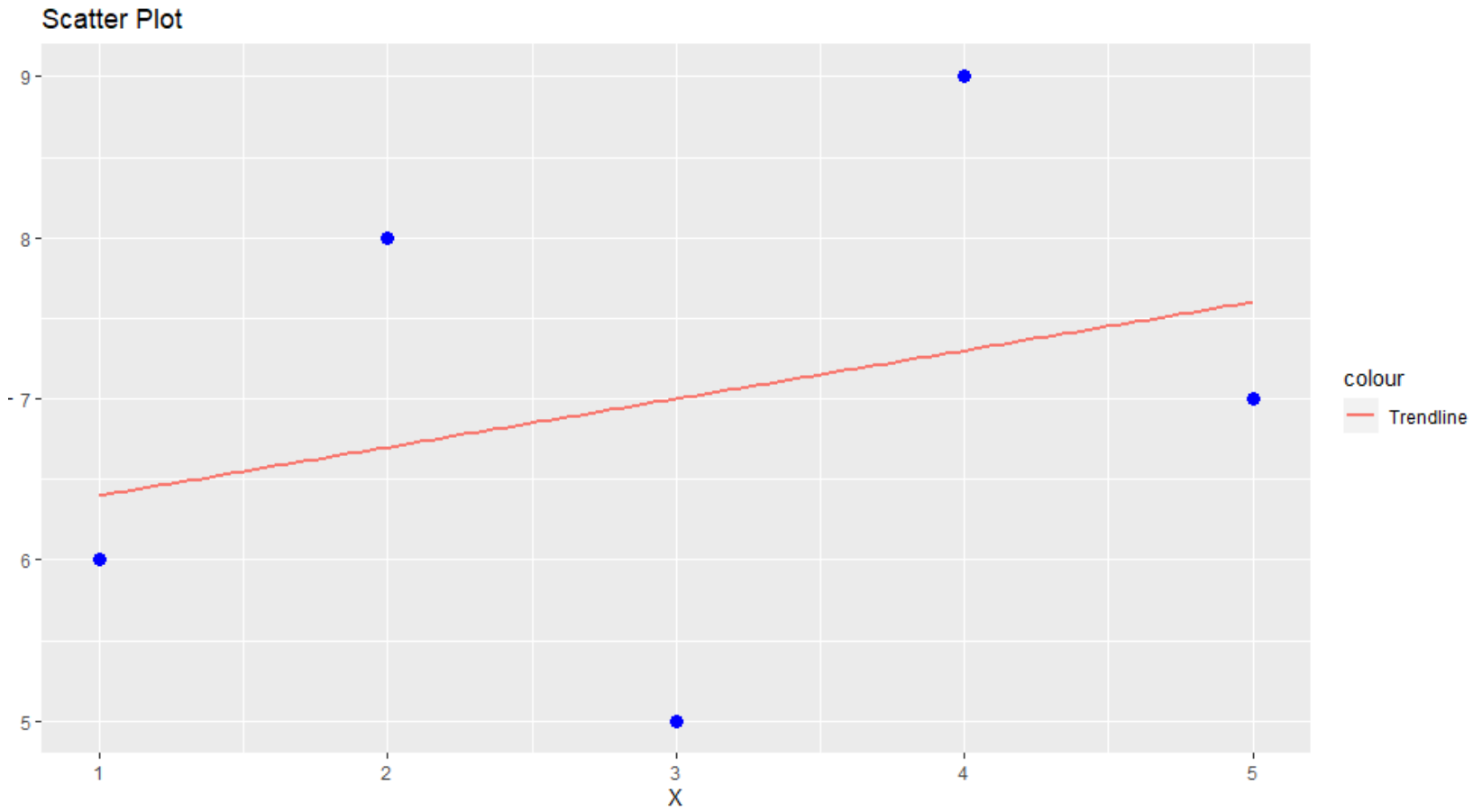
```
df <- data.frame(x = c(1, 2, 3, 4, 5), y = c(6, 8, 5, 9, 7))
df2 <- data.frame(x = c(5, 6, 7, 8, 9), y = c(18, 12, 16, 77, 63))

# Create the plot
ggplot() +
  geom_point(data = df, aes(x = x, y = y), color = "blue", size = 3) +
  geom_line(data = df, aes(x = x, y = y, color = "Connected Points"), size = 0.5) +
  geom_point(data = df2, aes(x = x, y = y), color = "blue", size = 3) +
  geom_line(data = df2, aes(x = x, y = y, color = "Connected Points 2"), size = 0.5) +
  labs(title = "Scatter Plot") +
  xlab("X") +
  ylab("Y")
```



## 기술통계-탐색적 데이터 분석(EDA)

### 탐색적 데이터 분석(EDA) – Continuous

- 산점도(Scatter Plot) : 산점도는 직교 좌표계를 이용해 좌표상의 점들을 표시함으로써 두 개 변수 간의 관계를 나타내는 그래프 방법

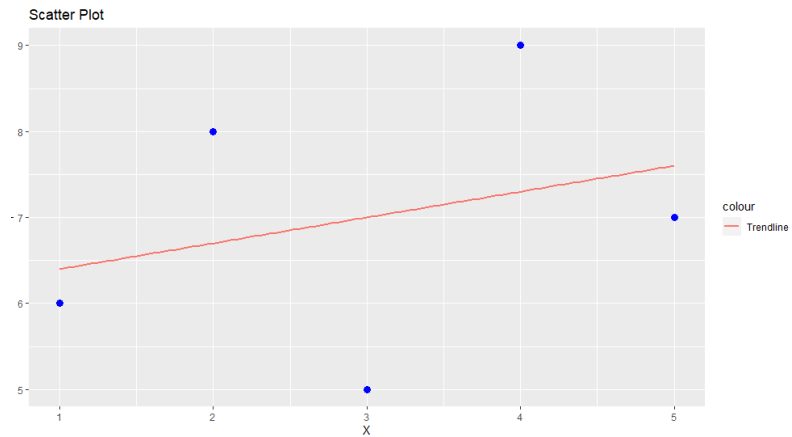


## 기술통계-탐색적 데이터 분석(EDA)

### 탐색적 데이터 분석(EDA) – Continuous

- 산점도(Scatter Plot) : 산점도는 직교 좌표계를 이용해 좌표상의 점들을 표시함으로써 두 개 변수 간의 관계를 나타내는 그래프 방법

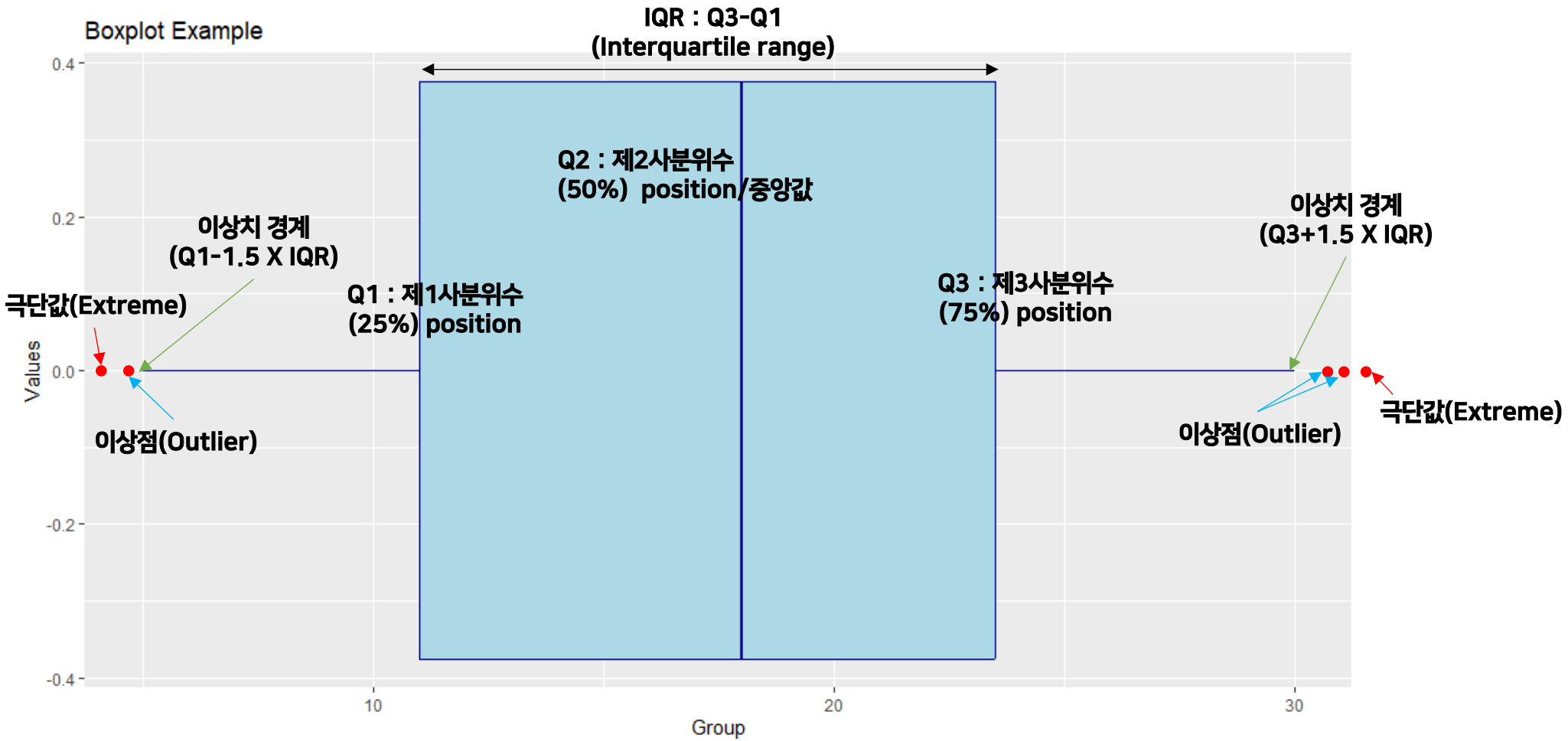
```
df <- data.frame( x = c(1, 2, 3, 4, 5),  y = c(6, 8, 5, 9, 7))

ggplot(df, aes(x = x, y = y)) +
        geom_point(color = "blue", size = 3) +
        geom_smooth(method = "lm", se = FALSE, aes(color = "Trendline")) +
        labs(title = "Scatter Plot") +
        xlab("X") +
        ylab("Y")
```



## 기술통계-탐색적 데이터 분석(EDA)

### 탐색적 데이터 분석(EDA) – Continuous & Categorical

- 상자수염그림(Boxplot) : box-and-whisker plot이라고도 하는 box plot은 데이터 집합의 분포를 요약하는 그래프 표현

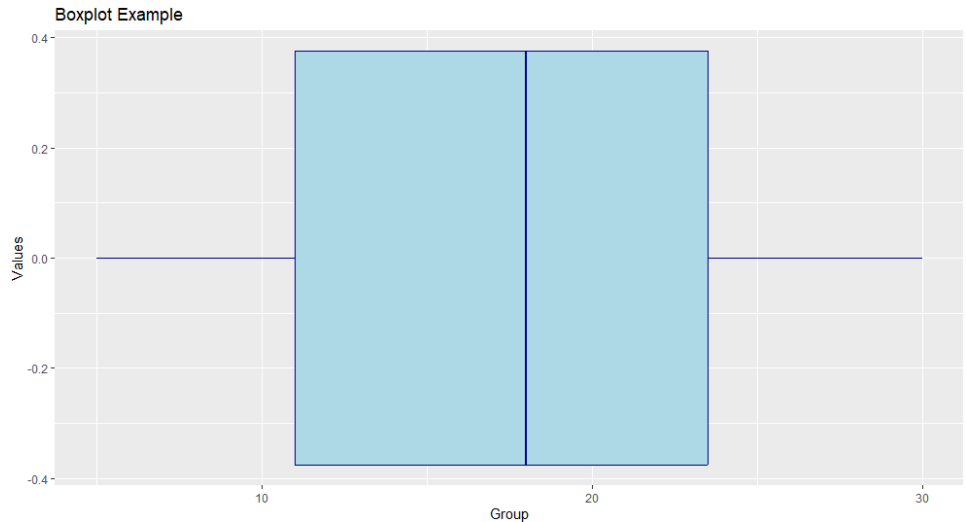


## 기술통계-탐색적 데이터 분석(EDA)

### 탐색적 데이터 분석(EDA) – Continuous & Categorical

- **상자수염그림(Boxplot) : box-and-whisker plot이라고도 하는 box plot은 데이터 집합의 분포를 요약하는 그래프 표현**

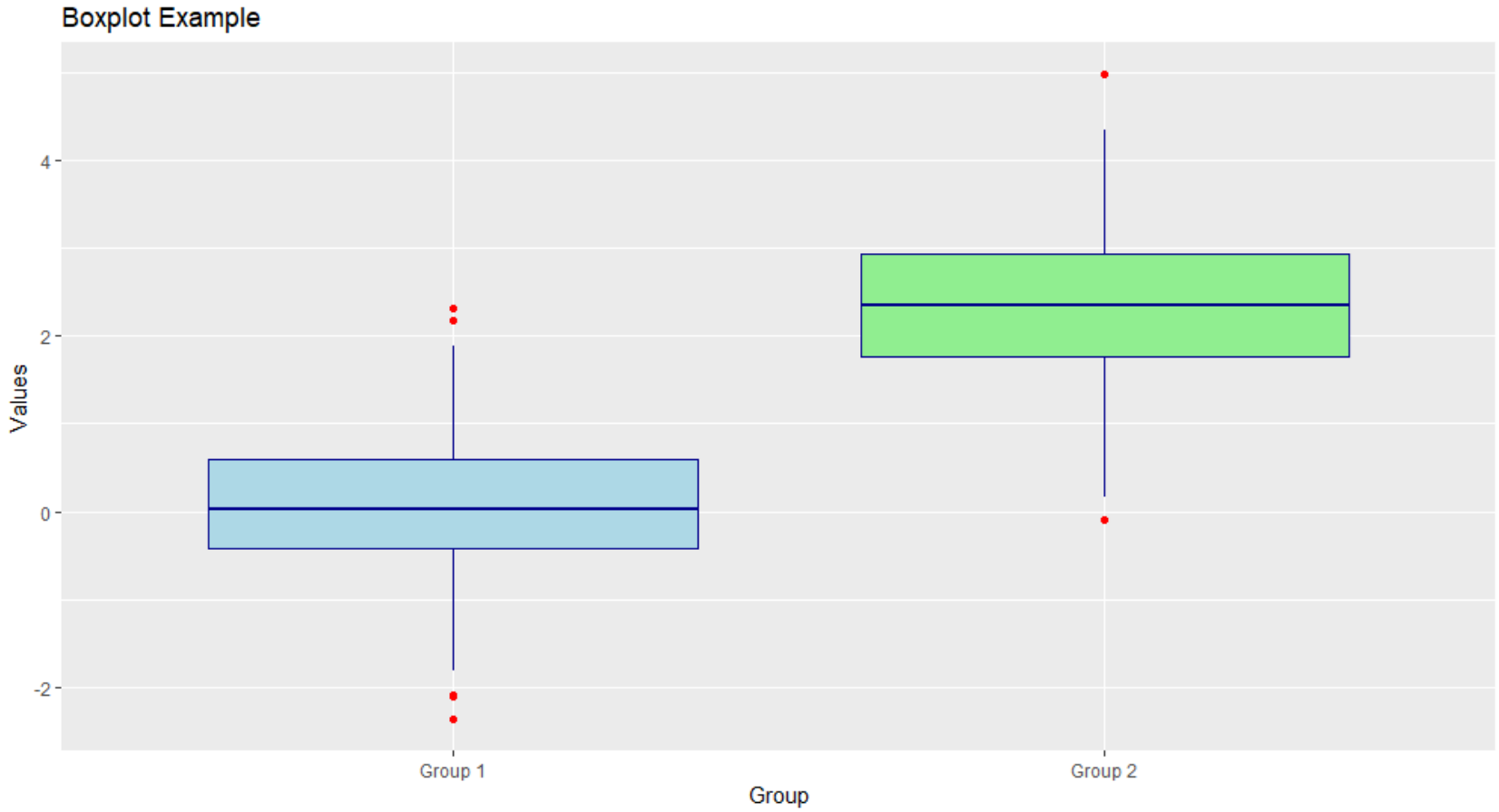
```
df <- data.frame(values = c(5, 7, 10, 12, 14, 18, 20, 22, 25, 27, 30))

ggplot(df, aes(x = values)) +
  geom_boxplot(binwidth = 5, fill = "steelblue", color = "white") +
  labs(title = "Histogram of Values") +
  xlab("Values") +
  ylab("Frequency")
```



## 기술통계-탐색적 데이터 분석(EDA)

### 탐색적 데이터 분석(EDA) – Continuous & Categorical

- 상자수염그림(Boxplot) : box-and-whisker plot이라고도 하는 box plot은 데이터 집합의 분포를 요약하는 그래프 표현

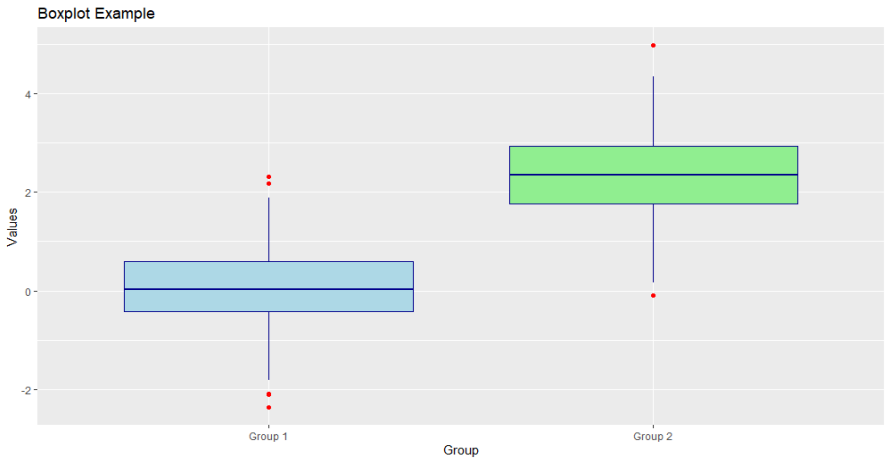


## 기술통계-탐색적 데이터 분석(EDA)

### 탐색적 데이터 분석(EDA) – Continuous & Categorical

- 상자수염그림(Boxplot) : box-and-whisker plot이라고도 하는 box plot은 데이터 집합의 분포를 요약하는 그래프 표현

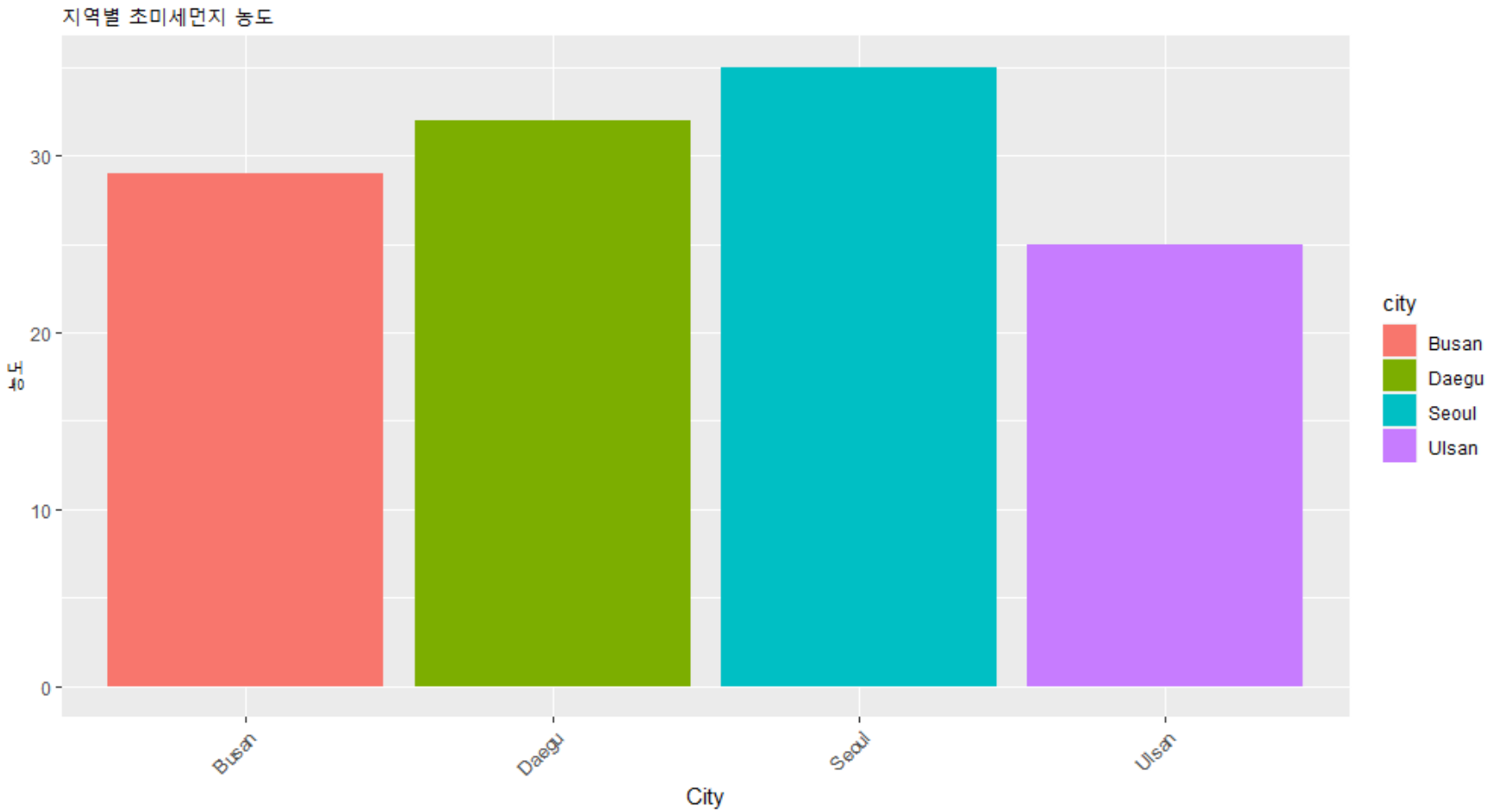
```
df <- data.frame(
  group = c(rep("Group 1", 60), rep("Group 2", 60)),
  values = c(rnorm(60, mean = 0, sd = 1), rnorm(60, mean = 2, sd = 1)))

ggplot(df, aes(x = group, y = values)) +
  geom_boxplot(fill = c("lightblue","lightgreen"), outlier.color = "red") +
  labs(title = "Boxplot Example") +
  xlab("Group") +
  ylab("Values")
```



## 기술통계-탐색적 데이터 분석(EDA)

### 탐색적 데이터 분석(EDA) – Categorical

- 막대그래프(Bar Chart) : 표현 값에 비례하여 높이와 길이를 지닌 직사각형 막대로 범주형 데이터를 표현하는 그래프

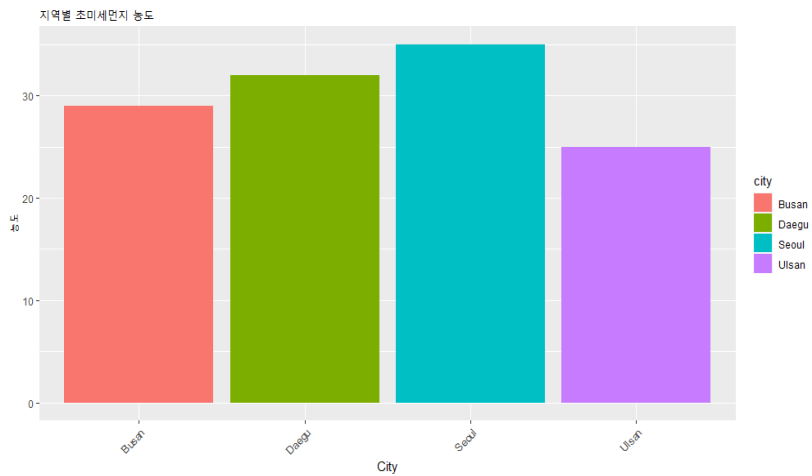


## 기술통계-탐색적 데이터 분석(EDA)

### 탐색적 데이터 분석(EDA) – Categorical

- 막대그래프(Bar Chart) : 표현 값에 비례하여 높이와 길이를 지닌 직사각형 막대로 범주형 데이터를 표현하는 그래프

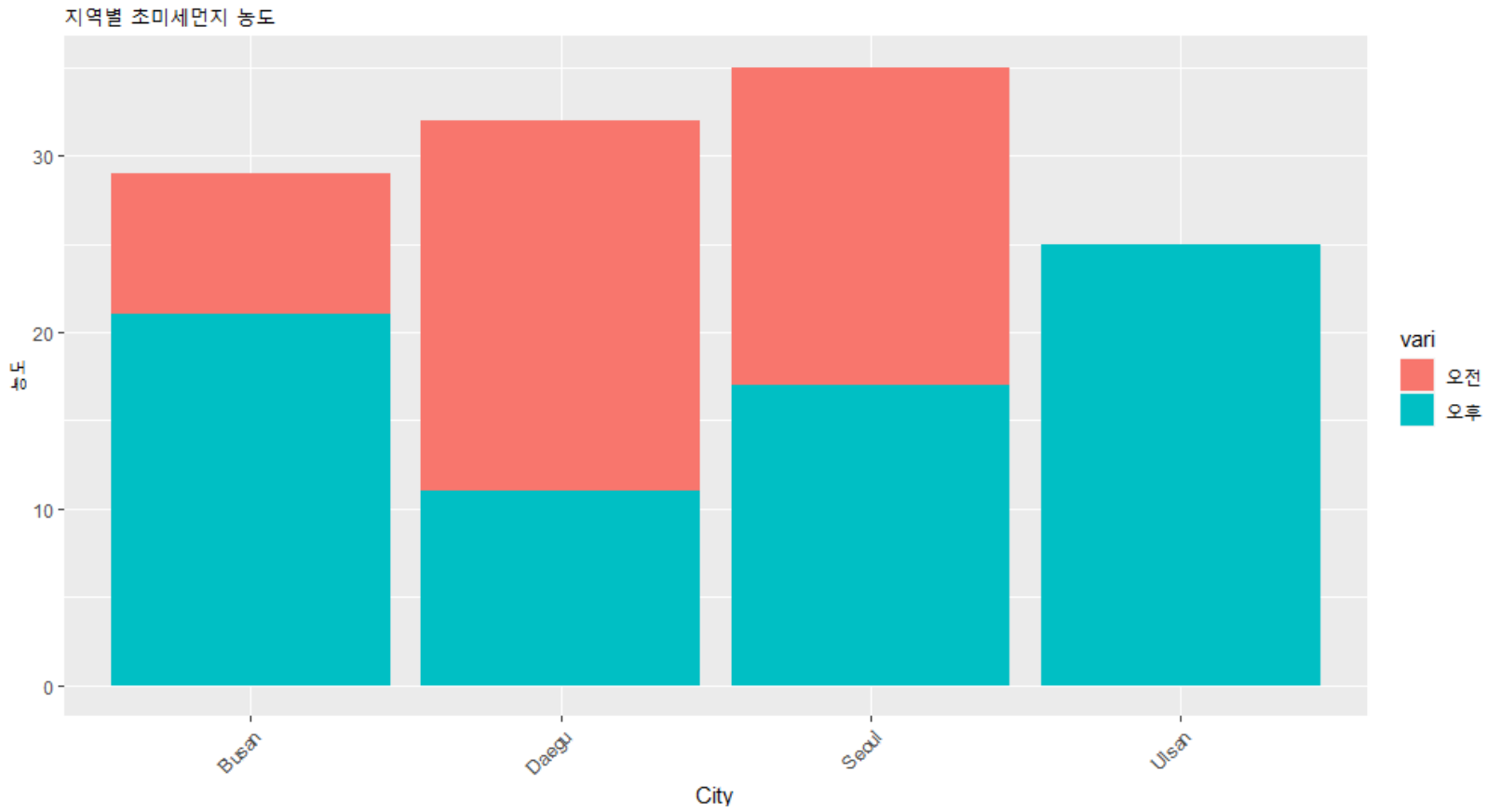
```
city <- c("Seoul", "Busan", "Daegu", "Seoul", "Busan", "Daegu", "Ulsan")
pm25 <- c(18, 21, 21, 17, 8, 11, 25)

df <- data.frame(city = city, pm25 = pm25)

ggplot(df, aes(x = city, y = pm25, fill = city)) +
  geom_bar(stat = "identity") +
  labs(title = "지역별 초미세먼지 농도") +
  xlab("City") +
  ylab("농도")
```



## 기술통계-탐색적 데이터 분석(EDA)

### 탐색적 데이터 분석(EDA) – Categorical

- 막대그래프(Bar Chart) : 표현 값에 비례하여 높이와 길이를 지닌 직사각형 막대로 범주형 데이터를 표현하는 그래프

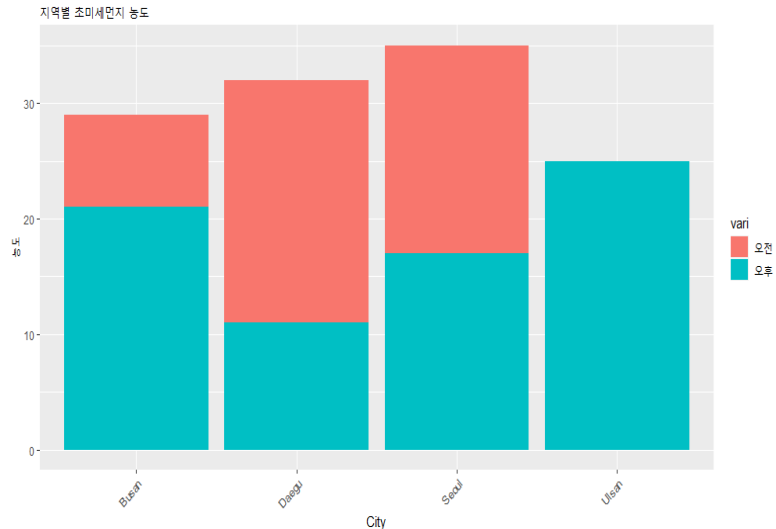


## 기술통계-탐색적 데이터 분석(EDA)

### 탐색적 데이터 분석(EDA) – Categorical

- 막대그래프(Bar Chart) : 표현 값에 비례하여 높이와 길이를 지닌 직사각형 막대로 범주형 데이터를 표현하는 그래프

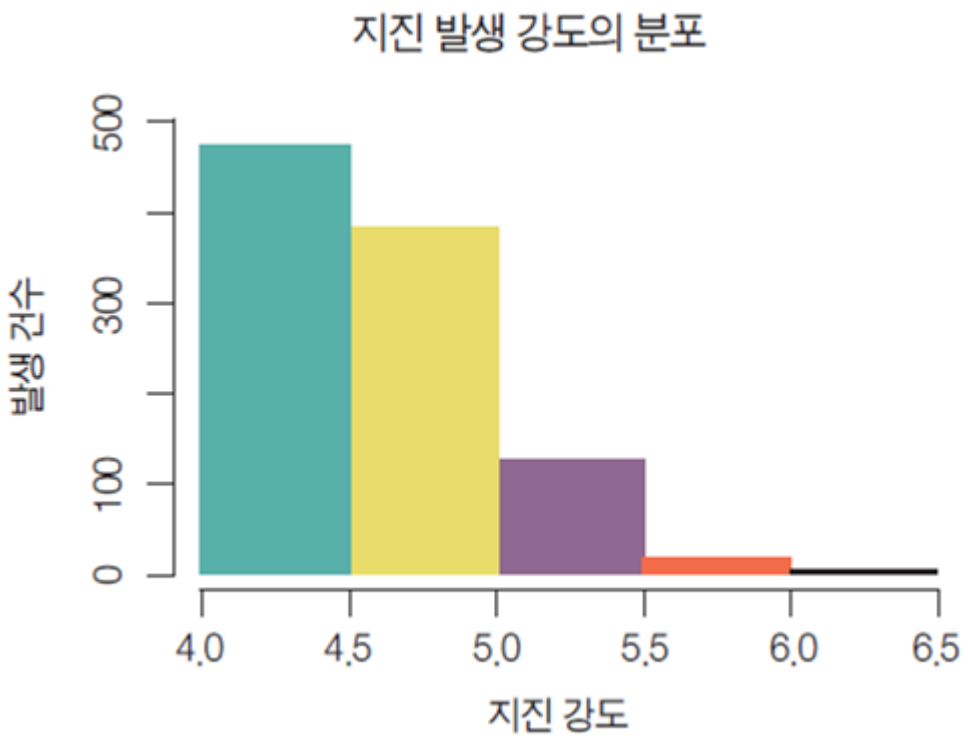
```
city <- c("Seoul", "Busan", "Daegu", "Seoul", "Busan", "Daegu", "Ulsan")
vari <- c("오전", "오후","오전","오후","오전","오후","오후")
pm25 <- c(18, 21, 21, 17, 8, 11, 25)

df <- data.frame(city = city, pm25 = pm25, vari=vari)

ggplot(df, aes(x = city, y = pm25, fill = vari)) +
  geom_bar(stat = "identity") +
  labs(title = "지역별 초미세먼지 농도") +
  xlab("City") +
  ylab("농도")
```

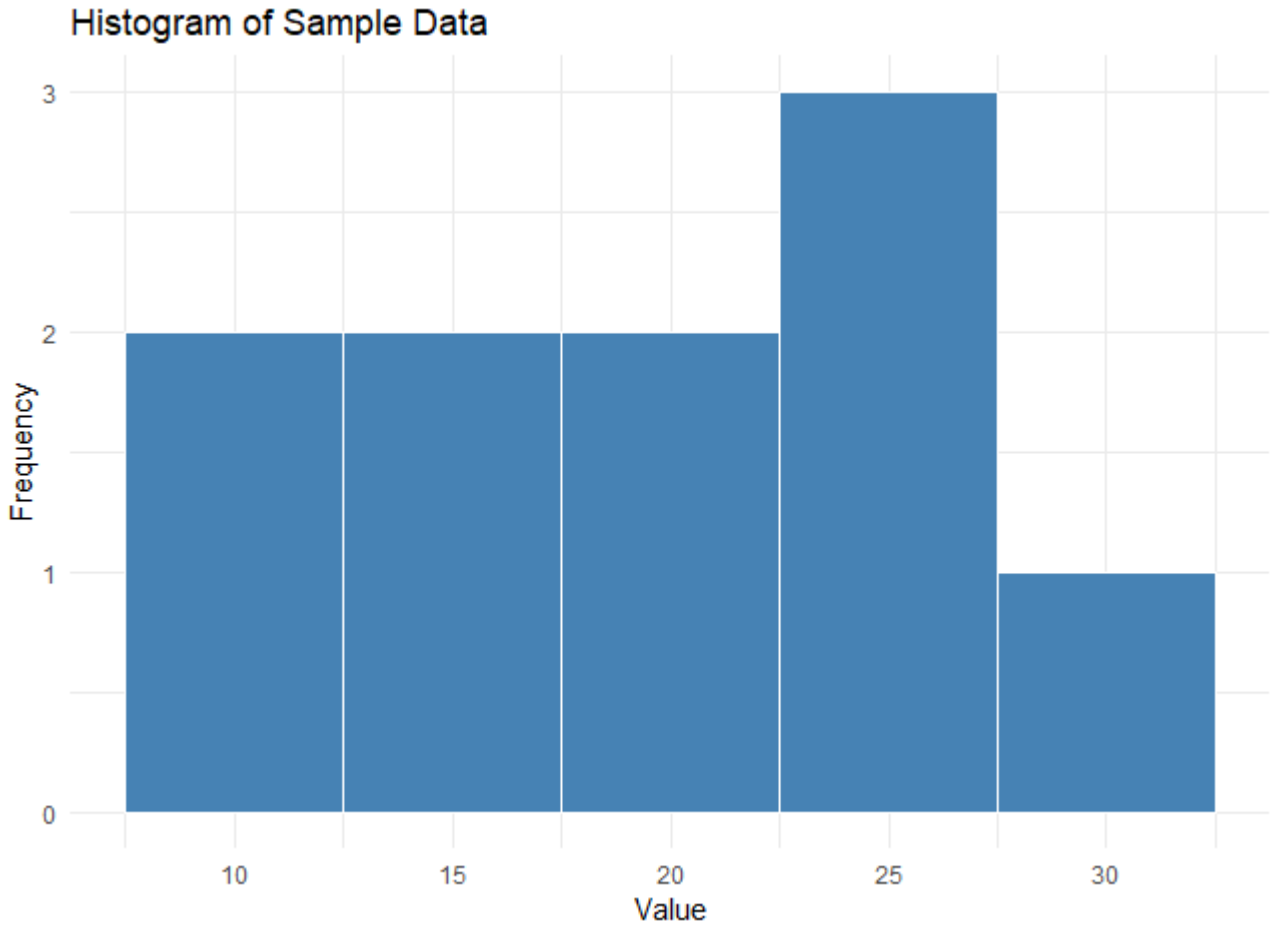


## 기술통계-탐색적 데이터 분석(EDA)

### 탐색적 데이터 분석(EDA) – Continuous

- 히스토그램(Histogram) : 히스토그램의 한 줄 요약은 데이터 분포에 대한 주요 정보를 제공하는 간결한 그래프



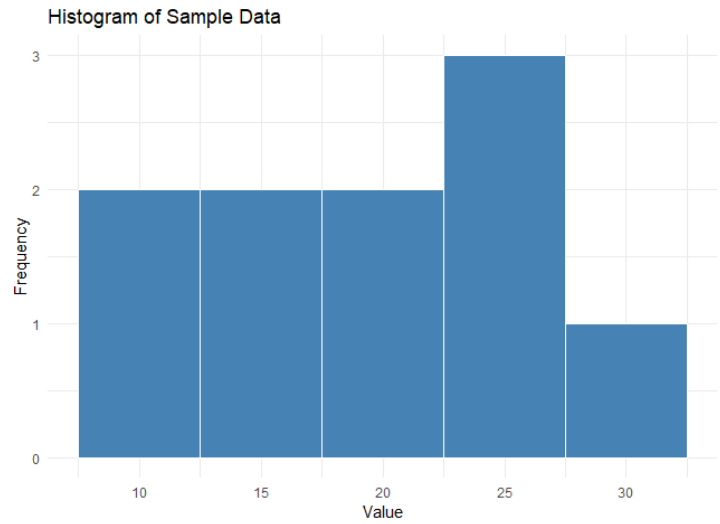



## 기술통계-탐색적 데이터 분석(EDA)

### 탐색적 데이터 분석(EDA) – Continuous

- 히스토그램(Histogram) : 히스토그램의 한 줄 요약은 데이터 분포에 대한 주요 정보를 제공하는 간결한 그래프

```
df <- data.frame(values = c(5, 7, 10, 12, 14, 18, 20, 22, 25, 27, 30))


# 히스토그램
ggplot(df, aes(x = values)) +
  geom_histogram(aes(y = ..density..), binwidth = 5, fill = "steelblue", color = "white") +
  labs(title = "Histogram") +
  xlab("Values") +
  ylab("Density")
```

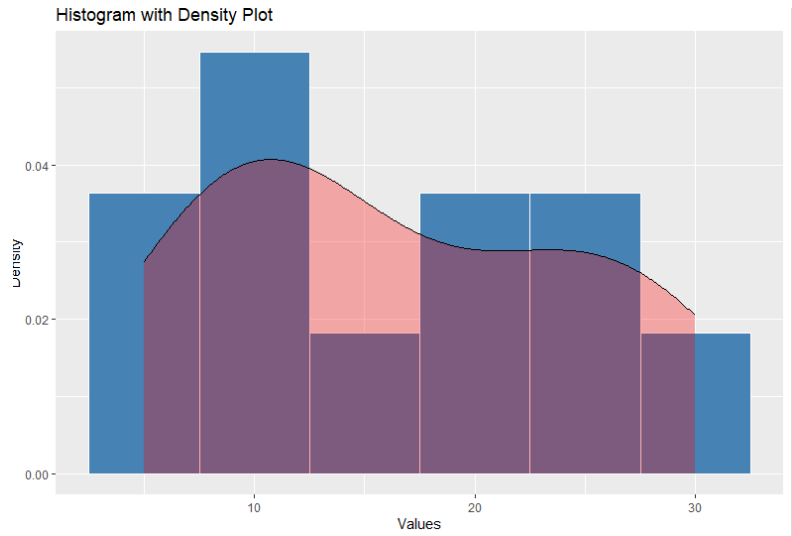


## 기술통계-탐색적 데이터 분석(EDA)

### 탐색적 데이터 분석(EDA) – Continuous

- 히스토그램(Histogram) : 히스토그램의 한 줄 요약은 데이터 분포에 대한 주요 정보를 제공하는 간결한 그래프

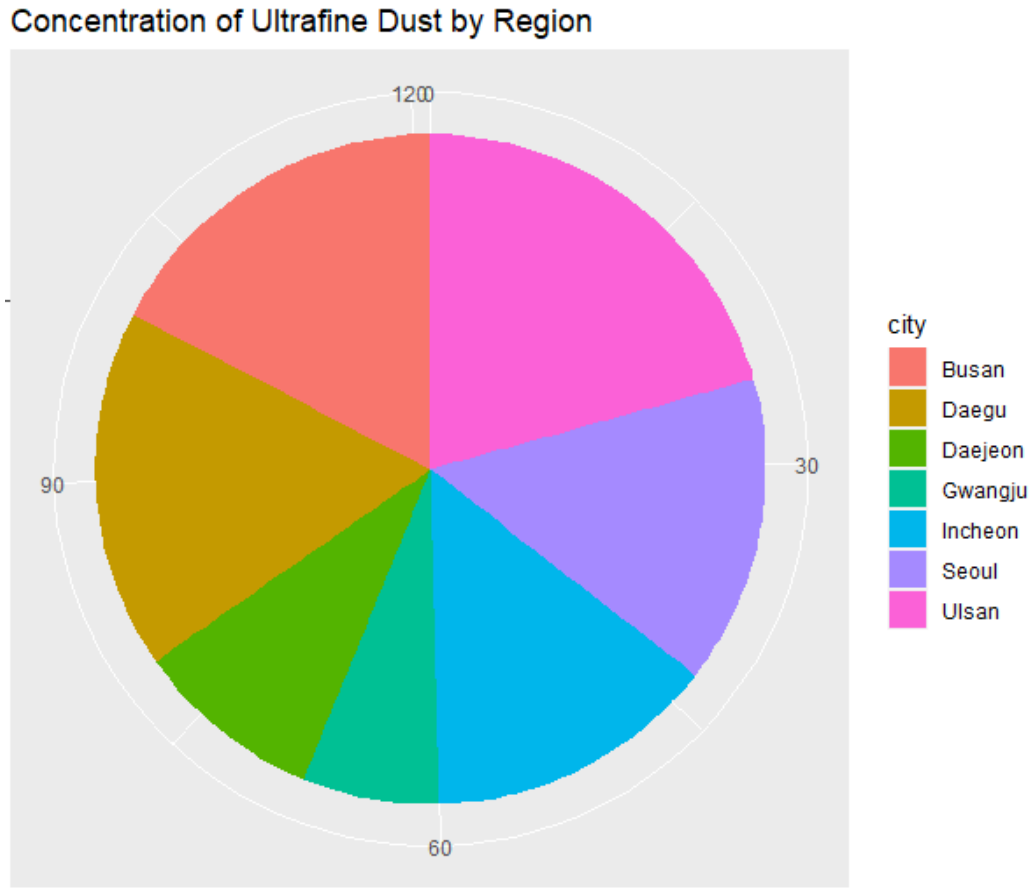
```
df <- data.frame(values = c(5, 7, 10, 12, 14, 18, 10, 22, 25, 27, 30))

# 히스토그램과 밀도 곡선 그리기
ggplot(df, aes(x = values)) +
  geom_histogram(aes(y = ..density..), binwidth = 5, fill = "steelblue", color = "white") +
  geom_density(alpha = 0.3, fill = "red") +  # 밀도 곡선 추가
  labs(title = "Histogram with Density Plot") +
  xlab("Values") +
  ylab("Density")
```



## 기술통계-탐색적 데이터 분석(EDA)

### 탐색적 데이터 분석(EDA) – Categorical

- 파이차트(pie Chart) : 원그래프는 전체에 대한 각 부분의 비율을 부채꼴 모양으로 백분율로 나타낸 그래프로 전체에서 차지하는 비율을 나타내며, 비율을 한눈에 볼 수 있다는 장점

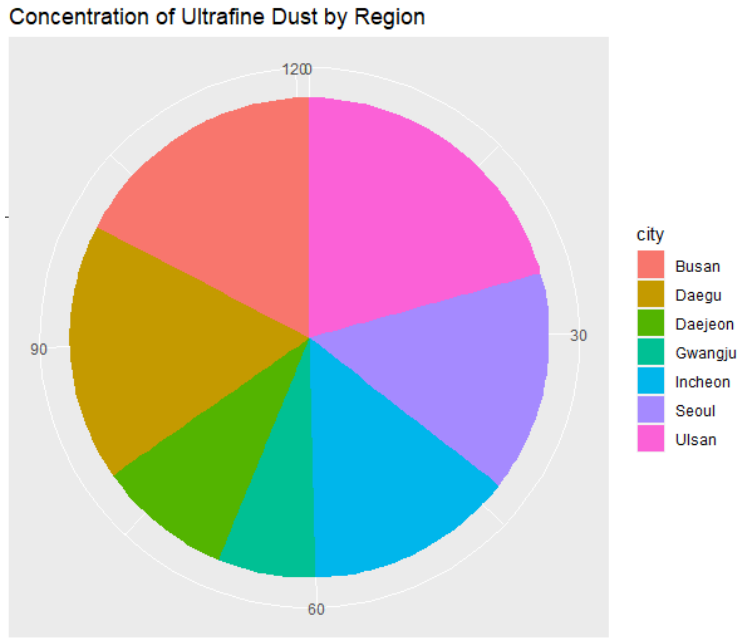


## 기술통계-탐색적 데이터 분석(EDA)

### 탐색적 데이터 분석(EDA) – Categorical

- 파이차트(pie Chart) : 원그래프는 전체에 대한 각 부분의 비율을 부채꼴 모양으로 백분율로 나타낸 그래프로 전체에서 차지하는 비율을 나타내며, 비율을 한눈에 볼 수 있다는 장점

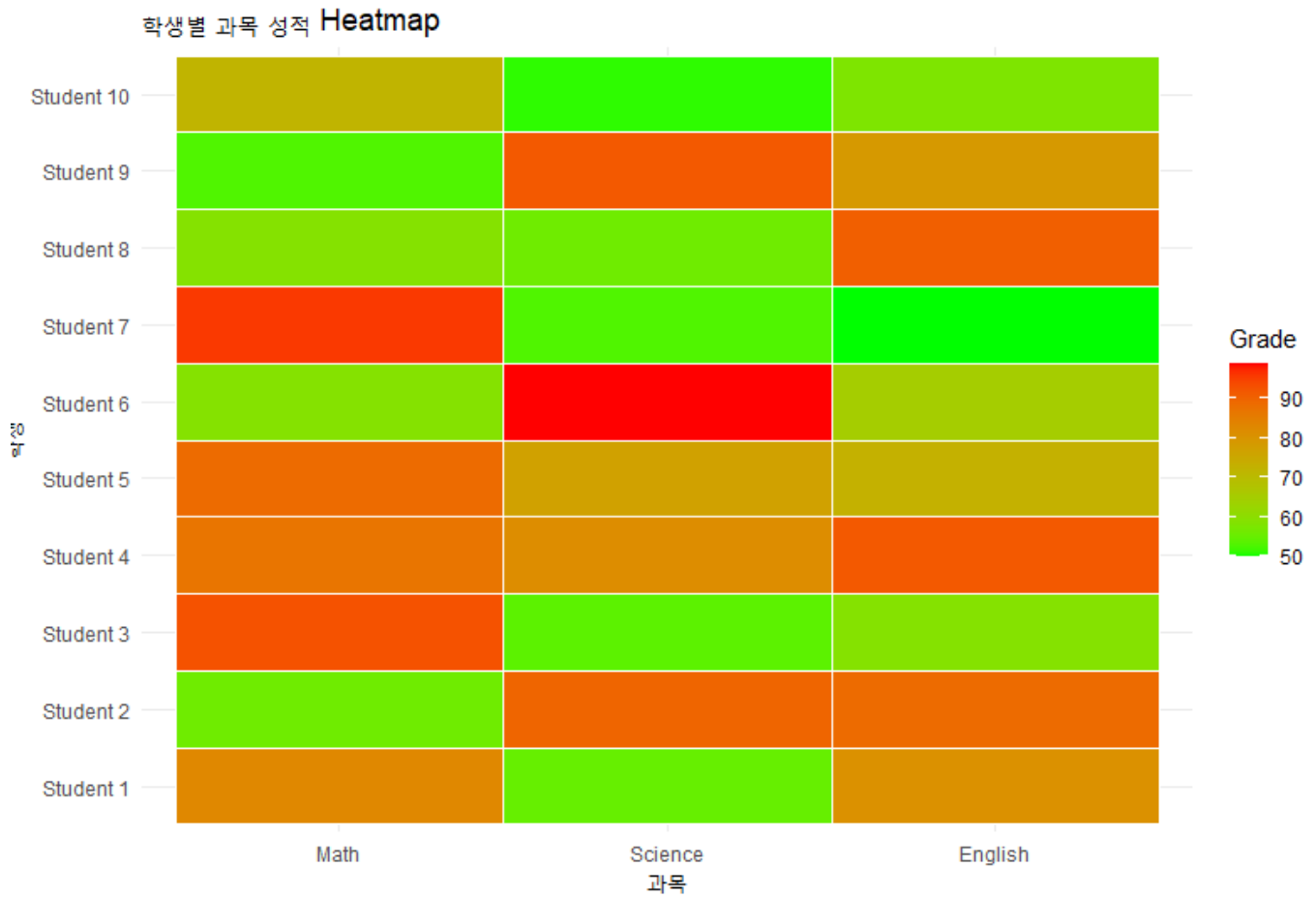
```
city <- c("Seoul", "Busan", "Daegu", "Incheon", "Gwangju", "Daejeon", "Ulsan")
pm25 <- c(18, 21, 21, 17, 8, 11, 25)
colours()
colors <- c("red", "orange", "yellow", "green", "lightblue", "blue", "violet")

df <- data.frame(city = city, pm25 = pm25, colors = colors)

ggplot(df, aes(x= "", y = pm25, fill = city)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  labs(title = "Concentration of Ultrafine Dust by Region") +
  xlab("") +
  ylab("")

 #컬러를 지정하기 위해서
 scale_fill_manual(values=colors)
```



## 기술통계-탐색적 데이터 분석(EDA)

### 탐색적 데이터 분석(EDA) – Categorical

- 히트맵(Heatmap) : 데이터를 색상의 변화를 표현하고, 복잡한 데이터의 패턴, 변화량, 밀도 등을 한눈에 파악하기에 유용



## 기술통계-탐색적 데이터 분석(EDA)

### 탐색적 데이터 분석(EDA) – Categorical

- **히트맵(Heatmap) : 데이터를 색상의 변화를 표현하고, 복잡한 데이터의 패턴, 변화량, 밀도 등을 한눈에 파악하기에 유용**

```
library(reshape2)

# 가상의 성적 데이터 생성
students <- paste("Student", 1:10) # 10명의 학생
subjects <- c("Math", "Science", "English") # 3개 과목
grades <- matrix(sample(50:100, 30, replace=TRUE), nrow=10, ncol=3, dimnames=list(students, subjects)) # 과목별 성적

# 데이터 프레임으로 변환
grades_melted <- melt(grades, id.vars = rownames(grades))
grades_melted

colnames(grades_melted) <- c("Student", "Subject", "Grade")

# Heatmap 생성
ggplot(grades_melted, aes(x=Subject, y=Student, fill=Grade)) +
  geom_tile(color = "white") + # 타일 테두리 추가
  scale_fill_gradient(low="green", high="red") + # 성적에 따른 색상 그라데이션 지정
  #theme_minimal() + # 미니멀한 테마 적용
  labs(title="학생별 과목 성적 Heatmap")+
  xlab("과목") +
  ylab("학생")
```