강원대학교
AI소프트웨어학과

재난안전 프로그래밍

Chapter 3. 기초통계

Chapter 3. 기초통계

3.1 데이터 분석이란?

- 데이터란?
  - 이론을 세우는 데 기초가 되는 사실, 또는 바탕이 되는 자료
  - 관찰이나 실험, 조사로 얻은 사실이나 자료
  - 컴퓨터가 처리할 수 있는 문자, 숫자, 소리, 그림 따위의 형태로 된 자료
  - 데이터는 신호,기호,숫자,문자 등으로 기록 됨
  - 정보를 위한 기초적인 자료를 말함
  - 정보는 데이터를 가공하지 않은 경우

Chapter 3. 기초통계

3.1 데이터 분석이란?

- 정보란? → 구성, 해석 및 맥락화 과정을 통해 데이터에서 파생

선수들의 수치

PLAYER	DPM	GOLDDIFFAT15	분당 K+A	GPM	분당 골드 차이	분당 데미지 차이
FAKER	375	232.21	0.220	396	11.732	36.575
SHOWMAKER	488	82.86	0.294	404	19.901	31.411
CHOVY	466	352.44	0.223	410	27.059	-30.201
BDD	461	-1.41	0.269	389	4.989	53.180
GORI	459	-400.44	0.239	397	3.552	-29.395
FATE	412	236.79	0.238	406	21.573	-8.689

선수들의 신체 조건에 따른 적성 진단

운동선수-일반인 기초스포츠적성진단 비교

자료: 국민체육진흥공단 체육과학연구원

운동선수 (일반인 평균)	체지방률 (16.1%)	팔굽혀펴기 (60.6회)	농구공던지기 (661.1cm)	왔동일으키기 (67.7회/2분)	제자리멀리뛰기 (224.7cm)	오래달리기 (465.5초/1600m)
역도	16.9	76.3	1010.7	66.3	285.0	480.9
야구	17.0	86.4	1004.2	88.0	268.2	357.3
축구	13.0	116.7	1025.6	97.2	258.8	311.0
핸드볼	12.2	108.6	1065.0	91.7	263.9	334.6
양궁	14.1	54.3	919.3	71.6	249.3	382.6
농구	18.2	61.7	1346.1	94.4	256.9	354.1
유도	15.2	118.5	926.1	82.7	246.7	347.8
수영	11.2	63.1	942.1	91.2	251.1	436.1
체조	8.8	65.2	745.3	96.1	233.9	459.6

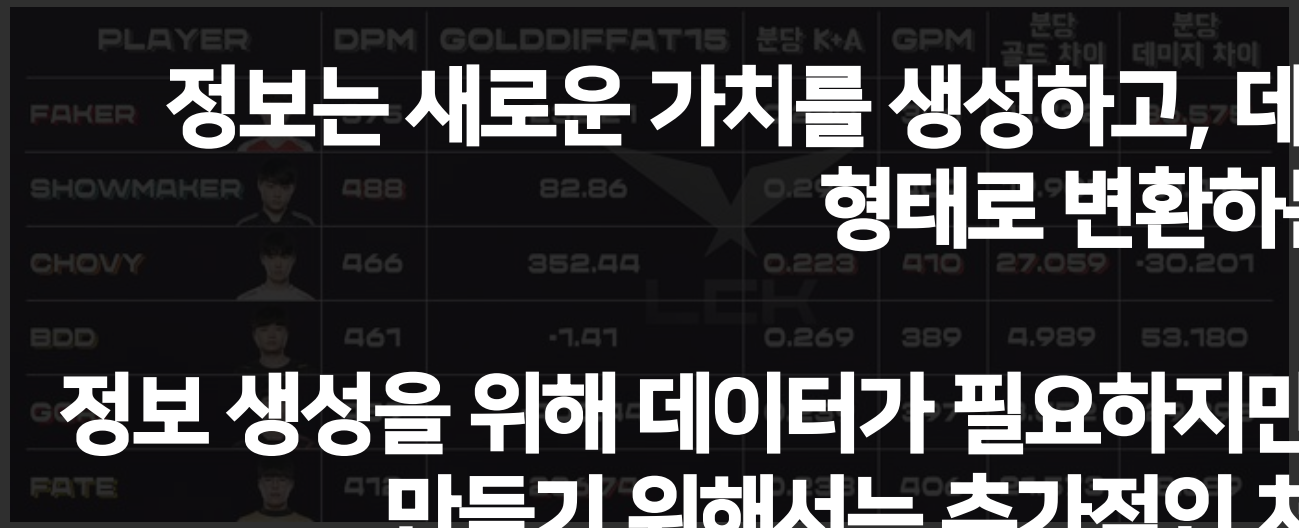
Chapter 3. 기초통계

3.1 데이터 분석이란?

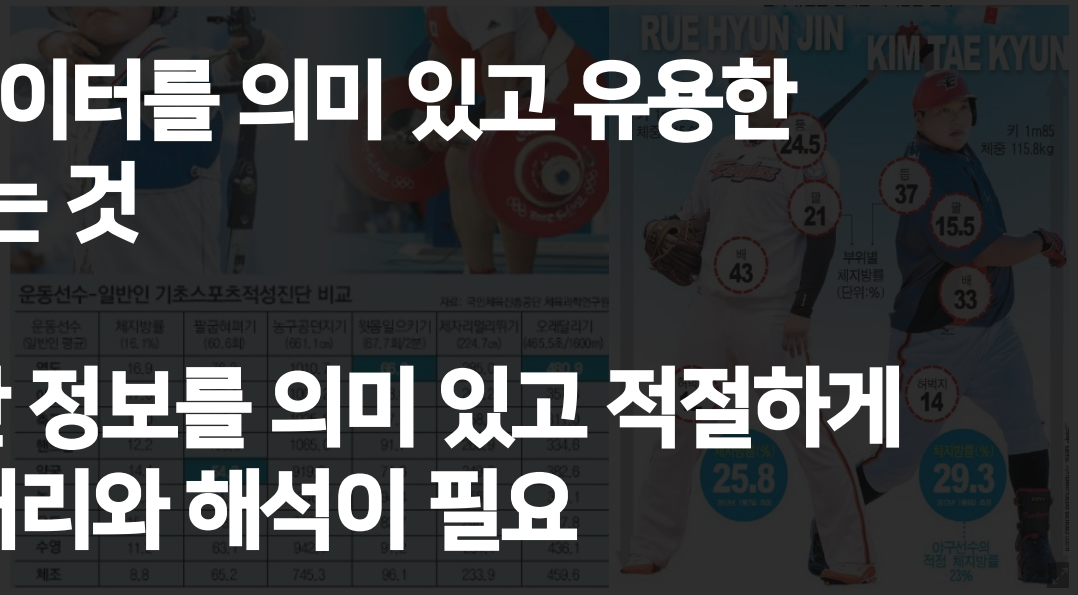
- 정보란? → 구성, 해석 및 맥락화 과정을 통해 데이터에서 파생

데이터는 정보가 생성되는 원재료

선수들의 수치



선수들의 신체 조건에 따른 적성 진단



정보는 새로운 가치를 생성하고, 데이터를 의미 있고 유용한 형태로 변환하는 것

정보 생성을 위해 데이터가 필요하지만 정보를 의미 있고 적절하게 만들기 위해서는 추가적인 처리와 해석이 필요

Chapter 3. 기초통계

3.1 데이터 분석이란?

통계란 무엇일까?

- 데이터 수집, 기술통계, 추론통계, 확률, 샘플링, 가설검정
- 데이터에서 유효한 결론을 도출하여 실제 문제를 해결하고 삶과 비즈니스의 다양한 측면을 개선하는 데 도움
- 통계의 모델의 적용을 통해 통찰력을 얻고 정보를 전달해 어떠한 문제의 결정을 내리는 것
- 통계는 데이터를 사용하여 결론을 도출하고 정보를 제공하여 문제를 해결하거나 다양한 현상에 대한 통찰력을 얻는 것

Chapter 3. 기초통계

3.1 데이터 분석이란?

통계가 생기게 된 계기

- 경험을 토대로 같거나 비슷한 문제가 발생했을 때, 이를 해결하기 위해 사용됨
- 각자의 상황에 대해 경험하고, 학습해 결과를 도출할 수 있음 → 기억에는 한계가 존재하고, 왜곡이 가능함
- 이러한 문제를 해결하기 위해 기록이란 것이 생겼고, 다양한 방법론들이 나오게 됨 → 수를 추정하기 위해
- 고대 문명: 가장 초기 형태의 통계는 이집트, 메소포타미아, 중국과 같은 고대 문명에서 찾을 수 있으며, 그곳에서 데이터는 세금, 인구 수, 토지 조사와 같은 목적으로 수집

Chapter 3. 기초통계

3.1 데이터 분석이란?

통계가 생기게 된 계기

- 통계를 통해 문제를 해결하려면 일반적으로 하나 이상의 가설을 설정해야 함
- 명확하고 검증 가능한 가설을 세우는 것은 분석을 수행하고 데이터에서 의미 있는 결론을 도출하기 위한 프레임워크를 제공
- 가설에는 샘플 정보를 기반으로 모집단에 대한 추론 또는 결론 도출과 관련된 문제를 해결하기 위한 데이터가 필요함

Chapter 3. 기초통계

3.1 데이터 분석이란?

통계의 학습 프로세스

- 가설
  - 확률이론, 가설검증, 통계적 추론 등 통계적 방법의 기초가 되는 이론적 개념과 원리

- 데이터
  - 통계 분석을 위한 입력으로 사용되는 정보 또는 관찰의 수집
  - 데이터는 설문 조사, 실험, 관찰 또는 기타 방법을 통해 수집할 수 있음

Chapter 3. 기초통계

3.1 데이터 분석이란?

통계의 학습 프로세스

- 모델
  - 일상속의 현상이나 프로세스의 수학적 또는 통계적 표현
  - 모델은 변수 간의 관계를 설명하고 데이터를 기반으로 예측하는 데 사용됨

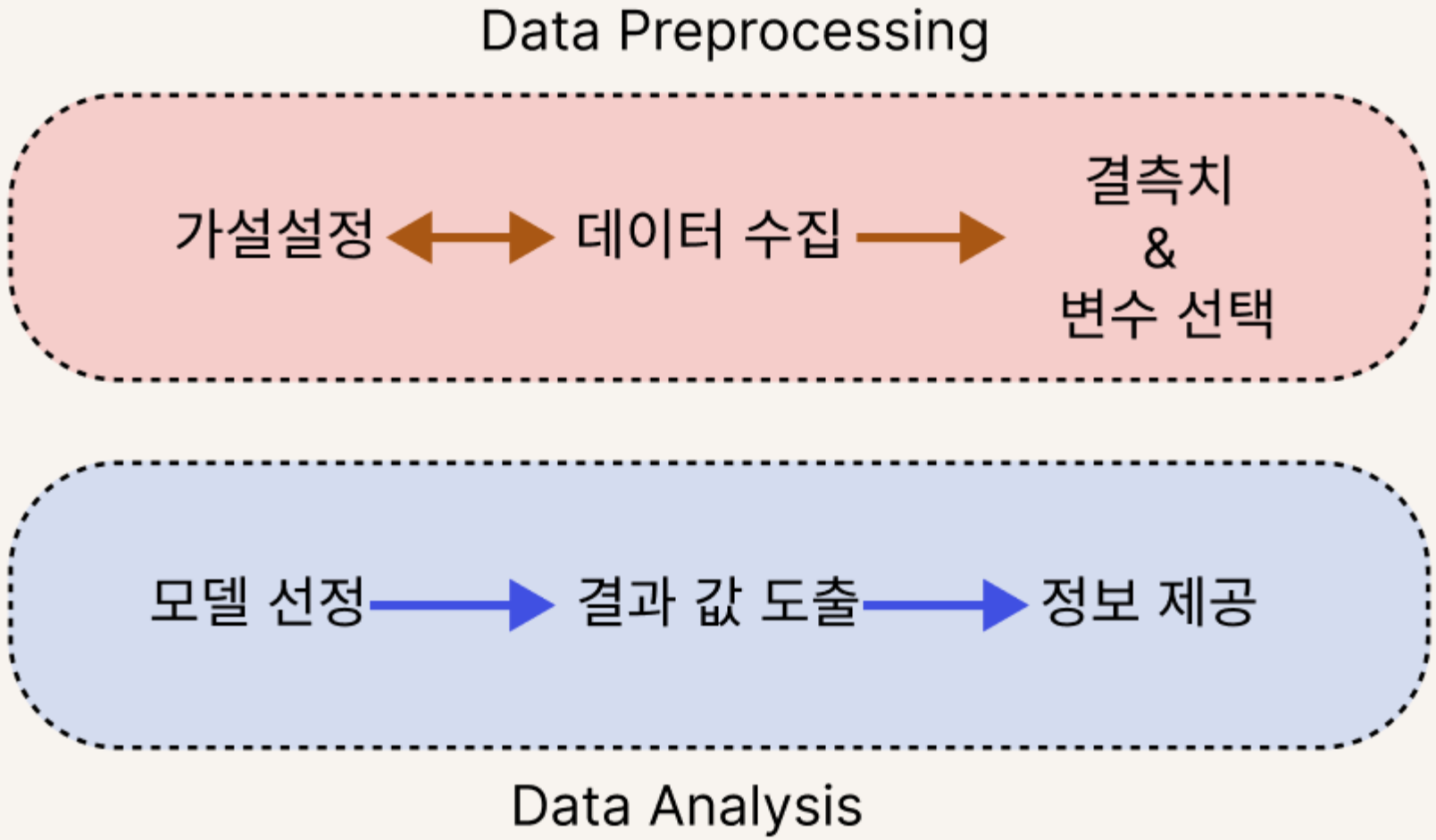
- 결과
  - 데이터의 통계 분석에서 얻은 결과 또는 발견
  - 결과는 수행된 데이터 및 분석을 기반으로 통찰력, 패턴 또는 결론을 제공

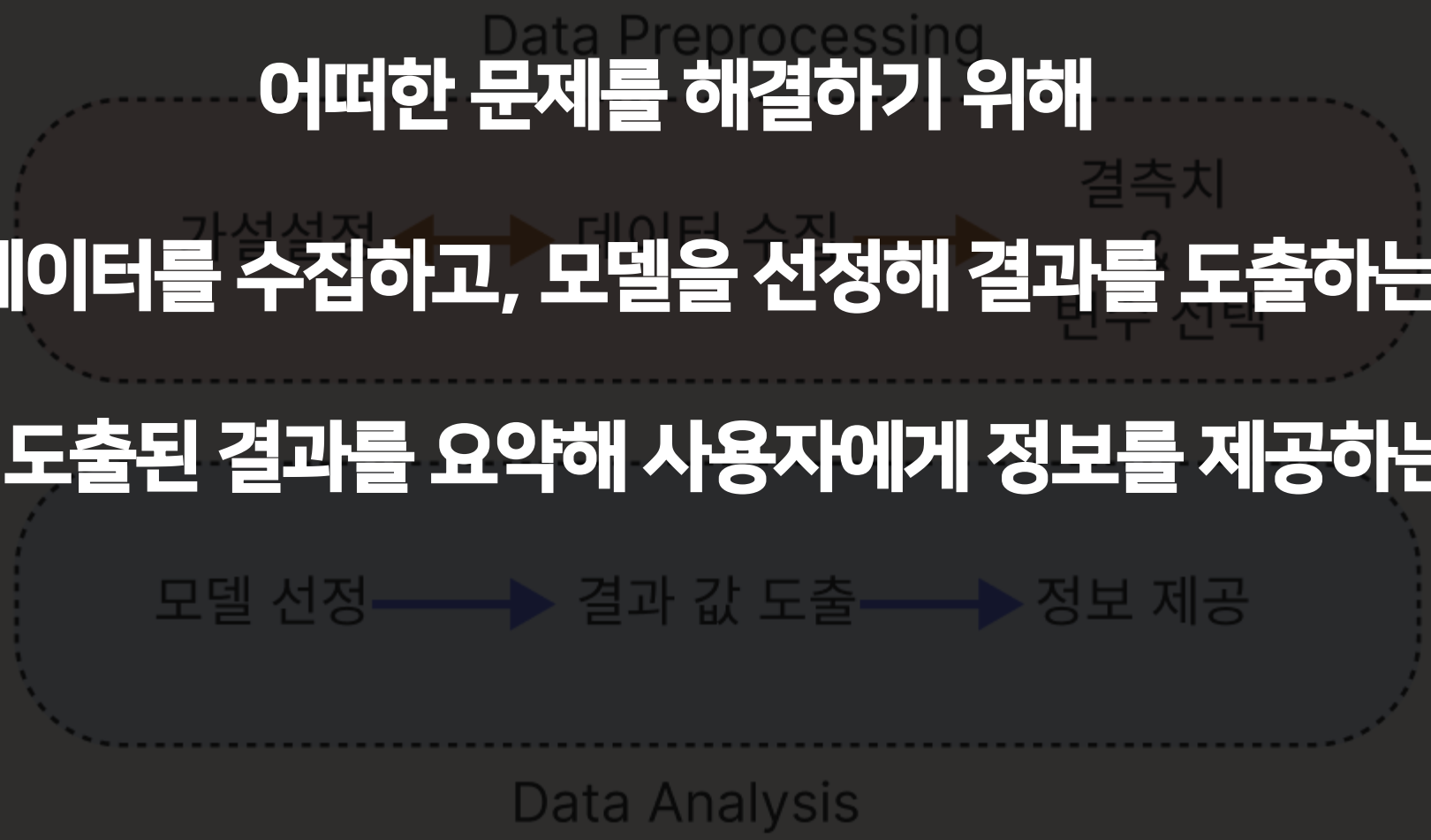
- 정보제공
  - 통계는 정보에 입각한 결정을 내리고, 결론을 도출하고, 개체군이나 현상에 대한 가설이나 주장을 뒷받침하기 위해 데이터를 분석하고 해석하여 정보를 제공

Chapter 3. 기초통계

3.1 데이터 분석이란?

통계의 학습 프로세스





어떠한 문제를 해결하기 위해

분석 : 데이터를 수집하고, 모델을 선정해 결과를 도출하는 과정

정보 : 도출된 결과를 요약해 사용자에게 정보를 제공하는 것

Chapter 3. 기초통계

3.1 데이터 분석이란?

데이터의 유형

- **변수란? → 데이터(data)를 저장하기 위해 프로그램에 의해 이름을 할당받은 메모리 공간**

데이터 종류	변수명	내용	예시
범주형 데이터	명목형 변수	순서나 순위를 암시하지 않고 데이터를 범주화 하는 변수	성별, 머리 색깔, 과일 종류
	순위형 변수	순서가 있는 범주이거나, 범주 간의 거리가 일정하지 않거나 알려지지 않은 변수	교육 수준, 설문 응답
수치형 데이터	이산형 변수	고유하고 개별적인 값을 갖는 계산가능한 숫자 변수	교통사고 발생수, 참가 인원
	연속형 변수	주어진 범위 내에서 무한한 수의 값을 가질 수 있는 숫자 변수	몸무게, 키, 평균 성적

Chapter 3. 기초통계

3.2 기술통계

중심 경향 측정

- 평균 : 데이터 세트에 있는 모든 데이터 포인트의 산술 평균
- 중앙값 : 데이터 세트에서 가장 작은 것부터 큰 순서로 정렬할 때 중간 값
- 최빈값 : 데이터 세트에서 가장 자주 발생하는 값
- 최대값/최소값 : 데이터 세트에서 가장 큰 값/ 데이터 세트에서 가장 작은 값

```
data <- c(10, 15, 20, 25, 30, 15, 20, 25, 25, 10)
mean_value <- mean(data)
median_value <- median(data)
max_value <-max(data)
min_value <-min(data)
mode_value <- table(data)
mode_values <- names(frequency_table)[frequency_table == max(frequency_table)]
```

Chapter 3. 기초통계

3.2 기술통계

변동성 측정

- 범위: 데이터 세트의 최대값과 최소값의 차이
- 사분위수 범위(IQR): 데이터의 중간 50%를 나타내는 첫 번째 사분위수(25번째 백분위수)와 세 번째 사분위수(75번째 백분위수) 사이의 값 범위

```
data <- c(10, 15, 20, 25, 30, 15, 20, 25, 25, 10)
range_value <- max(data) - min(data)
df <- data.frame(values = c(5, 7, 10, 12, 14, 18, 20, 22, 25, 27, 30))
Q1 <- quantile(data, 0.25)
Q3 <- quantile(data, 0.75)
iqr_value <- Q3 - Q1
```

Chapter 3. 기초통계

3.2 기술통계

변동성 측정

- 분산: 각 데이터 포인트와 평균 사이의 평균 제곱 차이
- 표준 편차: 데이터가 평균에서 얼마나 퍼져 있는지를 측정함

$$\text{Variance}(\sigma^2) = \frac{\sum((x_i - \mu)^2}{n} \qquad \textit{Standard Deviation } (\sigma) = \sqrt{\sigma^2}$$

$$Mean(\mu) = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

Chapter 3. 기초통계

3.2 기술통계

변동성 측정

- 분산 또는 표준 편차가 높을수록 데이터 포인트가 더 분산되어 더 큰 변동성 또는 분산을 나타냄
- 낮은 분산 및 표준 편차는 데이터 포인트가 평균에 가깝다는 것을 의미하며 더 일관되고 예측 가능한 데이터 세트를 나타냄
- 반대로 높은 분산 및 표준 편차는 더 많은 변동성과 낮은 일관성을 나타냄
- 높은 분산 및 표준 편차는 데이터에 이상값이 있음을 나타내는 지표로 사용가능 함

```
data <- c(10, 15, 20, 25, 30, 15, 20, 25, 25, 10)
mean_data <- mean(data)
squared_diff <- (data - mean_data)^2
variance <- sum(squared_diff) / length(data)
std_dev <- sqrt(variance)
```

Chapter 3. 기초통계

3.2 기술통계

변동성 측정

- 분산: 각 데이터 포인트와 평균 사이의 평균 제곱 차이
- 표준 편차: 데이터가 평균에서 얼마나 퍼져 있는지를 측정함

```
data <- c(10, 15, 20, 25, 30, 15, 20, 25, 25, 10)
sd_value <- sd(data)
var_value <- var(data)
```

Chapter 3. 기초통계

3.2 기술통계

데이터의 활용

- 중심 극한을 이루는 수치형 데이터에 주로 사용됨
- 분산이 너무 크면 결과를 저해할 수 있음

이상값&결측값

- 이상값 : 이상값은 데이터 세트의 다른 관찰에서 크게 벗어나는 데이터 포인트
- 나머지 데이터를 고려할 때 예상할 수 있는 것과 현저하게 다른 값 → 데이터 수집 또는 기록의 잠재적 이상 또는 오류의 결과
- 결측값 : 사용자가 잘못 입력하거나 누락한 값

극한값

- 극한값 : 데이터 세트의 최소값과 최대값을 나타냄 → 분포의 양쪽 끝에서 가장 극단적인 값
- 일반적으로 오류나 비정상적인 상황으로 인한 결과임을 암시하는 증거가 없는 한 데이터 세트에 유지됨

Chapter 3. 기초통계

3.3 Exploratory Data Analysis

데이터를 표현하는 방법 → 탐색적 데이터 분석(EDA)

- 바 차트
- 산포도
- 히스토그램
- 박스 플롯
- 파이 차트

Chapter 3. 기초통계

3.3 Exploratory Data Analysis

데이터를 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)
         )
```

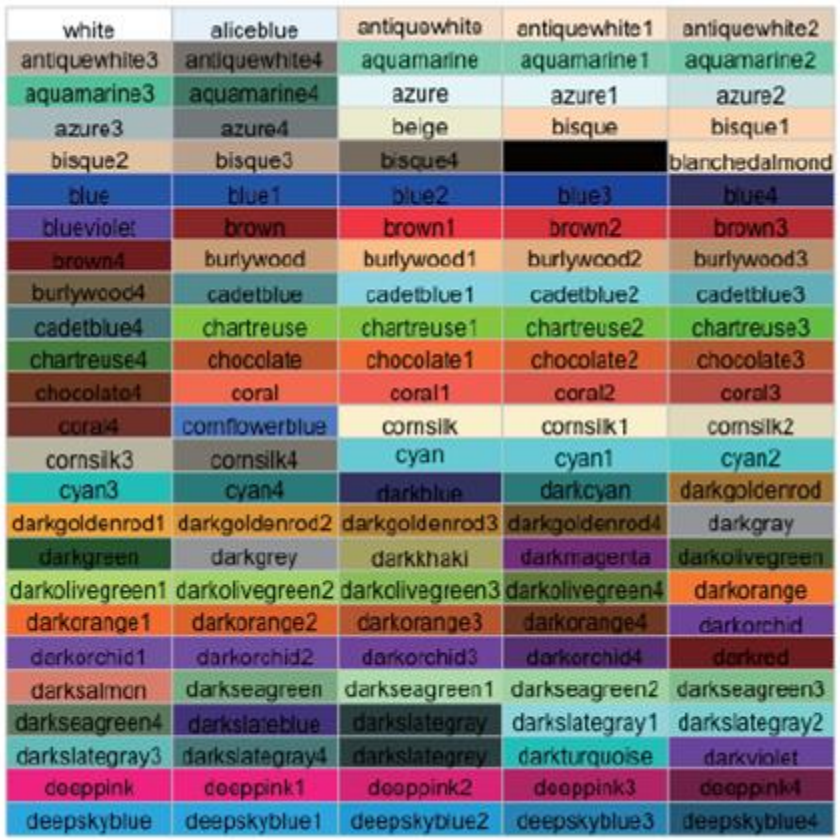
Chapter 3. 기초통계

3.3 Exploratory Data Analysis

데이터를 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"
```

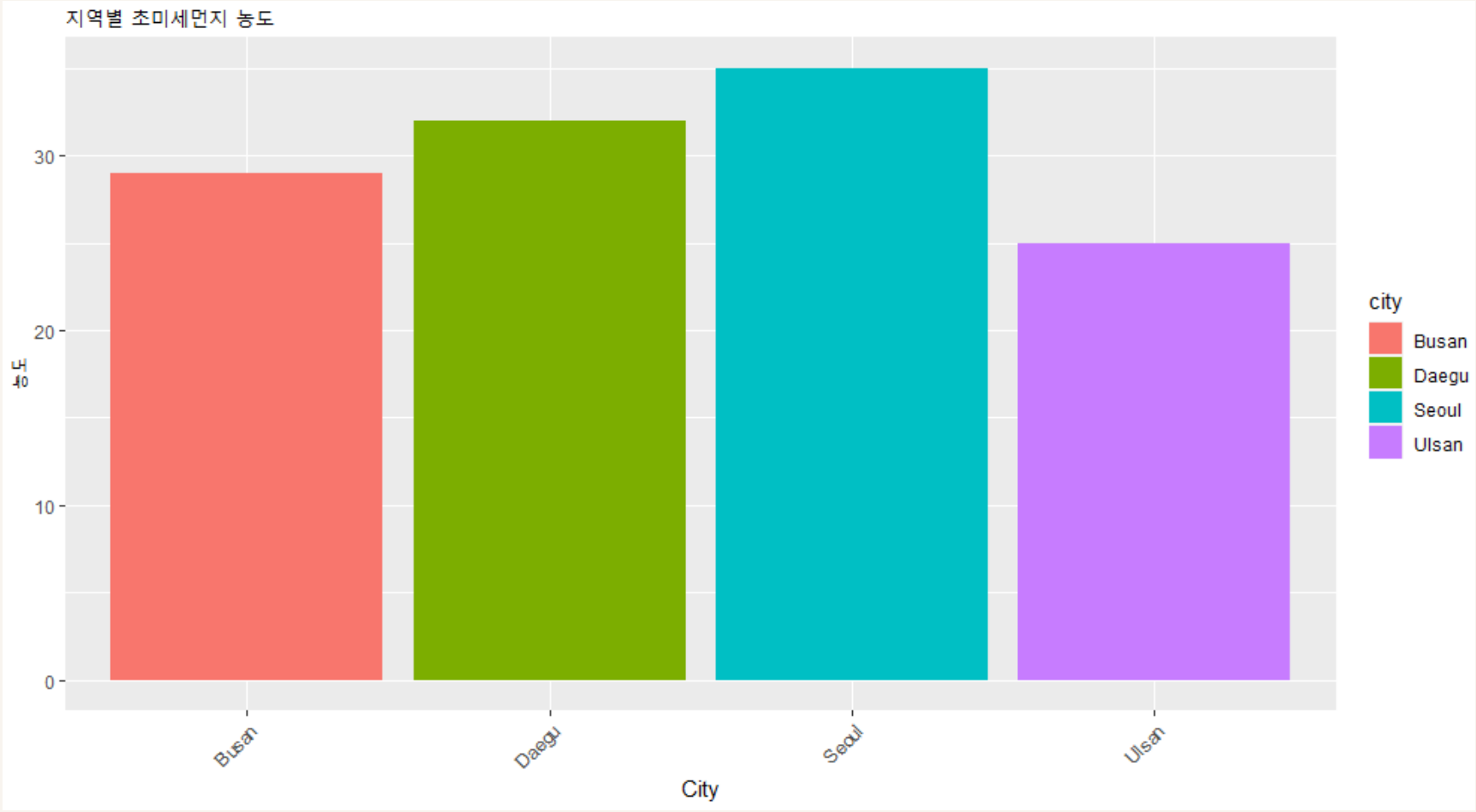


Chapter 3. 기초통계

3.3 Exploratory Data Analysis

데이터를 표현하는 방법

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



Chapter 3. 기초통계

3.3 Exploratory Data Analysis

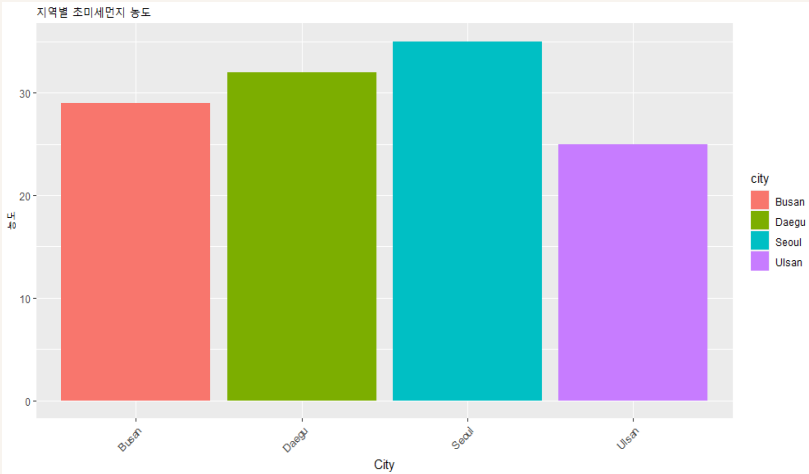
데이터를 표현하는 방법

- **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")
```

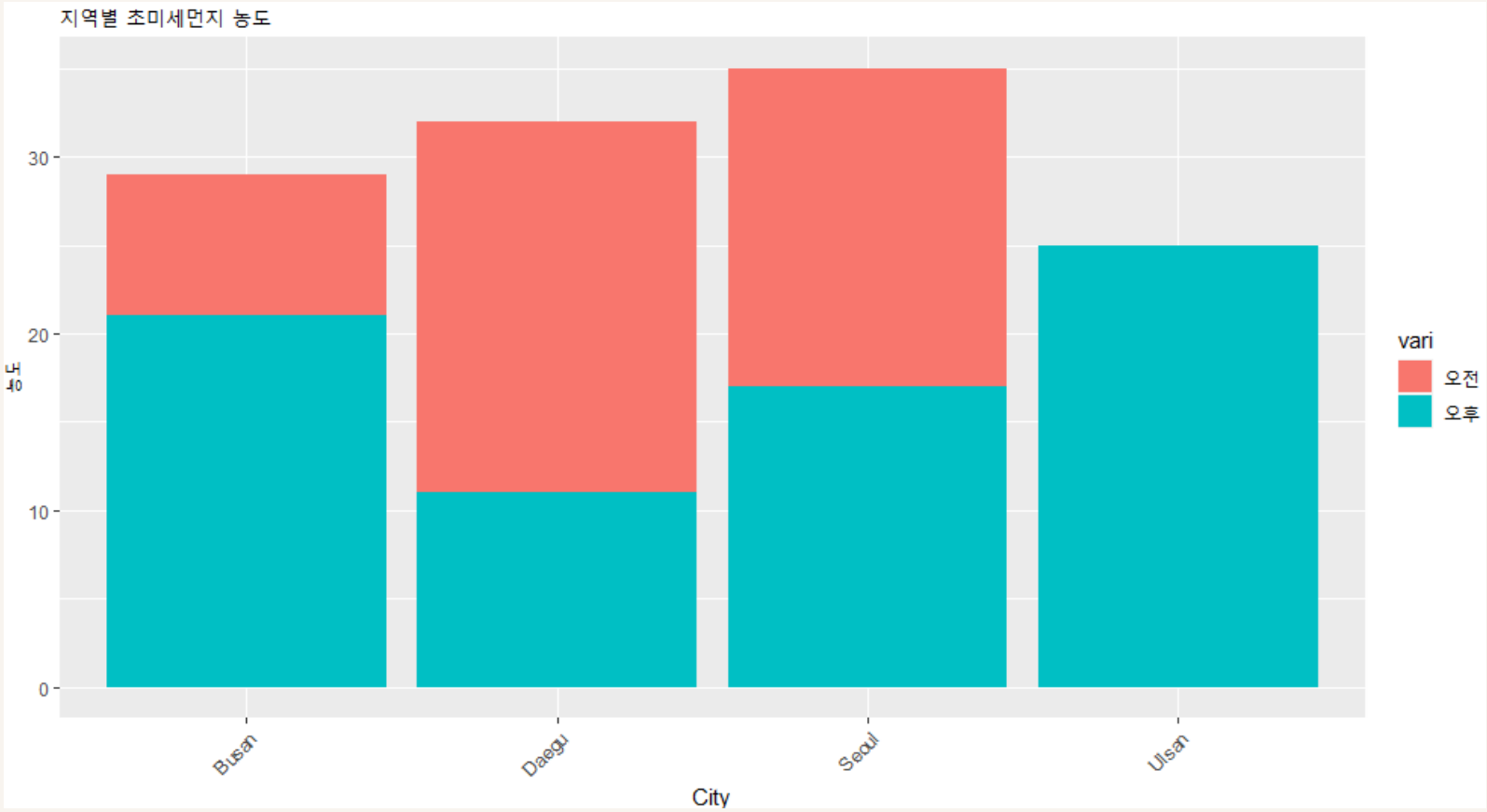


Chapter 3. 기초통계

3.3 Exploratory Data Analysis

데이터를 표현하는 방법

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



Chapter 3. 기초통계

3.3 Exploratory Data Analysis

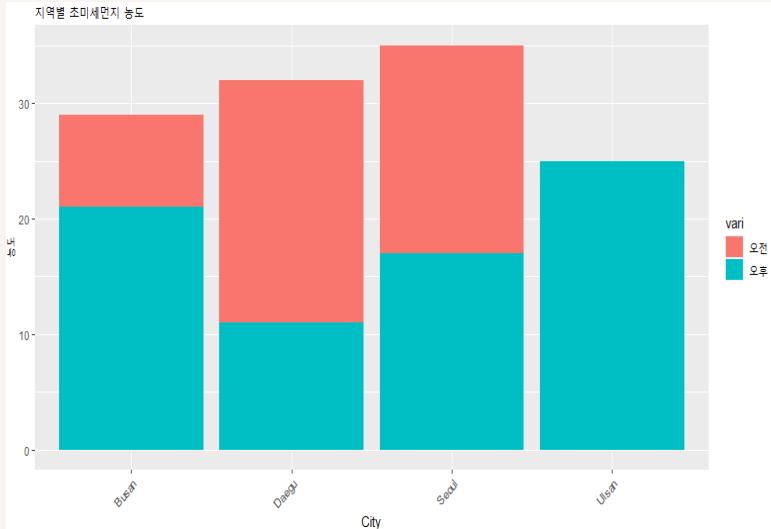
데이터를 표현하는 방법

- **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")
```

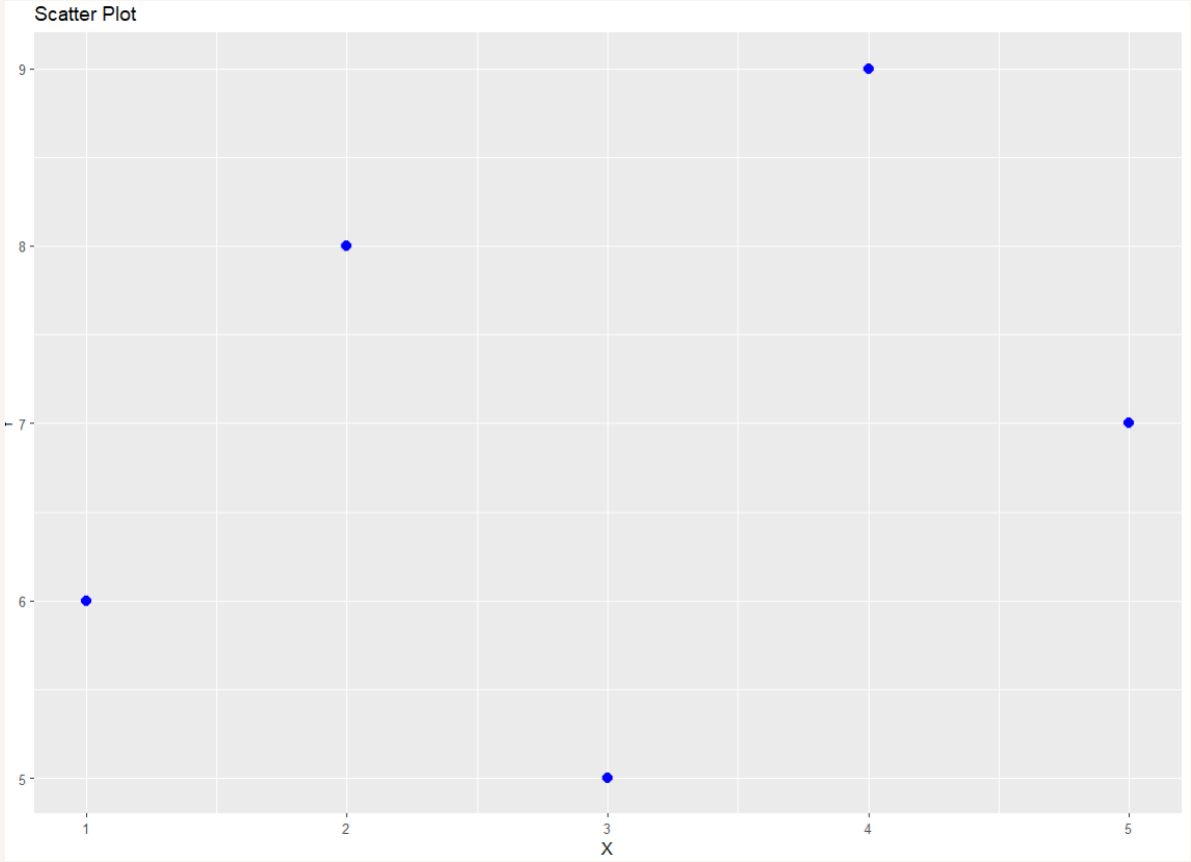


Chapter 3. 기초통계

3.3 Exploratory Data Analysis

데이터를 표현하는 방법

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



Chapter 3. 기초통계

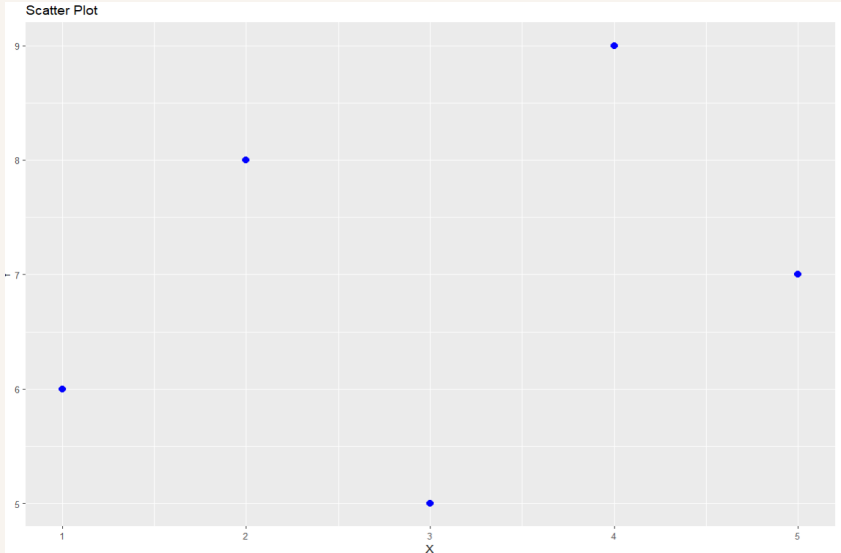
3.3 Exploratory Data Analysis

데이터를 표현하는 방법

- **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")
```

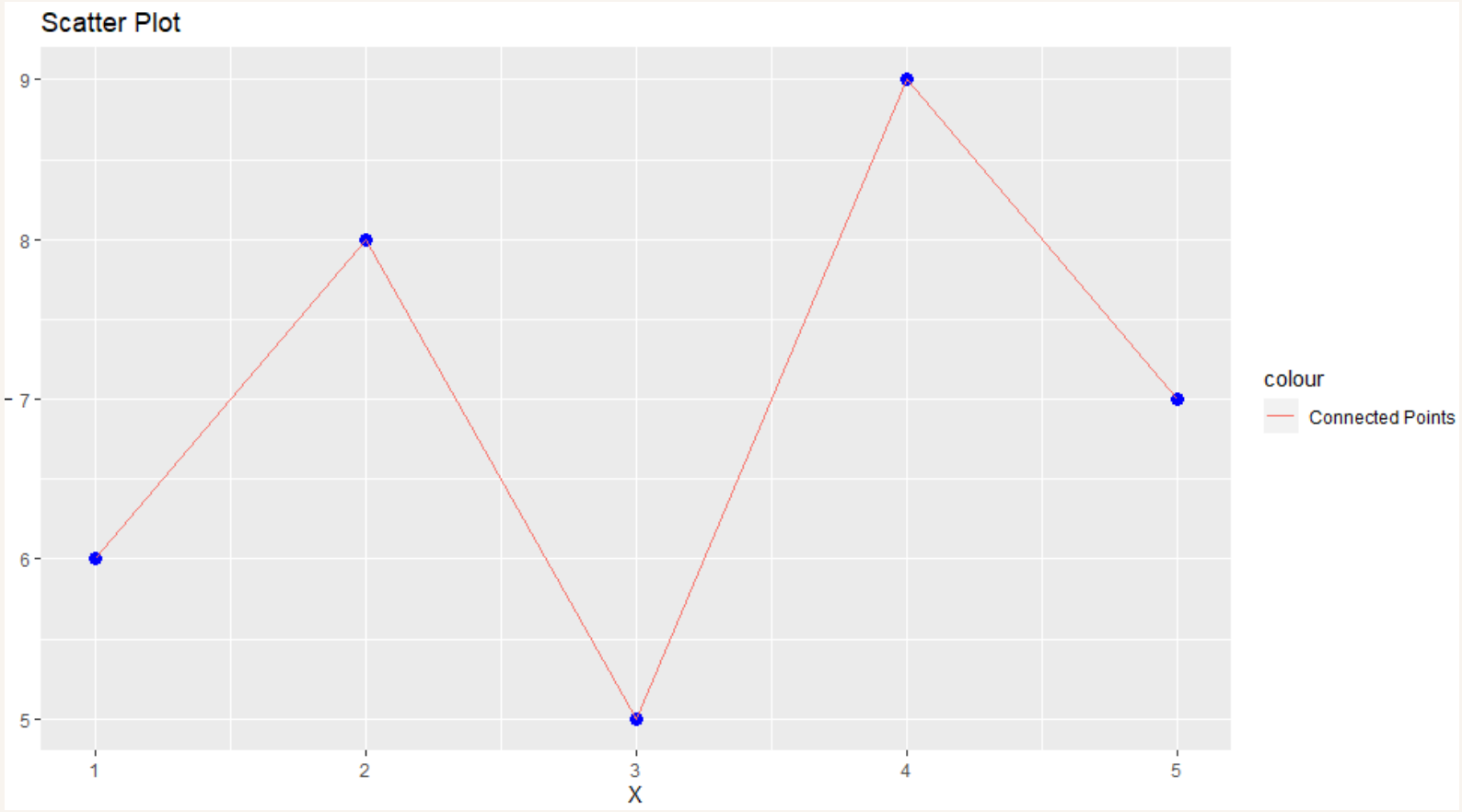


Chapter 3. 기초통계

3.3 Exploratory Data Analysis

데이터를 표현하는 방법

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



Chapter 3. 기초통계

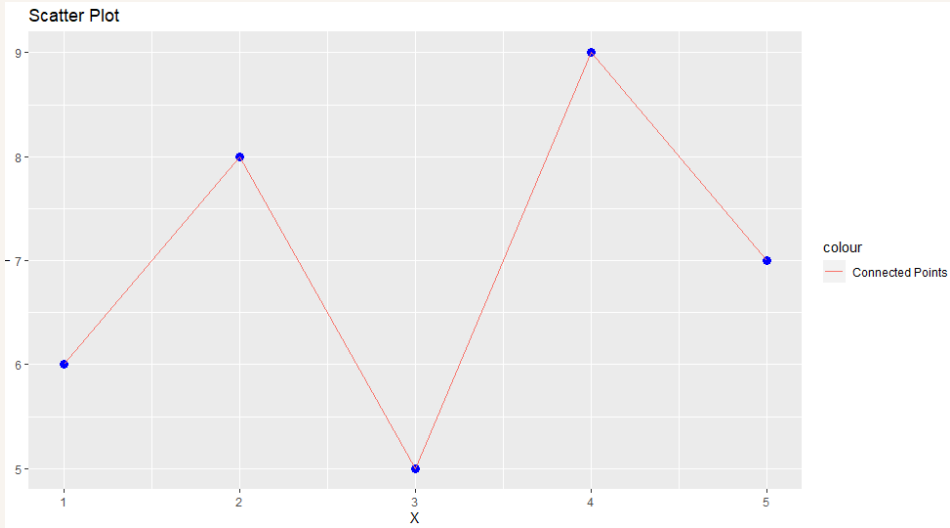
3.3 Exploratory Data Analysis

데이터를 표현하는 방법

- **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")
```

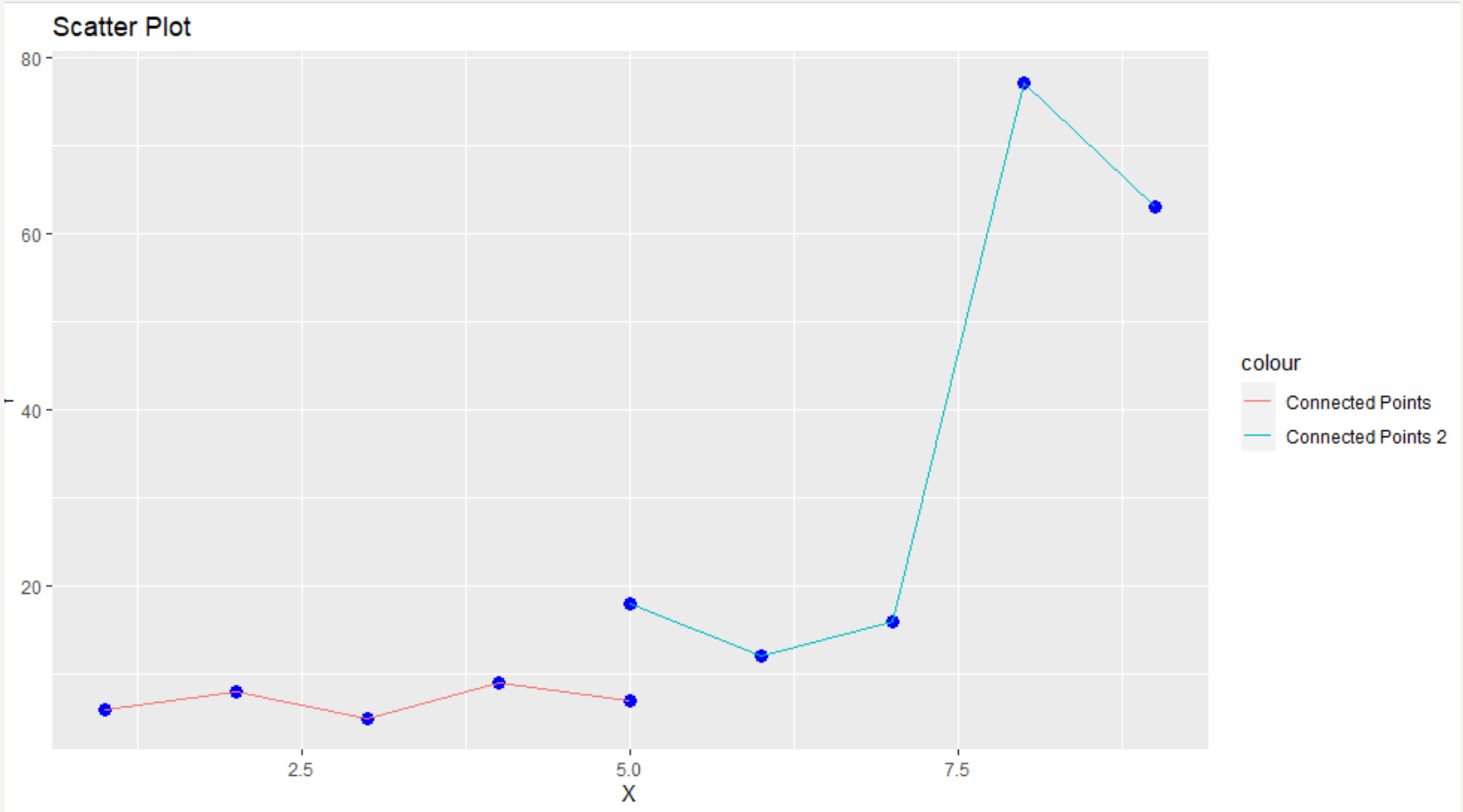


Chapter 3. 기초통계

3.3 Exploratory Data Analysis

데이터를 표현하는 방법

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



Chapter 3. 기초통계

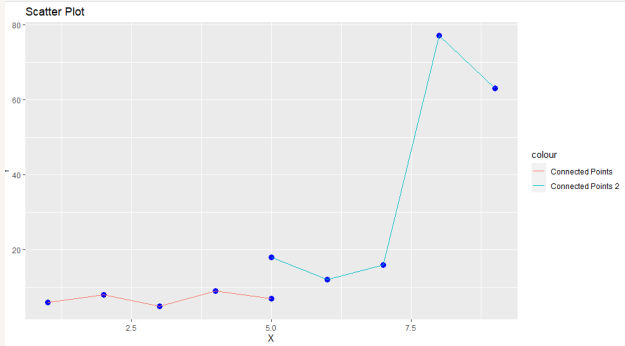
3.3 Exploratory Data Analysis

데이터를 표현하는 방법

- 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")
```

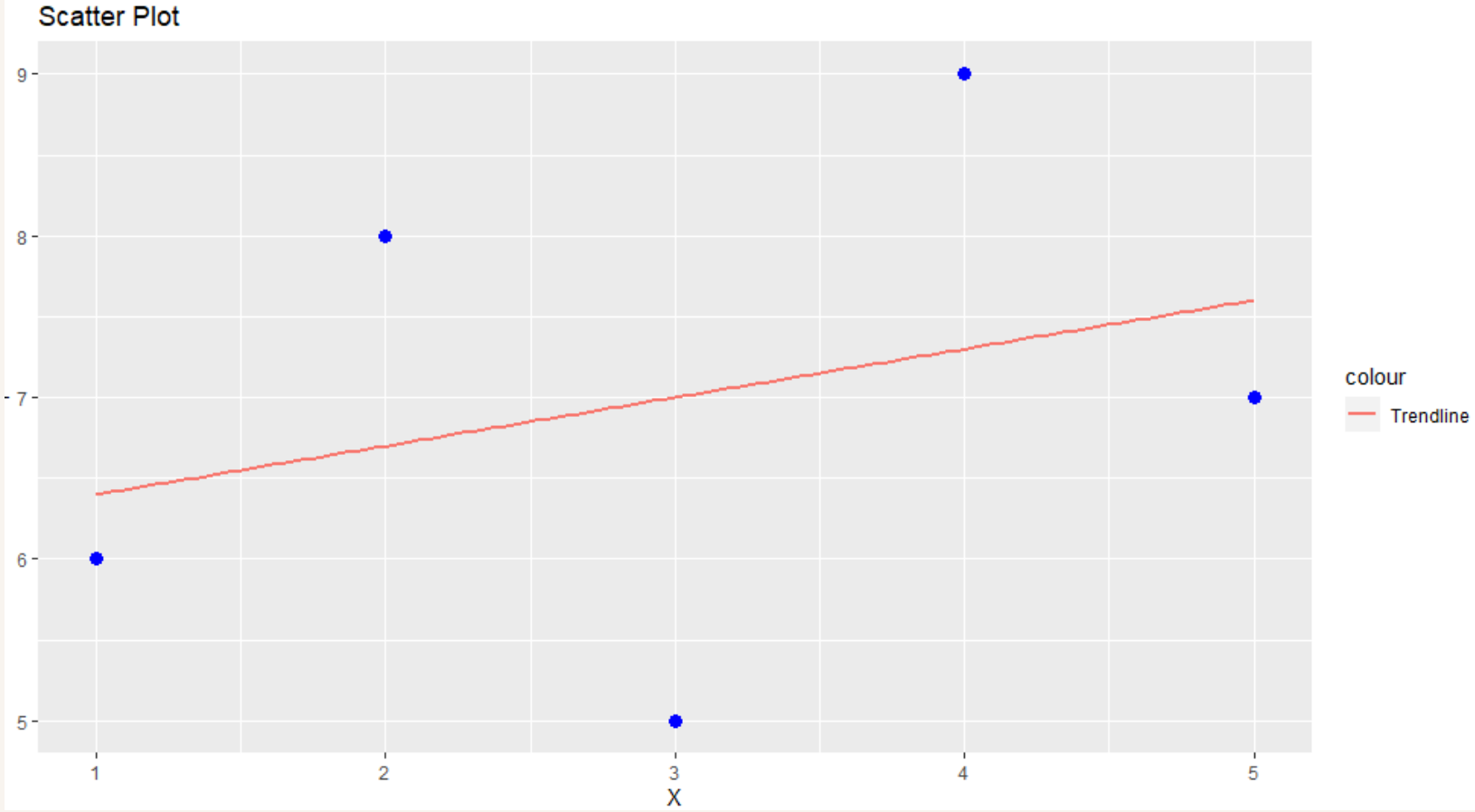


Chapter 3. 기초통계

3.3 Exploratory Data Analysis

데이터를 표현하는 방법

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



Chapter 3. 기초통계

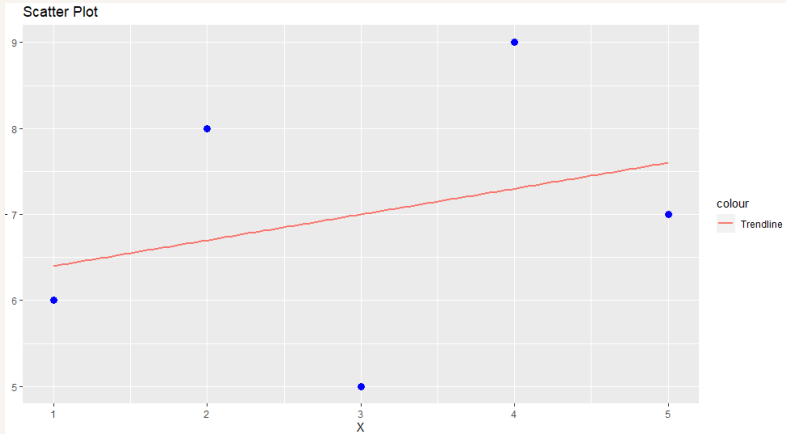
3.3 Exploratory Data Analysis

데이터를 표현하는 방법

- **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")
```

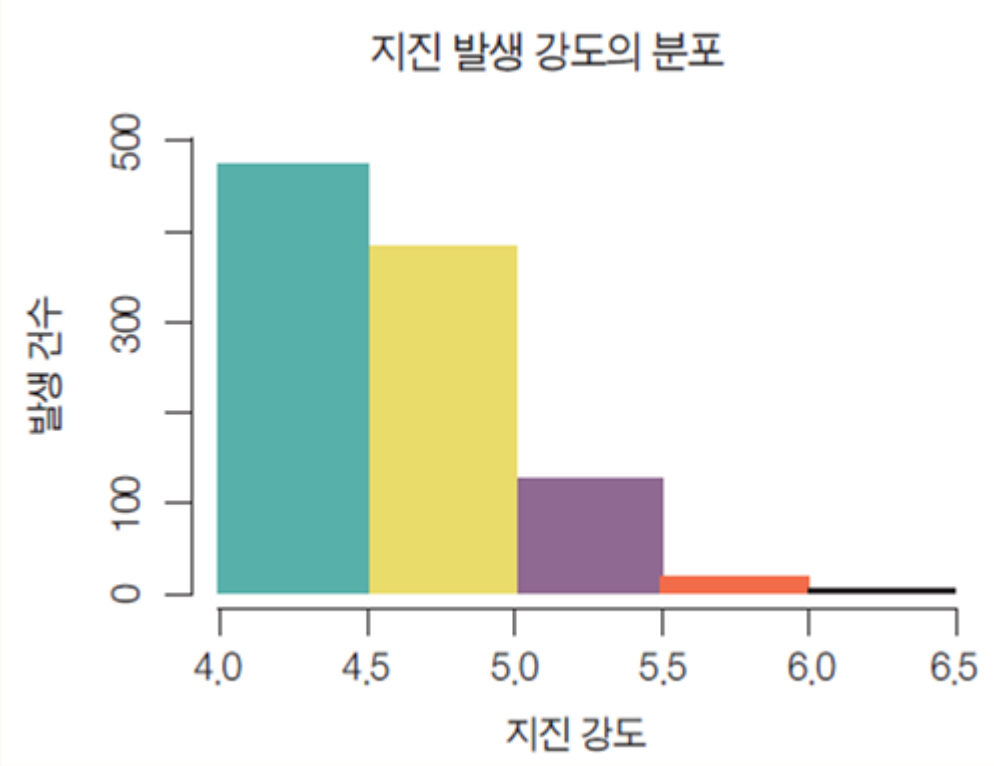


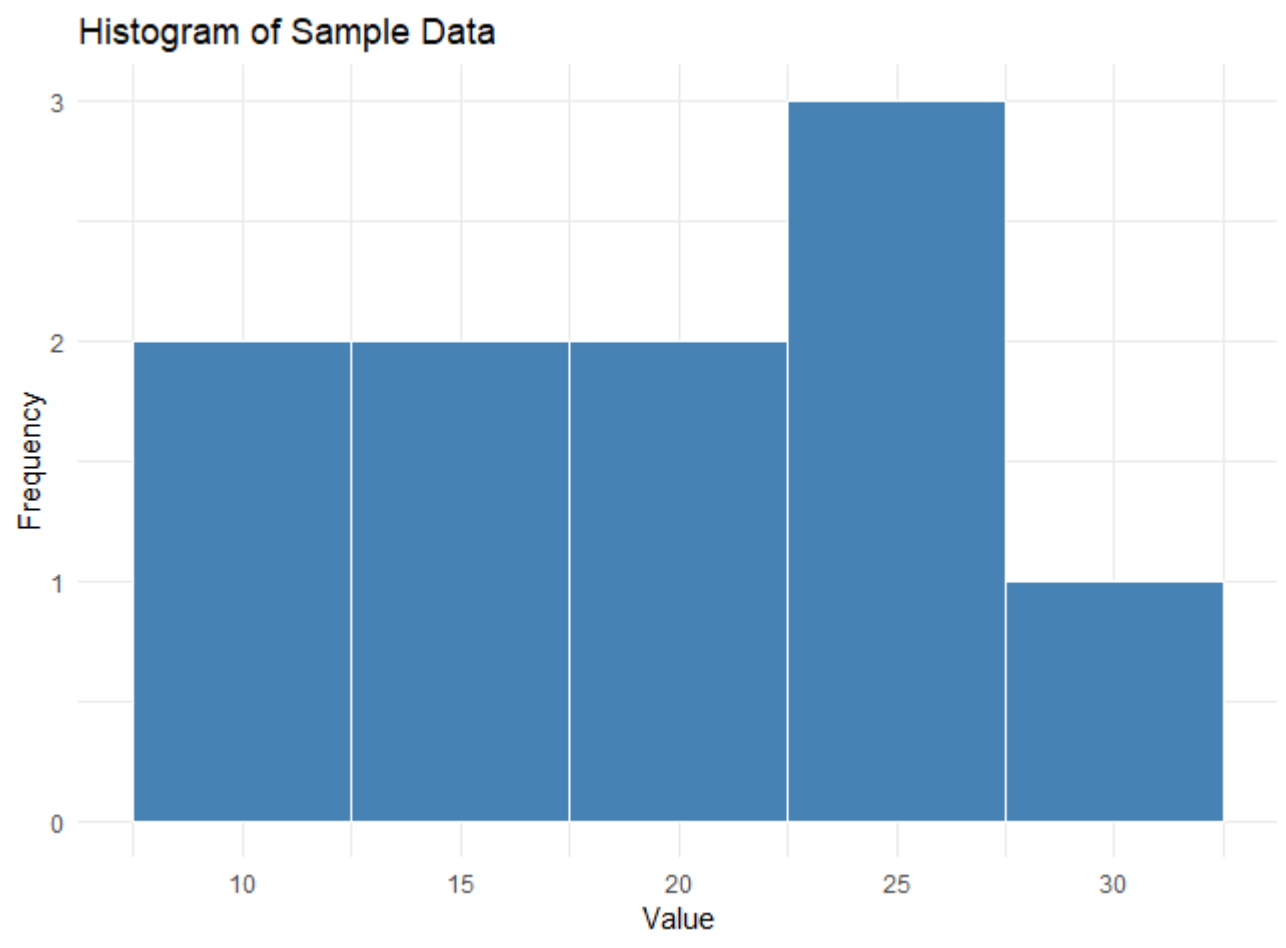
Chapter 3. 기초통계

3.3 Exploratory Data Analysis

데이터를 표현하는 방법

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





Chapter 3. 기초통계

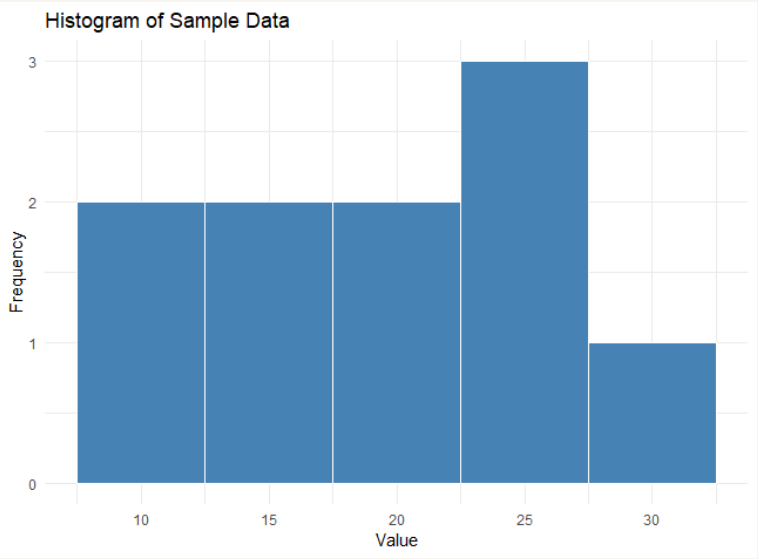
3.3 Exploratory Data Analysis

데이터를 표현하는 방법

- **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))

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

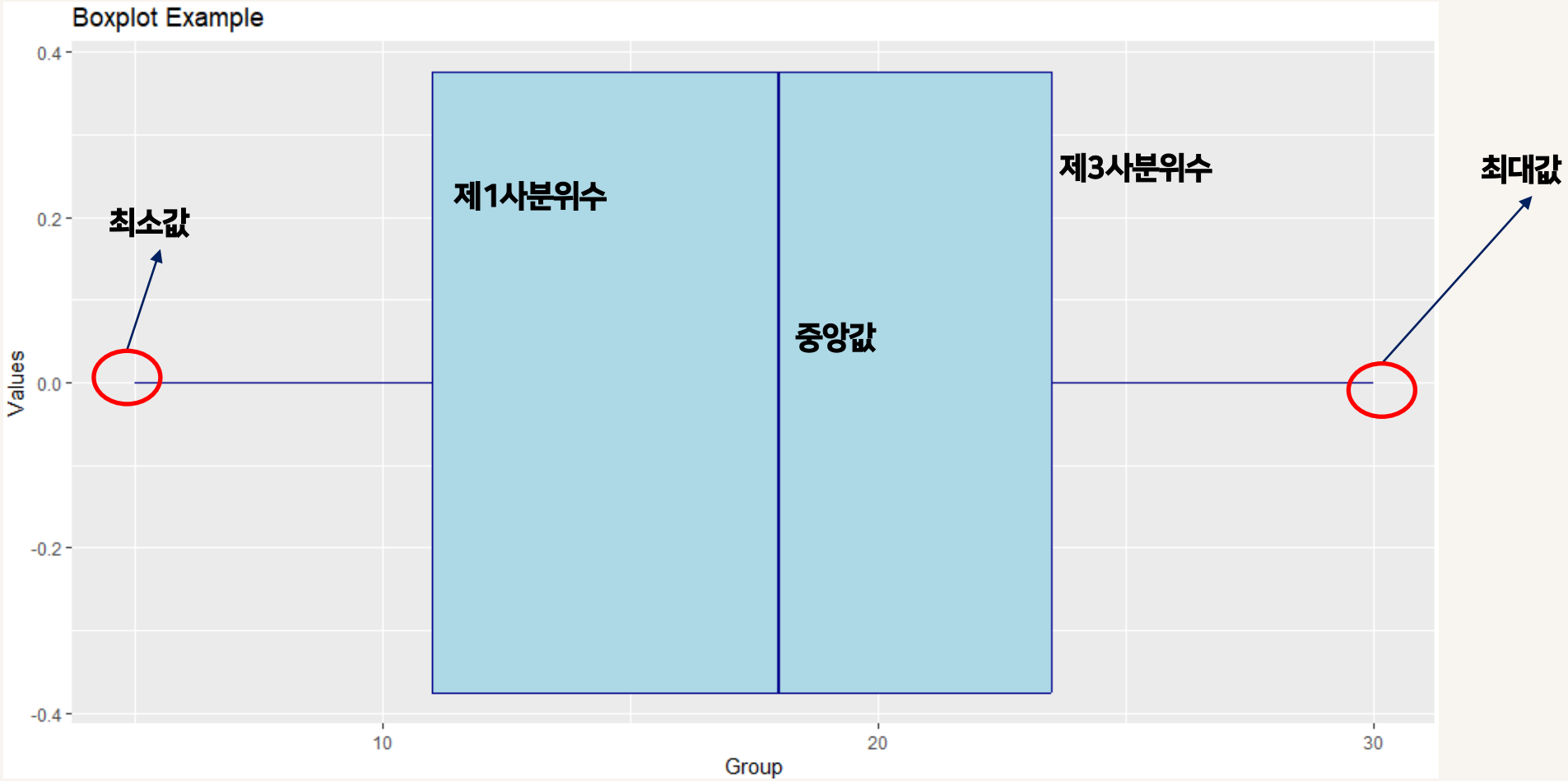


Chapter 3. 기초통계

3.3 Exploratory Data Analysis

데이터를 표현하는 방법

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



Chapter 3. 기초통계

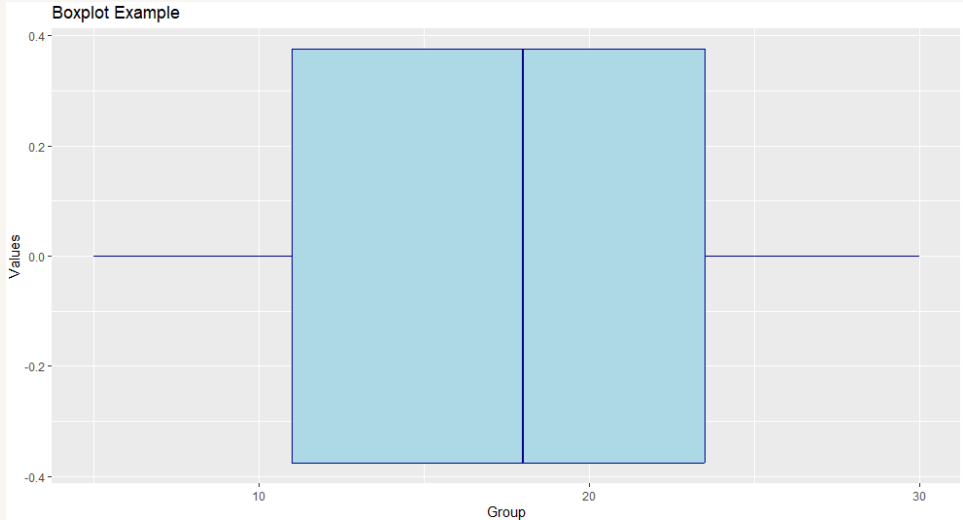
3.3 Exploratory Data Analysis

데이터를 표현하는 방법

- **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_histogram(binwidth = 5, fill = "steelblue", color = "white") +
  labs(title = "Histogram of Values") +
  xlab("Values") +
  ylab("Frequency")
```

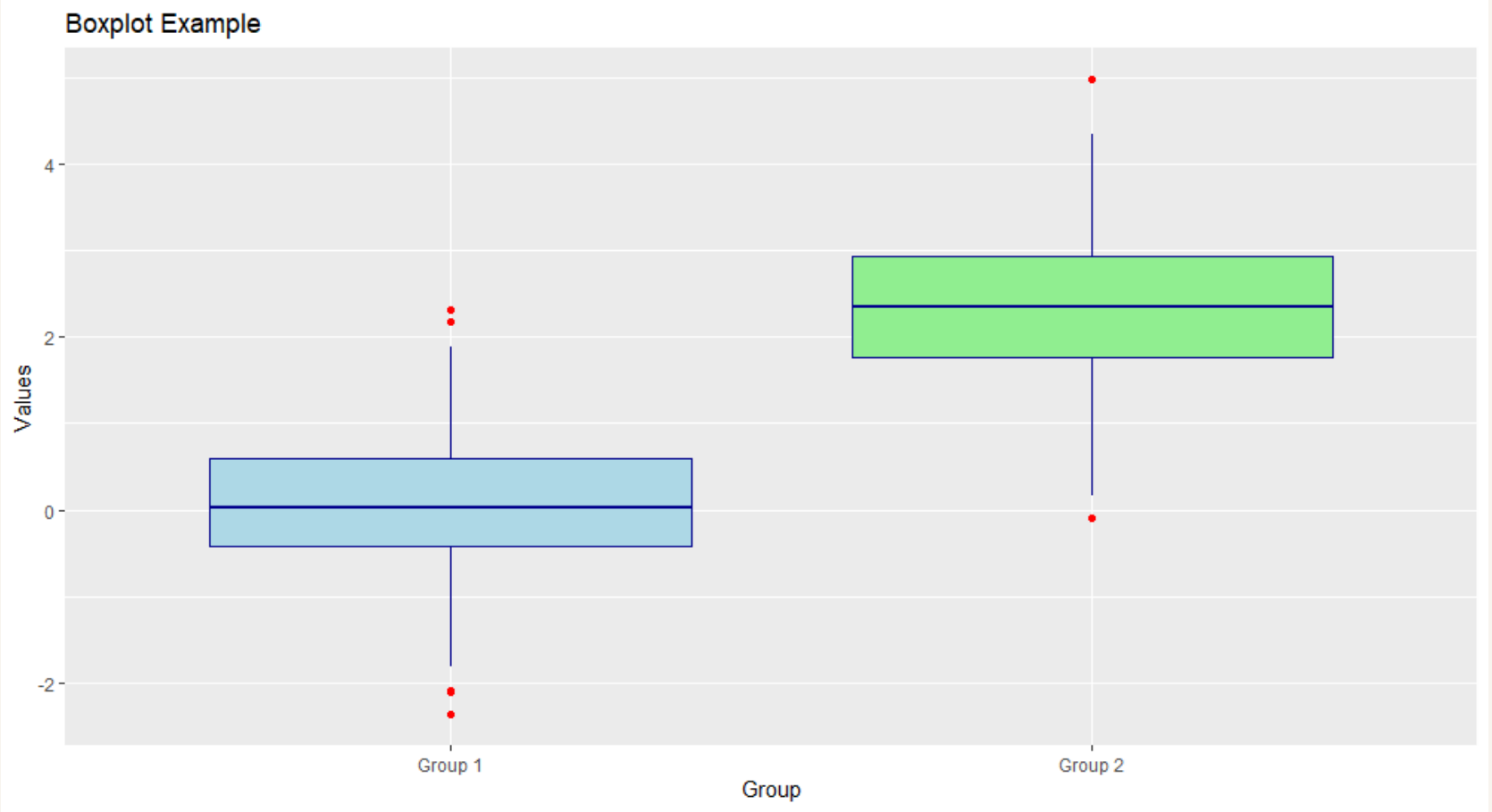


Chapter 3. 기초통계

3.3 Exploratory Data Analysis

데이터를 표현하는 방법

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



Chapter 3. 기초통계

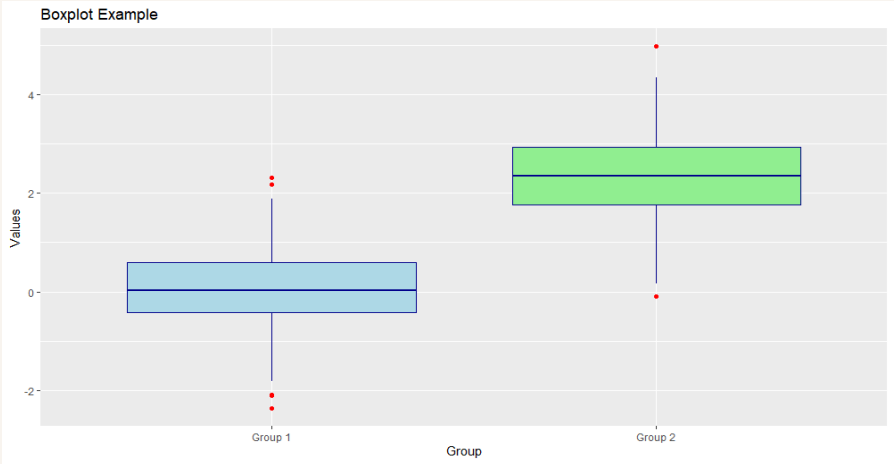
3.3 Exploratory Data Analysis

데이터를 표현하는 방법

- **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")
```

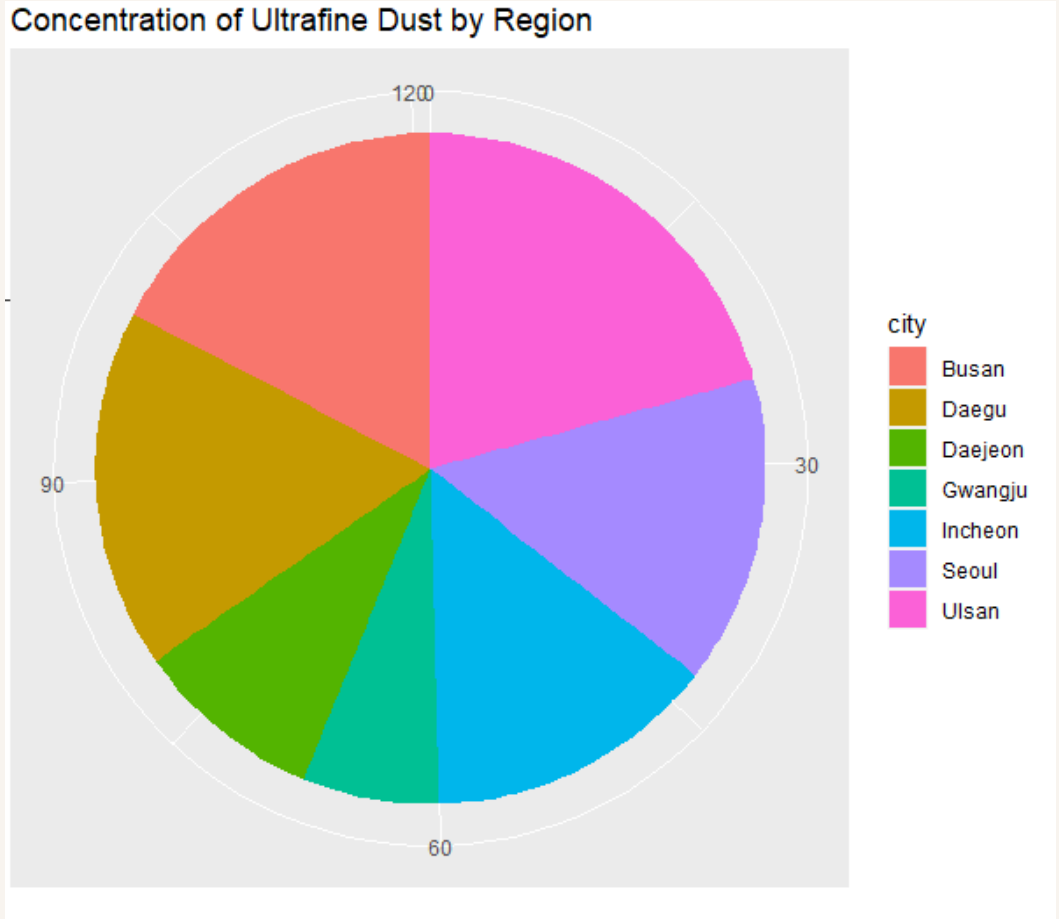


Chapter 3. 기초통계

3.3 Exploratory Data Analysis

데이터를 표현하는 방법

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



Chapter 3. 기초통계

3.3 Exploratory Data Analysis

데이터를 표현하는 방법

- **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("")
```