빅데이터의 이해와 활용

Understanding and Using Big Data

06

빅데이터 시각화

학습
내용

01 빅데이터 시각화의 개념

02 R의 그래프

03 텍스트 시각화

학습 목표

- 빅데이터 시각화의 개념과 필요성을 설명할 수 있다.
- R 프로그램을 이용하여 그래프를 작성하고 분석할 수 있다.
- R 프로그램을 이용하여 워드 클라우드를 작성할 수 있다.

생각
해보기

타이타닉호의 침몰

배에 승선한 2200 명 중에는 현재의 화폐 가치로
55,000달러 이상의 많은 요금을 내고 일등실에 탄 세계의
거부들도 포함되어 있었다.

최저 요금의 하층 객실에는 700명의 이민자들이 타고 있었다.

사망자 중 상당수는 최저 요금 객실에 있던 승객들이었다.
이들에게는 갑판 아래에 그대로 머물러 있으라는 명령이
내려졌다.

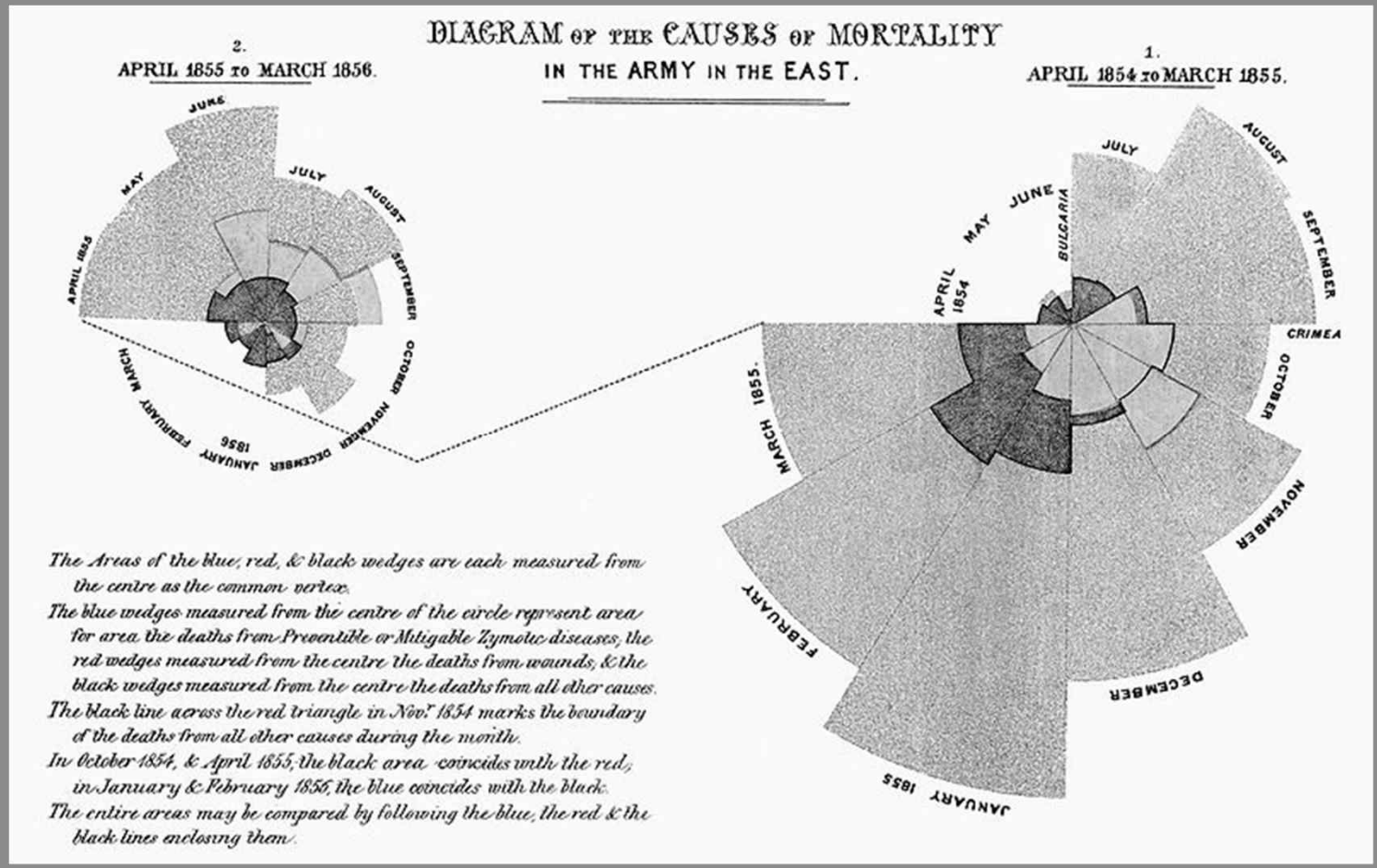
그 결과 배에 탔던 여성 중 1등실 여성 승객은 97 %가
생존했지만 2등실은 84 %, 3등실은 55 %만 생존하였다.

01

빅데이터 시각화의 개념

1) 역사 속의 시각화
2) 시각화의 개요
3) 시각화 과정

나이팅게일(Florence Nightingale) 폴라그래프(1858년)



2) 시각화의 개요

- 데이터 시각화(Data Visualization)란?

데이터 시각화

그 자체가 목적이 아니며 데이터로부터 유용한 정보와 인사이트를 얻어내기 위한 과정

→ 광범위하게 분산된 방대한 양의 자료를 분석해 한 눈에 볼 수 있도록 도표나 차트 등으로 정리하는 것

2) 시각화의 개요

- 데이터 시각화의 효과

01 많은 양의 데이터를 한 눈에 파악할 수 있음

02 인사이트(Insight)를 얻을 수 있음

03 평균적 경향과 특이점(이상치)를 발견할 수 있음

04 의사결정에 활용할 수 있음

2) 시각화의 개요

- 인사이트(Insight, 통찰, 洞察)

“ 예리한 관찰력으로
사물을 환히 꿰뚫어 봄 ”

데이터, 정보, 지식, 사람을 이해하고,
그들 사이의 관계를 파악함

지혜를 도출하는 일련의 과정과 그 결과물로 시각화의 도움을
받아 도출할 수 있음

2) 시각화의 개요

- 시각화 도구

- 스탠포드 대학 교수인 팻 한라한의 R&D 프로젝트에서 처음 탄생함
- 다양한 콘텐츠를 추가할 수 있고 시각적인 분석과 리포팅 도구를 제공하는 응용 프로그램

2) 시각화의 개요

- 시각화 도구

그래프, 차트 등 인터랙티브한 시각화를 쉽게
구현할 수 있는 자바 스크립트 라이브러리

2) 시각화의 개요

- 시각화 도구

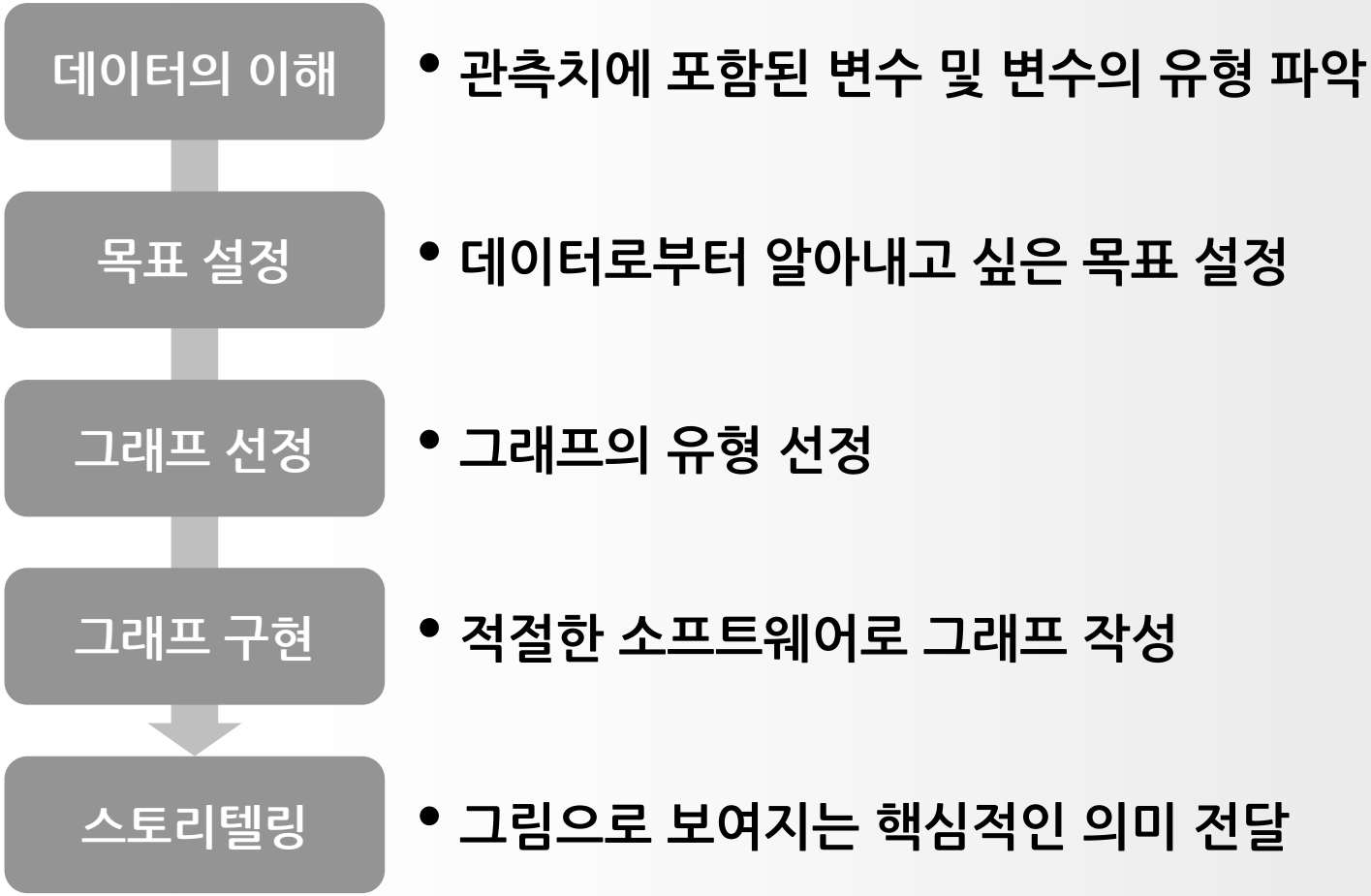
- 통계 분석의 기능 뿐만 아니라 간단한 명령어를 통해 그래프를 그릴 수 있음
- 다양한 패키지를 제공함

2) 시각화의 개요

- 시각화 도구

- 구글에서 제공하는 기능으로 그래프를 작성할 수 있는 대화식 웹 서비스
- 자바 스크립트로 작성함

3) 시각화 과정


데이터의 이해
• 관측치에 포함된 변수 및 변수의 유형 파악
목표 설정
• 데이터로부터 알아내고 싶은 목표 설정
그래프 선정
• 그래프의 유형 선정
그래프 구현
• 적절한 소프트웨어로 그래프 작성
스토리텔링
• 그림으로 보여지는 핵심적인 의미 전달

02

R의 그래프

1) 기본 그래프
2) 응용 그래프
3) 타이타닉 데이터 분석

1) 기본 그래프

● 원 그래프(Pie Chart)

원 그래프

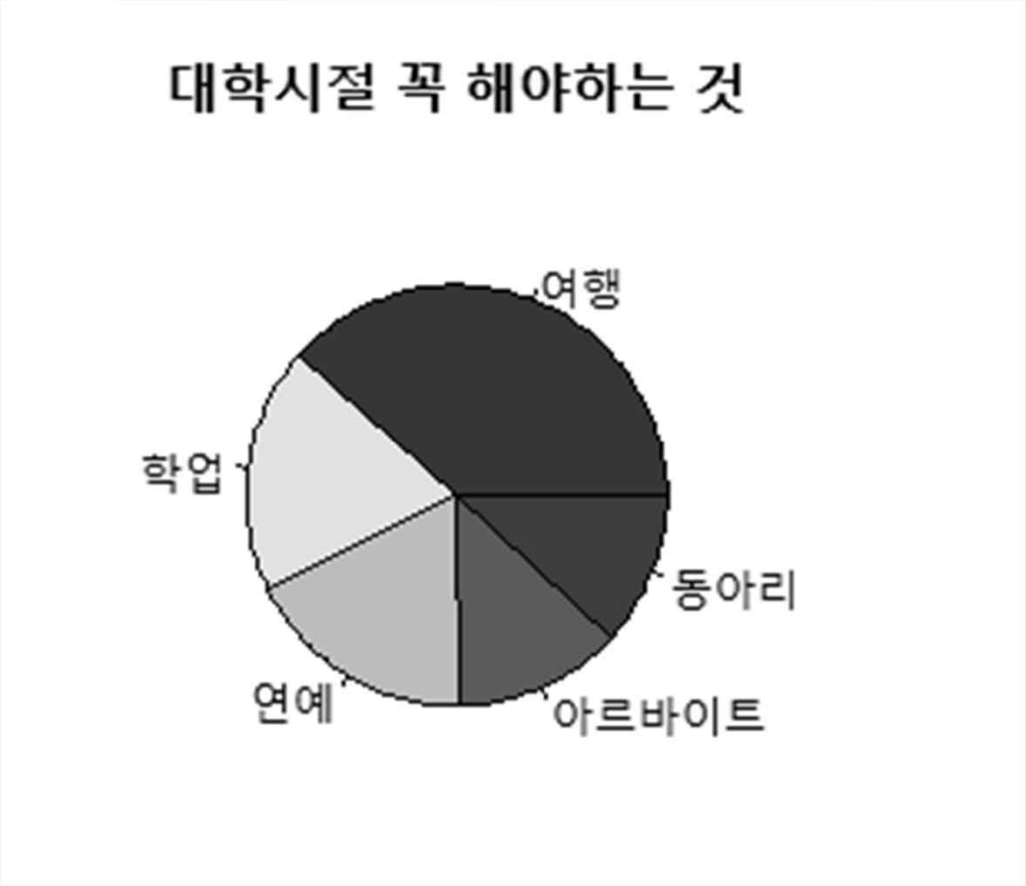
전체에서 각 항목이 차지하는 비율을 표시하는 그래프

- 서로 인접하지 않은 조각을 제대로 비교하기 어려움
- 최대한 구성 요소를 제한하고 내용을 설명하기 위한 텍스트와 비율을 포함하는 것이 좋음

1) 기본 그래프

- 원 그래프(Pie Chart)

구분	비율
여행	62.6
학업	30.8
연애	29.1
아르바이트	20.7
동아리	19.5



1) 기본 그래프

● 원 그래프(Pie Chart)

```
ans <- c(62.6, 30.8, 29.1, 20.7, 19.5)
```

▶ # 비율을 벡터로 작성

```
names(ans) <-
c("여행", "학업", "연애", "아르바이트", "동아리")
```

▶ # 객체의 이름을 설정

```
pie(ans,col=rainbow(5),main=
"대학시절 꼭 해야 하는 것")
```

▶ # 원 그래프 작성

1) 기본 그래프

● 막대 그래프(Bar Chart)

막대 그래프

항목별 도수를 막대의 상대적인 길이로 나타낸 그래프

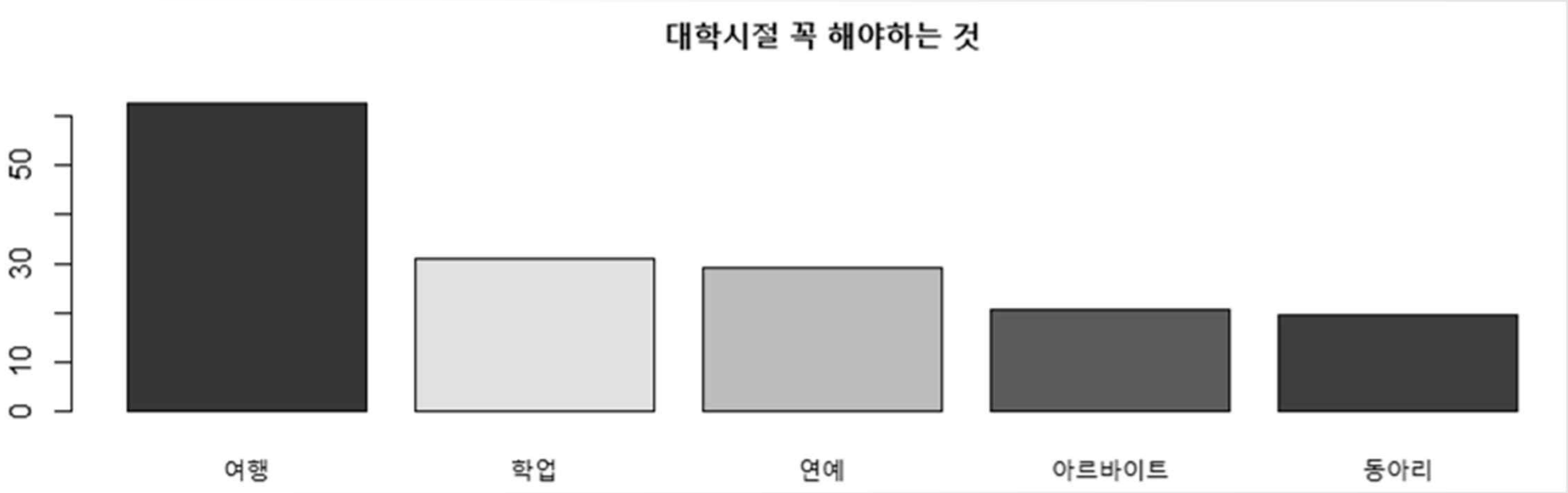
↳ 막대 값들의 차이가 미미하거나 표시할 막대의 수가 많은 경우에는 막대들을 비교하기가 어려움

1) 기본 그래프

● 막대 그래프(Bar Chart)

```
barplot(ans, col=rainbow(5),main
="대학시절 꼭 해야 하는 것")
```

▶ # 막대 그래프 작성



2) 응용 그래프

● 줄기 잎 그림(Stem-and Leaf Plot)

줄기 잎 그림

자료를 줄기와 잎이 달린 나무에 비유해서 나타난 그림

- 관측값을 줄기와 잎으로 구분함
- 자료의 분포형태와 실제 원자료를 알 수 있음

2) 응용 그래프

● 줄기 잎 그림(Stem-and Leaf Plot)

```
data("ChickWeight")
```
▶ # ChickWeight 데이터 셋을 로드함

```
head(ChickWeight)
```
▶ # 상위 6개 데이터 확인

```
attach(ChickWeight)
```
▶ # 변수를 이름만으로 접근

```
stem(weight)
```
▶ # 줄기 잎 그림 작성

```
stem(weight, scale=2)
```
▶ # 줄기를 2배로 작성

▶줄기 잎 그림의 결과

```
> stem(weight) # 줄기잎 그림 작성

  The decimal point is 1 digit(s) to the right of the |

   2 | 599999999
   4 | 000001111111111111111111122222222222222233334566788888888999999999999+38
   6 | 001111111222222223333344444555556666777778888889001111112222222333334+8
   8 | 001122233444444555555667777889999900012233335666667888888 9
  10 | 000011112223333333456666777888990112222344555566778 9
  12 | 000022233333444455556677888901134445555667888 9
  14 | 111234444555566666777888900112344445556667777777 89
  16 | 000022333344444667889900001344455557 89
  18 | 122444445556777822256777778889999
  20 | 0123444555557900245578
  22 | 0012357701123344556788
  24 | 08001699
  26 | 12344569259
  28 | 01780145
  30 | 355798
  32 | 12712
  34 | 1
  36 | 13
```

▶줄기 잎 그림의 결과

```
> stem(weight, scale=2) # 줄기를 2배로 작성

  The decimal point is 1 digit(s) to the right of the |

   3 | 599999999
   4 | 000001111111111111111111122222222222222233334566788888888999999999999
   5 | 00000011111111222233333444555556666777888899999
   6 | 00111111122222222333334444455555666677777888889
   7 | 0011111122222233333444444446667778889999
   8 | 001122233444444555555677778899999
   9 | 000122333356666678888889
  10 | 000011112223333333456666777888899
  11 | 01122223445555667789
  12 | 000022233333444455556778889
  13 | 0113444555566788889
  14 | 1112344445555666667778889
  15 | 001123444455566677777789
  16 | 0000223333444446678899
  17 | 0000134445555789
  18 | 1224444455567778
  19 | 2225677778889999
  20 | 01234445555579
  21 | 00245578
  22 | 00123577
  23 | 01123344556788
  24 | 08
  25 | 001699
  26 | 12344569
  27 | 259
  28 | 0178
  29 | 0145
  30 | 35579
  31 | 8
  32 | 127
  33 | 12
  34 | 1
  35 |
  36 | 1
  37 | 3
```

2) 응용 그래프

● 상자그림(Box Plot)

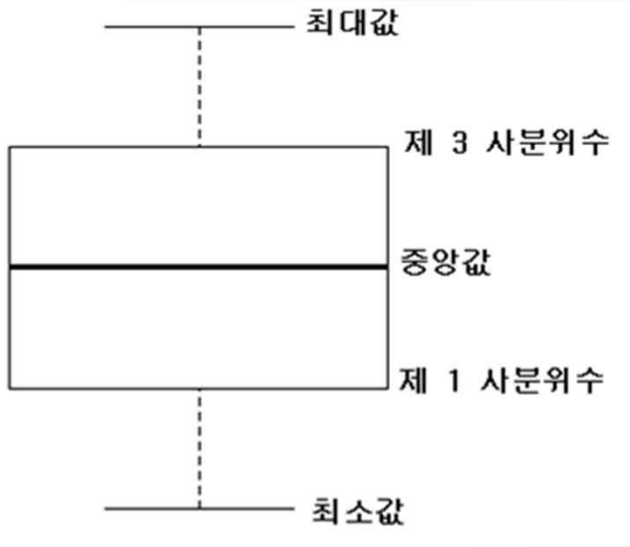
상자그림

요약통계량(최소값, 최대값, 제1사분위수, 제3사분위수, 중앙값)을 상자를 가지고 나타낸 그래프

2) 응용 그래프

- 상자그림(Box Plot)

  - 자료의 50%를 중앙의 상자에 담고 중앙값을 상자에 표시함
  - 특이점(이상치)을 상자 양 끝에 점으로 나타냄



2) 응용 그래프

- 상자그림(Box Plot)

```
boxplot(weight, col="yellow")
```

▶ # 상자 그림 작성

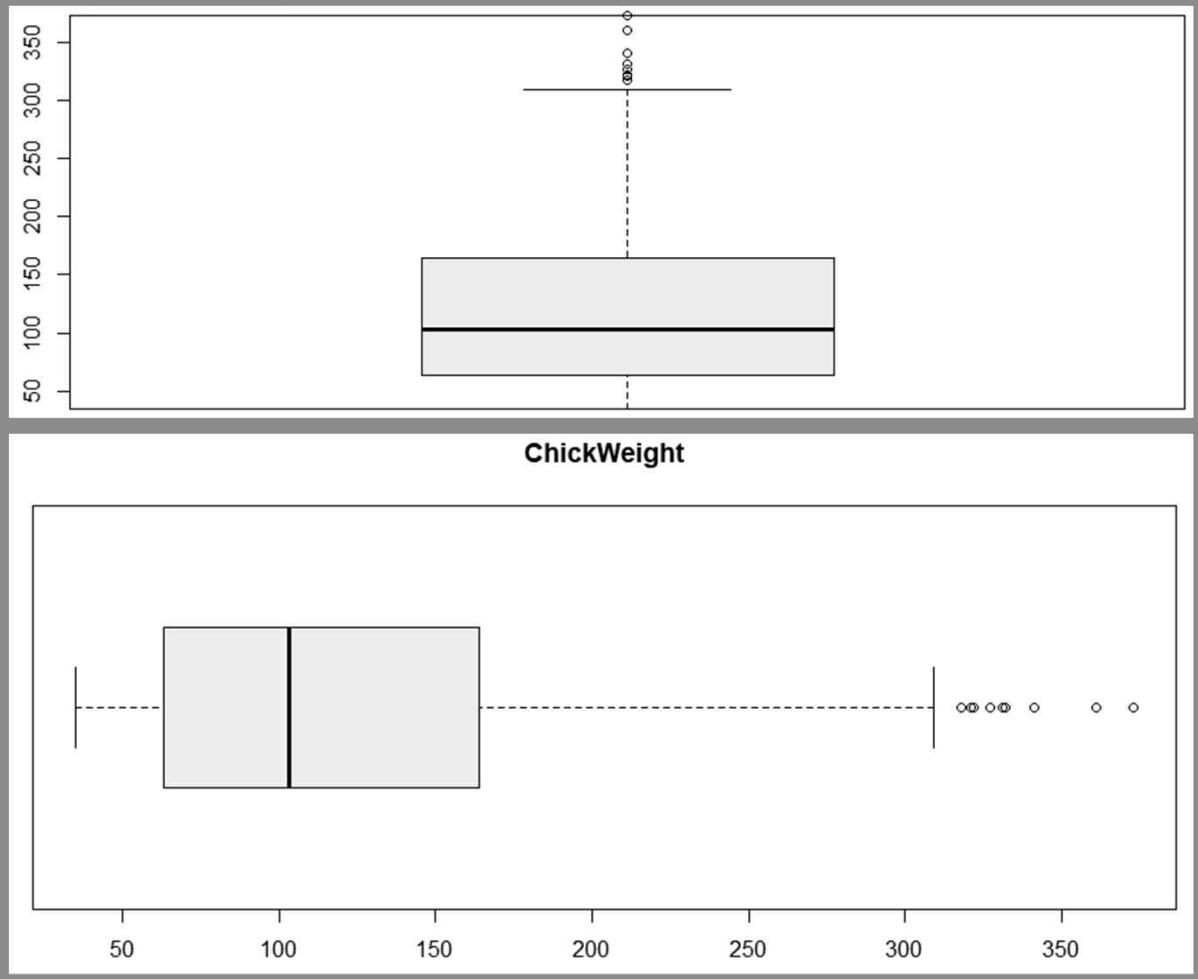
```
boxplot(weight, col="yellow",
horizontal = T, main="ChickWeight")
```

▶ # 상자 그림을 수평으로 작성

```
(boxstats <- boxplot(weight,
col="yellow"))
```

▶ # 상자 그림의 반환값 확인

▶ 상자그림의 결과



3) 타이타닉 데이터 분석

Titanic 데이터

1912년 4월 14일 타이타닉호

2,200여명의 승선자 중
에드워드 스미스(Edward Smith)선장을
포함한 1,500여 명과 함께 침몰함

Pclass	티켓 등급(1등급/2등급/3등급)
Survived	사망 0, 생존 1
Sex	성별
Age	나이
Fare	티켓요금

3) 타이타닉 데이터 분석

- Titanic 데이터

```
titanic <- read.csv("titanic.csv", header=T)
```

```
str(titanic)
```
▶ # 속성과 개수, 미리보기값 제공

```
head(titanic)
```
▶ # 처음 6행을 보여줌

```
tail(titanic)
```
▶ # 마지막 6행을 보여줌

```
summary(titanic)
```
▶ # 요약 통계를 보여줌

```
attach(titanic)
```
▶ # 변수를 이름만으로 접근

▶ Titanic 데이터

```
> summary(titanic) #변수별로 요약통계
     pclass          survived                                      name        sex
 Min.   :1.000   Min.   :0.000   Connolly, Miss. Kate               :   2        :  1
 1st Qu.:2.000   1st Qu.:0.000   Kelly, Mr. James                   :   2  female:466
 Median :3.000   Median :0.000                                     :   1  male  :843
 Mean   :2.295   Mean   :0.382   Abbing, Mr. Anthony               :   1
 3rd Qu.:3.000   3rd Qu.:1.000   Abbott, Master. Eugene Joseph     :   1
 Max.   :3.000   Max.   :1.000   Abbott, Mr. Rossmore Edward       :   1
 NA's   :1       NA's   :1       (Other)                           :1302
      age              sibsp           parch           ticket         fare
 Min.   : 0.1667   Min.   :0.0000   Min.   :0.000   CA. 2343:  11   Min.   :  0.000
 1st Qu.:21.0000   1st Qu.:0.0000   1st Qu.:0.000   1601    :   8   1st Qu.:  7.896
 Median :28.0000   Median :0.0000   Median :0.000   CA 2144 :   8   Median : 14.454
 Mean   :29.8811   Mean   :0.4989   Mean   :0.385   3101295 :   7   Mean   : 33.295
 3rd Qu.:39.0000   3rd Qu.:1.0000   3rd Qu.:0.000   347077  :   7   3rd Qu.: 31.275
 Max.   :80.0000   Max.   :8.0000   Max.   :9.000   347082  :   7   Max.   :512.329
 NA's   :264       NA's   :1        NA's   :1       (Other) :1262   NA's   :2
             cabin      embarked      boat           body                    home.dest
                :1015    :  3          :824   Min.   :  1.0                      :565
 C23 C25 C27    :   6   C:270    13    : 39   1st Qu.: 72.0   New York, NY       : 64
 B57 B59 B63 B66:   5   Q:123    C     : 38   Median :155.0   London             : 14
 G6             :   5   S:914    15    : 37   Mean   :160.8   Montreal, PQ       : 10
 B96 B98        :   4            14    : 33   3rd Qu.:256.0   Cornwall / Akron, OH:  9
 C22 C26        :   4            4     : 31   Max.   :328.0   Paris, France      :  9
 (Other)        : 271            (Other):308   NA's   :1189    (Other)            :639
```

생존한 승객

- 전체의 38.2%

전체 승객 중 남녀 수

- 여성 : 466명
- 남성 : 843명

나이

- 나이의 평균 29.9세
- 전체의 75%가 39세 이하

3) 타이타닉 데이터 분석

- 밀도 곡선

밀도 곡선

데이터의 분포를 살펴보는 그래프

→ 히스토그램과 달리 구간의 너비를 지정할 필요가 없음

3) 타이타닉 데이터 분석

- 밀도 곡선

```
library(ggplot2)
```
▶ # 시각화를 위한 패키지

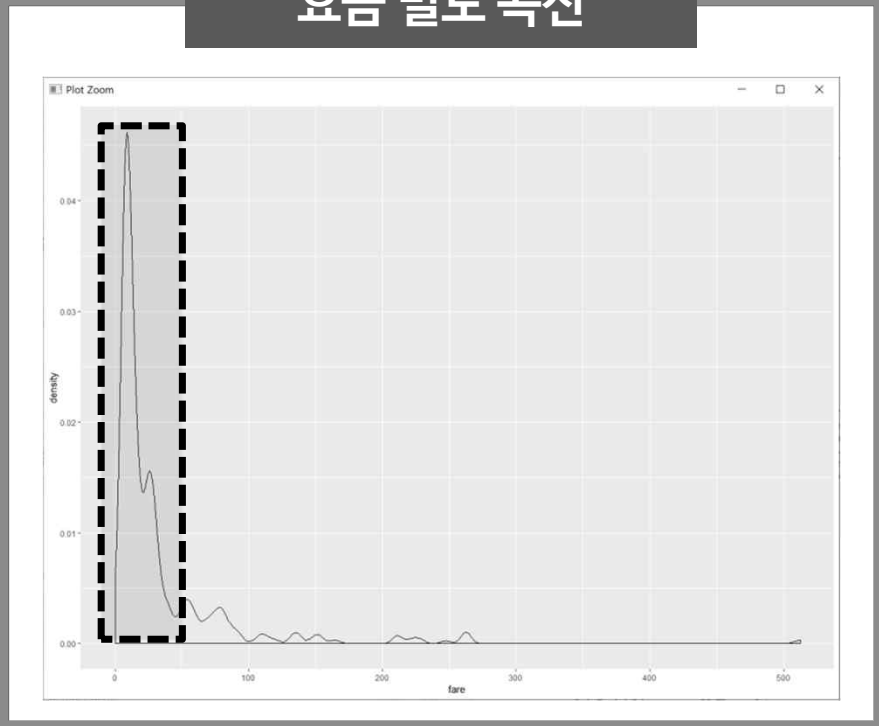
```
library(dplyr)
```
▶ # 데이터 처리 패키지

```
titanic %>% ggplot(aes(fare))+
geom_density( )
```
▶ # 요금의 밀도 곡선

```
titanic %>% ggplot(aes(age,col=sex))
+geom_density( )
```
▶ # 성별 밀도 곡선

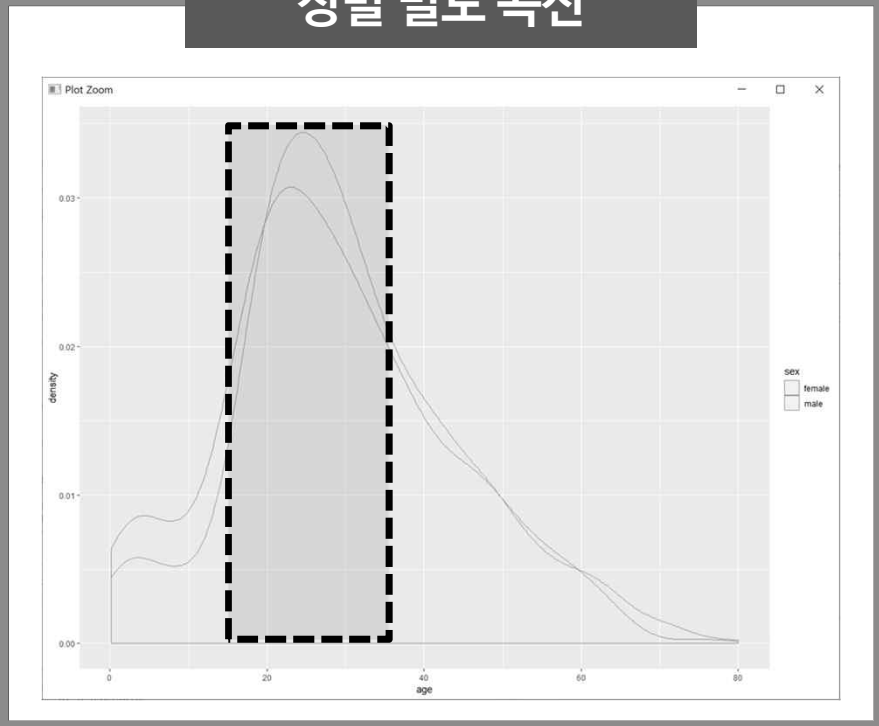
▶ Titanic 데이터 분석 : 밀도 곡선

요금 밀도 곡선



※ 달러 이하의 요금을 지불한 승객이 많음

성별 밀도 곡선



※ 세의 승객이 많으며,
어린 자녀가 함께 탑승한 것으로 추정됨

3) 타이타닉 데이터 분석

- 모자이크 플롯

모자이크 플롯

범주형 데이터를 나타내는 그래프

→ 각 사각형의 넓이가 범주에 속한 데이터의 수에 해당함

3) 타이타닉 데이터 분석

- 모자이크 플롯

```
par(mfrow=c(1,2))
```

▶ # 1행 2열로 분할

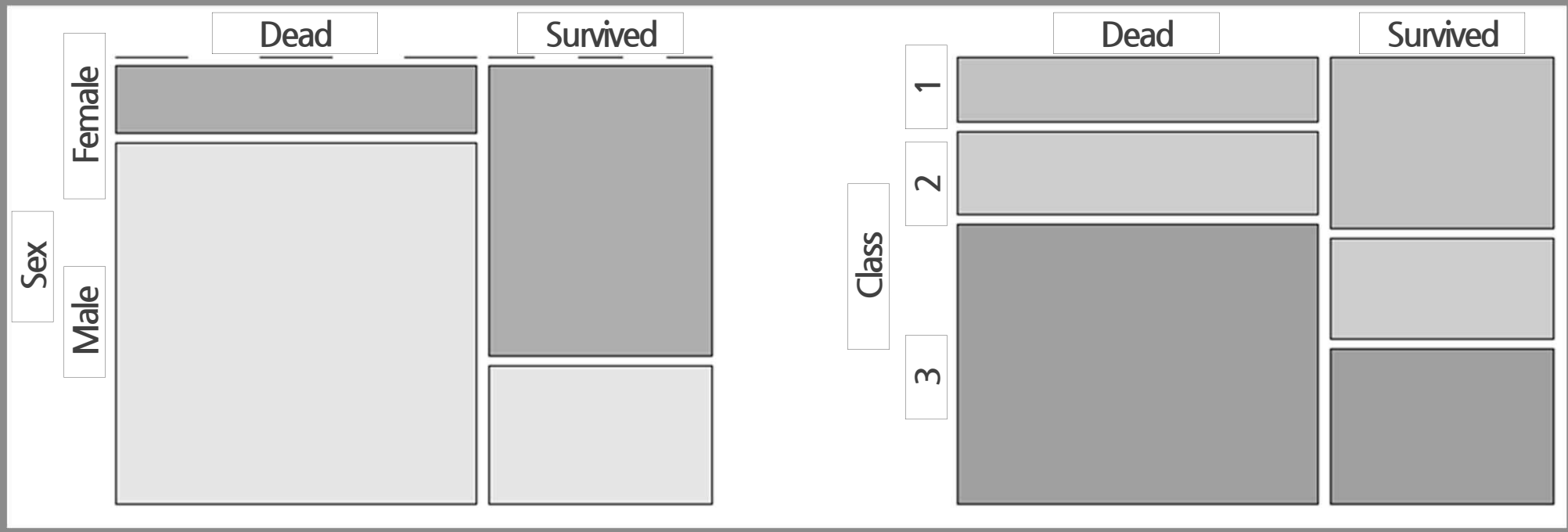
```
mosaicplot(table(ifelse(titanic$survived==1,"survived","Dead"),sex),main="",cex=1.2,color=TRUE)
```

▶ # 성별 생존 모자이크 플롯

```
mosaicplot(table(ifelse(titanic$survived==1,"Survived","Dead"),pclass),main="",cex=1.2,color=c("skyblue","pink","violet"))
```

▶ # 등급별 생존 모자이크 플롯

▶ Titanic 데이터 분석 : 모자이크 플롯



※ 사망자는 남성 > 여성의 순

※ 사망자는 3등급 > 2등급 > 1등급의 순

3) 타이타닉 데이터 분석

● 산점도

산점도

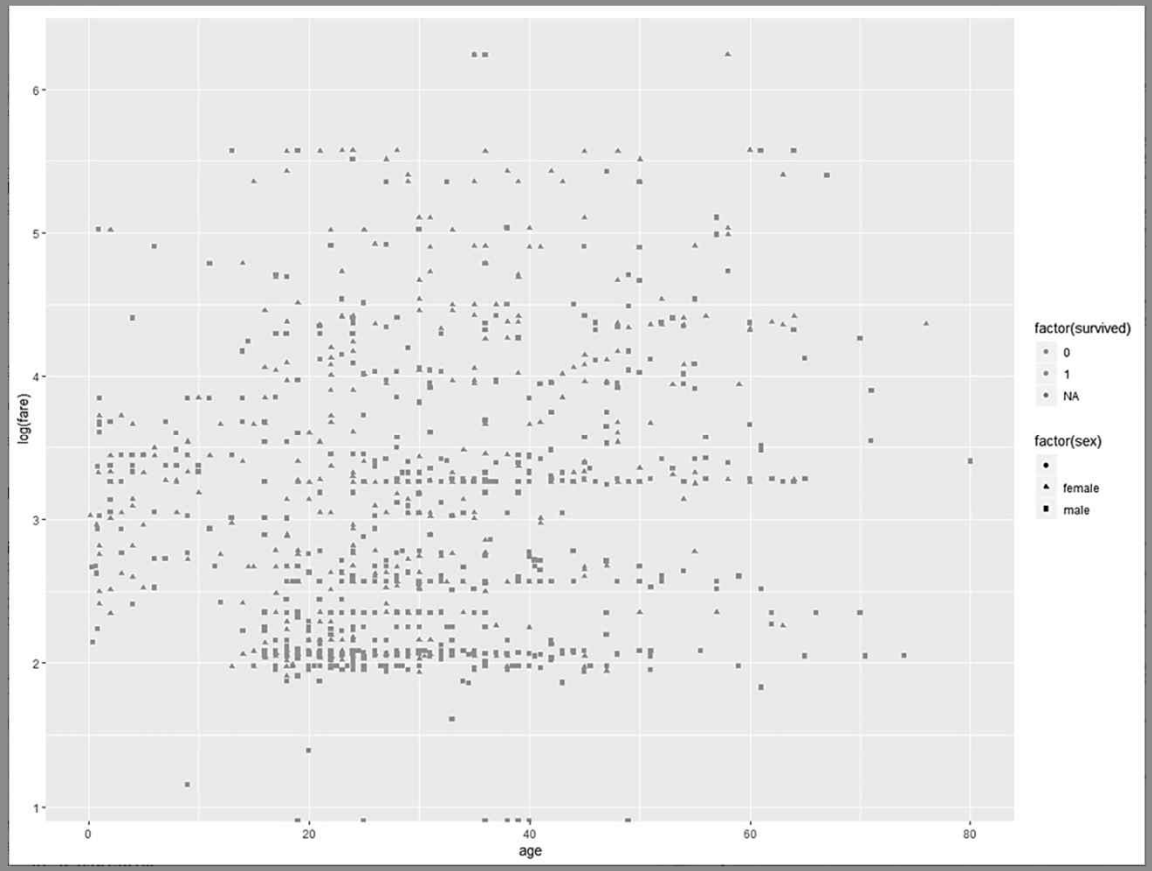
두 변수의 관계를 보여주는 자료 표시 방법

↳ 각 측정값은 두 변수를 의미하는 (x, y)의 점으로 표시함

```
titanic %>%
ggplot(aes(age,log(fare),color=factor(survived),shape
=factor(sex)))+geom_point()+geom_jitter()
```

▶ # 나이, 요금, 생존, 성별을 산점도로 확인

▶ Titanic 데이터 분석 : 산점도



연령	높아질수록 사망자가 많음
요금	높아질수록 생존자가 많음
어린이	등급과 상관없이 생존자가 많음
성별	여성 승객은 남성에 비해 생존자가 많음

03

텍스트 시각화

1) 텍스트 시각화
2) 워드 클라우드 작성

1) 텍스트 시각화

1) 텍스트 시각화

- 빅데이터 연구에서 텍스트 분석의 중요성이 점차 커짐
- 대부분 비정형 데이터이므로 기존의 데이터와는 다른 분석 방법이 요구됨
- 텍스트 데이터 안의 중요 단어를 추출한 후 단어의 출현 빈도, 단어와의 관계 등을 시각화함

1) 텍스트 시각화

- 태그 클라우드(Tag Cloud) 또는 워드 클라우드(Word Cloud)

태그 클라우드 또는 워드 클라우드

텍스트 데이터를 분석하여 단어의 출현 빈도나 중요도 등을 고려하여 중요 단어를 더 클 글자로 화면에 배치하여 시각화하는 방법

〈출처 : http://wordcloud.kr/〉

〈출처 : http://www.tagxedo.com/〉

2) 워드 클라우드 작성

- 패키지 설치

tidytext	텍스트 자료를 tidy data 형태로 정리하여 분석
readr	외부 데이터를 읽어오는 패키지
dplyr	데이터 처리에 특화된 패키지
ggplot2	강력한 시각화 패키지
wordcloud2	향상된 워드클라우드 패키지

2) 워드 클라우드 작성

- 패키지 설치

```
library(tidytext)
```

```
library(readr)
```

```
library(dplyr)
```

```
library(ggplot2)
```

```
library(wordcloud2)
```

2) 워드 클라우드 작성

- 데이터 불러오기

```
text <- read_file("trump.txt")
```

▶ # 파일을 읽음

```
text
```

```
text <- data_frame(text)
```

▶ # 자료를 data frame으로 바꿈

2) 워드 클라우드 작성

- 텍스트를 분해하고 불용어 제거

```
tidy_text <- text %>% unnest_tokens(word,text)
```

▶ # 텍스트를 개별 토큰으로 분해

```
tidy_text <- tidy_text %>% anti_join(stop_words)
```

▶ # 불용어 제거

2) 워드 클라우드 작성

- 빈도 그래프 작성

```
tidy_text %>% count(word, sort=T) %>%
```
▶ # 빈도수 계산

```
filter(n>1) %>%
```
▶ # 2번 이상 나온 단어를 대상으로

```
mutate(word=reorder(word,n)) %>%
```
▶ # 단어 빈도순으로 나열

```
ggplot(aes(word,n)) + geom_col() +
coord_flip()
```
▶ # 빈도를 나타낸 그래프를 그림

2) 워드 클라우드 작성

- 워드 클라우드 작성

```
tidy_text %>%
count(word,sort=T)%>%
filter(n>2) %>%
wordcloud2()
```

▶ # 워드 클라우드 작성

학습 평가

Q1 Q2

Q1

다음 중 상자 그림에서 제공하는 정보가 아닌 것은?

1 최소값
2 중앙값
3 이상치
4 평균

학습 평가

Q1 Q2

Q1

다음 중 상자 그림에서 제공하는 정보가 아닌 것은?

1 최소값

2 중앙값

3 이상치

4 ✓ 평균

정답 4 평균

해설 상자 그림에서는 최소값, 최대값, 제1사분위수, 제3사분위수, 중앙값을 표시합니다.

학습 평가

Q1 Q2

Q2

텍스트 분석에서 사용하지 않을 불용어를 제거하는 함수로 옳은 것은?

tidy_text <- tidy_text %>% (stop_words)

1 Mutate

2 anti_join

3 unnest_tokens

4 filter

학습 평가

Q1 Q2

Q2

텍스트 분석에서 사용하지 않을 불용어를 제거하는 함수로 옳은 것은?

```
tidy_text <- tidy_text %>% anti_join (stop_words)
```

1 Mutate
2 anti_join
3 unnest_tokens
4 filter

정답 2 anti_join

해설 tidytext에서는 anti_join( )으로 불용어를 제거합니다.

정리
하기

빅데이터 시각화의 개념

- ✓ 데이터 시각화
  - 광범위하게 분산된 방대한 양의 자료를 분석해 한 눈에 볼 수 있도록 도표나 차트 등으로 정리하는 것
  - 시각화는 인사이트를 부여할 수 있는 수단이며 빠른 의사 결정을 내릴 수 있게 해줌
  - 인사이트(Insight, 통찰, 洞察)
    : 데이터, 정보, 지식, 사람을 이해하고, 그들 사이의 관계를 파악해 지혜를 도출하는 일련의 과정과 그 결과물로 시각화의 도움을 받아 도출할 수 있음

정리
하기

R의 그래프

- ✓ R의 그래프
  - 원 그래프(Pie Chart)
    - : 전체에서 각 항목이 차지하는 비율을 표시하는 그래프
  - 막대 그래프(Bar Chart)
    - : 항목별 도수를 막대의 상대적인 길이로 나타낸 그래프
  - 줄기 잎 그림(Stem-and Leaf Plot)
    - : 자료를 줄기와 잎이 달린 나무에 비유해서 나타난 그림

정리
하기

R의 그래프

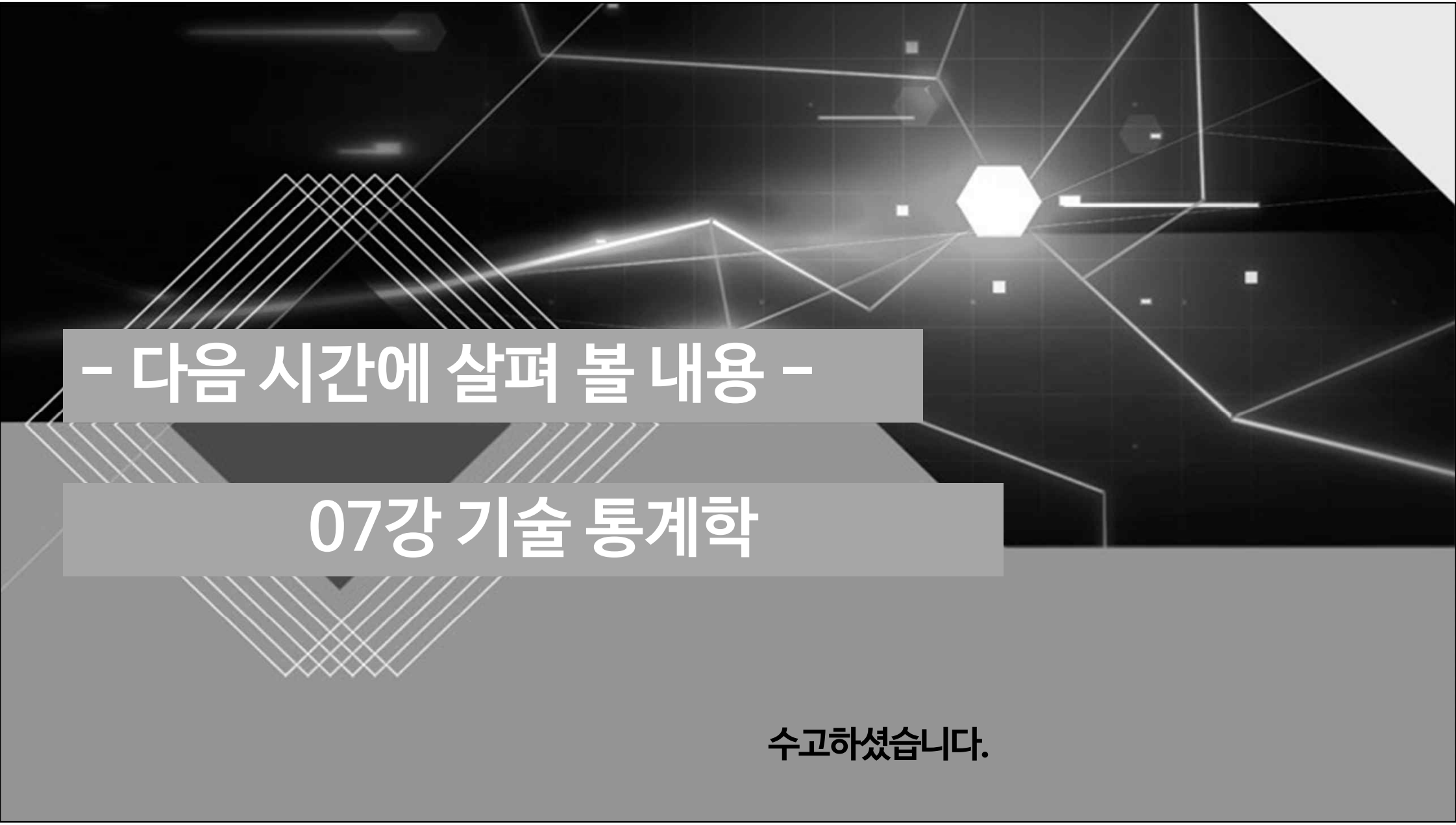
- ✓ R의 그래프
  - 상자그림(Box Plot)
    - : 요약통계량(최소값, 최대값, 제1사분위수, 제3사분위수, 중앙값)을 상자를 가지고 나타낸 그래프

정리하기

텍스트 시각화

- ✓ 텍스트 시각화
  - 대부분 비정형 데이터이므로 기존의 데이터와는 다른 분석 방법이 요구됨
  - 텍스트 데이터 안의 중요 단어를 추출한 후 단어의 출현 빈도, 단어와의 관계 등을 시각화

- ✓ 태그 클라우드(Tag Cloud) 또는 워드 클라우드(Word Cloud)
  - 텍스트 데이터를 분석하여 단어의 출현 빈도나 중요도 등을 고려하여 중요 단어를 더 클 글자로 화면에 배치하여 시각화하는 방법


- 다음 시간에 살펴 볼 내용 -
07강 기술 통계학
수고하셨습니다.