Chapter 7

데이터 정제하기(결측치 제거)

결측치란?

결측치(Missing Value)

- 누락된 값, 비어있는 값을 의미합니다
- 함수 적용 불가, 분석 결과 왜곡시키게 됩니다.
- 반드시 제거 후 분석 실시해야 합니다.

원자료

id	class	english	science
1	1	98	50
2	1	97	60
3	1	86	78
4	1	98	58
5		80	65
6	2	89	
7	2	90	45
8	2		99999
9	3	98	15
10	3	98	45
11	3	99999	65
12	3	85	32

정제하기

id	class	english	science
1	1	98	50
2	1	97	60
3	1	86	78
4	1	98	58
7	2	90	45
9	3	98	15
10	3	98	45
12	3	85	32

소제목

결측치 확인하기

```
data <-  data.frame(id = c(1:5),
         gender = c("M", "F", NA, "M", "M"),
         score = c(5,4,3,2,NA))

is.na(data) #na여부 확인

table(is.na(data) ) #전체 결측치 빈도 확인

table(is.na(data$gender)) #컬럼별 결측치 빈도 확인

##
  id gender score
1  1      M     5
2  2      F     4
3  3   <NA>     3
4  4      M     2
5  5      M    NA
```

소제목

결측치 제거하기

```
library(dplyr) # dplyr 패키지 로드

data %>% filter( !is.na(score) ) # 결측치 제거 filter()

data %>% filter( !is.na(score) & !is.na(gender) ) # 여러 결측치 제거


## 제거 후의 값
 id gender score
1  1      M     5
2  2      F     4
3  4      M     2
```

소제목

결측치가 하나라도 있다면 제거

- **분석에 필요한 데이터까지 손실 될 가능성 유의**
- **ex) 성별-소득 관계 분석하는데 지역 결측치까지 제거할 수 있습니다**

```
library(dplyr) # dplyr 패키지 로드

df_nomiss2 <- na.omit(df) # 모든 열에 결측치 없는 데이터 추출


## 제거 후의 값
 id gender score
1  1      M     5
2  2      F     4
3  4      M     2
```

소제목

함수에서 결측치 제거하고 사용하기 na.rm = T

```
library(dplyr) # dplyr 패키지 로드

mean(df$score, na.rm = T) # 결측치 제외하고 평균 산출

sum(df$score, na.rm = T) # 결측치 제외하고 합계 산출


# summarise에서 사용해보기
exam <- read.csv("data/excel_exam.csv")
exam
exam[ c(1, 3, 5, 6), "math" ] = NA # 1,3,5,6행 결측치 삽입

exam %>% summarise( score_mean = mean(math, na.rm = T),
                    score_sum = sum(math, na.rm = T) )
```

소제목

결측치 대체하기

- 결측치 많을 경우 모두 제외하면 데이터 손실 큼
- 대안: 다른 값 채워넣기

결측치 대체법(Imputation)

- 대표값(평균, 최빈값 등)으로 일괄 대체
- 통계분석 기법 적용, 예측값 추정해서 대체

소제목

평균값으로 결측치 대체하기

```
library(dplyr) # dplyr 패키지 로드

data = read.csv("data/excel_exam.csv")
data

data[ c(1,5,7,20), "math"] = NA #결측치 생성

#점수중 평균으로 대체하기
mean(data$math, na.rm = T) # 결측치 제외하고 평균 산출 (평균 57)
data$math <- ifelse(is.na(data$math), 57, data$math )

table(is.na(exam$math)) # 결측치 다시 확인
```

문제

ggplot2에 존재하는 mpg데이터를 사용합니다
mpg데이터를 다음 구문으로 불러와서 결측치를 생성하세요.

```
mpg <- as.data.frame(mpg)
mpg[c(65, 124, 131, 153, 212), "hwy"] <- NA
```

Q1
drv(구동방식) 별 hwy(고속도로연비) 평균이 어떻게 다른지 확인하려고 합니다.
결측치를 확인하고 drv, hwy 변수에 결측치가 몇 개 있는지 확인하세요.
Q2
filter() 를 이용해서 결측치를 제거하고 어떤 구동 방식 평균이 높은지 그룹별로 확인하고 차순정렬하세요

Chapter 7

데이터 정제하기(이상치 제거)

데이터 정제하기

이상치(Outlier) - 정상범주에서 크게 벗어난 값

- 이상치 포함시 분석 결과를 왜곡시키게 됩니다.
- 결측 처리 후 제외하고 분석해야 합니다.

이상치 종류	예	해결 방법
존재할 수 없는 값	성별 변수에 3	결측 처리
극단적인 값	몸무게 변수에 200	정상범위 기준 정해서 결측 처리

데이터 정제하기

이상치 제거하기

```
library(dplyr) # dplyr 패키지 로드

outlier <- data.frame(gender = c(1, 2, 1, 3, 2, 1),
                      score = c(5, 4, 3, 4, 2, 6))
outlier

table(outlier$gender) #이상치 확인

outlier$gender <- ifelse(outlier$gender == 3, NA, outlier$gender) #이상치를 NA로 변경

outlier %>%
  filter(!is.na(gender) & !is.na(score)) %>%
  group_by(gender) %>%
  summarise(mean_score = mean(score))
```

데이터 정제하기

이상치 제거하기2

- 정상범위 기준 정해서 벗어나면 결측 처리

판단 기준	예
논리적 판단	성인 몸무게 40kg~150kg 벗어나면 극단치
통계적 판단	상하위 0.3% 극단치 또는 상자그림 1.5 IQR 벗어나면 극단치

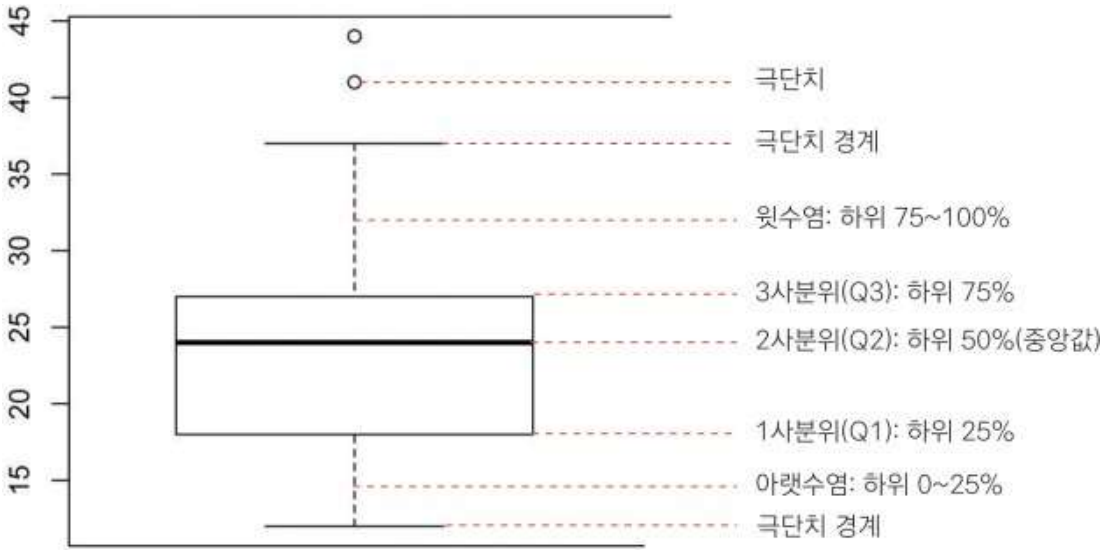
데이터 정제하기

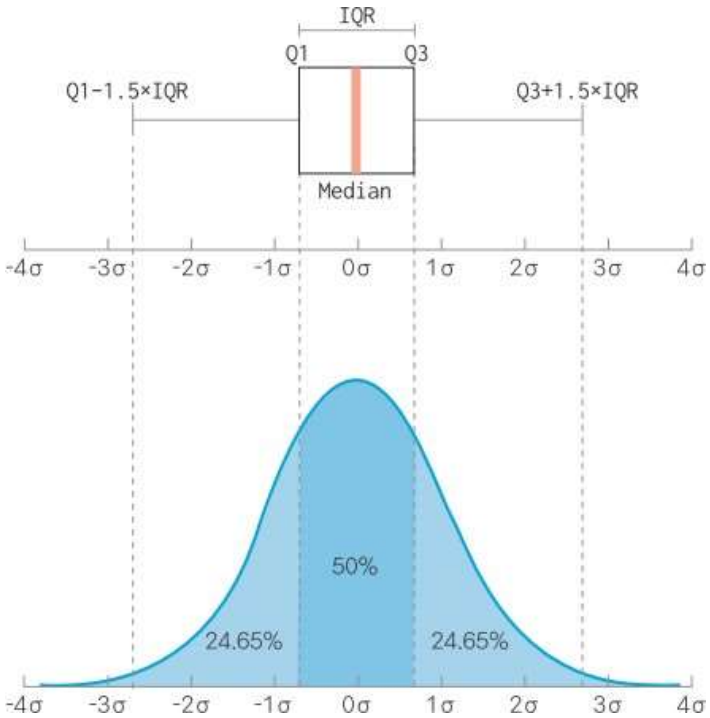
이상치 판별해보기

```
library(dplyr)
library(ggplot2)

mpg <- as.data.frame(mpg)
boxplot(mpg$hwy)
boxplot(mpg$hwy)$stats # 상자그림 통계치 출력
```

```
##      [,1]
## [1,]   12 이상치
## [2,]   18
## [3,]   24 평균
## [4,]   27
## [5,]   37 이상치
## attr(,"class")
```





데이터 정제하기

이상치 결측처리 후 분석

```
# 앞서 12와 37이 이상치 임을 확인 합니다

mpg$hwy <- ifelse(mpg$hwy < 12 | mpg$hwy > 37, NA, mpg$hwy) # 12 or 37을 넘으면 제거

table(is.na(mpg$hwy)) #확인

mpg %>%
  group_by(drv) %>%
  summarise(mean_hwy = mean(hwy, na.rm = T))
```

문제

ggplot2에 존재하는 mpg데이터를 사용합니다
mpg데이터를 다음 구문으로 불러와서 이상치를 생성하세요.

```
mpg <- as.data.frame(ggplot2::mpg)
mpg[ c(10, 14, 58, 93), "drv" ] <- "k"
mpg[ c(29, 43, 129, 203), "cty"] <- c(3,4,39,42)
```

drv의 k값과 cty에 극단적으로 낮은수가 이상치입니다

Q1
.drv에 이상치가 있는지 확인 합니다. 이상치를 결측치로 처리한 다음 확인하세요.

Q2
boxplot을 이용해서 cty의 이상치 범위를 확인하고 통계치를 이용해서 벗어난 값을 결측처리 한 후
다시 boxplot을 만들어서 확인하세요.

Q3
drv와 cty의 이상치를 결측처리 했다면, 결측치를 제외한 다음 drv별 cty평균이 어떻게 다른지 확인하세요.
파이프라인을 사용합니다. (그룹핑)

Chapter 7

수고하셨습니다