11 데이터 정제

학습목표

1. 데이터에서 결측값과 이상값을 찾아낼 수 있다.
2. 데이터 정제를 위해 결측값과 이상값을 처리할 수 있다.

강의내용

데이터 탐색

- 데이터의 구성 형태 전체적으로 살펴보기
  . 관측값과 변수의 개수, 변수의 자료형 확인
  . 데이터의 앞이나 뒤에서 확인하기
  . view로 데이터 살펴보기
- 데이터 특성을 파악하기 위해 각 속성값 살펴보기
  . 요약통계 및 빈도표
  . 시각자료(히스토그램, 상자그림)를 통해 값의 빈도 및 분포 확인

데이터 정제

- 데이터분석 방법이 아무리 뛰어나도 데이터분석에 사용되는 데이터의 품질, 즉 데이터 정제가 제대로 되어있지 않다면 분석 결과가 왜곡되거나 다른 결론이 나올 수밖에 없다.
- 외부에서 다른 형식의 데이터를 R로 읽어들일 때도 입력되지 않은 값이나 특정 문자를 R에서 인식하는 결측값 NA로 변경하여 읽어오도록 옵션을 지정하여야 한다.
  . na.strings= c(" ", "", ,"99999", NA)
- 데이터 탐색 등을 통해 확인된 결측값[1] 또는 이상값[2] 처리, 자료형 변환, 날짜 처리 등
  . 결측값 확인 : 요약통계 summary()나 is.na() 함수 등으로 확인
  . 이상값 확인 : 요약통계 summary(), 빈도표, 빈도그래프, 상자그림 등으로 확인

```
#  외부데이터 읽어올 때 특정 문자를 R의 결측값으로 인식
dfr <- read.csv("na_exam.csv", na.strings= c(" ", "", NA))

#  결측값 확인
mean(df$math)           # 결측값이 있는 경우 정상적인 연산 안됨
summary(df)             # 요약통계량을 통해 결측값 확인
is.na(df)               # TRUE로 표시된 값이 결측값
table(is.na(df))        # 결측값의 갯수 확인 (어떤 변수에 결측값이 있는지 알 수 없음)
table(is.na(df$math))
colSums(is.na(df))      # 열별 합계, 합계가 0보다 큰 열 : 결측치 존재

#  이상값 확인
summary(df)             # 요약통계량을 통해 이상값 확인
table(df$gender)        # 빈도표를 통해 값이 범위 확인
hist(df$math)           # 히스토그램을 통해 값의 범위(분포) 확인
boxplot(df$math)        # 상자그림을 통해 값의 범위(분포) 확인
qplot(df$class)         # 히스토그램을 통해 값의 범위(분포) 확인,  ggplot2패키지 로드
```

1) 결측값 : 누락된 값 또는 비어있는 값으로 NA로 표시
2) 이상값 : 논리적, 통계적으로 나올 수 없는 값이 입력된 경우로 분석에 심각한 문제를 일으키는 값

11 데이터 정제

- 결측값 처리 방법

  (1) 결측값이 포함된 행 제거 : filter()나 원소접근법을 이용하여 결측값이 없는 행만 추출

  (2) 특정 변수들에 결측값이 포함된 행 제거 : filter()와 논리연산자를 변수들에 모두 결축값이 없는 행 추출

  (3) na.omit() 이용하여 결측값이 하나라도 있으면 모두 제거 (※ 분석에 사용가능한 데이터까지 제거될 수 있음)

  (4) 다른 값으로 대체 : 평균값, 중간값, 계산값 등으로 결측값을 대체

  . 결측값 제거시 데이터 손실로 분석왜곡 발생할 수 있으므로 다른 값으로 대체하여 왜곡 문제 보완

  (※) 함수 사용 시 결측값을 제외하기

```
# (1) 결측값 제거 :  filter(), %>% 연산자 사용하여 결측값이 없는 행만 추출
library(dplyr)
df %>% filter(!is.na(math)) -> df_nomiss1

# (2) 결측값 제거 : filter()와 논리 연산자(&)  사용하여 여러 변수에 모두 결측값 없는 행 추출
df %>% filter( !is.na(math) & !is.na(class) ) -> df_nomiss2

# (3) 결측값 제거 : na.omit()  사용하여 결측값이 있는 모든 행 제거
df_nomiss3 <- na.omit(df)

# (*) 결측값 제거 : na.rm=TRUE 옵션 사용하여 함수에서 결측값 제외하고 연산
mean(df$math, na.rm=T)
```

```
# 결측값 대체 : 평균값으로 대체
mean_math <- mean(df$math, na.rm=T)   # df$math에서 NA 제거한 평균값 계산

# df$math가 NA인 경우만 평균값으로 대체
ifelse(is.na(df$math), mean_math, df$math) -> df$math
```

- 이상값 처리 방법

  (1) 이상값 제거 (이상값이 없는 행만 추출)

  (2) 결측값으로 변경

```
# 이상값이 없는 행만 추출
df <- df %>% filter( gender=='M' | gender=='F')    # gender변수가 M이나 F인 행만 추출

# gender 변수가 M도 아니고 F도 아닌 경우 NA 값으로 대체
df$gender <- ifelse( ((df$gender=="M")|(df$gender=="F")), df$gender, NA)
```

- 자료형 변환 ; as.factor() as.numeric() as.character() as.integer() as.logical()
- 변수이름 정리 : dplyr 패키지의 rename() 함수를 이용하여 변수의 이름 변경
- 문자열 조작 : stringr 패키지의 함수 (동영상 자료 참조)