데이터 베이스와 R

박찬엽

2017년 6월 27일

목차

과제 확인

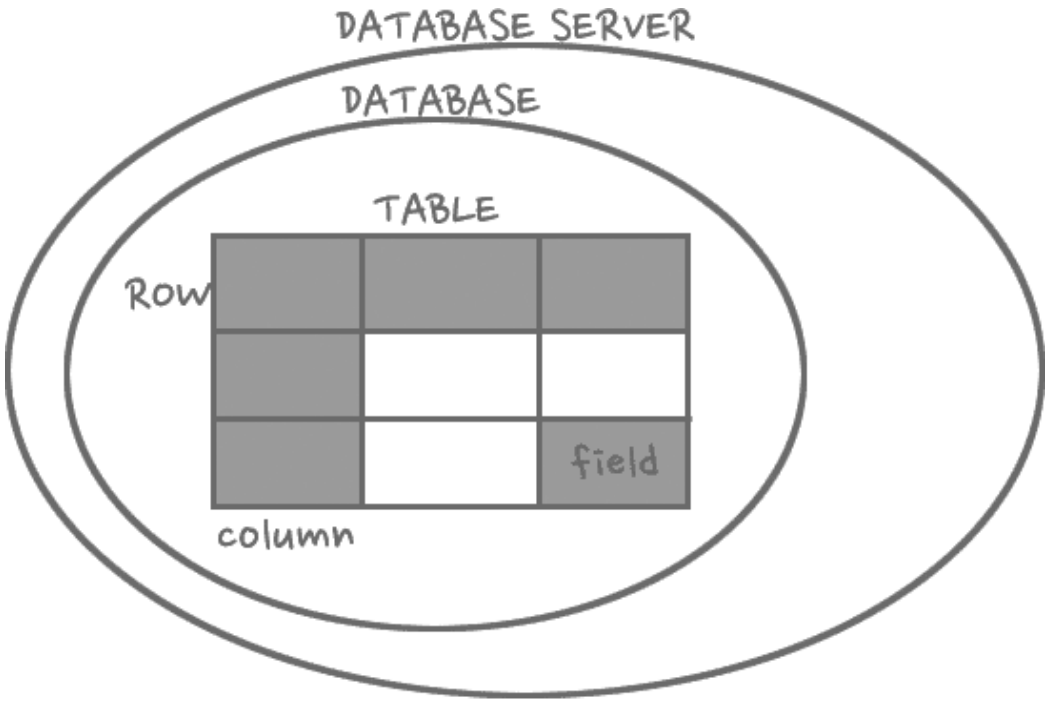
데이터베이스

데이터란

단순한 관찰이나 측정 등의 수단을 통해 현실 세계로부터 수집된 사실이나 값

의미있게 사용하기 위해서 구조화가 필요함
* 구조화: 체계적으로 조직하는 것

DBMS



* 이미지 출처: 생활코딩 MySQL 수업

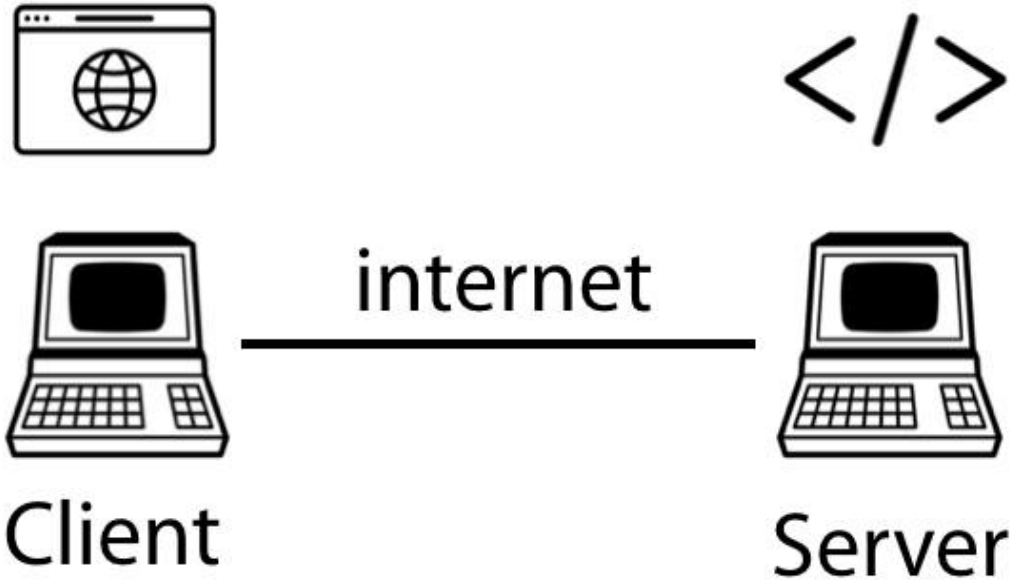
데이터베이스란

엑셀	DBMS
파일	데이터베이스
시트	테이블

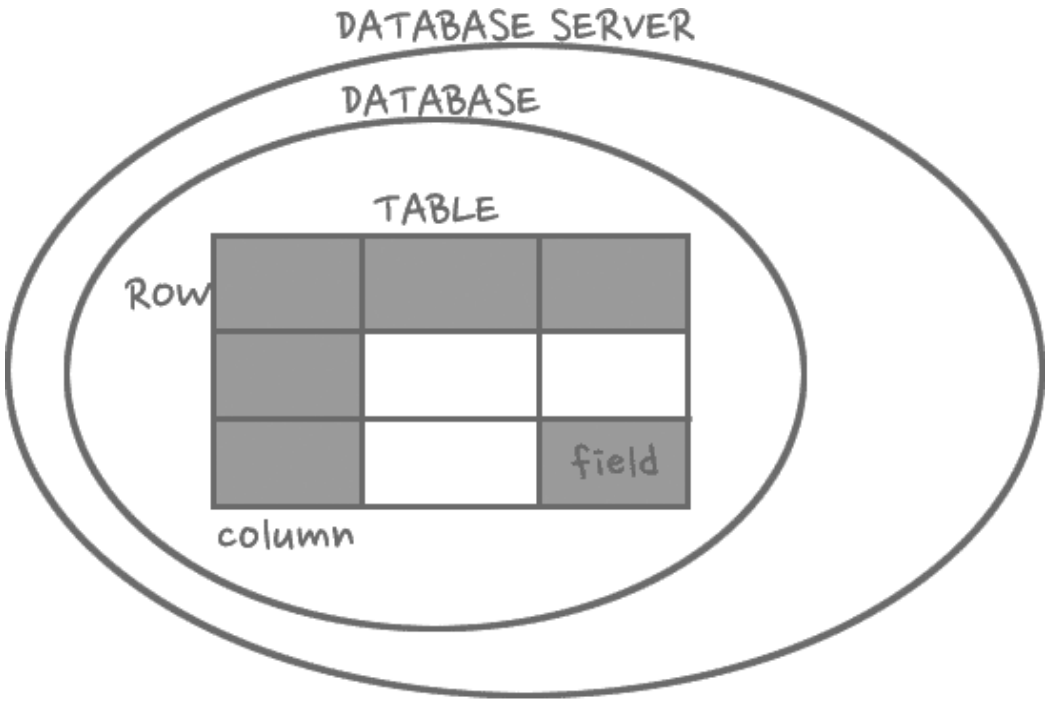
데이터베이스 클라이언트

- 대표적인 클라이언트
  - MySQL monitor
  - PHPmyAdmin
  - Navicat
  - HeidiSQL

서버와 클라이언트



테이블



* 이미지 출처: 생활코딩 MySQL 수업

data.frame

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

SQL

Structured Query Language

구조적 데이터 요청 언어

DBI

```
## Loading required package: devtools
```

```
## Loading required package: DBI
```

```
## Loading required package: RSQLite
```

SQLite

SQLite

SQLite is a self-contained, high-reliability, embedded, full-featured, public-domain, SQL database engine.

파일 하나로 구성하는 작고, 무료인 sql db

DBI

```
library(DBI)
library(RSQLite)
con <- dbConnect(RSQLite::SQLite(), dbname="class2.sqlite")

dbListTables(con)

dbWriteTable(con, "mtcars", mtcars, overwrite=T)
dbListTables(con)

dbReadTable(con, "mtcars")

dbRemoveTable(con,"mtcars")
dbListTables(con)

system.time(dbWriteTable(con, "member", "./recomen/membership.csv",row.names=F))
```

데이터

데이터 공유

Leek group에서 소개하는 데이터 공유 가이드

- 원시 데이터
- 정제후 데이터
- 코드북
- 변수 작성법
- 재현성

원시 데이터

최초 획득한 당시 그대로의 데이터

- 어떤 식으로든 수정을 가하지 않은 상태
- 수정을 가하는 과정을 함께 기록함으로써 신뢰성 확보
- 위 두 가지가 없는 경우 상황을 상상해야 함

정제후 데이터

해들리 위컴이 설명한 tidy data의 요건에 맞게 가공하여 데이터를 쉽게 다룰 수 있게 만든 상태

- 측정하는 각 변수는 하나의 열에 있어야 합니다.
- 측정하는 각 관찰은 하나의 행에 있어야 합니다.
- 각 종류의 변수에 대해 각 하나의 테이블이 있어야합니다.
- 여러 개의 테이블이있는 경우 테이블에 합치기 위한 기준 열을 포함해야합니다

코드북

데이터셋에 대해 필요한 설명을 담은 문서

- 정제후 데이터에 대해 추가적으로 필요한 설명이나 정보(단위 등)
- 정제 과정에서 사용한 방법의 설명과 사용한 이유
- 데이터가 사용된 분석에 대한 정보

데이터 원본

확보할 당시의 원시 데이터나, 항상 최신 상태를 유지하여 신뢰할 수 있는 데이터

- 커뮤니케이션 비용 감소
- 의사결정 및 활동의 기준
- 가공된 데이터의 신뢰성 확보

데이터 훑어보기

- tibble:
- head: 최초 6행의 데이터를 보여줌(행갯수 조절 가능)
- tail: 마지막 6행의 데이터를 보여줌(행갯수 조절 가능)
- summary: 각 컬럼의 자료형과 숫자라면 대표값을 함께 보여줌
- str: 각 컬럼의 자료형과 초기 값을 보여줌
- length: 데이터의 길이 출력(vector)
- nrow: 행 갯수 출력(data.frame)
- is.na: NA 인지 확인
- complete.cases: 값이 모두 있는지 행단위로 검사

추천 데이터 - chennel

- cusID : 5자리 숫자조합으로 구성된 고객ID
- chennel: 접속 체널
- useCnt : 사용횟수(건)

```
##      cusID          chennel             useCnt      
##  Min.   :    7   Length:8824        Min.   :  1.00  
##  1st Qu.: 6107   Class :character   1st Qu.:  2.00  
##  Median : 9506   Mode  :character   Median :  7.00  
##  Mean   : 9835                      Mean   : 13.57  
##  3rd Qu.:13812                      3rd Qu.: 19.00  
##  Max.   :19382                      Max.   :240.00
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    8824 obs. of  3 variables:
##  $ cusID  : int  7 14 42 74 74 94 112 112 122 123 ...
##  $ chennel: chr  "A_MOBILE/APP" "A_MOBILE/APP" "B_MOBILE/APP" "A_MOBILE/APP" ...
##  $ useCnt : int  4 1 23 1 30 14 16 1 27 10 ...
##  - attr(*, "spec")=List of 2
##   ..- attr(*, "class")= chr "col_spec"
```

추천 데이터 - competitor

- cusID : 5자리 숫자조합으로 구성된 고객ID
- partner : 제휴사
- competitor: 경쟁사
- useDate : 이용년월(YYYYDD)

```
##      cusID              partner            competitor          useDate
##  Length:28159       Length:28159       Length:28159       Min.   :201501
##  Class :character   Class :character   Class :character   1st Qu.:201504
##  Mode  :character   Mode  :character   Mode  :character   Median :201507
##                                                           Mean   :201507
##                                                           3rd Qu.:201510
##                                                           Max.   :201512
```

추천 데이터 - competitor

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    28159 obs. of  4 variables:
##  $ cusID     : chr  "00002" "00051" "00077" "00077" ...
##  $ partner   : chr  "D" "D" "D" "D" ...
##  $ competitor: chr  "D02" "D01" "D02" "D02" ...
##  $ useDate   : int  201507 201504 201503 201506 201507 201508 201511 201510 201511 201508 ...
##  - attr(*, "spec")=List of 2
##   ..- attr(*, "class")= chr "col_spec"
```

추천 데이터 - customer

- cusID: 5자리 숫자조합으로 구성된 고객ID
- sex : 성별
- age : 연령
  - 19세이하, 20세~24세, 25세~29세, 30세~34세, 35세~39세, 40세~44세, 45세~49세, 50세~54세, 55세~59세, 60세이상
- area : 거주지역

추천 데이터 - customer

```
##     cusID               sex                 age           
##  Length:19383       Length:19383       Length:19383      
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##      area          
##  Length:19383      
##  Class :character  
##  Mode  :character
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    19383 obs. of  4 variables:
##  $ cusID: chr  "00001" "00002" "00003" "00004" ...
##  $ sex  : chr  "M" "M" "M" "F" ...
##  $ age  : chr  "60세이상" "60세이상" "60세이상" "60세이상" ...
##  $ area : chr  "060" "100" "033" "016" ...
##  - attr(*, "spec")=List of 2
##   ..- attr(*, "class")= chr "col_spec"
```

추천 데이터 - item

- partner : 재휴사
- cate_1 : 대분류
- cate_2 : 중분류
- cate_3 : 소분류
- cate_2_name: 중분류명
- cate_3_name: 소분류명

추천 데이터 - item

```
##    partner             cate_1             cate_2
##  Length:4386        Length:4386        Length:4386
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##     cate_3          cate_2_name        cate_3_name
##  Length:4386        Length:4386        Length:4386
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
```

추천 데이터 - item

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    4386 obs. of  6 variables:
##  $ partner    : chr  "A" "A" "A" "A" ...
##  $ cate_1     : chr  "01" "01" "01" "01" ...
##  $ cate_2     : chr  "0101" "0101" "0101" "0101" ...
##  $ cate_3     : chr  "A010101" "A010102" "A010103" "A010104" ...
##  $ cate_2_name: chr  "일용잡화" "일용잡화" "일용잡화" "일용잡화" ...
##  $ cate_3_name: chr  "위생세제" "휴지류" "뷰티상품" "일용잡화" ...
##  - attr(*, "spec")=List of 2
##   ..- attr(*, "class")= chr "col_spec"
```

추천 데이터 - membership

- cusID : 5자리 숫자조합으로 구성된 고객ID
- memberShip: 멤버십명
- regDate : 가입년월

```
##     cusID           memberShip          regDate
##  Length:7456        Length:7456        Min.   :201210
##  Class :character   Class :character   1st Qu.:201311
##  Mode  :character   Mode  :character   Median :201407
##                                        Mean   :201412
##                                        3rd Qu.:201504
##                                        Max.   :201512
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    7456 obs. of  3 variables:
##  $ cusID     : chr  "00011" "00021" "00037" "00043" ...
##  $ memberShip: chr  "하이마트" "하이마트" "하이마트" "하이마트" ...
##  $ regDate   : int  201512 201506 201306 201403 201411 201312 201506 201404 201406 201311 ...
##  - attr(*, "spec")=List of 2
##   ..- attr(*, "class")= chr "col_spec"
```

추천 데이터 - tran

```
##    partner            receiptNum          cate_1          cate_2
##  Length:28593030    Min.   :       1   Min.   : 1.00   Min.   : 101
##  Class :character   1st Qu.: 3922474   1st Qu.: 4.00   1st Qu.: 401
##  Mode  :character   Median : 7167787   Median :11.00   Median :1102
##                     Mean   : 6447881   Mean   :18.37   Mean   :1840
##                     3rd Qu.: 9116336   3rd Qu.:18.00   3rd Qu.:1808
##                     Max.   :11096601   Max.   :92.00   Max.   :9206
##     cate_3              cusID         storeCode           date
##  Length:28593030    Min.   :    1   Min.   :  1.00   Min.   :20140101
##  Class :character   1st Qu.: 5206   1st Qu.: 16.00   1st Qu.:20140711
##  Mode  :character   Median :10104   Median : 44.00   Median :20150110
##                     Mean   : 9904   Mean   : 92.26   Mean   :20145817
##                     3rd Qu.:14638   3rd Qu.:110.00   3rd Qu.:20150703
##                     Max.   :19383   Max.   :593.00   Max.   :20151231
##       time            amount
##  Min.   : 0.00   Min.   :        1
##  1st Qu.:14.00   1st Qu.:     2050
##  Median :17.00   Median :     4290
##  Mean   :16.71   Mean   :    23678
##  3rd Qu.:19.00   3rd Qu.:     9900
##  Max.   :23.00   Max.   :101330000
```

추천 데이터 - tran

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    28593030 obs. of  10 variables:
##  $ partner   : chr  "B" "B" "B" "B" ...
##  $ receiptNum: int  8664000 8664000 8664000 8664000 8664001 8664001 8664002 8664002 8664002 8664003 ...
##  $ cate_1    : int  15 16 16 18 5 15 10 43 54 5 ...
##  $ cate_2    : int  1504 1601 1602 1803 509 1501 1003 4301 5403 504 ...
##  $ cate_3    : chr  "B150401" "B160101" "B160201" "B180301" ...
##  $ cusID     : int  17218 17218 17218 17218 17674 17674 14388 14388 14388 15773 ...
##  $ storeCode : int  44 44 44 44 44 44 44 44 44 44 ...
##  $ date      : int  20140222 20140222 20140222 20140222 20140222 20140222 20140222 20140222 20140222 20140222 ...
##  $ time      : int  20 20 20 20 22 22 23 23 23 21 ...
##  $ amount    : int  2420 1070 8060 6000 1120 1200 5290 5960 9900 970 ...
##  - attr(*, "spec")=List of 2
##   ..- attr(*, "class")= chr "col_spec"
```

- 클라우드 서비스 소개
- 구글 클라우드
- RMySQL 연결

클라우드 서비스