R tidyverse 마스터 클래스

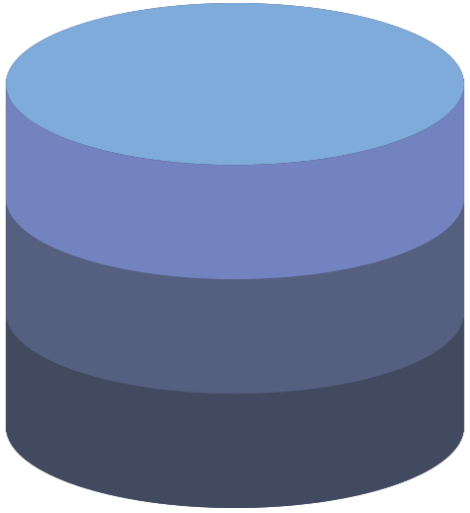
4강 - 우리만의 데이터 베이스 만들기

슬기로운통계생활

Issac Lee

서울시 이동데이터

데이터베이스 만들기



이동 데이터 다운받기

서울시 이동데이터

- 홈페이지 이동
- 자치구 단위 데이터에서 21년 1월 데이터를 다운받자. 자신의 컴퓨터 용량이 작다면 1월과 8월 데이터만 다운 받자.
- 링크 클릭하면 다음과 같은 2개 파일이 다운 받아진다.
  - `생활이동_자치구_202101.zip`

    ...
  - `생활이동_자치구_202108.zip`
- 압축을 풀게 되면, 각 월별 (폴더)에 24개의 파일 (시간별 기록)이 생기게 된다.

폴더 구조 및 자료 구조 보기

1개월 24개 파일
월별 1GB

```
list.dirs("./data/")
```

```
## [1] "./data/"
## [2] "./data//생활이동_자치구_202101"
```

```
list.files("./data/생활이동_자치구_202101/") %>% head()
```

```
## [1] "생활이동_자치구_202101_00시.csv"
## [2] "생활이동_자치구_202101_01시.csv"
## [3] "생활이동_자치구_202101_02시.csv"
## [4] "생활이동_자치구_202101_03시.csv"
## [5] "생활이동_자치구_202101_04시.csv"
## [6] "생활이동_자치구_202101_05시.csv"
```

Data base 만들기

R code

- 강의 페이지에서 R code를 긁어 복사 붙여넣기 해주세요.
- `folder_names`에는 다음과 같이 csv파일이 들어가 있는 폴더 이름들이 들어가 있다.

```
folder_names <- list.dirs("./data")[2]
folder_names
```

```
## [1] "./data/생활이동_자치구_202101"
```

- 코드는 `list.files()` 함수를 사용하여 파일 이름을 불러오고, 그 이름들을 사용하여 `csv` 파일을 하나씩 불러와 db 파일에 `seoul_moving_data` 테이블에 이어 붙이도록 되어있다.

전체 데이터 베이스 탐색

RSQLite를 사용한 DB 연결

```
library(DBI)
library(tidyverse)
library(magrittr)
con <- dbConnect(RSQLite::SQLite(),
                 "./data/movingdata_seoul.db")
dbListTables(con)[1:2]
```

```
## [1] "reference_data"         "seoul_moving_data_2021"
```

```
moving_db <- tbl(con, "seoul_moving_data_2021")
reference_db <- tbl(con, "reference_data")
```

월별 이동량

- 총 2억 천 7백만건 데이터

```
moving_db %>% tally()
```

```
moving_db %>%
  group_by(month) %>%
  tally()
```

```
## # Source:   lazy query [?? x 2]
## # Database: sqlite 3.36.0
## #   [C:\Users\issac\Documents\Teaching\R
##   month        n
##   <dbl>    <int>
## 1     1 24351892
## 2     2 25610731
## 3     3 27468233
## 4     4 27733676
## 5     5 28753287
## 6     6 28313847
## 7     7 27587624
## 8     8 27464405
```

내가 원하는 데이터 뽑아내기

서울시 출근 인구

```
commute_data <- moving_db %>%
    filter(trip_type == "HW" &
           departure_code < 20000 &
           arrival_code < 20000) %>% # 서울 내 이동
    mutate(generation =
               case_when(
                   between(age, 0, 25) ~ "청년층",
                   between(age, 25, 60) ~ "중년층",
                   age >= 65 ~ "장년층",
                   TRUE ~ as.integer(age)
               )) %>%
    group_by(month, generation) %>%
    summarise(population = sum(population, na.rm = TRUE)) %>%
    collect()
```

데이터 변환하기

dbplyr와 dplyr의 차이

추가적인 데이터 처리

- Factor level의 재조정 같은 경우 dbplyr에서 할 수 없음.
- 기반 데이터를 뽑아낸 후 알맞게 변형

```
commute_data %<>%
  mutate(population =
    round(population / 10000, digits = 1)) %>%
    mutate(generation = factor(generation,
                               levels = c("청년층", "중년층", "장년층"))
```

그래프 작성하기

Esquisse 패키지

- 기본 틀을 작성 후 수정

```
p <- ggplot(commute_data) +
    aes(x = month, y = population, fill = population) +
    geom_bar(stat="identity") +
    scale_fill_gradient(low = "#231AE4", high = "#F00202") +
    labs(
        x = "월 (2021년 기준)",
        y = "인구이동량 (단위: 만명)",
        title = "세대별 월별 인구 출근 이동량",
        fill = "인구이동량"
    ) +
    theme_minimal() +
    theme(legend.position = "right") +
    facet_wrap(vars(generation))
```

세대별 월별 출근 인구변화

인사이트?

같이 보면 좋은 책 추천

[1] R for Data Science

- 웹 상에 무료 공개된 책입니다.
- 위 교재의 한글 번역본 R 을 활용한 데이터과학도 있습니다.
- 도서 제목 클릭하셔서 구매하시면 저의 사리사욕을 충당하는데 도움이 됩니다.