따릉이 이용데이터 전처리

Jisu Kang

2025-11-23

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✔ ggplot2 3.5.1     ✔ purrr   1.0.2
## ✔ tibble  3.2.1     ✔ dplyr   1.1.4
## ✔ tidyr   1.3.1     ✔ stringr 1.5.1
## ✔ readr   2.1.3     ✔ forcats 0.5.2
```

```
## Warning: 패키지 'ggplot2'는 R 버전 4.2.3에서 작성되었습니다
```

```
## Warning: 패키지 'tibble'는 R 버전 4.2.3에서 작성되었습니다
```

```
## Warning: 패키지 'tidyr'는 R 버전 4.2.3에서 작성되었습니다
```

```
## Warning: 패키지 'purrr'는 R 버전 4.2.3에서 작성되었습니다
```

```
## Warning: 패키지 'dplyr'는 R 버전 4.2.3에서 작성되었습니다
```

```
## Warning: 패키지 'stringr'는 R 버전 4.2.3에서 작성되었습니다
```

```
## ── Conflicts ───────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(readr)
df2407 <- read_csv("C:/Users/강지수/iCloudDrive/2025-2/데이터시각화/기말과제/서울특별시 공공자전거 이용정보(일별)_2407.csv"
,locale = locale(encoding = "EUC-KR"))
```

```
## Rows: 2003560 Columns: 11
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (7): 대여소번호, 대여소, 대여구분코드, 성별, 연령대, 운동량, 탄소량
## dbl  (3): 이용건수, 이동거리(M), 이용시간(분)
## date (1): 대여일자
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
df2408 <- read_csv("C:/Users/강지수/iCloudDrive/2025-2/데이터시각화/기말과제/서울특별시 공공자전거 이용정보(일별)_2408.csv"
,locale = locale(encoding = "EUC-KR"))
```

```
## Rows: 2116470 Columns: 11
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (7): 대여소번호, 대여소, 대여구분코드, 성별, 연령대, 운동량, 탄소량
## dbl  (3): 이용건수, 이동거리(M), 이용시간(분)
## date (1): 대여일자
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
df2409 <- read_csv("C:/Users/강지수/iCloudDrive/2025-2/데이터시각화/기말과제/서울특별시 공공자전거 이용정보(일별)_2409.csv"
,locale = locale(encoding = "EUC-KR"))
```

```
## Rows: 2183685 Columns: 11
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (7): 대여소번호, 대여소, 대여구분코드, 성별, 연령대, 운동량, 탄소량
## dbl  (3): 이용건수, 이동거리(M), 이용시간(분)
## date (1): 대여일자
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
df2410 <- read_csv("C:/Users/강지수/iCloudDrive/2025-2/데이터시각화/기말과제/서울특별시 공공자전거 이용정보(일별)_2410.csv"
,locale = locale(encoding = "EUC-KR"))
```

```
## Rows: 2373305 Columns: 11
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (5): 대여소번호, 대여소, 대여구분코드, 성별, 연령대
## dbl  (5): 이용건수, 운동량, 탄소량, 이동거리(M), 이용시간(분)
## date (1): 대여일자
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
df2411 <- read_csv("C:/Users/강지수/iCloudDrive/2025-2/데이터시각화/기말과제/서울특별시 공공자전거 이용정보(일별)_2411.csv"
,locale = locale(encoding = "EUC-KR"))
```

```
## Rows: 1896324 Columns: 11
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (5): 대여소번호, 대여소, 대여구분코드, 성별, 연령대
## dbl  (5): 이용건수, 운동량, 탄소량, 이동거리(M), 이용시간(분)
## date (1): 대여일자
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
df2412 <- read_csv("C:/Users/강지수/iCloudDrive/2025-2/데이터시각화/기말과제/서울특별시 공공자전거 이용정보(일별)_2412.csv"
,locale = locale(encoding = "EUC-KR"))
```

```
## Rows: 1410916 Columns: 11
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (5): 대여소번호, 대여소, 대여구분코드, 성별, 연령대
## dbl  (5): 이용건수, 운동량, 탄소량, 이동거리(M), 이용시간(분)
## date (1): 대여일자
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
df2501 <- read_csv("C:/Users/강지수/iCloudDrive/2025-2/데이터시각화/기말과제/서울특별시 공공자전거 이용정보(일별)_2501.csv"
,locale = locale(encoding = "EUC-KR"))
```

```
## Rows: 1128074 Columns: 11
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (5): 대여소번호, 대여소, 대여구분코드, 성별, 연령대
## dbl  (5): 이용건수, 운동량, 탄소량, 이동거리(M), 이용시간(분)
## date (1): 대여일자
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
df2502 <- read_csv("C:/Users/강지수/iCloudDrive/2025-2/데이터시각화/기말과제/서울특별시 공공자전거 이용정보(일별)_2502.csv"
,locale = locale(encoding = "EUC-KR"))
```

```
## Rows: 1048575 Columns: 11
## ── Column specification ────────────────────────────────────────────────────
## Delimiter: ","
## chr  (6): 대여소, 대여구분코드, 성별, 연령대, 운동량, 탄소량
## dbl  (4): 대여소번호, 이용건수, 이동거리(M), 이용시간(분)
## date (1): 대여일자
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
df2503 <- read_csv("C:/Users/강지수/iCloudDrive/2025-2/데이터시각화/기말과제/서울특별시 공공자전거 이용정보(일별)_2503.csv"
,locale = locale(encoding = "EUC-KR"))
```

```
## Rows: 1709229 Columns: 11
## ── Column specification ────────────────────────────────────────────────────
## Delimiter: ","
## chr  (5): 대여소번호, 대여소, 대여구분코드, 성별, 연령대
## dbl  (5): 이용건수, 운동량, 탄소량, 이동거리(M), 이용시간(분)
## date (1): 대여일자
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
df2504 <- read_csv("C:/Users/강지수/iCloudDrive/2025-2/데이터시각화/기말과제/서울특별시 공공자전거 이용정보(일별)_2504.csv"
,locale = locale(encoding = "EUC-KR"))
```

```
## Rows: 1979115 Columns: 11
## ── Column specification ────────────────────────────────────────────────────
## Delimiter: ","
## chr  (5): 대여소번호, 대여소, 대여구분코드, 성별, 연령대
## dbl  (5): 이용건수, 운동량, 탄소량, 이동거리(M), 이용시간(분)
## date (1): 대여일자
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
df2505 <- read_csv("C:/Users/강지수/iCloudDrive/2025-2/데이터시각화/기말과제/서울특별시 공공자전거 이용정보(일별)_2505.csv"
,locale = locale(encoding = "EUC-KR"))
```

```
## Rows: 2101370 Columns: 11
## ── Column specification ────────────────────────────────────────────────────
## Delimiter: ","
## chr  (5): 대여소번호, 대여소, 대여구분코드, 성별, 연령대
## dbl  (5): 이용건수, 운동량, 탄소량, 이동거리(M), 이용시간(분)
## date (1): 대여일자
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
df2406 <- read_csv("C:/Users/강지수/iCloudDrive/2025-2/데이터시각화/기말과제/서울특별시 공공자전거 이용정보(일별)_2506.csv"
,locale = locale(encoding = "EUC-KR"))
```

```
## Rows: 2146663 Columns: 11
## ── Column specification ────────────────────────────────────────────────────
## Delimiter: ","
## chr  (5): 대여소번호, 대여소, 대여구분코드, 성별, 연령대
## dbl  (5): 이용건수, 운동량, 탄소량, 이동거리(M), 이용시간(분)
## date (1): 대여일자
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
# 1. 불러온 12개 데이터프레임을 리스트로 묶습니다.
data_list <- list(df2407, df2408, df2409, df2410, df2411, df2412,
                  df2501, df2502, df2503, df2504, df2505, df2406)
# 2. 모든 데이터프레임의 컬럼 타입을 통일하는 함수를 정의합니다.
standardize_df <- function(df) {
  df %>%
    mutate(
      # 문자형 (chr)으로 통일
      대여소번호 = as.character(대여소번호),
      대여소 = as.character(대여소),
      대여구분코드 = as.character(대여구분코드),
      성별 = as.character(성별),
      연령대 = as.character(연령대),

      # 숫자형 (dbl)으로 통일 (as.character()를 먼저 적용하여 NA를 유발하는 문자열을 처리)
      이용건수 = as.numeric(as.character(이용건수)),
      운동량 = as.numeric(as.character(운동량)),
      탄소량 = as.numeric(as.character(탄소량)),
      `이동거리(M)` = as.numeric(as.character(`이동거리(M)`)),
      `이용시간(분)` = as.numeric(as.character(`이용시간(분)`)),

      # 날짜형 (date)으로 통일 (만약 '대여일자'가 문자형으로 남아있다면 적용)
      대여일자 = as.Date(대여일자)
    )
}

# 3. purrr::map() 함수를 사용하여 리스트 내 모든 데이터프레임에 타입 통일 함수를 적용합니다.
df_standardized_list <- purrr::map(data_list, standardize_df)
```

```
## Warning: There were 2 warnings in `mutate()`.
## The first warning was:
## ℹ In argument: `운동량 = as.numeric(as.character(운동량))`.
## Caused by warning:
## ! 강제형변환에 의해 생성된 NA 입니다
## ℹ Run `dplyr::last_dplyr_warnings()` to see the 1 remaining warning.
## There were 2 warnings in `mutate()`.
## The first warning was:
## ℹ In argument: `운동량 = as.numeric(as.character(운동량))`.
## Caused by warning:
## ! 강제형변환에 의해 생성된 NA 입니다
## ℹ Run `dplyr::last_dplyr_warnings()` to see the 1 remaining warning.
## There were 2 warnings in `mutate()`.
## The first warning was:
## ℹ In argument: `운동량 = as.numeric(as.character(운동량))`.
## Caused by warning:
## ! 강제형변환에 의해 생성된 NA 입니다
## ℹ Run `dplyr::last_dplyr_warnings()` to see the 1 remaining warning.
## There were 2 warnings in `mutate()`.
## The first warning was:
## ℹ In argument: `운동량 = as.numeric(as.character(운동량))`.
## Caused by warning:
## ! 강제형변환에 의해 생성된 NA 입니다
## ℹ Run `dplyr::last_dplyr_warnings()` to see the 1 remaining warning.
```

```r
# 4. 타입이 통일된 리스트를 하나로 합칩니다.
df_final <- dplyr::bind_rows(df_standardized_list)

# 5. 최종 데이터프레임의 구조를 확인합니다.
print(dplyr::glimpse(df_final))
```

```
## Rows: 22,097,286
## Columns: 11
## $ 대여일자       <date> 2024-07-01, 2024-07-01, 2024-07-01, 2024-07-01, 2024-0…
## $ 대여소번호     <chr> "00739", "00743", "01027", "01044", "01271", "01447", "…
## $ 대여소         <chr> "739. 신월사거리", "743. 현대6차아파트 101동 옆", "1027…
## $ 대여구분코드   <chr> "정기권", "정기권", "정기권", "정기권", "정기권", "정기…
## $ 성별           <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…
## $ 연령대         <chr> "~10대", "~10대", "~10대", "~10대", "~10대", "~10대", "…
## $ 이용건수       <dbl> 1, 1, 1, 2, 3, 1, 1, 1, 2, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1…
## $ 운동량         <dbl> 54.44, 0.00, 57.01, 67.05, 151.10, 22.89, 49.76, 25.96,…
## $ 탄소량         <dbl> 0.48, 0.00, 0.51, 0.61, 1.51, 0.21, 0.48, 0.23, 4.18, 0…
## $ `이동거리(M)`  <dbl> 2051.70, 0.00, 2214.97, 2605.15, 6523.42, 889.46, 2060.…
## $ `이용시간(분)` <dbl> 11, 2, 17, 15, 31, 6, 14, 5, 104, 53, 5, 5, 11, 6, 2, 1…
## # A tibble: 22,097,286 × 11
##    대여일자   대여소번호 대여소 대여구분코드 성별  연령대 이용건수 운동량 탄소량
##    <date>     <chr>      <chr>  <chr>        <chr> <chr>     <dbl>  <dbl>  <dbl>
##  1 2024-07-01 00739      739. … 정기권       <NA>  ~10대         1   54.4   0.48
##  2 2024-07-01 00743      743. … 정기권       <NA>  ~10대         1    0     0
##  3 2024-07-01 01027      1027.… 정기권       <NA>  ~10대         1   57.0   0.51
##  4 2024-07-01 01044      1044.… 정기권       <NA>  ~10대         2   67.0   0.61
##  5 2024-07-01 01271      1271.… 정기권       <NA>  ~10대         3  151.    1.51
##  6 2024-07-01 01447      1447.… 정기권       <NA>  ~10대         1   22.9   0.21
##  7 2024-07-01 01532      1532.… 정기권       <NA>  ~10대         1   49.8   0.48
##  8 2024-07-01 01651      1651.… 정기권       <NA>  ~10대         1   26.0   0.23
##  9 2024-07-01 01656      1656.… 정기권       <NA>  ~10대         2  405.    4.18
## 10 2024-07-01 01720      1720.… 정기권       <NA>  ~10대         1    0     0
## # i 22,097,276 more rows
## # i 2 more variables: `이동거리(M)` <dbl>, `이용시간(분)` <dbl>
```

```r
# 운동량과 탄소량 컬럼의 NA 개수를 확인합니다.
df_final %>%
  summarise(
    na_운동량 = sum(is.na(운동량)),
    na_탄소량 = sum(is.na(탄소량)),
    total_rows = n()
  )
```

```
## # A tibble: 1 × 3
##   na_운동량 na_탄소량 total_rows
##       <int>     <int>      <int>
## 1     66731     66731   22097286
```

```r
summary(df_final)
```

```
##     대여일자           대여소번호           대여소          대여구분코드
##  Min.   :2024-07-01   Length:22097286    Length:22097286    Length:22097286
##  1st Qu.:2024-09-21   Class :character   Class :character   Class :character
##  Median :2024-12-10   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2024-12-26
##  3rd Qu.:2025-04-11
##  Max.   :2025-06-30
##
##      성별              연령대             이용건수           운동량
##  Length:22097286    Length:22097286    Min.   :  1.000   Min.   :    0.00
##  Class :character   Class :character   1st Qu.:  1.000   1st Qu.:   28.95
##  Mode  :character   Mode  :character   Median :  1.000   Median :   60.36
##                                        Mean   :  1.815   Mean   :  111.62
##                                        3rd Qu.:  2.000   3rd Qu.:  132.88
##                                        Max.   :151.000   Max.   :16536.42
##                                                          NA's   :66731
##      탄소량        이동거리(M)      이용시간(분)
##  Min.   :  0.00   Min.   :     0   Min.   :   0.00
##  1st Qu.:  0.26   1st Qu.:  1120   1st Qu.:   9.00
##  Median :  0.53   Median :  2303   Median :  20.00
##  Mean   :  0.97   Mean   :  4180   Mean   :  37.68
##  3rd Qu.:  1.16   3rd Qu.:  5015   3rd Qu.:  48.00
##  Max.   :160.34   Max.   :708798   Max.   :5967.00
##  NA's   :66731
```

```r
# lubridate 패키지를 사용해 월 정보를 추출합니다.
library(lubridate)
```

```
##
## 다음의 패키지를 부착합니다: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
df_na_check <- df_final %>%
  # 대여일자에서 월(month)을 추출하여 새로운 컬럼을 생성합니다.
  mutate(
    YearMonth = format(대여일자, "%Y-%m")
  ) %>%
  # 월별로 그룹화하여 운동량 또는 탄소량이 NA인 행의 개수를 셉니다.
  group_by(YearMonth) %>%
  summarise(
    NA_Count = sum(is.na(운동량) | is.na(탄소량)),
    Total_Rows = n()
  ) %>%
  # NA 비율을 계산합니다.
  mutate(
    NA_Ratio = (NA_Count / Total_Rows) * 100
  ) %>%
  arrange(desc(NA_Count))

print(df_na_check)
```

```
## # A tibble: 12 × 4
##    YearMonth NA_Count Total_Rows NA_Ratio
##    <chr>        <int>      <int>    <dbl>
##  1 2025-04       7109    1979115    0.359
##  2 2025-06       6695    2146663    0.312
##  3 2025-05       6613    2101370    0.315
##  4 2025-03       6405    1709229    0.375
##  5 2024-10       6186    2373305    0.261
##  6 2024-11       5806    1896324    0.306
##  7 2024-09       5547    2183685    0.254
##  8 2024-07       5458    2003560    0.272
##  9 2024-08       5333    2116470    0.252
## 10 2024-12       4775    1410916    0.338
## 11 2025-02       3482    1048575    0.332
## 12 2025-01       3322    1128074    0.294
```

```
library(dplyr)
df_clean <- df_final %>%
  drop_na(운동량, 탄소량)

# 3. 클리닝 후 최종 데이터셋의 행 수와 구조를 다시 확인합니다.
print(paste("클리닝 후 최종 데이터 행 수:", nrow(df_clean)))
```

```
## [1] "클리닝 후 최종 데이터 행 수: 22030555"
```

```
print(dplyr::glimpse(df_clean))
```

```
## Rows: 22,030,555
## Columns: 11
## $ 대여일자       <date> 2024-07-01, 2024-07-01, 2024-07-01, 2024-07-01, 2024-0…
## $ 대여소번호     <chr> "00739", "00743", "01027", "01044", "01271", "01447", "…
## $ 대여소         <chr> "739. 신월사거리", "743. 현대6차아파트 101동 옆", "1027…
## $ 대여구분코드   <chr> "정기권", "정기권", "정기권", "정기권", "정기권", "정기…
## $ 성별           <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…
## $ 연령대         <chr> "~10대", "~10대", "~10대", "~10대", "~10대", "~10대", "…
## $ 이용건수       <dbl> 1, 1, 1, 2, 3, 1, 1, 1, 2, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1…
## $ 운동량         <dbl> 54.44, 0.00, 57.01, 67.05, 151.10, 22.89, 49.76, 25.96,…
## $ 탄소량         <dbl> 0.48, 0.00, 0.51, 0.61, 1.51, 0.21, 0.48, 0.23, 4.18, 0…
## $ `이동거리(M)`  <dbl> 2051.70, 0.00, 2214.97, 2605.15, 6523.42, 889.46, 2060.…
## $ `이용시간(분)` <dbl> 11, 2, 17, 15, 31, 6, 14, 5, 104, 53, 5, 5, 11, 6, 2, 1…
## # A tibble: 22,030,555 × 11
##    대여일자   대여소번호 대여소 대여구분코드 성별  연령대 이용건수 운동량 탄소량
##    <date>     <chr>      <chr>  <chr>        <chr> <chr>     <dbl>  <dbl>  <dbl>
##  1 2024-07-01 00739      739. … 정기권       <NA>  ~10대         1   54.4   0.48
##  2 2024-07-01 00743      743. … 정기권       <NA>  ~10대         1    0     0
##  3 2024-07-01 01027      1027.… 정기권       <NA>  ~10대         1   57.0   0.51
##  4 2024-07-01 01044      1044.… 정기권       <NA>  ~10대         2   67.0   0.61
##  5 2024-07-01 01271      1271.… 정기권       <NA>  ~10대         3  151.    1.51
##  6 2024-07-01 01447      1447.… 정기권       <NA>  ~10대         1   22.9   0.21
##  7 2024-07-01 01532      1532.… 정기권       <NA>  ~10대         1   49.8   0.48
##  8 2024-07-01 01651      1651.… 정기권       <NA>  ~10대         1   26.0   0.23
##  9 2024-07-01 01656      1656.… 정기권       <NA>  ~10대         2  405.    4.18
## 10 2024-07-01 01720      1720.… 정기권       <NA>  ~10대         1    0     0
## # i 22,030,545 more rows
## # i 2 more variables: `이동거리(M)` <dbl>, `이용시간(분)` <dbl>
```