Lab 05

Joyce Gill

2026-02-10

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

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

```
library(nycflights13)

x <- data.frame(key = 1:3,
                val_x = paste0("x", 1:3))
y <- data.frame(key = c(1,2,4),
                val_y = paste0("y", 1:3))
z <- data.frame(KEY = c(2,4,6),
                val_z = paste0("z", 1:3))
```

```
library(dplyr)
library(ggplot2)

students <- data.frame(
  student_id = c("S001", "S002", "S003", "S004", "S005", "S007", "S008"),
  name = c("Alice", "Bob", "Charlie", "Diana", "Evan", "Fiona", "George"),
  major = c(
    "Statistics",
    "Mathematics",
    "Computer Science",
    "Statistics",
    "Physics",
    "Biology",
    "Mathematics"
  )
)

# enrollments.csv
enrollments <- data.frame(
  student_id = c("S001", "S001","S002", "S002", "S003",
                 "S006", "S007", "S008", "S008"),
  course_id = c("C101", "C102", "C101", "C103", "C103",
```

```
                "C101", "C104","C102", "C105"),
  semester = rep("Fall", 9)
)

# courses.csv
courses <- data.frame(
  course_id = c("C101", "C102", "C103", "C104", "C106"),
  course_name = c(
    "Intro to Statistics",
    "Linear Algebra",
    "Econometrics",
    "Intro to Political Science",
    "Statistical Computing"
  ),
  instructor = paste("Prof.", c("Wells", "Hazel", "Lee", "Jozwiak", "Friedrichsen"))
)
```

Question 1

```
q1 <- inner_join(enrollments, students, by = "student_id")
q1
```

```
##   student_id course_id semester    name            major
## 1       S001      C101     Fall   Alice       Statistics
## 2       S001      C102     Fall   Alice       Statistics
## 3       S002      C101     Fall     Bob      Mathematics
## 4       S002      C103     Fall     Bob      Mathematics
## 5       S003      C103     Fall Charlie Computer Science
## 6       S007      C104     Fall   Fiona          Biology
## 7       S008      C102     Fall  George      Mathematics
## 8       S008      C105     Fall  George      Mathematics
```

```
q1_full <- q1 %>%
  inner_join(courses, by = "course_id")
q1_full
```

```
##   student_id course_id semester    name            major
## 1       S001      C101     Fall   Alice       Statistics
## 2       S001      C102     Fall   Alice       Statistics
## 3       S002      C101     Fall     Bob      Mathematics
## 4       S002      C103     Fall     Bob      Mathematics
## 5       S003      C103     Fall Charlie Computer Science
## 6       S007      C104     Fall   Fiona          Biology
## 7       S008      C102     Fall  George      Mathematics
##                   course_name    instructor
## 1        Intro to Statistics   Prof. Wells
## 2             Linear Algebra   Prof. Hazel
## 3        Intro to Statistics   Prof. Wells
## 4               Econometrics     Prof. Lee
## 5               Econometrics     Prof. Lee
## 6 Intro to Political Science Prof. Jozwiak
## 7             Linear Algebra   Prof. Hazel
```

Question 2

```
q2 <- enrollments %>%
  inner_join(students, by = "student_id") %>%
  inner_join(courses, by = "course_id")
q2
```

```
##   student_id course_id semester    name            major
## 1       S001      C101     Fall   Alice       Statistics
## 2       S001      C102     Fall   Alice       Statistics
## 3       S002      C101     Fall     Bob      Mathematics
## 4       S002      C103     Fall     Bob      Mathematics
## 5       S003      C103     Fall Charlie Computer Science
## 6       S007      C104     Fall   Fiona          Biology
## 7       S008      C102     Fall  George      Mathematics
##                  course_name    instructor
## 1        Intro to Statistics   Prof. Wells
## 2             Linear Algebra   Prof. Hazel
## 3        Intro to Statistics   Prof. Wells
## 4               Econometrics     Prof. Lee
## 5               Econometrics     Prof. Lee
## 6 Intro to Political Science Prof. Jozwiak
## 7             Linear Algebra   Prof. Hazel
```

Question 3

```
q3 <- enrollments %>%
  inner_join(courses, by = "course_id") %>%
  filter(course_name == "Intro to Statistics")
q3
```

```
##   student_id course_id semester         course_name  instructor
## 1       S001      C101     Fall Intro to Statistics Prof. Wells
## 2       S002      C101     Fall Intro to Statistics Prof. Wells
## 3       S006      C101     Fall Intro to Statistics Prof. Wells
```

```
n_distinct(q3$student_id)
```

```
## [1] 3
```

Question 4

```
q4 <- left_join(students, enrollments, by = "student_id")
q4
```

```
##    student_id    name            major course_id semester
## 1        S001   Alice       Statistics      C101     Fall
## 2        S001   Alice       Statistics      C102     Fall
## 3        S002     Bob      Mathematics      C101     Fall
## 4        S002     Bob      Mathematics      C103     Fall
## 5        S003 Charlie Computer Science      C103     Fall
## 6        S004   Diana       Statistics      <NA>     <NA>
## 7        S005    Evan          Physics      <NA>     <NA>
## 8        S007   Fiona          Biology      C104     Fall
## 9        S008  George      Mathematics      C102     Fall
```

```
## 10       S008  George      Mathematics      C105     Fall
```

```
q4 %>%
  filter(is.na(course_id))
```

```
##   student_id  name      major course_id semester
## 1       S004 Diana Statistics      <NA>     <NA>
## 2       S005  Evan    Physics      <NA>     <NA>
```

Question 5

```
q5 <- courses %>%
  left_join(enrollments, by = "course_id") %>%
  left_join(students, by = "student_id")

q5
```

```
##   course_id               course_name         instructor student_id semester
## 1      C101        Intro to Statistics        Prof. Wells       S001     Fall
## 2      C101        Intro to Statistics        Prof. Wells       S002     Fall
## 3      C101        Intro to Statistics        Prof. Wells       S006     Fall
## 4      C102             Linear Algebra        Prof. Hazel       S001     Fall
## 5      C102             Linear Algebra        Prof. Hazel       S008     Fall
## 6      C103               Econometrics          Prof. Lee       S002     Fall
## 7      C103               Econometrics          Prof. Lee       S003     Fall
## 8      C104 Intro to Political Science      Prof. Jozwiak       S007     Fall
## 9      C106      Statistical Computing Prof. Friedrichsen       <NA>     <NA>
##      name            major
## 1   Alice       Statistics
## 2     Bob      Mathematics
## 3    <NA>             <NA>
## 4   Alice       Statistics
## 5  George      Mathematics
## 6     Bob      Mathematics
## 7 Charlie Computer Science
## 8   Fiona          Biology
## 9    <NA>             <NA>
```

Question 6

```
q6 <- anti_join(students, enrollments, by = "student_id")
q6
```

```
##   student_id  name      major
## 1       S004 Diana Statistics
## 2       S005  Evan    Physics
```

Question 7

```
anti_join(courses, enrollments, by = "course_id")
```

```
##   course_id           course_name         instructor
## 1      C106 Statistical Computing Prof. Friedrichsen
```

```
#courses %>%
  #left_join(enrollments, by = "course_id") %>%
 # filter(is.na(student_id))
```

idk

Question 8

Students and courses do not share a common key. You need the enrollments table as a bridge.

Question 9

```
q9 <- enrollments %>%
  left_join(courses, by = "course_id")

q9
```

```
##   student_id course_id semester               course_name    instructor
## 1       S001      C101     Fall        Intro to Statistics   Prof. Wells
## 2       S001      C102     Fall             Linear Algebra   Prof. Hazel
## 3       S002      C101     Fall        Intro to Statistics   Prof. Wells
## 4       S002      C103     Fall               Econometrics     Prof. Lee
## 5       S003      C103     Fall               Econometrics     Prof. Lee
## 6       S006      C101     Fall        Intro to Statistics   Prof. Wells
## 7       S007      C104     Fall Intro to Political Science Prof. Jozwiak
## 8       S008      C102     Fall             Linear Algebra   Prof. Hazel
## 9       S008      C105     Fall                       <NA>          <NA>
```

C105 doesn't have a course name

Question 10

```
full_join(enrollments, courses, by = "course_id")
```

```
##    student_id course_id semester               course_name         instructor
## 1        S001      C101     Fall        Intro to Statistics        Prof. Wells
## 2        S001      C102     Fall             Linear Algebra        Prof. Hazel
## 3        S002      C101     Fall        Intro to Statistics        Prof. Wells
## 4        S002      C103     Fall               Econometrics          Prof. Lee
## 5        S003      C103     Fall               Econometrics          Prof. Lee
## 6        S006      C101     Fall        Intro to Statistics        Prof. Wells
## 7        S007      C104     Fall Intro to Political Science      Prof. Jozwiak
## 8        S008      C102     Fall             Linear Algebra        Prof. Hazel
## 9        S008      C105     Fall                       <NA>               <NA>
## 10       <NA>      C106     <NA>      Statistical Computing Prof. Friedrichsen
```

Bad for counting because it includes unmatched rows but good for auditing because it shows students that are auditing too.

Question 11

Using an inner join is minimal so it may show less data

Question 12

Inner join removes students with no enrollments

Question 13

```
left_join(students, enrollments)
```

```
## Joining with `by = join_by(student_id)`
```

```
##    student_id    name            major course_id semester
## 1        S001   Alice       Statistics      C101     Fall
## 2        S001   Alice       Statistics      C102     Fall
## 3        S002     Bob      Mathematics      C101     Fall
## 4        S002     Bob      Mathematics      C103     Fall
## 5        S003 Charlie Computer Science      C103     Fall
## 6        S004   Diana       Statistics      <NA>     <NA>
## 7        S005    Evan          Physics      <NA>     <NA>
## 8        S007   Fiona          Biology      C104     Fall
## 9        S008  George      Mathematics      C102     Fall
## 10       S008  George      Mathematics      C105     Fall
```

NA means the student exists but has no matching enrollment record

Question 14

```
anti_join(enrollments, courses)
```

```
## Joining with `by = join_by(course_id)`
```

```
##   student_id course_id semester
## 1       S008      C105     Fall
```

```
semi_join(enrollments, courses)
```

```
## Joining with `by = join_by(course_id)`
```

```
##   student_id course_id semester
## 1       S001      C101     Fall
## 2       S001      C102     Fall
## 3       S002      C101     Fall
## 4       S002      C103     Fall
## 5       S003      C103     Fall
## 6       S006      C101     Fall
## 7       S007      C104     Fall
## 8       S008      C102     Fall
```

Anti_join returns enrollments with no matching courses Semi_join returns enrollments that ahve valid courses