Exercícios Cap 13

Rodrigo Giannotti

05/09/2020

Capitulo 13

Inicialização

```
library(tidyverse)
library(magrittr) # mais pipes, como %<>%
library(lubridate) # melhor manejo de datas
library(maps)
```

Para o capitulo 13 também utilizaremos a biblioteca de voos de NYC

```
library(nycflights13)
# ?flights
# View(flights)
head(flights)
```

```
## # A tibble: 6 x 19
##    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
## 1  2013     1     1      517            515         2      830            819
## 2  2013     1     1      533            529         4      850            830
## 3  2013     1     1      542            540         2      923            850
## 4  2013     1     1      544            545        -1     1004           1022
## 5  2013     1     1      554            600        -6      812            837
## 6  2013     1     1      554            558        -4      740            728
## # ... with 11 more variables: arr_delay <dbl>, carrier <chr>, flight <int>,
## #   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
## #   hour <dbl>, minute <dbl>, time_hour <dttm>
```

Exercícios

13.2 nycflights13

13.2.1

Imagine you wanted to draw (approximately) the route each plane flies from its origin to its destination. What variables would you need? What tables would you need to combine?

13.2.2

I forgot to draw the relationship between weather and airports. What is the relationship and how should it appear in the diagram?

13.2.3

weather only contains information for the origin (NYC) airports. If it contained weather records for all airports in the USA, what additional relation would it define with flights?

13.2.4

We know that some days of the year are “special”, and fewer people than usual fly on them. How might you represent that data as a data frame? What would be the primary keys of that table? How would it connect to the existing tables?

13.3 Keys

13.3.1

Add a surrogate key to flights.

13.3.2

Identify the keys in the following datasets

13.3.2.1 Lahman::Batting

13.3.2.2 babynames::babynames

13.3.2.3 nasaweather::atmos

13.3.2.4 fueleconomy::vehicles

13.3.2.5 ggplot2::diamonds

(You might need to install some packages and read some documentation.)

13.3.3

Draw a diagram illustrating the connections between the Batting, Master, and Salaries tables in the Lahman package. Draw another diagram that shows the relationship between Master, Managers, AwardsManagers.

13.3.4

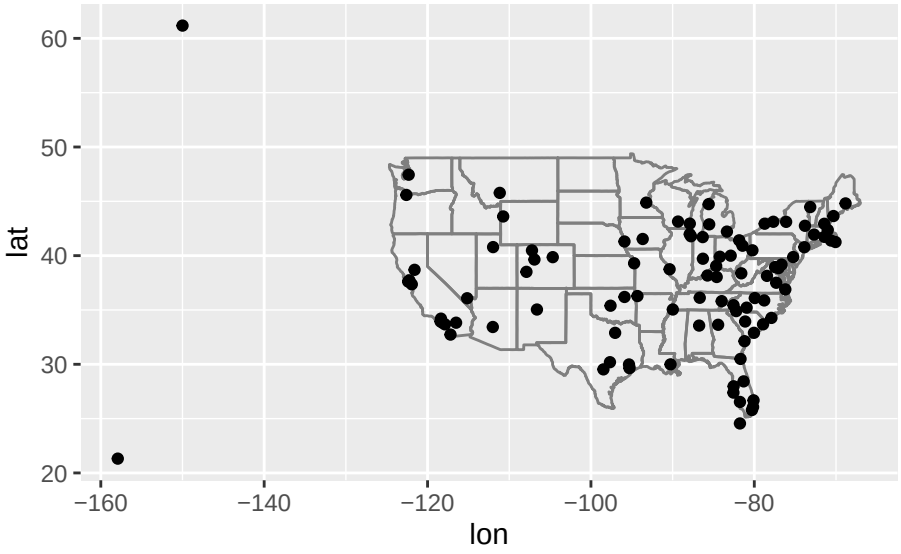
How would you characterise the relationship between the Batting, Pitching, and Fielding tables?

13.4 Mutating joins

13.4.1

Compute the average delay by destination, then join on the airports data frame so you can show the spatial distribution of delays. Here's an easy way to draw a map of the United States:

```
airports %>%
  semi_join(flights, c("faa" = "dest")) %>%
  ggplot(aes(lon, lat)) +
    borders("state") +
    geom_point() +
    coord_quickmap()
```




```
# Don't worry if you don't understand what semi_join() does - you'll learn about it next.
# You might want to use the size or colour of the points to display the average delay for each airport.
```

13.4.2

Add the location of the origin and destination (i.e. the lat and lon) to flights.

13.4.3

Is there a relationship between the age of a plane and its delays?

13.4.4

What weather conditions make it more likely to see a delay?

13.4.5

What happened on June 13 2013? Display the spatial pattern of delays, and then use Google to cross-reference with the weather.

13.5 Filtering joins

13.5.1

What does it mean for a flight to have a missing tailnum? What do the tail numbers that don't have a matching record in planes have in common? (Hint: one variable explains ~90% of the problems.)

13.5.2

Filter flights to only show flights with planes that have flown at least 100 flights.

13.5.3

Combine fueleconomy::vehicles and fueleconomy::common to find only the records for the most common models.

13.5.4

Find the 48 hours (over the course of the whole year) that have the worst delays. Cross-reference it with the weather data. Can you see any patterns?

13.5.5

What does anti_join(flights, airports, by = c("dest" = "faa")) tell you? What does anti_join(airports, flights, by = c("faa" = "dest")) tell you?

13.5.6

You might expect that there's an implicit relationship between plane and airline, because each plane is flown by a single airline. Confirm or reject this hypothesis using the tools you've learned above.