Chapter 13 - Relational Data

library(tidyverse)  
library(nycflights13)

# Airlines

flights

## # A tibble: 336,776 x 19  
## year month day dep\_time sched\_dep\_time dep\_delay arr\_time  
## <int> <int> <int> <int> <int> <dbl> <int>  
## 1 2013 1 1 517 515 2 830  
## 2 2013 1 1 533 529 4 850  
## 3 2013 1 1 542 540 2 923  
## 4 2013 1 1 544 545 -1 1004  
## 5 2013 1 1 554 600 -6 812  
## 6 2013 1 1 554 558 -4 740  
## 7 2013 1 1 555 600 -5 913  
## 8 2013 1 1 557 600 -3 709  
## 9 2013 1 1 557 600 -3 838  
## 10 2013 1 1 558 600 -2 753  
## # ... with 336,766 more rows, and 12 more variables: sched\_arr\_time <int>,  
## # 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>

airlines

## # A tibble: 16 x 2  
## carrier name   
## <chr> <chr>   
## 1 9E Endeavor Air Inc.   
## 2 AA American Airlines Inc.   
## 3 AS Alaska Airlines Inc.   
## 4 B6 JetBlue Airways   
## 5 DL Delta Air Lines Inc.   
## 6 EV ExpressJet Airlines Inc.   
## 7 F9 Frontier Airlines Inc.   
## 8 FL AirTran Airways Corporation  
## 9 HA Hawaiian Airlines Inc.   
## 10 MQ Envoy Air   
## 11 OO SkyWest Airlines Inc.   
## 12 UA United Air Lines Inc.   
## 13 US US Airways Inc.   
## 14 VX Virgin America   
## 15 WN Southwest Airlines Co.   
## 16 YV Mesa Airlines Inc.

airports

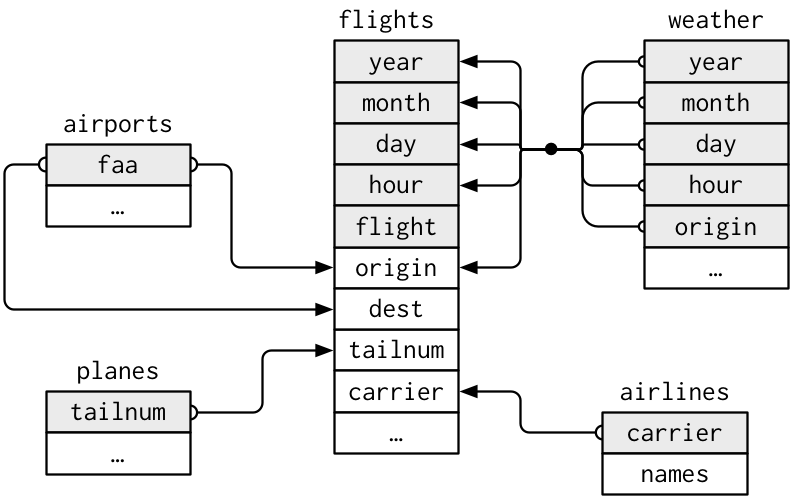
## # A tibble: 1,458 x 8  
## faa name lat lon alt tz dst tzone   
## <chr> <chr> <dbl> <dbl> <int> <dbl> <chr> <chr>   
## 1 04G Lansdowne Airport 41.1 -80.6 1044 -5 A America/New\_…  
## 2 06A Moton Field Municip… 32.5 -85.7 264 -6 A America/Chic…  
## 3 06C Schaumburg Regional 42.0 -88.1 801 -6 A America/Chic…  
## 4 06N Randall Airport 41.4 -74.4 523 -5 A America/New\_…  
## 5 09J Jekyll Island Airpo… 31.1 -81.4 11 -5 A America/New\_…  
## 6 0A9 Elizabethton Munici… 36.4 -82.2 1593 -5 A America/New\_…  
## 7 0G6 Williams County Air… 41.5 -84.5 730 -5 A America/New\_…  
## 8 0G7 Finger Lakes Region… 42.9 -76.8 492 -5 A America/New\_…  
## 9 0P2 Shoestring Aviation… 39.8 -76.6 1000 -5 U America/New\_…  
## 10 0S9 Jefferson County In… 48.1 -123. 108 -8 A America/Los\_…  
## # ... with 1,448 more rows

planes

## # A tibble: 3,322 x 9  
## tailnum year type manufacturer model engines seats speed engine  
## <chr> <int> <chr> <chr> <chr> <int> <int> <int> <chr>   
## 1 N10156 2004 Fixed wi… EMBRAER EMB-1… 2 55 NA Turbo…  
## 2 N102UW 1998 Fixed wi… AIRBUS INDUS… A320-… 2 182 NA Turbo…  
## 3 N103US 1999 Fixed wi… AIRBUS INDUS… A320-… 2 182 NA Turbo…  
## 4 N104UW 1999 Fixed wi… AIRBUS INDUS… A320-… 2 182 NA Turbo…  
## 5 N10575 2002 Fixed wi… EMBRAER EMB-1… 2 55 NA Turbo…  
## 6 N105UW 1999 Fixed wi… AIRBUS INDUS… A320-… 2 182 NA Turbo…  
## 7 N107US 1999 Fixed wi… AIRBUS INDUS… A320-… 2 182 NA Turbo…  
## 8 N108UW 1999 Fixed wi… AIRBUS INDUS… A320-… 2 182 NA Turbo…  
## 9 N109UW 1999 Fixed wi… AIRBUS INDUS… A320-… 2 182 NA Turbo…  
## 10 N110UW 1999 Fixed wi… AIRBUS INDUS… A320-… 2 182 NA Turbo…  
## # ... with 3,312 more rows

weather

## # A tibble: 26,115 x 15  
## origin year month day hour temp dewp humid wind\_dir wind\_speed  
## <chr> <dbl> <dbl> <int> <int> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 EWR 2013 1 1 1 39.0 26.1 59.4 270 10.4   
## 2 EWR 2013 1 1 2 39.0 27.0 61.6 250 8.06  
## 3 EWR 2013 1 1 3 39.0 28.0 64.4 240 11.5   
## 4 EWR 2013 1 1 4 39.9 28.0 62.2 250 12.7   
## 5 EWR 2013 1 1 5 39.0 28.0 64.4 260 12.7   
## 6 EWR 2013 1 1 6 37.9 28.0 67.2 240 11.5   
## 7 EWR 2013 1 1 7 39.0 28.0 64.4 240 15.0   
## 8 EWR 2013 1 1 8 39.9 28.0 62.2 250 10.4   
## 9 EWR 2013 1 1 9 39.9 28.0 62.2 260 15.0   
## 10 EWR 2013 1 1 10 41 28.0 59.6 260 13.8   
## # ... with 26,105 more rows, and 5 more variables: wind\_gust <dbl>,  
## # precip <dbl>, pressure <dbl>, visib <dbl>, time\_hour <dttm>



Check is there are any duplicate tailnumbers.

planes %>%   
 count(tailnum) %>%   
 filter(n > 1)

## # A tibble: 0 x 2  
## # ... with 2 variables: tailnum <chr>, n <int>

Duplicated?

weather %>%   
 count(year, month, day, hour, origin) %>%   
 filter(n > 1)

## # A tibble: 3 x 6  
## year month day hour origin n  
## <dbl> <dbl> <int> <int> <chr> <int>  
## 1 2013 11 3 1 EWR 2  
## 2 2013 11 3 1 JFK 2  
## 3 2013 11 3 1 LGA 2

Duplicates?

Note there are duplicated dates for flights and tailnum in the flights dataset. This may be a problem.

flights %>%   
 count(year, month, day, flight) %>%   
 filter(n > 1)

## # A tibble: 29,768 x 5  
## year month day flight n  
## <int> <int> <int> <int> <int>  
## 1 2013 1 1 1 2  
## 2 2013 1 1 3 2  
## 3 2013 1 1 4 2  
## 4 2013 1 1 11 3  
## 5 2013 1 1 15 2  
## 6 2013 1 1 21 2  
## 7 2013 1 1 27 4  
## 8 2013 1 1 31 2  
## 9 2013 1 1 32 2  
## 10 2013 1 1 35 2  
## # ... with 29,758 more rows

flights %>%   
 count(year, month, day, tailnum) %>%   
 filter(n > 1)

## # A tibble: 64,928 x 5  
## year month day tailnum n  
## <int> <int> <int> <chr> <int>  
## 1 2013 1 1 N0EGMQ 2  
## 2 2013 1 1 N11189 2  
## 3 2013 1 1 N11536 2  
## 4 2013 1 1 N11544 3  
## 5 2013 1 1 N11551 2  
## 6 2013 1 1 N12540 2  
## 7 2013 1 1 N12567 2  
## 8 2013 1 1 N13123 2  
## 9 2013 1 1 N13538 3  
## 10 2013 1 1 N13566 3  
## # ... with 64,918 more rows

Join airline name to the flights data.

flights2 <- flights %>%   
 select(year:day, hour, origin, dest, tailnum, carrier)  
flights2

## # A tibble: 336,776 x 8  
## year month day hour origin dest tailnum carrier  
## <int> <int> <int> <dbl> <chr> <chr> <chr> <chr>   
## 1 2013 1 1 5 EWR IAH N14228 UA   
## 2 2013 1 1 5 LGA IAH N24211 UA   
## 3 2013 1 1 5 JFK MIA N619AA AA   
## 4 2013 1 1 5 JFK BQN N804JB B6   
## 5 2013 1 1 6 LGA ATL N668DN DL   
## 6 2013 1 1 5 EWR ORD N39463 UA   
## 7 2013 1 1 6 EWR FLL N516JB B6   
## 8 2013 1 1 6 LGA IAD N829AS EV   
## 9 2013 1 1 6 JFK MCO N593JB B6   
## 10 2013 1 1 6 LGA ORD N3ALAA AA   
## # ... with 336,766 more rows

flights2 %>%  
 select(-origin, -dest) %>%   
 left\_join(airlines, by = "carrier")

## # A tibble: 336,776 x 7  
## year month day hour tailnum carrier name   
## <int> <int> <int> <dbl> <chr> <chr> <chr>   
## 1 2013 1 1 5 N14228 UA United Air Lines Inc.   
## 2 2013 1 1 5 N24211 UA United Air Lines Inc.   
## 3 2013 1 1 5 N619AA AA American Airlines Inc.   
## 4 2013 1 1 5 N804JB B6 JetBlue Airways   
## 5 2013 1 1 6 N668DN DL Delta Air Lines Inc.   
## 6 2013 1 1 5 N39463 UA United Air Lines Inc.   
## 7 2013 1 1 6 N516JB B6 JetBlue Airways   
## 8 2013 1 1 6 N829AS EV ExpressJet Airlines Inc.  
## 9 2013 1 1 6 N593JB B6 JetBlue Airways   
## 10 2013 1 1 6 N3ALAA AA American Airlines Inc.   
## # ... with 336,766 more rows

Simple examples.

x <- tribble(  
 ~key, ~val\_x,  
 1, "x1",  
 2, "x2",  
 3, "x3"  
)  
x

## # A tibble: 3 x 2  
## key val\_x  
## <dbl> <chr>  
## 1 1 x1   
## 2 2 x2   
## 3 3 x3

y <- tribble(  
 ~key, ~val\_y,  
 1, "y1",  
 2, "y2",  
 4, "y3"  
)  
y

## # A tibble: 3 x 2  
## key val\_y  
## <dbl> <chr>  
## 1 1 y1   
## 2 2 y2   
## 3 4 y3

x %>%   
 inner\_join(y, by = "key")

## # A tibble: 2 x 3  
## key val\_x val\_y  
## <dbl> <chr> <chr>  
## 1 1 x1 y1   
## 2 2 x2 y2

Duplicate keys.

x <- tribble(  
 ~key, ~val\_x,  
 1, "x1",  
 2, "x2",  
 2, "x3",  
 1, "x4"  
)  
x

## # A tibble: 4 x 2  
## key val\_x  
## <dbl> <chr>  
## 1 1 x1   
## 2 2 x2   
## 3 2 x3   
## 4 1 x4

y <- tribble(  
 ~key, ~val\_y,  
 1, "y1",  
 2, "y2"  
)  
y

## # A tibble: 2 x 2  
## key val\_y  
## <dbl> <chr>  
## 1 1 y1   
## 2 2 y2

left\_join(x, y, by = "key")

## # A tibble: 4 x 3  
## key val\_x val\_y  
## <dbl> <chr> <chr>  
## 1 1 x1 y1   
## 2 2 x2 y2   
## 3 2 x3 y2   
## 4 1 x4 y1

Both with duplicate keys.

x <- tribble(  
 ~key, ~val\_x,  
 1, "x1",  
 2, "x2",  
 2, "x3",  
 3, "x4"  
)  
x

## # A tibble: 4 x 2  
## key val\_x  
## <dbl> <chr>  
## 1 1 x1   
## 2 2 x2   
## 3 2 x3   
## 4 3 x4

y <- tribble(  
 ~key, ~val\_y,  
 1, "y1",  
 2, "y2",  
 2, "y3",  
 3, "y4"  
)  
y

## # A tibble: 4 x 2  
## key val\_y  
## <dbl> <chr>  
## 1 1 y1   
## 2 2 y2   
## 3 2 y3   
## 4 3 y4

left\_join(x, y, by = "key")

## # A tibble: 6 x 3  
## key val\_x val\_y  
## <dbl> <chr> <chr>  
## 1 1 x1 y1   
## 2 2 x2 y2   
## 3 2 x2 y3   
## 4 2 x3 y2   
## 5 2 x3 y3   
## 6 3 x4 y4