Applied Data Science with R: Working with Relational Data

Matthias Haber

06 March 2019

Last weeks Homework

```r
library(tidyverse)
library(nycflights13)
```

Carrier with longest average dep delay accounting for distance

```r
flights %>%
  group_by(carrier) %>%
  mutate(delay = dep_delay / distance) %>%
  summarise(delay = mean(delay, na.rm =T)) %>%
  arrange(desc(delay)) %>%
  head(n=1)
```

```
## # A tibble: 1 x 2
##   carrier  delay
##   <chr>    <dbl>
```

Prerequisites

Packages

```r
library(tidyverse)
library(dbplyr) # install.packages("dbplyr")
library(DBI)
library(RSQLite) # install.packages("RSQLite")
library(nycflights13)
library(readr)
```

Data

```r
url <- paste0("https://raw.githubusercontent.com/mhaber/",
              "AppliedDataScience/",
              "master/slides/week5/data/")
films <- read_csv(paste0(url, "films.csv"))
people <- read_csv(paste0(url, "people.csv"),
```

Relational data

Relational data

Data analysis rarely involves only a single table, but many tables of data, and you must combine them to answer the questions that you're interested in. Collectively, multiple tables of data are called relational data because it is the relations that are important.

The most common place to find relational data is in a relational database management system (or RDBMS).

Databases

What is a database:

> *A collection of information organized to afford efficient retrievel.*
>
> *"When people use the word database, fundamentally what they are saying is that the data should be self-describing and it should have a schema." (Jim Gray)*

SQL

SQL

Structured Query Language is a language for interacting with databases. SQL is over 40 years old, and is used by pretty much every database in existence.

- A query is a request for data from a database table (or combination of tables)
- SQL can be used to query but also to create and modify databases.

SELECT

In SQL, you can select data from a table using a SELECT statement. For example, the following query selects the name column FROM the people table:

```
SELECT name
```

SQL in R

SQL vs R

SQL is not designed to do data analysis. For example, calculate the median arrival delay per carrier.

Using dplyr:

```
flights %>%
  dplyr::group_by(carrier) %>%
  dplyr::summarize(delay = median(arr_delay, na. rm =TRUE))
```

SQL vs R

PostgreSQL:

```
WITH ordered_flights AS (
 SELECT arr_delay,
        row_number() OVER (order by id) AS row_id,
```

Group Exercise

Group exercise

In today's exercise you'll be working with a database containing information on almost 5000 films. Get together in groups of two or three and

1. set up our own (SQLite) database
2. write the four objects `films`, `people`, `review`, `role` as tables into your database and use `dbListTables()` to make sure you did it correctly

Group exercise

3. complete the following tasks using either SQL or `dplyr`'s language:
   - Get the title, release year and country for every film
   - Get all the different type of film roles

Homework Exercises

Homework Exercises

For this week's homework exersises go to Moodle and answer the Quiz posted in the Relational Data section.

Deadline: Tuesday, March 12.

That's it for today. Questions?