Databases and dbplyr

David Gerard

2019-02-20

Learning Objectives

- Using dplyr-like syntax for databases.
- Introduction to dbplyr

dbplyr

- SQL is a language used to query from relational datasets.
- dplyr basically implements the most common actions in SQL (but SQL can do more).
- We'll use a soccer dataset to demonstrate how to use dplyr (instead of SQL) syntax when interacting with a database. Download and unzip the soccer database from https://dcgerard.github.io/stat_412_612/data.html.
- We'll use the dbplyr package to interact with databases.

```
install.packages("dbplyr")
```

```
library(tidyverse)
library(dbplyr)
```

- dbplyr allows you to work with databases as if you are using dplyr.
- You'll also need to install the RSQLite package. There are different ways to create/access/update/delete data from relational databases, and RSQLite provides an R interface for one of these ways.

```
install.packages("RSQLite")
```

```
library(RSQLite)
```

- If your database uses a different engine, you'll need to download other packages to interact with it (see Introduction to dbplyr)
- First, we'll tell R where the database is using `dbConnect()`, (you might need to change the path).

```
con <- dbConnect(drv = SQLite(), dbname = "../../data/soccer/soccer.sqlite")
```

- Now we'll list the data frames available in the connection we just created.

```
dbListTables(con)
```

```
## [1] "Country"            "League"             "Match"
## [4] "Player"             "Player_Attributes"  "Team"
## [7] "Team_Attributes"    "sqlite_sequence"
```

- Use `tbl()` to make a reference to the tables in `con`.

```
Team_db    <- tbl(con, "Team")
Team_at_db <- tbl(con, "Team_Attributes")
Country_db <- tbl(con, "Country")
```

```r
League_db  <- tbl(con, "League")
Match_db   <- tbl(con, "Match")
```

- We can now interact with all of these data frames mostly like if they were in memory (with some limitations).

```r
head(Country_db)
```

```
## # Source:   lazy query [?? x 2]
## # Database: sqlite 3.22.0
## #   [/home/david/Dropbox/teaching/stat_412_612/data/soccer/soccer.sqlite]
##      id name
##   <int> <chr>
## 1     1 Belgium
## 2  1729 England
## 3  4769 France
## 4  7809 Germany
## 5 10257 Italy
## 6 13274 Netherlands
```

```r
head(Match_db)
```

```
## # Source:   lazy query [?? x 115]
## # Database: sqlite 3.22.0
## #   [/home/david/Dropbox/teaching/stat_412_612/data/soccer/soccer.sqlite]
##      id country_id league_id season stage date  match_api_id
##   <int>      <int>     <int> <chr>  <int> <chr>        <int>
## 1     1          1         1 2008/~     1 2008~       492473
## 2     2          1         1 2008/~     1 2008~       492474
## 3     3          1         1 2008/~     1 2008~       492475
## 4     4          1         1 2008/~     1 2008~       492476
## 5     5          1         1 2008/~     1 2008~       492477
## 6     6          1         1 2008/~     1 2008~       492478
## # ... with 108 more variables: home_team_api_id <int>,
## #   away_team_api_id <int>, home_team_goal <int>, away_team_goal <int>,
## #   home_player_X1 <int>, home_player_X2 <int>, home_player_X3 <int>,
## #   home_player_X4 <int>, home_player_X5 <int>, home_player_X6 <int>,
## #   home_player_X7 <int>, home_player_X8 <int>, home_player_X9 <int>,
## #   home_player_X10 <int>, home_player_X11 <int>, away_player_X1 <int>,
## #   away_player_X2 <int>, away_player_X3 <int>, away_player_X4 <int>,
## #   away_player_X5 <int>, away_player_X6 <int>, away_player_X7 <int>,
## #   away_player_X8 <int>, away_player_X9 <int>, away_player_X10 <int>,
## #   away_player_X11 <int>, home_player_Y1 <int>, home_player_Y2 <int>,
## #   home_player_Y3 <int>, home_player_Y4 <int>, home_player_Y5 <int>,
## #   home_player_Y6 <int>, home_player_Y7 <int>, home_player_Y8 <int>,
## #   home_player_Y9 <int>, home_player_Y10 <int>, home_player_Y11 <int>,
## #   away_player_Y1 <int>, away_player_Y2 <int>, away_player_Y3 <int>,
## #   away_player_Y4 <int>, away_player_Y5 <int>, away_player_Y6 <int>,
## #   away_player_Y7 <int>, away_player_Y8 <int>, away_player_Y9 <int>,
## #   away_player_Y10 <int>, away_player_Y11 <int>, home_player_1 <int>,
## #   home_player_2 <int>, home_player_3 <int>, home_player_4 <int>,
## #   home_player_5 <int>, home_player_6 <int>, home_player_7 <int>,
## #   home_player_8 <int>, home_player_9 <int>, home_player_10 <int>,
## #   home_player_11 <int>, away_player_1 <int>, away_player_2 <int>,
## #   away_player_3 <int>, away_player_4 <int>, away_player_5 <int>,
## #   away_player_6 <int>, away_player_7 <int>, away_player_8 <int>,
## #   away_player_9 <int>, away_player_10 <int>, away_player_11 <int>,
## #   goal <chr>, shoton <chr>, shotoff <chr>, foulcommit <chr>, card <chr>,
## #   cross <chr>, corner <chr>, possession <chr>, B365H <dbl>, B365D <dbl>,
## #   B365A <dbl>, BWH <dbl>, BWD <dbl>, BWA <dbl>, IWH <dbl>, IWD <dbl>,
## #   IWA <dbl>, LBH <dbl>, LBD <dbl>, LBA <dbl>, PSH <dbl>, PSD <dbl>,
## #   PSA <dbl>, WHH <dbl>, WHD <dbl>, WHA <dbl>, SJH <dbl>, SJD <dbl>,
## #   SJA <dbl>, VCH <dbl>, ...
```

```r
Match_db %>%
  select(id:away_team_goal)
```

```
## # Source:   lazy query [?? x 11]
## # Database: sqlite 3.22.0
## #   [/home/david/Dropbox/teaching/stat_412_612/data/soccer/soccer.sqlite]
##       id country_id league_id season stage date  match_api_id
##    <int>      <int>     <int> <chr>  <int> <chr>        <int>
##  1     1          1         1 2008/~     1 2008~       492473
##  2     2          1         1 2008/~     1 2008~       492474
##  3     3          1         1 2008/~     1 2008~       492475
##  4     4          1         1 2008/~     1 2008~       492476
##  5     5          1         1 2008/~     1 2008~       492477
##  6     6          1         1 2008/~     1 2008~       492478
##  7     7          1         1 2008/~     1 2008~       492479
##  8     8          1         1 2008/~     1 2008~       492480
##  9     9          1         1 2008/~     1 2008~       492481
## 10    10          1         1 2008/~    10 2008~       492564
## # ... with more rows, and 4 more variables: home_team_api_id <int>,
## #   away_team_api_id <int>, home_team_goal <int>, away_team_goal <int>
```

```r
names(Match_db) ## won't work
```

```
## [1] "src" "ops"
```

- Once you select the variables you want and the observations you want, you should use `collect()` to get the data frame into memory so that you can have all of the functionality of R (e.g., `gather()` and `spread()` will only work on in-memory data frames).

```r
Match_db %>%
  select(id:away_team_goal) %>%
  collect() ->
  Match
Team_db %>%
  collect() ->
  Team
Country_db %>%
  collect() ->
  Country
```

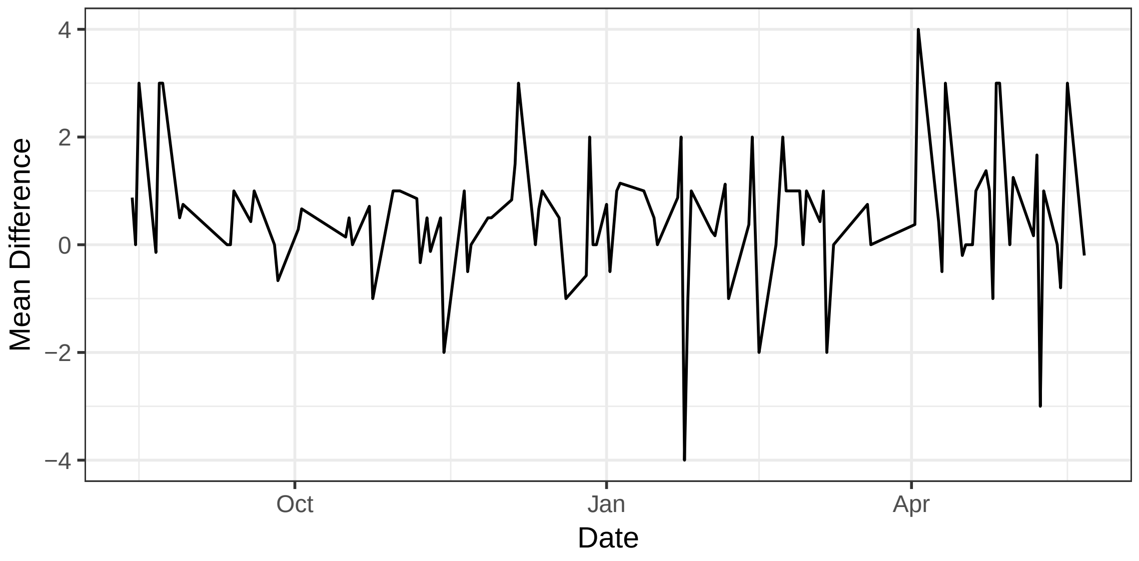
- The following will return a data frame telling you where each team is from.

```r
Match %>%
  select(country_id, home_team_api_id, away_team_api_id) %>%
  gather(-country_id, key = "home_away", value = "team_api_id") %>%
  select(-home_away) %>%
  distinct() %>%
  left_join(Team, by = "team_api_id") %>%
  left_join(Country, by = c("country_id" = "id")) %>%
  select(team_long_name, team_short_name, name) %>%
  rename(country_name = name)
```

```
## # A tibble: 299 x 3
##    team_long_name    team_short_name country_name
##    <chr>             <chr>           <chr>
##  1 KRC Genk          GEN             Belgium
##  2 SV Zulte-Waregem  ZUL             Belgium
##  3 KSV Cercle Brugge CEB             Belgium
##  4 KAA Gent          GEN             Belgium
##  5 FCV Dender EH     DEN             Belgium
##  6 KV Mechelen       MEC             Belgium
##  7 KSV Roeselare     ROS             Belgium
##  8 Tubize            TUB             Belgium
##  9 KVC Westerlo      WES             Belgium
## 10 Club Brugge KV    CLB             Belgium
## # ... with 289 more rows
```

- **Exercise**: Extract all matches from the `England Premier League` and calculate the mean team difference (average of home team goals minus away team goals) each day in the `"2010/2011"` season. Plot this proportion against time. (hint: you'll need separate date and time. You'll also need to use before you plot `parse_date()`).

  Your plot should look like this:


Mean Difference
Oct
Jan
Apr
Date