Lab/HW 5: Data Transformation, SQL

Your lab/homework must be submitted in with two files: (1) R Markdown format file; (2) a pdf or html file. Other formats will not be accepted. Your responses must be supported by both textual explanations and the code you generate to produce your result.

For this homework, we will look trends in baseball team payrolls between the years 1985 and 2010. The data come from the Baseball Databank and is based in part on Lahman's Baseball Database. You will need to download the SQLite database file baseball.db from our course webpage to your computer.

Part I - `dplyr`

The following exercises use the `flights` data set from the `nycflights13` package, you are asked to use `dplyr` functions and verbs. In Both questions you should use the pipe (`%>%`) operator to compute this in one chain.

[20 pt] 1. Which destinations are served by at least three airlines? The final answer should include only the destination and the number of carriers. It should be sorted first in a descending order of the number of carriers, and the in ascending order of destination names.

[20 pt] 2. Which carrier has the highest number of delayed departures and which the lowest (and what are the corresponding number of delays)? The final output should only contain the airline symbol and number of delays.

Part II - SQL

[20 pt] 1. Here we will import payroll data from the database.

a. Using `DBI` and `RSQLite`, setup a connection to the `SQLite` database stored in `baseball.db`. Use `dbListTables()` to list the tables in the database.
b. Use the table that contains salaries and compute the payroll for each team in 2010. Use `dbReadTable()` to grab the entirety of the table, then manipulate using `dplyr` verbs. Which teams had the highest payrolls (that is, sum of all paid salaries)?
c. Repeat the previous step, but now do this using only `dbGetQuery()` and SQL. Are your answers `identical()`? Why or why not? Are their values `all_equal()`?
d. Repeat again Step b., using this time the `dbplyr` functions. After you show the results, show the SQL query that was used behind the scenes.
e. Modify the SQL statement to compute the payroll for each team for each year from 1985 to 2010.
f. Do the same with dbplyr.

[20 pt] 2. Write a function that accepts three inputs: minimal total salary, first year and last year. The function returns the number of players whose total salary between the first year and last year (inclusive) exceeded the minimal total salary.

The function should execute a SQL query directly.

The output for (100000000, 1995, 2005) is 15. The output for (200000000, 1995, 2010) is 3.

[20 pt] 3. Write a function that takes as input a year, a name of a team, and minimal AB value, and returns all playerIDs in the input team that had at least the minimal AB value in the input year.

The function should execute a SQL query directly.

The output for (2010, “CHA”, 300) is: konerpa01
kotsama01
vizquom01
pierzaj01
pierrju01
riosal01
quentca01
ramiral03
beckhgo01