Prep9

Sebastian

Due by midnight, Monday April 25

Reminder: Prep assignments are to be completed individually. Upload a final copy of the .Rmd and renamed .pdf to your private repo, and submit the renamed pdf to Gradescope.

Reading

It's our FINAL prep!

The associated reading for the week is Chapter 15 on SQL.

Practice 9 will contain questions about SQL and last week's content on iteration and simulation (Chapters 7 and 13).

1 - Chapter Basics

> part a - In your own words, explain what a relational database is, and why using one may be better than using a `flat file`.

Solution: Flat files are one dimensional, consisting of just a single sheet with columns and rows. The problem with these files is that computers are limited in their RAM and so cannot work with really big versions of these files at once. Relational databases contain all their data stored separately on a disk but can have different parts of their data accessed at once, meaning we can utilize very large databases without needing the entire portion of those files to take up RAM.

> part b - What R package have we been using all semester that was structured to be similar to SQL?

Hint: We have usually not loaded this package directly, but it has been loaded when we load `tidyverse`.

Solution: The DBI package.

> part c - What two arguments are required for a SQL `select` query to run?

Hint: Many arguments can be provided in a `select` query. This is asking about the required two that a `select` query will not run without.

Solution: SELECT and FROM.

> part d - Comparing R and SQL, based on the arguments in the reading, which is better for data analysis? Which is better for data management?

Solution: R appears to be better for data analysis, allowing for a huge amount of functions to use on small pieces of data. SQL is better for data management, as it allows you to access large amount of data efficiently.

2 - Airline Flights in SQL

Learning SQL requires having a SQL server set up to access. Run the code below to get access to a server with the airline flights data. Then, use the provided code below to get a sense of the data and address a few questions.

```
# SQL commands
con <- dbConnect_scidb("airlines")
```

part a - How many tables are present?

```
query1 <- "SHOW TABLES"

dbGetQuery(con, query1)
```

```
##   Tables_in_airlines
## 1           airports
## 2           carriers
## 3            flights
## 4             planes
```

Solution: There are 4 tables.

part b - What variables are present in the flights data? List some that may be of interest to you to explore.

```
query2 <- "DESCRIBE flights"

dbGetQuery(con, query2)
```

```
##             Field        Type Null Key Default Extra
## 1            year smallint(4)  YES MUL    <NA>
## 2           month smallint(2)  YES        <NA>
## 3             day smallint(2)  YES        <NA>
## 4        dep_time smallint(4)  YES        <NA>
## 5  sched_dep_time smallint(4)  YES        <NA>
## 6       dep_delay smallint(4)  YES        <NA>
## 7        arr_time smallint(4)  YES        <NA>
## 8  sched_arr_time smallint(4)  YES        <NA>
## 9       arr_delay smallint(4)  YES        <NA>
## 10        carrier  varchar(2)   NO MUL
## 11        tailnum  varchar(6)  YES MUL    <NA>
## 12         flight smallint(4)  YES        <NA>
## 13         origin  varchar(3)   NO MUL
## 14           dest  varchar(3)   NO MUL
## 15       air_time smallint(4)  YES        <NA>
## 16       distance smallint(4)  YES        <NA>
## 17      cancelled  tinyint(1)  YES        <NA>
## 18       diverted  tinyint(1)  YES        <NA>
## 19           hour smallint(2)  YES        <NA>
## 20         minute smallint(2)  YES        <NA>
## 21      time_hour    datetime  YES        <NA>
```

Solution: Some variables included are dest (flight destination) ,origin (flight origin), arr_delay (arrival delay), and day.

part c - How many flights went from Hartford (BDL) to Chicago (ORD) in 2012?

```
query3 <- "SELECT COUNT(*) as N
  FROM flights
  WHERE dest = 'ORD' AND year = 2012 AND origin = 'BDL'
"

dbGetQuery(con, query3)
```

```
##      N
## 1 2897
```

Solution: 2897 flights went from Hartford to Chicago in 2012.

part d - Your turn! How many flights went from Chicago to Hartford in 2009?

Solution: 2927 flights went from Chicago to Hartford in 2012.

```
query4 <- "SELECT COUNT(*) as N
  FROM flights
  WHERE dest = 'BDL' AND year = 2012 AND origin = 'ORD'
"

dbGetQuery(con, query4)
```

```
##      N
## 1 2927
```

part e - Use more date info. How many domestic flights flew into Portland, Oregon (PDX) on May 14, 2012?

Solution: 150 flights flew into Portland on May 14, 2012.

```
query5 <- "SELECT COUNT(*) as N
  FROM flights
  WHERE day = 14 AND year = 2012 AND dest = 'PDX' and month = 5
"

dbGetQuery(con, query5)
```

```
##     N
## 1 150
```

part f - Design your own query. You can continue pulling from flights or use another table. Explain what you wanted the query to show (i.e. what question is it helping to answer?) and then provide an answer.

Solution: I wanted to figure out how many flights going into Hartford in the year 2012 were delayed by more than 60 minutes. The query shows that there are 1126 of these flights.

```
query6 <- "SELECT COUNT(*) as N
  FROM flights
  WHERE dep_delay > 60 AND year = 2012 AND dest = 'BDL'
"

dbGetQuery(con, query6)
```

```
##      N
## 1 1126
```