Prework Assignment

David Dorr/Tracy Edinger/Ted Laderas

June 28, 2016

Purpose: To build and test your data warehouse from extracted files.

Background: In the analytics class, you will have a miniature data warehouse that will be accessed by R and SQLite. In order to do this, you will load a set of 7 flat files into a SQLite database, which will serve as your data warehouse.

Load the flat files `patient_encounter_hosp.txt` and `patient.txt` into a SQLite database using R; these files are both delimited by a pipe "|" and contain variable names in the header. Use a SQL query to return the number of readmissions within 30 days contained in the dataset; note that there is not a variable for readmissions, but the file is limited to initial hospitalizations and readmissions. Use an R query or SQL to summarize the readmission rate, either by count or percent or both.

Some hints: The `patient_encounter_hosp` file should have one to two observations per patient, each observation a specific hospitalization. The analyst has subset the file for you to assure there are only INDEX admissions and READMISSIONS. They do not look different, however, so you'll have to do some date differences across patientid to find readmissions.

Suggested Steps

If you have been following the week by week R/SQL assignments, you will already have done most of this.

1) Download the 7 files (zip file) if you haven't already.
2) Write code that loads all the flat files (all with header rows and with a pipe (`|`) delimiter).
3) Store all the data frames into a database in R and SQLite. Note that you'll need to install the RSQLite package using `install.packages()` and load it using the `library()` command.
4) Create a new table in your SQLite database by using querying SQL to identify just the index admissions (the first admission dates) and call this variable `indexadmit`. *Hint*: which table should you use for this? How can you use a self-join to identify the earliest dates for each patient?
5) Next, identify those admission dates that are readmissions within 30 days from the index admission. Name the readmission variable `Readmit30`. Note that the file only has the index admission and readmission within 30 days; it does so by limiting rows for individual patients to maximum 2 admissions (either a single admission or the admission and the readmission within 30 days).
6) Report the Count of unique patients and the number of readmissions in the file.

Deliverables:

1) R Code that loads the data warehouse; and
2) R Code that encodes index admissions with whether a readmission follows (and discarding readmissions).
3) Data Dictionary for each table (excel file is fine).

Please post on the Week 6 forums if you are having any issues, or email me (laderast@ohsu.edu) if you are stuck.