sql, databases

adam okulicz-kozaryn

adam.okulicz.kozaryn@gmail.com

this version: Thursday 28th April, 2016 13:34

outline

misc

why sql?

outline

misc

why sql?

yet another reason to use python

- stata is super clunky with databases
- python works much better here
- in fact, python works much better with any other software
- than stata; or for that matter any other software
- Python is a glue language–it is made for working with others!

outline

misc

why sql?

data management

- this class is called “data management”
- but really it is “data management” for soc sci
- they also have such a class at business schools
  - especially in MIS program
  - and that class is quite different
- in corporate world (in real world outside of academia)
  - “data management” = databases (+ often SQL)
- so it would be highly inadequate if in data management
  - class we don’t talk about databases at all!

databases

- ◇ databases usually mean relational databases
- ◇ it is a database with a bunch of tables/spreadsheets
- ◇ and relations between these tables are defined
- · so called layout
- ◇ layout is just definitions of merge variables in tables
- · draw picture; say amzn products, customers, wish list)
- · that is how tables can be merged together
- ◇ databases handle TB, even PB of data
- ◇ you can't have TBs in Stata, Python, etc
- ◇ big dataset ($>$ 100TB), you have to have some database

popular databases

- Oracle: developed/complex
- MSSQL: a Microsoft product
- PostgreSQL: open source
- MySQL: open source–we'll use this one
- NoSQL, Hadoop, etc etc

SQL

◇ SQL: Structured Query Language

$$$

- knowledge of databases and SQL can get you a job
- a job paying $50k-over $100k
- and it is not complicated at all!
- knowing Stata well as you do, SQL is easy
- and SQL is more popular and marketable than Stata

what do you do with databases/SQL

- build websites
- Facebook runs on MySQL https://www.percona.com/blog/2014/03/27/a-conversation-with-5-facebook-mysql-gurus
- any modern website runs on a database
- when you browse stuff on Amazon it produces SQL query that plucks stuff out from the database
  when you buy stuff it adds items to your table that relates to products table
- again, big data (TBs)–you have to have a database
- sometimes you can get data only from database

what's in there

- ◇ there are databases
- ◇ and in databases there are tables (spreadsheets)
- ◇ and the are 'related'. i.e. you can merge them
- ◇ e.g. Amazon's list of products, and Amazon users's list of purchases
  - · if you kept everything in one table it would be huge with almost all missings...
- ◇ run code