Data Analysis using R and SQL

Irina Nikiforova, inikiforova@hse.ru

Course Objectives

Open source tools such as MySQL database systems and *R* language for statistical computing became industry standards and the core of many scientific projects, both in academia and industry. These solutions are platform independent and work on Windows, Linux and Mac OS. The course is designed for researchers familiar with R and engaged in data analysis in various disciplines but who lack the time to explore MySQL databases, integrate with *R* and troubleshoot issues. The course provides the fundamentals of managing relational databases such as MySQL, running SQL queries and performing explorative data analyses/data mining in *R*. It also discusses the portability solutions of working in Excel, SPSS or MySQL and running *R* statistical tools.

Course Structure

Seminar 1. Introduction to databases and database management systems; installation and management of MySQL and SQLite

Seminar 2. Database navigation; SQL language and queries

Seminar 3. Advanced SQL queries

Seminar 4. Data manipulation and portability

Seminar 5. Data Analysis in R: Exploratory data analysis and introduction to data mining

Prerequisites

- Familiarity with *R* and running statistical analyses in *R*
- Experience working with databases

- Personal laptop with installed *R* (RGui/Rstudio), Excel, SPSS (recommended).

Homework

Participants are expected to work independently and will be given a set of exercises to try at home.

References

Bessant, Conrad, Shadforth, Ian, & Oakley, Darren. (2009). Building Bioinformatics Solutions: with Perl, R and MySQL. Oxford, UK: Oxford University Press.

Spector, Phil. (2008). Data Manipulation with R (Use R!). New York: Springer.

Torgo, Luis. (2010). Data Mining with R: Learning with Case Studies. Chapman & Hall/CRC.

Welling, Luke, & Thomson, Laura. (2003). MySQL Tutorial. Indianapolis, Indiana, USA: MySQL Press.

Course Outline

Seminar 1. Introduction to databases and database management systems; installation and management of MySQL and SQLite

Database systems. Data models and classification. Structure of Relational Databases. Entity-Relationship(E-R) model and diagrams. Database design. Relations among tables and fields. Keys. Design of E-R database schema. MySQL and SQLite installation and setup. Connections. User privilleges.

Seminar 2. Database navigation; SQL language and queries

Using the Query editor. Interface phpMyAdmin. Creating a database and its tables. Pupulating the database. Removing data and tables from the database. Creating database views and exploring INFORMATION_SCHEMA. Navigation commands. Views. Introduction to SQL (Structured Query Language). Data Types. Querying the database.

Composing SQL statements. Basic structure: SELECT, FROM, WHERE. Insertion, updates, and deletion of data.

Seminar 3. Advanced SQL queries

Advanced statements using IN, BETWEEN, LIKE, HAVING, GROUP BY, ANY, ALL, SOME, EXISTS, UNION, ORDER BY, and regular expressions. String operations. Ordering. Set operations: UNION, INTERSECT and EXCEPT operations. Table aggregation.

Seminar 4. Data manipulation and portability

Input and outputs methods. Redirecting output. Portability to/from other applications: Excel, SPSS, R, MySQL, SQLite. ODBC drivers, RODBC, SQLDF и DBI пакеты. Using SQL in R.

Seminar 5. Data Analysis in R: Exploratory data analysis and introduction to data mining

Using SQL commands in R. Exploratory data analysis. Introduction to «data mining». Analysis of data structures: visualization, principle component, and classification (cluster) analysis.