R for Software Developers and Data Analysts

When: June 28, 2014
Where: Microsoft NERD, Cambridge, M
Cost: $179 through May 20
$239 May 21 - June 3
$309 June 4 - June 24
$339 after June 24

Details and registration: www.gbcacm.org

Speakers

Dr. Kabacoff is a seasoned researcher, with 30 years of experience in data analysis and data visualization.

As Vice President of Research for Management Research Group (1997-present), he consults widely with academic, government, and corporate organizations throughout North America, Western Europe, and the Pacific Rim.

As a Professor in the Center for Psychological Studies at Nova Southeastern University (1987-1997), he taught numerous graduate courses on multivariate statistics, statistical consulting, and research computing.

Dr. Kabacoff created and maintains the popular tutorial website Quick-R. The second edition of his popular book R in Action: Data Analysis and Graphics with R, is due out this year.

R and Big Data Analytics

If you are looking at this workshop, you probably have some data that you need to collect, summarize, transform, explore, model, visualize, or present. If so, then R is for you! R has become the world-wide language for statistice, predictive analytics, and data visualization. It offers the widest range of methodologies for understanding data, from the most basic to the most complex and bleeding edge.

One of the hottest topics today is Big Data. Much of the publicity around Big Data focuses on interactive query operations, but the greatest value comes from Big Data Analytics – statistical analysis and visualization of the data. The R language is widely used for Big Data Analytics, and has become one of the most popular languages for data analysis and visualization in general. Like many popular Big Data tools, R is free software – it is available at no charge under an open source license. This makes R a very attractive tool to learn and use.

SEMINAR TOPICS

This workshop provides a practical introduction to R. You will learn to import data into R from a variety of sources; clean, recode, and restructure data; and apply R's many functions for summarizing, modeling, and graphing data. Both basic and more advanced forms of data analysis and graphics will be covered. Additional topics include navigating R's comprehensive help systems, practical advice for processing data, common programming mistakes to avoid, and useful functions for data mining.

I. Introduction – An introduction to R: R syntax and data structures; working interactively and in batch; alternative IDEs and GUIs; adding functionality through packages; common programming mistakes; getting unstuck – were to find answers to your questions.

II. Data Management – Importing, cleaning, and reformatting data: transforming and recoding variables; subsetting, merging, and aggregating data; control structures; user-written functions.

III. Graphics – Taking advantage of R's powerful graphics: creating basic and advanced graphs; customizing and combining graphs; innovative methods for visualizing complex data.

IV. Statistical Analysis and Data Mining – Using R for description, prediction, and classification: descriptive statistics and multi-way tables; ANOVA variants; regression (e.g., linear, logistic, poisson), classification trees, cluster analysis, and other multivariate methods; dealing effectively with missing data; going further.

Why you should attend this seminar

Robert Kabacoff notes: *"I think that there are two reasons why R can be challenging to learn quickly."*

First, while there are many introductory tutorials, none alone are comprehensive. In part, this is because much of the advanced functionality of R comes from hundreds of user contributed packages. Hunting for what you want can be time consuming, and it can be hard to get a clear overview of what procedures are available.

The second reason is more ephemeral. As users of statistical packages, we tend to run one prescribed procedure for each type of analysis. We carefully set up the run with all the parameters and options that we need. When we run the procedure, the resulting output may be a hundred pages long. We then sift through this output pulling out what we need and discarding the rest.

The paradigm in R is different. Rather than setting up a complete analysis at once, the process is highly interactive. You run a command, take the results and process it through another command (say a set of diagnostic plots), take those results and process it through another command (say cross-validation), etc. The cycle may include transforming the data, and looping back through the whole process again. You stop when you have fully analyzed the data.

Details and registration: www.gbcacm.org