Tools for Analytics Lab - R-track

Name	Gergely Daroczi
Department	Department of Economics, CEU Business School
Semester	Winter, 2016
Course level	MA
Credits	2 credits

Course Description

This is an introductory course on how to use the R programming language and software environment for data manipulations, exploratory data analysis, statistical modeling, machine learning, data visualizations, automated reporting and self-service dashboards.

Course Requirements

Assessment is through in-class exam (50%) and take-home assignments (50%).

Outline

Jan 28, 13:00-14:30 pm (60 min): Introduction to R and General Programming

- Downloading and installing R
- History of R, general ecosystem, R packages, CRAN
- R community, R-bloggers, StackOverflow, Coursera, DataCamp
- R User Groups & meetups
- RStudio
- git, GitHub
- R Syntax: Variables, functions, descriptive statistics
- Loading data from text files

- Univariate statistics
- Statistics with two variables: cross-tables, correlation, ANOVA
- Simpson's paradox

Jan 28, 15:00-16:00 pm (60 min): First Steps with Data Visualization

- Plots outside of Excel: `dotchart` and `violinplot` examples
- The Grammar of Graphics in R with `ggplot2`
- Using interactive JavaScript libraries with `htmlwidgets`

Jan 29, 9:30-12:30 (150 min + 30 min break): Data Preparation

- Loading data from relational databases – using `sqlite`
- Filtering and summarizing data with `plyr`, `dplyr` and `data.table`
- Merging and left/right/inner/outer/anti-joining data
- Wide and long tables with `reshape2` and `tidyr`
- Regular expressions
- Dealing with date and time
- Other dirty data problems (missing values, data imputation, outliers)

Jan 29, 13:00-16:00 (150 min + 30 min break): An R Introduction to Modeling

- Correlation, causality
- Linear regression
- Goodness of fit
- Polynomial regression
- Overfitting
- Confounders
- Poisson/negative binomial regression

Feb 4, 9:30-11:00 (90 min): Supervised and Unsupervised ML Methods for Qualitative Data

- Hierarchical and k-means clustering
- Training and test dataset, cross validation, confusion matrix, accuracy, ROC curve
- Logistic regression
- k-Nearest Neighbors algorithm
- Decision trees with `rpart`, `caret` and `C50`

Feb 4, 11:30-12:30 (60 min): Dimension Reduction

- Principal Component Analysis
- Discriminant Analysis
- Multidimensional Scaling
- Handling geospatial data in R

Feb 4, 13:00-16:00 (150 min + 30 min break): Supervised ML Methods for Quantitative Data

- Decision trees
- Random forest with `randomForest` and `H2O`
- Gradient boosting
- Consulting on exam exercises

Feb 5, 9:30-11:00 (90 min): Exam

Develop R code to load, analyze, model and visualize data.

Feb 5, 11:30-12:30 (60 min): Dynamic Reports and Reproducible Research

- Introduction to markdown
- First steps with `knitr`
- Markdown in R with `pander`
- Chunk options and document formats in `rmarkdown` and `kintr`

Feb 5, 13:00-16:00 (150 min + 30 min break): Interactive Data Analysis and Dashboards

- Introduction to `shiny` and `shinydashboard`
- `htmlwidgets` examples, including `networkd3d`, `leaflet`, `dygraphs`, `DT`, `treemap`
- Consulting on take-home assignments