Microsoft

Intelligent Cloud Conference

Copenhagen May 29-30

An R primer for SQL folks

Thomas Hütter

Event Sponsors

INSPARI

Microsoft

it-Craft SQL Specialisterne

Expo Sponsors

kapacity Business Analytics

solvo it Microsoft Partner

The Software Experts Crayon

CT Global

ADVECTAS Excellence in your decisions

Expo Light Sponsors

CLOUDEYN

SUPERUSERS

PROACTIVE

Intelligent Cloud Conference
København 2018

An R primer for SQL folks

Thomas Hütter

An R primer for SQL folks

Thomas Hütter, Diplom-Betriebswirt

- Application developer, consultant, accidental DBA
- Worked at consultancies, ISVs, end user companies
- Speaker at SQL events around Europe
- SQL Server > 6.5, Dynamics Nav > 3.01, R > 3.1.2

@DerFredo https://twitter.com/DerFredo

de.linkedin.com/in/derfredo

www.xing.com/profile/Thomas_Huetter

Agenda

- History: what is R, how did R come to be,
  what does the R ecosystem look like today
- Introduction: R IDE, RStudio, basic data types / objects,
  packages, in-/output, data analysis, visualization
- Business case demo:
  - Extracting ‘sales’ data from a Nav DB on SQL Server
  - Basic analysis and visualization
  - Advanced visualization using the Shiny framework
- Example: data science going wrong, round-up, resources
- This is an introductory walk-through, no deep dive -
  so no fancy predictions, regression, big data science :-(

History: R - then and now

- Programming language for statistical computing, analysis and visualization, widely used by statisticians, data miners, analysts, data scientists
- Created by **R**oss Ihaka and **R**obert Gentleman, Uni Auckland, in 1993 as an open source implementation of the (1970s) S language
- GNU project, maintained by the R Foundation for Statistical Computing, compiled builds für Mac OS, Linux, Windows, supported by R Consortium
- Extensible through user-created packages, > 12.500 available on CRAN
- Commercial support, e.g. since 2007 by Revolution Analytics, acquired by Microsoft in 2015, now provide Microsoft R Open, R Server
- IDEs: R.App, RStudio, MS R Tools for Visual Studio
- Support for R now in SQL Server, Power BI, Azure ML, Data science VM

Introduction: data objects

- Data types
  - numeric, integer, complex
  - character
  - logical
  - factor
  - Posix types for date/time

  - NA = Not available

- Data structures
  - vector: 1 dim, 1 data type
  - matrix: 2 dim rect, 1 data type
  - list: collection of other objects
  - table: > 2 dimensions
  - data frame
    2 dim rect, cols = vectors

▶ `DemoBasics1`

Introduction: packages

- Extensions to the R base system, containing code, data, documentation. Key factor to the success of R; flexible, user contributable. -> CRAN
- `installed.packages()` lists all *installed* packages incl. versions, dependencies, license and other info
- `search()` lists currently *attached* packages
- `install.packages()` downloads and installs packages
- `library()` loads/attaches packages, also `require()`
- Hadley Wickham, chief scientist at RStudio, professor of statistics
  „Tidyverse“: dplyr, tidyr, lubridate, readr, httr, ggplot2
  + many more: hadley.nz

▶ `DemoBasics2`

Introduction: basic data in-/output

- Generic functions `read.table` and `write.table`
  - `read.csv` / `read.csv2` comma/semicolon delimited
  - `read.delim` / `read.delim2` Tab delimited, decimal point/comma
  - `read.fwf` fixed width format
- Some additional I/O packages
  - `reader` functions flexibly load multiple formats fast
  - `foreign` reads data from Minitab, S, SAS, SPSS, Stata, dBase…
  - `DBI/ODBC` database access via ODBC
  - `xlsx` and `readxl` read and write Excel 97/XP/200X files
  - `XML` reads XML and tables from http web sites

Introduction: basic data analysis + visualization

- Analyzing (numeric) data:
  `str()` structure = data types and ranges
  `summary()` Min, max, mean, median, quartiles;
  for factors: count of levels
  `head()`/`tail()` shows top/bottom n rows (default = 6)
- Distribution of values:
  `hist()` shows frequency distribution,
  `boxplot()` for min, max, quartiles, outliers,
  `mosaicplot()` contingency mosaic

▶ `DemoBasics3`

Continued… data analysis + visualization

- Libraries: `tidy` for data tidying/reshaping, `ggplot2` implements grammar of graphics, `raster` for geo data
- `apply()` family of functions applies functions to the margins of an array or a matrix
- `gather()/spread()` convert between wide/long format
- `ggplot()` very powerful plot function, plots point, line or bar geometrics etc with versatile parameters

▶ `DemoBasics4`

Business case demo

- We are the distributor for all German petrol stations,
  with two subsidiaries: NorthTank and SouthFuel
- Business calls „We need some analysis of our 2015 Diesel sales“,
  preferably some visualizations, and „maybe something is wrong…“
- Of interest: distribution by post code zones
- Source: Dynamics Nav ERP database, on the customer card (table „Customer“) there’s a field called „Sales (LCY)“ (= Local currency)
- Publicly available shape- and data files for post code zones

Extracting data & first analysis

- Using ODBC and the DBI package
  (also available: JDBC, RODBC and others)
- `dbConnect()` to establish a connection,
  then `dbGetQuery()` to query the database
- Calculate aggregates (sums) using `ddply()`
- Bar plot: `ggplot() + geom_bar()`
- Line diagram: `ggplot() + geom_line()`

Analysis & visualization

- Calculate intervals for sales sums: `cut()`
- libraries `raster`, `rgeos` for visualizing geospatial data
- shapefiles: open vector data format for GIS software, describes points, lines or polygons in these files: .shp shapes, .shx shape index, .dbf attributes, .prj projection
- merge shape and sales data: `merge()`
- plot maps, colouring post code zones according to sales

▶ `DemoTankData`

Use of Shiny framework

- Framework for interactive web applications in R
  apps consist of `server.R` and `ui.R` or just `app.R`
- `ui` defines screen appearance & controls
- `server` handles any data processing, plotting etc.
- apps can be run in web browser

- `DemoShiny/app`

Example: data science going wrong?

- Anscombe's quartet:
- 4 data sets, each with 11 completely different x-y pairs
- yet nearly identical statistical properties
  - Mean of x = 9
  - Mean of y = 7.5
  - Correlation between x and y = 0.816
  - Linear regression y = 3 + 0.5 x

Anscombe

Round-up / conclusions

- With R, a lot is possible in terms of analysis and visualization
- There's probably always a package for that

But please:

- Know your data
- Look at your data
- Think - does it make sense?
- Consider the influence of outliers
- Don't blindly rely on R 'doing the trick'

Resources online

- https://en.wikipedia.org/wiki/R_(programming_language)
- https://www.r-project.org/ -> Mirrors of CRAN = Comprehensive R Archive Network
- https://www.r-consortium.org/
- http://www.r-bloggers.com/
- www.kdnuggets.com
- www.rseek.org Pimped Google search for R-related subjects

- Twitter hashtag #rstats
- LinkedIn groups R Developers und Users Group, R Programming, The R Project for…

- www.swirlstats.com „Learn R, in R“
- www.coursera.org Data Science specialization (10 courses) MOOC
- www.edx.org

Resources offline

- Beginning R, The statistical programming language
  Dr. Mark Gardener, Wrox/Wiley, ISBN 978-1118164303
- R Cookbook, Paul Teetor, O'Reilly, ISBN 978-0596809157
- R Graphics Cookbook, Winston Chang, O'Reilly,
  ISBN 978-1449316952
- R in a Nutshell, Joseph Adler, O'Reilly, ISBN 978-1449312084
- Practical Data Science with R, Nina Zumel + John Mount,
  Manning publications, ISBN 978-1617291562

Credits

- Titanic data set: www.kaggle.com/c/titanic/data
- SQL Database structure:
  mbs.microsoft.com Dynamics Nav 2016 demo database
- Customer and „sales“ data: www.tankerkoenig.de (license CC BY 4.0)
- Shape files:
  - www.suche-postleitzahl.org (Open database license, © OpenStreetMap contributors)
  - Bundesamt für Kartographie und Geodäsie, Frankfurt am Main, 2011
- Some icons made by:
  http://www.flaticon.com/authors/hanan (license CC BY 3.0)
- Anscombe’s quartet: Francis J. Anscombe 1973

An R primer for SQL folks

Time for some Q & A:

That is: questions that might be of common interest, and their answers might fit into the remaining time :-)

An R primer for SQL folks

Thank you for your interest & keep in touch:

- @DerFredo https://twitter.com/DerFredo
- de.linkedin.com/in/derfredo
- www.xing.com/profile/Thomas_Huetter

Slides and scripts to this presentation will be at
https://github.com/SQLThomas/Conferences/tree/master/ICC2018