SQL Saturday 666
Holland 2017

Next first steps -
selected applications of R

Thomas Hütter

PASS SQLSATURDAY

Thanks to our sponsors!

avanade

DATADOGS

ilionx
Your partner in digital business

CLOUDDBA

GOVERNOR®

COZYROC™
Go to the next level

VALID
STAY AHEAD

PURESTORAGE

dbWatch
DATABASE CONTROL

Rubicon
Consulting & Technology

Visser ✦ Van Baars
BUSINESS INTELLIGENCE RECRUITMENT

VX
VX Company
IT Services

Microsoft

SentryOne™

SQL SUPPORT

Ai
Axis into ict

Next first steps - selected applications of R

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 - 2017, Nav 3.01 - 2017, R 3.1.2 - 3.4.1

@DerFredo https://twitter.com/DerFredo

de.linkedin.com/in/derfredo

www.xing.com/profile/Thomas_Huetter

Agenda

- Recap: the R ecosystem; a light-weight installation
- How to create dynamic T-SQL using R functions
- Visualizations in R based on shape files, choropleth techniques
- Applying Benford's law for analysis & fraud detection
- Round-up; resources; credits; Q&A

Recap: The R ecosystem

- Programming language for statistical computing 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, > 11.500 available at CRAN

- Commercial support, e.g. since 2007 by Revolution Analytics, acquired by Microsoft in 2015, now provide Microsoft R Open, R Server

- Support for R now in SQL Server, Power BI, Azure ML…
  IDEs: R.App, RStudio, MS R Tools for VisualStudio

Recap: A light-weight installation

- Follow www.swirlstats.com -> „Learn"
  This works equally well for Windows, Mac and Linux

- Necessary: Get and install the R base system

- Recommended: Download and install the RStudio IDE

- Useful: check for updates

- Optional: Also install Swirl

- Let the fun begin! 😉

- Remember:
  `install.packages("packagename")` to download a new package
  `library(packagename)` to start using it

Dynamic T-SQL: basics

The exact SQL statement is composed at runtime, because

- it depends on parameters or conditions
- which may be determined interactively or from data
- and can influence filtering, columns or even tables used

Different conditions result in different SQL statements

Pro & con: flexibility vs complexity / security (SQL injection)

▷ `Dynamic1`

- Packages used: DBI, RSQLServer

Dynamic T-SQL and R: applied

- Microsoft Dynamics Nav database (multi-company), meaning:
- (almost) all > 1000 tables exist once per company

- Goal: aggregate sales data over all companies
- number of companies may change over time

▷ **Dynamic2**

Packages used: DBI, RSQLServer

Visualizations in R based on shapefiles

What a shapefile is:

- open file format standard for handling geospatial vector data
- developed and maintained by GIS software vendor Esri
- actually „a shapefile" consists of 3 mandatory files:
  - ▸ `.shp` - shape format, the actual geo objects
  - ▸ `.shx` - shape index, to allow seeking forwards/backwards
  - ▸ `.dbf` - attributes/payload data for each shape (dBase IV format)
  - ▸ plus optional metadata files, projection, geocoding index…

Visualizations in R based on shapefiles

Where to get shapefiles:

• all over the internet :-) e.g. government or open data organizations

• or even „roll your own" using GIS software

What to do with them:

• visualize all kinds of data that are geo-related

• location of places, distribution of measures

▷ **Shape1**

 Packages used: raster, rgeos, foreign

Choropleth techniques

- A choropleth map is a thematic map in which areas are coloured/shaded/ patterned depending on some measure to be analyzed

- Measures can be populations, election results, sales figures…

- Auto-shading using RColorBrewer, for sequential (light to dark), diverging (around mid-range) or qualitative (max. difference) palettes

Shape2

Packages: GISTools(maptools, sp, RColorBrewer, rgeos), plyr, XML

Benford's law: basics

- Aka „Newcomb-Benford-Law" or „First-Digit-Law"
- Simon Newcomb 1881, credited to Frank Benford 1938
(a fact which follows Stigler's law, discovered by Merton ;-))
- It's an observation about the distribution of leading digits in naturally occurring collections of numerical data
- Intuition: digits are evenly distributed
- Observed: In logarithm tables, the earlier pages were more worn
- Conclusion: leading digits are more likely to be small

Benford's law: maths

- Evenly distributed digits:
  $P = 1/9 \approx 0.1111$

- First digit D_1 according to Benford:
  $P(D_1=d) = \log_{10}(d+1) - \log_{10}(d)$
  $= \log_{10}(1 + 1/d)$

Even more math on Benford:
en.wikipedia.org/wiki/Benford%27s_law

d	evenly	Benford
1	0.1111	0.3010
2	0.1111	0.1761
3	0.1111	0.1249
4	0.1111	0.0969
5	0.1111	0.0792
6	0.1111	0.0669
7	0.1111	0.0580
8	0.1111	0.0512
9	0.1111	0.0458

Applying Benford's law

- Determine the data / measure to examine
- extract first digits, regardless of magnitude
- calculate the table of relative density
- compare to Benford's table
- visualize

DemoBenford1+2

Packages: DBI, XML, ggplot2

Round-up

Dynamic SQL
- possible realization in R using apply functions
- know the blessings *and* the curse of your dynamic SQL

Shapefiles & choropleth
- showing data relating to geographic instances
- keep data order, have a balanced colour / shading scheme

Benford's law
- works for natural or transactional data, the bigger the better
- does not work for numbers influenced by human rules

Resources on- and offline

- www.swirlstats.com „Learn R, in R"

- www.r-project.org/ -> Mirrors of CRAN = Comprehensive R Archive Network

- www.sommarskog.se/dynamic_sql.html The Curse and Blessings of Dynamic SQL

- www.suche-postleitzahl.org/downloads Shapefiles post codes Germany

- www.geodatenzentrum.de Shapefiles federal states

- www.mygeoposition.com Geocoding

- R Cookbook, Paul Teetor, O'Reilly, ISBN 978-0596809157

- R Graphics Cookbook, Winston Chang, O'Reilly, ISBN 978-1449316952

- Datendesign mit R, Thomas Rahlf, Open Source Press, (German)
  ISBN 978-3955390945, ~~Out of press - hurry!~~ now: Springer-Verlag

Credits

Data:
- mbs.microsoft.com Cronus database
- statisticstimes.com/index.php Country data (UN, Worldbank, IMF)
- www.tankerkoenig.de Base for „sales" data (CC BY 4.0)

Shape files:
- www.suche-postleitzahl.org (Open database license, © OpenStreetMap)
- www.geodatenzentrum.de GeoBasis-DE / BKG 2016
- thematicmapping.org Bjørn Sandvik (CC Attribution-Share Alike)
- www.imergis.nl BRK Kadaster Nederland (CC BY)

Some icons made by:
- www.flaticon.com/authors/hanan (CC BY 3.0)

Next first steps – selected applications of R

Time for some Q & A:

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

Next first steps - selected applications of R

Thank you for your interest & keep in touch:

🐦 @DerFredo https://twitter.com/DerFredo

in. de.linkedin.com/in/derfredo

X www.xing.com/profile/Thomas_Huetter

This file and all demo scripts can be found at:
https://github.com/SQLThomas/Conferences/tree/master/SQLSat666