SQL Server 2017: Data Science with Python or R?

Dejan Sarka

Introduction

- Dejan Sarka (dsarka@solidq.com, dsarka@siol.net, @DejanSarka)
  - 30+ years of experience
  - SQL Server MVP, MCT,…
  - 17 books
  - 15+ courses
  - Focus:
    - Data modeling
    - Data mining
    - Data quality

Introduction

- SQL Server 2016 started supporting R
- SQL Server 2017 added support for Python
- In both cases, target applications were mainly data science projects
  - The current big thing
- The question: which one would you use?

Introducing R

- The R statistical programming language is a free open source package based on the S language developed by Bell Labs
- R written as a research project by Ross Ihaka and Robert Gentleman
  - Published in 1995
  - Now developed by a group of statisticians called 'the R core team', with a home page at www.r-project.org
- R is available free of charge and is distributed under the terms of the Free Software Foundation's GNU General Public License
  - Available for Windows, Mac OS X, and Linux

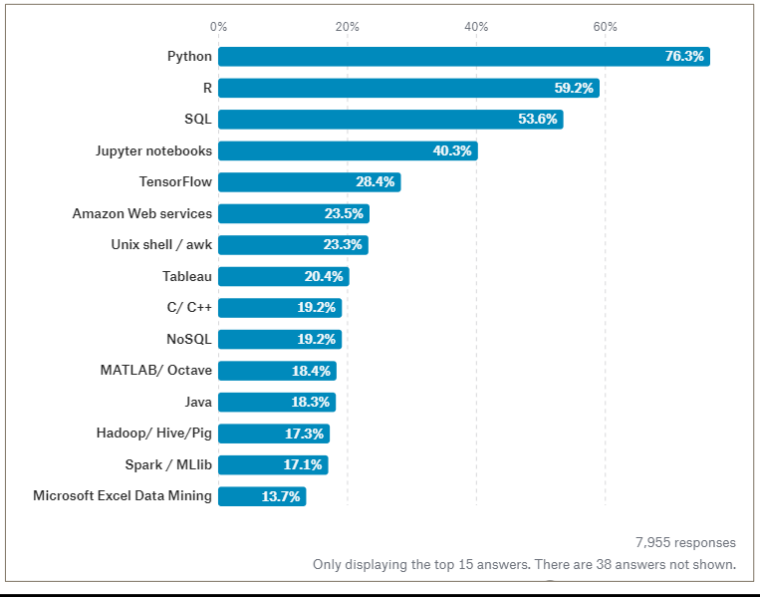
Introducing Python

- Python is one of the most popular programming languages
  - It is a general purpose high level language
  - Created by Guido van Rossum, released in 1991
  - It is an interpreted language, working on multiple platforms
- Python Software Foundation (PSF) takes care about Python advances
  - Manages the open source licensing for Python version 2.1 and later and own and protect the trademarks associated with Python

Purpose and Main Usage

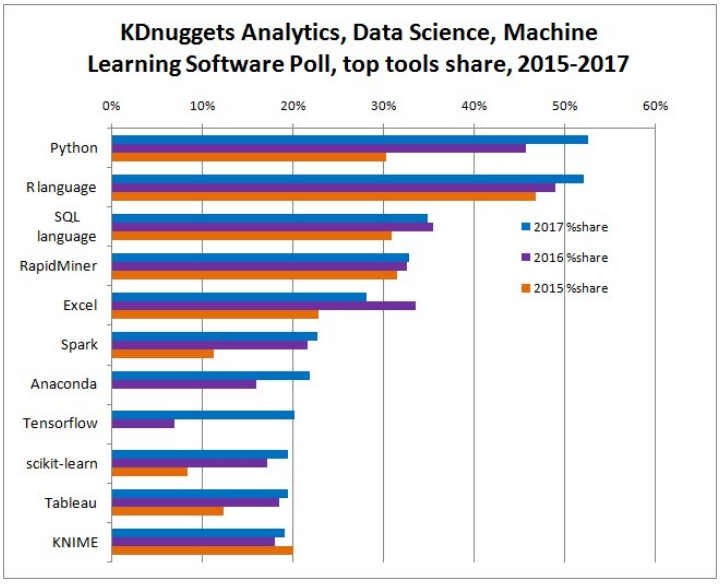
- R focuses on data analysis, statistics, and visualization
- Main users are in academics and research communities
  - Expanding also in the enterprise market
- The philosophy of the Python language is about the code readability and productivity
- Used primarily by developers that turn to data science
  - As a general purpose language, it is already present in the enterprise market

Popularity: Kaggle Competition



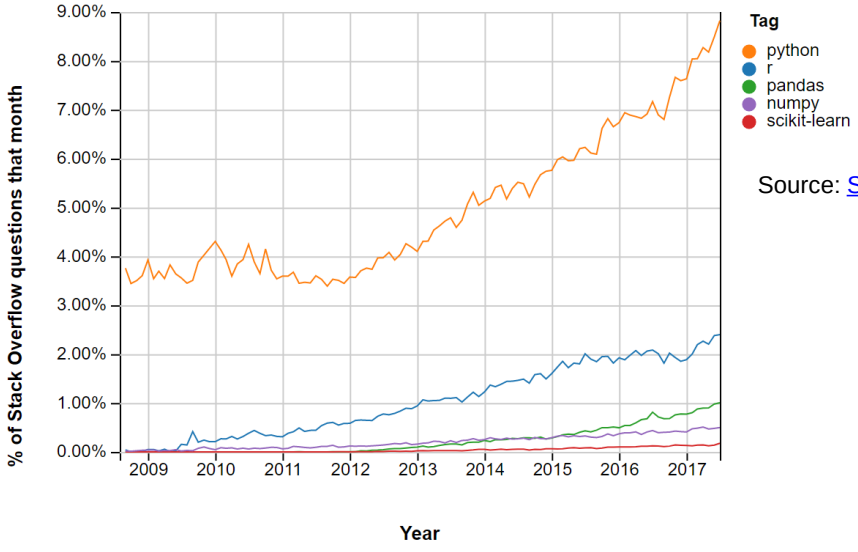
Source: [Kaggle survey of competitors 2017]

Popularity: KDnuggets



Source: [18th annual KDnuggets Software Poll]

Popularity: Stack Overflow Trends



Source: [Stack Overflow Trends]

Support

- R – huge community support
  - Mailing lists
  - Stack Overflow
  - Community documentation
- Python – good support for general coding
  - Mailing lists
  - Stack Overflow
  - Community code and documentation

Syntax

- R - statistical models in few lines
  - However, no strict coding style
  - Can write the same functionality in many different ways
  - Easy to use complex formulas and algorithms
- Python – coding and debugging generally easier
  - Nice, indented code
  - Write same functionality always in the same way
  - Easy to write something new

Learning Curve

- R has a steep learning curve
  - Once beyond basics, it is simple to do advanced stuff
  - For data science, easy to test different ideas quickly
- Python learning curve low and gradual
  - Good for starting programmers
  - Might sooner need to develop custom code for data science

Code Repositories

- CRAN – Comprehensive Archive R Network
  - Huge… huge… huge repository of packages
  - Simple to contribute
  - A package can be installed with a single line
- PyPI – Python Package Index
  - Repository of packages
  - Complicated to contribute
  - Take care about package dependencies when installing

Data Science - R

- R clear leader
  - Basic data analysis without additional packages; huge number of additional packages
  - Naming convention not enforced
- Some of the most popular packages include:
  - dplyr, plyr, data.table, stringr, zoo for data manipulation
  - moments, descry for descriptive statistics
  - vcd, corrgram for intermediate statistics
  - ggplot2, lattice for graphics
  - psych, e1071, caret, party, rpart.plot for DM / ML

Data Science with Python

- The vase majority of the packages you need for data science comes with basic installation
  - The lower number of available packages might not be a problem for many users
- Some of the most popular libraries include:
  - numpy, pandas for data manipulation
  - scipy for scientific computing
  - matplotlib, seaborn for visualization
  - sklearn (scikit-learn) for DM / ML

Miscellaneous

- R pros: lingua franca for statistics, excellent visualizations, huge community, Rstudio IDE
- R cons: limited for non-ML apps, inconsistent syntax, many times slow, weird naming
- Python pros: multi-purpose language, clear syntax, IPython (Jupiter) Notebook
- Python cons: visualizations more convoluted, limited number of ML packages, naming convention sometimes not enforced as well

MS Libraries

- R
  - **RevoScaleR** – this is a set of parallelized scalable R functions for processing data, data overview and preliminary analysis, and machine learning models
  - **RevoPemaR** – this package allows you to write custom parallel external algorithms
  - **MicrosoftML** – added in december 2016, with many additional scalable machine learning algorithms implemented
- Python
  - **revoscalepy** – adapted from R, not all functions
  - **microsoftml** – adapted from R, not all functions
  - Both bring R syntax in Python (inconsistency)

VS Integration

- R Tools for Visual Studio
  - Download for VS 2015, integrated in VS 2017
  - Empty R script only
  - Similar to RStudio
  - Installs MS R Client (includes MS R libraries)
- Python projects
  - Plethora of templates
  - Machine learning templates in VS 2017
  - Follows the RStudio paradigm
  - Installs Anaconda Python (no MS Python libraries)
  - Installs additional popular tools (IPython, Jupiter Notebook, Spider…)

Usage in BI Suite

- R integrated in:
  - Azure ML (Execute R Script and Create R Model)
  - Power BI (R Script Data Connector and R Script Visual)
- Python integrated in:
  - Azure ML (Execute Py Script)
    - Plots not automatically redirected to images as they are when using R
    - Explicitly save any plots to PNG files to be returned back to Azure ML
  - Power BI (Py Script Data Connector and Py Script Visual)
- Both integrated via SQL Server only in SSRS and SSIS

Conclusion

- No clear winner
- If you want to focus on data science, then R might be a better choice
  - Packages…
- If you are more developer type of a person, Python might be a better choice
  - Readability…
- In SQL Server 2017, you have both of them!

Book Promo

For the attendees only – discount for the Data Science with SQL Server Quick Start Guide

- Promo codes:
  - EWXHVI15 – 15% p-book
  - GBWZJK50 – 50% e-book
- Valid from Jan 17th to Mar 10th
- Create a login on the Packt site www.packtpub.com and add the book to the cart
  - Use the promo codes above