Introduction to R Programming

Bok, Jong Soon
javaexpert@nate.com
https://github.com/swacademy/R

What is R?

R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R.

What is R?

- Is a scripting programming language for statistical data manipulation and analysis.
- Is a software environment for statistical analysis, graphics representation and reporting.
- Is world’s most widely used statistics programming language.
- Provides a wide variety of graphical techniques, linear and nonlinear modelling, data analysis, and time-series analysis with great extent.

What is R? (Cont.)

- Was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand.
- First appearance in 1993.
- Is currently developed by the R Development Core Team.
- Was named R, based on the first letter of first name of the two R authors (Robert Gentleman and Ross Ihaka), and partly a play on the name of the Bell Labs in AT&T Language S(for statistics).

Licensing

- R is free software released under the Free Software Foundation's General Public License.
- This means that R is free of any restrictions on how it can be disseminated.
- Versions of R can be obtained without charge and can be redistributed to others.
- The license prevents the creation of encumbered derived works (i.e. commercial versions).

Brief History

Developed by
John Chambers,
Rick Becker,
Allan Wilks

	태동	전파	확산
	Bell Lab	Commercial	GNU/Open source
O/S	• UNIX	• BSD/System V • HP, IBM, SUN	• LINUX • Application
Analysis System	• The S system	• S-PLUS	• R • Packages
	1976년 Bell Lab 태생	1988년 라이센스 양도	1993년도 오픈소스화

Brief History (Cont.)

The R Foundation for Statistical Computing (R Development Core Team)

organization

The Comprehensive R Archive Network (CRAN)

distribution

Windows
UNIX
OS X

5,093 Packages
(2014/01/14)
3,290 Packages
(2011/09/18)

Related Projects

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS
Analysis genomic data
More 749 Packages

..

sourceforge

github
SOCIAL CODING

Early History - 1990

- **Ross Ihaka joins the Department of Statistics at the University of Auckland.**
- **Robert Gentleman spends sabbatical from the University of Waterloo.**
- **During a chance encounter in the corridor, the following exchange takes place**
  - Gentleman: “Let’s write some software.”
  - Ihaka: “Sure, that sounds like fun.”
- **The initial goal is to build a testbed for trying out ideas and to publish a paper or two.**

The Initial Language

```
> (set x (seq 10))
(1 2 3 4 5 6 7 8 9 10)
> (sum x)
55
> (set factorial (lambda (x)
    (if (< x 1)
        1
        (* x (factorial (- x 1))))))
<closure>
> (factorial 5)
120
```

Early History - 1992

- **Robert Gentleman joins the department at Auckland.**
- **A decision is made to develop enough of a language to teach introductory statistics courses at Auckland.**
  - It is decided to adopt the syntax of the S language developed at Bell Laboratories.
  - As a joke, the name “R” is coined for the language (standing for Robert and Ross).

Early History - 1994

- An initial version of the language is complete.
- Colleagues overseas encourage us to release the language as “free software.”
- A little thought convinces us that there are limited prospects for the software as a commercial product.
- We adopt the Free Software Foundation GPL as our license and begin to make releases via the internet.
- We start a small email list so that we and our users can discuss R.

The original R developers plotting world domination.

Early History - 1996

- By 1996 we were becoming victims of our own success.
- We were being supplied with a continual stream of bug reports and suggestions for improvement.
- Maintaining the mailing list was becoming problematic.
- It was beginning to be clear that the project was getting close to the limit of what two of us could handle.

Success! - 1997

- The mailing list turned out to be very successful and our user base increased enormously (to nearly 100!).
- The list was so successful that was split into the present r-help and r-devel lists.
- Kurt Hornik and Fritz Leisch established the CRAN archive at TU Vienna as a repository for user contributions.
- We became so deluged with patches and requests for enhancements that we decided to open up the development process by giving a selected “core” of developers direct access to the CVS archive.

R Becomes A GNU Project

From: Richard Stallman <rms@gnu.ai.mit.edu>
To: ihaka@stat.auckland.ac.nz
cc: rms@gnu.ai.mit.edu
Subject: Re: Seen on your wishlist
Date: Tue, 16 Sep 1997 21:56:06 -0400

So [explicitly], yes we would like R to be considered as a GNU program.

I hereby dub R GNU software!

A Free Software Project

- Since we opened up the project, it has gone ahead in leaps and bounds.
- On February 29, 2000, the software was deemed fully featured enough and stable enough for the 1.0 release to take place.
- There are now nearly 20 core developers maintaining and extending the language interpreter and its basic functionality.
- The group includes a number of well-known researchers in Statistical Computing.
- The software now has a regular six-monthly release cycle and will shortly see the release of version 2.10.

The intense software development effort leading up to R version 1.

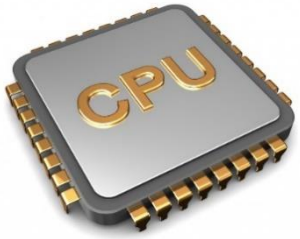
Features of R

- R is an Open source or Free Software.
- Is freely available under the GNU General Public License.
- It is available on a wide variety of platforms such as: Mac OS, UNIX, and Windows.
- Easy coding.
- Wide number of packages.
- Object-Oriented Programming.
- Functional Programming.

Limitations

- **Two major complaints are:**
  - "It's too slow for my analysis."
  - "It can't handle my multigigabyte data set."

- In-Memory로 데이터 처리
- 32Bit Machine ➔ $\frac{2^{32}}{1024^2} = 4G$
- 64Bit Machine(64Bit OS)



- 연산에 1 Core만 사용함
- 자원의 낭비
- Multi-core 지원 Packages로 해결
- R 2.14.0 Version에서 병렬처리 지원 (parallel package 기본 탑재)

3rd Party Packages로 단점의 보완 (like RHive)

Reference Sites

- **R Home Page**
  - http://www.r-project.org
- **한국 R 사용자 그룹**
  - http://www.r-project.kr
- **한국 R 사용자 커뮤니티**
  - http://r.fossa.kr/
- **R Studio Page**
  - http://www.rstudio.com
- **R Wikipedia**
  - http://en.Wikipedia.org/wiki/R_(programming_language)

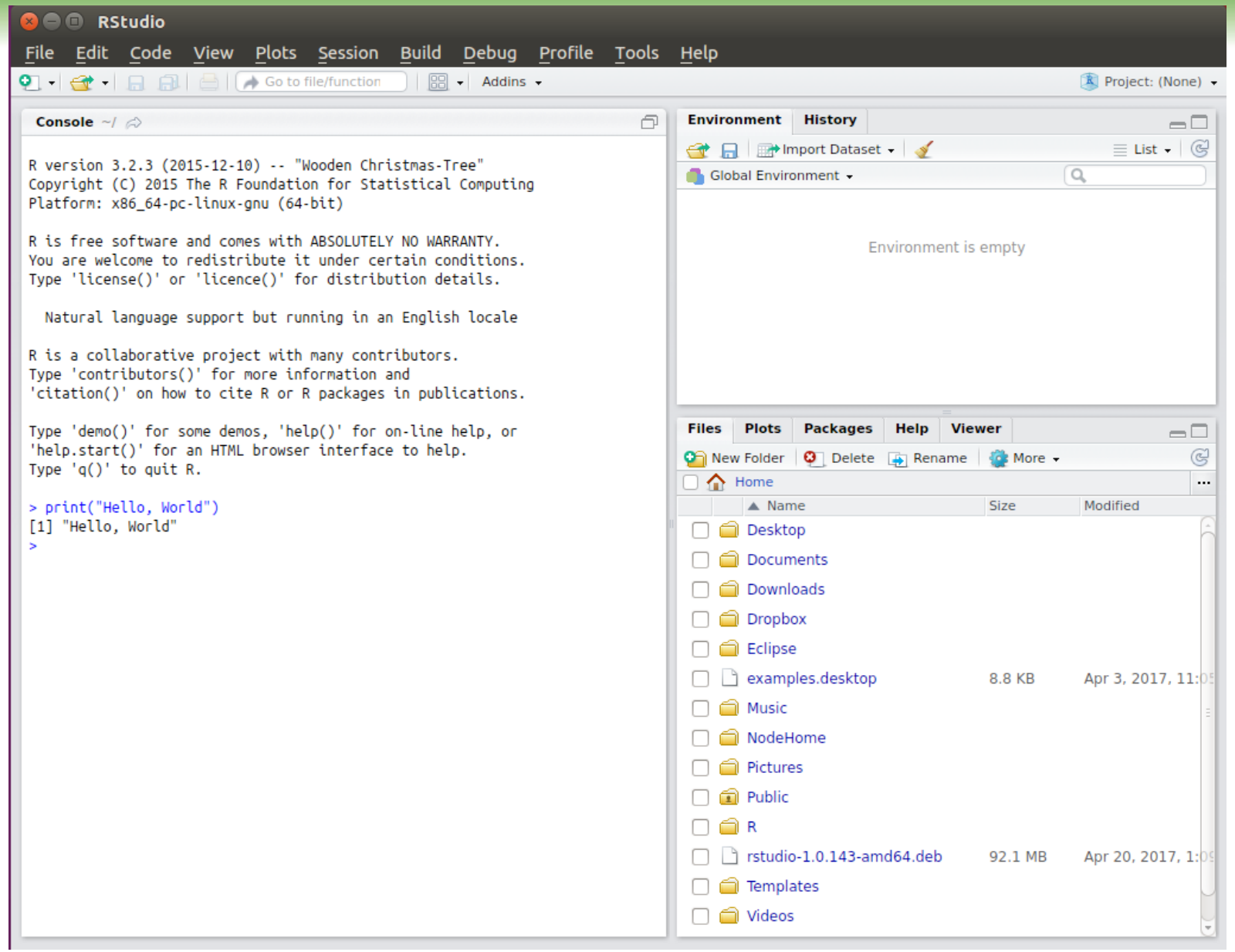
Reference Sites (Cont.)

- **Rseek**
  - http://rseek.org
- **The R Journal**
  - https://journal.r-project.org/
- **R Commander**
  - http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/
- **R GUI Projects**
  - https://www.r-statistics.com/tag/r-gui/

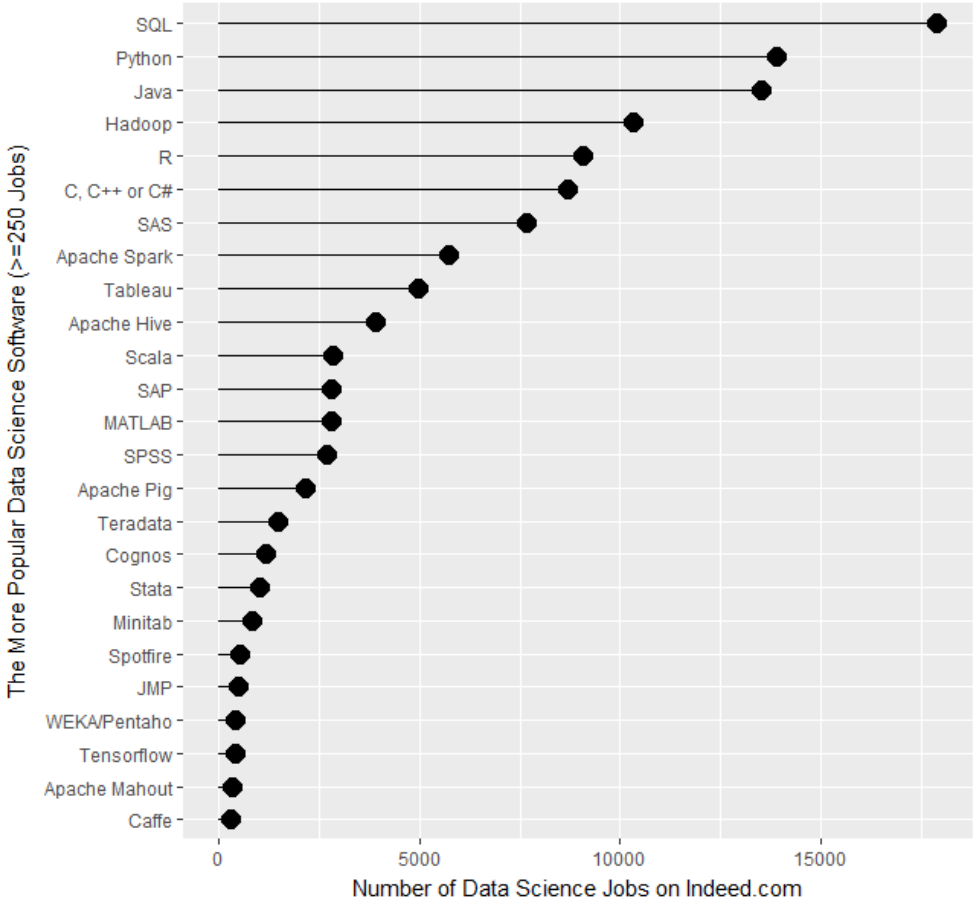
Open source or free tools

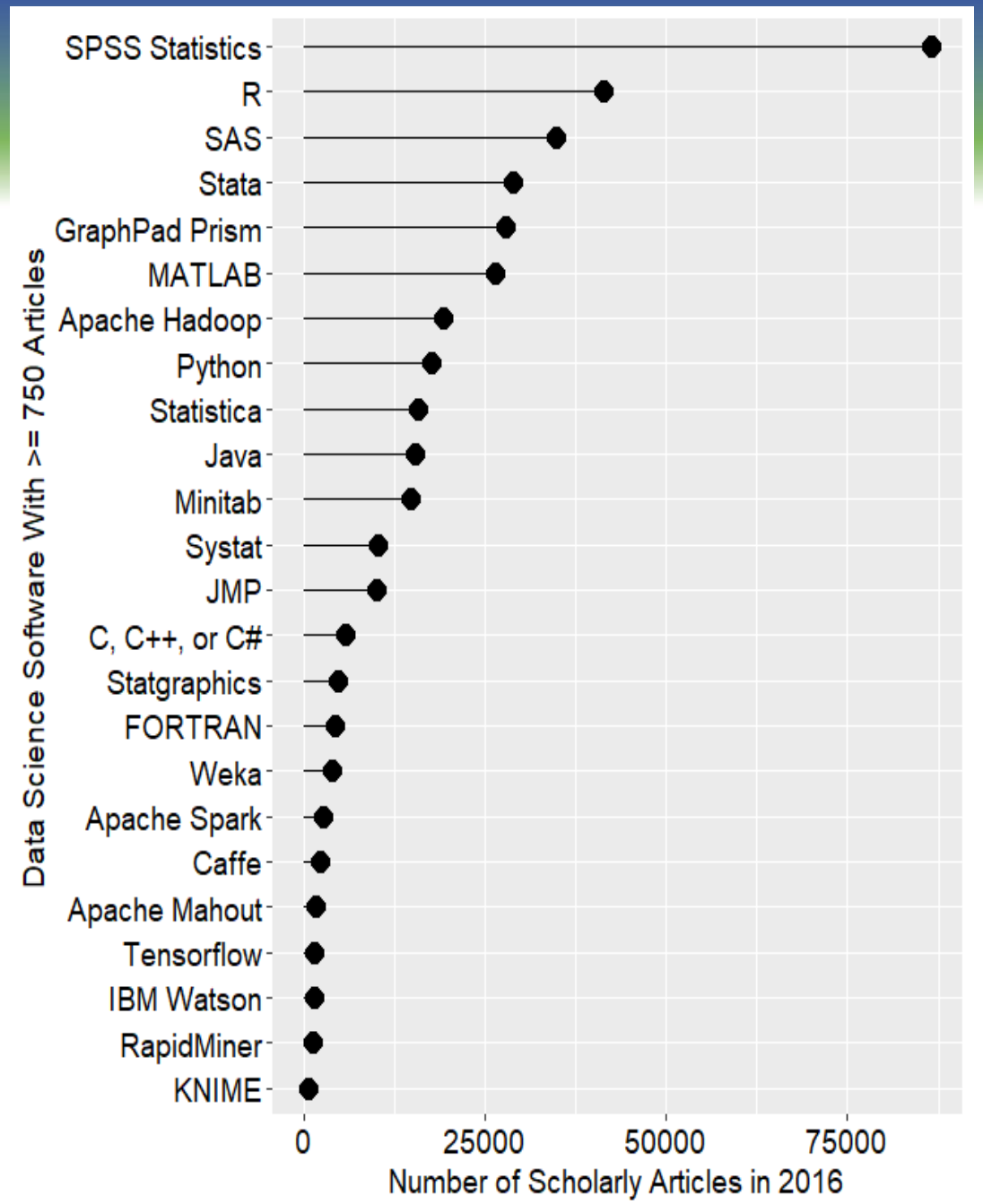
- RStudio, http://www.rstudio.org/
- StatET, http://www.walware.de/goto/statet/
- ESS (Emacs Speaks Statistics), http://ess.r-project.org/
- R Commander: John Fox, “The R Commander: A Basic-Statistics Graphical Interface to R,” Journal of Statistical Software 14, no. 9 (2005):1–42.
- JGR (Java GUI for R), http://cran.r-project.org/web/packages/JGR/index.html

Open source or free tools (Cont.)



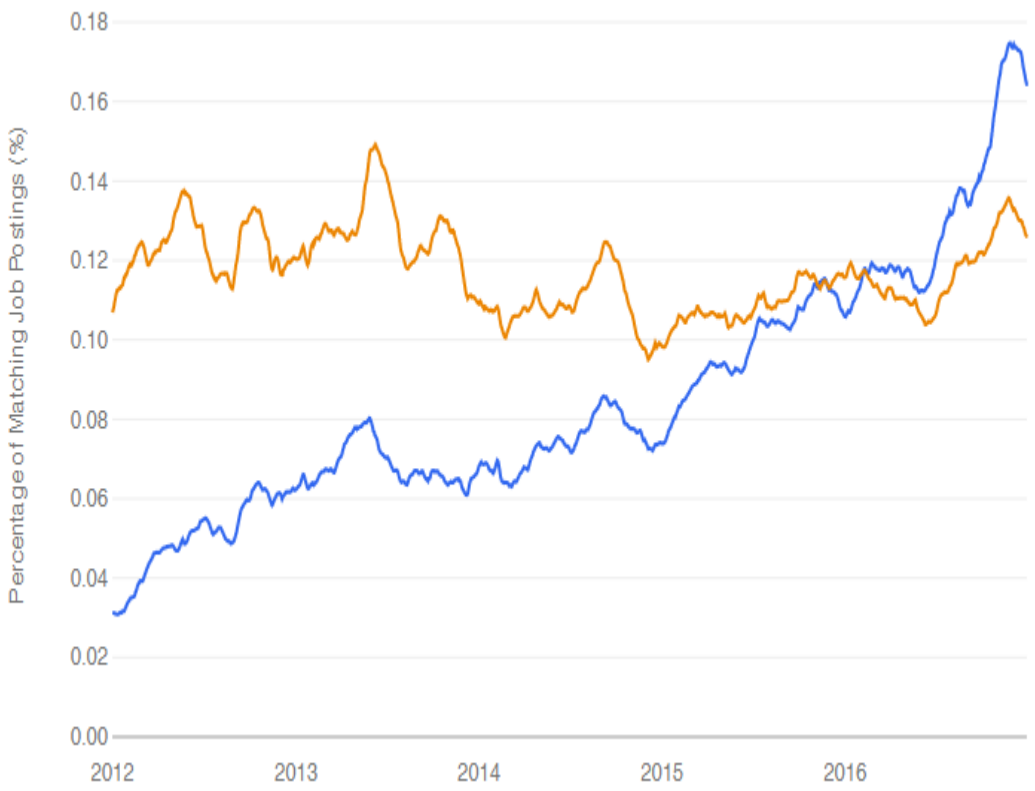
Big Data Period & R



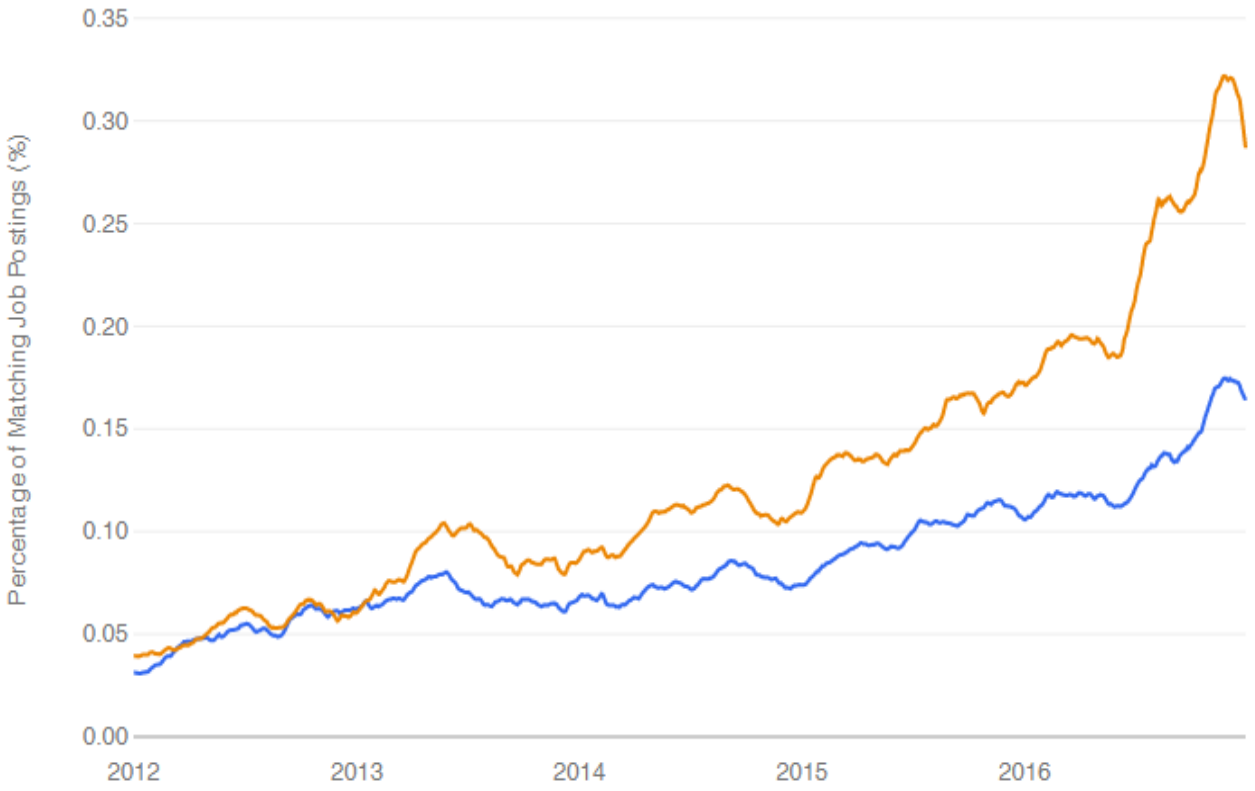


Source : http://r4stats.com/articles/popularity/

Big Data Period & R (Cont.)



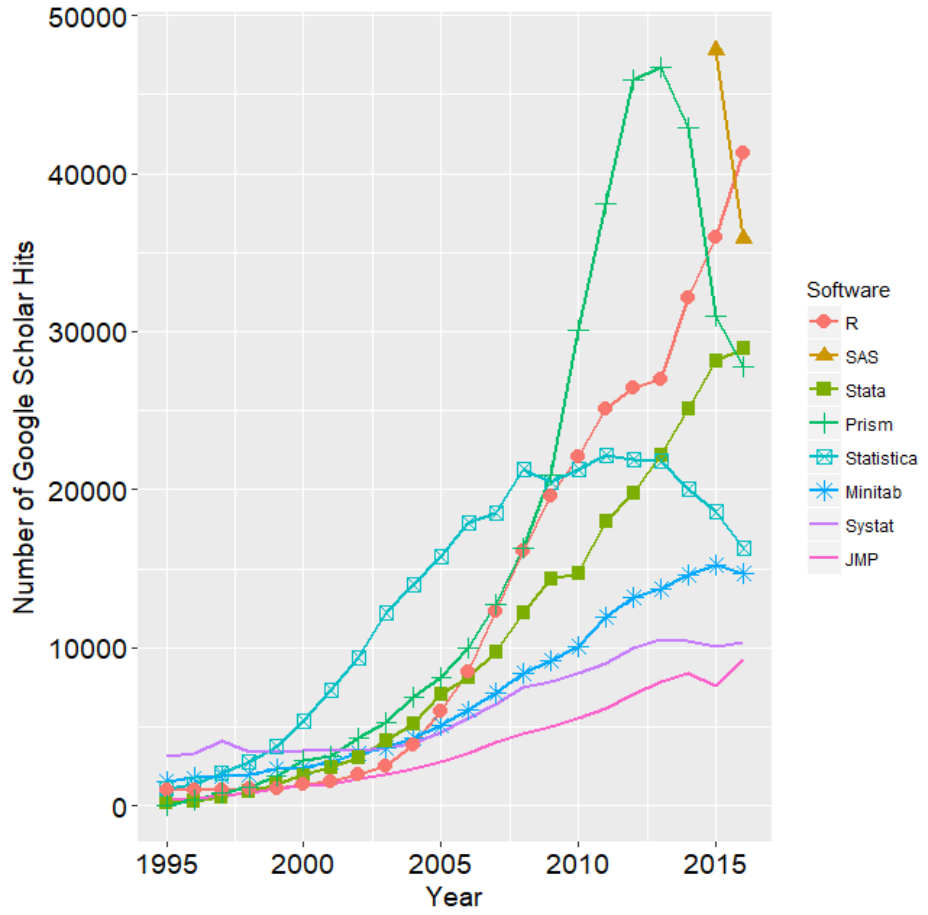
Data science job trends for R (blue) and SAS (orange).

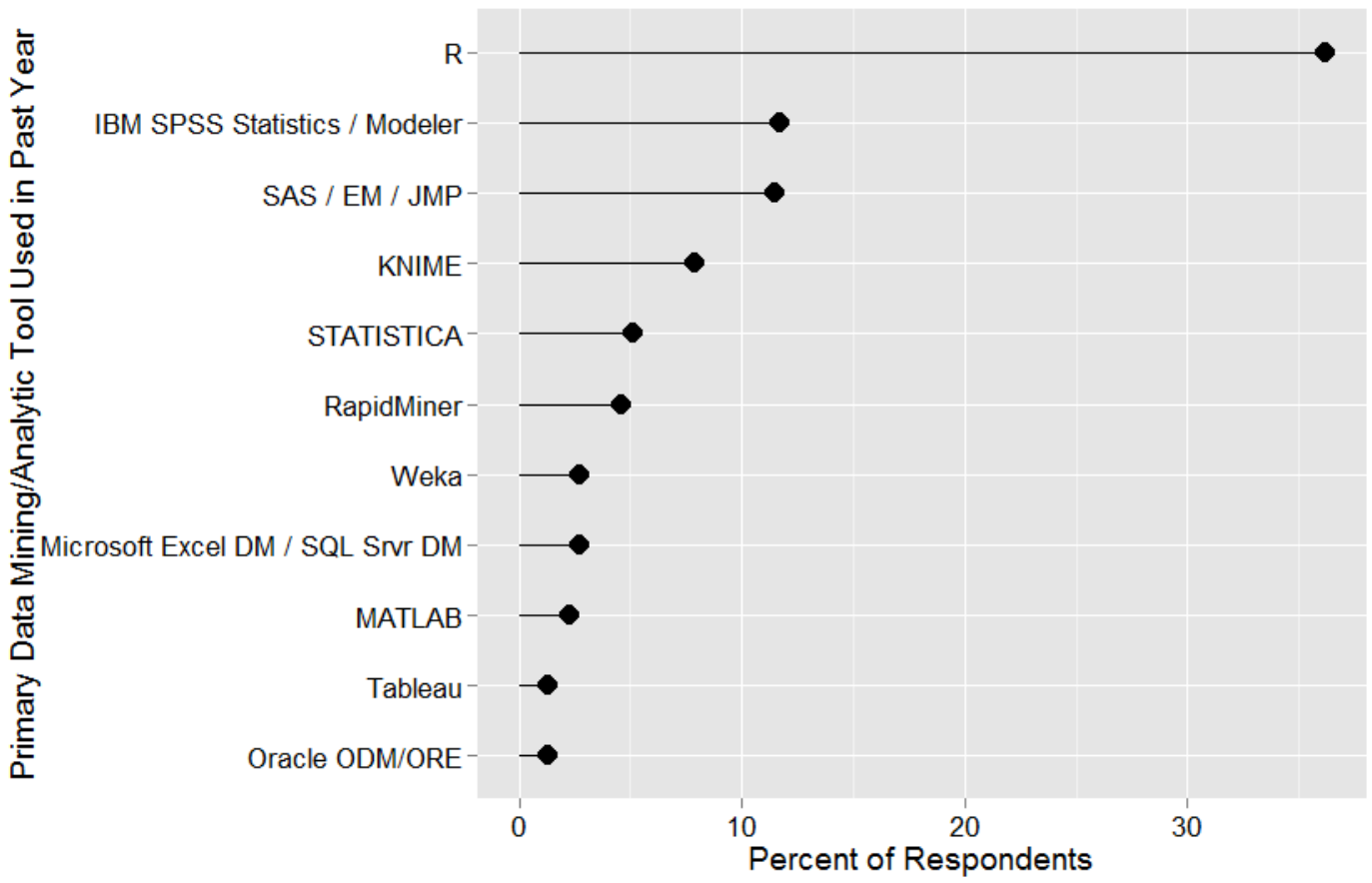


Jobs trends for R (blue & lower) and Python (orange & upper)

Source : http://r4stats.com/articles/popularity/

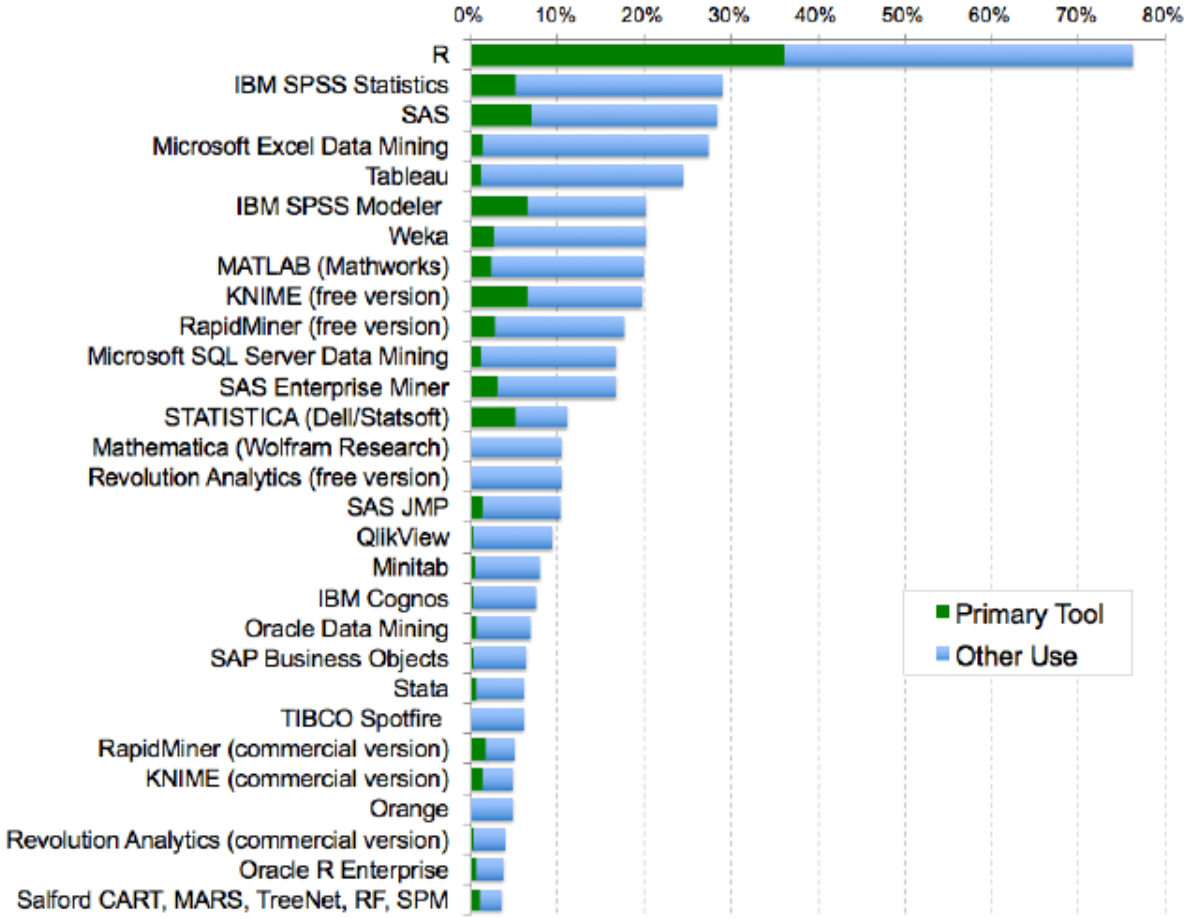
Big Data Period & R (Cont.)



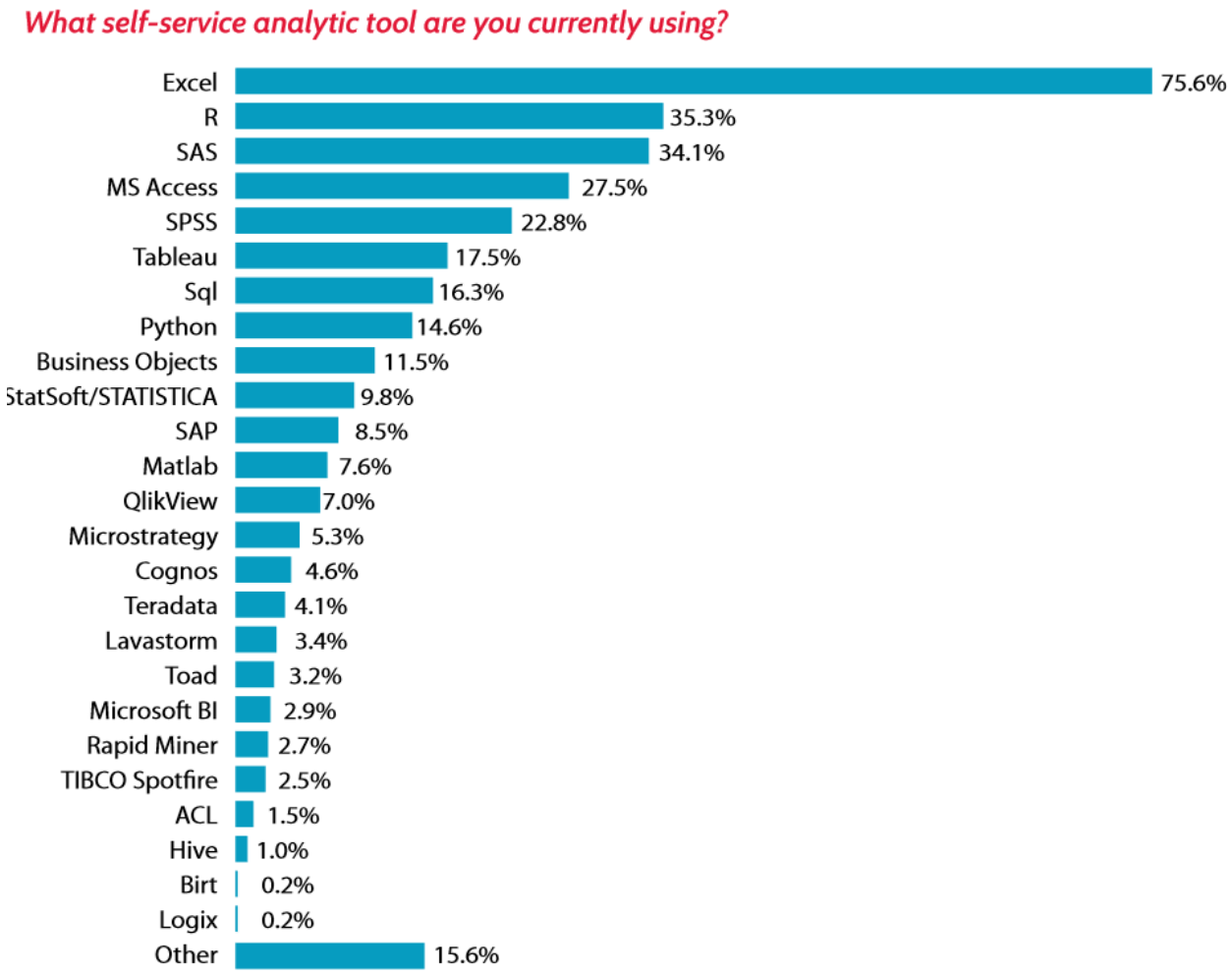


Source : http://r4stats.com/articles/popularity/

Big Data Period & R (Cont.)



Analytics tools used by respondents to the 2015 Rexer Analytics Survey

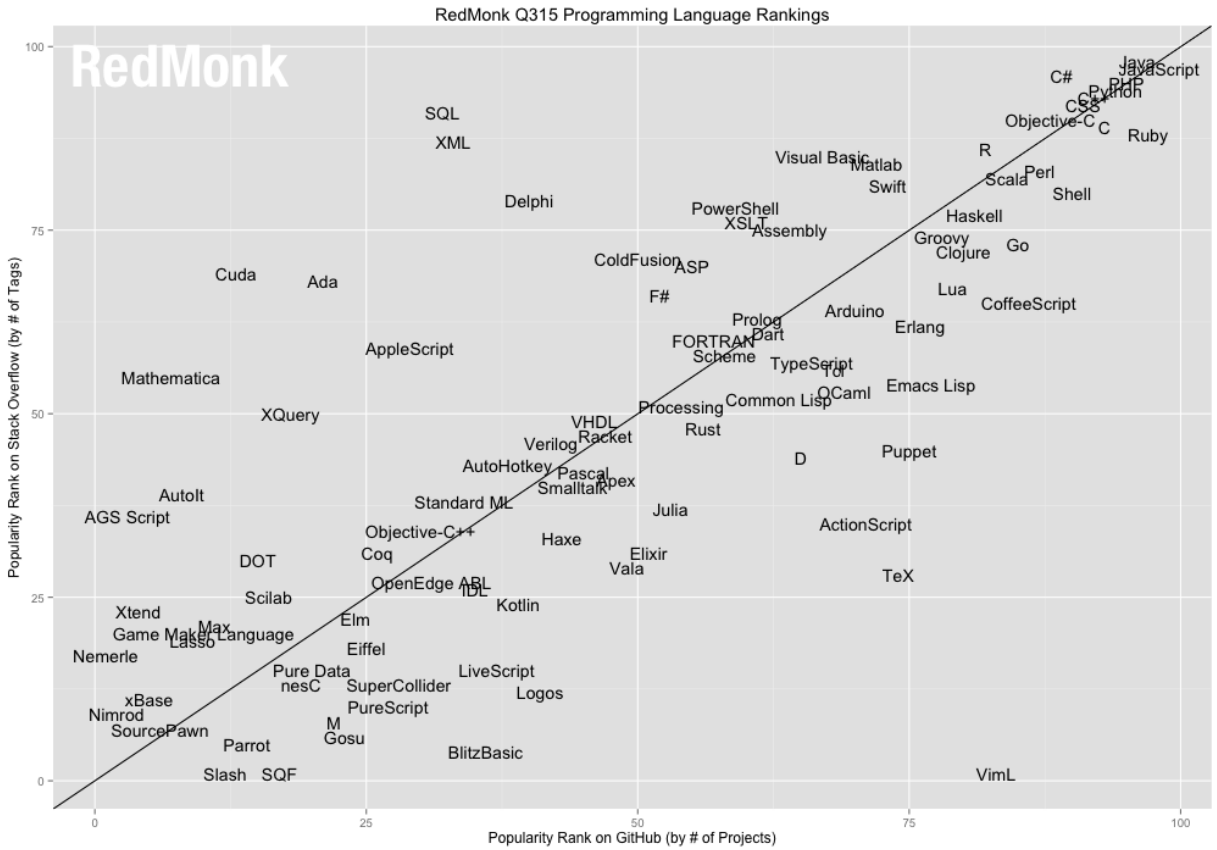


Source : http://r4stats.com/articles/popularity/

Big Data Period & R (Cont.)

Language Rank	Types	Spectrum Ranking	Spectrum Ranking
1. Java		100.0	100.0
2. C		99.9	99.3
3. C++		99.4	95.5
4. Python		96.5	93.5
5. C#		91.3	92.4
6. R		84.8	84.8
7. PHP		84.5	84.5
8. JavaScript		83.0	78.9
9. Ruby		76.2	74.3
10. Matlab		72.4	72.8

IEEE Spectrum language popularity rankings



Source : http://r4stats.com/articles/popularity/

Big Data Period & R (Cont.)

Jun 2017	Jun 2016	Change	Programming Language	Ratings	Change
1	1		Java	14.493%	-6.30%
2	2		C	6.848%	-5.53%
3	3		C++	5.723%	-0.48%
4	4		Python	4.333%	+0.43%
5	5		C#	3.530%	-0.26%
6	9	^	Visual Basic .NET	3.111%	+0.76%
7	7		JavaScript	3.025%	+0.44%
8	6	v	PHP	2.774%	-0.45%
9	8	v	Perl	2.309%	-0.09%
10	12	^	Assembly language	2.252%	+0.13%
11	10	v	Ruby	2.222%	-0.11%
12	14	^	Swift	2.209%	+0.38%
13	13		Delphi/Object Pascal	2.158%	+0.22%
14	**16**	^	**R**	**2.150%**	**+0.61%**
15	48	^^	Go	2.044%	+1.83%
16	11	vv	Visual Basic	2.011%	-0.24%
17	17		MATLAB	1.996%	+0.55%
18	15	v	Objective-C	1.957%	+0.25%
19	22	^	Scratch	1.710%	+0.76%
20	18	v	PL/SQL	1.566%	+0.22%

Source : https://www.tiobe.com/tiobe-index/

Statistical Computing

주요 통계계산 기능	통계량/기초통계	• EDA(Exploratory Data Analysis) • Summary
	통계분석	• 전통적인 통계분석 방법론 • 최신 통계분석 방법론, Spatial, Bayesian 통계 등
	마이닝 분석	• Decision Tree, SVM, Clustering, … • WEKA interface
	시뮬레이션	• 모형 시뮬레이션 • Operation Research
	수치해석	• 미분, 적분, 행렬대수 • 근사값 계산, Optimization
교육	대학/대학원 교육	• 대학 및 대학원에서의 통계 교육의 표준으로 사용
업계의 활용	분석업무 활용	• Google : Google Analytics(SaaS)에 R을 사용 • Facebook, Yahoo 등 회사에서 내부 분석용 도구로 활용
	제품 개발	• Oracle, Teradata, EMC 등 업체의 DBMS 내 분석툴로
활용 프로젝트	Bioinformatics 프로젝트	• BioConductor Project – 749 이상의 Packages • 게놈, Bio, 신약연구 등 • Bioinformatics의 표준 통계분석 언어
	Finmatrics 프로젝트	• 금융 예측분석에 사용, 여러 가지 금융 예측모형 구현

Appliance DBMS for Big Data Analytics

Vendor	Products	Analytics Engine
Oracle	Big Data Appliance Exadata	Oracle R Enterprise ®
IBM	InfoSphere BigInsights Netezza Appliance	Revolution R, SAS, SPSS 연동
Teradata	Aster Discovery Platform	SQL-Map/Reduce, SAS, R
EMC	Greenplum Data Computing Appliance	Java, R
SAP	HANA (In memory Appliance) – Not Big Data	R 연동 사례

- **Appliance DBMS & Hadoop**
  - Hadoop 보다 Appliance DBMS 에 치중 예상
- **Analytics**
  - Analytics Product을 DBMS Product 내부에 포함시키고 있음.
  - Analytics Engine은 공통적으로 R 사용

References

- Ihaka, Ross. "The R Project: A Brief History and Thoughts About the Future"
- Norman Matloff. "THE ART OF R PROGRAMMING". 2011.
- Joshua F. Wiley. "Beginning R". 2015.
- http://www.r-graph-gallery.com/