

R의 저변확대를 위한 노력

2012/10/24

유충현 - kt cloudware (NexR) KRUG 대표

1.2011 Conference 설문 분석

R User Conference 소개 Survey 문항의 분석 및 시사점

2. CRAN Download 분석

CRAN Download 분석 CRAN Download 분석 및 시사점

3.R의 저변확대를 위한 노력

KRUG의 활동 소개 R의 활성화를 위한 방안

1. 2011 Conference 설문 분석

R User Conference 소개

Survey 문항의 분석

R 사용자와 R에 관심이 많은 잠재적 사용자들의 정보 공유의 장으로 2011년 시작

- R User Conference 2011
 - ✓ New Wind in Korea with GNU R
 - ✓ 12개 세션
 - ✓ Dr. Duncan, John Fox의 축하 메시지
 - ❖ R Development Core Team
 - ✓ Tal Galili 세션
 - ❖ R-Bloggers.com 운영자
- R User Conference 2012
 - ✓ R의 활용 사례 공유 및 가능성 진단
 - ✓ 5개 주제의 11개 세션
- 주체
 - ✓ Korea R Users Group
 - ✓ kt clouldware (NexR)
- 협찬
 - ✓ Begas, RTechCenter, 서울과학종합대학원
 - ✔ 자유아카데미, 인사이트
- 세션발표자
 - ✔ 고려대학교 허명회 교수님 외 10분



1. 2011 Conference 설문 분석

Survey 문항의 분석

Survey 문항의 분석

R User Conference 2011의 참가자를 대상으로 Survey 수행

- 대상: R User Conference 2011 참석자
- 인원:65명
- 문항: 13문항
 - ✔ R관련 9문항
 - ❖ R 경험 여부 (3 문항)
 - ❖ 직업 (1 문항)
 - ❖ R 관심분야 등 (3 문항)
 - ❖ R의 장점 (1 문항)
 - ❖ 사용 분석툴 (1 문항)
 - ✓ Conference 관련 2문항
 - ❖ Conference 인지 경로 (1 문항)
 - ❖ R의 미래 (1 문항)
 - ✓ 다음 Conference의 희망 주제
 - ✓ 다음 Conference의 발표 희망자 조사

R 이용에 대한 사용자 질의서

안녕하세요?

R 사용에 대한 Conference 참가자들의 의견을 알아보고자 합니다.

설문을 통해 더 나은 행사의 기획 및 R의 활용을 위한 R 사용자들의 온라인/오프라인의 교류에 대한 방향성을 진단하려고 합니다.

바쁘시더라도 잠시 시간을 내 주셔서 응답해 주시면 감사하겠습니다.

R 관련 질문

Q1.R을 사용해 본 경험은 얼마나 되나요?(한 달에 1회이상 사용해 본 것을 기준으로)

(1) 사용해 보지 않았다.

(2) 1년 미만

3) 1년이상 2년미만

(4) 2년이상 3년미만

(5) 3년이상

Q2. 현재 하시는 일은 무엇입니까?

- (1) 학생 및 교직원, 교수
- (2) 데이터 분석 (통계, 확률)
- (3) 소프트웨어 개발 엔지니어
- (4) 경영 지원 및 비니지스데이터 분석

(5) 기타

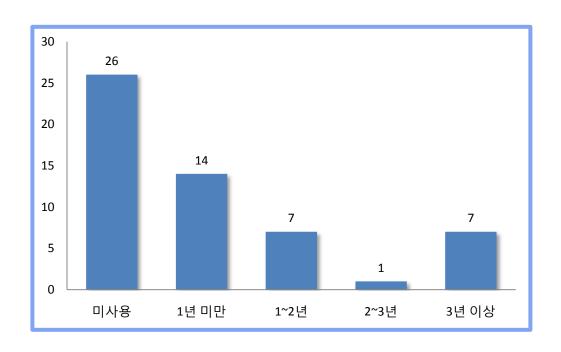
Q3.R을 현업의 프로젝트 적용해 본 횟수는?

- (1) 아직 안 해봤지만 해볼 생각이 있다.
- (3) 2~3호

(4) 4회 이상

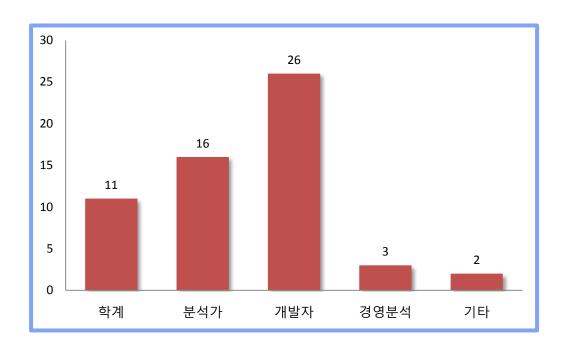
(5) 적용 계획 없음

미사용자의 비율이 가장 높았으며, 1년 미만의 경험자를 포함 초심자의 비율이 높음



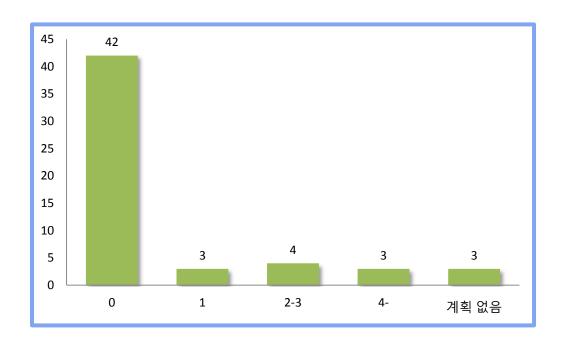
■ R을 사용해보지 않은 참석자가 26명으로 47.27%를 차지. 그 다음으로는 1년 미만>3년 이상>1~2년 사용>2~3년 사용자 순.

SI 관련 업체 종사자의 비율이 가장 높았음. Big Data 분석 엔진으로서 R이 회자되는 영향이 컸던 것으로 판단



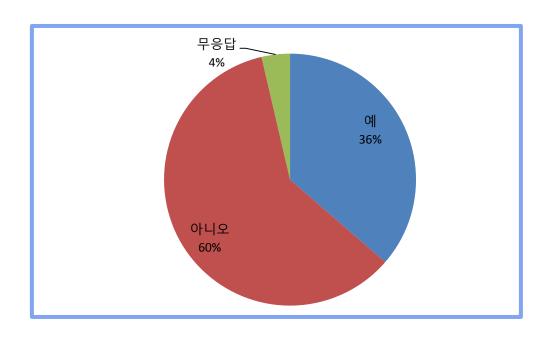
- 개발자가 전체 58개의 응답 중 26명으로 44.82%의 비율로 가장 높게 나타남
- 개발자 > 분석가 > 학계 > 경영분석 > 기타 순으로 참석자의 직업군 분류
- 기타: UX researcher, IT consultant

아직은 현장에서의 R의 업무 적용의 사례는 많지 않지만, 일부 현장에서는 R을 업무에 사용하는 사례가 생겨나고 있는 추세임



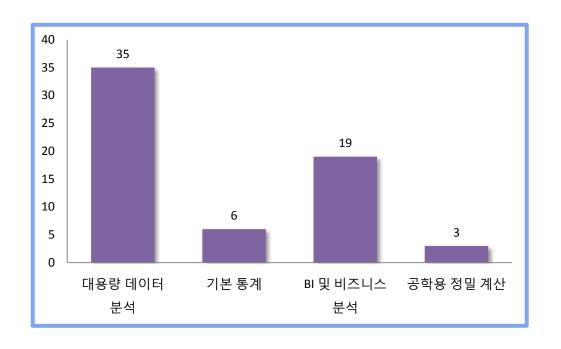
- 의향은 있으나 아직 적용하지 않은 사례가 42명으로 76.36%의 높은 비율을 나타남
- 4회 이상 적용한 경우는 3건으로 낮게 나타남

의외로 학교에서 R을 접해 보지 못한 경우가 높아, 대학에서의 R을 이용한 통계학습의 방법이 시급해 보임



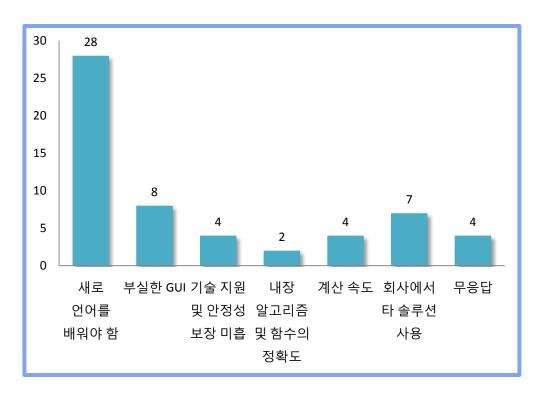
- R을 학교에서 접하지 못한 경우가 60%로 높게 나타남
- 현재 학계에 있는 11명 중에서도 R을 접하지 못한 경우가 4건 확인됨

관심 활용 분야에서는 대용량 데이터 분석의 비중이 가장 높게 나타남



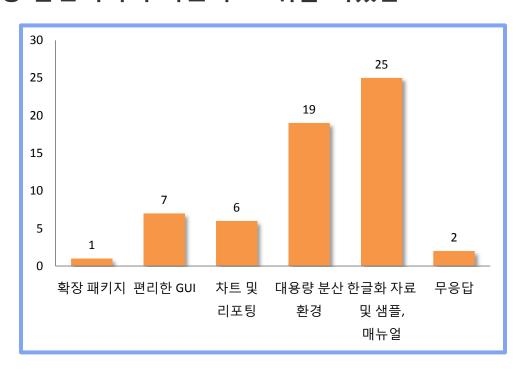
- 총 63건의 대답 중 대용량 데이터 분석이 35명(55.56%)으로 가장 높게 나타났으며, 그 다음으로는 BI 및 비즈니스 분석 > 기본 통계 > BI 및 비즈니스 분석 순으로 나타남
- 중복 응답 중 "BI 및 비즈니스 분석+대용량 분석"의 경우가 가장 높음

R을 도입하는데 가장 장애가 되는 것은 새로운 언어를 배워야 한다는 부담으로 나타남



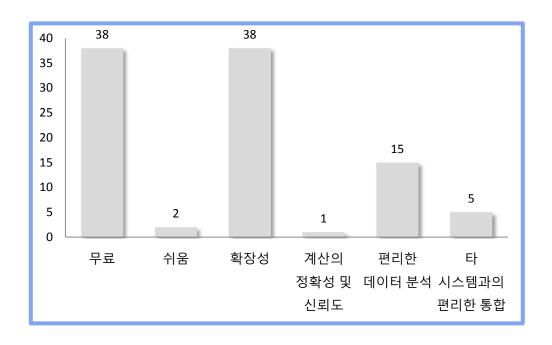
- 총 57 응답건수 중 새로 언어를 배워야 하는 부담이 28건(49.12%)으로 가장 높게 나타남
- 그 외에는 GUI 문제-회사의 타 솔루션 사용 문제가 높게 나타남

R의 확산을 위해서는 한글화된 자료 및 매뉴얼이 가장 시급한 문제로 꼽혔으며, 대용량 분산처리의 지원이 그 뒤를 이었음



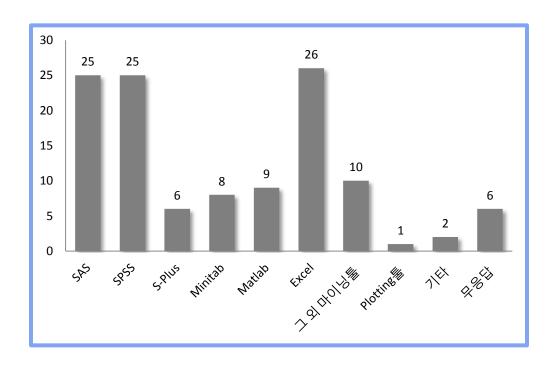
- 총 60 응답건수 중 한글화 자료 및 샘플, 매뉴얼 제공을 필요로 한다는 대답이 25건 (41.67%)으로 가장 높았음
- 또한 대용량 분산 환경 또한 19건으로 높게 나타남

R의 장점으로는 무료 소프트웨어라는 점과 확장성을 가장 높게 평가하였음



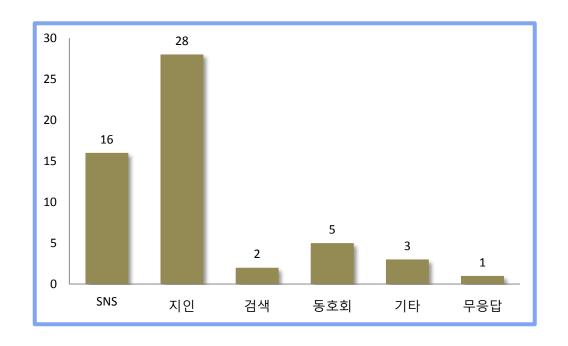
- 응답 건수 99건 중 무료, 확장성을 장점으로 꼽는 의견이 각 38건(38.38%, 명수 대비 69.09%)으로 높게 나타남
- 전체 55명 중 3개를 대답한 경우가 10명, 2개를 대답한 경우가 25명, 1개를 대답한 경우가 10명으로 집계

현장에서는 Excel, SAS, SPSS 등을 많이 사용하고 있음



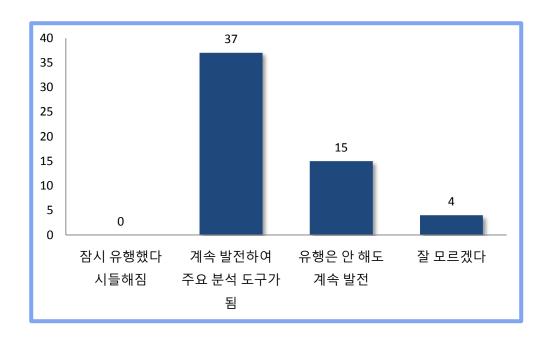
- 총 응답자 55명, 총 응답건수 118건 중 Excel이 26건(22.03%, 응답자수 대비 47.27%)으로 가장 높게 나타났고, SAS와 SPSS가 25건으로 거의 유사하게 높은 비율로 나타남
- 기타: GAUSS, Spotfire

지인과 SNS 등과 같은 Social Network를 통해서 Conference의 개최를 인지하였음



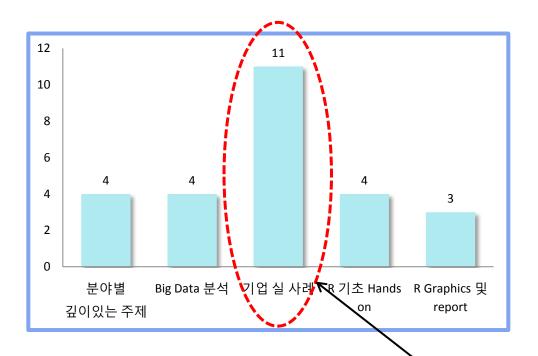
- 전체 55명 중 28명(50.91%)이 지인을 통해 Conference를 알게 되었음
- 그 다음으로 SNS의 비율이 높게 나타남

R의 미래에 대해서는 많은 사람들이 계속해서 발전할 것으로 기대했으며, 잠시 유행했다가 시들 것이라는 기대를 한 사람은 한 명도 없었음



■ 대부분이 긍정적인 답변을 해주셨고, 계속 발전하여 주요 분석도구가 된다는 의견이 55명 중 37명으로 67.27%로 나타남

기업의 실 사례에 대한 니즈가 가장 많았는데, 이는 대부분의 응답자가 현재 R 도입예정 혹은 희망하는 경우가 많은 것으로 판단됨



■ 기업의 실 사례에 대한 주제의 선호도가 가장 높게 나타남

2012년 Conference의 주제를 "R의 활용사례 공유 및 가능성 진단" 으로 선정한 사유

Survey 결과의 Issue로부터 몇 개의 Key Findings를 도출함

주요 Issue

R의 경험에 대한 이슈

- 1년 미만의 경험자를 포함한 미경험자의 비율이 가장 높음
- R의 사용 의향은 있으나 70%이상이 업무에 미 사용중임
- 대학에서 R을 접해보지 않은 사람이 60%로 높게 나타남
- R의 진입의 약 50%가 새로운 언어를 배워야 한다는 부담을 들었음

R이 업무에 활용되기 위한 이슈

- 데이터 분석가나 학회보다는 시스템 개발자의 비율이 높음
- R의 필요한 지원 요소에서 대용량 데이터 분석의 지원의 비중이 높음
- 확장성이 높은 R의 장점
- 무료 소프트웨어의 장점
- 기업의 실 사례를 검증 후 도입하려는 경향

R의 제약 및 부족한 인프라에 대한 이슈

- R의 관심 활용 분야에서 대용량 데이터 분석이 가장 높게 나타남
- R의 필요한 지원 요소에서 한글화된 매뉴얼, 자료 등이 가장 높았음
- Excel, SAS, SPSS 등을 많이 사용함
- 계속 발전을 하여 주요한 분석 도구가 될 것이다.

Key Finings

배우려고 시도하는 초심자를 위한 학습 인프라의 필요성

현장에서 활용하려는 예비 사용자를 위한 지원 체계의 필요성

상용 툴 대비 대용량 데이터 분석에서의 제약 및 교육 리소스 부족

분석의 확장성 및 유연한 시스템 통합 기능의 장점 활용

낙관적이고, 우호적인 관심과 기대

2. CRAN Download 분석

CRAN Download 분석

CRAN(The Comprehensive R Archive Network)의 패키지 Download 로그를 분석함

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- . Download R for Linux
- Download R for MacOS X
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2012-06-22, Roasted Marshmallows): <u>R-2.15.1.tar.gz</u>, read <u>what's new</u> in the latest version.
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are <u>available here</u>. Please read about <u>new</u> <u>features and bug fixes</u> before filing corresponding feature requests or bug reports.
- Source code of older versions of R is available here.
- Contributed extension packages

■ 대상 : http://cran.nexr.com의 WEB access log

■ 기간 : 2012-03-08 ~ 2012-09-18 (약 6개월), 약 2,021,227 Download

■ 방법 : Linux Shell 및 R을 이용한 전처리, R을 이용한 분석

■ 기타 : 일부 어뷰징(Abusing)패턴도 존재함

R Packages Download 국가는 다양하며, 한국, 중국, 일본에서 가장 많음

• 84개 국가에서 Download함

(상위 10개국)

한국	중국	일본	미국	인도	호주	홍콩	영국	독일	캐나다
1,779,93	8 168,430	45,987	9,599	2,608	1,621	1,410	1,388	1,017	935

Ranking	국가명	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
1	한국	R	zoo	car	mvtnorm	colorspace	Hmisc	rgl	Rcmdr	rJava	e1071
2	중국	R	survival	car	alr3	Z00	plm	leaps	Imtest	Rcpp	MASS
3	일본	R	mvtnorm	colorspace	MASS	rgl	Rcmdr	Z00	ggplot2	Hmisc	car
4	미국	R	zoo	colorspace	car	Hmisc	Rcmdr	caTools	mvtnorm	rgl	Imtest
5	인도	R	foreign	e1071	Z00	Rcmdr	car	colorspace	gdata	Hmisc	leaps
6	호주	R	ggplot2	munsell	Rglpk	scales	colorspace	digest	e1071	gbm	hexbin
7	홍콩	R	colorspace	MASS	plyr	psych	zoo	digest	ggplot2	coda	RColorBrewer
8	영국	R	mvtnorm	SparseM	akima	colorspace	Hmisc	multcomp	Z00	gplots	quantreg
9	독일	R	lme4	nlme	plyr	RColorBrewer	bitops	caTools	gdata	ggplot2	gplots
10	캐나다	R	effects	colorspace	coda	car	Imtest	mvtnorm	Z00	abind	alr3

대학(원)에서 가장 많은 Download의 숫자를 보임

• 1,037 조직에서 Download함 (중복 조직 존재)

(상위 50개 조직)

Rank	1	2	3	4	5	6	7	9	9	0
0	숭실대학교	주식회사 케 이티	주식회사 엘 지유플러스	서울대학교	에스케이브로드 밴드주식회사	서강대학교	충북대학교	고려대학교	(의)대진의료 재단분당재 생병원	충북본부장
1	연세대학교	한국과학기 술원	엘지유플러 스	한신대학교	엔씨소프트	단국대학교	대아레저산 업(주)	주식회사 케 이티수도권 고객센터	행정자치부	(주)엘지씨엔에스
2	충남대학교	강원대학교	한국직업능 력개발원	(주)HCN충북 방송	(주)큐릭스광진 성동방송	성균관대학 교	서울대학교 병원	(주)신한은행	(주)엘지파워 콤	한양대학교
3	삼성네트웍 스	일진엔터프 라이즈	엘지파워콤	한국해양연 구원	한국산업은행	제일증권	(주)티브로드 중부방송	혜천대학	티엔에스미 디어	현대모비스(주)
4	국민대학교	한국개발연 구원	한국항공대 학교	파워콤	넥스지	이화여자대 학교정보전 산원	엘지씨엔에 스	씨앤앰강남 케이블티브 이	건국대학교	분당서울대학교 병원

(상위 50개 조직 분류 집계)

대학(원)	기업	공공(연구기관)	병원	금융	ISP
18	9	4	3	3	13

대학(원)에서 가장 많은 Download의 숫자를 보임

OS는 MS-Windows가 대부분을 차지함

os	MS-Windows Mac OS		Linux	기타	
건수	1,983,745	18,376	17,421	169	
비율	98.2	0.9	0.9	0.0	

- 일 평균 10,357.5건 다운로드
- 월 집계 현황

월	2012-03	2012-04	2012-05	2012-06	2012-07	2012-08	2012-09
건수	280,449	246,918	243,985	239,613	220,693	467,122	320,931
일평균건	11,685.4	8,230.6	7,870.5	7,987.1	7,119.1	15,068.5	17,829.5

Package Download Top 15

Ranking	Package	Description	Keyword	category
1	ZOO	Regular and Irregular Time Series	Time Series	Statistical Computing
2	car	Companion to Applied Regression	Regression	Statistical Computing
3	mvtnorm	Multivariate Normal and t Distributions	Distributions	Statistical Computing
4	colorspace	Color Space Manipulation	Color	Visualization
5	Hmisc	many functions useful for data analysis	Statistics	Statistical Computing
6	rgl	3D visualization device system (OpenGL)	3D visualization	Visualization
7	Rcmdr	basic-statistics GUI for R	GUI	UX
8	rJava	Low-level R to Java interface	Java interface	System Integration
9	MASS	'Modern Applied Statistics with S'	Statistics	Statistical Computing
10	e1071	Misc Functions of the Department of Statistics (e1 071), TU Wien	Statistics	Statistical Computing
11	RODBC	An ODBC database interface	ODBC	System Integration
12	Imtest	Testing Linear Regression Models	Regression	Statistical Computing
13	multcomp	Simultaneous Inference in General Parametric Mo dels	Multiple Compa risons	Statistical Computing
14	RColorBrewer	ColorBrewer palettes	Color	Visualization
15	plyr	Tools for splitting, applying and combining data	tools	Data Manipulation

2. CRAN Download 분석

CRAN Download 분석의 시사점

R Packages Download 결과의 Issue로부터 몇 개의 Key Findings를 도출함

주요 Issue

Key Finings

R의 사용자에 대한 이슈

• 다양한 국가에서 Korea CRAN으로부터 Package를 Download 함

R 활용의 다양성에 대한 이슈

- Mac OS나 Linux에서 R을 사용하는 유저가 있음
- rJava, RODBC 등 시스템 통합을 위한 패키지의 Download가 많음

R의 가능성에 대한 이슈

- 대학에서 교육 용도로 사용하는 미래의 현업 데이터 분석가가 많음
- Download 회수가 증가하는 추세임

R을 사용하는 해외의 거주(유학)자들도 있을 수 있음

Platform Independent의 특징으로 다양한 OS에서 사용함

SI관련 역량을 가진 사용자의 존재

교육 현장에서의 R의 학습

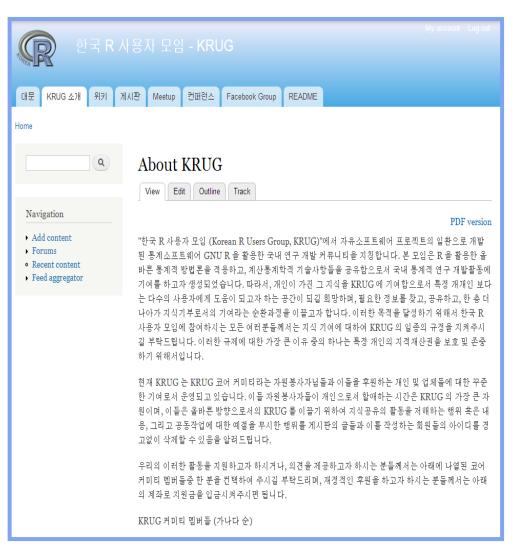
R 사용자의 증가 추세

3. R의 저변확대를 위한 노력 KRUG의 활동 소개

KRUG의 활동 소개

KRUG는 Korea R Users Group의 약어로 한국 R 사용자 모임을 의미합니다.

- Korea R Users Group
 - ✓ 한국 R 사용자 모임
 - ✓ http://r-project.kr
 - ❖ 오픈소스 CMS인 Drupal로 만들어진 온라인 홈 페이지
 - ❖ 396명 가입 (10/23 현재)
 - √ http://r-project.kr/wiki/
 - ❖ 한글화된 R 사용자 가이드
 - ✓ Meetup 8회 개최 (2011/12~)
 - ❖ 지식공유 및 기부, 기여 지식들에 대한 토론의 오프라인 모임
 - ✓ R User Conference
 - ❖ kt cloudware(NexR)과 공동 개최
 - ❖ 2012/10/24 2회 개최
 - ✓ Thanks
 - ❖ 이철희
 - ➤ KRUG 커미터
 - R Translation Teams(Korean)



KRUG의 활동 소개

Wiki Book은 KRUG 회원의 R 관련 지식 기부의 채널입니다.



- R Wiki Book
 - ✓ http://r-project.kr/wiki/
 - ❖ 한글화된 R 사용자 가이드
 - ✓ wikipedia와 같이 R 사용자가 자신의 지식을 기부하는 장
 - ✓ 지식 기부의 방법
 - ❖ 본인이 자주 쓰는 패키지에 대한 튜토리얼을 위키를 활용하여 작성
 - ❖ 본인이 개발한 패키지의 소개서와 사용설명서를 위키를 활용하여 작성
 - ❖ R 에 구현된 소스코드들에 대한 해석 및 분석에 대한 의견을 작성
 - ❖ 기타 등등
 - ✓ 현재 Wiki 카테고리 (대분류 및 중분류)
 - ❖ R 프로그래밍
 - ▶ R 초급
 - ▶ R 중급
 - ➤ R 고급
 - ➤ R 패키지
 - ❖ R과 통계학
 - ▶ 기초통계학
 - ▶ 일반선형모형
 - ▶ 시뮬레이션 스터디
 - ▶ 수리통계학
 - ▶ 빅데이터
 - ❖ 함께 알면 좋은...
 - ▶ 우분투 리눅스
 - ▶ 페도라 리눅스와 R
 - ➤ CentOS 리눅스와 R
 - ➤ LaTeX
 - > C/C++
 - ▶ 다양한 프로그래밍 언어들

- Translating R Messages (R 메시지 번역)
 - √ http://translation.r-project.kr
 - ✓ R의 영문 에러 메시지 등을 자국어의 메시지로 변환하는 프로젝트
 - ❖ 커맨드 라인의 영문 메시지를 한글로 볼 수 있게 해주는 프로젝트
- Translating R Documentation (R 기술문서 번역)
 - ✓ http://r-project.kr/wiki/R_고급/기술문서
 - ✓ R의 설치와 함께 제공하는 고급 기술문서의 한글 번역 프로젝트
 - ❖ An Introduction to R
 - R language definition
 - Writing R Extensions
 - ❖ R Data Import/Export
 - * R Installation and Administration
 - R Internals
 - ❖ The R Reference Index
- Korean R Help Files (한국어 도움말 프로젝트)
 - ✓ http://r-project.kr/wiki/Help-ko
 - ✓ R의 대표적인 Packages의 한글 번역 작업
 - ❖ help(ls) 등
 - base, stats, graphics, ...

- KRUG Newsletter
 - ✓ KRUG의 활동 및 Meetup 소식 및 R의 팁 등을 담은 Newsletter 발행
 - ❖ R의 Newsletter를 답습
- R 사용자 교육 세션 개발
 - ✓ 기초적인 R 학습을 위한 콘텐트의 교육
 - ❖ 프로그래밍
 - Graphics
 - Statistics
- R Study group 지원
 - ✓ R을 활용하는 데이터 분석 Study group 지원
 - ❖ Study의 결과를 Meetup에 발표하고, Wiki에 정리
- R 교육 동영상 KRUG 홈페이지에 게시
 - ✓ 2분 동영상 91개
 - ❖ 동영상 기여 한글 자막 본으로 R Tech Center에서 번역 후 지원 (감사합니다)
- KRUG는 비 영리 커뮤니티로 여러 분의 자발적인 기여로 운영됩니다.

KRUG의 Issue로부터 몇 개의 Key Findings를 도출함

주요 Issue

Key Finings

R의 지식 기부에 대한 이슈

- •지식 공유 및 기부의 채널은 다양하게 열려있으나 참여가 저조함
- 배우려고만 하고, 정보를 공유하려는 노력이 부족함
- 경험자의 참여가 절대적으로 필요함

커뮤니티 활동에 대한 이슈

- 커뮤니티의 허리를 담당할 R을 사랑하는 R 경험자의 참여가 부족함
- 커뮤니티 활성화를 위한 아이디어와 운영 방법의 경험 부족
- 비 영리로 여가 시간을 쪼개서 활동하는 것에 대한 애로사항

커뮤니티 활동의 도구에 이슈

• 홈페이지가 지식의 기부를 중심으로 설계되고 채택되어 불편함 (국내에서는 게시판 중심의 채널에 익숙해져 있음) 지식의 기부와 공유의 문화가 확산되어야 함

R 커뮤니티의 로열티 제고가 필요함

다양한 아이디어 및 운영에 제안의 수렴이 필요함 3. R의 저변확대를 위한 노력

R의 활성화를 위한 방안

R의 활성화를 위한 방안

Key Finings 노력의 방향 □ 배우려고 시도하는 초심자를 위한 학습 인프라의 필요성 □ 현장에서 활용하려는 예비 사용자를 위한 지원 체계의 필요성 한글화된 사용자 가이드 Survey □ 상용 툴 대비 대용량 데이터 분석에서의 제약 및 교육 리소스 부족 리소스의 체계적인 축적 □ 분석의 확장성 및 유연한 시스템 통합 기능의 장점 활용 □ 낙관적이고, 우호적인 관심과 기대 R 커뮤니티의 체계적인 활동 및 지원 체계 구축 □ R을 사용하는 해외의 거주(유학)자들도 있을 수 있음 □ Platform Independent의 특징으로 다양한 OS에서 사용함 CRAN **Localization R Package** □ SI관련 역량을 가진 사용자의 존재 개발의 지원 □ 교육 현장에서의 R의 학습 □ R 사용자의 증가 추세 R 커뮤니티의 활성화로 R □ 지식의 기부와 공유의 문화가 확산되어야 함 지식의 공유 필요 **KRUG** □ R 커뮤니티의 로열티 제고가 필요함 □ 다양한 아이디어 및 운영에 제안의 수렴이 필요함

감사합니다.

bdboy@r-project.kr
http://r-project.kr