



Again wind in Korea with GNU R

R의 부활을 꿈꾸며

유충현

(한화생명, Tidyverse Korea)

2021-11-19

발표 개요

한국에서 R 사용자를 위한 첫 컨퍼런스인, '**R User Conference 2011**'이 성황리에 개최된 지 10년이 흘렀습니다. 강산이 변한다는 10년동안 국내외 데이터 분석 필드는 많은 변화와 발전을 이루었으며, 디지털 경제전환(Digital Transformation)의 가속화의 핵심에는 데이터와 데이터 분석 기술이 있다는 중론이 있습니다.

10년 전 컨퍼런스를 호스트한 R 사용자로서의 감회는, **데이터 분석 필드의 염원과 기대를 R의 대중화로 뿌리내리지 못한 것에 대한 아쉬움입니다**. 다시 R 사용자의 염원을 담아서 **R에 대한 관심과 대중화를 위한 몇 가지 시도를 소개**하면서, R 사용자의 관심을 요청합니다.

1. 과거를 회상하며
2. 미래를 설계하며
3. 아카데미를 위한 R 환경 개선
4. 엔터프라이즈를 위한 R 환경 개선
5. 마무리

과거를 회상하며

New Wind in Korea with GNU R

R User Conference 2011

New wind in Korea with GNU R

일시 : 2011년 10월 28일 (금) 09:00 ~ 18:30 | 장소 : 역삼 포스틸타워 이벤트홀 | 주최 : **NEXR**, R User's Group in Korea | 후원 : 자유아카데미, RevolutionAnalytics, Begas

- 10년 전에는
 - 빅데이터 분석 도구로서의 R 활용에 대한 기대감
 - SI, 인터넷, 통신, 게임업체 등 기업의 높은 관심도
 - Statistical Computing > Big Data > Visualization > Bioinformatics 니즈
- "New Wind in Korea with GNU R"
 - 제 1회 R User Conference 슬로건
 - Dr. Duncan, Dr. John Fox, Dr. Friedrich Leisch, Tal Galili
- 조기 마감, 210여명 오프라인 행사 참석
 - 컨퍼런스 공간의 한계로 선착순 수용

2011년도 컨퍼런스 설문지 분석

Key Findings

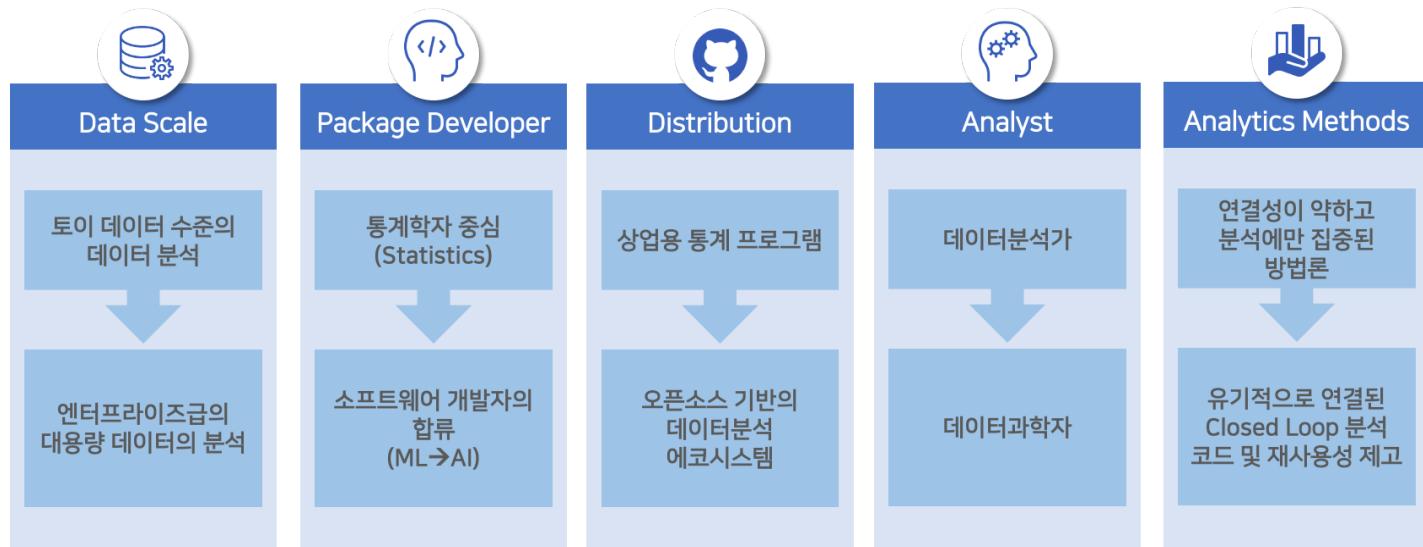
Survey 결과의 Issue로부터 몇 개의 Key Findings를 도출함

| 주요 Issue | Key Findings |
|--|--------------------------------------|
| R의 경험에 대한 이슈 | 배우려고 시도하는 초심자를 위한 학습 인프라의 필요성 |
| <ul style="list-style-type: none"> 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 | 현장에서 활용하려는 예비 사용자를 위한 지원 체계의 필요성 |
| ROI 업무에 활용되기 위한 이슈 | 상용 툴 대비 대용량 데이터 분석에서의 제약 및 교육 리소스 부족 |
| <ul style="list-style-type: none"> 데이터 분석이나 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 | 분석의 확장성 및 유연한 시스템 통합 기능의 장점 활용 |
| R의 제약 및 부족한 인프라에 대한 이슈 | 낙관적이고, 우호적인 관심과 기대 |
| <ul style="list-style-type: none"> R의 관심 활용 분야에서 대용량 데이터 분석이 가장 높게 나타남 R의 필요한 지원 요소에서 한글화된 메뉴얼, 자료 등이 가장 높았음 Excel, SAS, SPSS 등을 많이 사용함 계속 발전을 하여 주요한 분석 도구가 될 것이다. | |

R 대중화를 위한 솔루션 은 무엇이었던가?

- 진입 장벽을 낮출, 초보자를 위한 R 학습 인프라
- 일선 현장에서의 활용을 위한 지원 체계
- 엔터프라이즈 환경에서의 활용 사례 발굴
- R의 장점을 살린 커뮤니티

데이터 분석 필드에서의 패러다임 전환



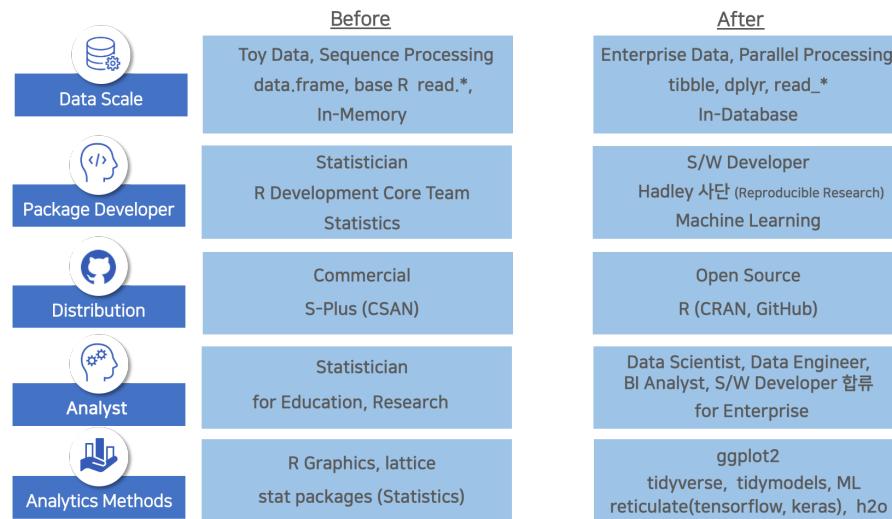
python의 대중화와 R의 소외감

- 공급부족 데이터 분석시장에 개발자 캐릭터 분석가의 진입 → 개발자 선호 툴
- CRM, 빅데이터, 거쳐 시장에서의 딥러닝의 불업 → 딥러닝 python 라이브러리

엔지니어링의 중요성 부각

- 오픈소스 생태계의 여러 솔루션을 다룰 수 있는 멀티플레이어 요구

R 생태계의 패러다임 전환



독립 시스템에서 에코 시스템으로 전환

- base R을 넘어 Tidyverse, Tidymodels, ...

재현가능 연구를 위한 생태계 완비

- 데이터 분석 경험의 공유 및 핸즈온 교육 용이

엔터프라이즈 시장에서의 활용 가능

- 부족했던 R 성능의 캐치업 → 수행속도, 데이터 핸들링 용량 등

미래를 설계하며

Again Wind in Korea with GNU R

한국 R 컨퍼런스 2021

코로나19로 촉발된 뉴노멀 시대 디지털 경제전환과 함께하는 애자일 R !

Date and Location

1. Date: 2021년 11월 19일(금) 10:00 ~ 17:00

2. Location: 온라인 라이브

◦ 연사분 촬영장소: 롯데월드타워 35층 원티드랩 (서울 송파구 올림픽로 300)

- "Again Wind in Korea with GNU R"
- 10년 전처럼, 오늘 행사가 **R에 대한 기대감 염원**에 불을 다시 지피기는 계기
- 개별 세션들을 통해서 시청자들이 많은 경험을 습득하는 컨퍼런스가 되길 기대
- 개인적으로는
 - 20여년 R 사용을 통해 꿈 꾸었던 빅 빅처를 소개하고,
 - 역동적인 한국의 R 생태계 조성을 위한 협력이 늘어나기를 기대

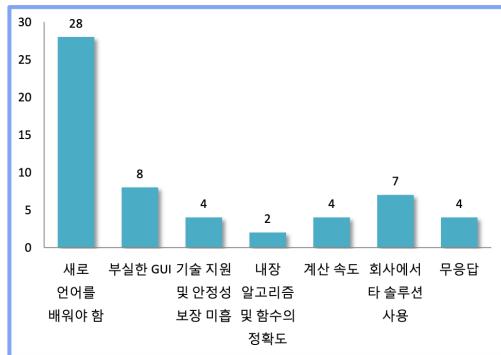
아카데미를 위한 R 환경 개선

다시 꺼낸 Survey - 진입장벽

R의 진입 장벽

Survey 결과

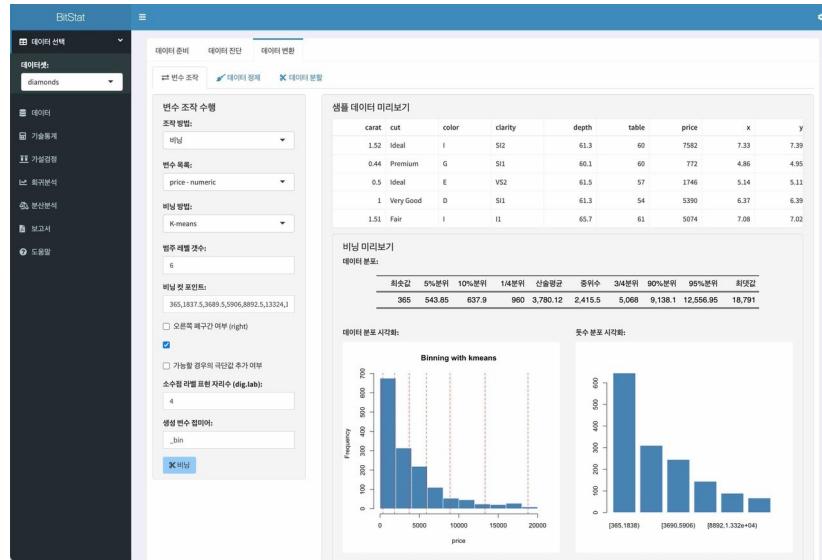
R을 도입하는데 가장 장애가 되는 것은 새로운 언어를 배워야 한다는 부담으로 나타남



- 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

- 아카데미, 공공기관, 중소기업의 R 진입 장벽은 무엇인가?
 - 새로운 언어를 배워야 한다는 부담감
 - 이미 익숙한 상용 데이터분석 소프트웨어를 사용함
 - 비용 절감 목적으로 접근하려는데, 취약한 R 교육훈련 인프라
- 오늘 이야기할 첫번째 주제 - 발상의 전환
 - "R이 아니라, R로 만들어진 데이터 분석 소프트웨어를 만들어 보자."

오픈소스 통계분석 시스템 개발

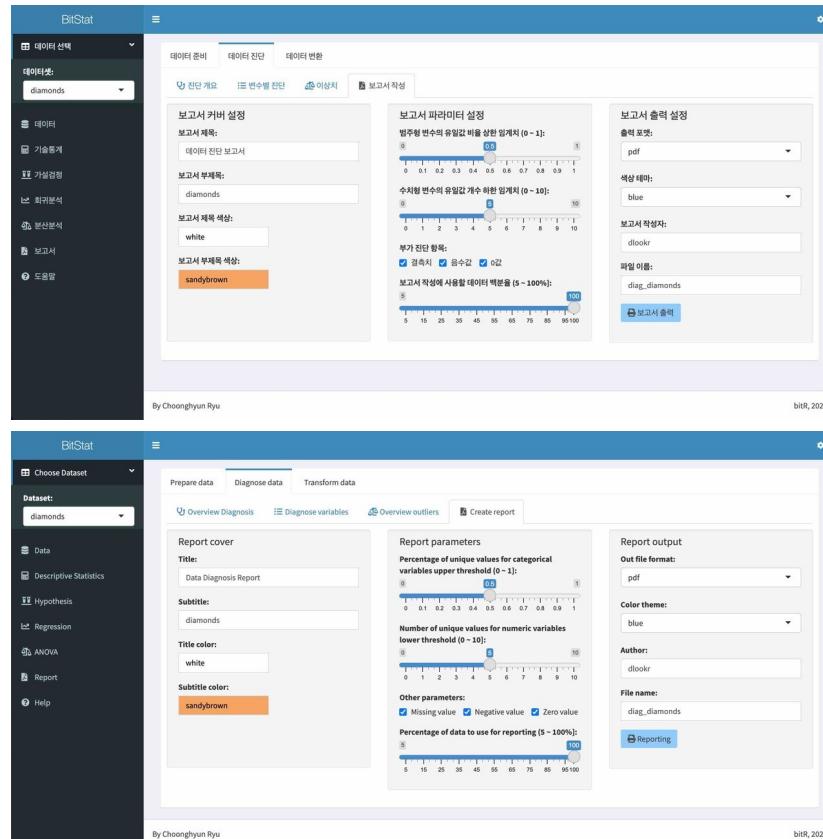


- 오픈소스 통계분석 시스템이란?
 - R/Shiny로 개발한 오픈소스 기반의 R 패키지, 현재 프로토타이핑 중
- 아카데미, 중소기업, 공공기관, 일반인 대상
 - 통계 비전공자들을 위한 데이터 분석 시스템
 - R을 모르더라도 사용가능, R 사용자는 더욱 쉽게 활용
- 기초통계 및 머신러닝 분석 기능 제공

오픈소스 통계분석 시스템 특징

다국어 지원(i18n)

- 국문과 영문 메뉴 및 메시지



The image displays two screenshots of the BitStat software interface, illustrating its multilingual support (i18n). Both screenshots show the same configuration screen with different language settings.

Screenshot 1 (Top): Report Generation Settings

- Left Panel:** Dataset dropdown set to "diamonds".
- Report Cover Section:**
 - Report title: 보고서 제목 (Report Title)
 - Report subtitle: 보고서 부제 (Report Subtitle)
 - Color settings: 보고서 색상 (Report Color) - white, 보고서 부제색 (Report Subtitle Color) - sandybrown.
- Report Parameters Section:**
 - Percentage of unique values for categorical variables upper threshold (0 ~ 1): 0.5
 - Number of unique values for numeric variables lower threshold (0 ~ 10): 5
 - Other parameters: 결측치 (Missing value), 음수값 (Negative value), 음수값 (Zero value).
- Output Section:**
 - Report output: 보고서 출력 설정 (Report Output Settings)
 - Color theme: blue
 - File name: diag_diamonds
 - Report button: 보고서 출력 (Generate Report)

By Choonghyun Ryu bitR, 2021

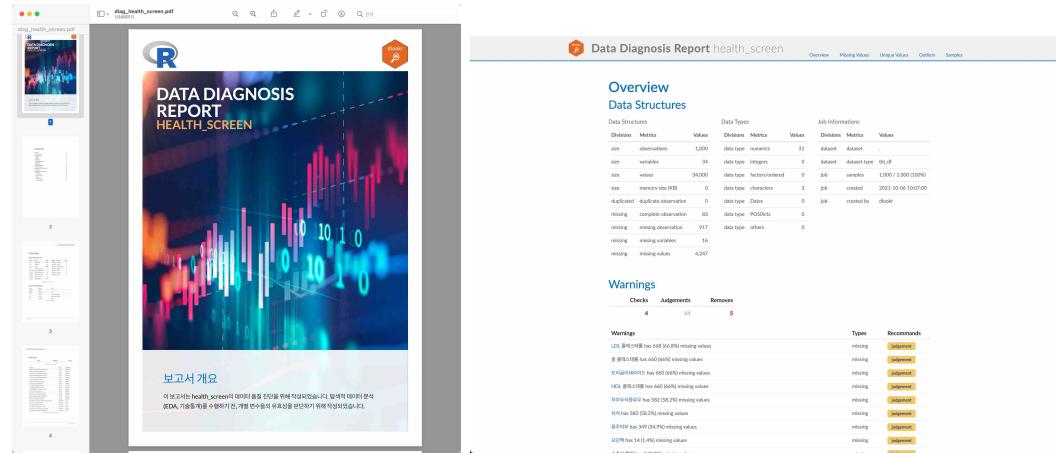
Screenshot 2 (Bottom): Diagnostic Report Configuration

- Left Panel:** Dataset dropdown set to "diamonds".
- Report Cover Section:**
 - Report title: Report cover
 - Report subtitle: Data Diagnosis Report
 - Color settings: Title color (white), Subtitle color (sandybrown).
- Report Parameters Section:**
 - Percentage of unique values for categorical variables upper threshold (0 ~ 1): 0.5
 - Number of unique values for numeric variables lower threshold (0 ~ 10): 5
 - Other parameters: Missing value, Negative value, Zero value.
- Output Section:**
 - Report output: Report output
 - Color theme: blue
 - Author: dlookr
 - File name: diag_diamonds
 - Reporting button: Reporting

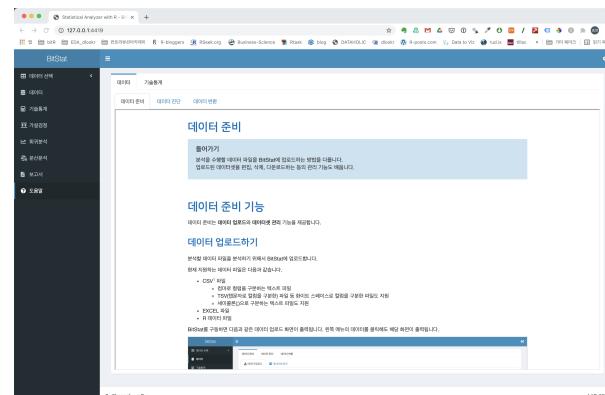
By Choonghyun Ryu bitR, 2021

오픈소스 통계분석 시스템 특징

데이터 분석 보고서 지원 → PDF 포맷 보고서와 HTML 포맷 보고서



도움말 및 튜토리얼 지원



왜 통계분석 시스템인가?

뿌리깊은 나무는 바람에 아니 뭘세

- 머신러닝, 딥러닝(AI)의 학문적 백드라운드인 통계학
- 데이터를 이해하고 인사이트를 발굴할 수 있는 보편적인 통계적 방법론 적용

예쁘지만 만질 수 없는 그림 속의 장미꽃

- 팬시한 딥러닝 사례는 만지만, 현실속에서 활용한 사례는 제한적
- 만질 수 있는 안개꽃이 현실적 (데이터 한계, 리소스 한계, 분석 목적에 부합하는 방법)

대중을 위한 보편적인 기능에 충실하자

- 엔터프라이즈 시장이 대상이 아닌 아카데미, 소규모 연구 및 분석 조직 타겟팅
- Digital Transformation 시대에 소외된 계층 지원
- 파레토 법칙 → 80%는 통계적 방법론에 부합하고, 20%가 머신러닝/딥러닝이 필요?

Win-Back을 기대하며

- 오픈 통계분석 시스템 사용자가 R 사용자로 전향하는 사례 기대
- R은 목적이 아닌 수단이지만, 그래도 이왕이면 R 사용자

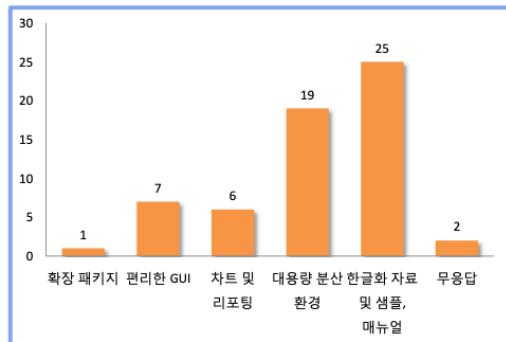
엔터프라이즈를 위한 R 환경 개선

다시 꺼낸 Survey - 대중화의 걸림돌

R에서의 지원 필요 요소

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

Survey 결과

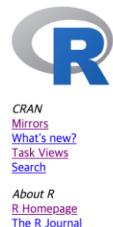


- 총 60 응답건수 중 한글화 자료 및 샘플, 매뉴얼 제공을 필요로 한다는 대답이 25건 (41.67%)으로 가장 높았음
- 또한 대용량 분산 환경 또한 19건으로 높게 나타남

- 엔터프라이즈의 R 진입 장벽은 무엇인가?
 - 대용량 분산 환경 니즈 → 대용량 데이터처리의 성능 구현 사례
 - 한글화 자료, 샘플, 매뉴얼의 니즈 → 한글화된 가이드 제공
- 오늘 이야기할 두번째 주제 - 빅 픽쳐, 욕심 내보기
 - "All-in-One R 데이터분석 방법론을 만들어 보자."
 - CRISP-DM, SEMMA, Tidiverse 데이터과학 프로세스 접목

대용량 데이터 처리의 니즈

- 10년 전의 질문 - **R로 대용량 데이터 분석이 가능한가요?**
- CRAN Task Views
 - High-Performance and Parallel Computing with R



CRAN Task View: High-Performance and Parallel Computing with R

Maintainer: Dirk Eddelbuettel

Contact: [Dirk.Eddelbuettel at R-project.org](mailto:Dirk.Eddelbuettel@R-project.org)

Version: 2021-11-08

URL: <https://CRAN.R-project.org/view=HighPerformanceComputing>

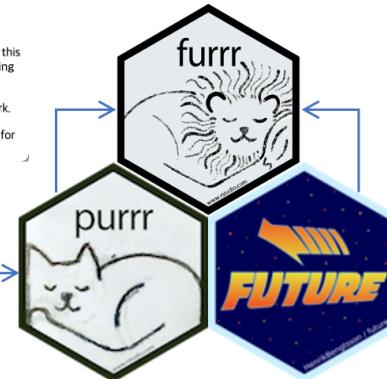
This CRAN task view contains a list of packages, grouped by topic, that are useful for high-performance computing (HPC) with R. In this context, we are defining 'high-performance computing' rather loosely as just about anything related to pushing R a little further: using compiled code, parallel computing (in both explicit and implicit modes), working with large objects as well as profiling.

Unless otherwise mentioned, all packages presented with hyperlinks are available from CRAN, the Comprehensive R Archive Network.

Several of the areas discussed in this Task View are undergoing rapid change. Please send suggestions for additions and extensions for this task view to the [task view maintainer](#).



future + purrr



functional
program

parallel/Distributed
processing

- 대용량 처리 관련, 몇개의 다건 데이터 처리 관련 코드를 공유하고자 합니다.

대용량 데이터 처리를 위한 솔루션 - 예제

준비

for loop apply 계열 purrr furrr

```
# 테스트용 데이터 생성
distribution <- tibble::tibble(
  uniform = runif(4),
  normal = rnorm(4),
  student_t = rt(4, df = 3)
))
```

```
# A tibble: 4 x 3
  uniform   normal student_t
  <dbl>     <dbl>     <dbl>
1 0.0852 -0.0929    -0.203
2 0.508   -0.216     0.726
3 0.618    0.575    -0.745
4 0.764   -0.145    -0.651
```

```
# 사용자 정의 함수
minmax <- function(x) {
  (x - min(x)) / diff(range(x))
```

대용량 데이터 처리를 위한 솔루션 - 사례

- 수십 GB JSON 파일을 tibble 객체로 불러오는 사례

```
parsing_log <- seq(nrow(ga_record)) %>%
  future_map_dfr(function(x) {
    sucess <- TRUE
    msg <- "OK"

    result <- try(ga_record[x, ] %>%
      pull %>%
      fromJSON)

    if (class(result) == "try-error") {
      msg <- result

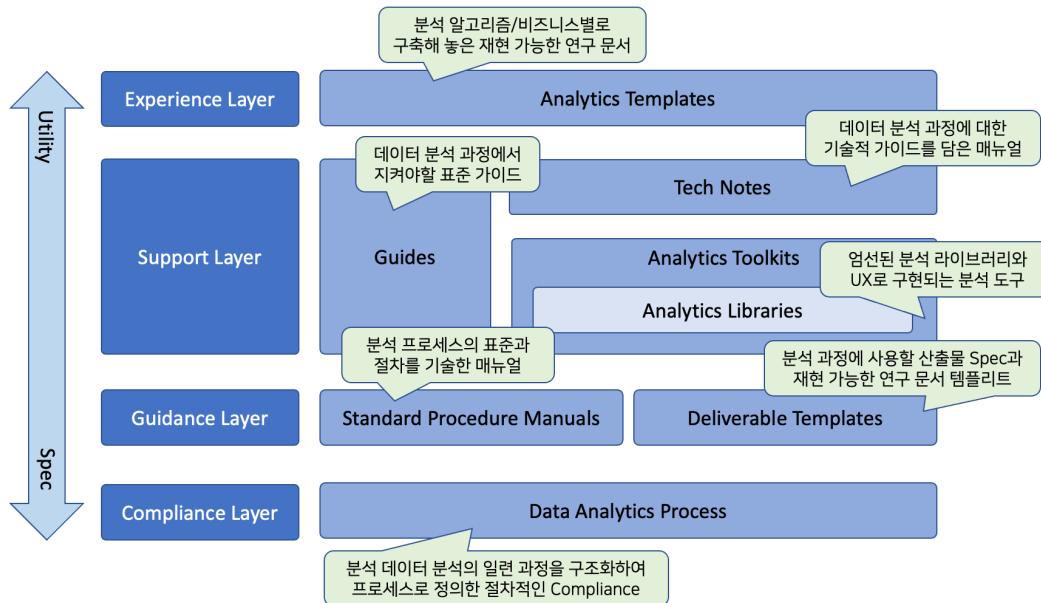
      result <- list("error")
      sucess <- FALSE
    }

    tibble(result = list(result), sucess = sucess, msg = msg)
  }, .progress = TRUE)
```

문제는 연장이 아니다

- 다건 처리의 진화, 당신의 위치는 어디인가요?
 - loop → apply → purrr → furrr
- 의외로 분석가는 습득한 기술만 고집한다
 - 빠른 신기술 개발 주기
 - 분화되고 전문화되는 데이터 분석 기술
- R에 반젤리스트 필요성
 - 새로운 유용한 R 기술이 나왔을 경우, 새로 창조된 가치를 R 분석가에 전파
 - R 커뮤니티의 역할
- 자율적 선택으로는 모자라!!, 동기를 부여하자.
 - 데이터분석 방법론에 R 테크놀로지를 접목하자.
 - 그러나, 방법론의 Compliance 요소 보다는 효용성을 전파하자.
- R Data Analytics Methodology
 - 데이터 분석 방법론
 - Reproducible Research 기반의 데이터 분석 경험 공유
 - 아직은 함께 만들어가야 할 방린이 (방법론 + 어린이)

R Data Analytics Methodology



- RDAM(R Data Analytics Methodology)
 - RDAM은 데이터 분석의 절차 및 방법의 표준을 제시하고, 분석의 생산성과 품질 향상을 위한 여러 콤포넌트로 구성된 R 기반의 데이터분석 방법론 Eco System
- RDAM 목적
 - 분석가의 경험과 무관하게, 만족할 수준 이상의 분석 결과를 얻을 수 있도록 노력
- RDAM 구성
 - 4개 layers와 8개의 components로 구성되며, 각각의 layers와 components는 상호 유기적으로 결합되어 운용

RDAM 컴포넌트 목록

8개 컴포넌트의 목록 및 기능

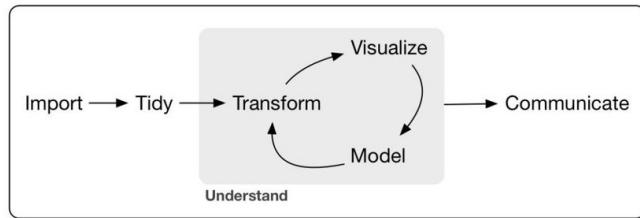
| Layers | Components | 기능 |
|------------------|----------------------------|----------------------|
| Compliance Layer | Big Data Analytics Process | 데이터 분석 프로세스의 정의 |
| Guidance Layer | Standard Procedure Manulas | 데이터 분석의 표준과 절차 제시 |
| Guidance Layer | Deliverable Templates | 표준 산출물 템플릿 제공 |
| Support Layer | Guides | 데이터 분석 과정의 표준 가이드 제시 |
| Support Layer | Analytics Library | 데이터 분석 R 라이브러리 제공 |
| Support Layer | Analytics Toolkits | 데이터 분석 R 툴박스 제공 |
| Support Layer | Tech Notes | 데이터 분석 R 테크니컬 노트 제공 |
| Experience Layer | Analytics Templates | 데이터 분석 R 템플릿 제공 |

- 오늘 소개할 컴포넌트

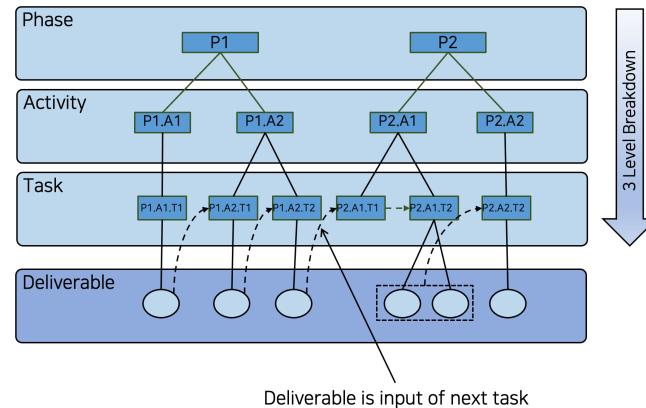
- Data Analytics Process
 - 데이터 분석 과정을 구조화하여 프로세스로 정의한 절차적 Compliance
- Standard Procedure Manuals
 - Data Analytics Process 개별 과정의 표준과 절차
- Guides
 - 데이터 분석 과정에서 수행하는 특정 작업의 표준을 위해 제정한 지침
- Analytics Library / Toolkit
 - 데이터 분석을 위한 R 패키지
 - 라이브러리를 자동화하거나 UX로 구성한 분석 자동화 툴
- Tech Notes
 - 데이터 분석 기법과 활용을 기술적으로 집대성한 R 테크니컬 가이드

Data Analytics Process

- 데이터 분석 과정을 구조화하여 프로세스로 정의한 절차적 Compliance



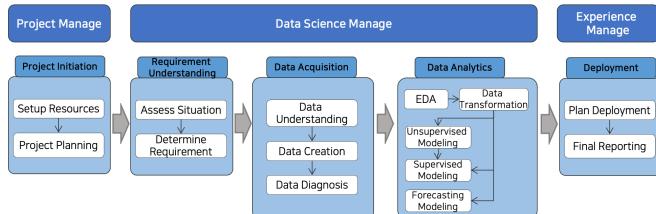
R for Data Science 분석 프로세스



RDAM에 분석 프로세스의 구조

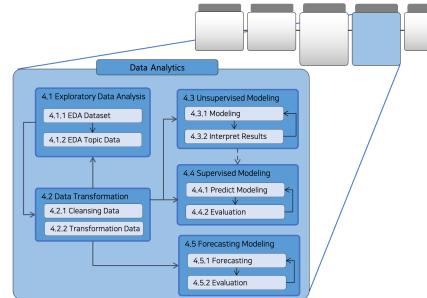
- RDAM에서는 데이터 분석 수행 단계뿐만 아니라,
 - 데이터 분석 프로젝트를 셋업하고,
 - 분석 결과를 시스템에 적용(DataOps)하는 프로젝트를 마무리 등 전 과정 정의
- 5 Phases, 12 Activities, 23 Tasks

Data Analytics Process



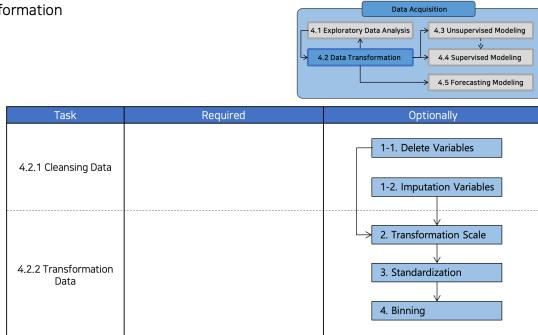
Phase 레벨 프로세스

□ 4. Data Analytics Process



Phase 레벨 프로세스 예시

□ 4.2 Data Transformation



Task 레벨 프로세스 예시

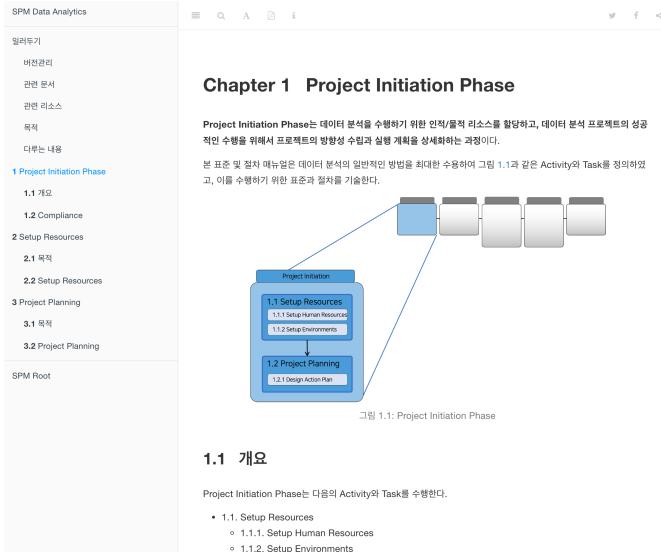
| Phase | Activity | Task | Input | Deliverable |
|-----------------------------|---------------------------|---------------------|-------------------------|--------------------------|
| Analytic Project initiation | Setup Human Resources | | | Assigned Human Resources |
| | Setup Environment | | | Project Setup |
| Requirement Understanding | Project Planning | Setup Environment | PMP, WBS, Kick-Off 발표자료 | |
| | Assess Situation | Business Process 평가 | 인력수급, 현황분석서 | |
| | Determine Requirement | 미래사업 분야 현황분석서 | 미래사업 분야 현황분석서 | 미래사업 분야 현황분석서 |
| Data Acquisition | Data Understanding | 현황분석서 | 현황분석서 | 현황분석서 |
| | Data Creation | 분석 데이터 정의 설정 | 조사사항 정의서 | 분석데이터정의서 |
| | Data Transformation | 분석 데이터 구조화 | Raw 데이터 | Raw 데이터 및 서비스 |
| | Data Diagnosis | Raw 데이터 | Raw 데이터 | Raw 데이터 및 서비스 |
| Data Analytics | Exploratory Data Analysis | Raw 데이터 및 서비스 | 데이터 및 서비스 | 데이터 및 서비스 |
| | Big Data Analytics | 설계 문서 | 설계 문서 | 설계 문서 |
| | Modeling | 설계 문서 | 설계 문서 | 설계 문서 |
| | Model Evaluation | 설계 문서 | 설계 문서 | 설계 문서 |
| | Data Analytics | 설계 문서 | 설계 문서 | 설계 문서 |
| | Big Data Aggregate | 설계 문서 | 설계 문서 | 설계 문서 |
| | Modeling | 설계 문서 | 설계 문서 | 설계 문서 |
| | Model Evaluation | 설계 문서 | 설계 문서 | 설계 문서 |
| Deployment | Plan Deployment | 설계 문서 | 설계 문서 | 설계 문서 |
| | Create Final Report | 설계 문서 | 설계 문서 | 설계 문서 |
| | Final Presentation | All | All | Final Presentation |

프로세스 목록

Standard Procedure Manuals

- Data Analytics Process 개별 과정의 표준과 절차를 정의한 표준 및 절차 매뉴얼

SPM Data Analytics



Chapter 1 Project Initiation Phase

Project Initiation Phase는 데이터 분석을 수행하기 위한 인적/물적 리소스를 활용하고, 데이터 분석 프로젝트의 성공적인 수행을 위해서 프로젝트의 방향성 수립과 실행 계획을 상세화하는 과정이다.

본 표준 및 절차 매뉴얼은 데이터 분석의 일반적인 방법을 최대한 수용하여 그림 1.1과 같은 Activity와 Task를 정의하였고, 이를 수행하기 위한 표준과 절차를 기술한다.

그림 1.1: Project Initiation Phase

```

graph TD
    PI[Project Initiation] --> SR[1.1 Setup Resources]
    SR --> SR1[1.1.1 Setup Human Resources]
    SR --> SR2[1.1.2 Setup Environments]
    SR1 --> PP[1.2 Project Planning]
    SR2 --> PP
    PP --> DAP[1.2.1 Design Action Plan]
  
```

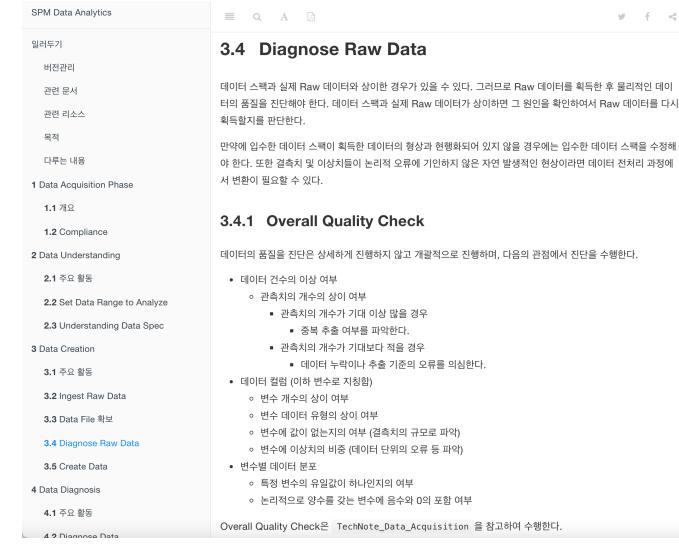
1.1 개요

Project Initiation Phase는 다음의 Activity와 Task를 수행한다.

- 1.1. Setup Resources
 - 1.1.1. Setup Human Resources
 - 1.1.2. Setup Environments

Project Initiation 스크린샷

SPM Data Analytics



3.4 Diagnose Raw Data

데이터 스펙과 실제 Raw 데이터와 상이한 경우가 있을 수 있다. 그러므로 Raw 데이터를 획득한 후 물리적인 데이터의 품질을 진단해야 한다. 데이터 스펙과 실제 Raw 데이터가 상이하면 그 원인을 확인하여서 Raw 데이터를 다시 획득할지를 판단한다.

만약에 입수한 데이터 스펙이 획득한 데이터의 형상과 현행화되어 있지 않을 경우에는 입수한 데이터 스펙을 수정해야 한다. 또한 결측치 및 이상치들이 논리적 오류에 기인하지 않은 자연 발생적인 현상이라면 데이터 전처리 과정에서 변환이 필요할 수 있다.

3.4.1 Overall Quality Check

데이터의 품질을 진단은 상세하게 진행하지 않고 개괄적으로 진행하며, 다음의 관점에서 진단을 수행한다.

- 데이터 간수의 이상 여부
 - 관측치의 개수가 기대 이상 많을 경우
 - 증폭 축출 여부를 파악한다.
 - 관측치의 개수가 기대보다 적을 경우
 - 데이터 누락이나 추출 기준의 오류를 의심한다.
- 데이터 칠坑 (어학 번수로 지칭함)
 - 번수 개수의 상이 여부
 - 번수 티파마터 유형의 상이 여부
 - 번수에 값이 있는지의 여부 (관측치의 규모로 파악)
 - 번수에 이상치의 비중 (데이터 단위의 오류 등 파악)
- 번수별 데이터 분포
 - 특정 번수의 유일값이 하나인지의 여부
 - 논리적으로 양수를 갖는 번수에 음수와 0의 포함 여부

Overall Quality Check은 TechNote_Data_Acquisition을 참고하여 수행한다.

Data Acquisition 스크린샷

- 5개 Phase별, 표준 및 절차 매뉴얼 개발
- bookdown으로 작성된 웹 문서로 배포하며, PDF 문서 파일로도 다운로드 가능함

Guides

- 데이터 분석 과정에서 수행하는 특정 작업의 표준을 위해 제정한 지침



R 코딩 가이드 북 스크린샷

- 분석서버 운영 가이드
 - RStudio Server 운영 가이드
 - Shiny Server 운영 가이드
- R 코딩 가이드
 - R 코딩 가이드북
 - R 코딩 템플리트
- 프로젝트 커뮤니케이션 가이드
 - Reproducible Research 이용한 분석결과 리뷰 템플리트

Analytics Library / Toolkit

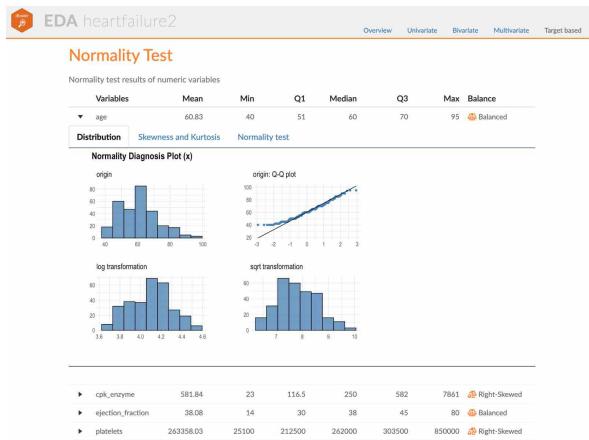
- 데이터 분석을 위해 엄선한 R 패키지와 자동화하거나 UX로 구성한 자동화 툴
- dlookr (<https://choonghyunryu.github.io/dlookr/>)
 - support Data Diagnosis, EDA, Data Transformation Activity
- alookr (<https://choonghyunryu.github.io/alookr/>)
 - support Supervised Modeling Activity
- mlookr (미개발)
 - support Plan Deployment Activity



dlookr / alookr

dlookr

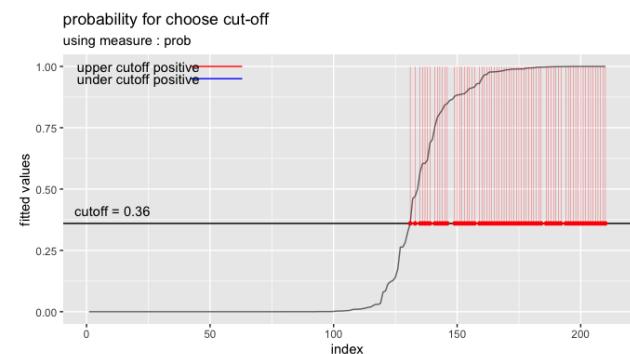
- 데이터 품질 진단
- 탐색적 데이터 분석
- 변수변환, 결측치 및 이상치 처리
- 자동화된 보고서 3종
 - 웹 보고서, PDF 보고서
- data.frame, DBMS의 테이블 지원



dlookr 스크린샷

alookr

- 모델용 데이터 분할 및 정제
- 대표 이진분류 모델 자동 적합
- 모델 성능 평가 및 최적 모델 선택
- 이진 분류 전 과정 지원
 - auto-ML
- h2o 및 python 모델 수용 계획



alookr 스크린샷

Tech Notes

- 데이터 분석 기법과 활용을 기술적으로 집대성한 **R 테크니컬 가이드**

다른 내용

작성한 경기

1 모색(Summary)

1.1 주제

1.2 주제

1.3 단행본 번수

1.4 단행본은 강정

2 학습(Discovery)

2.1 개념

2.2 이전에 번수

2.3 다른번 번수

2.4 단행 번수 단행

2.5 target 번수와 predictor 번수와의 관계

2.5.1 target 번수와 predictor 번수와의 EDA

2.5.2 (아이디)

2.5.3 target 번수 정리

2.5.4 target 번수와 밀접한 번수인 경우

2.5.5 target 번수가 소수인 경우 예상되는 ...

3 EDA toolkit

3.1EDA toolkit 사용법

3.2EDA toolkit 결과 보기

3.3EDA toolkit 결과 해석

3.4 dooder 폐지자료를 사용하는 방법

References

데이터와 그 상황모형의 예측값 사이의 관계를 산정도로 나타낸 것이다. 두 변수 사이에 선형 관계가 있는 경우 관측치의 설정 또는 빨간색 대각선에 수렴한다.

`> plot(num_nurs)`

Sales's scatter plot by Price

Predicted vs Observed (Sales)

5.5.2 예측변수가 범주형 번수인 경우

따음 예제는 ShelveLoc 가 target 변수 Sales 사이의 관계를 보여준다. 예측변수인 ShelveLoc 은 범주형 번수다. target = predictor 과정은 one-way ANOVA 를 수행한 결과를 보여준다. 결과는 분산분석의 관점에서 표현된다. sumerry() 정하는 관측변수와 각 계약에 대한 평균 총수 를 보여준다. 다시말해 target = predictor 관계의 단순 회귀식에 대한 성질 정리를 보여준다.

> # If the variable of interest is a categorical variable
> num_cat <- related(num_nurs, ShelveLoc)
> num_cat

Analysis of Variance Table

Response: Sales

Exploratory Data Analysis 스크린샷

일자리기
비자금관리
제작
다리는 내용
작성 환경

1 분류모델(Classification Model)

- 1.1 분류모델(classification)의 정의
- 1.2 분류모델의 데이터와 개념
- 1.3 분류모델의 기법
- 1.4 분류모델의 활용
- 1.5 분류모델 절차

2 데이터와 훈련(Split Dataset)

- 2.1 개요
- 2.2 training/test sets 분할
- 2.3 분할방법 데이터의 보정

3 logistic regression

- 3.1 개요

3.2 binomial logistic regression

4 tree model

- 4.1 개요
- 4.2 tree model의 이해
- 4.3 tree 제작기준 이용한 모델학습
- 4.4 rpart 제작기준 이용한 모델학습
- 4.5 party 제작기준 이용한 모델학습

col = "red")

```
# 예제로 성공자수가 최종값을 때 기본값을 출력합니다.
text(x = minCutoff,
     y = -0.5 + minCutoff),
label = stringr::str_c("cut-off:", minCutoff),
pos = 3)
```

최적의 분류 기준검 찾기

기준점 (cut-off)

misclassification rate (%)

cut-off:0.56

그림3.2: 데이터의 상관계수를 이용한 최적의 cut-off 찾기

train set으로 cut-off 구할수는 있지만, train set으로 적합한 모델을 train set으로 평가해서는 안된다. 왜냐하면, 모델을 쓰면서 데이터로 폐하한 모델의 능력을 볼수는 없기 때문이다. 그 이유는 해당 데이터에 반영되어서 즉 모델이 해당 데이터들의 특성이 반영되었기 때문이다. 그래서 모델을 평가하기 위해서는 모델에 사용하지 않은 데이터를 평가해야 한다. 앞서 준비한 test set이 바로 모델을 평가하기 위해서 준비한 데이터이다.

Supervised Modeling 스크린샷

- 7개 Activity별 테크니컬 노트 개발
 - bookdown으로 작성된 웹 문서로 배포하며, PDF 문서 파일로도 다운로드 가능함
 - 앞에서 다룬 R에서의 대용량 데이터 핸들링 방법 등이 테크니컬 노트에 수록

마무리

- R 커뮤니티의 역할 기대
 - 중급자를 위한 R 커뮤니티의 에반젤리스트화
 - 초보자를 위한 R 커뮤니티의 경험 전수
- 오픈 통계 분석기의 대중화
 - 오픈 통계 분석기의 성공적인 개발
 - SPSS/SAS 사용자의 대체제
- R 데이터 분석 방법론 전파
 - 데이터 분석가의 경험 수렴을 통한 개발과 개선
 - 데이터 분석의 효율성과 생산성 제고 기여
- Again wind in Korea with GNU R
 - 한국 R 생태계의 활성화 기대
 - 오픈 통계 분석기, 방법론의 협업 방법 구체화 계획

경청해 주셔서
감사합니다.

유충현

Tidyverse Korea

choonghyun.ryu@gmail.com



Tidyverse
Korea