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.

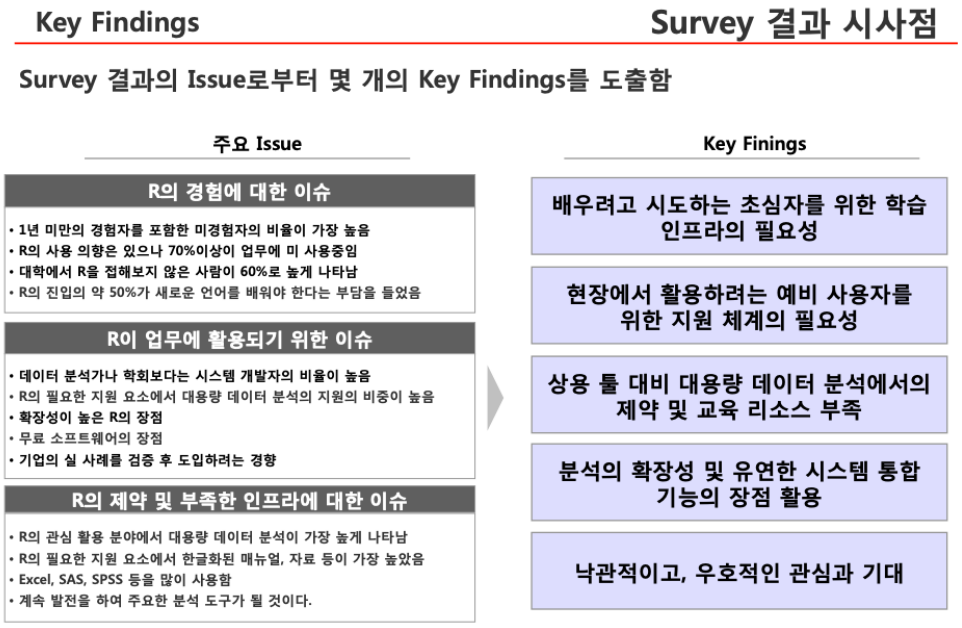
10년 전 컨퍼런스를 호스트한 R 사용자로서의 감회는, **데이터 분석 필드의 염원과 기대를 R의 대중화로 뿌리내리지 못한 것에 대한 아쉬움**입니다. 다시 R 사용자의 염원을 담아서 **R에 대한 관심과 대중화를 위한 몇 가지 시도를 소개**하면서, R 사용자의 관심을 요청합니다.

과거를 회상하며

New Wind in Korea with GNU R

- 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

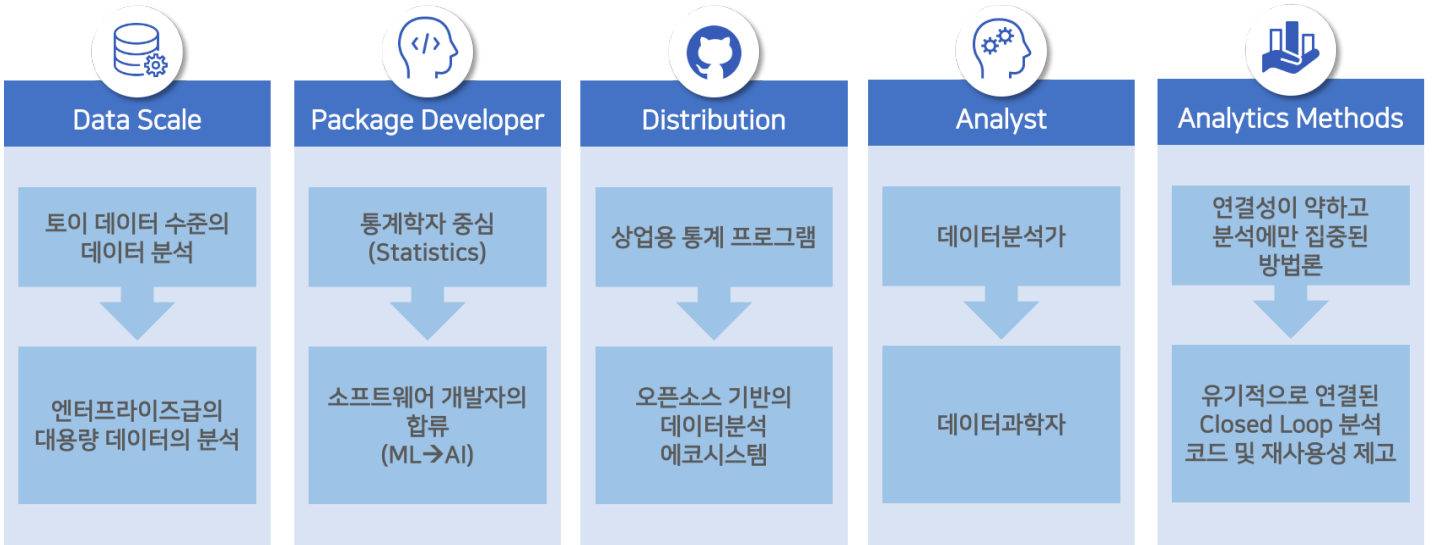
2011년도 컨퍼런스 설문지 분석



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 킬러 컨텐츠

데이터 분석 필드에서의 패러다임 전환



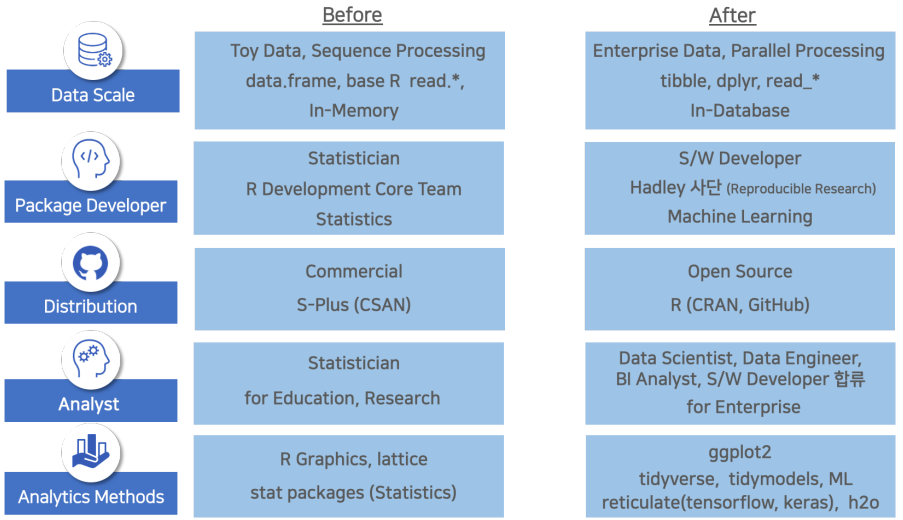
python의 대중화와 R의 소외감

- 공급부족 데이터 분석시장에 개발자 캐릭터 분석가의 진입 → 개발자 선호 툴
- CRM, 빅데이터, 거쳐 시장에서의 딥러닝의 붐업 → 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

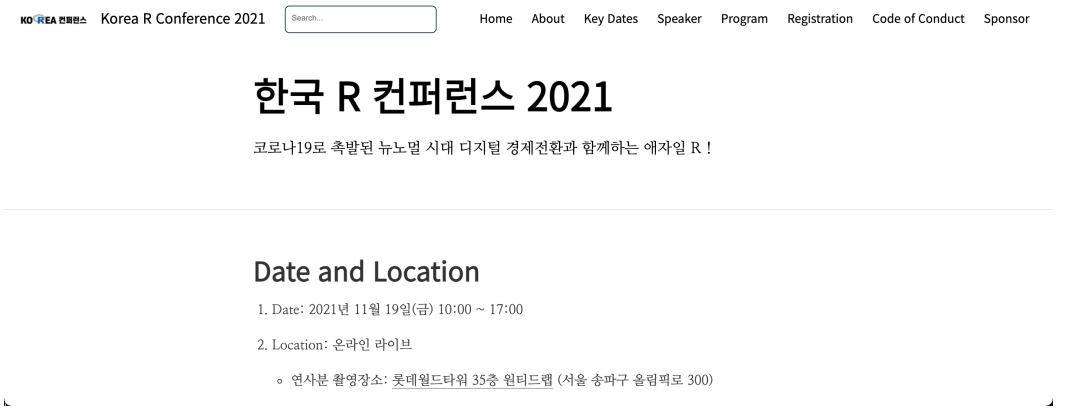
- 데이터 분석 경험의 공유 및 핸즈온 교육 용이

엔터프라이즈 시장에서의 활용 가능

- 부족했던 R 성능의 캐치업 → 수행속도, 데이터 핸들링 용량 등

미래를 설계하며

Again Wind in Korea with GNU R



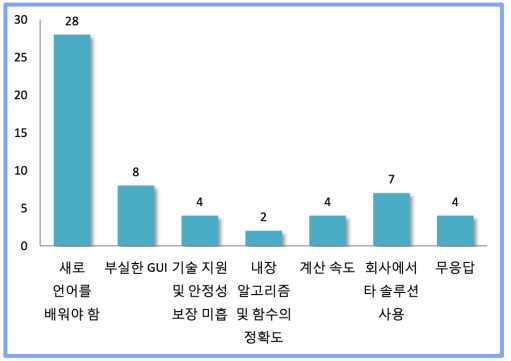
- "Again Wind in Korea with GNU R"
- 10년 전처럼, 오늘 행사가 **R에 대한 기대감 염원**에 불을 다시 지피기는 계기
- 개별 세션들을 통해서 시청자들이 많은 경험을 습득하는 컨퍼런스가 되길 기대
- 개인적으로는
  - 20여년 R 사용을 통해 꿈 꾸었던 빅 빅처를 소개하고,
  - 역동적인 한국의 R 생태계 조성을 위한 협력이 늘어나기를 기대

다시 꺼낸 Survey - 진입장벽

KOREA Conference

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

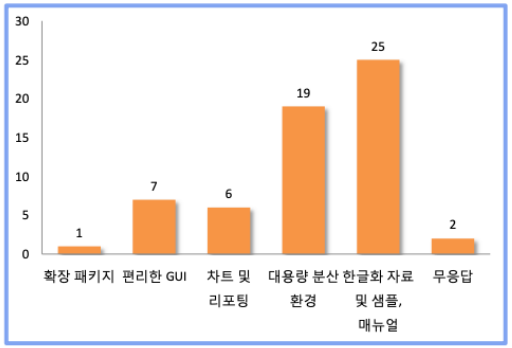
- 아카데미, 공공기관, 중소기업의 R 진입 장벽은 무엇인가?
  - **새로운 언어를 배워야 한다는 부담감**
  - 이미 익숙한 상용 데이터분석 소프트웨어를 사용함
    - 비용 절감 목적으로 접근하려는데, 취약한 R 교육훈련 인프라

- 오늘 이야기할 첫번째 주제 - 발상의 전환
  - "R이 아니라, R로 만들어진 데이터 분석 소프트웨어를 만들어 보자."

다시 꺼낸 Survey - 대중화의 걸림돌

R에서의 지원 필요 요소 **Survey 결과**

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

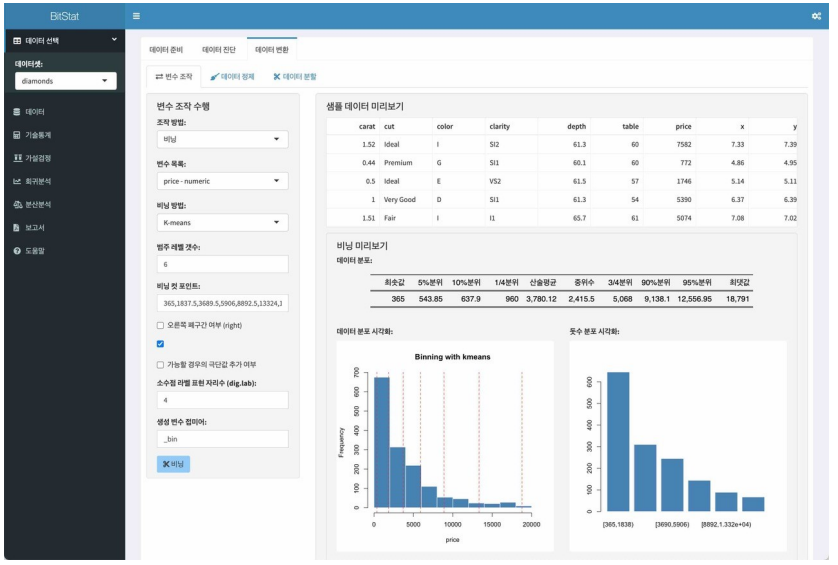


- **총 60 응답건수 중 한글화 자료 및 샘플, 매뉴얼 제공을 필요로 한다는 대답이 25건 (41.67%)으로 가장 높았음**
- **또한 대용량 분산 환경 또한 19건으로 높게 나타남**

- 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

아카데미를 위한 R 환경 개선

오픈소스 통계분석 시스템 개발

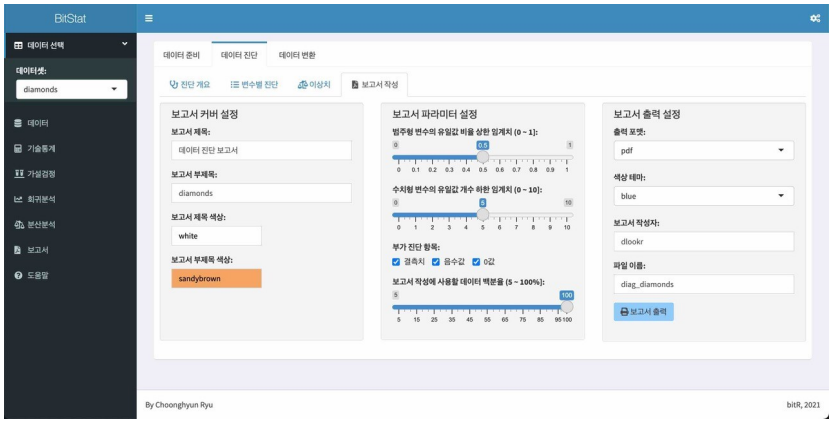


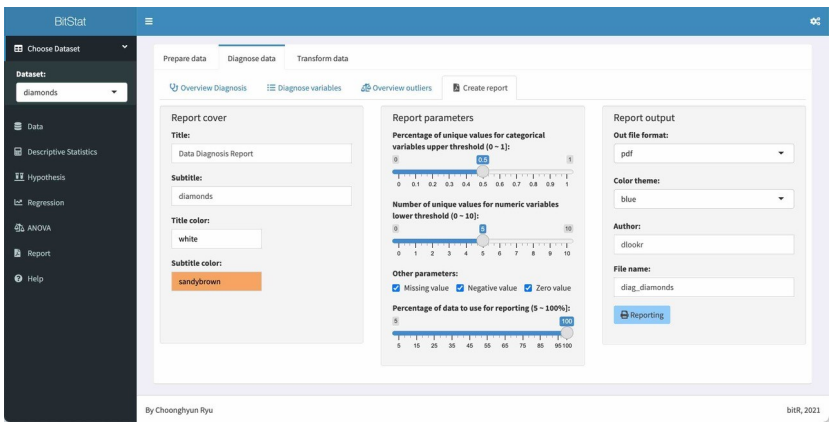
- 오픈소스 통계분석 시스템이란?
  - R/Shiny로 개발한 오픈소스 기반의 R 패키지, 현재 **프로토타이핑** 중
- 아카데미, 중소기업, 공공기관, 일반인 대상
  - 통계 비전공자들을 위한 데이터 분석 시스템
  - R을 모르더라도 사용가능, R 사용자는 더욱 쉽게 활용
- 기초통계 및 머신러닝 분석 기능 제공

오픈소스 통계분석 시스템 특징

다국어 지원(i18n)

- 국문과 영문 메뉴 및 메시지





왜 통계분석 시스템인가?

KOREA Conference

뿌리깊은 나무는 바람에 아니 뮐세

- 머신러닝, 딥 러링(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

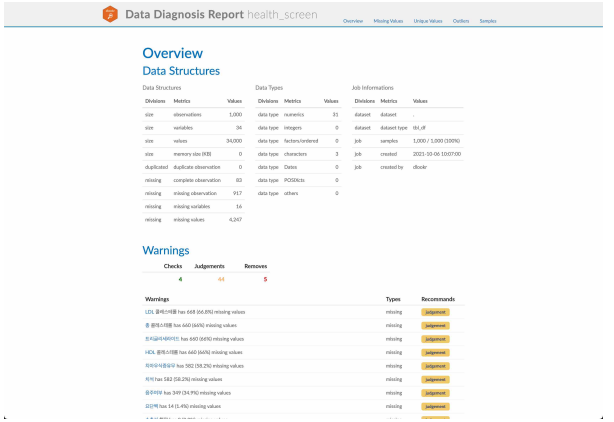
- 엔터프라이즈 시장이 대상이 아닌 아카데미, 소규모 연구 및 분석 조직 타켓팅
- Digital Transformation 시대에 소외된 계층 지원
- 파레토 법칙 → 80%는 통계적 방법론에 부합하고, 20%가 머신러닝/딥러닝이 필요?

Win-Back을 기대하며

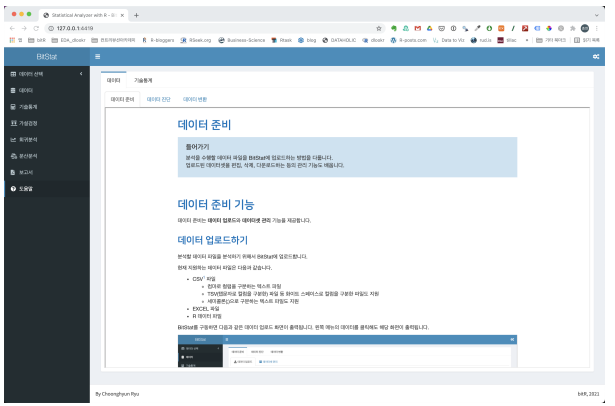
- 오픈 통계분석 시스템 사용자가 R 사용자로 전향하는 사례 기대
- R은 목적이 아닌 수단이지만, 그래도 이왕이면 R 사용자

오픈소스 통계분석 시스템 특징

데이터 분석 보고서 지원 → PDF 포맷 보고서와 HTML 포맷 보고서



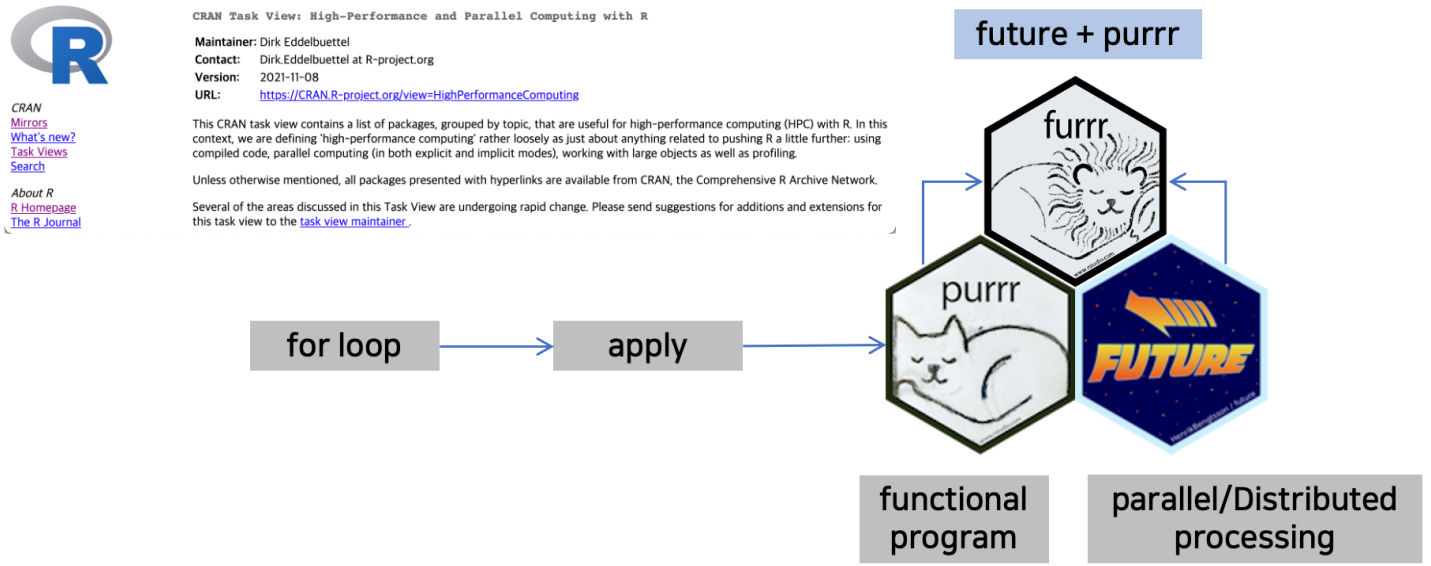
도움말 및 튜토리얼 지원



엔터프라이즈를 위한 R 환경 개선

대용량 데이터 처리의 니즈

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



- 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 0.542    1.30     -0.341
2 0.00124 -1.09      0.689
3 0.00862 -0.686     0.695
4 0.376   -0.907    -0.460
```

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

대용량 데이터 처리를 위한 솔루션 - 사례

KOREA Conference

- 수십 GB JSON 파일을 tibble 객체로 불러오는 사례

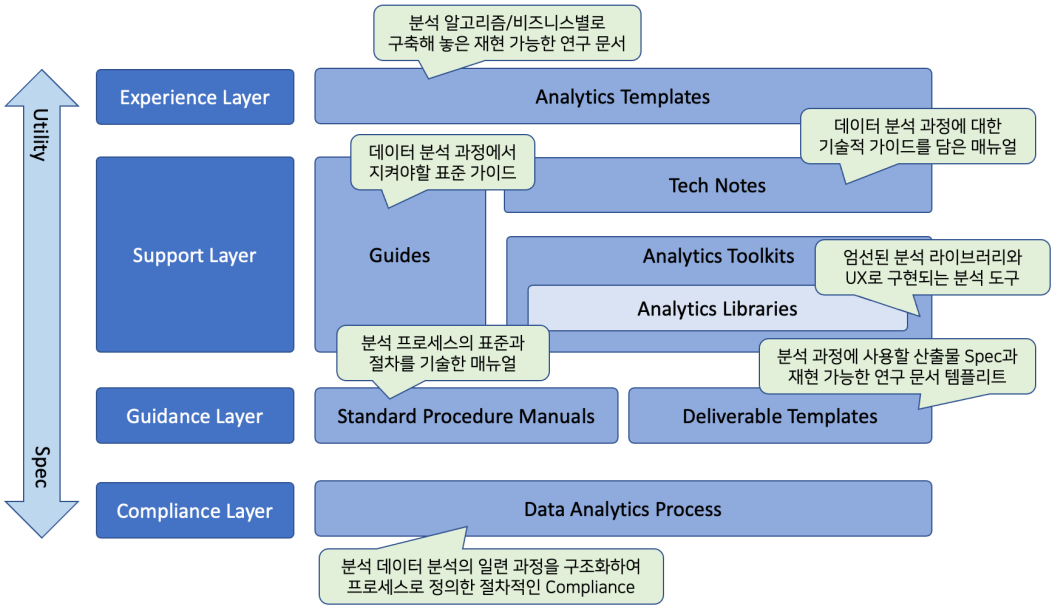
```
parsing_log <- seq(nrow(ga_record)) %>%
  future_map_dfr(function(x) {
    sucess <- TRUE
    msg <- "OK"

    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

KOREA Conference

- 다건 처리의 진화, 당신의 위치는 어디인가요?
  - loop → apply → purrr → furrr
- 의외로 분석가는 습득한 기술만 고집한다
  - 빠른 신기술 개발 주기
  - 분화되고 전문화되는 데이터 분석 기술
- R 에반젤리스트 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 만들어가야할 방린이 (방법론 + 어린이)

R Data Analytics Methodology



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

RDAM 컴포넌트 목록

KOREA Conference

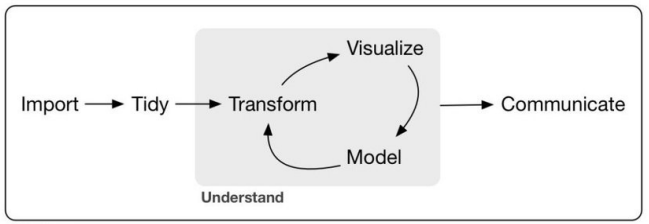
8개 컴포넌트의 목록 및 기능

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 템플리트 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 템플리트 제공

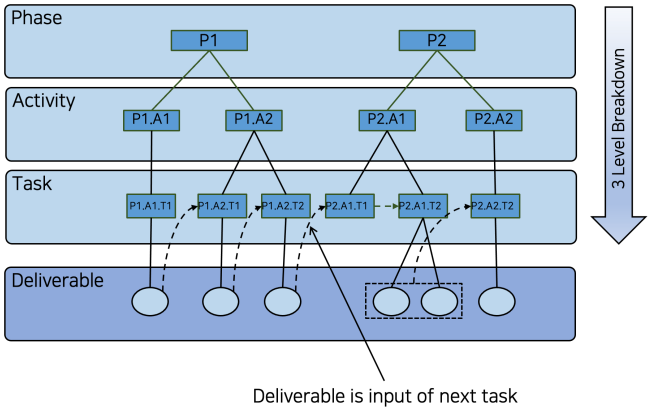
- 오늘 소개할 컴포넌트
  - Data Analytics Process
    - 데이터 분석 과정을 구조화하여 프로세스로 정의한 절차적 Compliance
  - Standard Procedure Manuals
    - Data Analytics Process 개별 과정의 표준과 절차
  - Guides
    - 데이터 분석 과정에서 수행하는 특정 작업의 표준을 위해 제정한 지침
  - Analytics Library / Toolkit
    - 데이터 분석을 위한 R 패키지
    - 라이브러리를 자동화하거나 UX로 구성한 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



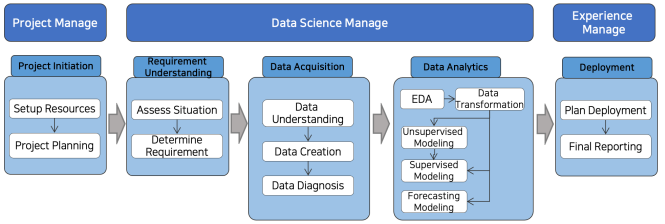
R for Data Science 분석 프로세스



RDAM에 분석 프로세스의 구조

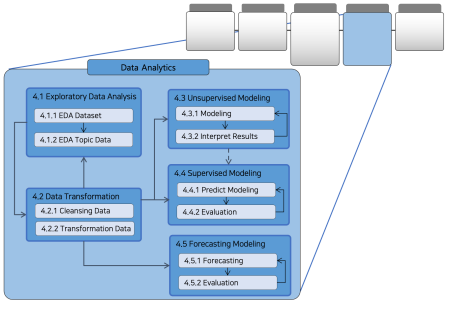
- RDAM에서는 데이터 분석 수행 단계분만 아니라,
  - 데이터 분석 프로젝트를 셋업하고,
  - 분석 결과를 시스템에 적용(DataOps)하는 프로젝트를 마무리 등 전 과정 정의
- 5 Phases, 12 Activities, 23 Tasks

Data Analytics Process



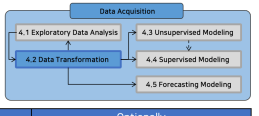
Phase 레벨 프로세스

☐ 4. Data Analytics Process



Phase 레벨 프로세스 예시

☐ 4.2 Data Transformation



Task	Required	Optionally
4.2.1 Cleansing Data		1-1. Delete Variables 1-2. Imputation Variables
4.2.2 Transformation Data		2. Transformation Scale 3. Standardization 4. Binning

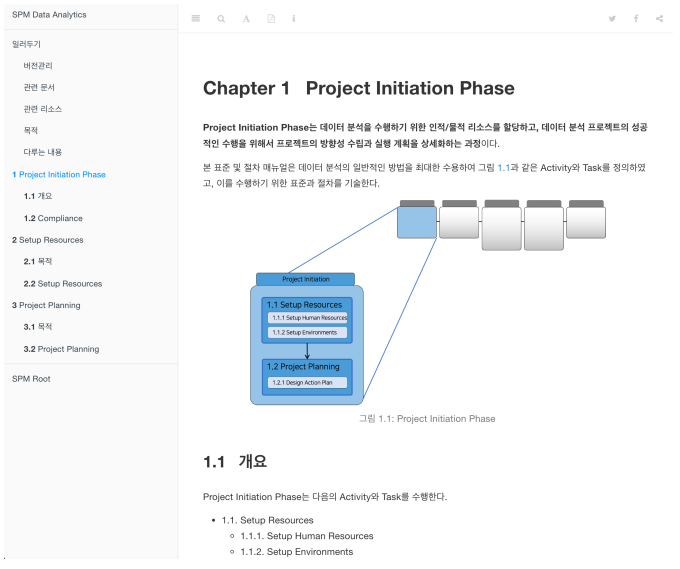
Task 레벨 프로세스 예시

Phase	Activity	Task	Input	Deliverable
Analytic Project initiation	Setup Resources	Setup Human Resources		Assigned Human Resources
		Setup Environment		Project Setup
	Project Planning	Design Action Plan		PMP, WBS, Kick-Off 발표자료
Requirement Understanding	Assess Situation	Business Process 분석		인터뷰결과, 현황분석서
		데이터시스템 현황분석		인터뷰결과, 현황분석서
	Determine Requirement	데이터분석 요구사항정의	현황분석서	요구사항정의서
Data Acquisition	Data Understanding	분석 데이터 범위 설정	요구사항정의서	분석데이터목록
		DataSpec Understanding	분석데이터목록	Raw 데이터 명세서
	Data Creation	Ingest Raw Data	분석데이터목록	Raw 데이터
		Diagnose Raw Data	Raw 데이터	데이터품질진단리포트
		Create Data	Raw 데이터 명세서	데이터, 데이터명세서
	Data Diagnosis	Diagnose Data	데이터	데이터품질진단리포트
		Cleansing Data	데이터품질진단리포트	정제된 데이터
Data Analytics	Exploratory Data Analysis	대상 데이터 셋 EDA	데이터명세서	데이터탐색보고서
		분석 주제 EDA	데이터명세서, 요구사항정의서	데이터탐색보고서
	Big Data Analytics	Modeling	데이터탐색보고서	모델정의서
		Model Evaluation	모델정의서	모델정의서, 분석결과 보고서
	Data Analytics	Big Data Aggregate	데이터탐색보고서	Aggregate Data Set
		Modeling	데이터탐색보고서, Aggregate Data Set	모델정의서
		Model Evaluation	모델정의서	모델정의서, 분석결과 보고서
Deployment	Plan Deployment	Design Deployment Plan	모델정의서, 분석결과보고서	전개계획서
	Final Report	Create Final Report	All	완료보고서
		Final Presentation		

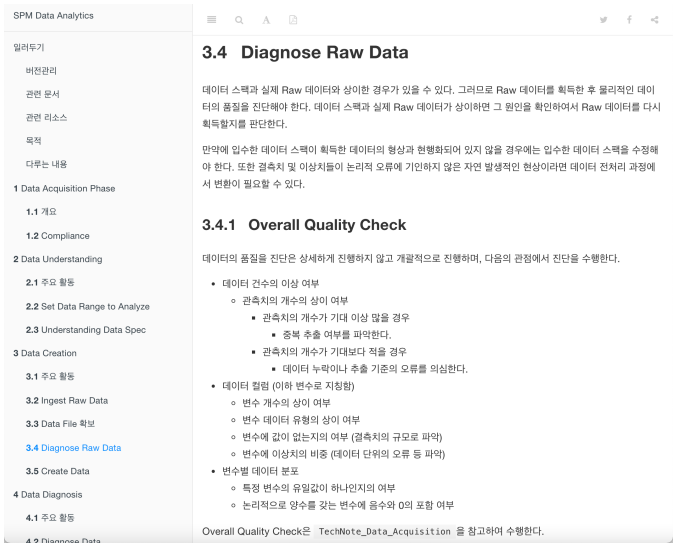
프로세스 목록

Standard Procedure Manuals

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



Project Initiation 스크린샷



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 스크린 샷

- 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

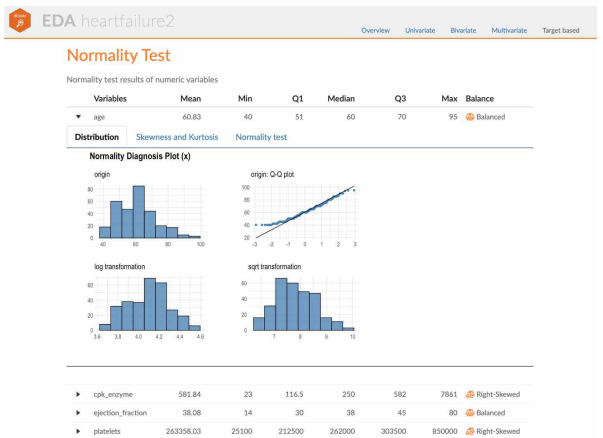
- 데이터 분석을 위해 엄선한 R 패키지와 **자동화하거나 UX로 구성한 자동화 툴**
- dlookr (https://choonghyunryu.github.io/dlookr/)
  - support **Data Diagnosis**, **EDA**, **Data Transformation** Activity
- alookr (https://choonghyunryu.github.io/alookr/)
  - support **Supervised Modeling** Activity
- mlookr (미개발)
  - support **Plan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

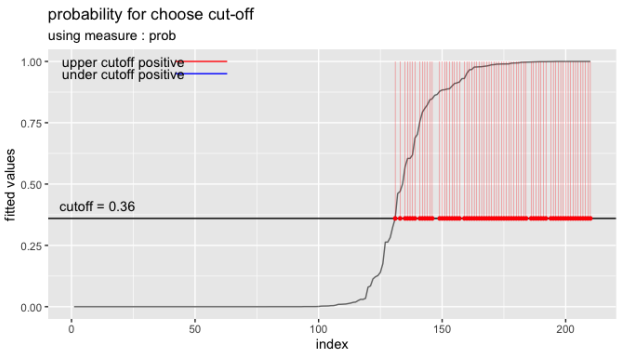


dlookr 스크린샷

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

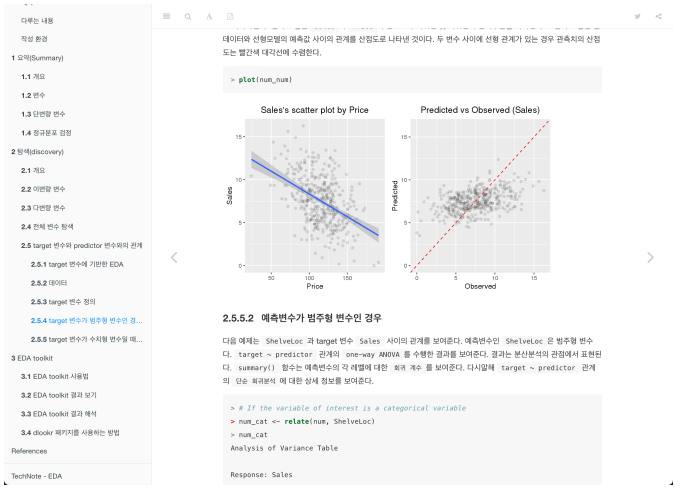
- h2o 및 python 모델 수용 계획



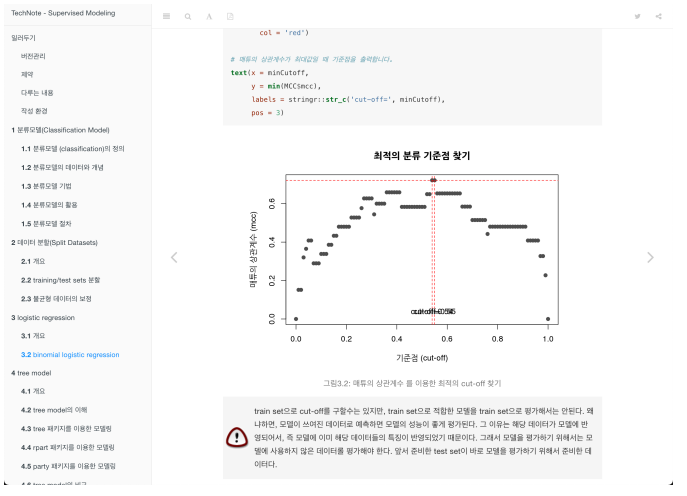
alookr 스크린샷

Tech Notes

- 데이터 분석 기법과 활용을 기술적으로 집대성한 R 테크니컬 가이드



Exploratory Data Analysis 스크린샷



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

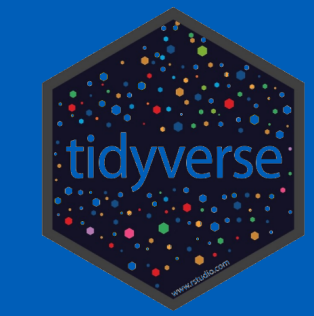
- R 커뮤니티의 역할 기대
  - 중급자를 위한 R 커뮤니티의 에반젤리스트화
  - 초보자를 위한 R 커뮤니티의 경험 전수
- 오픈 통계 분석기의 대중화
  - 오픈 통계 분석기의 성공적인 개발
  - SPSS/SAS 사용자의 대체제
- R 데이터 분석 방법론 전파
  - 데이터 분석가의 경험 수렴을 통한 개발과 개선
  - 데이터 분석의 효율성과 생산성 제고 기여
- Again wind in Korea with GNU R
  - 한국 R 생태계의 활성화 기대
  - 오픈 통계 분석기, 방법론의 협업 방법 구체화 계획

경청해 주셔서
감사합니다.

유충현

Tidyverse Korea

choonghyun.ryu@gmail.com



Tidyverse
Korea