

R 프로그래밍

(1주차)

2016. 03. 05(토)

장운호

(ADP 002-0004)

목차

※ 강사 소개

I. R 소개

II. R 프로그래밍

III. R의 장단점

IV. 강의 계획 소개

Wrap-up. 효과적인 R 학습법



※. 강사소개

데이터 분석 전문가 (ADP)

R을 활용한 실기시험을 거치는 국내 최초의 자격 시험으로,
2016년에 국가 공인 자격으로 전환되었음.



Statistical Computing / Data Science / Machine Learning



장 운 호

이사 / ADP (자격번호 002-0004)
국민대 겸직교수(R프로그래밍 강의)

[주]에스엔피솔루션

133-728 서울특별시 성동구 성수2로 118
성수아카데미타워 13층 1306호

Tel : 02-466-1218 Fax : 02-466-1219

Mobile : 010-8122-2201

Email : support@snpsolution.co.kr

unho.chang@gmail.com

<http://www.snpsolution.co.kr>

데이터분석 입문 계기

2007년경 요금청구서에 들어가 있는 광고 삽지(DM)의 효과를 어떻게 하면 정량적으로 측정할 수 있을까?를 고민하면서 데이터분석에 관심을 가지게 되었음.



“항상 ‘로그인’이 제일 어렵다.”



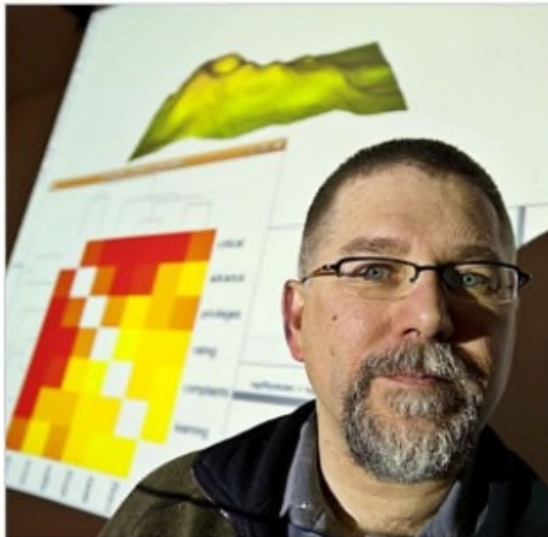


I . R 소개

1. R

통계 계산과 그래픽을 위한 **프로그래밍 언어이자 소프트웨어 환경이다.**
뉴질랜드 오클랜드 대학의 로스 이하카와 로버트 젠틀만에 의해 시작되어,
현재는 R 코어 팀이 개발하고 있음.

- R은 통계 소프트웨어 개발과 자료 분석에 널리 사용되고 있으며,
패키지 개발이 용이하여 통계학자들 사이에서 통계 소프트웨어 개발에
많이 쓰이고 있다. 위키피디아 R(프로그래밍)



Robert Gentleman

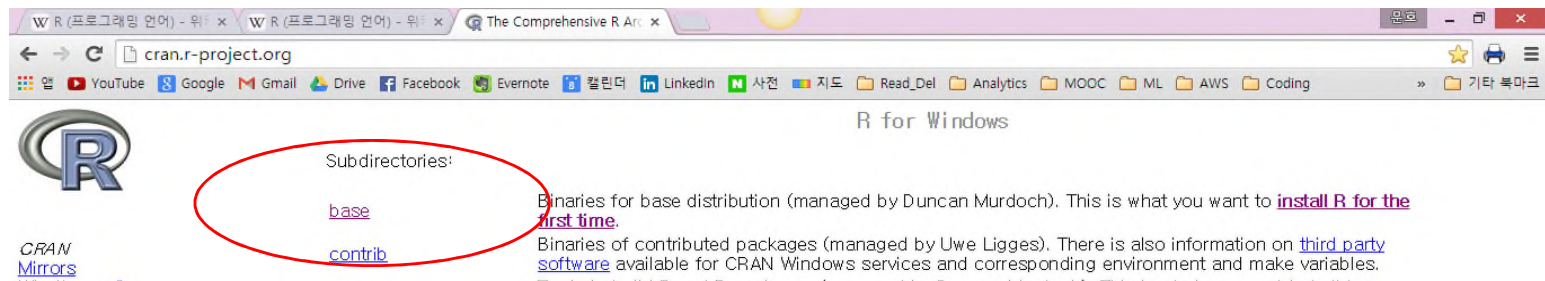
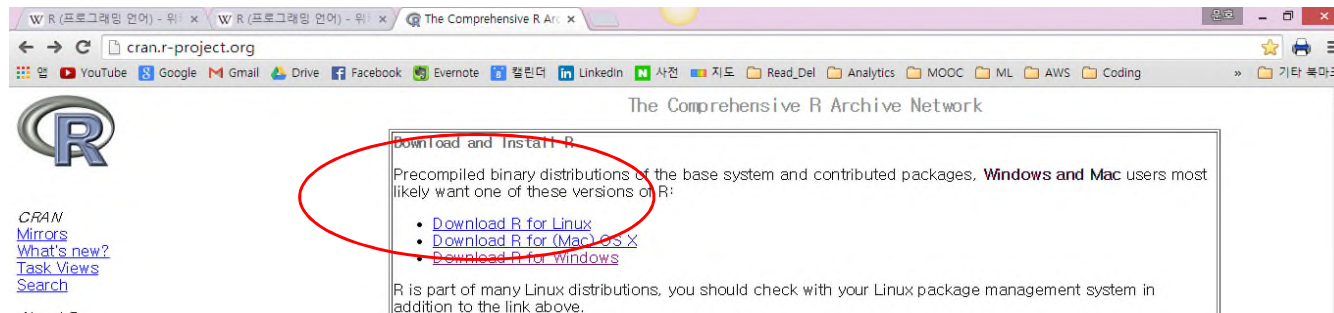


Ross Ihaka



2. R 설치

R은 오픈소스로 제공되어 무료로 설치 및 사용가능 (R 설치 : cran.r-project.org)



R-3.2.2 for Windows (32/64 bit)

[Download R 3.2.2 for Windows](#) (62 megabytes, 32/64 bit)

[Installation and other instructions](#)



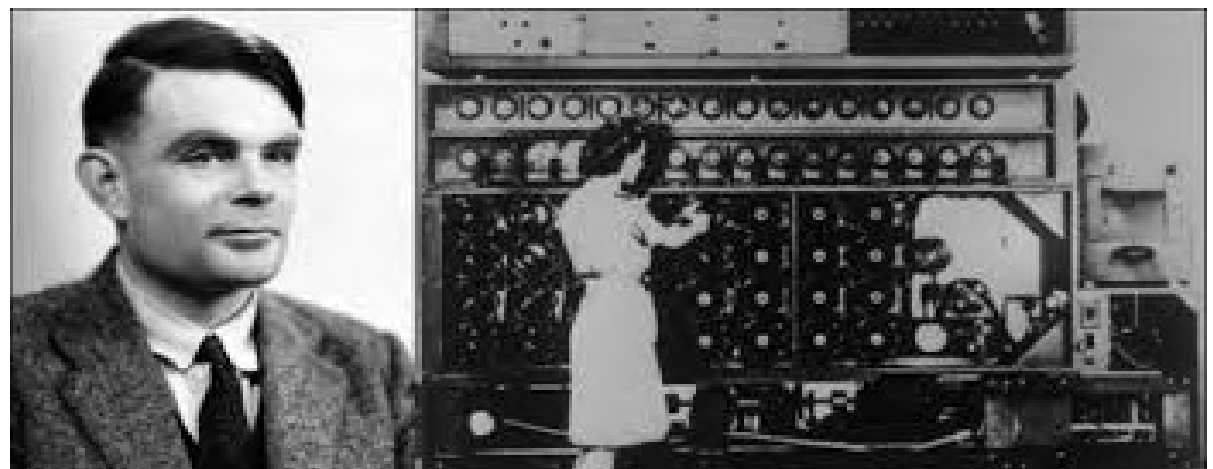
II. R 프로그래밍

1. 컴퓨터의 시작 (김현철, “데이터로 표현하는 세상 : 정보적 사고의 시작”)

앨런 튜링은 인간이 해결하는 어떤 문제라도

그 해결 과정을 기계가 처리할 수 있는 단위까지 아주 잘게 분해하여
순서대로 나열해 놓고,

그 기계에게 차례대로 처리하도록 할 수만 있다면,
인간의 문제를 자동으로 처리하는 기계를 설계할 수 있다고 생각하였음.



2. 알고리즘 (김현철, "데이터로 표현하는 세상 : 정보적 사고의 시작")

문제를 최대한 분해하여 조각들로 나눈 다음,
분해된 조각 들을 절차적인 순서대로 재구성함으로써,
순서대로 실행하면 문제가 해결될 수 있도록 하는 것을 **알고리즘**이라 칭함.

[일반적인 문제해결 과정]

① 문제 분석 및 표현

- 관련된 정보를 바탕으로 문제를 글, 표, 그림 등으로 표현하거나 문제를 잘게 나뉘서 세부단위로 분해하고 재구조화하는 것이 효과적임.

② 문제 해결방법 찾기

③ 실행

④ 평가

[컴퓨터에서의 문제해결 과정]

① 문제 분석 및 표현

- 문제를 해결하기 위해 알아야 하는 정보들과 요구사항을 분석하는 단계

② 문제 해결방법 찾기

③ 알고리즘 설계

- 문제 해결 절차나 방법을 알기 쉽게 절차적인 순서로 기술

④ 프로그램 작성

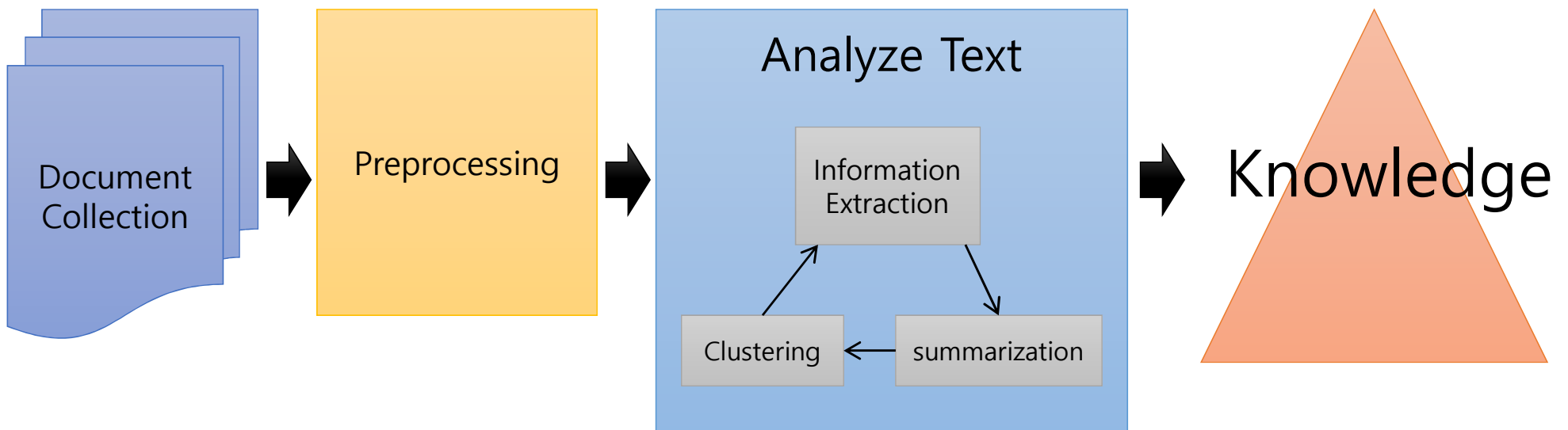
- 설계된 알고리즘을 바탕으로 컴퓨터에서 실행될 수 있도록 프로그래밍 언어로 Coding

⑤ 실행 및 수정 (debugging)

3. R 프로그래밍

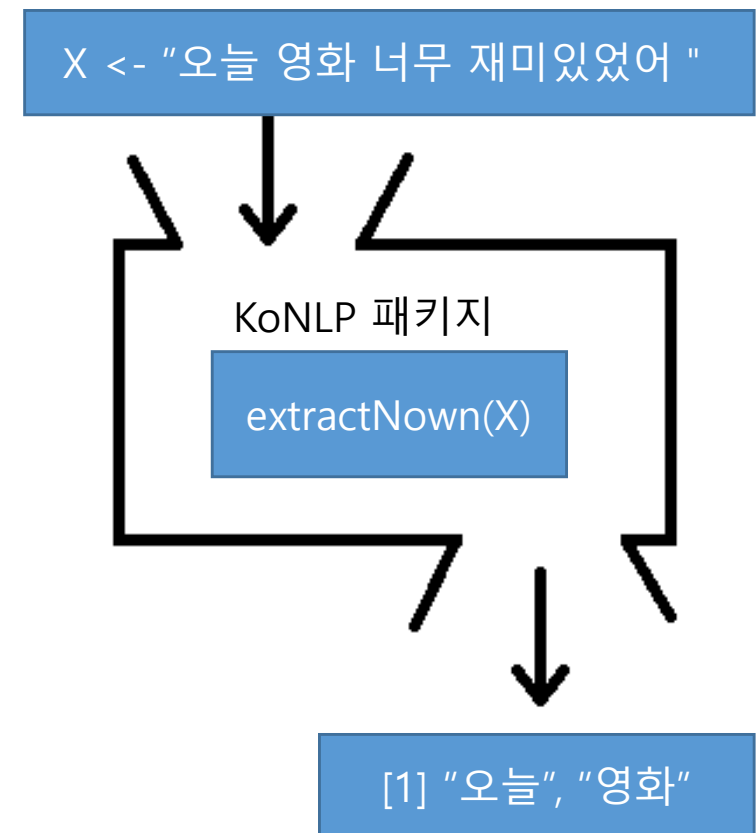
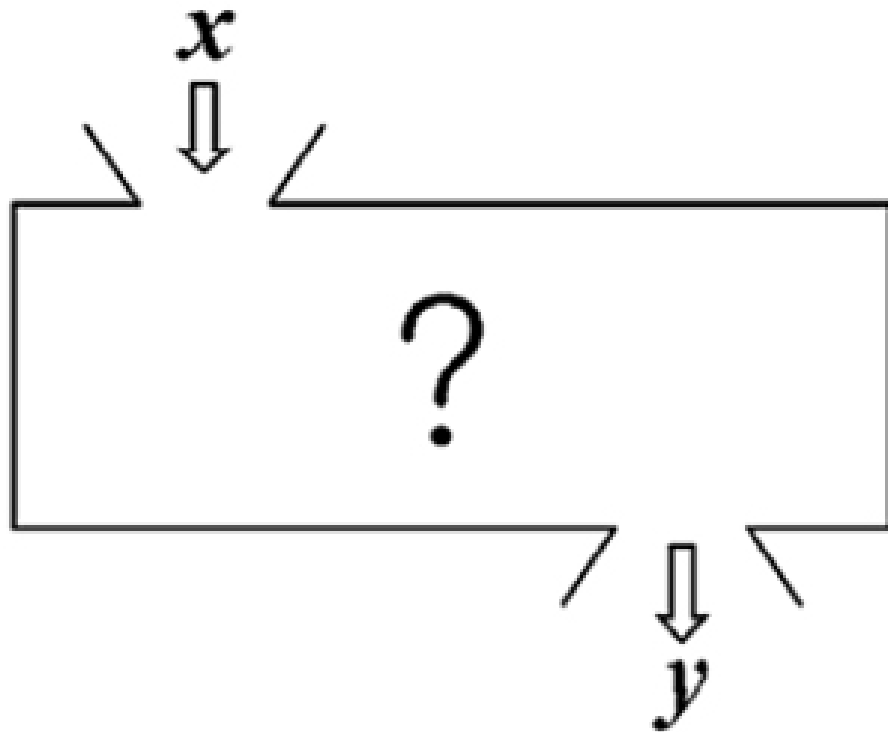
- ① 주어진 과제 혹은 당면한 문제를 해결하기 위하여 필요한 세부적인 절차를 정하고,
- ② 이를 **R이라는 언어**에서 정의된 명령어들을 사용하여,
- ③ 적절한 순서대로 이들 명령어를 나열함과 아울러 컴퓨터에 전달함으로써,
- ④ 컴퓨터가 그 나열된 순서대로 명령어들을 차례대로 수행하면,
- ⑤ 결국 최종적으로 그 과제, 혹은 문제가 해결되게 하는 것이 "**R 프로그래밍**"임.

[VoC 텍스트 분석(또는 텍스트 마이닝) 절차]



3. R 프로그래밍

R 프로그래밍의 핵심은 문제를 해결하는데 필요한 기능을 구현해 놓은 패키지를 정확히 파악하고, 이들 패키지에 내장되어 있는 함수들을 활용하여 적절한 분석이 이루어 질 수 있도록 **데이터와 함수들을 연결시켜주는** 것임.





Ⅲ. R의 장단점

1. R의 장점

전세계 모든 통계학 계열의 교수 및 학생들이 경쟁적으로 패키지를 만들어 올리고 있어서, 첨단환상적인 알고리즘들을 실시간으로 활용 가능.

Daniela M. Witten

[home](#) [people](#) [research](#) [software](#) [teaching](#) [news](#)



Daniela Witten
Associate Professor of Biostatistics and Statistics
University of Washington

Department of Biostatistics
F-649, Health Sciences Building, Box 357232
Seattle, WA 98195-7232

Department of Statistics
B-221, Padelford Hall, Box 354322
Seattle, WA 98195-4322

E-mail: dwitten@u.washington.edu






















Looking for a post-doc position? E-mail me your CV and the names of three references.

1. [참조 페이지](#)
2. [다이엘라 위튼 교수가 개발한 R패키지 소프트웨어 리스트 페이지](#)

1. R의 장점

전문 프로그래머 층에서도 점점 더 인기가 높아지고 있음.



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

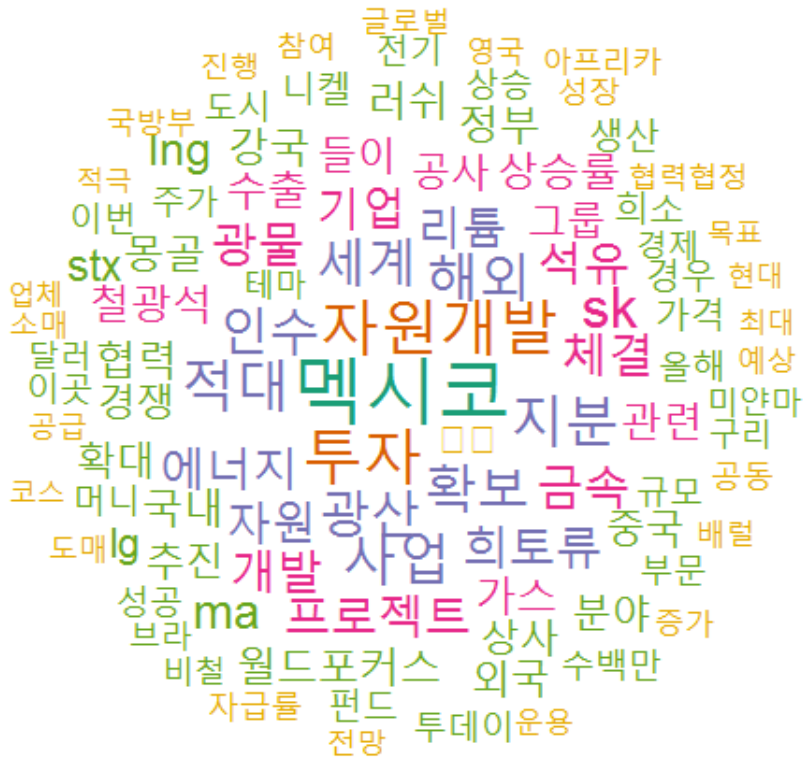
자료 : <http://spectrum.ieee.org/computing/software/the-2015-top-ten-programming-languages>

100

다양한 데이터의 Handling가능하고, 손쉽게 고품질의 그래픽 생성이 가능함.

[자원개발 : 2012년 이전 빈발단어 Word Cloud]

[자원개발 : 2014년 이후 빈발단어 Word Cloud]



자료) Radian6 (maintained by Salesforce.com)

1. R의 단점

다만, 가용한 패키지의 대부분이 해당 분야의 전문가들이 “전문가”들을 대상으로 만들었기 때문에 초보자들의 경우, 이를 배워서 활용하기가 쉽지 않음.

- “낮설다”는 (보이지 않는) 진입장벽이 있다고 볼 수 있으나, 극복이 불가능하지는 않음.

[R을 잘 다루기 위해서 미리 배워두면 도움이 될 선수 과목 리스트]

구분	컴퓨터 공학	통계학	수학
기본	1. 프로그래밍 개론 및 실습 2. 자료구조론 3. 패턴 인식	1. 통계학 개론 2. 회귀분석 3. 수리통계	1. 기초 Algebra 2. 미/적분(Calculus) 3. 선형대수
응용	1. 운영체제(리눅스/윈도우) 2. Hardware 3. Application Software	1. Supervised Learning 2. Unsupervised Learning 3. Reinforced Learning	1. 미분 방정식 (최적화) 2. 확률(과정)론 3. 수치해석 (Matlab 등 활용)



IV. 강의 계획

1. 강의 계획

목표는 R의 기본 자료 구조를 이해하고, R을 다양한 패키지를 활용하여 프로그래밍적으로 일정 수준의 문제를 해결 하실 수 있도록 해드리는 것임.

[강의 계획]

구분	강의	실습
3주차	1. R 프로그래밍 개념 이해 2. 객체/함수/변수 개념 이해 3. 연산자/기초내장함수	주간 e-mail 연습문제 (gmail 필수)
10주차	4. 5대 주요 자료구조 5. Data Indexing 6. Data Handling/Aggregation	워싱턴주 BikeSharing Dataset 활용, 장바구니 데이터 활용
16주차	7. dplyr / reshape 패키지 8. 정규 표현식 9. Visualization (ggplot2)	베스트셀러 크롤링 데이터 등을 활용한 실습

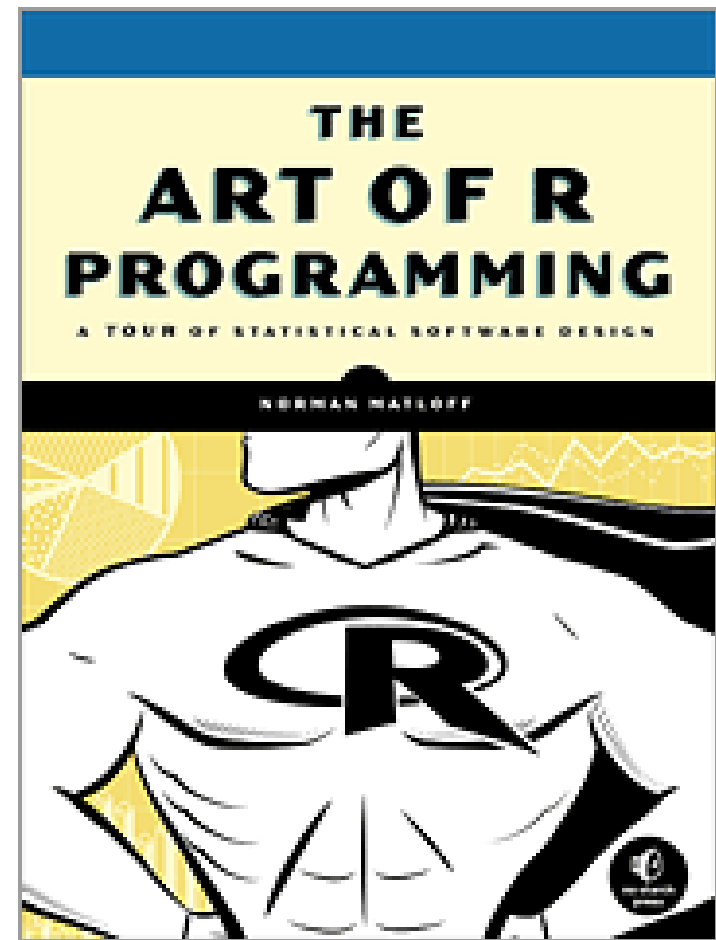
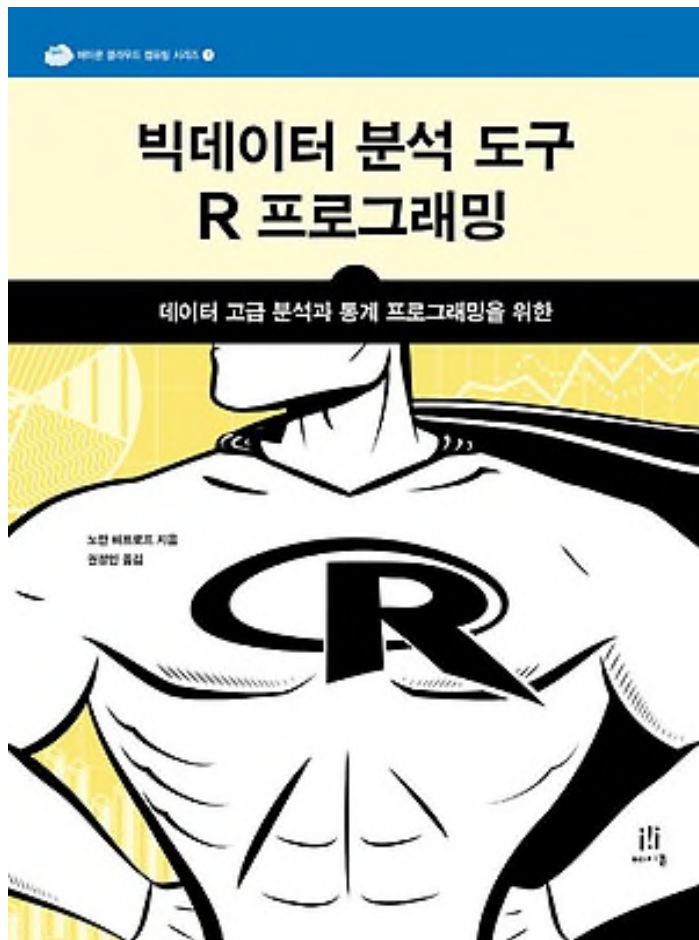
[평가 계획 및 비중]

평가 계획	비중
수업 참여도 30% (출석, Q&A참여 등)	A 90%이상 B 80%이상 C 70%이상 D 60%이상
Quiz/Report 40%	(Quiz/Report/test 중 최저점수 1회는 최종 평균 계산시 제외)
중간고사 15% 기말고사 15%	

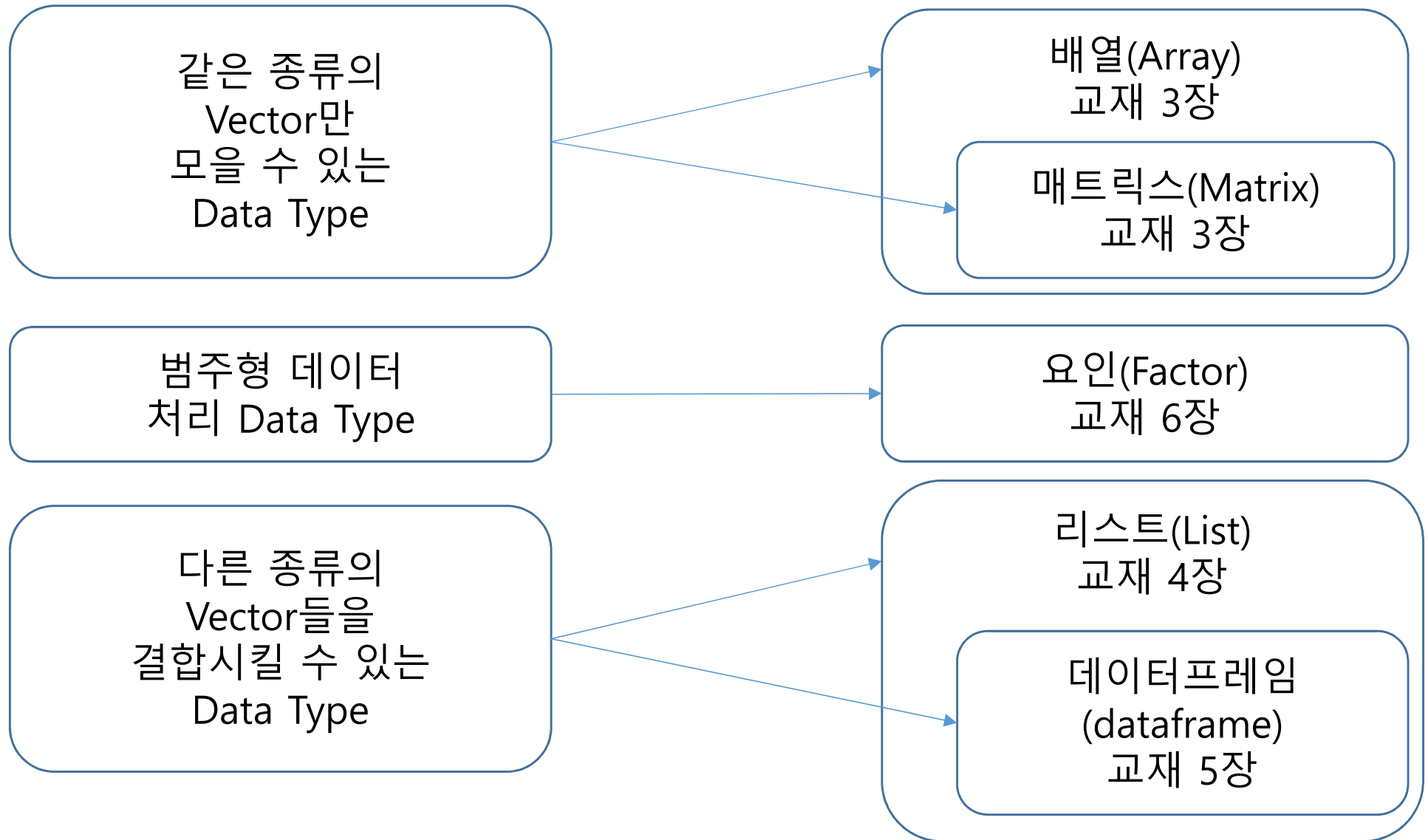
2. 교재 소개


R을 활용한 프로그램을 작성하는 방법을 집중적으로 훈련할 수 있는 교재 선정.

- 교재에 나와 있는 코드들이 쉬우면서도 프로그래밍 역량 개발에 도움이 되도록 구성되어 있음.



3. 교재 주요 내용





Wrap-up. 효과적인 R 학습법

의도적 수련 (김창준, "블로그 : 애자일 이야기")

내가 부족한 점을 찾아, 이를 중점적으로 보완하기 위한 자발적 학습을 일컬음.

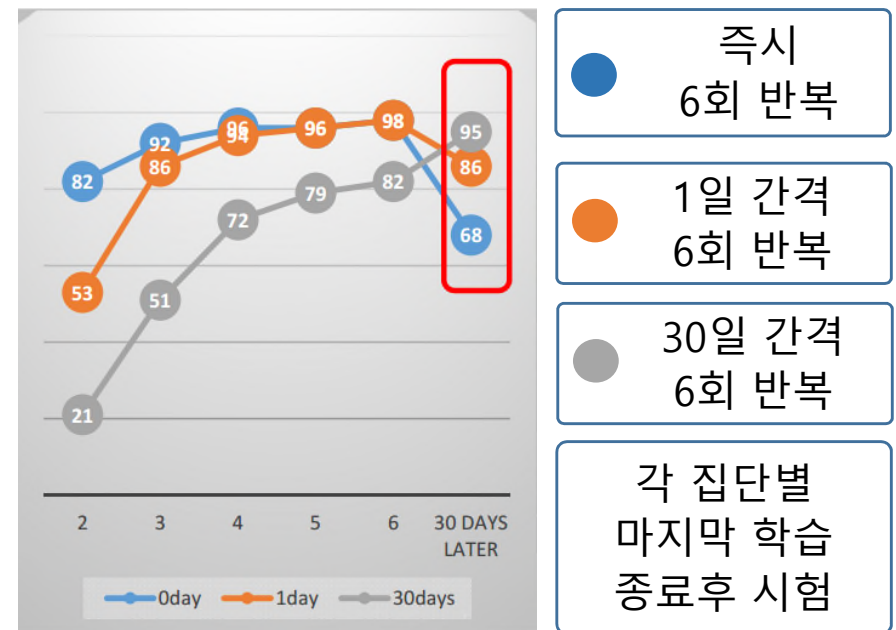
- 프로그래밍 능력 향상을 위한 의도적 수련에 필요한 전제조건

- ① 잘 정의된 작업, ② 적절한 난이도, ③ Feedback (Q&A), ④ 반복과 실수 교정의 기회

[학습법의 효용성(Utility) 평가 실험결과]

학습법	평가 결과
자세한 물음	보통
스스로 설명	보통
요약하기 / 밑줄 긋기 / 칠하기	낮음
교재를 마음에 그리기	낮음
다시 읽기	낮음
연습 문제 풀기	높음
분산 연습	높음

[분산연습]



자료: 서울대 심리학과 김정오 명예교수님 강의자료
에서 재인용 (2014, 청소년을 위한 심리학 교실)

자료: 경북대학교 교수학습센터

※ 데이터 분석 관련 8대 기회요인

1	데이터 분석업무는 더 이상 분석가의 업무에서 벗어나, 현업담당자가 직접 데이터를 분석하고, 의사결정하는 방향으로 진화할 것.	데이터분석 업무 지원 도구 필요
2	전문적인 데이터 사이언티스트 수요가 급증하고 있는데, 공급은 턱없이 모자라고 단기간에 수혈될 수 있는 성격도 아니라, 대책 필요	분석가 양성 서비스가 필요
3	기업 내부 데이터 분석 역량 부족으로 전문기업 외부의 활용이 늘어나고 있고, 유능한 데이터 분석가를 보유한 회사가 각광받음	효율적인 분석 아웃소싱 회사 필요
4	기업의 데이터 분석 업무 증가에 대응하는 시스템 비용 부담증가로 클라우드 기반의 분석서비스를 이용하려는 트렌드 활성화 예상	클라우드 기반의 분석 서비스 필요
5	데이터 분석이 고도화 되면서 단순 리포트가 아닌 실행가능한 분석 결과를 원하고, 각 버티컬 영역별로 특화된 분석 서비스 요구	Vertical 영역 전문 분석 서비스 필요 (SaaS)
6	IoT와 웨어러블 등의 센서데이터가 늘어날 것으로 전망되고, 이를 잘 분석하여 서비스화할 수 있는 분석역량이 요구됨.	센서데이터에 최적화된 분석 서비스가 필요
7	데이터 분석의 목적이 현황파악이나 가설 검증을 넘어, 미래를 예측하고 의사결정에 도움을 주는 방향으로 발전	예측분석을 통해 고객의사결정을 도와주는 서비스 필요
8	데이터 분석가 업무의 대부분이 단순데이터 전처리 작업으로 인해 낭비되고 있는데, 이를 해결하고자 하는 Needs가 강함.	데이터 전처리 작업 지원 도구 필요

자료) 한재선(전 KT NexR CEO)님 세미나 발표자료 : Data business 성공전략 (2014-06-17)

End of Document.

감사합니다.