



R 통계패키지 & 알고리즘 기반

빅 데이터 분석과 추론 & 예측

소 속 : 에이콘 아카데미

담 당 : 김 진 성



교육내용

1

PART-I. R 설치 및 기초 문법

2

PART-II. 탐색적 데이터 분석과 전처리

3

PART-III. 추론통계 분석

4

PART-IV. 예측분석(기계학습)



PART-I. R 설치 및 기초 문법

1. R 프로그램 개요 및 기초문법

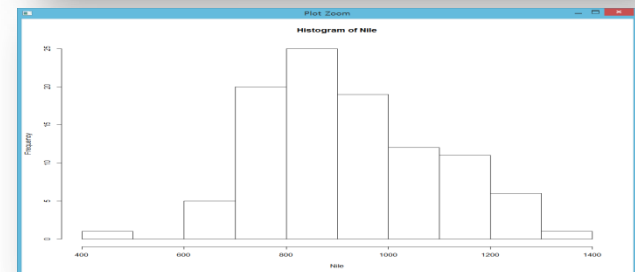
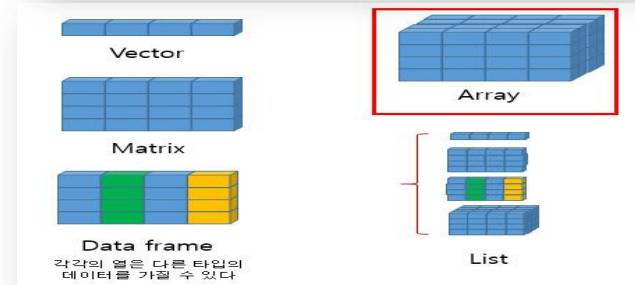
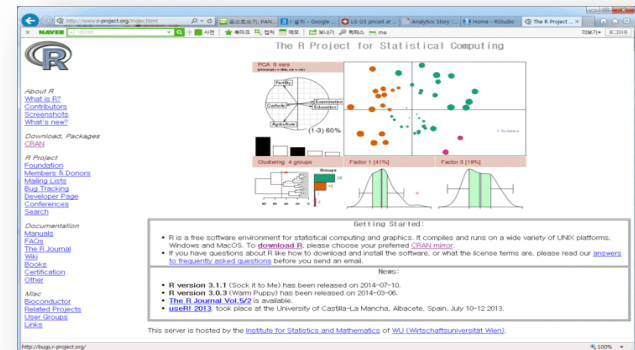
- ▶ R/R Studio 설치 및 기본메뉴 실습
- ▶ 패키지 설치 및 사용법

2. 데이터의 유형 및 자료구조 이해

3. 데이터 입출력 및 파일 처리

4. 제어문과 함수

- ▶ 제어문(조건문,반복문)
- ▶ 사용자 함수와 내장 함수





PART-II. 탐색적 데이터 분석과 전처리

5. 데이터 분석을 위한 시각화

▶ 이산변수와 연속변수 시각화

6. 데이터 조작(핸들링)

7. 데이터 분석을 위한 전처리

▶ 데이터 특성 분석

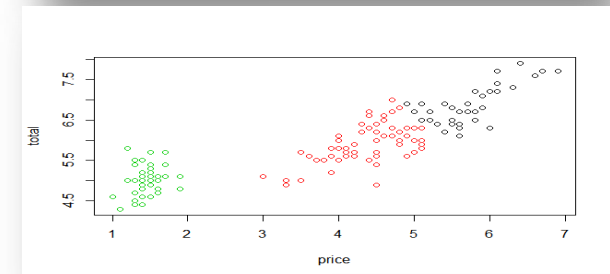
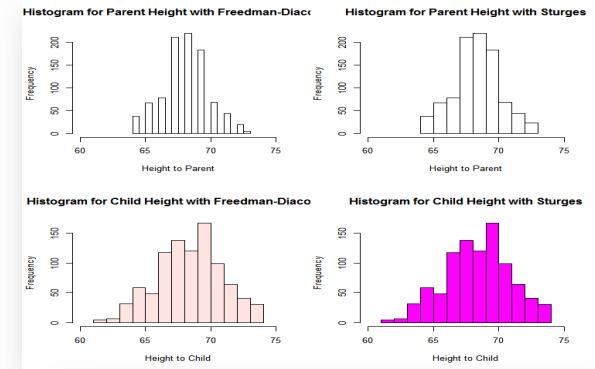
▶ 탐색적 분석을 위한 시각화

8. 고급 시각화 분석

9. 정형 데이터와 비정형 데이터 처리

▶ 정형 데이터 처리(SQL 데이터 처리)

▶ 비정형 데이터 처리(워드 클라우드)





PART-III. 추론통계 분석

10. 분석절차와 통계지식

11. 기술통계분석

12. 교차분석과 Chi-square 분석

▶ 교차분석과 교차표 작성

▶ chi-square 분석 및 검정

13. 집단 간 차이 분석

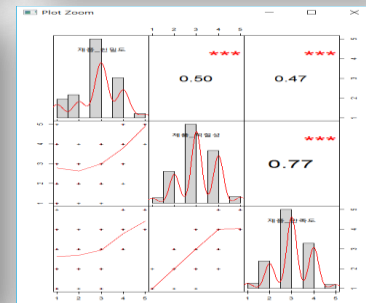
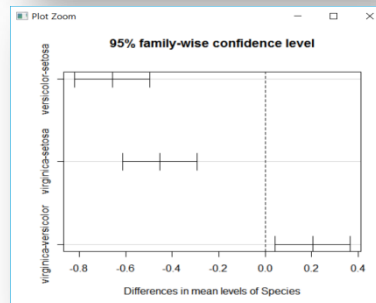
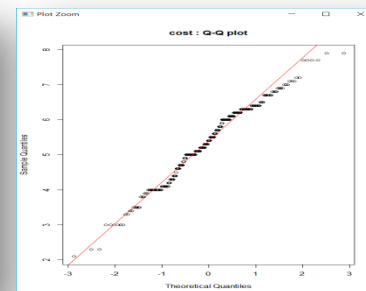
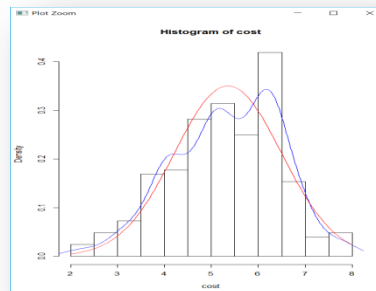
▶ 집단별 평균 차이 검정(T-test, Anova)

14. 요인분석과 상관관계 분석

▶ 요인분석, 변수정제

▶ 피어슨의 상관계수 r

학력수준		실패	진학	X-squared	유의확률(p)
고졸	관찰빈도 기대빈도	40 36	49 54	2.766951	0.2507057
대졸	관찰빈도 기대빈도	27 33	55 49		
대학원졸	관찰빈도 기대빈도	23 21	31 32		





PART-IV. 예측분석(지도/비지도 학습)

15. 지도학습(Supervised Learning)

▶ 회귀분석(Regression Analysis)

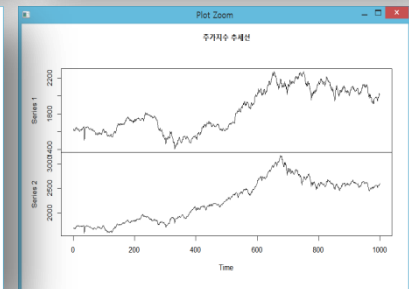
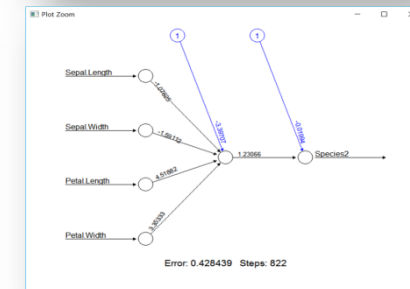
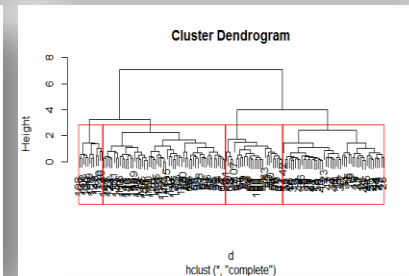
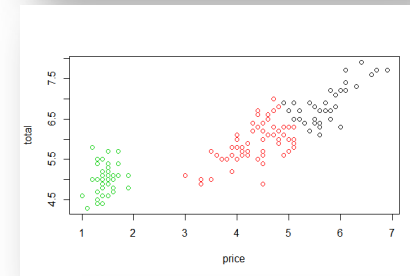
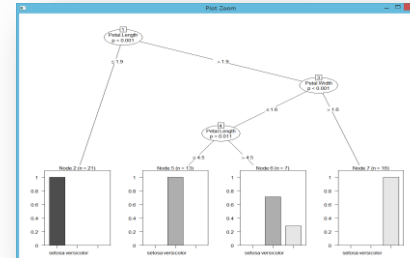
▶ 분류분석(DT, RF, ANN)

16. 비지도학습(unSupervised Learning)

▶ 군집분석(Clustering Analysis)

▶ 연관분석(Association Rule)

17. 시계열 분석(Time Series Analysis)





PART-IV. 예측분석(지도/비지도 학습)

18. 지도 학습(Supervised Learning)

▶ 회귀분석(Regression Analysis)

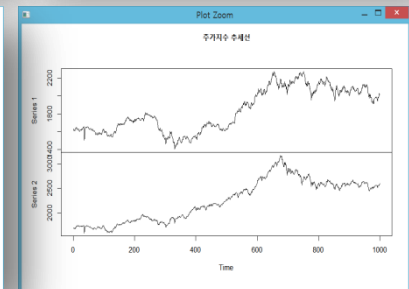
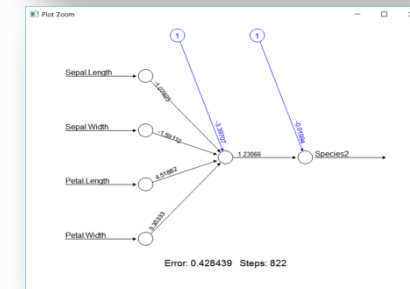
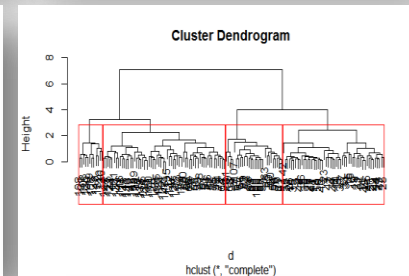
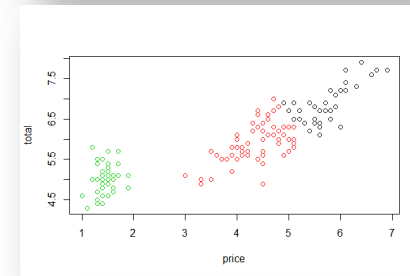
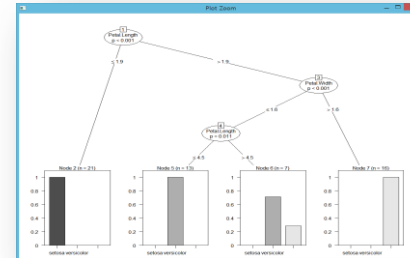
▶ 분류분석(DT, RF, ANN)

16. 비지도 학습(unSupervised Learning)

▶ 군집분석(Clustering Analysis)

▶ 연관분석(Association Rule)

17. 시계열 분석(Time Series Analysis)





BIG 데이터 개요

● 빅 데이터의 3V(가트너 그룹)

1. 용량(Volume)

✓ 데이터 용량의 증가 → 대 단위 분석 이슈

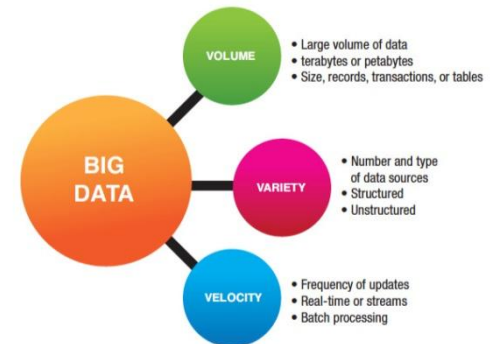
2. 속도(Velocity)

✓ 빠른 데이터 생성 → 빠른 데이터 처리

3. 다양성(Variety)

✓ 소셜미디어, 모바일, DB → 다차원 의사결정

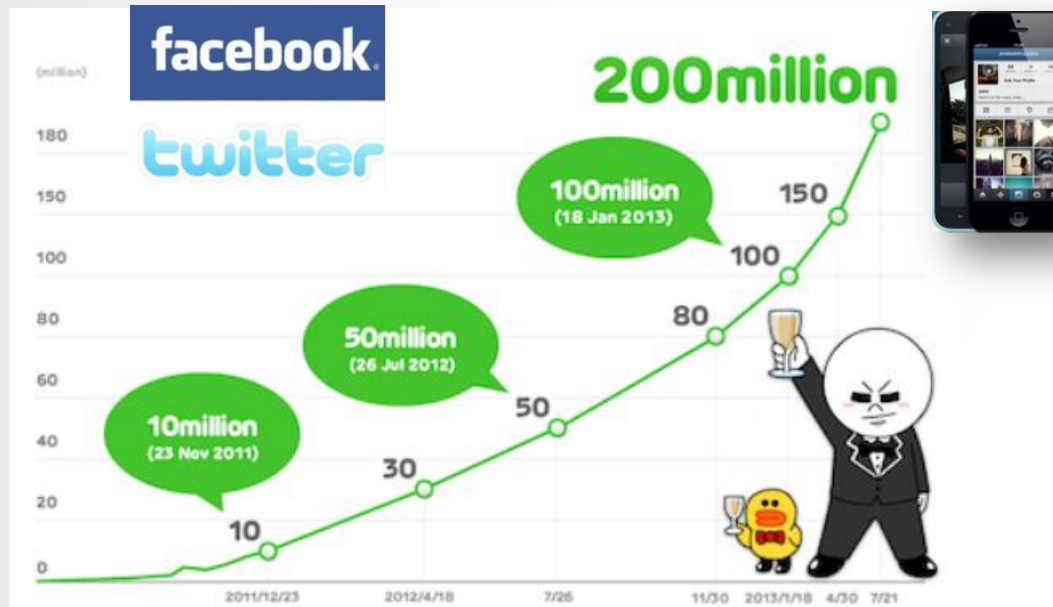
The Three Vs of Big Data





BIG 데이터 등장배경

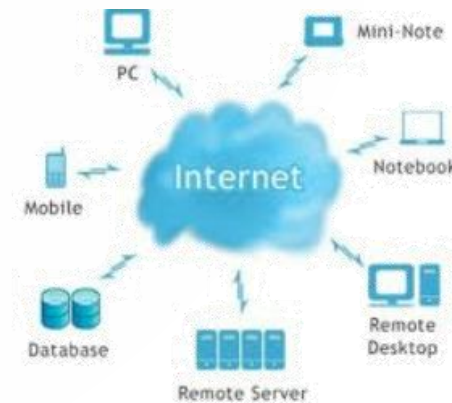
1. 스마트 폰 및 IT 디지털 기기 → 사용량 증가
2. SNS 확산 → 비정형 데이터 증가
3. 비정형 데이터 처리 기술 향상





BIG 데이터 정의 및 특징

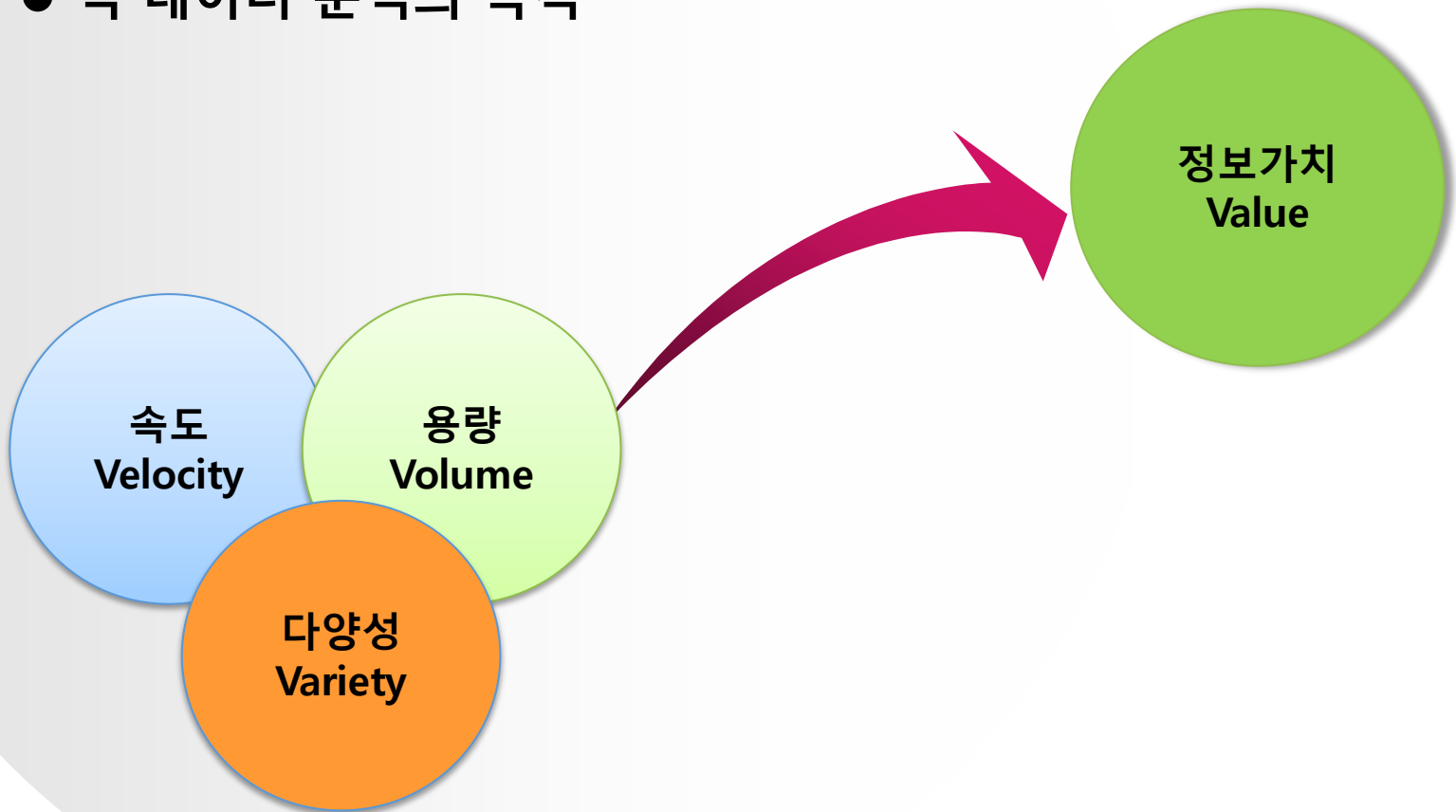
1. 일반적으로 사용되는 데이터 수집, 관리, 처리 소프트웨어의 수용 한계를 초과한 데이터
2. 기존 데이터베이스의 역량을 넘어서는 대량의 정형 또는 비정형 데이터의 집합
3. 기존 데이터 규모에서 불가능했던 새로운 통찰이나 새로운 형태의 가치를 추출 및 예측 할 수 있는 데이터 집합체





BIG 데이터 정의 및 특징

- 빅 데이터 분석의 목적





BIG 데이터 분석 과정

● 정형/비정형 데이터 분석과정

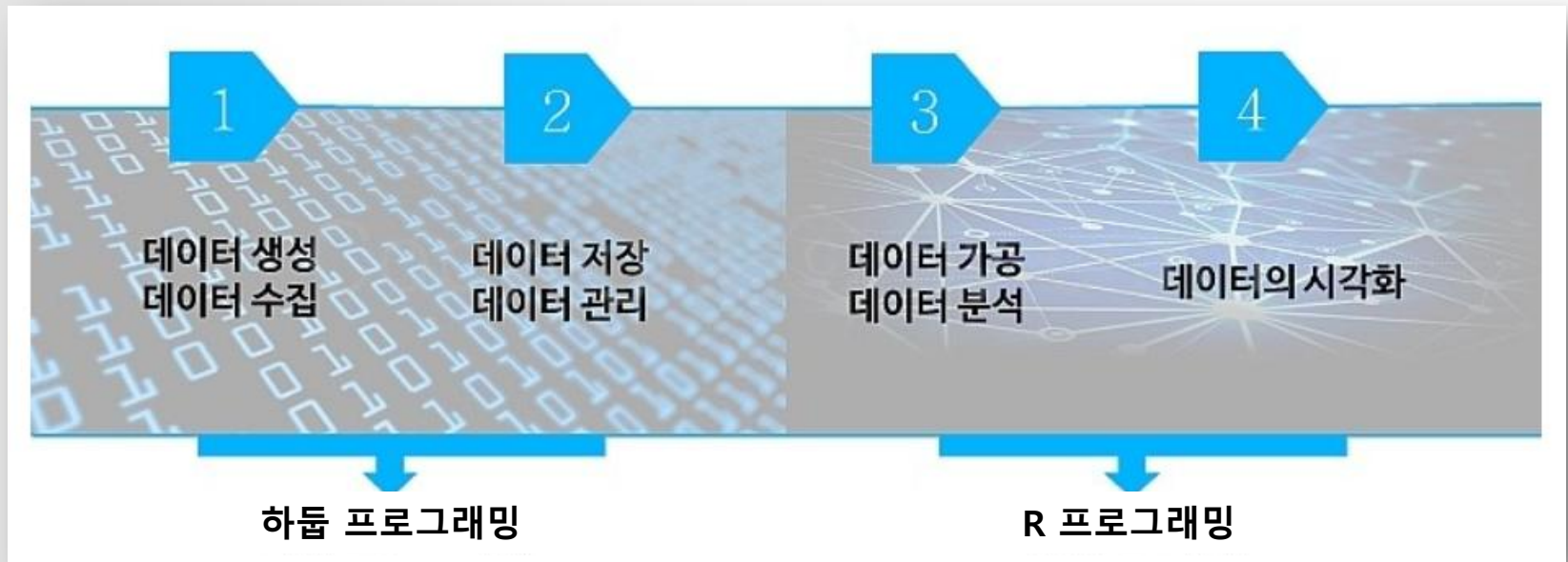


✓실 세계의 현상을 관찰하여 수량화하고 이를 통해서 패턴을 분석하여
추론.예측하는 일련의 과정



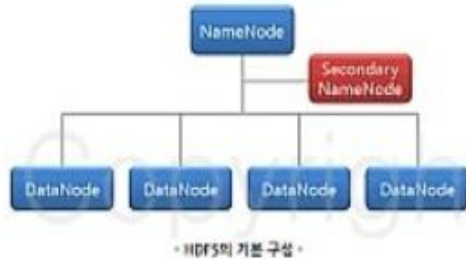
BIG 데이터 분석과정

● 하둡과 R 프로그래밍





BIG 데이터 분석과정

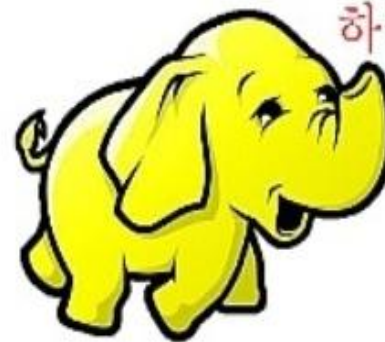


여러 대의 컴퓨터를 병렬구조로 하둡 클러스터라는 여러 대의 서버를 하나의 서버로 묶는 방식으로 방대한 양의 데이터를 분산처리하여 빠른 시간 내에 결과를 제공하는 데이터 관리 기술. 핵심기술은 분산파일시스템 (HDFS)와 맵리듀스 (MapReduce)가 있다.



- 맵리듀스의 Word Count Processing 예제 -

[출처: H대 기술 연구소]



하둡 프로그래밍



BIG 데이터 분석과정

3

데이터 가공-데이터 분석

4

데이터 시각화



[출처: 말성에서 발견하는 분석이야기 AnalyticsStory 블로그]



R프로그래밍

수 많은 데이터를 통계로 활용 할 수 있도록 하기 위한 특화된 프로그래밍 언어이자 소프트웨어 환경으로 자료 분석에 많이 활용된다. 직관적이고 사용자 친화적인 인터페이스와 활용도 높은 기능, 오픈소스라는 점 덕분에 R프로그래밍이 보편적으로 많이 활용되고 있다.

Per Capita Income: Per capita income in the past 12 months (in 2011 inflation-adjusted dollars)



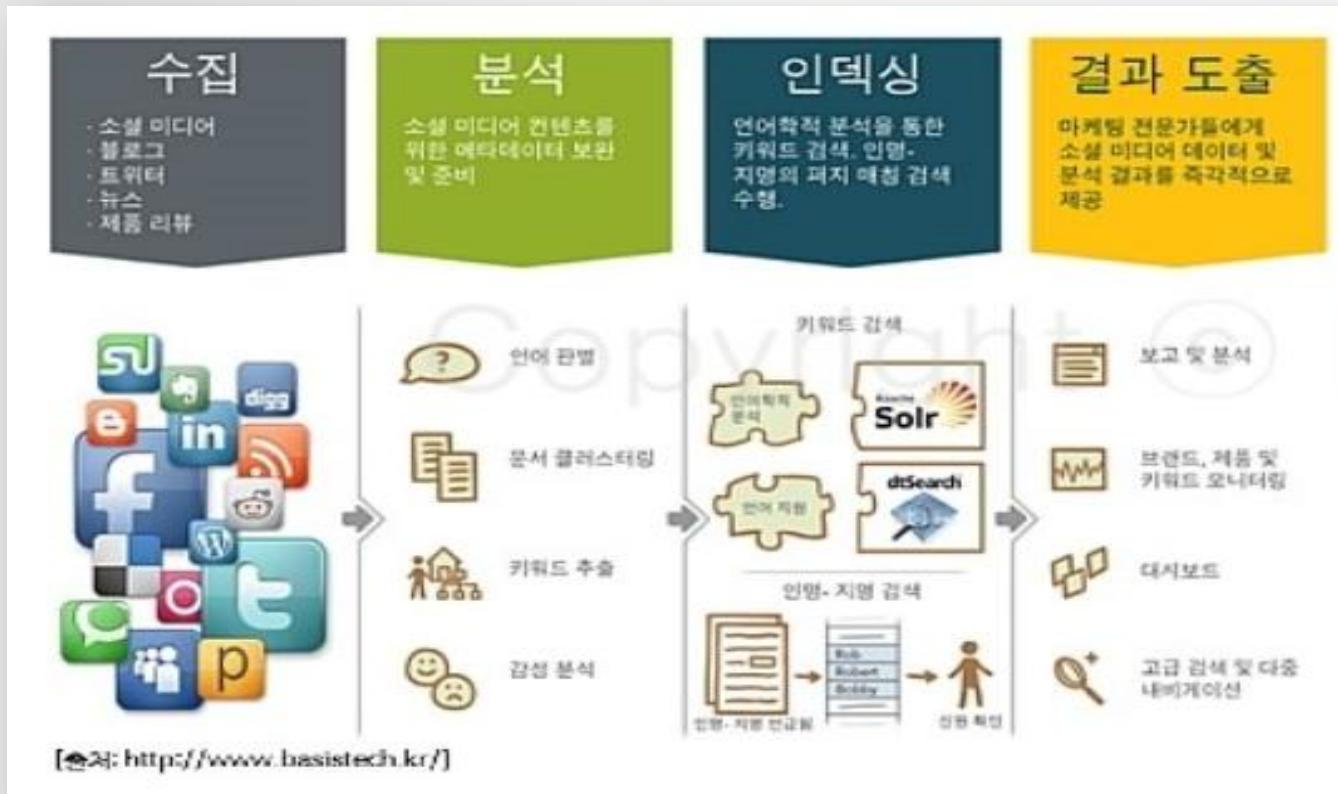
[Trulia미국 부동산 블로그, 소득-전망이 연관성 시각화 예시]

11



BIG 데이터 분석과정

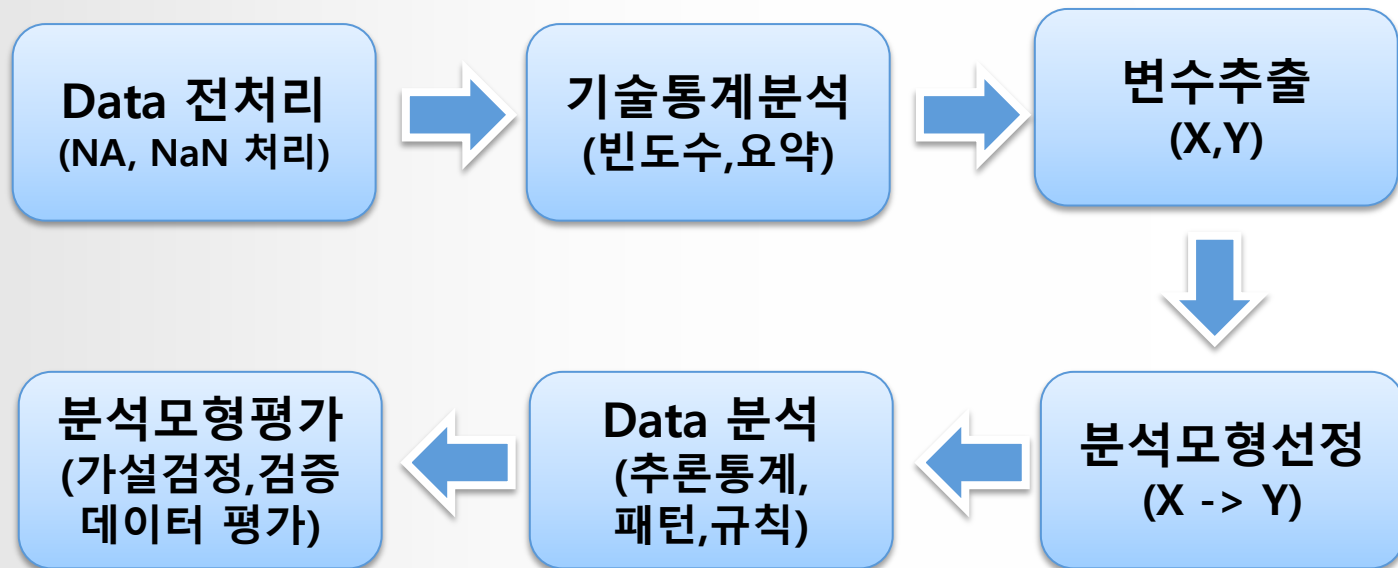
● 일반적인 분석절차





BIG 데이터 분석과정

● 일반적인 분석 절차





BIG 데이터 분석방법

● 분석유형

1. 기술통계분석 : 분석 데이터의 특성 분석
2. 추론 통계분석 : 통계 및 수학 기초 분석 및 검정
3. 텍스트 마이닝 : 키워드 및 연관어 분석
4. 데이터 마이닝
 - ✓ 지도학습 : 인과관계 기반 예측분석
 - ✓ 비지도학습 : 패턴 기반 예측분석



BIG 데이터 분석방법

기술통계분석

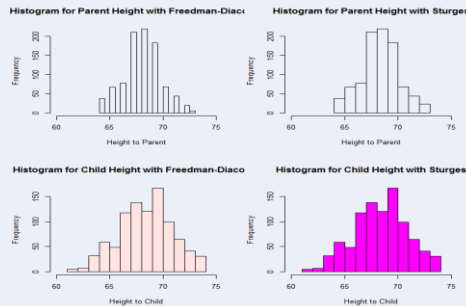
■ 데이터 특성분석

학력수준		실패	진학
고졸	관찰빈도	40	49
	기대빈도	36	54
대졸	관찰빈도	27	55
	기대빈도	33	49
대학원졸	관찰빈도	23	31
	기대빈도	21	32

- 합계, 평균, 빈도수, 비율, 표준편차, 분산, 교차분석 등
- 데이터의 특성 분석

추론통계분석

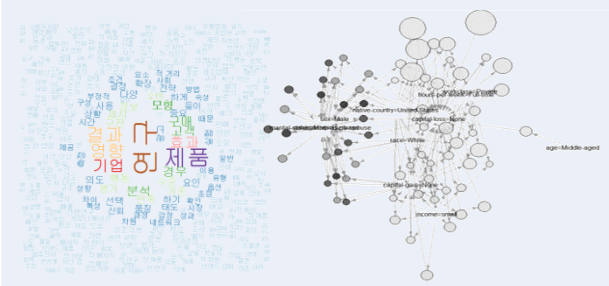
■ 모집단을 표본으로 분석하여 추론



- 카이제곱검정, 비율검정, 평균차이검정, 상관분석, 분산분석
- 집단간 차이 분석

데이터마이닝

■ 변수의 관계와 패턴 분석으로 미래 예측



- 텍스트분석, 예측, 분류, 군집, 연관분석
- 패턴 및 규칙을 이용한 의사결정



BIG 데이터 파급효과

1. 기존 사실에 대한 체계적인 객관적 근거 제시
2. 다변화된 현대 사회를 더욱 정확하게 예측 및 대응 가능
3. 개인화된 현대 사회 구성원 마다 맞춤형 정보 제공, 관리, 분석 가능
4. 정치, 사회, 경제, 문화, 과학 기술 등 전 영역에 걸쳐서 사회와 인류에게 가치 있는 정보 제공(중요성 부각)





BIG 빅데이터 활용 사례

●빅데이터 활용 사례





BIG 빅데이터 활용 사례

● 빅데이터 활용 사례

경제적 효과

	계수	2012	2015	2020
생산유발효과(억원)	1.32	95,573	119,736	171,296
부가가치 유발효과 (억원)	0.99	71,680	89,802	128,472
인력창출(명)	1.24	89,781	112,479	160,914

*유통업계 재고관리에 효율적인 빅데이터 사용 사례

[파리바게트의 날씨 마케팅]

27도 이상의 맑은 날씨: 샌드위치 판매 상승

20도 이하의 쌀쌀 날씨: 피자 빵

[출처: 한국통신학회지 (정보와통신) 제29권제11호, 48-54 (7 pages), 최기훈(빅데이터) 등장에 따른
경제적 파급효과 및 범(규제) 연구, 이규철, 김희선, 2012.10.]





BIG 빅데이터 활용 사례

- 빅데이터 활용 사례

공공측면(재난대응)



싱가포르의 국가위험관리, 일본은 자연재해방지



관련분야	빅 데이터 활용 사례
공공분야	▪ 교통, 부정부패, 세수 증감 데이터 분석
복지분야	▪ 자살 예보시스템, 실버 계층 의료 개선
기상분야	▪ 날씨 분석과 재난 예고
정치분야	▪ 소셜 데이터를 통한 선거 전략 수립, 맞춤형 캠페인
의료분야	▪ 인간 유전자 데이터 분석과 희귀병 치료
금융분야	▪ 주가지수 예측, 거시 변수 예측
의류분야	▪ 유행 디자인 사전 파악 및 시장 선호도 분석
스포츠분야	▪ 선수 부상예측, 상대팀 전술 분석 및 대응