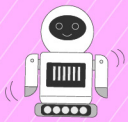


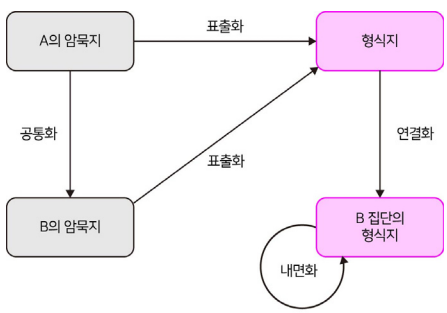
1 과목

빅데이터 분석 기획

데이터(본문 19, 21페이지)

데이터의 형태	형태	예
정량적 데이터	숫자, 수치, 도형	75kg, 12,000원, -15도
정성적 데이터	문자, 언어	인터뷰, 메모, 메일, 자료 영상

암묵지와 형식지의 상호작용: 공통화 〉 표출화 〉 연결화 〉 내면화



실전 Tip

지식창조 메커니즘은 [공 표 연 내] 공포였네!

상호작용	예
공통화	대화, OJT, 회의, 제자 양성
표출화	회의록 작성, 매뉴얼 작성
연결화	공식 업무방법서 배포, 출간
내면화	습득

DIKW 피라미드(지식의 피라미드)

실전 Tip

지식의 피라미드 순서는 [데정 지혜] 데정이가 지혜를 좋아한다.

구분	내용
데이터(Data)	현실에서 관찰된 값으로 다른 데이터와 상관관계가 없는 가공하기 전의 순수한 수치나 기호 예) A마트는 노트가 1,000원, B슈퍼는 노트가 3,000원이다.
정보(Information)	데이터를 처리, 가공하여 데이터 간의 연관 관계와 의미가 도출된 요소 예) A마트의 노트가 B슈퍼보다 저렴하다.
지식(Knowledge)	상호 연결된 정보를 이해하여 이를 토대로 예측한 결과물 예) A마트에서 노트를 구매해야겠다.
지혜(Wisdom)	근본 원리에 대한 이해를 바탕으로 도출되는 창의적 아이디어로 상황이나 맥락에 맞게 규칙을 적용하는 요소 예) A마트의 다른 상품도 B슈퍼보다 저렴할 것이다.

데이터베이스(본문 22, 25페이지)

데이터베이스의 특징

- 공용 데이터(Shared data): 여러 사용자가 서로 다른 목적으로 데이터베이스의 데이터를 공동으로 이용한다.
- 통합된 데이터(Intergrated data): 동일한 데이터가 중복되어 저장되지 않는다.
- 저장된 데이터(Stored data): 컴퓨터가 접근할 수 있는 저장매체에 저장한다.
- 변화되는 데이터(Changed data): 데이터는 현시점의 정확한 데이터를 유지하면서 지속적으로 갱신한다.

실전 Tip

데이터베이스의 특징은 [공통 저변] 공통점이 저변에 깔려 있다.

데이터베이스의 시스템

구분	OLTP	OLAP
구조	복잡	단순
갱신	동적으로 순간적	정적으로 주기적
응답시간	수 초 이내	수초에서 몇 분 사이
데이터 범위	수십 일 전후	오랜 기간
성격	정규적 핵심 데이터	비정규적 읽기전용 데이터
크기	수 기가 바이트	수 테라 바이트
내용	현재 데이터	요약된 데이터
특성	트랜잭션 중심	주제 중심
액세스 빈도	높음	보통
자료 예측	주기적, 예측 가능	예측 어려움

데이터 산업(본문 26페이지)

데이터 산업은 처리 - 통합 - 분석 - 연결 - 권리 시대로 구분할 수 있다.

실전 Tip

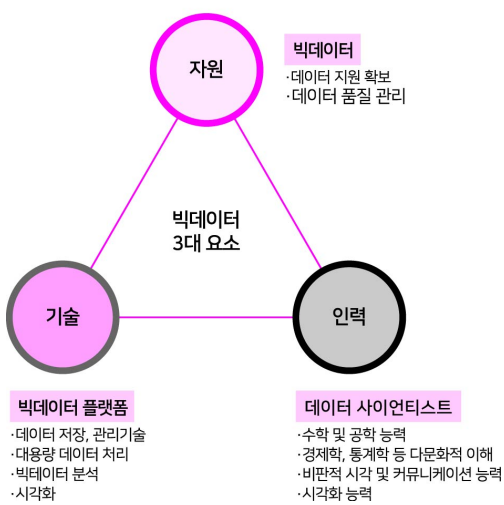
데이터 산업의 각 시대별 주요한 특징을 키워드로 파악합니다.
처리시대 – 데이터는 업무처리 대상
통합시대 – 데이터 웨어하우스
분석시대 – 빅데이터 등장
연결시대 – 오픈API
권리시대 – 마이데이터, 경제적 자원

빅데이터(본문 29, 30페이지)

구분	특징	설명
3V	규모 (Volume)	과거 텍스트 데이터부터 사진, 동영상 등 계속해서 생성되는 다양한 데이터의 인식으로 방대한 규모를 가진다.
	다양성 (Variety)	정형, 비정형, 반정형 데이터 등 형태의 다양성이 있다.
	속도 (Velocity)	• 실시간 정보 생성 및 전달의 속도가 증가하여 실시간 처리 등의 요구가 확대되었다. • 가치 있는 정보를 위해 데이터 처리 및 분석 속도가 중요하다.
4V	가치 (Value)	• 각 데이터에 내재되어 있는 가치를 인식해야 한다. • 빅데이터의 가치는 데이터의 정확성 및 시간성과 관련된다.
7V	정확성 (Veracity)	• 질 높은 데이터를 활용한 정확한 분석 수행이 없다면 의미가 없다. • 데이터가 타당한지 정확한지에 대한 여부는 의사결정의 중요한 요소이다.
	휘발성 (Volatility)	데이터가 의미 있는 기간으로 장기적인 관점에서 유용할 필요가 있다.
	신뢰성 (Veracity)	데이터 기반 분석에서 잘못된 결론 도출을 방지하기 위하여 데이터의 품질에 대한 신뢰성과 정확성이 필요하다.

빅데이터 활용을 위한 3요소

빅데이터 활용을 위한 3대 요소



실전 Tip

빅데이터 활용을 위한 요소는 [자기인] 자기 전에 인사하자!

빅데이터 조직 및 인력(본문 34페이지)

구분	내용
기능구조	• 일반적인 형태로 별도 분석조직이 없고 해당 부서에서 분석 수행 • 전사적 핵심 분석이 어려우며 과거에 국한된 분석수행
집중구조	• 전사의 분석 업무를 별도의 분석 전담 조직에서 담당함 • 전략적 중요도에 따라 분석조직이 우선순위를 정해서 진행 가능 • 현업 업무부서의 분석 업무와 중복 및 이원화 가능성
분산구조	• 분석조직 인력들을 현업 부서로 직접 배치해 분석 업무를 수행 • 전사 차원의 우선순위 수행 • 분석 결과에 따른 신속한 피드백이 나오고 베스트 프랙티스 공유가 가능 • 업무 과다와 이원화 가능성이 존재할 수 있어 부서 분석 업무와 역할 분담이 명확해야 함 • 다수의 데이터분석 엔지니어가 필요

빅데이터 플랫폼(본문 38, 40, 41페이지)

주요 기술

단계	주요 기술
수집	• ETL • EAI • 크롤링
저장	• 분산파일시스템 • RDBMS • Nosql
처리, 분석	• SQL • 머신러닝 • 통계분석
분석	• R • Python • BI • Open API

하둡 분산파일시스템(HDFS)

구분	내용
네임노드	파일시스템의 네임스페이스를 관리하며, 파일에 속한 모든 블록이 어느 데이터 노드에 있는지 파악
데이터 노드	클라이언트나 네임노드의 요청이 있을 때 블록을 저장하고 탐색하며, 저장하고 있는 블록의 목록을 주기적으로 네임노드에 보고

하둡에코시스템(Hadoop Ecosystem)

기능	기술	설명
정형 데이터수집	스쿱(Sqoop)	관계형 데이터스토어 간 대량 데이터를 효과적으로 전송하는 도구
	히호(Hiho)	대용량 데이터 전송 솔루션으로, 하둡에서 데이터를 가져오기 위한 SQL을 지정할 수 있으며, JDBC 인터페이스를 지원
비정형 데이터수집	척와(Chukwa)	분산환경에서 생성되는 데이터를 HDFS에 안정적으로 저장하는 플랫폼
	플럼 (Flume)	• 대용량의 로그 데이터를 효과적으로 수집, 집계, 이동시키는 분산 서비스 제공 솔루션 • 분산된 서버에 에이전트가 설치되고, 에이전트로부터 데이터를 전달받는 콜랙터로 구성
	스크라이브 (Scribe)	• 페이스북에서 개발한 데이터 수집 플랫폼 • 데이터를 중앙 집중 서버로 전송하는 방식
분산데이터 저장	HDFS	• 수십 테라바이트 또는 페타바이트 이상의 대용량 파일을 분산된 서버에 저장하고, 그 저장된 데이터를 빠르게 처리할 수 있게 하는 파일시스템 • 네임노드와 데이터노드로 구현
분산 데이터 베이스	HBASE	• HDFS의 칼럼 기반 데이터베이스 • 실시간 랜덤 조회 및 업데이트가 가능하며, 각각의 프로세스들은 개인의 데이터를 비동기적으로 업데이트 가능
분산 데이터 처리	맵리듀스 (Map-reduce)	• 구글에서 제작한 소프트웨어 프레임워크 • 분산 병렬컴퓨팅에서 대용량 데이터 처리
리소스 관리	얀 (Yarn)	• 리소스 관리와 컴포넌트 처리를 분리한 자원관리 플랫폼 • 대규모 데이터 처리 어플리케이션들을 실행하는 운영체제 역할을 수행
데이터 마이닝	머하웃 (Mahout)	하둡 기반 데이터 마이닝 알고리즘을 구현한 오픈소스
데이터 가공	피그 (Pig)	• 데이터처리를 위한 병렬처리 언어이며 아파치 하둡의 서브프로젝트 • Pig Latin이라는 자체 언어를 제공 • Map과 Reduce 두 단계로 이루어진 단순한 병렬모델 • 코드 길이가 짧고 개발 시간도 단순해짐
	하이브 (Hive)	• 페이스북에서 개발한 데이터 웨어하우징 인프라 • SQL 기반의 쿼리언어와 JDBC를 지원
실시간SQL질의	임팔라 (Impala)	• 클라우드데라에서 개발한 하둡 기반의 실시간 SQL 질의 시스템 • 맵리듀스를 사용하지 않고, 자체 개발한 엔진으로 분석과 트랜잭션 처리를 모두 지원
	타조 (Tajo)	하둡 기반의 대용량 데이터를 SQL 형태의 명령을 통해 분산 분석 작업을 지원하는 대용량 데이터 웨어하우스
워크 플로우 관리	우지 (Ooozie)	하둡의 작업을 관리하는 워크플로우 및 코디네이터 시스템
분산 코디네이션	주키퍼 (Zookeeper)	• 분산 환경에서 서버들간에 상호 조정이 필요한 다양한 서비스를 제공하는 시스템 • 하나의 서버에만 서비스가 집중되지 않도록 서비스를 알맞게 분산하여 동시에 처리하게 해 줌

개인정보(본문 50페이지)

프라이버시 보호모델

모델	기능	설명
k-익명성 (k-anonymity)	특정인임을 추론할 수 있는지 검토한다.	동일한 값을 가진 레코드들이 최소 k개 이상 존재하도록 하여, 개인을 식별할 확률이 1/k이다.
l-다양성 (l-divrsity)	특정인임을 추론이 불가하지만, 민감정보의 다양성을 높혀 추론 가능성을 낮춘다.	각 레코드들은 최소 l개 이상의 다양성을 가져 동질성에 의한 추론을 방지한다.
t-근접성 (t-closeness)	민감정보의 분포를 낮춰 추론가능성을 낮춘다.	특정 정보의 분포와 전체 데이터의 정보분포 차이를 t 이하로 하여 추론을 방지한다.

개인정보 법과 제도(본문 52페이지)

실전 Tip

데이터 3법은 [정신개]이다.

데이터 3법	주요 내용
정보통신망법	온라인 상 개인정보보호 감독주체를 개인정보보호위원회로 변경
신용정보법	• 가명정보 도입으로 빅데이터 분석 및 이용의 법적 근거 마련 • 가명정보는 통계작성, 연구, 공익적 기록 보존 등을 위해 신용정보 주체의 동의 없이 이용 및 제공 가능 • 마이 데이터 산업 도입으로 개인정보 보호 강화
개인정보 보호법	• 개인정보의 범위 명확화 • 가명정보 활용범위 명확화와 이에 따른 안정장치 마련 • 동의 없이 처리할 수 있는 개인정보의 인정 • 개인정보 보호체계 일원화

분석마스터 플랜과 로드맵 설정(본문 75페이지)

분석마스터 플랜 수립 기준

구분	기준	설명
우선순위 설정	전략적 중요도	전략적 필요성과 시급성 고려 • 비즈니스 전략적 목표에 직접적인 연관 관계 여부 • 사용자 요구사항 또는 업무능률 향상에 얼마나 시급히 수행되어야 하는지를 확인
	비즈니스 성과	비즈니스 성과에 따른 투자 여부 판단
	실행 용이성	투자 용이성 및 기술 용이성을 통해 실제로 프로젝트 추진이 가능한지 여부를 분석 • 기간 및 인력, 비용 투입 용이성 • 기술의 안정성, 개발의 성숙도
로드맵 수립	업무 내재화 적용 수준	업무에 내재화하거나 별도의 분석화면으로 적용할 것인지 결정
	분석 데이터 적용 수준	내부 데이터 및 외부 데이터 범위 결정
	기술 적용 수준	분석 기술의 범위 및 방식을 고려

분석 ROI 요소 4V

구분	ROI요소	특징	내용
3V	투자비용 요소	데이터 크기(Volume)	양과 규모
		데이터 형태(Variety)	종류와 유형
		데이터 속도(Velocity)	생성속도 및 처리속도
4V	비즈니스 효과	새로운 가치(Value)	분석 결과가 창출하는 가치

데이터 확보 계획(본문 79페이지)

데이터 확보 계획 수립절차

순서	단계	업무	내용
1	목표 정의	• 성과 목표 정의 • 성과 지표 설정	• 비즈니스 도메인 특성 적용 • 구체적인 성과목표 정의 • 성과측정을 위한 지표 도출
2	요구사항 도출	데이터 및 기술 지원 등과 관련된 요구사항 도출	• 필요 데이터 확보 및 관리 계획 • 데이터 정제 수준, 데이터 저장 형태 • 기존 시스템 및 도구 활용 여부 • 플랫폼 구축 여부
3	예산안 수립	자원 및 예산 수립	데이터 확보, 구축, 정비, 관리 예산
4	계획 수립	• 인력 투입 방안 • 일정 관리 • 위험 및 품질관리	• 프로젝트 관리 계획 수립 • 범위, 일정, 인력, 의사소통 방안 수립

분석절차와 작업 계획(본문 85페이지)

절차	설명
문제 인식	• 비즈니스 문제와 기회를 인식하고 분석 목적을 정의 • 분석 주제 정의, 문제는 가설의 형태로 정의
연구 조사	• 목적 달성을 위한 각종 문헌 조사 • 조사 내용을 해결방안에 적용
모형화	• 분석 문제를 단순화하여 수치와 변수 사이의 관계로 정의함 • 많은 변수가 포함된 현실 문제를 특징적 변수로 정의
자료 수집	• 데이터 수집, 변수 측정 • 기존 데이터 수집 가능성 확인 및 대체 데이터 확인
자료 분석	• 수집된 자료 내에서 의미 및 변수들 간 관계 분석 • 기초 통계부터 데이터 마이닝 기법 활용
분석결과 공유	• 변수 간의 관련성을 포함한 분석결과 제시 • 의사결정자와 결과 공유 • 표, 그림, 차트를 활용하여 가시화

실전 Tip

문제 인식 〉 연구조사 〉 모형화 〉 데이터 수집 〉 데이터 분석 〉 분석 결과 제시

데이터 분석 준비도(Readiness) (본문 100페이지)

조직 분석 성숙도 단계

단계	도입단계	활용단계	확산단계	최적화단계
설명	데이터 분석을 시작하여 환경 및 시스템 구축하는 단계	분석결과를 실제 업무에 적용하는 단계	전사 차원에서 분석 관리 및 공유단계	분석을 진화시켜 혁신 및 성과 향상에 기여하는 단계
조직 역량 부문	• 일부 부서에서 수행 • 담당자 역량에 의존	• 담당 부서에서 수행 • 분석 기법 도입	• 전사 모든 부서 시행 • 분석 전문가 조직 운영 • 데이터 사이언티스트 확보	• 데이터 사이언스 그룹 • 경영진 분석 활용 및 전략연계
비즈니스 부문	• 실적분석 및 통계 • 정기보고	• 미래결과예측 • 시뮬레이션	• 전사성과 실시간 분석 제공 • 분석규칙 및 이벤트 관리	• 외부 환경분석 활용 • 최적화 업무 적용
IT 부문	• Data Warehouse • Data Mart • ETL/EAI • OLAP	• 실시간 대시보드 • 통계분석환경	• 빅데이터 관리환경 • 시뮬레이션 최적화 • 비주얼 분석 • 분석전용 서버	• 분석 협업환경 • 분석 Sandbox

데이터 수집(본문 113페이지)

구분	원천시스템	종류
내부 데이터	서비스 시스템	ERP, CRM, 정보계, 포털, 원장정보시스템, 인증/과금시스템, 거래시스템 등
	네트워크 데이터	방화벽, 백본, 스위치, IPS
	마케팅 데이터	고객 방문 로그, VOC 데이터 등
외부 데이터	소셜 데이터	리뷰, 메신저, 인스타그램 등
	네트워크 데이터	장비 발생 로그, 센서 데이터 등
	공공데이터	정부 공개 데이터, 의료, 지역 정보, 공공정책, 지리, 환경, 통계 등

데이터 품질 검증(본문 130, 131페이지)

정형데이터 품질기준

품질기준	설명	세부 품질기준
완전성	필수항목에 누락이 없으며, 칼럼 값이 항상 존재해야 한다.	개별완전성, 조건완전성
유일성	데이터 항목은 유일해야 하며 중복되어서는 안 된다.	단독, 조건 유일성
일관성	데이터의 할 구조, 값, 형태가 일관되게 정의되는 것으로 신뢰를 보장하는 척도이다.	기준코드 일관성, 참조 무결성, 데이터 흐름 일관성, 칼럼 일관성
정확성	• 현실에 존재하는 객체의 표현 값이 정확히 반영되어야 한다. • 사용 목적에 따라 데이터 정확성의 기준은 달라질 수 있다.	선후 관계 정확성, 계산/집계 정확성, 최신성, 업무규칙 정확성
유효성	데이터는 정해진 데이터 유효범위 및 도메인을 충족해야 한다.	범위, 날짜, 형식 유효성

비정형 데이터 품질기준

품질기준	설명	세부 품질기준
기능성	해당 데이터가 특정 조건에서 사용될 때, 명시된 요과 내재된 요구를 만족하는 기능 제공 여부	적절성, 정확성, 상호 운영성, 기능 순응성
신뢰성	데이터가 규정된 신뢰 수준을 유지하거나 사용자가 오류를 방지할 수 있도록 하는 정도	성숙성, 신뢰 순응성
사용성	데이터가 사용될 때, 사용자가 이해 가능하며 선호하는 정도	이해성, 친밀성, 사용 순응성
효율성	데이터가 사용될 때, 사용되는 자원의 양에 따라 요구된 성능을 제공하는 정도	시간 효율성, 자원 효율성, 효율 순응성
이식성	해당 콘텐츠가 다양한 환경과 상황에서 실행될 가능성	적응성, 공존성, 이식 순응성

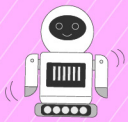
데이터 저장(본문 135페이지)

데이터 웨어하우스

실전 Tip

데이터 웨어하우스 특징 4가지 [주, 통, 시 비] 주통에게 시비를 걸었다.

특징	내용
주제 지향적 (Subject Oriented)	기업의 기능이나 업무가 아닌 주제 중심적으로 구성되는 특징
통합성(Integrated)	여러 소스에서 데이터를 조합해 전사적 관점에서 하나로 통합되는 특징
시계열성(Time-variant)	시간 변이키가 존재해 시간에 따른 변경을 반영하고 있다는 특징
비휘발성(Non-Volatile)	정기적 데이터 변경을 제외하고 검색 작업만 수행되는 읽기 전용의 데이터를 유지함



2 과목

빅데이터 탐색

데이터 정제(본문 148, 149페이지)

실전 Tip

데이터 오류는 [이 결 노] 이 결혼 안 된다.

데이터 오류의 종류

종류	설명	처리 방법
이상값	데이터의 범위에서 많이 벗어난 아주 작은 값이나 아주 큰 값	하한값 또는 상한값 대체
결측값	존재하지 않거나 관측되지 않는 값	중심 경향값 넣기
노이즈	실제는 입력되지 않았지만 입력되었다고 잘못 판단된 값	평균값 또는 중간값 대체

데이터 정제의 과정

단계	수행 내용
데이터 수집	• 데이터의 수집 방법 및 기준 설정 • 입수경로 구조화 • 집계 및 저장소 결정
데이터 변환	• 데이터를 분석이 가능한 형태로 변환 • ETL, 일반화, 정규화 등
데이터 교정	결측치, 이상치, 노이즈 값 처리
데이터 통합	데이터 분석이 용이하도록 기존 또는 유사 데이터와의 통합

데이터 이상값 처리(본문 157페이지)

데이터 이상값 검출 방법

통계기법	내용
ESD	평균으로부터 3 표준편차 떨어진 값을 이상값으로 판단
기하평균	기하평균으로부터 2.5 표준편차 떨어진 값을 이상값으로 판단
사분위 수	제1사분위, 제3사분위를 기준으로 사분위 간 범위의 1.5배 이상 떨어진 값을 이상값으로 판단
표준화 점수 Z score	• 서로 다른 척도 등으로 비교하기 어려운 데이터 추적에 유용 • 평균이 μ이고 표준편차가 σ인 정규분포를 따르는 관측치 간의 차이의 비율을 활용해 이상값 여부를 검정하는 방법 $z = \frac{x-\mu}{\sigma}$
딕슨의 Q 검정 (Dixon's Q-test)	• 오름차순으로 정렬된 데이터에서 범위에 대한 관측치 간의 차이의 비율을 활용해 이상값 여부를 검정하는 방법 • 데이터 수가 30개 미만인 경우 적절함
그럽스 T-검정 (Grubbs T-test)	정규분포를 만족하는 단변량 자료에서 이상값을 찾는 통계적 검정
카이제곱 검정 (chi-square test)	• 카이제곱 검정은 데이터가 정규분포를 만족하나, 자료의 수가 적은 경우에 이상값을 검정하는 방법 • 두 범주형 변수 사이의 독립성을 검정하는 데 사용
마할라노비스 거리 (Mahalanobis distance)	• 다변량 이상치 검출, 불균형 데이터셋에서의 분류 등에서 유용 • 모든 변수 간에 선형관계를 만족하고, 각 변수들이 정규분포를 따르는 경우 적용할 수 있는 접근법 • 데이터의 분포를 고려하여 데이터의 형태를 잘 반영함

변수 선택(본문 164페이지)

래퍼 기법(Wrapper)

선택방법	내용
전진 선택법	• 기존 모형에 가장 설명력이 좋은 변수를 하나씩 추가하는 방법이다. • 모형에서 단순 상관관계수의 절댓값이 가장 큰 변수를 분석모형에 포함시킨다. • 한번 추가된 변수는 제거하지 않는다.
후진 선택법	• 모든 변수가 포함된 모형에서 설명력이 가장 적은 변수를 제거하는 방법이다. • 단순상관계수의 절댓값이 가장 작은 변수를 분석모형에서 제외시킨다. • 한번 제거된 변수는 추가하지 않는다.
단계적 선택법	전진선택법을 선택하여 설명력이 좋은 변수를 추가한 다음 후진선택법을 통해 유의하지 않은 변수를 제거한다.

임베디드기법(Emdedded)

기법	설명
라쏘(L1규제) Lasso	• 중요한 몇 개의 변수를 선택하고 나머지 변수들의 영향력을 0으로 만든다. • 가중치의 절댓값의 합을 최소화하는 것을 제약조건으로 하는 방법이다. $J(\theta) = MSE(\theta) + \alpha \sum_{i=1}^{n} \lvert\theta_i\rvert$
릿지(L2규제) Ridge	• 전체 변수를 유지하면서 각 변수의 계수 크기만 조절하는 것이다. • 가중치들의 제곱합을 최소화하는 것을 추가적인 제약조건으로 하는 방법이다. $J(\theta) = MSE(\theta) + \alpha \frac{1}{2} \sum_{i=1}^{n} \theta_i^2$
엘라스틱 넷 Elastic Net	• 릿지와 라쏘규제를 결합하여 만든 모델이다. • 가중치의 절대값의 합과 제곱합을 동시에 추가적인 제약조건으로 하는 방법이다. $J(\theta) = MSE(\theta) + r\alpha \sum_{i=1}^{n} \lvert\theta_i\rvert + \frac{1-r}{2} \alpha \sum_{i=1}^{n} \theta_i^2$

데이터 탐색(본문 182페이지)

개별변수 탐색 방법

① 범주형 데이터

종류	설명	예시
명목형 데이터	변수나 변수의 크기가 순서와 상관없고, 의미가 없이 이름만 의미를 부여할 수 있는 경우	성별, 지역, 학교명
순서형 데이터	어떤 기준에 따라 변수에 순서를 부여할 수 있는 경우	키, 상태, 소득 수준

② 수치형 데이터

종류	설명	예시
이산형 데이터	변수가 취하는 값을 셀 수 있는 경우	설문조사 참여인원, 지역 내 동의 개수
연속형 데이터	변수가 어떤 구간 내의 모든 값을 가질 수 있는 경우	키, 몸무게 등

비정형 데이터 탐색(본문 197페이지)

비정형데이터 분석

종류	내용	특징
데이터 마이닝	대규모의 데이터 안에서 체계적이고 자동으로 통계적 규칙이나 패턴을 분석해 가치 있는 정보를 추출하는 과정이다.	• 신용평점시스템, 사기탐지시스템, 장바구니 분석 등과 같이 다양한 산업 분야에서 사용된다. • 자료가 현실을 충분히 반영하지 못한 상태에서 정보를 추출한 모형을 개발할 경우 잘못된 모형을 구축하는 오류를 범할 수 있다.
텍스트 마이닝	자연어처리 방식을 이용해 대규모 문서에서 정보추출, 연계성 파악, 분류 및 군집화 등을 통해 데이터의 숨겨진 의미를 발견하는 기법이다.	정보검색, 문서 자동분류, 신문 기사 클러스터링, AI 등에 활용된다.
오피니언 마이닝	• 텍스트 마이닝의 한 분류로 특정 주제에 대한 사람들의 의견을 통계화하여 객관적 정보로 바꾸는 기술이다. • 감정분석이라고도 불린다.	소비자의 반응, 시장예측 등에 활용된다.
웹마이닝	웹자원으로부터 의미 있는 패턴 및 추세 등을 도출해 내는 것이다.	• 대량의 로그기록을 기반으로 정보를 수집하여 패턴 등을 도출해 낸다. • 로그기록을 바탕으로 마케팅 기획, 결과분석 등 다양한 분야에 활용된다.

확률분포(본문 211페이지)

확률분포의 종류

종류	내용
이산확률변수	• 유한하거나 셀 수 있는 범위에서만 값을 가지는 확률변수를 나타낸다. • 각 값에 대한 확률이 명확하게 정의되며, 이산적인 값에 대해서만 확률이 존재한다.
연속확률변수	• 특정 범위 안에서 값이 나타날 확률이 정의된다. • 확률 밀도 함수를 통해 값을 계산하며, 연속적인 값에 대해서도 확률이 존재한다.

점추정(본문 220페이지)

점추정의 조건

실전 Tip

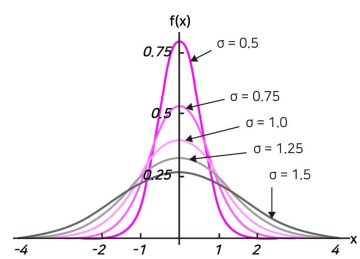
점 추정의 조건 4가지는 [불 효 일 충]이다.

① 불편성(unbiasedness)

- 표본으로부터 구한 통계량의 기대치가 추정하려 하는 모수의 실제 값에 같거나 가까워지는 성질이다.
- 표본에서 얻은 추정량의 기댓값은 모집단의 모수와 차이가 없다.

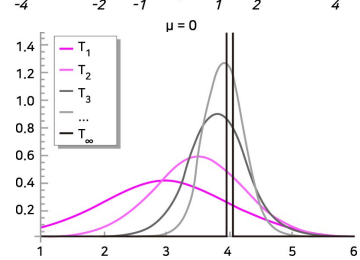
② 효율성(efficiency)

추정량의 분산이 작을수록 좋다.



③ 일치성(consistency)

표본의 크기가 커지면, 추정량이 모수와 거의 같아진다.

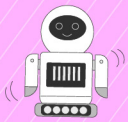


④ 충분성(sufficiency)

- 추정량이 모수에 대해서 가장 많은 정보를 제공해야 한다.
- 충분추정량을 사용해 충분성을 측정한다.

가설검정(본문 225페이지)

판단 기준	내용
제1종 오류	귀무가설이 참일 때, 귀무가설을 기각하도록 결정하는 오류
제2종 오류	귀무가설이 거짓일 때, 귀무가설을 채택할 오류



3과목

빅데이터 모델링

분석 모형의 종류(본문 238, 241페이지)

예측 모델(Prediction Model)

기법	설명
회귀분석	관찰된 연속형 변수들로 두 변수 사이의 모형을 구한 뒤 적합도를 측정해 내는 분석 방법
의사결정나무	의사결정 규칙을 트리구조로 도표화하여 분류와 예측을 수행하는 방법
인공신경망	인간 두뇌의 뉴런이 전기신호를 전달하는 모습을 모방한 예측 모델
시계열 분석	시계열로 관측되는 자료를 분석하여 미래를 예측하는 분석 방법

머신러닝 기반 분석 모형 선정

① 지도학습(supervised learning)

기법	설명
로지스틱 회귀분석	반응변수가 범주형인 경우 적용되는 회귀분석 모형
의사결정나무	분할 기준 속성을 선정하고 이에 따라 트리 형태로 모델링하는 분류 및 예측 모델
랜덤 포레스트	배깅과 부스팅보다 더 많은 무작위성을 주어 약한 학습기들을 생성한 후 이를 선형 결합하여 최종 학습기를 만드는 방법
인공신경망 분석	인간의 뉴런 구조를 모방하여 만든 기계학습 모델
서포트벡터 머신	데이터를 분리하는 초평면 중에서 데이터들과 거리가 가장 먼 초평면을 선택해 분리하는 분류 모델

② 비지도학습(Unsupervised Learning)

기법	설명
Clustering	주어진 관측값들 사이의 거리(distance) 또는 유사성을 이용하여 전체를 몇 개의 집단으로 그룹화하여 각 집단의 성격을 파악하고 데이터에 대한 이해를 돕고자 하는 분석 방법
K-Means	군집의 수를 사전에 정하고, 각 개체를 가까운 초깃값에 할당해 군집을 형성하고 각 군집의 평균을 재계산하여 초깃값을 갱신하는 과정을 반복하여 k개의 최종군집을 형성하는 방법
DBSCAN	밀도 기반 군집분석으로 서로 인접한 데이터들은 같은 군집 내에 있다는 것을 가정한 알고리즘

분석모형 정의(본문 243페이지)

변수 선택법	내용
전진 선택법	상관관계가 큰 변수부터 순차적으로 모형에 추가하여 변수를 추가하는 방법
후진 제거법	모든 독립변수가 추가된 전체 모형에서 상관관계가 작은 변수부터 제거해나가는 방법
단계적 선택법	전진 선택법으로 상관관계가 높은 변수를 추가하면서 중요도가 작은 변수를 후진제거법으로 제거하는 혼합방식

회귀분석(본문 260, 264페이지)

선형회귀분석의 기본가정

가정	설명
선형성	독립변수와 종속변수의 관계가 선형 $E(Y_i) = E(\alpha + \beta x + \epsilon_i) = \alpha + \beta x + E(\epsilon_i)$
등분산성	독립변수와 무관하게 잔차들의 분산이 일정 $Var(\epsilon_i) = \sigma^2$
독립성	입력변수와 오차는 관련 없음 $Cov(\epsilon_i, \epsilon_j) = 0 \ (i \neq j)$
비상관성	• 오차들끼리 상관이 없음 • 오차항들은 서로 독립적이며 그들의 공분산은 0
정규성	잔차항이 정규분포를 따름 $Y \sim N(\alpha + \beta x, \sigma^2)$

분류모델 성능 평가 방법

		Predicted class: Positive	Predicted class: Negative
Actual class	Positive	TP	FN
	Negative	FP	TN

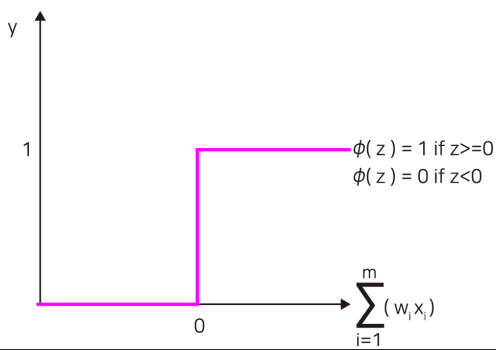
* TP(True Positive): 실제값과 예측치 모두 True
* TN(True Negative): 실제값과 예측치 모두 False
* FP(False Positive): 실제값은 False, 예측은 True
* FN(False Negative): 실제값은 True, 예측은 False

인공신경망(본문 273, 275페이지)

구조	설명
가중치	신경계 노드와의 연결계수 $w_0 \sim w_n$
활성함수	• 순입력함수로부터 전달받은 값을 출력값으로 변환해 주는 함수 • 입력받은 값을 얼마나 출력할지 결정하고, 출력된 신호의 활성화 여부를 결정
입력값	입력 데이터 $w_0 \sim w_n$
순입력함수	입력값에 가중치를 곱한 값을 모두 더해 하나의 값으로 만드는 함수

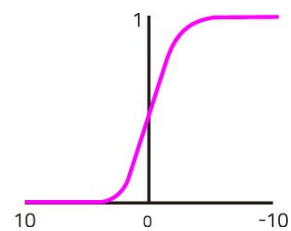
뉴런의 활성화 함수

① 계단(Step) 함수

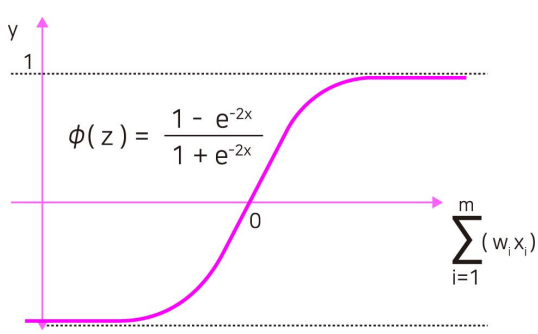


② 시그모이드(Sigmoid) 함수

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

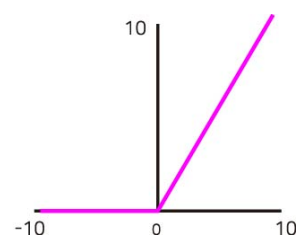


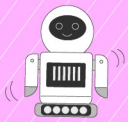
③ 하이퍼블릭 탄젠트(Hyperbolic Tangent) 함수



④ 렐루(ReLU: Rectified Linear Unit) 함수

max(0,x)





4과목

빅데이터 결과 해석

평가지표(본문 329페이지)

혼동행렬(confusion matrix)

실전 Tip

혼동행렬을 활용한 평가지표는 빈출되는 문제입니다.

실제값(Reference)

예측값(Reference)	Y	N
Y	True Positive(TP)	False Positive(FP)
N	False Negative(FN)	True Negative(TN)

구분	내용
TP(True Positive)	옳은 것을 옳다고 예측한 것
TN(True Negative	틀린 것을 틀리다고 예측한 것
FP(False Positive)	틀린 것을 옳다고 예측한 것
FN(False Negative)	옳은 것을 틀리다고 예측한 것

혼동행렬을 통한 분류모형의 평가지표

평가지표	공식	설명
정확도 (Accuracy)	$\frac{TP+TN}{TP+TN+FN+FP}$	전체 중 True를 True라고 옳게 예측한 경우와 False를 False라고 예측한 경우 예측모형의 전체적인 정확도를 평가
재현율(Recall) = 민감도(sensitivity)	$\frac{TP}{TP+FN}$	실제 True인 것 중에서 모델이 True라고 예측한 비율
정밀도 (Precision)	$\frac{TP}{TP+FP}$	모델이 True라고 분류한 것 중에서 실제 True인 비율
특이도 (Specificity)	$\frac{TN}{TN+FP}$	실제 False인 data 중에서 모델이 False라고 예측한 비율
거짓 긍정률 (False Positive Rate)	$\frac{FP}{TN+FP}$	실제 False인 data 중에서 모델이 True라고 예측한 비율
F1-Score	$2\times\frac{precision\times recall}{precision+recall}$	정밀도와 재현율의 조화평균으로 어느 한쪽으로 치우치지 않는 수치를 나타낼 때 F1 Score는 높은 값을 가짐

적합도 검정(본문 340페이지)

적합도 검정기법의 종류

실전 Tip

카이제곱 검정의 분류는 [동독 적] 동독의 적이다.

검정	내용
동질성 검정(Test of homogeneity)	두 집단의 분포가 동일한지 검정
독립성 검정(Test for independence)	두 개 이상의 변수가 독립인지 또는 상관 있는지 검정하는 방법
적합도 검정(Goodness of fit test)	어떤 모집단의 표본이 그 모집단을 대표할 수 있는지 검정하는 방법

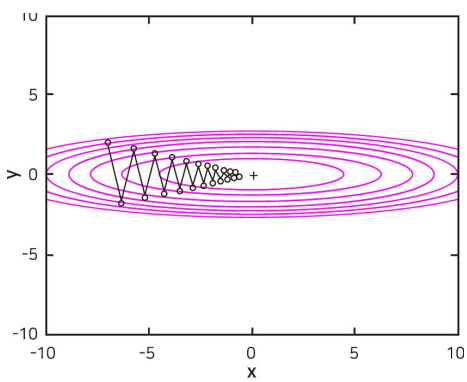
매개변수 최적화(본문 345페이지)

실전 Tip

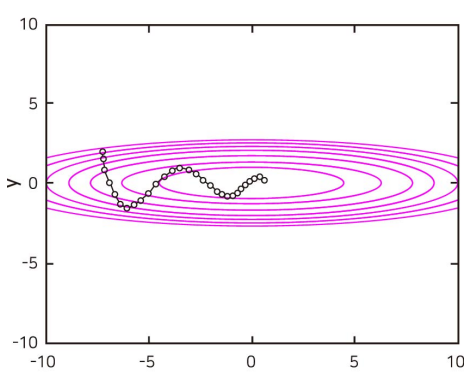
매개변수의 종류

- 가중치: 각 입력값에 각기 다르게 곱해지는 수치
- 편향: 하나의 뉴런에 입력된 모든 값을 다 더한 값에 더해 주는 상수

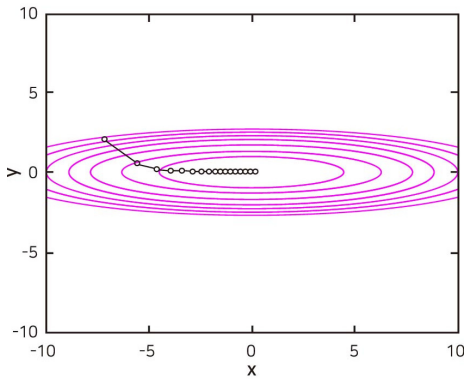
① 확률적 경사 하강법(SGD: Stochastic Gradient Descent)



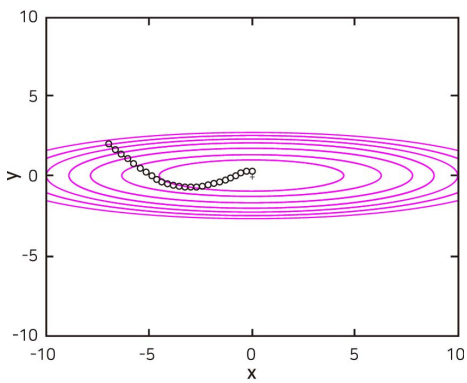
② 모멘텀(momentum)



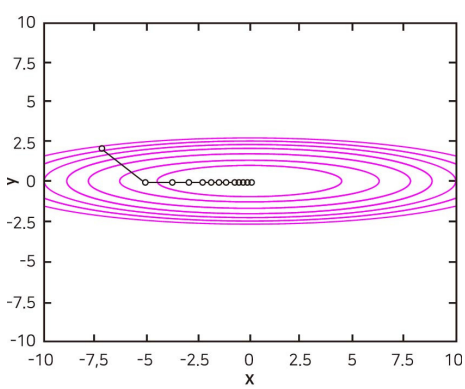
③ AdaGrad(Adaptive gradient)



④ Adam(Adaptive Moment Estimation)

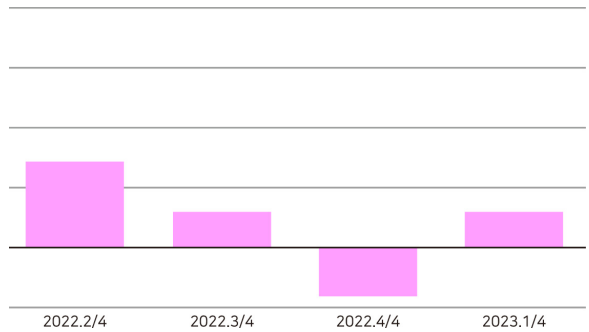
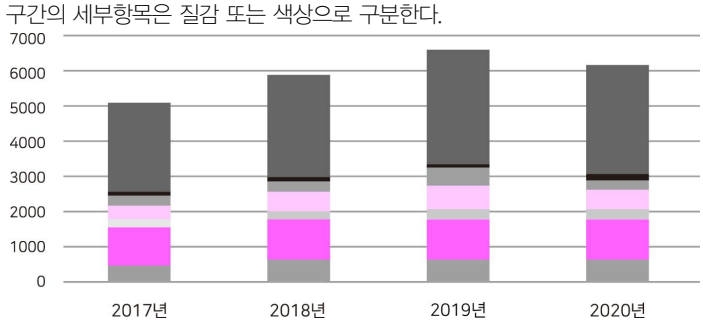
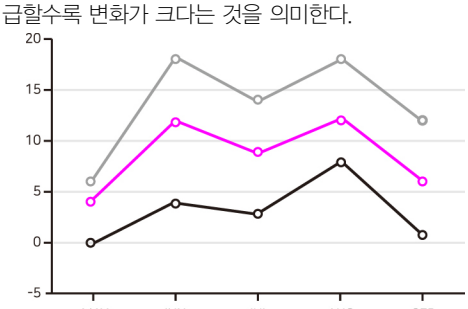
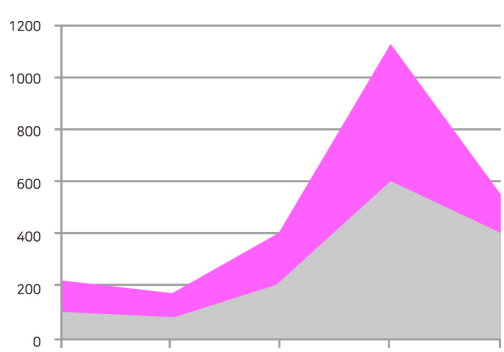


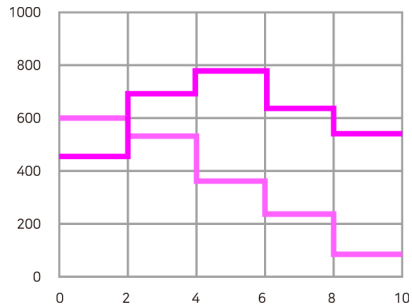
⑤ RMSProp(Root Mean Square Propatatio)



시간 시각화(본문 362페이지)

시간 시각화의 유형

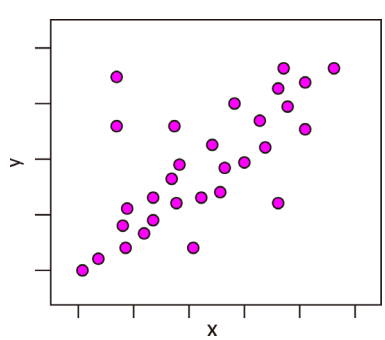
종류	내용
막대 그래프 (Bar Graph)	• 수치를 길이로 표현해 여러 값의 상대적인 차이를 알아보는 그래프이다. • 시간축은 시간 순서대로 정렬된 시간의 특징시점을 나타내며, 값 축은 그래프의 크기 범위를 나타낸다. 
누적 막대 그래프 (Stacked Bar Graph)	• 한 구간이 몇 개의 세부 항목으로 나뉘면서도 전체의 합이 있을 때 사용한다. • 한 구간의 세부항목은 질감 또는 색상으로 구분한다. 
선 그래프 (Line Graph)	• 연속적인 데이터의 끊임없이 변화하는 현상의 추이를 볼 수 있다. • 선의 기울기가 급할수록 변화가 크다는 것을 의미한다. 
영역차트 (Area Chart)	선 그래프와 막대 그래프를 결합하여 시간 경과에 따른 수량 변화를 표시하는 그래프이다. 

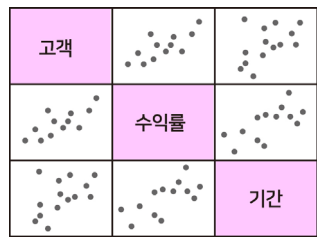
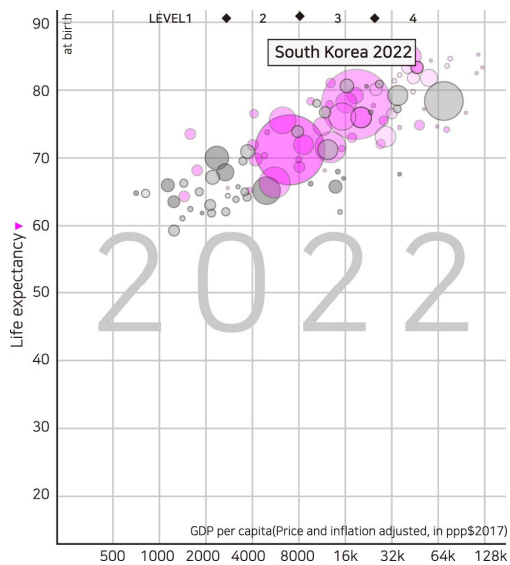
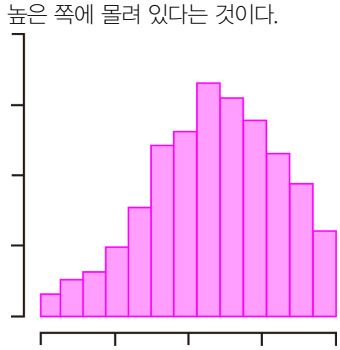
계단식 그래프 (Step Line Graph)	x축에 대한 y축의 변화 모습을 나타낼 때 효과적인 그래프이다. 

분포 시각화(본문 364페이지)

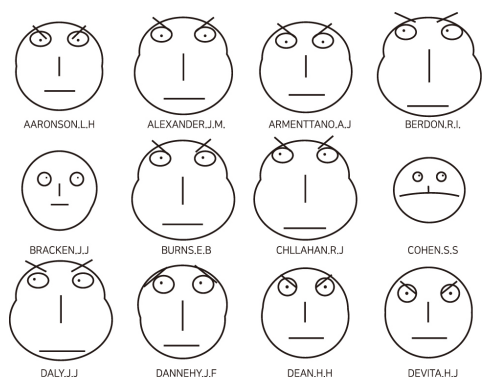
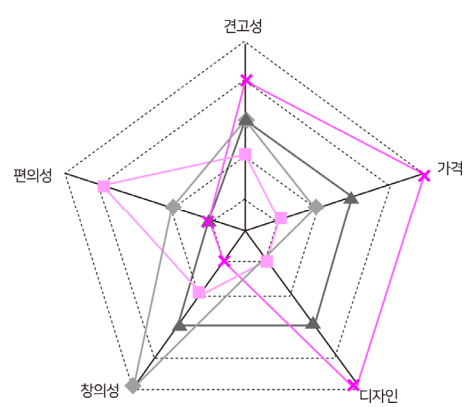
종류	내용
원그래프 (pie chart)	• 부분과 전체, 부분과 부분 간의 비율을 알아보는 데에 사용된다. • 분포의 정도를 총합 100%로 나타내서 부분 간의 관계를 보여주며, 면적으로 값을 보여주며, 수치를 각도로 표현한다.
도넛차트 (donut chart)	중심부를 잘라낸 원 모양으로 조각에 해당하는 수치는 조각의 길이로 표시된다.
트리맵 (Treemap)	• 영역기반의 시각화 방법으로, 각 사각형의 크기가 수치를 나타낸다. • 위계 구조가 있는 데이터나 트리 구조의 데이터를 표시할 때 사용된다.
누적 연속그래프	• 가로축은 시간을 나타내며 세로축은 데이터값을 나타낸다. • 한 시점의 세로 단면을 보면 그 시점의 분포를 확인할 수 있다.

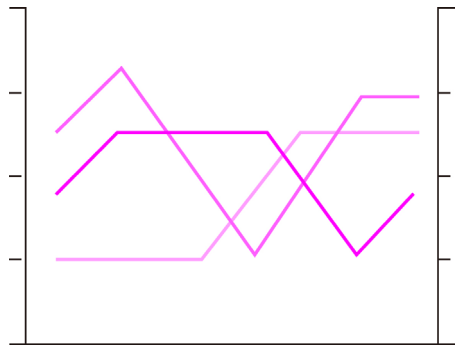
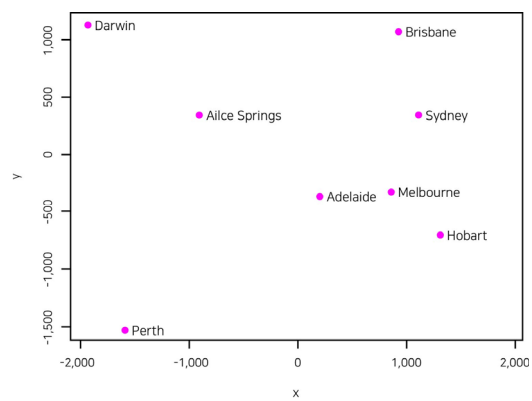
관계 시각화(본문 364페이지)

종류	내용
산점도 (Scatter Plot)	• 각 점은 관측치를, 점의 위치는 관측값을 나타낸다. • 각 데이터들의 상관성 여부를 파악하는 데 유용하다. 

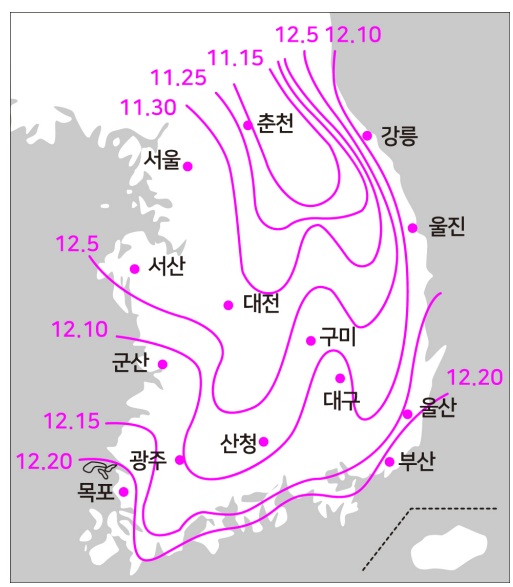
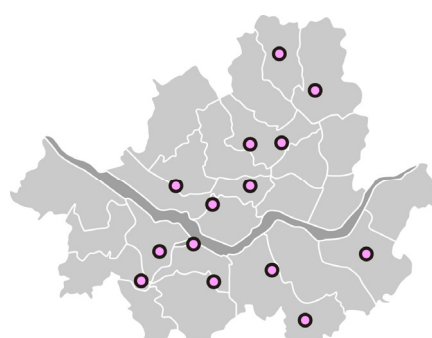
산점도 행렬 (Scatter Plot Matrix)	다변량 변수 데이터에서 가능한 모든 변수 쌍에 대한 산점도를 행렬 형태로 표현한 것이다. 
버블차트 (Bubble Chart)	• 세 가지 요소의 상관관계를 표현할 수 있는 방법이다. • 가로 및 세로축의 위치와 버블의 면적으로 표현된다. *출처: https://www.gapminder.org/ 
히스토그램 (Histogram)	• 세로축은 데이터의 분포 정도를 표현하고, 가로축은 특정 변수의 구간 폭을 의미한다. • 왼쪽으로 치우쳤다면 데이터가 전체 범위에서 수치가 낮은 쪽에 있고, 오른쪽에 치우쳤다면 높은 쪽에 몰려 있다는 것이다. 

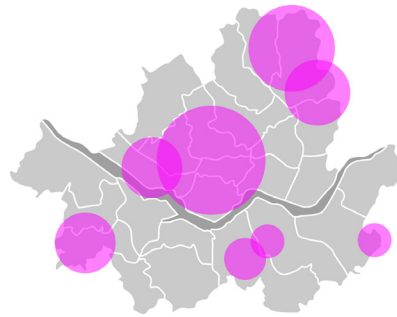
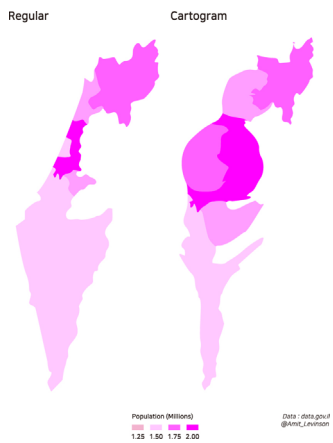
비교 시각화(본문 366페이지)

종류	내용
히트맵 (Heatmap)	• 한 칸의 색으로 데이터값을 표현한 그래프이다. • 주로 웹로그 분석 등에 활용된다. 
체르노프 페이스 (Chernoff Faces)	• 데이터를 사람의 얼굴 이미지로 표현한 그래프이다. • 얼굴의 가로 및 세로 길이, 눈, 코, 입 등 각 분위를 변수로 대체하여 데이터의 속성을 파악할 수 있다. 
스타차트 (Star chart)	• 차트 중앙에서 외부 링까지 이어지는 몇 개의 축을, 전체의 공간에서 하나의 변수마다 축 위의 중앙으로부터의 거리로 수치를 나타낸다. • 거미줄 차트 또는 방사형 차트라고도 한다. 

평행좌표계 (Parallel Coordicates)	• 한 축에서 윗부분은 변수값 범위의 최댓값을 아래는 변수값 범위의 최솟값을 나타낸다. • 대상이 많은 데이터에서 집단적 경향성을 쉽게 알아보게 해 준다. 
다차원 척도법 (Multidimensional scaling,MDS)	• 데이터상의 거리를 바탕으로 이들 간의 관계 구조를 시각적으로 표현하는 통계 데이터 분석기법이다. • 대상 간의 유사성측도에 의거하여 대상을 다차원 공간에 배치시키는 것으로 유사성이 작으면 멀리, 크다면 가까이 배치한다. 

공간 시각화(본문 368페이지)

종류	내용
단계 구분도 (Choropleth Map)	• 여러 지역에 분포되어 있는 정보를 나타낼 때 데이터가 분포된 지역별로 색을 다르게 나타낸 지도이다. • 데이터값의 크기에 따라 채도 및 밝기에 차등을 준다. 
등치선도 (Isometric Map)	• 데이터 왜곡이 될 수 있는 등치지역도의 결점을 극복한 그래프이다. • 데이터값 크기에 따라 색상 농도 변화를 활용하여 표현한다. 
도트맵 (Dot Map)	• 데이터를 지도위에 점으로 표현한 그래프이다. • 시간 경과에 따른 확산 등을 나타낼 때 사용한다. 

버블맵 Bubble Map	지도 위에 데이터를 그 크기에 따라 서로 다른 크기의 원형으로 표시한 그래프이다. 
카토그램 (Cartogram)	데이터값의 변화에 따라 지도의 면적을 인위적으로 왜곡하여 나타낸 것이다. 

분석결과 활용 시나리오 개발(본문 374페이지)

서브퀄모형(SERBQUAL)

- 확신성
- 신뢰성
- 반응성
- 공감성
- 유형성

실전 Tip

서브퀄 모형의 평가지표는 [확신 반 공유] 확신할 수 없어서 반만 공유했다.