Data Science

데이터 분석 과정 2-(1)

2023년 5월 3일~5월 27일 (4주)

권오준

(ojkwon@deu.ac.kr)

내용

- 데이터 분석 개요
- 탐색적 분석
- 데이터 전처리
- 클러스터링
- 선형 회귀
- 선형 분류
- 로지스틱 회귀

데이터 분석 개요

데이터 분석의 목적

- 데이터 분석
  - 컴퓨터가 데이터를 분석하여 중요한 의미를 추출하거나 미래를 예측하는 모델을 만드는 기술

- 데이터 분석을 사용하는 목적
  - 예측 (Prediction)
  - 설명 (Description)
  - 추천 (Recommendation)

예측 (predictive 분석)

- 새로운 샘플에 대한 미래 값을 예측
  - 회귀 (regression)
  - 분류 (classification)

- 회귀 : 수치를 예측
  - 내일의 날씨 예측
  - 주가 예측
  - 병에 걸릴 확률 예측
  - 다음달 가게의 매출 예측

- 분류 : 주어진 샘플이 어느 카테고리에 속하는 지를 예측
  - 수신한 메일이 스팸인 지 여부 예측
  - 이번 은행 대출이 부도가 날지 아닐지 여부 예측
  - 누가 우수 고객인지 예측

설명 (Descriptive 분석)

- 어떤 현상의 원인을 데이터 분석을 통해 설명
  - 고객, 비즈니스 프로세스, 성과 등 데이터를 이해하는 것
  - 예)
    - 어떤 상품이 많이 팔렸다면 그 이유를 파악
    - 고객 리서치
    - 고객의 행동전환 파악
    - 탐색적 분석

- 슈퍼마켓에서 어떤 품목들이 자주 같이 판매되는 지 패턴을 찾아 내는 것은 일종의 서술형 모델
  - 유사한 특성을 가진 항목들을 함께 묶는 군집화

- 결과
  - 새로운 인사이트를 얻는 것

추천 (Recommendation)

- 최종적으로 의사결정을 돕는 것
  - 단순히 정보(인사이트)를 주는 것을 넘어서 최적의 추천
  - 다양한 설명 및 예측 모델을 종합적으로 활용

- 예
  - 약의 처방, 네비게이터, 검색엔진, 상품/영화/음악 추천
  - 자율차의 운행
  - 알파고와 같은 게임 플레이어,
  - 보험 사기청구 거절 등

심슨 패러독스(Simpson's Paradox)

두 도시에서의 A사, B사의 불량률 비교와 전체 불량률 비교

도시	A사	B사
서울	정상품 90, 불량품 10 (불량률 10%)	정상품 920, 불량품 80 (불량률 8%)
춘천	정상품 980, 불량품 20 (불량률 2%)	정상품 99, 불량품 1 (불량률 1%)
전체	A사 총 불량률 30/1,100 = 3%	B사 총 불량률 81/1,100 = 8%

- **같은 데이터를 가지고도 통계 분석을 통한 결과 해석 방법에 따라 서로 상반된 결과를 얻을 수 있다.**
  예) 각 그룹의 데이터에서 나타나는 특징이 결합되었을 때 달라지는 현상

데이터 분석의 유형

- 지도 학습(Supervised learning)
- 비지도 학습 (Unsupervised learning)
- 강화 학습 (Reinforcement learning)

데이터 분석의 유형 – 지도 학습

- 지도 학습(Supervised learning)
  - 입력 값(x)과 정답(y, label)를 포함하는 훈련용 데이터(training data)를 이용하여 학습하고, 그 학습된 결과를 바탕으로 미지의 데이터(test data)에 대해 미래 값을 예측(predict)하는 방법
  - 회귀나 분석 등 예측 모델은 시간이 지나면 정답을 확인할 수 있고, 모델의 성능에 대한 정확한 평가가 가능
  - 정답에 해당하는 값 : 목적변수(target variable), 레이블(label)
    - 회귀 : 수치 값
    - 분류 : 카테고리 변수
  - 예) 스팸 메일 분류기의 학습
    - 수집한 데이터로부터 어떤 메일이 스팸이었는 지 정답 샘플도 같이 주어져야 한다.

데이터 분석의 유형 – 지도 학습

- 회귀 분석
  - 수치를 예측하는 것

- 회귀 분석의 응용
  - 경제지표 예측
  - 사회학 연구
  - 마케팅
  - 의학에서 치료효과 분석

- 회귀 분석 알고리즘
  - 선형회귀
  - KNN
  - SVM
  - 로지스틱 회귀
  - 랜덤 포레스트
  - 신경망

데이터 분석의 유형 – 지도 학습

- 분류
  - 어떤 항목(item)이 어느 그룹에 속하는지를 판별
  - 이진 분류(binary classification)
    - 두 가지 카테고리를 나누는 작업
  - 다중 분류(multiclass classification)
    - 세 개 이상의 클래스를 나누는 작업

- 분류의 응용
  - 스팸 메일/우수 고객/충성심 높은 신입사원/투자할 좋은 회사 구분
  - 매장 입장 고객의 타입 분류
    - 물건을 구매, 단순히 구경, 항의 고객인지 판단하여 적절한 대응
  - 과거의 구매 이력/SNS 등을 분석하여 구매 확률이 높은 고객 구분
    - 광고 안내문, 기념품을 잠재 고객에게 보낼 때 필요

데이터 분석의 유형 – 비지도 학습

- 비지도 학습(Supervised learning)
  - 정답(label)은 없고 입력 데이터만 있는 훈련용 데이터(training data)를 이용한 학습을 통해 정답을 찾는 것이 아닌 입력 데이터의 패턴, 특성 등을 발견하는 방법
  - 데이터의 특성을 기술하는 서술형 모델

- 기법
  - 군집화(clustering)
    - 유사한 항목들을 같은 그룹으로 묶는다
  - 시각화
    - 데이터의 속성을 명확하게 시각화하기 위해서 고차원의 특성 값들을 2차원이나 3차원으로 차원을 축소하는 작업
  - 데이터 변환
    - 데이터를 분석하기 좋게 다른 형태로 변환
  - 주성분 분석(PCA)
    - 머신 러닝에 사용할 특성의 수를 줄인다.

데이터 분석의 유형 – 비지도 학습

- **연관 분석**
  - 어떤 사건이 다른 사건과 얼마나 자주 동시에 발생하는지 파악
  - 자주 발생하는 패턴 찾기(상품의 연관성, 취향의 연관성 등 분석)
  - 같이 구매한 상품 분석(market basket analysis, 장바구니 분석)
  - 상품의 진열 배치 및 상품 프로모션(쿠폰 발행 등)에 활용

데이터 분석의 유형 – 강화 학습

- 강화 학습(Reinforcement learning)
  - 입력 샘플마다 정답이 있어 답을 알려주는 것이 아니라 일정 기간 동안의 행동(action)에 대해 보상(reward)을 해 줌으로써 어느 방향으로 학습해야 하는 지 방향성만 알려주는 학습 방법

- 응용 예
  - 게임의 경우 매 입력시마다 답을 주지는 못하지만, 게임을 이기고 있는 지, 지고 있는 지를 알려 줌
    - 스스로 게임을 잘 수행하는 방법을 터득
  - 로봇이 혼자 그네 타는 방법, 바둑 두는 방법을 터득
  - Alphago(바둑 프로그램)

데이터 분석의 특징

- 예전에는 컴퓨터는 프로그래머가 코딩한 대로 동작
- 데이터 분석에서는 컴퓨터가 데이터를 보면서 점차 성능을 향상
  - 컴퓨터가 데이터를 보고 스스로 기능을 향상시키는 방법으로 학습
- 데이터 분석은 예측이나 설명을 위해 모델을 사용
- 모델 예
  - 스팸을 찾아내는 모델
  - 누가 게임에 이길지 예측하는 모델
  - 내일 날씨를 예측하는 모델

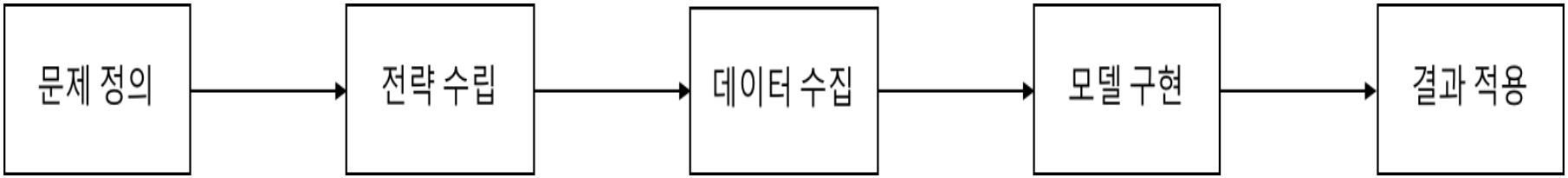
데이터 분석의 유형 – 머신 러닝

	머신러닝 유형	알고리즘
지도학습	분류	kNN, 베이즈, 결정 트리, 랜덤 포레스트, 로지스틱 회귀, 그라디언트부스팅, 신경망
	회귀	선형 회귀, SVM, 신경망
비지도학습	군집화	k-means, DBSCAN
	데이터 변환	스케일링, 정규화, 로그변환
	차원축소	PCA, 시각화

데이터 분석의 특징

- 예전에는 컴퓨터는 프로그래머가 코딩한 대로 동작
- 데이터 분석에서는 컴퓨터가 데이터를 보면서 점차 성능을 향상
  - 컴퓨터가 데이터를 보고 스스로 기능을 향상시키는 방법으로 학습
- 데이터 분석은 예측이나 설명을 위해 모델을 사용
- 모델 예
  - 스팸을 찾아내는 모델
  - 누가 게임에 이길지 예측하는 모델
  - 내일 날씨를 예측하는 모델

데이터 분석 프로세스



- 문제 정의 – 해결하려는 문제를 명확히 정의하는 것
- 전략 수립 – 문제 해결을 위해 어떤 데이터를 어떻게 사용할지를 정함
- 데이터 수집 – 머신 러닝에 필요한 데이터를 수집하는 것
- 모델 구현 – 분류, 회귀, 설명, 추천 등을 위한 머신 러닝 모델을 구현
- 결과 적용 – 머신 러닝 모델을 실제 상황에 적용하고 성능을 개선하는 것

데이터 분석 프로세스 – 문제 정의

- 가장 먼저 수행, 가장 중요한 단계
- 문제를 명확하게 구체적으로 정의하는 것
- 주어진 문제가 데이터를 확보, 분석하면 과연 해결할 수 있는 문제인 지를 먼저 파악해야
  - 문제가 아닌 것을 해결하려고 시도하는 경우가 의외로 많음
  - 아무리 데이터를 수집해도 답을 얻을 수 없는 문제도 있다
- 문제는 조건, 조직, 시기, 장소 등에 따라 다름
- 최종적으로 해결해야 할 큰 문제를 정의하는 것이 필요하지만, 데이터 분석 모델이 해결할 수 있는 작은 문제로 나누어서 접근하는 전략 필요
  - 예) 기업의 수익 감소 문제 : 한번에 해결이 아닌 현실적인 작은 문제로 나눔
    - 배송이 늦어지는 이유
    - 고객 불만이 발생하는 원인
    - 반품이 많은 이유

데이터 분석 프로세스 – 전략 수립

- 수행 전략 수립을 위한 사전 파악
  - 관련 부서는 어디인 지?
  - 프로젝트는 누가 주도할 지?
  - 필요한 데이터는 누가 가지고 있는 지?
  - 주어진 예산과 기간은 얼마인 지?
- 전략을 잘 세우는 첫 단계 : 문제를 다시 검토하는 것
  - 왜 그것이 문제인지? 문제의 본질을 다시 생각
  - 고객의 근본적인 고민이 무엇인지
  - 현재는 어떤 방법을 사용하고 있는 지
  - 현재 방법이 왜 잘 동작하지 않는 지
- 문제 해결의 목표를 명확히 구체적으로 설정
  - 막연히 비용 절감, 좋은 아이디어 도출, 서비스 개선 등 불분명한 목표가 아닌
  - 비용을 몇 % 줄일 지 등 구체적으로 목표 설정

데이터 분석 프로세스 – 전략 수립

- 가용자원 파악
  - 사용할 수 있는 데이터에는 무엇이 있는 지
  - 인적 자원은 어떤 지
  - 데이터 수집 비용
  - 데이터를 가지고 있는 부서와 협조 현황
  - 기타 프로젝트 수행 관련 현황 및 정보
- 관련 데이터 목록 작성
  - 직접 필요한 데이터 뿐만 아니라 조금이라도 관련된 데이터 목록
  - 그들 데이터에는 어떤 것들이 들어 있는 지
- 의외로 예상하지 못한 데이터로부터 유용한 도움 받을 수도
- 직접 도움이 되는 데이터를 구하기 어려운 경우
  - 우회적으로 어떤 데이터로부터 원하는 정보를 추정할 수 있는 지

데이터 분석 프로세스 – 데이터 수집

- 데이터 분석에 필요한 데이터를 실제로 준비하는 과정
  - 어떤 데이터가 수집 가능한 지
  - 실제로 수집되는 데이터가 쓸만한 지
  - 이 정도 데이터 분량이면 충분한 지
  - 데이터의 품질은 만족할만 한 지
- 전체 과정에서 70~80%의 시간을 소모함
- 핵심 데이터를 확보했는지 여부
- 데이터 품질 상태
- 잘못된 데이터 사용은 잘못된 결과를 도출

데이터 분석 프로세스 – 데이터 분석 모델 구현

- 최종 데이터 분석 모델 구조를 결정하기 전에 일부 샘플 데이터를 사용해서 여러 가지 데이터 분석 모델을 시도할 필요
- 처음부터 완벽한 모델을 만들려고 하지 말고, 기본적인 동작만 수행해 보고 상세한 모델은 더 많은 데이터를 학습하면서 개선
- 초기 결과를 보고 세부 내용 결정
  - 수집해야 할 데이터의 종류가 달라질 수도 있고
  - 얼마나 상세한 데이터가 필요한 지도 달라질 수도 있다
    - 예) 하루에 한 번의 평균값만 있으면 될 분석에 1분마다 측정한 데이터를 사용할 필요는 없다

데이터 분석 프로세스 – 결과 적용

- 데이터 분석의 최종 목적 : 모델을 실제 상황에 적용하는 것
- 모델을 실제 어떻게 이용할 지도 미리 시뮬레이션해 봐야
- 데이터 분석 모델
  - 한번에 만족할 만한 성과를 내지 못하는 경우 많다
  - 실전에 적용하면서, 즉 새로운 데이터가 계속 추가 입력되면서 성능이 진화
  - 모델이 진화하도록 설계되지 않았다면 제대로 설계, 구현되지 못한 것

탐색적 분석

탐색적 분석 정의

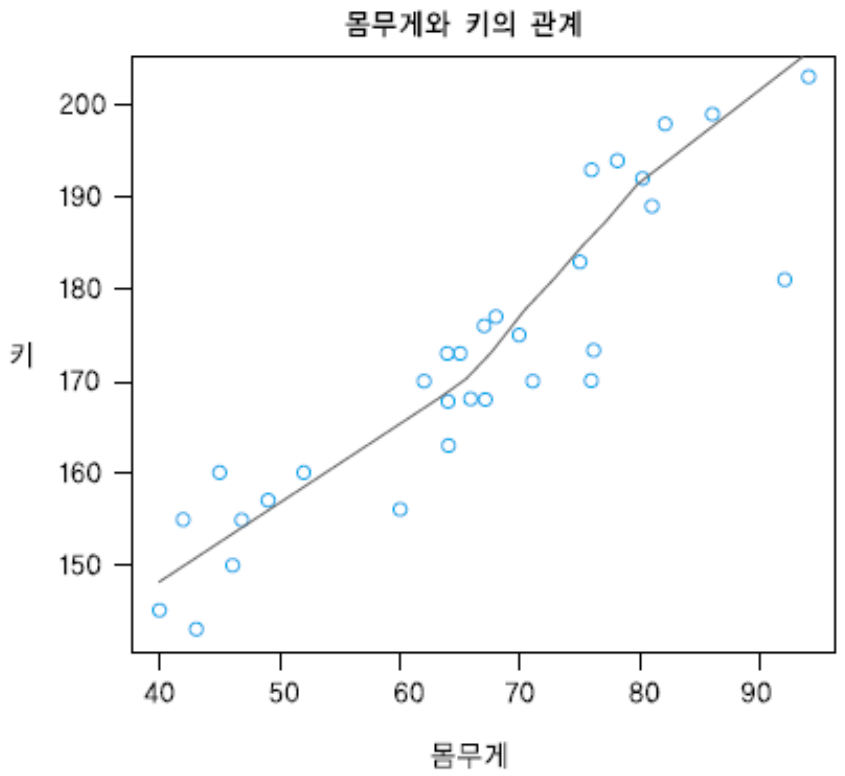
- 본격적인 데이터 분석에 앞서 수집한 데이터가 분석에 적절한지 알아보는 과정
- 수집한 데이터의 전체적인 특성을 분석
  - Exploratory Data Analysis: EDA
- 데이터를 시각화하여 그래프로 그려보는 방법을 기본으로 사용
  - 시각화 도구 이용
  - 히스토그램, 박스 플롯, 막대 그래프, 스캐터 플롯 등
- 기본적인 통계적 특성 파악
  - 숫자형 데이터의 평균, 최대값, 최소값, 표준편차, 분산 등

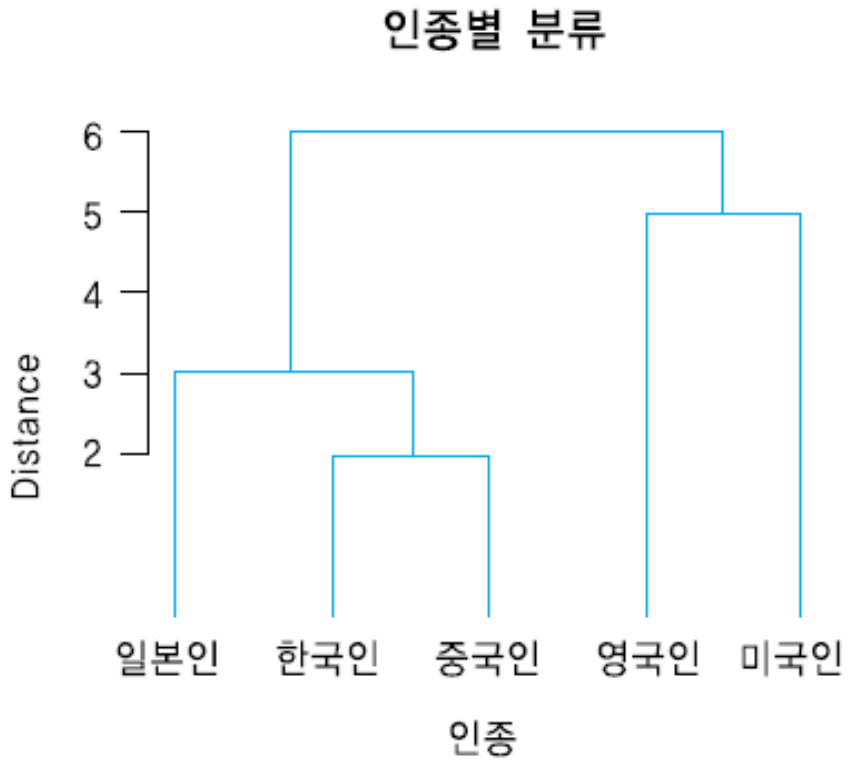
탐색적 분석 정의 – 데이터 시각화

- 데이터 시각화(visualization)
  - 그래프, 도표, 도형 등을 이용 하여 데이터의 특징을 파악하게 하는 것
- 숨어 있던 새로운 의미를 찾아낼 수 있음
- 데이터 탐색 뿐만 아니라 분석 결과를 고객에게 설명할 때에도 필수
- 위치, 길이, 각도, 방향, 형태, 면적, 부피, 명암, 색상 정보를 활용

탐색적 분석 정의 – 데이터 시각화

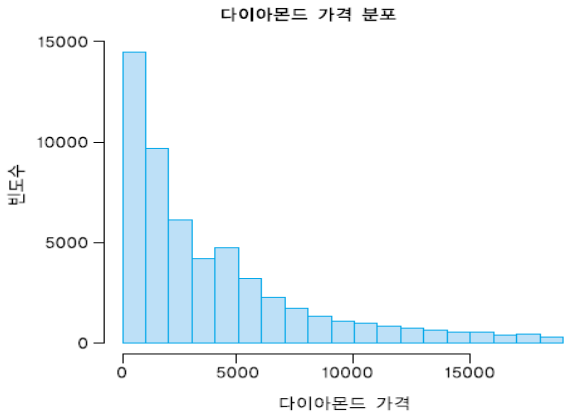
- **기본적 시각 모형 – 위치**
  - **산포도, 덴드로그램**

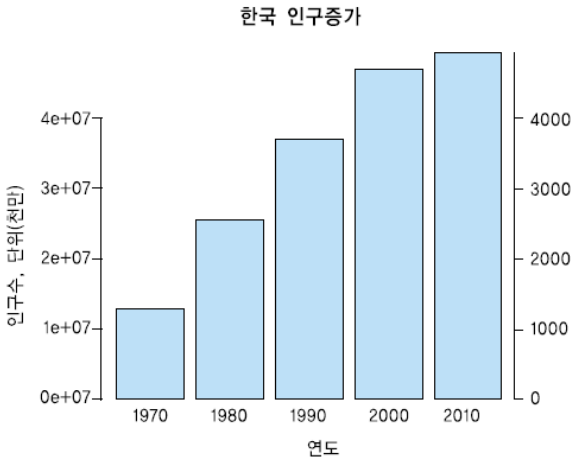


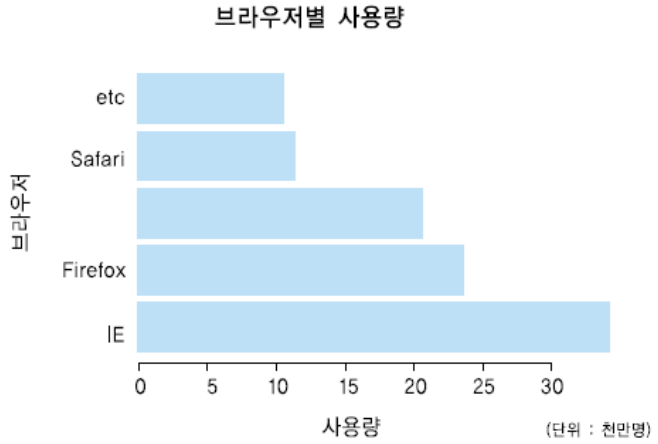


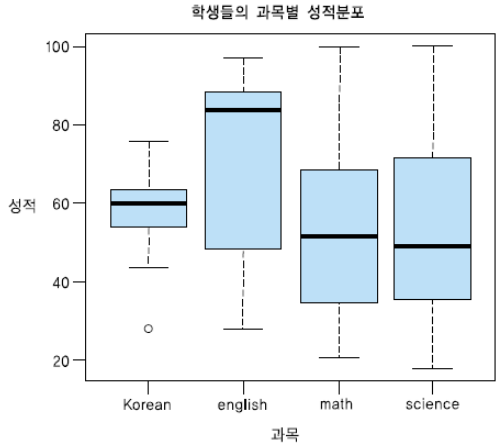
탐색적 분석 정의 – 데이터 시각화

- **기본적 시각 모형 – 길이**



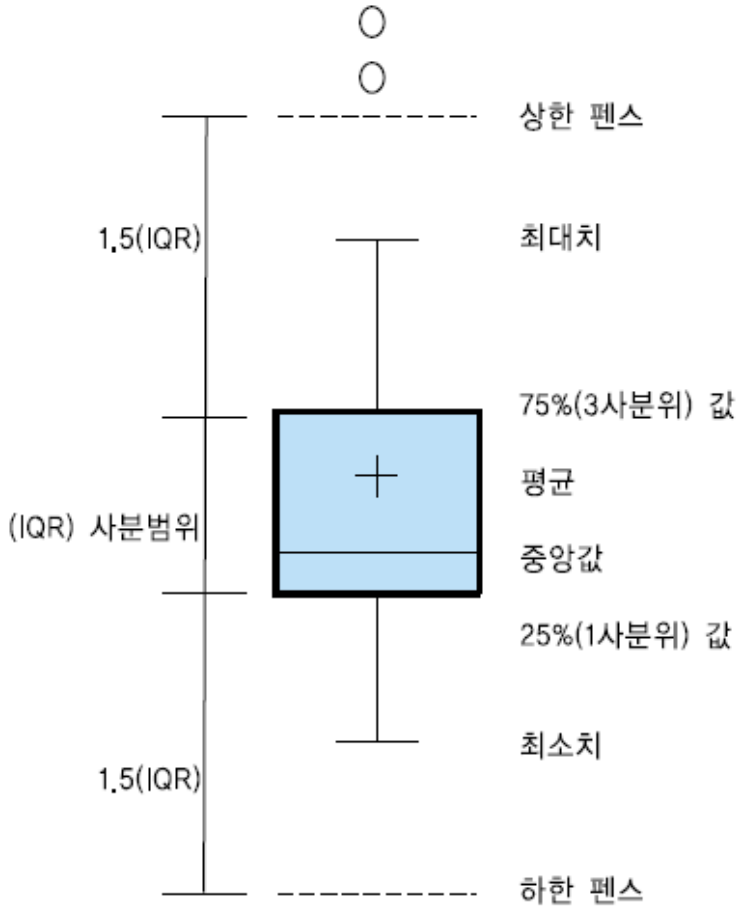






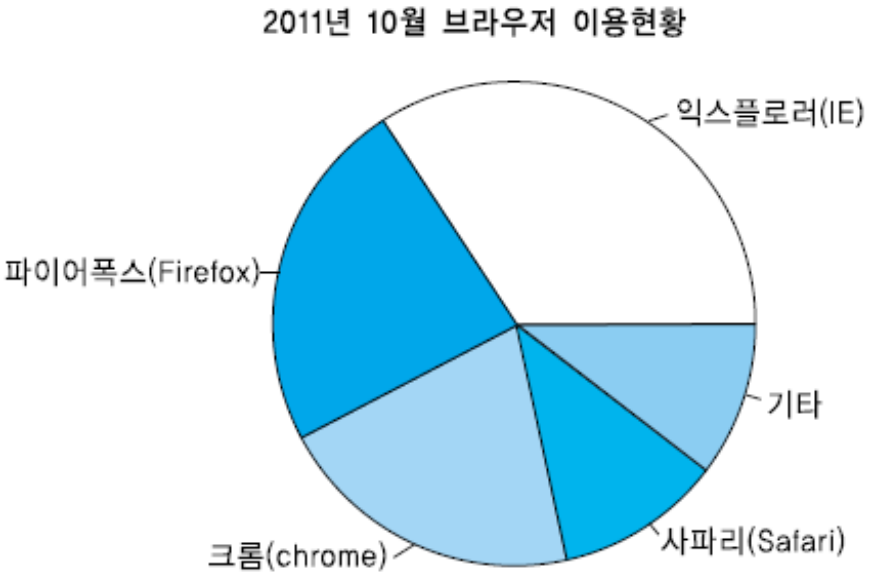
탐색적 분석 정의 – 데이터 시각화

- **기본적 시각 모형 – 박스 플롯**



탐색적 분석 정의 – 데이터 시각화

- **기본적 시각 모형 – 각도, 면적/부피**



우리나라 인구 밀도 히트맵 (출처: http://brainage.egloos.com/5769171)

탐색적 분석 정의 – 데이터 타입

- 데이터 : 4 가지 타입
  - 문자형: “Hello World”, “대한민국”, …
    - 이름, 주소, 텍스트 본문 등
  - 수치형: 1, 5, 10, 3.14, 0.9, …
    - 나이, 키, 금액, 온도, 습도 등 센서 측정값
  - 바이너리형: 0100100101010101…
    - 오디오, 비디오, 실행 파일 등 비트 단위로 구성된 파일
  - 논리형: True/False

탐색적 분석 정의 – 데이터 타입

- 수치형 데이터
  - 범주형(categorical)
    - 문자형으로도 표현되지만, 편의상 숫자로 대체하여 표현
    - (월요일=1, 화요일=2…), (남성=1, 여성=0)
    - 특정 클래스를 지칭 : 덧셈/뺄셈 등 연산은 의미가 없다
  - 순서형(ordinal)
    - 순서가 의미를 가진다
    - 여성의 옷 사이즈 (44, 55, 66 ..), 달력 (1일, 2일, 3일 ..)
    - 덧셈/뺄셈 등 연산은 의미가 없다
  - 연속형(continuous)
    - 숫자의 양이 어떤 의미를 가지는 데이터
    - 무게, 길이, 온도, 압력, 속도, 화폐 단위
    - 덧셈/뺄셈 등 연산의 결과가 동일한 연속형 데이터로 의미를 갖는다

탐색적 분석 정의 – 데이터 타입

- **정형과 비정형 데이터**

형식	내용
정형 (structured)	• 데이터의 포맷이 정해져 있는 데이터 • 서식이 정해진 데이터(엑셀의 표 등) • CSV(comma separated value) 파일 등 포맷이 일정
비정형 (unstructured)	• 미리 정해진 포맷을 가지지 않는 데이터 • 블로그, 트위터 데이터 등 임의의 문장 등으로 구성 • 오디오, 비디오 데이터
반정형 (semi-structured)	• 데이터 내부에는 논리적인 형식을 가지고 있으나 외형상으로는 데이터 포맷이 정형 데이터처럼 완전하게 정의되어 있지는 않은 데이터 • 센서 데이터, 웹 사용 기록 등

수고하셨습니다

Q & A