데이터 분석 - 개념

데이터란?

데이터의 개념

- 사실이나 관찰된 값, 측정 결과 또는 경험
- 기록하고 분석해 의미있는 정보를 도출할 수 있습니다.
- 데이터 그 자체로는 의미가 모호할 수 있으나, 올바르게 수집, 분석 활용할 때 유용한 지식으로 변환
- 여러 형태를 가집니다.
  - 숫자
  - 문자
  - 이미지
  - 소리
  - 기타...

데이터의 분류

구조에 따른 분류

- 구조화된 데이터
  - 데이터베이스와 같이 행과 열로 정리된 데이터
- 비구조화된 데이터
  - 영상, 이미지, 문서 등 정형화되지 않은 데이터
- 반구조화된 데이터
  - XML, JSON 등 형식 문서로 작성되었으나, 고정된 형식이 아닌 데이터

속성에 따른 분류

- 정성적 데이터
  - 질적 정보

    - 숫자가 아닌 설명이나 특징
    - “한국은 잘 사는 나라이다.”
- 정량적 데이터
    - 양적 정보
    - 숫자나 구체적 지표로 표현되는 설명이나 특징
    - 한국의 GDP
- 혼합 데이터
    - 질적 정보와 양적 정보가 혼재된 데이터
    - 대부분의 복합 데이터는 혼합 데이터에 속함

일반적인 데이터 분석 프로세스

1. 문제 정의
2. 데이터 수집
3. 데이터 정제
4. 데이터 탐색
5. 데이터 엔지니어링
6. 모델링
7. 데이터 분석
8. 결과 해석
9. 결과 도출
10. 결과 커뮤니케이션

도구들

소프트웨어 또는 분석 플랫폼

- 스프레드시트 (Excel, Google Sheets 등)
- 데이터 시각화 도구 (Tableau 등)
- SAS (Statistical Analysis System)

- 기타 다양한 소프트웨어

프로그래밍 언어

- Python
- R
- SQL

데이터 처리 및 분석 라이브러리

- Python
  - Pandas - 테이블 형식의 데이터를 효율적으로 처리하고 조작하는데 사용
  - NumPy - 다차원 배열을 다루기 위한 라이브러리로 효율적인 수학 연산을 처리
  - Scikit-learn - 머신러닝, 통계 등 데이터 과학 분야의 풍부한 기능을 제공하는 라이브러리
- R
  - dplyr - 데이터 프레임을 조작할 수 있는 R 패키지
  - tidry - 데이터 정리 및 변환 작업을 도와주는 R 패키지

데이터 시각화 도구

- Matplotlib, Seaborn (Python)
  - 파이썬에서 데이터를 시각화하는데 사용되는 대표적인 도구
  - Matplotlib은 기본적인 그래프 생성에 사용
  - Seaborn은 더 정교한 시각화, 통계적 시각화에 자주 사용
- ggplot2 (R) - R의 시각화 라이브러리
- Tableau - 프로그래밍 없이 대화형 대시보드와 보고서를 만들 수 있음

데이터베이스 및 쿼리 도구

- RDBMS
  - 관계형 연산을 지원하는 데이터 관리 도구
  - MySQL, PostgreSQL 등의 오픈소스 DBMS와 여러 종류의 상용 DBMS
- NoSQL DB
  - 키-값 형태의 데이터를 저장, 관리하는 데이터 관리 도구

    - MongoDB 등 비정형 데이터를 효율적으로 저장하고 처리할 수 있는 오픈소스 도구가 있음

빅데이터 분석 도구

- Hadoop - 대용량 데이터를 저장하고 분산 처리하는 오픈 소스 프레임워크
- Spark - 실시간 데이터 처리가 가능한 대용량 분석 도구
- Hive - Hadoop에서 SQL을 사용해 데이터를 처리할 수 있는 도구

클라우드 기반 데이터 분석 도구

- Google BigQuery - 구글 클래우드 플랫폼에서 제공하는 데이터 웨어하우스 솔루션
- AWS Redshift - 아마존 웹 서비스에서 제공하는 데이터 웨어하우스 솔루션
- MS Azure Synapse Analytics - 데이터 통합과 분석에 장점이 있는 MS Azure 클라우드 기반 솔루션

머신러닝 및 AI 도구

- TensorFlow - Google에서 개발한 오픈 소스 딥러닝 라이브러리
- Keras - TensorFlow 기반으로 동작하는 고수준 딥러닝 라이브러리
    - 요즘 TensorFlow를 사용한다는 말은 일반적으로 Keras로 개발하고 Tensorflow가 학습, 추론한다는 말과 동일한 의미
- PyTorch - Facebook에서 개발한 오픈소스 딥러닝 라이브러리

데이터의 품질

데이터 품질의 영향

- 데이터 분석의 정확성, 신뢰성에 직접적으로 영향을 미침
- 데이터 품질에 따라 잘못된 분석 결과가 초래될 수 있고
- 의사 결정에 부정적인 영향을 미칠 수 있음

데이터 품질 평가 지표

- 정확성 - 실제 현실을 얼마나 잘 반영하고 있는지를 의미
- 완전성 - 필요한 데이터가 누락없이 수집되어 있는지를 의미

- 일관성 - 동일한 데이터를 다양한 출처로부터 수집할 때, 그 값이 일관된 기준을 유지하는지를 의미
- 유효성 - 데이터가 사전에 약속된 형식, 규칙에 맞는지 의미
- 적시성 - 데이터가 얼마나 최신 정보를 반영하고 있는지를 의미
- 접근성 - 필요한 데이터를 사용할 수 있는지를 의미
- 중복성 - 중복된 데이터가 존재하는지를 의미

데이터의 품질을 확보하는 방법

- 문제 분석 및 데이터 확보 노력 (기본 중의 기본!)
- 데이터 정제
  - 데이터의 오류를 수정
  - 누락된 데이터 보충
  - 데이터의 형식을 변환
  - 중복 데이터의 제공
  - 유효성의 검증
- 정기적인 데이터 업데이트
- 데이터 검증 프로세스 확보

데이터의 선택

데이터 선택의 중요성

- 목표 달성 - 분석 목적에 맞는 데이터를 선택해야 정확하고 의미 있는 결과를 도출할 수 있음
- 데이터 품질 보장 - 정확하고 일관된 데이터를 선택해야 목표 품질 확보가 가능
- 분석 효율성 - 잘못된 데이터 선택은 분석 과정에 불필요한 작업이 발생
- 대표성 확보 - 분석 대상을 가장 잘 설명할 수 있는 데이터가 좋은 분석에 필수
- 결과의 신뢰성 강화 - 양과 질 측면에서 확보 가능한 좋은 데이터 만이 신뢰성있는 설명이 가능

데이터 선택에서 고려할 주요 요소

- 목표와의 연관성
  - 분석 목표에 맞는 데이터를 선택해야 함
  - 예) 범죄 발생률과 아이스크림 판매량
- 데이터 가용성
  - 아무리 좋은 데이터라도 사용할 수 있어야 함
  - 예) 작전 대상의 통화 내용
- 데이터의 품질
  - 정확성, 완전성, 일관성 등 데이터의 품질 기준을 충족하는 데이터 여부를 고려
- 대이터의 대표성
  - 선택한 데이터가 분석 대상을 대표할 수 있어야 함
  - 편향성 발생의 원인이 됨
  - 예) 연말 평가에서 회의 참석 횟수 데이터를 사용하는 경우
- 데이터의 양
  - 데이터의 양이 너무 적으면 분석의 신뢰성이 떨어질 수 있음
  - 예) 대한민국 국민을 전수 조사하면 완전하게 신뢰 가능한 분석이 가능할까?
- 데이터의 시계열
  - 시계열 데이터 - 데이터가 측정 시각의 함수로 표현되는 형태의 데이터
  - 특정 분석의 경우 시간의 흐름에 따른 상황의 변화를 파악하는 것이 중요
  - 예) 일기예보, 주가 예측, 장비 고장 예측
- 특징 선택
  - 데이터 자체의 선택 뿐 아니라, 데이터의 어떤 요소를 사용할 지도 중요
  - 기계 학습의 예
  - 예) MBTI
- 중복 및 불필요한 데이터 제외
  - 선택한 데이터 중 분석을 통하지 않고도 명백히 중복되었거나 불필요한 데이터는 미리 제외할 수 있음
  - 중복 또는 불필요하다고 판단되더라도, 경우에 따라서는 그렇지 않은 경우가 있으므로 주의

    - 예) 도로의 차량 통행량과 도로의 평균 속도 데이터

IQR을 사용한 이상 데이터 찾기

사분위 수

- 데이터를 어떤 값 기준으로 25%씩 나누었을 때 기준이 되는 4개의 숫자
    - 각각 Q1, Q2, Q3, Q4라고도 부르며, Q4는 가장 큰 값
- 1등부터 101등까지 있는 경우
    - 1등의 점수 - Q4
    - 51등의 점수 - Q2
    - 25등 또는 26등의 점수 - Q3
    - 76등 또는 77등의 점수 - Q1

사분위 범위 및 이를 활용한 이상 데이터 판정

- IQR = Q3 - Q1 - 데이터의 중간 50% 구간의 범위
- 일반적으로 Q1 - 1.5 * IQR 이하, Q1 + 1.5 * IQR 이상을 이상 데이터로 판정
    - 존 튜키가 제안 (Box Plot과 함께 제시)
    - “경험적으로 적절하더라”
    - 확인해보니 정규분포 기준으로 약 99.3%의 데이터가 이 범위 내에 위치

정규분포 기반 데이터 분석

- 데이터가 정규 분포를 따르는지 확인
- 정규 분포를 따르는 경우 이를 활용한 데이터 분석

정규 분포란?

- 대부분의 데이터가 평균 근처에 모여 있고
- 양쪽으로 갈 수록 빈도가 적어지는 대칭적인 분포
- 68-95-99.7의 법칙

    - 평균에서 `- 표준편차`와 `+ 표준편차` 사이에 데이터의 약 68%가 분포
    - 평균에서 `- 2표준편차`와 `+ 2표준편차` 사이에 데이터의 약 95%가 분포
    - 평균에서 `- 3표준편차`와 `+ 3표준편차` 사이에 데이터의 약 99.7%가 분포

정규 분포의 활용

- 가설 검정
    - 모집단의 평균에 대한 검정을 수행할 때 데이터가 정규 분포를 따른다는 가정 하에 검정을 수행
    - 평균이 주어진 모집단의 평균과 유의미한 차이가 있는지 판단하는 경우등에 활용
- 신뢰 구간
    - 데이터가 정규 분포를 따른다는 가정 하에 모집단의 평균이 특정 구간에 존재할 확률
    - 예를 들어 평균에 대한 95% 신뢰 구간은 `평균 - 1.96 * 표준편차 ~ 평균 + 1.96 * 표준편차`로 구할 수 있음
- 표준 정규 분포
    - 평균이 0이고 표준 편차가 1인 정규 분포
    - 데이터를 정규화해서 얻어지는 분포

정규 분포를 활용한 이상 데이터 찾기

- Z-점수
    - $z = \frac{x-\mu}{\sigma}$
- Z 점수가 -1 이하 또는 +1 이상이면 68% 기준 이상치
- Z 점수가 -2 이하 또는 +2 이상이면 95% 기준 이상치
- Z 점수가 -3 이하 또는 +3 이상이면 99.7% 기준 이상치
- 단, 양측 검정 및 정규분포 전제

상관관계 분석에 대해서

개념

- 두 변수 간의 연관성의 방향과 강도를 나타내는 관계

- 다음 세 가지 관계를 가짐
  - 양의 상관관계 - 한 변수가 증가할 때 다른 변수도 증가하는 경우의 관계
    - 키와 몸무게의 관계
  - 음의 상관관계 - 한 변수가 증가할 때 다른 변수는 감소하는 경우의 관계
    - 가격과 수요의 관계
  - 상관관계 없음
    - 키와 성적의 관계

상관계수

- 두 변수 간의 관계를 수치화
- 최대 1, 최소 -1의 값을 가지며 두 변수가 완전히 동조될 때 1, 완전히 반대로 동조될 때 -1의 값을 가짐 (피어슨 상관계수 기준)
- 대표적인 상관계수 계산 법
  - 피어슨 상관계수 - 두 변수 간의 선형 관계
  - 스피어만 상관계수 - 순위를 기준으로 한 상관계수
    - 선형 관계가 아니더라도 함께 증가, 함께 감소하는 경우를 파악하는데 유용
  - 켄달의 타우 - 두 변수의 순위의 일치 정도를 측정

상관관계와 인과관계

- 상관이 인과를 의미하지 않습니다.
- 인과는 상관을 포함할 수는 있습니다.
- 상관의 예 - 아이스크림 판매량과 익사율
- 인과의 예 - 흡연과 폐암 사이의 관계
- 다음 관계는 무엇일까요?
  - 특정 복권 판매점의 복권 판매 매출액과 고액권 당첨 횟수