

생성형AI

Day 5

데이터 분석



목차

1. 데이터 분석 개요
2. 데이터의 종류와 속성
3. 데이터 탐색(EDA)
4. 기초통계
5. 상관관계와 인과관계
6. 가설검정과 A/B test
7. 실습과제



데이터 분석 개요

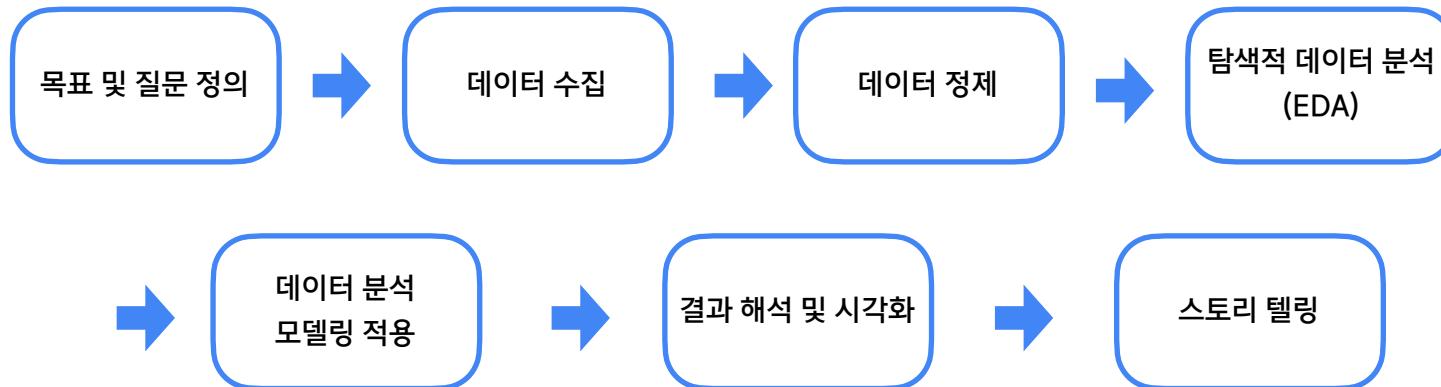
데이터 직군명	데이터 엔지니어	데이터 분석가	데이터 사이언티스트
역할	데이터 분석을 위한 환경구축	과거와 현재 데이터를 분석	과거와 현재 데이터를 기반으로 미래를 예측
주요업무	빅데이터 수집 및 관리 및 시스템 품질 관리 데이터 웨어하우스, 데이터 베이스 구축 및 관리 데이터 파이프라인 구축 SQL 튜닝, 성능 최적화 대용량, 실시간 서비스 개발	데이터를 분석 및 처리하고 비즈니스 인사이트 도출 분석 도구를 활용한 데이터 시각화 분석보고서 설계 및 작성	데이터를 분석 및 처리하고 비즈니스 인사이트 도출 과거 패턴으로부터 미래를 예측 비즈니스에 여러 알고리즘을 적용시켜 새로운 분석모델 및 머신러닝 모델 개발

데이터 분석 개요

데이터 분석 이란?

- 데이터로부터 유의미한 정보를 추출하고 결과를 분석하여 결정을 지원하는 과정
- 데이터를 통해 통찰력을 얻고 비즈니스 결정을 내리는 과학적 접근법

데이터 분석 과정



데이터 분석 개요

데이터 분석의 중요성

경쟁 우위 확보

- 데이터 주도 결정으로 시장 변화에 빠르게 대응하고 경쟁자보다 앞서 나갈 수 있습니다.

의사결정 개선

- 데이터 분석을 통한 정량적 접근은 주관적 판단의 오류를 줄이고 더 객관적인 결정을 내릴 수 있도록 돕습니다.

효율성 향상

- 과정의 비효율적인 부분을 식별하고 최적화하여 전체적인 비즈니스 효율성을 높일 수 있습니다.

고객 이해 및 서비스 개선

- 고객 데이터 분석을 통해 고객의 필요와 행동을 더 깊이 이해하고, 이를 기반으로 맞춤형 서비스를 제공할 수 있습니다.

데이터 분석 개요

데이터 분석과 실무

마케팅 최적화

- 고객 세분화, 타겟 마케팅, 캠페인 효과 분석 등 데이터를 기반으로 한 정밀 마케팅 실행

재무 관리

- 신용 평가, 리스크 관리, 사기 감지 등 재무 관련 의사결정에서 데이터 분석 활용

운영 최적화

- 공급망 관리, 재고 최적화, 제조 공정 개선 등 운영 효율성을 높이는 데이터 분석 적용

실무 사례

넷플릭스

- 사용자 기반을 AI 분석을 사용하여 다양한 그룹으로 나누어 시청 기록 및 선호도를 포함한 사용자 행동을 조사
- 이를 통해 영화 및 웹 시리즈에 대한 맞춤형 제안을 제공

paypal

- 사기 활동의 패턴을 식별

월마트

- 재고 수준, 제품 수요 및 운송 물류에 대한 세심한 분석을 통해 월마트는 운영 효율성을 높이고 비용을 절감

데이터 분석 개요

The Types of Data Analysis



<https://www.datacamp.com/blog/what-is-data-analysis-expert-guide#rdl>

데이터의 종류와 속성

정량적 데이터 (Quantitative Data)

- 수치로 표현되는 데이터
- 나이, 소득, 판매량 등의 데이터
- 통계 분석, 예측 모델링 등에 사용

정성적 데이터 (Qualitative Data)

- 문자, 동영상 등 의미와 특성으로 분류되는 데이터
- 댓글, 동양상, 음성 등의 데이터
- 통계 분석 어려움



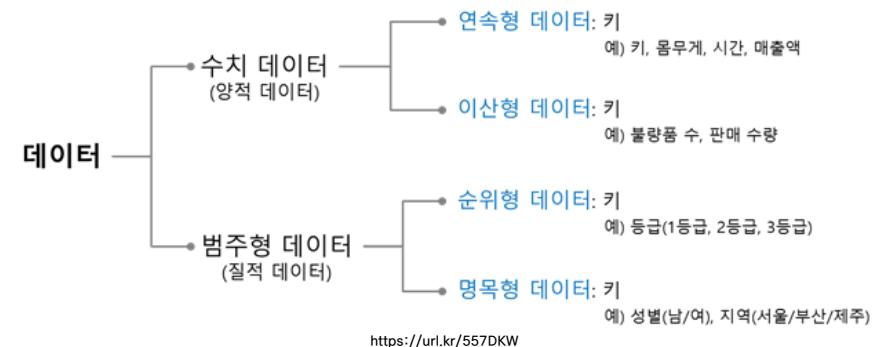
데이터의 종류와 속성

수치형 데이터 (Numerical Data)

- 값의 범위가 무한하고, 측정 가능
- 온도, 무게, 거리 등의 데이터
- 연속 변수에서의 경향성 분석, 상관관계 파악 등에 사용
- 연속형, 이산형 데이터

범주형 데이터 (Categorical Data)

- 제한된 범위의 값을 가지며, 일반적으로 레이블 형태
- 결혼 상태, 학력, 직업 유형 등의 데이터
- 인구 통계학적 분석, 고객 세분화 등에 사용
- 순위형, 명목형 데이터



데이터의 종류와 속성

데이터 품질의 요소

- 정확성, 완전성, 일관성, 타당성

데이터 전처리의 중요성

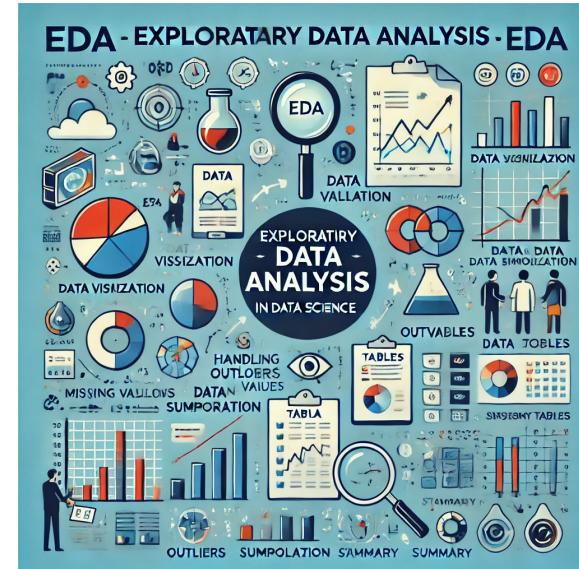
- 높은 품질의 데이터는 분석 결과의 신뢰성을 보장
- 전처리 과정: 결측치 처리, 이상치 탐지 및 제거, 데이터 정규화
- 또한 정확한 데이터 전처리는 비즈니스 인사이트의 정확성과 직결



데이터 탐색(EDA)

EDA란?

- 데이터 분석 초기 단계에서 수행되는 데이터의 시각화와 요약을 통해 중요한 특성과 패턴을 발견하는 과정
- 데이터의 구조, 예외, 패턴, 및 기초 통계적 요약 제공
- 데이터에 대한 이해도를 향상
- 모델링 방향을 설정하는 데 기여



데이터 탐색(EDA)

EDA 중요성

- 데이터 분석의 방향과 품질을 결정지을 수 있는 핵심 단계로
- 잠재적 문제를 사전에 파악하고 수정할 기회를 제공
- 데이터의 질과 구조를 이해하여 데이터 분석 결과의 정확성 향상

데이터 분석에서 EDA의 역할

- 데이터의 결측치, 이상치 및 분포를 파악하여 데이터 정제 및 전처리 계획 수립
- 관련성이 높은 변수를 식별하여 더 효과적인 모델을 구축할 수 있도록 지원

데이터 탐색(EDA)

데이터 시각화

- 데이터의 분포와 관계를 시각적으로 표현
- 히스토그램, 박스 플롯, 산점도, 히트맵, 파이 차트

기술 통계

- 통계적 수치를 계산하여 데이터의 경향을 요약
- 중심 경향 측정: 평균, 중앙값, 최빈값 등
- 분산도 측정: 표준편차, 분산, 범위, 사분위수 범위 (Q1, Q3) 등
- 요약 통계: 여러 기술 통계를 요약하여 제공

다변량 분석

- 여러 변수 간의 관계를 분석하여 인사이트를 도출
- 상관 분석: 피어슨, 스피어만, 켄달 등
- 주성분 분석 (PCA): 다차원 데이터의 차원을 축소하여 중요한 변수를 추출
- 요인 분석: 변수들의 상호 관련성을 소수의 요인(factor)으로 추출

기초통계

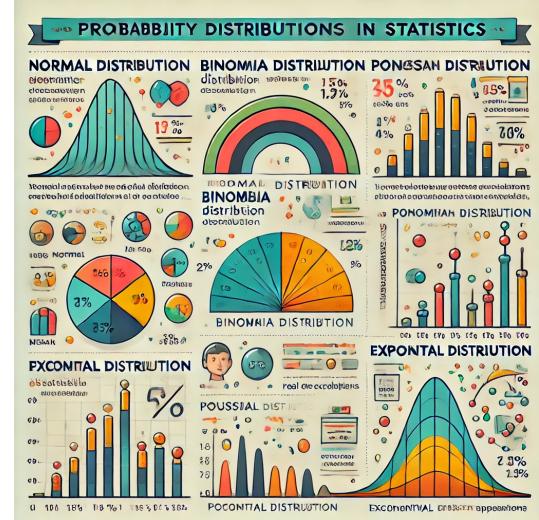
기술통계

- 데이터 집합의 중심 경향, 분산도 및 전반적인 분포를 요약하는 통계적 수치
- 중심 경향성: 평균, 중앙값, 모드
- 분산성: 범위, 분산, 표준편차, 사분위수
- 형태: 왜도(비대칭도)와 첨도(봉우리의 높이)
- 데이터의 일반적인 형태와 특성을 빠르게 파악하는 데 사용
- EX) 데이터의 ‘정상 범위’를 설정하기 위해 데이터의 평균과 표준편차를 사용

기초통계

확률 분포

- 데이터 포인트가 발생할 확률을 설명하는 수학적 모델
- 이산 확률 분포: 이항 분포, 포아송 분포 등
- 연속 확률 분포: 정규 분포, 지수 분포, t-분포, F-분포 등
- 모델링에서 데이터를 어떻게 가정하고 처리할지 결정하는 데 중요한 기준을 제공
- 많은 통계적 검정과 머신러닝 알고리즘은 데이터가 ‘정규 분포’를 따른다고 가정



상관관계 & 인과관계

상관관계 (Correlation)

- 정의: 두 변수 간의 관계에서 한 변수의 변화가 다른 변수의 변화와 어떻게 연관되어 있는지 나타내는 지표
- 피어슨 상관 계수, 스피어만 순위 상관 계수 등을 사용해서 측정

인과관계 (Causality)

- 정의: 한 변수(원인)의 변화가 다른 변수(결과)의 변화를 유발하는 관계
- 실험 설계, 회귀 분석, 경로 분석 등을 통해 추론

상관관계 & 인과관계

상관관계와 인과관계는 뭐가 다른가?

- 상관관계는 두 변수 간의 관계를 수치적으로 설명하지만, 인과관계를 설명하지는 않음
- 인과관계는 변수 A의 변화가 변수 B의 변화를 직접적으로 유발한다는 것을 의미

실제 예시

- 상관관계: 아이스크림 판매량과 수영장 사고 건수 간의 상관관계
- 인과관계: 흡연과 폐암 발생률 간의 인과관계

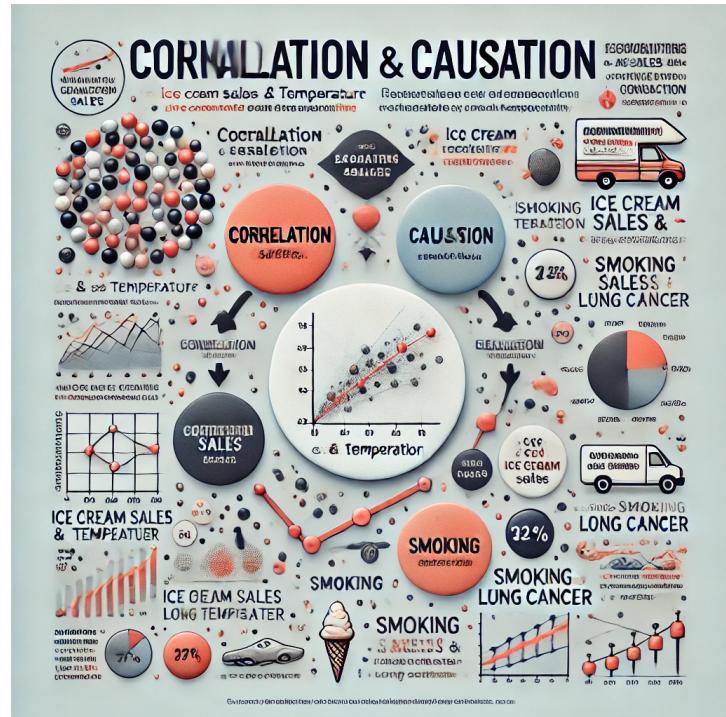
상관관계 & 인과관계

비즈니스 의사결정에서의 응용

- 상관관계를 통해 비즈니스 데이터에서 패턴과 트렌드 식별
- 인과관계를 통해 마케팅 캠페인, 정책 변경 등의 효과 분석

실무 예시

- 소비자 데이터 분석을 통한 제품 개발과 마케팅 전략 수립
- 고객 만족도 조사를 통한 서비스 개선 및 매출 증대



가설 검정과 A/B 테스트

가설 검정의 기초

- 가설 검정은 표본 데이터를 사용하여, 모집단에 대한 통계적 가설이 타당한지를 판단하는 과정
- 연구 가설의 지지 여부를 결정하기 위해 통계적 증거를 평가

프로세스:

- 귀무 가설(H_0): 기본 가설, 변화가 없음을 주장
- 대립 가설(H_1): 연구 가설, 변화를 주장
- 결정 규칙: 통계적 유의성을 평가하여 가설을 기각하거나 채택(귀무 가설의 기각 여부 결정)

가설 검정과 A/B 테스트

T검정

- 목적: 두 그룹 간의 평균 차이가 통계적으로 유의미한지 검정
- 데이터가 정규 분포를 따르고 두 집단의 샘플 크기가 비교적 작을 때 사용
- 유형: 독립 표본 T검정, 대응 표본 T검정, 단일 표본 T검정

ANOVA (분산 분석)

- 목적: 세 개 이상의 그룹 간 평균의 차이가 통계적으로 유의미한지 검정
- 응용: 여러 집단의 데이터를 비교할 때 일반적으로 사용, 예를 들어, 다양한 마케팅 전략의 효과 비교

가설 검정과 A/B 테스트

비모수적 방법

- 데이터가 정규 분포를 따르지 않을 때 사용하는 통계적 검정
- 많은 데이터가 정규 분포의 가정을 만족하지 않기 때문에 필수적인 방법론
- 크루스칼-왈리스 검정: ANOVA의 비모수적 대안, 여러 독립된 표본 그룹을 비교
- 맨-휘트니 U 검정: 두 독립 표본 간의 차이 검정
- 월콕슨 부호 순위 검정: 두 관련 표본 간의 차이 검정

가설 검정과 A/B 테스트

A/B 테스트

- 두 가지 이상의 버전(예: 웹 페이지, 제품, 서비스)을 대상으로 동시에 실행하여 어느 것이 더 효과적인지를 결정하는 실험적 접근 방식
- 사용자 경험, 제품 성능, 마케팅 전략 등을 최적화하여 최상의 결과를 도출하기 위함

A/B 테스트의 중요성

- 비즈니스 의사결정 개선: 실제 데이터에 기반해서 의사결정을 내리므로, 가정이나 추측보다 신뢰 가능
- 혁신적인 아이디어 검증: 새로운 아이디어나 기능을 안전하게 테스트하고, 실제 효과를 평가 가능
- 사용자 만족도 향상: 사용자 경험을 개선하고, 사용자의 요구에 더 잘 맞는 서비스를 제공 가능

시나리오

여러분은 온라인 쇼핑몰을 운영하는 'A'의 데이터 분석가입니다.

A는 최근 웹사이트의 사용자 인터페이스(UI)를 새롭게 변경했습니다.

이 변경이 실제로 사용자들의 구매 전환율에 어떤 영향을 미쳤는지 분석하고자 합니다.

수집한 데이터: UI 변경 전과 변경 후의 사용자 상호작용 로그, 구매 이력, 그리고 사용자 피드백 데이터

시나리오

데이터 종류와 속성 탐색

- 수집된 데이터의 종류와 속성을 파악
- 사용자 피드백에서 얻은 텍스트 데이터를 분석하여 정성적인 피드백과 정량적인 데이터를 구분
- 연속형 데이터(체류 시간, 구매 금액)와 범주형 데이터(사용자 성별, 사용 UI 타입)를 식별

기초 통계 및 EDA (Exploratory Data Analysis):

- 데이터의 분포, 중심 경향, 분산 등을 파악하고, 초기 인사이트 획득
- 체류 시간과 페이지 뷰의 히스토그램을 그려 데이터의 분포를 파악
- 구매 전환율과 체류 시간의 관계를 산점도로 나타내 상관관계를 탐색
- 박스 플롯을 사용하여 체류 시간의 이상치를 식별하고 처리 방안을 고려

시나리오

상관관계 분석:

- 각 변수들이 구매 전환율과 갖는 상관관계와 인과관계를 분석
- 피어슨 상관 계수를 계산하여 체류 시간과 구매 전환율 간의 선형 관계를 파악
- 인과관계를 유추하기 위해 추가적인 변수들(예: 페이지 뷰, 세션 중 발생한 이벤트 수)을 고려하여 다중 회귀분석을 수행

가설 검정 및 A/B 테스트:

- 변경된 UI가 구매 전환율에 긍정적인 영향을 미쳤는지를 통계적으로 검증
- 사용자를 두 그룹(기존 UI와 새 UI 사용자)으로 나누고, 이들 간의 구매 전환율을 비교
- 사용자를 무작위로 두 그룹(A: 기존 UI, B: 새 UI)으로 할당
- 각 그룹의 구매 전환율을 계산하고, 두 그룹 간의 차이를 평가하기 위해 독립 표본 T검정을 수행
- p-값과 유의 수준(0.05)을 비교하여 귀무 가설(두 UI 간에 차이가 없다)의 기각 여부를 결정
- 통계적 검정을 통해 얻은 결과를 바탕으로 UI 변경의 효과를 평가

이론 예제

- 데이터 분석 이론 예제 코드

<https://colab.research.google.com/drive/1nssJ5T89XjXC2JNISU4zSCWPn8gkLK3c?usp=sharing>

실습 과제

데이터 분석 실습

데이터는 Kaggle 등에서 본인이 원하는 데이터를 수집하여 사용합니다.

1. 데이터의 종류와 속성

- 주어진 데이터셋에서 범주형 및 연속형 데이터 열을 식별하고, 각 열의 기술 통계와 빈도수를 출력해보세요.

2. 데이터 탐색 (EDA)

- 제공된 데이터셋의 결측치와 중복을 파악하고, 적절한 처리 방안을 제안해보세요.

3. 기초통계

- 데이터셋의 왜도와 첨도를 계산하고, 그 의미에 대해 설명해보세요.

4. 상관관계와 인과관계

- 선택한 두 변수 간의 상관관계를 계산하고, 그 결과를 해석해보세요. 이를 바탕으로 가능한 인과관계를 논의해보세요.

5. 가설검정과 A/B 테스트

- 주어진 데이터를 사용하여 A/B 테스트를 설계하고, 가설을 설정한 후 통계적 검정을 수행하세요. 결과를 해석하고 결론을 도출해보세요.
- 본인의 수집한 데이터에서 가설검정, A/B 테스트를 수행해보기 어렵다면 어떤 가설을 설정하고 A/B 테스트를 어떻게 설계했을지 생각해보세요.

실습 진행