

생성형AI

Day 6

데이터 분석



목차

1. 시계열 데이터 분석
2. 다변량 분석
3. 실시간 데이터 분석
4. 데이터 기반 의사결정
5. 데이터 윤리 및 법적 고려사항
6. 우리는 어떤 방식으로 데이터를 분석해야 할까?
7. 실습과제



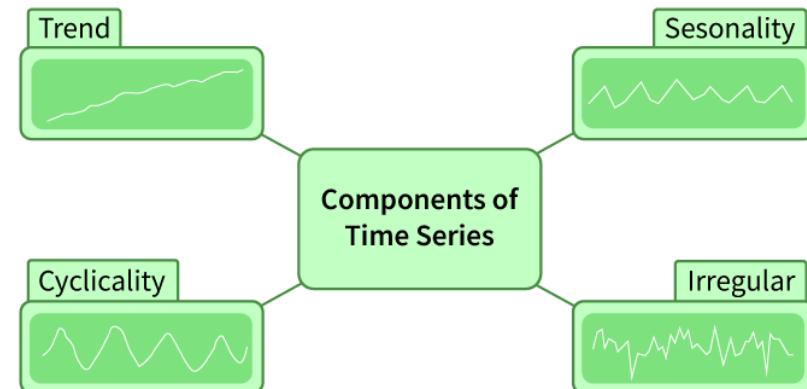
시계열 데이터 분석

시계열 데이터

- 시간 순서대로 정렬된 데이터 포인트의 연속

시계열 데이터 특성

- 추세(Trend): 장기적인 데이터 증가 또는 감소 경향
- 계절성(Seasonality): 특정 시간 패턴이 반복되는 현상 (예: 월별, 주별, 일별 패턴)
- 주기성(Cyclical): 불규칙적인 간격으로 반복되는 변동
- 잡음(Noise): 데이터에 포함된 불규칙한 변동



시계열 데이터 분석

시계열 분석 방법

시계열 분해

- 모델: 시계열 = 추세 + 계절성 + 순환성 + 잡음
- 가법 모형: 개별 요인의 효과를 구분하고 함께 더하여 데이터를 모형화하는 데이터 모형
- 승법 모형: 데이터가 증가하면 계절 패턴도 증가한다고 가정하는 모형

통계적 방법

- 이동 평균(Moving Average): 데이터의 단기 변동을 평활화
- 지수 평활(Exponential Smoothing): 최근 관측값에 더 큰 가중치를 두는 방법

시계열 예측

- ARIMA 모델: 자기회귀 통합 이동평균 모델로 시계열 데이터 예측



시계열 데이터 분석

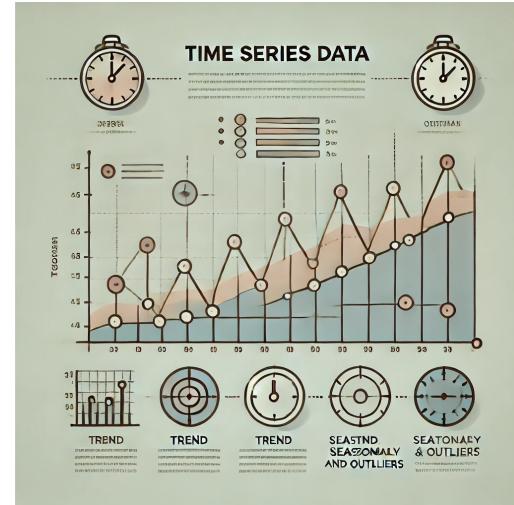
시계열 데이터를 다룰 때 발생하는 주요 문제와 해결책

문제

- 결측치(Missing Values): 데이터의 연속성이 끊기는 문제, 특정 시점의 데이터가 없음
- 이상치(Outliers): 예상 범위를 벗어나는 데이터 포인트, 일관성 없는 데이터 포인트

해결책

- 결측치는 보간법(interpolation), 평균으로 대체하여 처리
- 이상치 탐지 및 처리는 Z-점수, IQR(Interquartile Range) 방법 사용
- 차분(differencing)을 통한 추세 및 계절성 제거를 통한 안정화



다면량 분석

다면량 분석

- 여러 현상이나 사건에 대한 측정치를 개별적으로 분석하지 않고 동시에 한번에 분석하는 통계적 기법

다면량 분석의 이해

- 두 개 이상의 변수를 동시에 분석하는 기법으로, 변수들 간의 관계를 파악하고, 패턴을 예측
- 단변량 분석에서 간과할 수 있는 변수들 간의 상호작용과 복잡한 관계를 포착 가능
- Ex) 소비자 데이터에서 구매 패턴, 고객 세분화



다면량 분석

상관분석

피어슨 상관계수 (Pearson Correlation)

- 두 변수 간의 선형 관계의 강도와 방향을 측정하는 통계적 방법
- 연속적인 수치 데이터 간의 선형 관계를 측정
- 연속적이고 정규 분포를 따르는 수치 데이터에 적합

스피어만 상관계수 (Spearman Correlation)

- 두 변수의 순위에 기반하여 관계를 측정하는 비모수적 방법
- 순위 기반의 비선형 관계를 측정
- 데이터가 정규 분포를 따르지 않거나 순위형 데이터일 때 유용

	데이터 요구사항	이상치 민감도	적용범위
피어슨 상관계수	선형 관계와 정규 분포	이상치에 민감	선형적인 상관 관계 분석에 적합
스피어만 상관계수	순위 데이터나 비 선형 관계	이상치의 영향에 덜 민감	단조적인 관계(선형일 필요는 없음)를 측정하는 데 적합

다변량 분석

주성분 분석(PCA, Principal Component Analysis)

- 고차원 데이터의 차원을 축소하여 가장 중요한 특성을 추출
- 데이터의 분산이 최대가 되는 방향을 찾아, 중요 정보를 유지하면서 차원 축소

주성분 분석의 단계

- 데이터 표준화: 각 변수의 평균을 0, 표준편차를 1로 맞추어 모든 변수의 비중을 맞춤
- 공분산 행렬 계산: 표준화된 데이터의 변수들 간의 선형 관계를 나타내는 공분산 행렬을 계산
- 고유값 분해: 공분산 행렬의 고유값과 고유벡터를 계산(고유벡터: 데이터 분산이 최대인 방향, 고유값: 분산의 크기)
- 주성분 선택: 가장 큰 고유값에 해당하는 고유벡터부터 순서대로 주성분으로 선택
- 새로운 특성 공간으로 데이터 투영: 선택된 주성분에 원래 데이터를 투영하여 차원을 축소

주성분 분석의 응용

- 데이터 시각화: 고차원 데이터를 2D나 3D 공간으로 시각화하여 데이터의 패턴을 이해
- 노이즈 제거: 데이터의 주요 구조만을 포착하고 잡음을 제거
- 특성 추출과 데이터 압축: 대량의 변수를 가진 데이터셋에서 가장 정보가 풍부한 특성을 추출해 저장 공간을 절약

다면량 분석

요인 분석

- 변수들 사이의 관계를 분석하여 몇 가지 잠재적인 요인으로 요약하는 통계적 방법
- 관측된 변수들 뒤에 숨어 있는 잠재적 요인을 발견
- 변수들이 하나 이상의 비관측된 잠재 변수(요인)에 의해 영향을 받는다는 가정
- PCA와 유사하지만 요인 분석은 데이터 내 잠재적 구조를 모델링 하는데 초점
- 관측된 변수들의 변동성을 설명할 수 있는 공통 요인을 찾아냄으로써 변수의 수를 줄이고 데이터의 구조를 이해하는 데 도움
- 주로 심리학, 사회과학, 마케팅, 기타 연구 분야에서 설문지 데이터의 구조를 분석하는 데 사용

요인분석의 목적

- 변수 축소: 여러개의 변수들이 하나의 요인으로 묶임
- 불필요한 변수 제거: 요인에 포함되지 않거나 포함 되더라도 중요도가 낮은 변수 탐색 가능
- 변수 특성 파악: 관련된 변수들이 묶임으로서 요인들의 상호 독립적인 특성을 파악

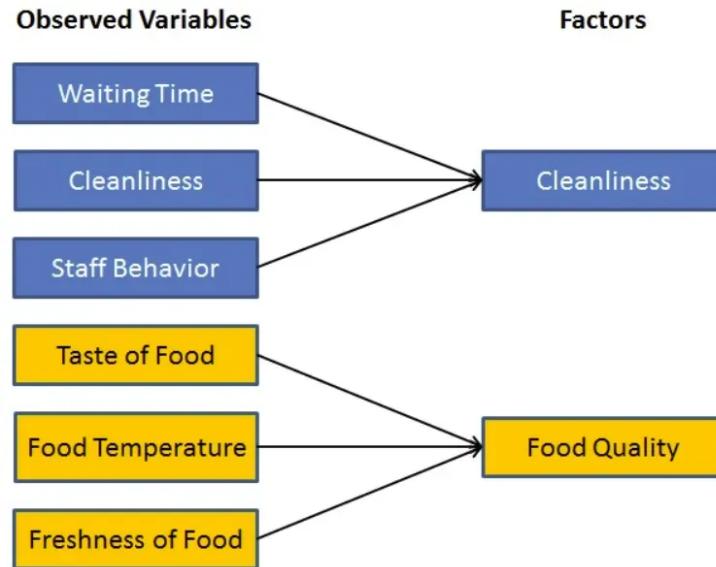
다면량 분석

요인 분석

- 그림에서와 같이 여러 변수들을 새로운 잠재변수로 묶는 분석

요인 분석 프로세스

- 적합성 검사:
 - 모든 변수가 요인분석에 적합한지를 평가(ex. KMO)
- 요인 추출
 - 초기 요인을 추출하기 위해 설명된 분산이 높은 주성분부터 순차적으로 요인을 선택
 - 요인 수 결정
- 요인 회전
 - 추출된 요인 간의 해석을 용이하게 하기 위해 Varimax 회전 등을 실시
 - 요인 적재를 최대화하여 각 요인에 대한 변수들의 기여도 분석
- 요인 해석:
 - 각 요인에 높게 적재된 변수들을 분석하여, 그 요인이 무엇을 의미하는지 해석



<https://domino.ai/data-science-dictionary/factor-analysis>

실시간 데이터 분석

실시간 데이터(스트리밍 데이터)

- 데이터가 생성됨과 동시에 지속적으로 분석되는 데이터 스트림
- 연속성: 데이터가 중단 없이 지속적으로 생성/수집
- 변동성: 데이터의 양과 속성이 시간에 따라 변동 가능
- 속도: 빠른 속도로 데이터가 수집 및 처리되어야 함
- Ex) 소셜 미디어 피드, 주식 시장 데이터, IoT(사물인터넷) 센서 데이터

실시간 분석

- 즉각적인 의사결정: 실시간 반응이 필요한 상황에서 중요
- 성능 모니터링: 시스템이나 프로세스의 실시간 모니터링을 통해 문제를 사전에 감지
- Apache Kafka, Apache Storm, Apache Flink, ...



실시간 데이터 분석 – 도전과제 및 사례

도전 과제

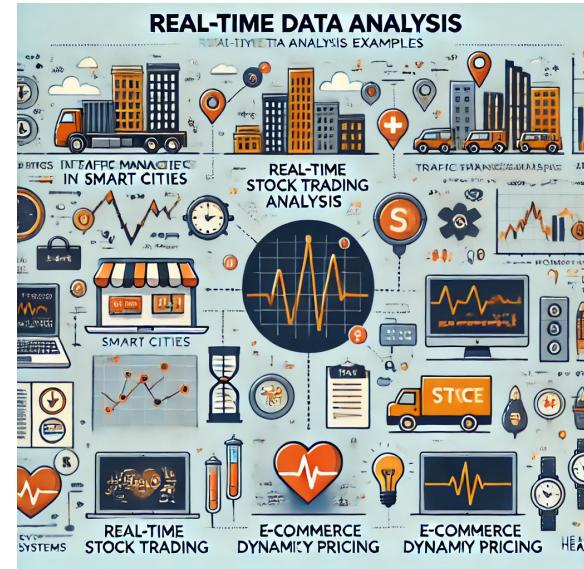
- 데이터 볼륨: 막대한 양의 데이터를 효율적으로 처리해야 함
- 데이터 품질: 불완전하거나 오류를 포함할 수 있는 데이터의 처리
- 보안 문제: 실시간 데이터 처리 시 보안과 프라이버시 유지

사례

- 금융 거래 감시: 실시간으로 거래를 모니터링하여 사기를 탐지
- 교통 관리 시스템: 실시간 교통 데이터를 분석하여 교통 흐름 최적화

실무에서의 실시간 데이터 분석 응용

- e-커머스: 고객의 온라인 행동을 실시간으로 분석하여 개인화된 광고 제공
- 헬스케어: 환자의 실시간 건강 데이터를 모니터링하여 즉각적인 의료 개입 가능
- 스마트 시티: 도시의 각종 센서로부터 실시간 데이터를 수집 및 분석하여 에너지 관리, 안전 강화



데이터 기반 의사결정

데이터 기반 의사결정

- 데이터 기반 의사결정은 객관적 데이터를 활용하여 의사결정 과정에서 주관성을 최소화하고, 최적의 결과를 도출하기 위한 접근 방식

핵심 요소

- 데이터 수집: 정확하고 신뢰할 수 있는 데이터의 수집, 통합된 데이터 관리, 데이터 품질 보장
- 데이터 분석: 수집된 데이터를 분석하여 인사이트 도출, 적절한 분석 도구 활용
- 결정 실행: 분석 결과를 바탕으로 전략 수립 및 실행, 데이터 주도 문화 조성, 피드백 루프 구축
- 효과: 오류 감소, 효율성 증가, 객관적 결정 가능

데이터 기반 의사결정

데이터 기반 의사결정의 실무 예시

- 예시 1: 재고 관리 – 소매업에서 판매 데이터와 고객 수요 예측을 분석하여 재고 수준을 최적화
- 예시 2: 고객 세분화 – 고객 데이터를 분석하여 다양한 고객 그룹을 식별하고 맞춤형 마케팅 전략을 개발
- 중요성: 신속한 의사결정이 요구되는 비즈니스 환경에서 데이터 기반 의사결정은 리스크를 줄이고 경쟁 우위를 확보하는 데 필수

데이터 기반 의사결정의 Challenge

- 데이터 품질 문제: 부정확하거나 불완전한 데이터
- 분석 오류: 잘못된 분석 방법이나 기술의 사용
- 윤리적 고려: 데이터의 사용이 개인의 프라이버시 침해와 같은 윤리적 문제 야기
- 과도한 의존: 데이터에 지나치게 의존하면 혁신을 저해하고 유연성이 감소

데이터 윤리 및 법적 고려사항

데이터 보안과 개인정보 보호

- 데이터 보안: 데이터의 무단 접근, 사용, 공개, 파괴 및 변경을 막는 모든 조치를 포함
- 개인정보 보호: 개인의 데이터를 적절하게 관리하고 보호
- 중요성: 사이버 공격과 데이터 유출이 증가함에 따라, 강력한 데이터 보안 및 개인정보 보호 정책이 필수적

데이터 보안의 전략

- 데이터 암호화
- 접근 제어
- 네트워크 보안
- 데이터 백업



데이터 윤리 및 법적 고려사항

데이터 윤리

- 데이터의 적절한 사용을 위한 원칙과 기준
- 개인정보 보호, 데이터 접근성, 데이터의 정확성 및 투명성

윤리적 데이터 사용의 중요성

- 신뢰성 증진: 윤리적인 데이터 관리는 고객 및 이해관계자의 신뢰를 쌓는 데 기여
- 리스크 관리: 데이터 유출이나 오용으로 인한 법적 및 명예적 리스크를 방지
- 규제 준수: 데이터 보호 규정(GDPR, HIPAA 등)을 준수하여 법적 제재 방지

데이터 분석 관련 법적 규제

- 법적 규제: 데이터 보호 및 개인정보 보호에 관한 법률, GDPR, HIPAA(보건보험 이동성 및 책임에 관한 법), CCPA(캘리포니아 소비자 프라이버시 법) 등
- 준수 필요성: 법률을 준수하지 않는 경우 높은 벌금, 기업의 명성 손상 등의 결과를 초래할 수 있음
- 예시: 헬스케어 분야에서 데이터 분석을 수행할 때 HIPAA 규정을 준수하여 환자의 건강 정보를 보호해야 함

우리는 어떤 방식으로 데이터를 분석해야 할까?

1. 질문하기
2. 작은 단위에서 큰 단위로
3. 여러가지 분석과 결과 분석하기
4. 현장을 이해하고 방향 만들기
5. 스토리 만들기
6. 피드백 준비하기



시나리오

여러분들은 집앞에 있는 작은 카페의 아르바이트생입니다.

그런데 최근 카페의 매출이 떨어져 여러분들의 아르바이트 자리가 위협받고 있습니다.

최근 전세계적으로 어려운 경제 때문일까요? 그런데 왜 우리 학교앞 카페는 항상 사람이 붐비는 것 같습니다.

아르바이트 자리가 위험한 여러분은 **카페의 매출이 떨어지는 문제를 해결하기 위해 카페 데이터를 분석하고자 합니다.**



시나리오

여러분들은 집앞에 있는 작은 카페의 아르바이트생입니다.

그런데 최근 카페의 매출이 떨어져 여러분들의 아르바이트 자리가 위협받고 있습니다.

최근 전세계적으로 어려운 경제 때문일까요? 그런데 왜 우리 학교앞 카페는 항상 사람이 붐비는 것 같습니다.

아르바이트 자리가 위험한 여러분은 **카페의 매출이 떨어지는 문제를 해결하기 위해 카페 데이터를 분석하고자 합니다.**

1. 질문하기

- 데이터를 향해 질문해보기
- 원하는 바를 향해 질문해보기
- 지식, 생각을 채워나갈 수 있는 질문 해보기



시나리오

여러분들은 집앞에 있는 작은 카페의 아르바이트생입니다.

그런데 최근 카페의 매출이 떨어져 여러분들의 아르바이트 자리가 위협받고 있습니다.

최근 전세계적으로 어려운 경제 때문일까요? 그런데 왜 우리 학교앞 카페는 항상 사람이 붐비는 것 같습니다.

아르바이트 자리가 위험한 여러분은 **카페의 매출이 떨어지는 문제를 해결하기 위해 카페 데이터를 분석하고자 합니다.**

2. 작은 단위에서 큰 단위로

- 질문을 작은 단위로 쪼개서 데이터로 바꾸기
- 작은 단위의 분석부터 시작하기



시나리오

여러분들은 집앞에 있는 작은 카페의 아르바이트생입니다.

그런데 최근 카페의 매출이 떨어져 여러분들의 아르바이트 자리가 위협받고 있습니다.

최근 전세계적으로 어려운 경제 때문일까요? 그런데 왜 우리 학교앞 카페는 항상 사람이 붐비는 것 같습니다.

아르바이트 자리가 위험한 여러분은 **카페의 매출이 떨어지는 문제를 해결하기 위해 카페 데이터를 분석하고자 합니다.**

3. 여러가지 분석과 결과 분석하기

- 여러가지 데이터로 적절한 분석 실행
- 분석결과를 종합해서 다시 분석하기



시나리오

여러분들은 집앞에 있는 작은 카페의 아르바이트생입니다.

그런데 최근 카페의 매출이 떨어져 여러분들의 아르바이트 자리가 위협받고 있습니다.

최근 전세계적으로 어려운 경제 때문일까요? 그런데 왜 우리 학교앞 카페는 항상 사람이 붐비는 것 같습니다.

아르바이트 자리가 위험한 여러분은 **카페의 매출이 떨어지는 문제를 해결하기 위해 카페 데이터를 분석하고자 합니다.**

4. 현장을 이해하고 방향 만들기

- 분석 결과를 통해 방향, 전략 수정하기
- 수정한 방향과 전략이 현장에 적합한지 파악하기



시나리오

여러분들은 집앞에 있는 작은 카페의 아르바이트생입니다.

그런데 최근 카페의 매출이 떨어져 여러분들의 아르바이트 자리가 위협받고 있습니다.

최근 전세계적으로 어려운 경제 때문일까요? 그런데 왜 우리 학교앞 카페는 항상 사람이 붐비는 것 같습니다.

아르바이트 자리가 위험한 여러분은 **카페의 매출이 떨어지는 문제를 해결하기 위해 카페 데이터를 분석하고자 합니다.**

5. 스토리 만들기

- 현장에 대한 이해를 바탕으로 데이터 분석에 스토리 입히기
- 이해관계자에 따라 스토리 구성을 다르게 하기



시나리오

여러분들은 집앞에 있는 작은 카페의 아르바이트생입니다.

그런데 최근 카페의 매출이 떨어져 여러분들의 아르바이트 자리가 위협받고 있습니다.

최근 전세계적으로 어려운 경제 때문일까요? 그런데 왜 우리 학교앞 카페는 항상 사람이 붐비는 것 같습니다.

아르바이트 자리가 위험한 여러분은 **카페의 매출이 떨어지는 문제를 해결하기 위해 카페 데이터를 분석하고자 합니다.**

6. 피드백 준비하기

- 변화한 전략을 평가/수정하기 위해 지속적으로 데이터 수집 전략 세우기
- 피드백 전략 수립 후 분석하기



실습 과제

1. 데이터 분석하기 (<https://www.kaggle.com/datasets/nikhil7280/weather-type-classification>)

- Kaggle의 날씨 데이터
- 데이터 분석
- 어떤 인사이트를 얻을 수 있까? 어떤 서비스를 개발할 수 있을까?

2. 일상생활에서 겪을 만한 사건을 시나리오로 설정하고 데이터 분석

1. 질문하기
2. 작은 단위에서 큰 단위로
3. 여러가지 분석과 결과 분석하기
4. 현장을 이해하고 방향 만들기
5. 스토리 만들기
6. 피드백 준비하기

실습 진행