데이터 분석 개요

R이나 Python을 활용하여 데이터 분석을 진행하기 위해서는 데이터 분석에 필요한 기본 개념들에 대한 숙지가 필요합니다. 이 포스팅에서는 데이터 분석 전에 숙지해야 할 기본 용어들과 개념을 소개합니다.

#01. 데이터 수집(측정)

1) 측정이란?

- 표본조사나 실험을 하는 과정에서 추출된 원소나 **관측자료**를 얻는 것

> 관측자료 = 데이터

2) 변수(variable) = Data

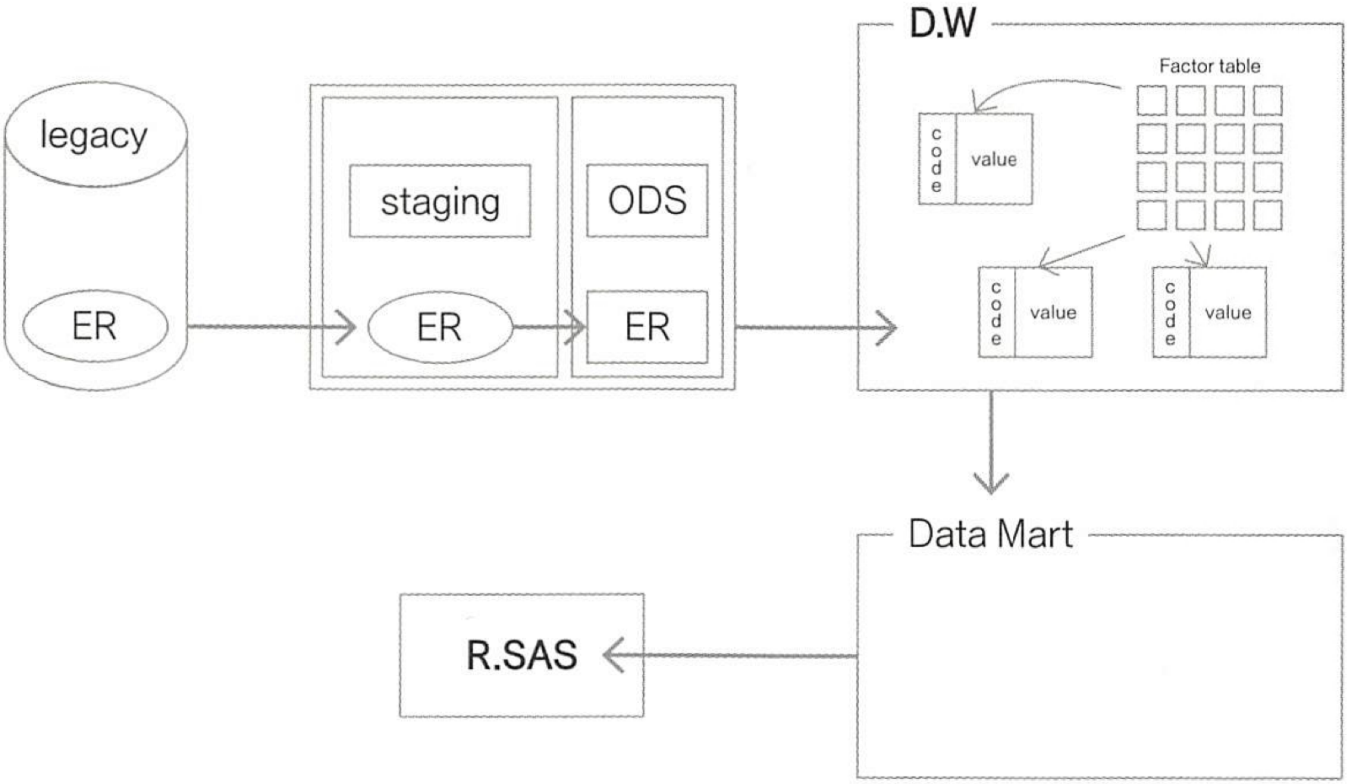
각 단위에 대해 관측되는 특성

#02. 데이터 전처리

1) 개요

분석에 적합한 형태로 데이터를 가공하는 것

용어	설명
legacy	기업 업무에 필요한 기존 운영 시스템 예) 마케팅, 세일즈, ERP, CMS
staging	데이터를 가공 없이 로딩 하는 영역 예) 비정형 데이터
ODS	로딩된 데이터를 DW에 저장하기 전에 임시로 저장하는 영역 예) 비정형 -> 정형 데이터로 처리
데이터웨어하우스(DW)	DW란 운영계 데이터를 사용자 관점에서 주제별로 통합하여, 별도의 장소에 저장해 놓은 통합 데이터베이스.
데이터마트(DM)	분석의 편의성을 높이고자, DW의 데이터를 주제별, 업무별로 요약하여 구성한 데이터 저장소. 일반적으로 각 부서별로 다양한 예측과 분석을 목표로 만들어진다. 예) DataFrame



2) 활용

- 데이터웨어하우스(DW), 데이터마트(DM)를 통해 분석 데이터를 가져와 사용
- Legacy System이나 Staging Area에서 데이터를 가져와서 DW에서 가져온 내용과 결합하여 사용 가능하지만 보안에 취약한 방법이기 때문에 거의 이루어지지 않음
- 클린징 영역인 ODS(Operation Data Store)에서 데이터의 전처리를 해서 DW나 DM과 결합하여 사용하는 것이 이상적

3) 최종 데이터 구조로 가공

분류값과 입력 변수들을 연관시켜 인구통계, 요약변수, 파생변수 등을 산출

구분	설명
비정형 데이터	DBMS에 저장됐다가 텍스트 마이닝을 거쳐 데이터 마트와 통합
관계형 데이터	DBMS에 저장되어 사회 신경망분석을 거쳐 분석 결과 통계값이 마트와 통합되어 활용

#03. 데이터 분석(=통계)

특정한 집단이나 불확실한 현상을 대상으로 자료를 수집해 대상 집단에 대한 정보를 구하고, 적절한 통계분석 방법을 이용해 의사결정을 하는 과정

- 특정집단을 대상으로 수행한 조사나 실험을 통해 나온 결과에 대한 요약된 형태의 표현
- 일기예보, 물가/ 실업률, 정당 지지도 , 의식조사와 사회조사 분석 통계 , 임상실험 등의 실험 결과 분석 통계

1) 데이터 분석 구분

	탐색적 데이터 분석(EDA)	확증적 데이터 분석(CDA)
활동	**기술통계**	**추론통계**
정의	데이터를 요약해 설명하는 기법	단순히 숫자를 요약하는 것을 넘어 어떤 값이 발생할 확률을 계산하는 통계 기법. 모집단에서 샘플링한 표본을 가지고 모집단의 특성을 추론하고 그 결과가 신뢰성이 있는지 검정하는 것이다.
예시	사람들이 받는 월급을 집계해 전체 월급 평균을 구한다.	수집된 데이터에서 성별에 따라 월급에 차이가 있는 것으로 나타났을 때 이런 차이가 우연히 발생할 확률을 계산한다.
과정	데이터 수집 > 시각화 탐색 > 패턴 도출 > 인사이트 발견	가설 설정 > 데이터 수집 > 탐색적 데이터 분석 > 추론통계 > 가설 검증

2) 탐색적 데이터 분석 (EDA)

개요

- 다양한 차원과 값을 조합
- 특이한 점이나 의미있는 사실을 도출
- 분석의 최종 목적을 달성해가는 과정
- 데이터의 특징과 내재하는 구조적 관계를 알아내기 위한 기법들의 통칭
- 프린스톤 대학의 튜키교수가 1977년 발표한 저서에서 소개

기술통계

모집단으로부터 표본을 추출하고 표본이 가지고 있는 정보를 쉽게 파악할 수 있도록 데이터를 정리하거나 요약하기 위해 하나의 숫자 또는 그래프의 형태로 표현하는 절차

- 주어진 자료로부터 어떠한 판단이나 예측과 같은 주관이 섞일 수 있는 과정을 배제하여 통계집단 들의 여러 특성을 수량화하여 객관적인 데이터로 나타내는 통계분석 방법론
- Sample에 대한 특성인 평균 , 표준편차 , 중위수 , 최빈값 , 그래프 , 왜도 , 첨도 등을 구하는 것을 의미

4대 주제

- 저항성의 강조
- 잔차 계산
- 자료변수의 재표현
- 그래프를 통한 시각화

저항성

자료의 일부가 파손되었을 때 영향을 적게 받는 성질

자료의 파손은 자료 일부가 뜬금없는 값으로 대체되는 경우를 의미함.

저항성이 있다면 이러한 자료의 변동에 민감하지 않다.

> ex) 평균은 저항성에 민감, 중앙값은 저항성에 민감하지 않음.

잔차

실제 관측된 데이터와 예측된 값 사이의 차이를 나타내는 오차.

즉, 주어진 데이터를 이용하여 회귀모델을 통해 예측한 값과 실제 값 간의 차이

자료 변수의 재표현

원래의 변수를 적당한 척도로 바꾸는 것.

보통 로그 변환이나 제곱근 변환을 통해 수행함

이를 통해 분포의 대칭성, 선형성, 분산 안정성등 데이터의 구조를 파악하는데 도움을 얻을 수 있음.

탐색적 데이터 분석의 과정

단계	주요 활동
데이터이해 단계	변수의 분포와 특성 파악
변수생성 단계	분석 목적에 맞도록 데이터 요약 및 파생변수 생성
변수선택 단계	목적변수에 의미있는 후보 변수 선택
데이터 시각화	데이터의 상태를 한눈에 확인할 수 있도록 요약하여 그래픽으로 표현

데이터 시각화

- 가장 낮은 수준의 분석이지만 때로는 복잡한 분석보다도 더 효율적
- 빅데이터에서 시각화는 필수
- 특히, 탐색적 분석의 결과를 요약하는데 효과적

3) 확증적 데이터 분석 (CDA)

추론통계

모집단으로부터 추출된 표본의 표본통계량으로 부터 모집단의 특성인 모수에 관해 통계적으로 추론하는 절차

> 실질적인 데이터 분석(통계)를 의미하는 과정

- 자료의 정보를 이용해 집단에 관한 추측, 결론을 이끌어내는 과정
- 수집된 자료를 이용해 대상 집단(모집단)에 대한 의사결정을 하는 것으로 Sample을 통해 모집 단을 추정하는 것을 의미
- 제한된 표본을 바탕으로 모집단에 대한 일반적인 결론을 유도하려는 시도이므로 본질적으로 불확실성을 수반함.

구분	설명
모수추정	- 표본집단으로부터 모집단의 특성인 모수 (평균 , 분산 등)를 분석하여 모집단을 추론 - 전수조사가 불가능하면 모집단에서 표본을 추출하고 표본을 근거로 확률론을 활용하여 모집단의 모수들에 대해 추론
점추정	표본의 정보로부터 모집단의 모수를 하나의 값으로 추정하는 것

구분	설명
구간추정	일정한 크기의 신뢰구간으로 모수가 특정한 구간에 있을 것이라고 선언하는 것. (= 신뢰구간)
가설검정	대상집단에 대해 특정한 가설을 설정한 후에 그 가설이 옳은지 그른지에 대한 채택여부를 결정하는 방법론
예측	미래의 불확실성을 해결해 효율적인 의사결정을 하기 위해 활용한다 . (예 :회귀분석, 시계열분석)
모집단의 추론	- 모집단의 변동성 또는 퍼짐의 정도에 관심이 있을 경우, 모분산이 추론의 대상이 된다. - 모집단이 정규분포를 따르지 않더라도 중심극한정리를 통해 정규모집단으로부터의 모분산에 대한 검정을 유사하게 시행할 수 있다. - 이 표본에 의한 분산비 검정은 두 표본의 분산이 동일한지를 비교하는 검정으로 검정통계량은 F분포를 따른다.

활용분야

분야	예시
정부의 경제 정책 수립과 근거자료	실업률, 고용률, 물가지수
농업	가뭄,수해,병충해 등에 강한 품종 개발/개량
의학	임상실험 결과 분석
경영	제품 개발, 품질관리, 시장조사, 영업관리 등
스포츠	선수들의 체질향상 및 개선, 경기 분석, 전략 분석, 선수 평가/기용 등

#04. 공간분석(GIS)

공간적 차원과 관련된 속성들을 시각화 하는 분석으로 지도 위에 관련 속성들을 크기, 모양, 선 굵기, 색상 등으로 표시한다.

> 도시공학 분야에서 활발히 사용한다.

#05. 데이터마이닝(머신러닝)

1) 개요

- 대표적인 고급 데이터 분석법
- 대용량의 자료로부터 정보를 요약
- 미래에 대한 예측
- 관계, 패턴, 규칙 등을 탐색
- 모형화
- 유용한 지식을 추출

2) 방법론

데이터베이스에서의 지식탐색

- 데이터웨어하우스에서 데이터마트를 생성
- 각 데이터들의 속성을 사전분석을 통해 지식을 얻음

기계학습

- 컴퓨터가 학습할 수 있도록 알고리즘과 기술을 개발하는 분야
- 인공신경망, 의사결정나무, 클러스터링, 베이지안분류, SVM 등

패턴인식

- 원자료를 이용해서 사전지식과 패턴에서 추출된 통계 정보를 기반으로 자료 또는 패턴을 분류
- 장바구니 분석, 연관규칙 등

3) 활용분야

분야	**예시**
데이터베이스 마케팅	방대한 고객의 행동정보를 활용 예) 목표 마케팅, 고객세분화, 장바구니 분석, 추천 시스템 등
신용평가 및 조기경보 시스템	신용카드 발급, 보험, 대출 업무 등
생물정보학	유전자 분석, 질병 진단, 치료법/신약 개발
텍스트마이닝	전자우편, SNS 등 디지털 텍스트 정보를 통한 고객 성향 분석, 감성 분석, 사회관계망 분석 등