[3주차]EDA와 활용사례

데이터 분석의 단계적 절차

문제파악 → 자료수집 → 자료요약 → 자료분석

1) 문제파악
→ 문제제기에 합당한 데이터를 얻기 위한 디자인 단계
→ 어떤 자료를 수집할지, 대상은 어떻게 정할지, 대상이 문제제기에 부합된 목표 모집단인지를 명확히 구분해야 함

2) 자료수집
→ 자료수집 단계는 다음의 3가지 점에 주의
1. 문제제기에 의한 결과를 해석하고 적용하는 대상의 전체인 목표모집단 설정⇒ **“모집단” 정의와 범위의 중요성**

2. 모집단을 잘 대표하는 표본을 데이터로 구성하며, 직접 수집 또는 측정 가능하도록 구체적인 데이터 추출 설계 (예, 표본추출방법, 실험디자인방법, 데이터마이닝 기법 등) ⇒ **“랜덤성”의 중요성**

3. 기획, 전략, 정책 방향의 근거(evidence-based)가 될 수 있도록 문제제기에 부합된 데이터의 정제(cleaning) 단계 및 데이터의 타당성 검증
⇒ **“데이터마이닝”의 중요성**

→ 자료요약은 데이터 탐색 단계이며, 첫 문제제기의 방향 및 문제제기의 방향을 수정 보완 가능한 단계 ⇒ **데이터 “패턴탐색”의 중요성**

3) 자료요약 (패턴의 시각화 및 기초통계화)
1. 데이터 탐색 단계는 숫자(표)를 통한, 그림을 통한 방법이 있으며, 데이터의 특성 및 성질에 따라 적합한 방법으로 요약 및 탐색 가능

2. 탐색 단계에서 두 가지 요인이나 변수들의 상관성을 살펴보고자 할 경우에는, 데이터의 특성이 관찰연구에 의해 수집된 디자인인지, 실험에 의한 환경 통제가 가능하여 제 3의 요인의 영향을 배제된 디자인에서 추출된 데이터인지를 명확히 구분하여, 두 변수의 상관성을 이해하여야 함.

3. 상관성이 없다고 하여, 두 변수간의 유의미한 연관성이 없는 것이 아니며, 상관성이 높다고 하여, 두 변수간의인과성이 있는 것은 아님.

4. 상관성을 살펴보는 것은 단지, 두 변수의 관계가 선형적 관계(linear relationship)를 갖는지 여부이기 때문이므로 해석에 조심하여야 함

3-1) 데이터를 대표하는 숫자 제공하기
목적 : 문제제기된 연구대상(모집단, 또는 데이터에서의 정보를 설명하는 대상의 범위)을 대표하여 설명하는 기술통계(descriptive statistics)값 구하기
기술통계량 :
데이터의 중심을 표현하는 대푯값 : 평균(mean), 중앙값(median), 최빈수(mode)

데이터의 중심으로부터 산포된 정도를 설명하는 대푯값 : 분산(variance), 표준편차(standard deviation)
데이터의 상대적 위치를 표현하는 대푯값 : 범위 (range), 백분위수 (percentile score), 사분위수 (quartile score)

4) 자료분석
1. 두 집단의 평균값 차이에 대한 분석 : two-sample z-test, t-test
2. 세 집단의 평균값 차이에 대한 분석 : 분산분석(ANOVA-test)
3. 빈도에 차이에 대한 분석 : 카이제곱분석, 범주형자료분석, trend for p-value
4. 두 변수간의 상관성 분석 : 상관계수, 상관분석
5. 설명변수에 의한 종속변수간의 관계 분석 : 일반화 선형모형, 비선형 모형 단순회귀분석, 다중회귀분석, 로짓분석 등
6. 모집단의 분포의 모수가 결정되지 않을 경우 : 비모수적 분석 (예, 부호검정 등)
7. 빅데이터의 패턴 및 관계 분석 : 데이터 마이닝 기법 (연관성 규칙, 판별분석 및 군집화, 의사결정나무모형, 선형회귀모형, 신경망분석, 기계학습 및 딥러닝 등

데이터의 구성 요소

단위(unit)/케이스(case) : 데이터를 구성하는 가장 기본이 되는 개체
- 단위 : 관심의 대상인 모집단을 구성하는 개별 조사대상으로 조사나 관찰의 대상 (예, 관심의 대상이 사람인 경우에는 단위는 바로 '개인')

변수(variable) : 각 단위에 대해 측정되는 특성 또는 속성을 말함
- 소비자의 구매 선호도 조사 : 소비자를 대상으로 '성별', '구매성향', '나이', '직업', '교육수준' 등이 변수

- 각 단위에서 정의된 변수는 오직 하나의 값만을 갖게 되어 변수 값이 '남자'이면서 '여자'라든가 또는 '기혼'이면서 미혼'이라는 중복된 범위를 동시에 갖지
못함

데이터 = 하나 이상의 변수에 대한 관찰 값의 모임

분포(distribution) : 모집단의 특성 값이 흩어져 있는 상태를 합이 1이 되는 양수로써 나타낸 것을 모집단의 분포.

변수의 종류

질적변수 : 조사대상을 특성에 따라 범주로 구분하여 측정된 변수

사칙연산을 할 수 없고, 명목형, 순위형으로 구분

양적변수

1. 이산형 변수 : 변수가 취하는 값이 셀 수 있는 경우(이산적 정수의 값)
2. 연속형 변수 : 변수가 구간 안의 모든 값을 가질 수 있는 경우(연속적 실수의 값)

데이터의 유형에 따라 표현이 달라진다

데이터 유형에 맞는 통계와 시각화

목적: 데이터의 특성(양적, 질적변수)에 따라 숫자로서 자료를 탐색하고 요약하는 방법에 대한 이해도를 높임

데이터의 유형: 정형, 비정형

1. 정형데이터 : 질적, 양적 데이터 특성을 지닌 형태가 정리되고, 가공된 데이터를 의미. 데이터베이스 등과 같이 행과 열에 맞게 정리된 자료 형태
2. 비정형데이터 : 정형화 되지 않은 상태의 데이터 (사진, 영상)
3. 반정형데이터: 비정형 데이터만큼 상태 그대로는 아니지만, 일반적인 통계분석에 바로 사용할 수 있을 만큼 정제되어있지 않는 데이터(서적의 텍스트, 신문기사)

데이터 살펴보기

데이터의 내외적 타당성

외적 : 데이터의 항목 수, 속성 목록(칼럼정의서, 코딩북), 결측값, 속성이 갖는 데이터 유형

데이터 정제 단계에서의 비표본오차, 결측 등과 함께 데이터 확인

내적 : 각 항목들이 목표 모집단으로 설계된 범위와 정의에 합당한지 확인

기초통계량을 통해 변수별 개체단위, 구성 비율 등의 오류 확인

분석기법에 합당한 데이터 셋인지를

체크 방법은 분석법에 필요한 변수들의 행과 열이 정제되어 있는지

데이터양이 많은 경우는 랜덤표본추출하여 사전분석 결과가 문제제기한 결과로 해석이 가능할지, 분석기법에 데이터 정렬이 잘 되어 있는지 사전확인

이상치 분석

이상치 발견 방법 : 개별 데이터인 내외적 데이터를 관찰하여 전체적인 패턴과 추세, 그리고 특이사항을 잘 관찰함

방대한 양의 데이터에서의 발견 방법 : 드물게 나타나는 이상치는 비표본오차 또는 단위의 착오 등에서 나타날 수 있음. head, tail 등의 무작위 추출을 통한 표본을 통해 관찰

이상치를 발견하는 통계기법 : 수치형 변수의 경우에는 상대적 위치가 반영되는 box-plot을 통해 관찰가능

빅데이터 머신러닝 기법 : K-means, Static based detection, Deviation based method, Distance based detection 등을 통해 이상치 발견

항목 간의 관계 분석

관계분석을 통해 상호 의미있는 상관관계의 항목 조합을 찾아냄

이산형변수 : Heatmap이나 scatterplot을 통해 두 속성 간의 연관성을 시각화

이산형범주변수: 카테고리별 통계치를 범주형태로 나누어서 관찰, 상자그림, 주성분 분석으로 시각화

범주형변수: 범주속성에 해당하는 값의 빈도, 분포를 관찰 pie chart, mosaic plot

*PCAplot(주성분)

여러 변수 중에 주성분이 높은 변수로 축소하는 차원 축소방법

Screeplot을 이용하여 주성분의 수를 정하고, 이를 기반하여 PCAplot. Biplot을 통해 주성분 간의 관계를 확인