Data Science

데이터 분석

데이터 탐색

강사 양석환

탐색적 데이터 분석 (EDA)

탐색적 데이터 분석 (EDA)

- 탐색적 데이터 분석(Exploratory Data Analysis, EDA)이란?
  - 수집한 데이터가 들어왔을 때, 다양한 방법을 통해서 자료를 관찰하고 이해하는 과정
  - 본격적인 데이터 분석 전에 자료를 직관적인 방법으로 통찰하는 과정
  - 데이터의 분포 및 값을 검토함으로써 데이터가 표현하는 현상을 이해하는 과정

탐색적 데이터 분석 (EDA)

- **탐색적 데이터 분석(EDA)의 필요성**
  - **데이터의 분포 및 값을 검토함으로써 데이터가 표현하는 현상을 이해하며 내재된 잠재적 문제에 대해 인식하고 해결안을 도출할 수 있음**
    - 문제점 발견 시 본 분석 전에 데이터의 수집 의사를 결정할 수 있음
  - **다양한 각도에서 데이터를 살펴보는 과정을 통해 문제 정의 단계에서 인지하지 못한 새로운 양상·패턴을 발견할 수 있음**
    - 새로운 양상 발견 시 초기 설정 문제의 가설을 수정하거나 또는 새로운 가설을 수립할 수 있음

탐색적 데이터 분석 (EDA)

- **분석 과정 및 절차**

1. **분석 목적과 변수 확인**
   - 개별 변수의 이름과 특성을 확인함
2. **데이터의 문제성 확인**
   - 결측치와 이상치 유무 등을 확인함
   - 분포상의 이상 형태를 확인함(Head 또는 Tail 부분 확인)
3. **데이터의 개별 속성값 분포 확인**
   - 기초 통계량을 통해 데이터가 예상한 범위와 분포를 가지는지 확인함
4. **데이터 사이의 관계 확인**
   - 개별 속성에서 보이지 않는 상관관계 등을 확인함

탐색적 데이터 분석 (EDA)

- 개별 데이터 관찰
  - 데이터 값을 눈으로 살펴보면서 전체적인 추세와 특이사항을 관찰함
  - 데이터의 앞/뒤 부분 관찰, 무작위 표본 추출 등을 사용함
  - 분석목적과 변수를 파악함

탐색적 데이터 분석 (EDA)

- **데이터의 문제성 확인**
  - **결측치와 이상치 유무를 확인함**
  - **데이터 문제성 확인 방법**
    - 결측치 발견 방법: 개별 데이터 관찰, 관련 함수 활용, 상관관계 활용 등
    - 이상치 발견 방법: 개별 데이터 관찰, 통계값 활용, 시각화 활용, 머신러닝 기법 활용 등
  - **결측치와 이상치가 왜 발생했는지 의미를 파악하는 것이 중요함**
  - **어떻게 대처해야 할지(제거, 대체, 유지 등)를 판단함**
    - 결측치 대치 방법: 단순대치법, 다중 대치법 등
    - 이상치 대치 방법: 제거, 대체, 유지 등

개별 데이터 관찰

- 데이터 값을 눈으로 살펴보면서 전체적인 추세와 특이사항을 관찰함
- 패턴이 어디에서 나타날지 알 수 없으므로 데이터가 많다고 앞부분만 보거나 하면 발견하지 못할 수 있음
  - 데이터의 앞과 뒤를 함께 관찰함
  - 무작위로 표본을 추출하여 관찰함
    -> 주의점: 이상치의 경우, 표본의 크기가 작다면 나타나지 않을 수 있음

탐색적 데이터 분석 (EDA)

- **데이터의 개별 속성 값 분포 확인**
  - **적절한 요약 통계지표를 사용해서 데이터를 이해할 수 있음**
    - 데이터의 중심: 평균(Mean), 중앙값(Median), 최빈값(Mode)
    - 데이터의 분산: 범위(Range), 분산(Variance), 표준편차(Standard Deviation)
  - **사분위범위(Inter Quantile Range, IQR) 방법 사용**
    - 전체 데이터를 오름차순으로 정렬한 후, 4등분하여 75%지점의 값과 25%지점의 값의 차이를 IQR로 정의
      - 최대값 = 3사분위수 + 1.5 x IQR
      - 최소값 = 1사분위수 - 1.5 x IQR
    - 결정된 최대값보다 크거나 최소값보다 작은 값을 이상치로 간주

탐색적 데이터 분석 (EDA)

- **데이터의 개별 속성 값 분포 확인**
  - **정규분포 활용 방법**
    - 평균과 분산을 이용한 이상치 제거 방법
    - 예시: $(\mu - 2\sigma)$ ~ $(\mu + 2\sigma)$ 또는 $(\mu - 1.5\sigma)$ ~ $(\mu + 1.5\sigma)$ 구간을 벗어나는 값은 이상치로 판단함
      - μ : 평균, σ : 표준편차
  - **시각화를 통해 주어진 데이터의 개별 속성 파악**
    - 확률밀도 함수, 히스토그램, 박스 플롯, 산점도 등
    - 워드 클라우드, 시계열 차트, 지도 등
  - **머신러닝 기법 활용**
    - K-Means 기법 등

탐색적 데이터 분석 (EDA)

- **데이터의 속성 간 관계 파악**
  - **상관관계 분석**
    - 두 변수 간에 선형적 관계가 있는지 분석하는 방법
    - 관계가 없으면 독립적인 관계, 관계가 존재하면 상관된 관계(correlation)임
    - 단순상관분석: 2개의 변수가 어느 정도 강한 관계에 있는지 측정
    - 다중상관분석: 3개 이상의 변수 간의 관계 강도를 측정
    - 상관분석의 기본 가정
      - 선형성, 동변량성(=등분산성, Homoscedasticity), 두 변인의 정규분포성, 무선독립표본

상관분석의 기본 가정

- 선형성: 두 변인 X, Y가 직선관계인가 확인 (산점도 활용)
- 동변량성: X의 값에 관계없이 Y의 흩어진 정도가 같은 것
- 두 변인의 정규분포성: 두 변인의 측정치 분포가 모집단에서 모두 정규분포를 이루는 것
- 무선독립표본: 모집단에서 표본을 뽑을 때, 표본 대상이 확률적으로 선정됨