1. 데이터 탐색

수집한 데이터를 분석하기 전에 통계적인 방법을 이용하여 다양한 각도에서 데이터의 특징을 파악하고 자료를 직관적으로 바라보는 방법

2. 탐색적 데이터 분석(EDA)의 4가지 주제

저항성(Resistance)

수집된 자료에 오류점, 이상값이 있을 때에도 영향을 적게 받는 성질

잔차 해석(Residual)

관찰값들이 주 경향으로부터 벗어난 정도

자료 재표현(Re-expression)

데이터 분석과 해석을 단순화할 수 있도록 원래 변수를 적당한 척도로 바꾸는 것

현시성(Graphic Representaion)

데이터 분석 결과를 이해할 수 있도록 시각적으로 표현하고 전달하는 과정

3. 상관 관계 분석

변수 사이에 존재하는 상호 연관성의 존재 여부와 연관성의 강도를 측정하여 분석하는 방법

4. 상관 관계의 표현 방법

삼전도

직교 좌표계를 이용해 변수 간의 관계를 나타내는 방법

공분산

2개의 변수 사이의 상관 정도를 나타내는 값

상관 계수

두 변수 사이의 연관성을 수치적으로 객관화하여 방향성과 강도를 표현

5. 박스 플롯

많은 데이터를 그림을 이용하여 집합의 범위와 중앙값을 빠르게 확인할 수 있으며, 통계적으로 이상값이 있는지 확인이 가능한 시각화 기법

구성 요소

하위 경계, 최소값, 제 1·2·3 사분위, 최대값, 상위 경계, 수염, 이상값