탐색적 데이터 분석 (Exploratory Data Analysis, EDA)은 데이터 분석의 첫 번째 단계로, 데이터를 깊이 이해하고 데이터의 구조, 패턴, 이상치, 상관 관계 등을 파악하는 과정입니다. EDA는 모델링을 위한 기반을 마련하고, 데이터의 품질을 평가하며, 분석의 방향성을 정하는 데 매우 중요한 역할을 합니다.

EDA는 주로 시각화와 기술적 통계를 통해 이루어지며, 데이터의 주요 특성, 분포, 상관 관계, 이상치 등을 파악하는 데 유용합니다. 이제 EDA의 주요 단계와 기법을 좀 더 자세히 설명하겠습니다.

---

1. EDA의 목표

EDA의 주요 목표는 데이터에 대한 직관적인 이해를 돕고, 분석에 필요한 의사 결정을 내리기 위한 정보를 제공하는 것입니다. 이를 통해:

- 데이터의 분포와 특징을 파악합니다.
- 이상치나 결측치를 탐지하고 처리 방법을 결정합니다.
- 변수들 간의 상관 관계를 확인하여 모델링 방향을 설정합니다.
- 데이터의 패턴과 구조를 확인하여 적합한 분석 방법을 선택합니다.

---

2. EDA의 주요 단계

A. 기술 통계 분석 (Descriptive Statistics)

기술 통계는 데이터셋의 주요 특성을 요약하고, 기초적인 통계적 값을 제공합니다. 이를 통해 데이터를 간략하게 이해할 수 있습니다.

- 기술 통계량:
    - 평균 **(Mean)**: 데이터의 평균값
    - 중앙값 **(Median)**: 데이터의 중간값
    - 최대값 **/** 최소값 **(Max / Min)**: 데이터의 최대값과 최소값
    - 표준편차 **(Standard Deviation)**: 데이터가 평균값에서 얼마나 퍼져 있는지를 나타냄
    - 사분위수 **(Quartiles)**: 데이터의 분포를 4등분하는 값 (Q1, Q2, Q3)
    - 왜도 **(Skewness)**: 데이터의 비대칭 정도
    - 첨도 **(Kurtosis)**: 데이터의 분포가 중심으로부터 얼마나 뾰족한지를 나타냄

B. 결측치 확인 및 처리

결측치는 분석에 큰 영향을 미칠 수 있으므로, 이를 적절하게 처리해야 합니다.

- 결측치 확인:
  데이터셋에 결측치가 있는지 확인하고, 결측치가 있는 변수와 행을 찾아냅니다.

- 결측치 처리 방법:
  - 삭제: 결측치가 있는 행이나 열을 삭제
  - 대체: 평균, 중앙값, 최빈값 등으로 결측치를 대체
  - 예측 모델링: 결측치를 다른 변수들로 예측하여 채우기

C. 이상치 탐지 **(Outlier Detection)**

이상치는 데이터 분석의 품질을 저하시킬 수 있으므로 이를 확인하고 처리하는 것이 중요합니다.

- 이상치 탐지 방법:
  - 박스플롯**(Box Plot)**: IQR(Interquartile Range)을 활용해 이상치를 시각적으로 식별
  - **Z-Score**: 표준편차 기준으로 3배 이상 벗어난 값을 이상치로 정의
  - **IQR (Interquartile Range)**: 상위 사분위수와 하위 사분위수 간의 범위를 계산하여 그 외의 값을 이상치로 판별

D. 데이터 분포 분석

데이터가 어떻게 분포되어 있는지 시각적으로 확인하여 변수 간의 특성을 파악합니다.

- 히스토그램 **(Histogram)**:
  연속형 변수의 분포를 시각적으로 확인할 수 있는 도구입니다. 예를 들어, 연봉이나 경력 연수의 분포를 파악할 수 있습니다.

- 확률 밀도 함수 **(Density Plot)**:
  데이터가 연속적일 때, 히스토그램과 함께 사용하여 데이터의 분포를 매끄럽게 시각화할 수 있습니다.

- 박스 플롯 **(Box Plot)**:
  데이터의 중앙값, 사분위수, 이상치 등을 시각적으로 보여주는 그래프입니다. 데이터의 분포와 이상치를 쉽게 확인할 수 있습니다.

E. 변수 간 상관 관계 분석 **(Correlation Analysis)**

변수들 간의 관계를 파악하여, 분석 또는 모델링 시 중요한 변수 간의 상호작용을 이해할 수 있습니다.

- 상관 행렬 **(Correlation Matrix)**:
  변수 간의 상관 관계를 계산하고 이를 시각적으로 표현한 행렬입니다. 상관 계수는 -1에서 1까지의 값으로 표현되며, 1에 가까운 값은 강한 양의 상관 관계를, -1에 가까운 값은 강한 음의 상관 관계를 의미합니다.

- 산점도 **(Scatter Plot)**:
  두 변수 간의 관계를 시각적으로 표현합니다. 예를 들어, 연봉과 경력 연수 간의 관계를 산점도로 확인할 수 있습니다.

- **Heatmap**:
  상관 행렬을 색깔로 나타내어 변수 간의 관계를 한눈에 파악할 수 있습니다.

F. 시각적 탐색

EDA에서 중요한 것은 시각화입니다. 데이터를 시각적으로 나타내면 숨겨진 패턴이나 트렌드를 더 쉽게 발견할 수 있습니다.

- 히스토그램 **(Histogram)**, 산점도 **(Scatter Plot)**, 상자 수염 그림 **(Box Plot)**, 산점도 행렬 **(Pair Plot)** 등의 다양한 시각화를 사용해 데이터를 탐색합니다.

- 파이 차트 **(Pie Chart)**: 범주형 데이터의 분포를 확인할 때 유용합니다. 예를 들어, 채용 공고가 어느 산업군에 속하는지 시각적으로 나타낼 수 있습니다.

---

3. EDA 도구와 기법

- **Python** 라이브러리:
  - **Pandas**: 데이터의 요약 통계 및 결측치 확인, 데이터 조작 등 기본적인 작업에 사용됩니다.
  - **Matplotlib / Seaborn**: 시각화 라이브러리로, 다양한 그래프를 그릴 수 있습니다.
  - **NumPy**: 수치 계산을 위한 라이브러리로, 기술 통계와 배열 연산에 사용됩니다.
  - **Scikit-learn**: 데이터 전처리와 머신러닝을 위한 라이브러리로, 모델링 전 데이터 분석에 유용합니다.
- **R**:
  - **ggplot2**: 강력한 시각화 라이브러리로, 데이터의 다양한 측면을 시각적으로 표현하는 데 유용합니다.
  - **dplyr**: 데이터 조작 및 처리에 유용한 라이브러리입니다.

---

4. EDA의 최종 목표

EDA의 주요 목표는 데이터에 대한 직관적인 이해를 통해 분석의 방향성을 잡고**,** 모델링에 적합한 데이터를 준비하는 것입니다. 이를 통해, 채용 정보 분석 플랫폼에서 사용자가 더 쉽게 데이터를 이해하고, 의미 있는 패턴과 인사이트를 도출할 수 있게 됩니다.