Pandas(판다스)란?

- 파이썬의 데이터 분석 라이브러리
- 데이터 조작, 정제, 분석, 시각화 등을 위한 다양한 기능을 제공
- 판다스는 시리즈(Series)와 데이터프레임(DataFrame)이라는 자료형을 이용하여 데이터 처리함
- 2008년 금융 데이터 분석용으로 처음 개발되어 데이터 수집, 정리에 최적화된 도구
- 파이썬 기반의 무료 오픈소스
- 통계, 데이터과학(80~90% 업무 처리 가능), 머신러닝 분야에서 중요하게 사용
- 코랩 사용시 별도의 설치가 필요 없다 (설치: pip install pandas)

데이터 분석 단계

1. 문제 정의

> 데이터 분석의 목적을 명확히 이해하고, 분석이 필요한 문제를 정의합니다. 예를 들어, 고객 이탈률 예측, 매출 증대 전략 수립, 제품 품질 향상 등의 문제를 정의할 수 있습니다.

2. 데이터 수집

> 분석에 필요한 데이터를 수집합니다. 이 데이터는 내부 시스템, 외부 데이터베이스, API, 웹 스크레이핑 등 다양한 소스에서 가져올 수 있습니다.

3. 데이터 전처리

> 수집한 데이터를 정제하고 전처리합니다. 이 단계에서는 결측치 처리, 이상치 제거, 데이터 형식 변환, 데이터 정규화 등의 작업을 수행합니다.

4. 탐색적 데이터 분석 (EDA)

> 데이터를 탐색하고 시각화하여 데이터의 특성을 파악합니다. 이를 통해 데이터 간의 관계를 이해하고 패턴을 발견할 수 있습니다.

5. 모델링

> 데이터에 적합한 모델을 선택하고 학습시킵니다. 이 단계에서는 회귀 분석, 분류, 군집화, 시계열 예측 등의 다양한 머신러닝 알고리즘을 사용할 수 있습니다.

6. 모델 평가

> 학습된 모델의 성능을 평가합니다. 이를 통해 모델의 예측 능력을 검증하고 필요한 경우 모델을 수정하거나 다른 모델을 시도할 수 있습니다.

7. 결과 해석

> 분석 결과를 해석하고 비즈니스에 적용 가능한 인사이트를 도출합니다. 이를 통해 의사 결정을 지원하고 비즈니스 목표를 달성하는 방안을 제시합니다.

8. 배포 및 유지보수

> 최종 모델을 배포하고 사용자에게 제공합니다. 또한 모델의 성능을 모니터링하고 필요에 따라 모델을 업데이트하거나 유지보수합니다.

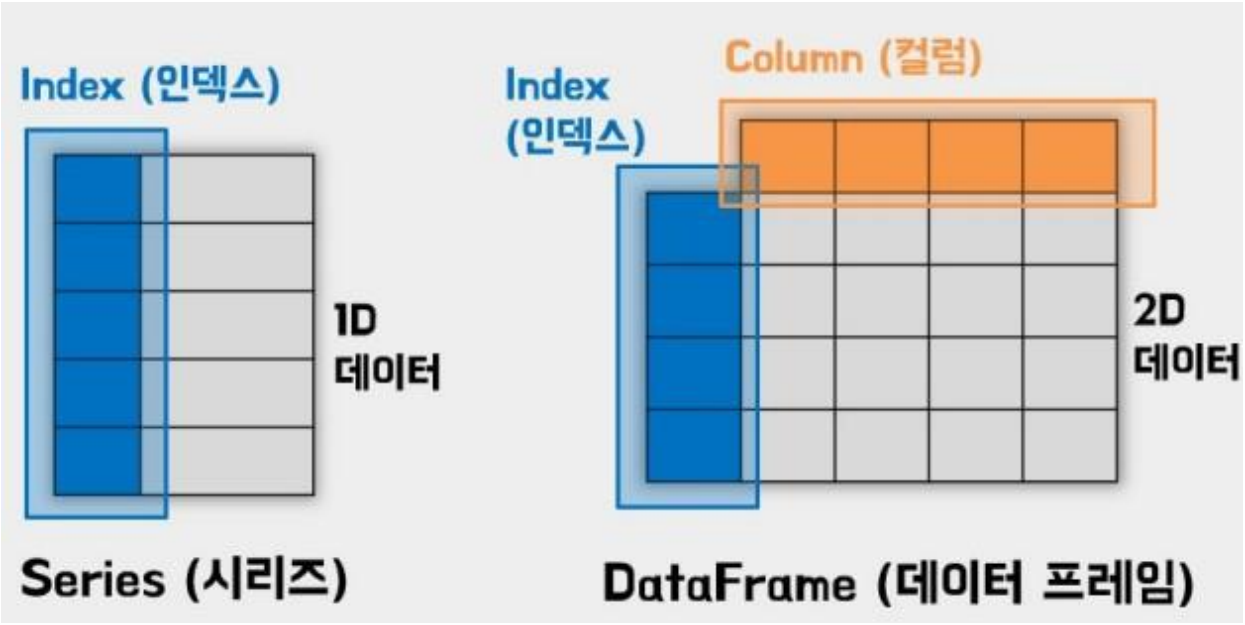
Pandas의 자료 구조

- 시리즈(Series):
  - 시리즈는 1차원 배열과 유사한 데이터 구조입니다.
  - 각 요소는 값과 인덱스로 구성됩니다.
  - 값은 데이터의 실제 내용을 나타내고, 인덱스는 각 값에 대한 레이블이 됩니다.
  - 시리즈는 판다스의 Series 클래스를 사용하여 생성됩니다.

    > 예를 들어, 주가 데이터의 경우 날짜가 인덱스이고 해당 날짜의 주가가 값으로 저장될 수 있습니다.

- 데이터프레임(DataFrame):
  - 데이터프레임은 2차원 표 형식의 데이터 구조입니다.
  - 행과 열로 구성되며, 각 열은 시리즈로 구성됩니다.
  - 데이터프레임은 판다스의 DataFrame 클래스를 사용하여 생성됩니다.
  - 데이터프레임은 여러 유형의 데이터를 저장할 수 있으며, 행과 열에 대한 레이블을 사용하여 데이터에 쉽게 접근할 수 있습니다.
  - 엑셀 스프레드시트나 데이터베이스의 테이블과 유사한 형태를 가지고 있어 데이터의 구조를 이해하기 쉽습니다.

시리즈와 데이터프레임


Index (인덱스)
1D
데이터
Series (시리즈)
Column (컬럼)
Index
(인덱스)
2D
데이터
DataFrame (데이터 프레임)