데이터 분석 실습

사용하는 도구

- Python
- Jupyter Notebook
- pandas
- numpy
- matplotlib
- seaborn
- scipy
- statsmodels

pandas와 dataframe

기초 개념

- pandas 라이브러리
  - 파이썬에서 데이터를 쉽게 다루고 분석할 수 있도록 설계된 오픈 소스 라이브러리
  - 데이터 분석과 관련된 다양한 작업을 간편하게 수행할 수 있음
  - 특히 테이블 형식의 데이터를 다루는데 최적화
  - 엑셀이나 SQL과 유사하게 데이터를 조작할 수 있는 직관적인 프로그래밍 기능을 제공
- 데이터프레임과 시리즈
  - 데이터프레임 - Pandas의 기본 데이터 구조로 행과 열로 구성된 2차원 데이터 (테이블)
  - 시리즈 - 1차원 배열의 구조로 데이터 프레임의 열 하나를 구성하는 구조

Pandas 사용법

- 제공되는 노트북 (pandas.ipynb 파일)을 참고 바랍니다.

matplotlib

기초 개념

- 데이터 시각화에 사용할 수 있는 강력한 라이브러리
- 주피터 노트북과 결합해서 매우 편리하게 사용할 수 있음

matplotlib으로 만들 수 있는 시각화

- 선 그래프
- 막대 그래프
- 히스토그램
- 산포도
- 파이차트
- 박스플롯
- 히트맵
- 컨투어 플롯 (등고선 그래프)
- 밀도 그래프
- 3D 그래프
- 레이더 차트 - 여러 변수의 값을 동일한 척도로 비교
- 스택 막대 그래프

matplotlib 사용법

- 제공되는 노트북 (matplotlib.ipynb 파일)을 참고 바랍니다.

CSV 파일에 대해서

- 반 정형화된 데이터 파일
- 컴마로 구분된 텍스트 데이터 (CSV : Comma-Seperated Value)
    - 컴마가 있는 데이터는 데이터를 따옴표로 감싸는 형태로 사용
- 행과 열이 있는 데이터

- 제목행이 존재할 수도 존재하지 않을 수도 있음
  - 제목행이 없을 때는 `pandas`에서 csv 파일을 읽어들일 때 `header = None` 옵션을 사용
  - `df = pd.read_csv(filename, header=None)`

```
이름,나이,성별
도널드덕,20,남
"D,Jr",30,여
```

- VS Code의 경우 Extension을 활용하면 좋음!