pandas 과제 - 백준호

1.

pandas는 데이터 조작 및 분석을 위한 파이썬 라이브러리로, 다양한 데이터 형식을 손쉽게 다루게 해준다. 특히 DataFrame과 Series 객체를 통해 데이터를 효율적으로 관리하고 조작할 수 있다. 공식 문서를 통해 발견한 흥미로운 점 중 하나는 pd.merge 함수의 다양한 사용법이었다. 이 함수는 SQL의 JOIN 기능과 유사하며, 데이터프레임 간의 복잡한 조인 연산을 매우 간단하게 처리하게 해준다.

2.

```python
import pandas as pd
import json
```

✓ 0.0s

```python
# JSON 파일 경로
file_path = 'results.json'

# JSON 파일을 읽어 데이터프레임으로 변환
with open(file_path, 'r', encoding='utf-8') as file:
    data = json.load(file)

# 데이터를 DataFrame으로 변환
df = pd.DataFrame(data)
```

✓ 0.1s

```python
df.head()
```

✓ 0.0s Python

	date	date_edit	href	title	content
0	2024.07.25 23:50	2024.07.25 23:50	https://www.hankyung.com/article/202407250313i	딜리버스, 146억 투자 유치...국제드론쇼 참가한 파블로항공 [Geeks' Briefing]	한국경제신문의 프리미엄 스타트업 미디어 플랫폼 긱스(Geeks)가 25일 스타트업 ...
1	2024.07.25 23:05	2024.07.25 23:05	https://www.hankyung.com/article/202407250284Y	美 2분기 성장률 2.8%...고금리 지속에도 탄탄한 성장세 (종합)	소비·재고투자 힘입어 1분기보다 더 상승...시장 전망도 웃돌아\n재고투자 '반짝 효과...
2	2024.07.25 22:31	2024.07.25 22:31	https://www.hankyung.com/article/202407250251Y	위메프 1천400명 환불 완료..."여행상품 이어 일반상품도 환불"(종합3보)	현장 창구 없는 티몬 사무실, 환불요청 고객 점거..."환불 진행 중"\n위메프 대표 ...
3	2024.07.25 22:00	2024.07.25 22:45	https://www.hankyung.com/article/202407258535g	"돈 다 냈는데 여행사가 책임지고 보내줘야죠"...불만 터졌다 [일파만파 티메프]	25일 오후 서울 강남구 티몬 본사 사옥 앞에서 '정산 지연 사태'로 환불을 요구하...
4	2024.07.25 21:39	2024.07.25 21:50	https://www.hankyung.com/article/202407250217Y	美 2분기 경제성장률 2.8%...시장 전망 상회	미 상무부는 2분기 미국의 국내총생산(GDP) 증가율(속보치)이 2.8%(직전분기 ...

```
df.info()
```
✓ 0.0s

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8255 entries, 0 to 8254
Data columns (total 5 columns):
 #   Column     Non-Null Count  Dtype 
---  ------     --------------  ----- 
 0   date       8255 non-null   object
 1   date_edit  8255 non-null   object
 2   href       8255 non-null   object
 3   title      8255 non-null   object
 4   content    8255 non-null   object
dtypes: object(5)
memory usage: 322.6+ KB
```

3.

4. **pickle**

• 파이썬 객체 구조를 직렬화하여 파일로 저장할 수 있게 해주는 포맷으로, 데이터 분석 작업 후 객체를 재사용하고자 할 때 유용하며, 파이썬 객체를 저장하고 불러올 때 사용된다.

2. **CSV, TSV**

• 쉼표(Comma)나 탭(Tab)으로 구분된 값들로 이루어진 텍스트 파일 포맷으로, 데이터 교환의 표준 포맷이며, 데이터베이스, 스프레드시트 등에서 데이터를 내보내고 불러올 때 사용된다.

3. **JSON**

• 경량의 데이터 교환 포맷으로, 사람이 읽고 쓰기 쉬우며, 기계가 분석하고 생성하기 쉬운 형식으로, 웹 애플리케이션에서 주로 사용되며 API 응답, 설정 파일 등으로 사용된다.

4. **HTML**

• 하이퍼텍스트 마크업 언어로 웹 페이지를 작성하는데 사용되며, 웹 페이지를 만들 때 필요하고 웹 데이터를 저장하고 불러올 때 사용된다.

5. **XML**

• 데이터의 계층 구조를 표현할 수 있는 마크업 언어로, 데이터 교환과 저장에서 많이 사용되며, 설정 파일, 데이터 교환에 사용된다.

6. **Parquet**

• 열 지향 저장 포맷으로, 대용량 데이터 처리에 효율적이며, 빅데이터 처리와 분석에 사용되며, Hadoop, Spark 등의 빅데이터 플랫폼에서 주로 사용된다.

7. **YAML**

• 사람이 읽기 쉬운 데이터 직렬화 포맷으로, 설정 파일이나 데이터 저장에서 주로 사용되며, 설정 파일, 데이터 저장에 사용된다.

8. **TOML**

• 간결하고 명확한 구문을 가진 설정 파일 포맷으로, 설정 파일 작성에 적합하며, 설정 파일에 사용된다.

pickle에 대한 추가 조사

• 객체의 상태를 저장하거나 전송할 수 있도록 바이트 스트림으로 변환하는 과정인 직렬화(serialization)와 바이트 스트림을 다시 객체로 변환하는 과정인 역직렬화(deserialization)가 필요한데, 이는 네트워크 전송, 파일 저장 등에서 객체 상태를 보존하고 전송하기 위해 필요하다.