7회차 과제 - Pandas

Date	@2024년 8월 6일 오후 11:59
Tag	과제

1번 과제

pandas.DataFrame.rename

- column이나 index의 label을 변경하는 메서드
- 일반적으로 mapper / column / index 에 dictionary를 전달하는 방식으로 데이터프레임의 column이나 index label을 수정하지만,

```
air_quality_renamed = air_quality.rename(
     columns={
         "station_antwerp": "BETR801",
         "station_paris": "FR04014",
         "station_london": "London Westminster",
     }
)
```

  mapper / column / index 는 함수도 인자로 받을 수 있습니다.

```
air_quality_renamed = air_quality_renamed \
.rename(columns=str.lower)
```

pandas.pivot, pandas.melt

- pandas.pivot을 통해 엑셀의 피벗테이블과 같은 기능을 사용할 수 있으며, 피벗테이블을 원본 테이블로 되돌리기 위해서는 pandas.melt를 사용할 수 있습니다.

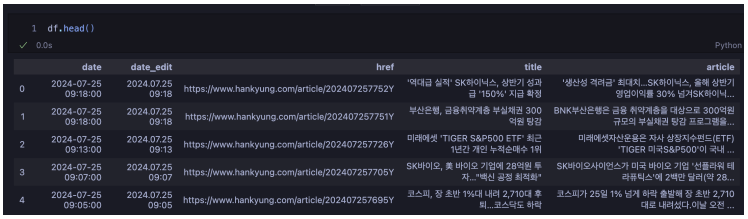
pandas.DataFrame.plot

- matplotlib 등 다른 라이브러리 호출 없이도 데이터를 시각화할 수 있는 pandas 내부 시각화 함수입니다. 다만 내부적으로는 matplotlib을 사용한다고 합니다.

```
ser = pd.Series([1, 2, 3, 3])
plot = ser.plot(kind='hist', title="My plot")
```

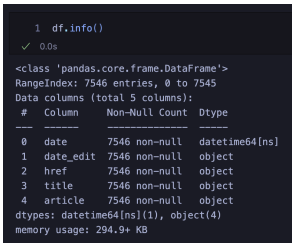
2번 과제

- DataFrame.head()



	date	date_edit	href	title	article
0	2024-07-25 09:18:00	2024.07.25 09:18	https://www.hankyung.com/article/202407257752Y	'역대급 실적' SK하이닉스, 상반기 성과급 '150%' 지급 확정	'생산성 격려금' 최대치...SK하이닉스, 올해 상반기 영업이익률 30% 넘겨SK하이닉...
1	2024-07-25 09:18:00	2024.07.25 09:18	https://www.hankyung.com/article/202407257751Y	부산은행, 금융취약계층 부실채권 300억원 탕감	BNK부산은행은 금융 취약계층을 대상으로 300억원 규모의 부실채권 탕감 프로그램을...
2	2024-07-25 09:13:00	2024.07.25 09:13	https://www.hankyung.com/article/202407257726Y	미래에셋 'TIGER S&P500 ETF' 최근 1년간 개인 누적순매수 1위	미래에셋자산운용은 자사 상장지수펀드(ETF) 'TIGER 미국S&P500'이 국내 ...
3	2024-07-25 09:07:00	2024.07.25 09:07	https://www.hankyung.com/article/202407257705Y	SK바이오, 美 바이오 기업에 28억원 투자..."백신 공정 최적화"	SK바이오사이언스가 미국 바이오 기업 '선플라워 테라퓨틱스'에 2백만 달러(약 28...
4	2024-07-25 09:05:00	2024.07.25 09:05	https://www.hankyung.com/article/202407257695Y	코스피, 장 초반 1%대 내려 2,710대 후퇴...코스닥도 하락	코스피가 25일 1% 넘게 하락 출발해 장 초반 2,710대로 내려섰다.이날 오전 ...

- DataFrame.info()



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7546 entries, 0 to 7545
Data columns (total 5 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   date       7546 non-null   datetime64[ns]
 1   date_edit  7546 non-null   object
 2   href       7546 non-null   object
 3   title      7546 non-null   object
 4   article    7546 non-null   object
dtypes: datetime64[ns](1), object(4)
memory usage: 294.9+ KB
```

3번 과제

1. Pickle

- **특징**: Python 객체를 직렬화 및 역직렬화하는 바이너리 포맷
  - 직렬화: Python 객체 계층 구조를 바이트 스트림으로 변환
  - 역직렬화: 바이트 스트림을 객체 계층 구조로 복원
- **필요성**: Python 객체를 그대로 저장하고 나중에 복원하는 데 유용
- **주된 사용처**: 모델 학습 결과, 객체 데이터 저장 등 Python 환경에서의 데이터 저장 및 복원

2. CSV (Comma-Separated Values) / TSV (Tab-Separated Values)

- **특징**: 텍스트 형식으로, 각 행은 레코드이고 열은 콤마(CSV) 또는 탭(TSV)으로 구분됨
- **필요성**: 단순한 테이블 형식 데이터 저장 및 교환
- **주된 사용처**: 스프레드시트, 간단한 데이터 교환, 데이터 분석 초기 단계

3. JSON (JavaScript Object Notation)

- **특징**: 경량의 데이터 교환 포맷으로, 사람이 읽기 쉽고 기계가 해석하고 생성하기 쉬운 텍스트 형식
- **필요성**: 데이터 교환의 표준 포맷으로 다양한 언어 및 환경에서 사용 가능
- **주된 사용처**: 웹 API, 구성 파일, 로그 데이터

4. HTML (HyperText Markup Language)

- **특징**: 웹 페이지를 작성하기 위한 마크업 언어로, 텍스트와 멀티미디어를 구조화
- **필요성**: 웹 콘텐츠 작성 및 배포
- **주된 사용처**: 웹 사이트 및 웹 애플리케이션 인터페이스

5. XML (eXtensible Markup Language)

- **특징**: 사용자 정의 태그를 사용하여 데이터를 구조화하는 마크업 언어
- **필요성**: 다양한 시스템 간의 데이터 교환
- **주된 사용처**: 웹 서비스, 구성 파일, 문서 저장

6. Parquet

- **특징**: 컬럼 지향의 저장 포맷으로, 효율적인 데이터 압축 및 스캔 가능
- **필요성**: 대용량 데이터의 저장 및 고속 처리
- **주된 사용처**: 빅 데이터 처리 및 분석 (ex.Apache Hadoop, Spark)

7. YAML (YAML Ain't Markup Language)

- **특징**: 사람이 읽기 쉬운 데이터 직렬화 표준으로, 들여쓰기를 통해 계층 구조를 나타냄
- **필요성**: 구성 파일 및 데이터 직렬화
- **주된 사용처**: 설정 파일 (ex. `docker-compose.yml`)

8. TOML (Tom's Obvious, Minimal Language)

- **특징**: 단순하고 명확한 구성 파일 형식으로, 데이터 타입을 명확히 표현
- **필요성**: 읽기 쉽고 명확한 구성 파일 작성
- **주된 사용처**: 프로젝트 설정 파일 (ex.: Python 프로젝트의 `pyproject.toml`)

Pandas I/O로서의 각 데이터 저장 포맷

Format Type	Data	Reader	Writer
binary	Pickle	read_pickle	to_pickle
text	CSV	read_csv	to_csv
text	JSON	read_json	to_json
text	HTML	read_html	to_html
text	XML	read_xml	to_xml
binary	Parquet	read_parquet	to_parquet