

# pandas report

## 공식문서 내용: copy on write

```
In [1]: df = pd.DataFrame({"foo": [1, 2, 3], "bar": [4, 5, 6]})
In [2]: subset = df["foo"]
In [3]: subset.iloc[0] = 100
In [4]: df
Out[4]:
```

	foo	bar
0	100	4
1	2	5
2	3	6

과거에는 DataFrame에서 추출된 view를 수정하면 원본 DataFrame의 값도 변하게 동작이 되었습니다. 이는 딱 봐도 찾기 어려운 오류를 유발하는 형태였습니다. pandas 3.0부터 기본적으로 동시에 2개 이상의 pandas 객체를 수정하지 못도록 copy on write 원칙을 도입했습니다.

```
In [5]: pd.options.mode.copy_on_write = True
In [6]: df = pd.DataFrame({"foo": [1, 2, 3], "bar": [4, 5, 6]})
In [7]: subset = df["foo"]
In [8]: subset.iloc[0] = 100
In [9]: df
Out[9]:
```

	foo	bar
0	1	4
1	2	5
2	3	6

## 크롤링 json pandas로 불러오기

```
In 3 | df.head()
Executed at 2024.08.01 21:01:00 in 16ms

Out 3 |
date      | date_edit | href | title | article
2024-07-22 22:30:00 | 2024.07.22 22:30 | https://www.hankyung.com/article/2024072215051 | 글로벌 IT대란 '크라우드스트라이크' 투자와 사건=게티이미지저널 |
2024-07-22 21:54:00 | 2024.07.22 21:54 | https://www.hankyung.com/article/2024072214725 | 대한항공, 보잉 항공기 최대 50대 구매...30 대한항공이 미국 보잉 |
2024-07-22 21:20:00 | 2024.07.22 21:20 | https://www.hankyung.com/article/2024072214545 | 지속은행 기업 마중 수수료율 하향...공시도 | 상환이 없었던 지속은 |
2024-07-22 21:18:00 | 2024.07.22 21:18 | https://www.hankyung.com/article/2024072214515 | 최상목, '전국민 25만원'에 "부작용 우려도 최상목 부총리 겸 기 |
2024-07-22 21:11:00 | 2024.07.22 21:11 | https://www.hankyung.com/article/2024072214211 | 테슬라, 민간형로봇 옵티머스 내년부터 공장 테슬라 휴머노이드 로 |

In 4 | df.info
Executed at 2024.08.01 21:01:30 in 7ms

Out 4 |
article
0   사건=게티이미지저널 주 결합있는 보안 소프트웨어 업데이트로 전세계 윈도우 운영체제를...
1   대한항공이 미국 보잉의 최형단 증대형 항공기인 777-9와 787-10을 도입하여 ...
2   상환이 없었던 지속은행 기업 한도대출(마이너스대출) 수수료가 합리적으로 개선되고,...
3   최상목 부총리 겸 기획재정부 장관최상목 부총리 겸 기획재정부 장관은 22일 다보어만,...
4   테슬라 휴머노이드 로봇 옵티머스, 사건=테슬라테슬라의 최고경영자(CEO) 일론 머스...
...
7674 엔비디아의 주가가 이를 연속으로 35대의 하락세를 보이고 있다.21일(현지시간) 마...
7675 이번 주 국내 주유소 휘발유와 경유의 주간 평균 판매 가격이 동반 하락세를 보였다....
7676 뉴욕증시는 '인공지능(AI) 선두주자'인 엔비디아가 하락세를 보이면서 기술주 투자 ...
7677 이번 주 국내 주유소 휘발유와 경유의 주간 평균 판매 가격이 동반 하락세를 보였다....
7678 5월 들어 미국의 주택거래가 3개월 연속 하락한 가운데 집값은 최고 수준을 경신했다...

[7679 rows x 5 columns]>
```

## 데이터 저장 포맷

- pickle
  - 파이썬 객체를 직렬화, 역직렬화 즉 객체와 파일형태로 변환에서 쓰입니다.
  - python에서 객체를 로드하거나 학습된 모델을 저장할때 많이 사용됩니다.
- CSV, TSV
  - 텍스트 파일 형식으로, 데이터를 csv는 쉼표 tsv는 탭으로 구분하여 저장합니다. 간단한 형식으로, 각 줄이 한 레코드를 나타내며, 각 필드는 쉼표로 구분됩니다.
  - 데이터분석에서 널리활용됩니다.
- JSON
  - key-value형태로 데이터를 저장합니다. 가독성이 xml같은 것에 비해 좋습니다
  - 웹api, db저장 포맷으로 많이 쓰입니다.
- HTML
  - 웹사이트의 마크업 언어입니다. 웹사이트의 구조와 내용을 정의합니다.
  - 웹개발에서 필수적입니다
- XML
  - 가독성이 많이 떨어집니다. 태그를 사용하여 데이터를 계층적으로 표현합니다.
  - 설정파일, 웹api 에서 포맷으로 많이 활용됩니다.
- Parquet

- 하둡에서 사용되는 방식이다. 열기반으로 저장하는 포맷이다. 열에는 유사한 데이터들이 있기 때문에 압축에 유리하다!
- 빅데이터 처리(하둡)에서 많이 쓰이는 포맷이
- YAML
  - 인간의 가독성에 신경쓴 포맷입니다. 들여쓰기(2칸)으로 계층적으로 데이터를 표현합니다.
  - 설정파일, 문서화의 포맷으로 자주 활용됩니다.
- TOML
  - 구문이 간단해서 YAML과 비슷하게 가독성이 좋지만 유연성은 부족합니다.
  - 설정파일의 포맷으로 자주 활용됩니다.