

[ C U A I - 금 융 스 터 디 ]

# 웹 스크레이핑을 사용한 데이터 분석

---

20193878 최은서



[한국거래소]

종목코드 구하기

[네이버금융 - 웹 스크래이핑]  
일자별 주가데이터  
수집

[- 캔들 차트]

주가데이터 그리기

# 종목코드 구하기

- 홈페이지에서 상장법인목록 다운로드(excel)

Cf. 파일읽기(확장자 .xls)

HTML형태=>read\_html()

엑셀로 개봉->.xlsx로 다시 저장하면 read\_excel()함수 이용가능.

- Jupyter notebook 이용

□ 리스트의 첫 번째 원소를 인덱싱해 데이터 프레임으로 출력.

□ map() 함수를 이용해 종목코드 자리 맞추기(6자리)

cf. 주석은 url을 데이터프레임으로 -> □ 수행 -> 종목코드 오름차순 정렬 by sort\_values함수.

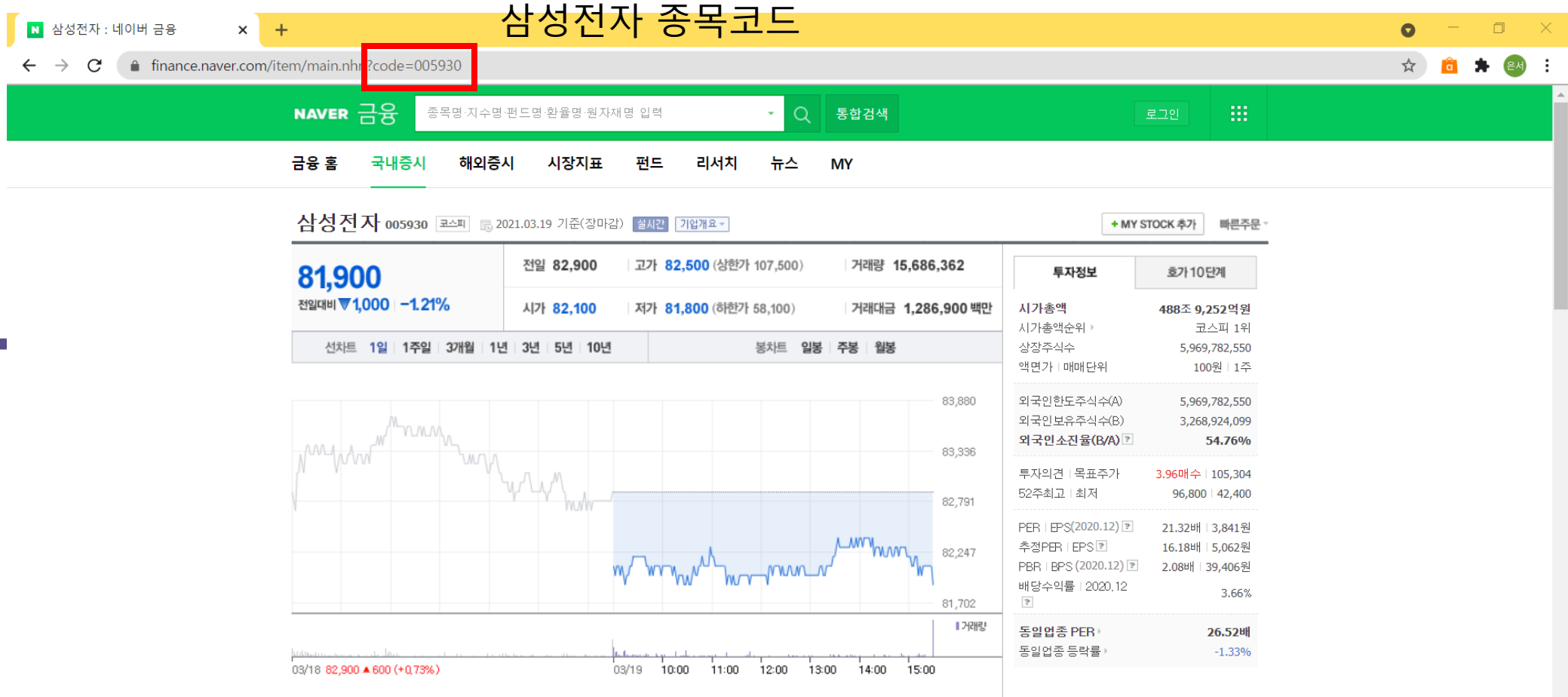
다른 문서 연결    텍스트가 화면에서 어떻게 보여야 하는지 정의하는 기법. <>: 태그이용

# HTML

H y p e r T e x t   M a r k u p   L a n g u a g e

- 웹 페이지: 다양한 언어로 작성-(렌더링)->HTML로 변환된 결과
  - 주요태그-책 참고.
  - 기본 : <head>(:HTML 페이지 정보)+<body>(:보이는 부분)
-

# 웹에서 일별 시세 구하기



# 네이버 금융 일별 시세 분석

일별시세

날짜	종가	전일비	시가	고가	저가	거래량
2021.01.19	87,000	▲ 2,000	84,500	88,000	83,600	39,895,044
2021.01.18	85,000	▼ 3,000	86,600	87,300	84,100	43,227,951
2021.01.15	88,000	▼ 1,700	89,800	91,800	88,000	33,431,809
2021.01.14	89,700	0	88,700	90,000	88,700	26,393,970
2021.01.13						
2021.01.12						
2021.01.11						
2021.01.08						
2021.01.07						
2021.01.06						

뒤로	Alt+왼쪽 화살표
앞으로(F)	Alt+오른쪽 화살표
새로고침	Ctrl+R
다른 이름으로 저장...	Ctrl+S
인쇄	Ctrl+P
전송...	
이 페이지의 QR 코드 생성	
한국어(으)로 번역	
페이지 소스 보기	Ctrl+U
프레임 소스 보기	
프레임 새로고침	
검사	Ctrl+Shift+I

인기검색

- 1 삼성
- 2 한국
- 3 카카
- 4 현대
- 5 SK
- 6 LG
- 7 HMA
- 8 SK
- 9 유한
- 10 셀트

시세정

코스I  
코스I  
상한  
거래  
고가  
시가  
외국

# 소스코드에서 링크 주소 검색

```
← → ↻ ① view-source:https://finance.naver.com/item/sise_day.nhn?code=005930&page=5
324 </td>
325 <td>
326   <a href="/item/sise_day.nhn?code=005930&page=6" >6</a>
327 </td>
328 <td>
329   <a href="/item/sise_day.nhn?code=005930&page=7" >7</a>
330 </td>
331 <td>
332   <a href="/item/sise_day.nhn?code=005930&page=8" >8</a>
333 </td>
334 <td>
335   <a href="/item/sise_day.nhn?code=005930&page=9" >9</a>
336 </td>
337 <td>
338   <a href="/item/sise_day.nhn?code=005930&page=10" >10</a>
339 </td>
340
341 <td class="pgR">
342   <a href="/item/sise_day.nhn?code=005930&page=11" >
343   다음
344   </a>
345 </td>
346
347 <td class="pgRR">
348   <a href="/item/sise_day.nhn?code=005930&page=622" >맨뒤</a>
349   
350   </a>
351 </td>
352
```

# 뷰티풀 수프로 일별 시세 읽어오기

- 웹 크롤러 : 웹에서 링크된 페이지들을 돌아다니며 데이터 읽음.
- 웹 스크레이퍼 : 크롤링해서 모은 데이터에서 원하는 정보를 추출
- HTML -> XML형태의 파이썬 객체로 변환

(책 참고)


- 파서 종류

lxml파서 이용(속도)

- find\_all() vs find()

[limit=개수]로 찾을 개수 지정 가능, 한 개이면 find() 함수 사용





# Jupyter notebook 참고

- 맨 뒤 페이지 구하기
  - 전체 페이지 읽어오기
-

# O H L C 와 캔들 차트

- OPEN HIGH LOW CLOSE(시가, 고가, 저가, 종가)

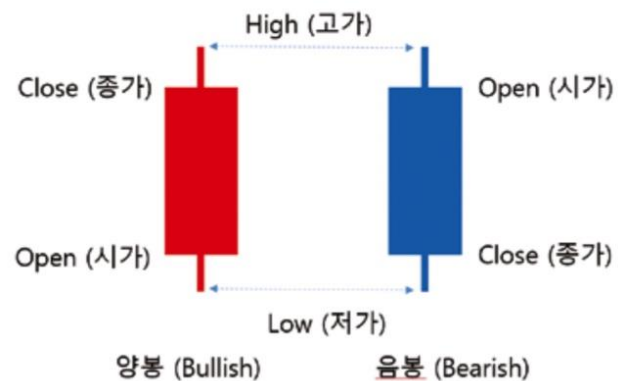
-색깔 : 시가-종가

-실선 : 저가-고가

- OHLC 차트(미국) vs 캔들차트(대한민국)



OHLC 차트

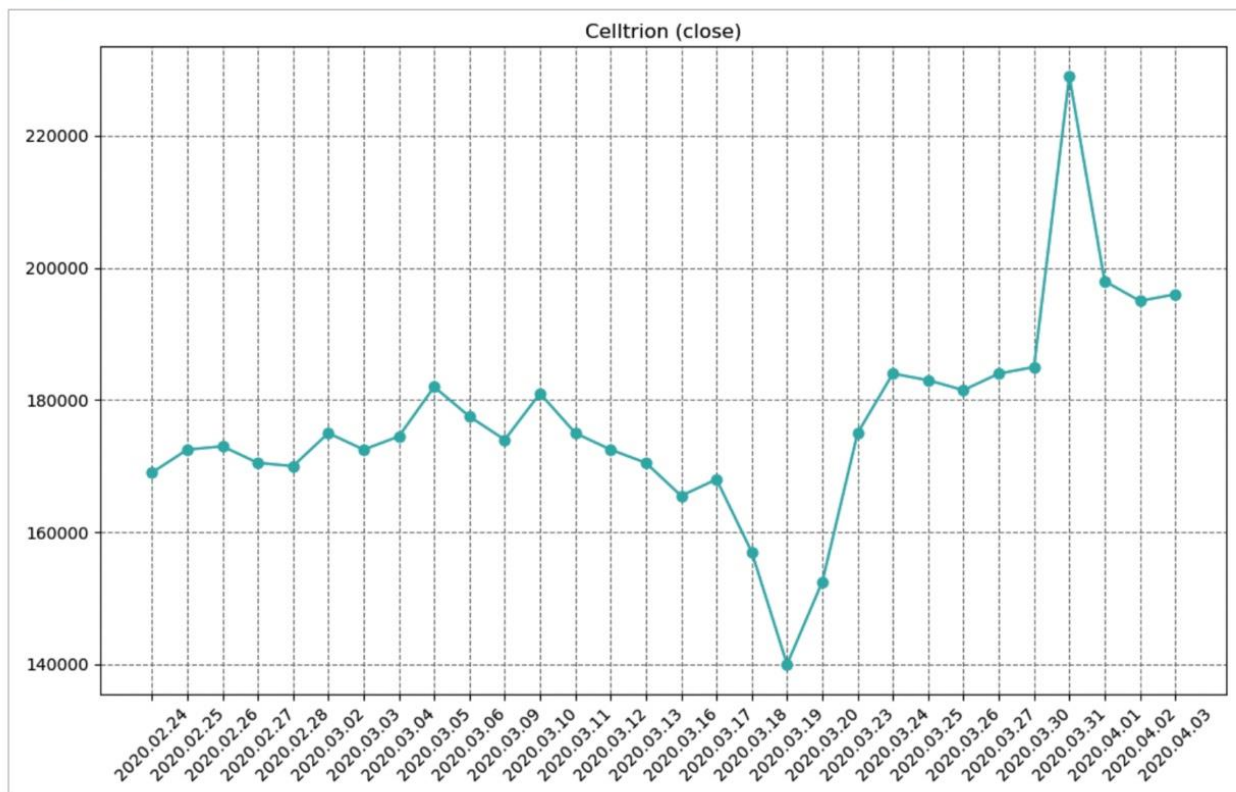


캔들 차트

# Jupyter notebook 참고

- 종가만으로 가격 변동을 표시

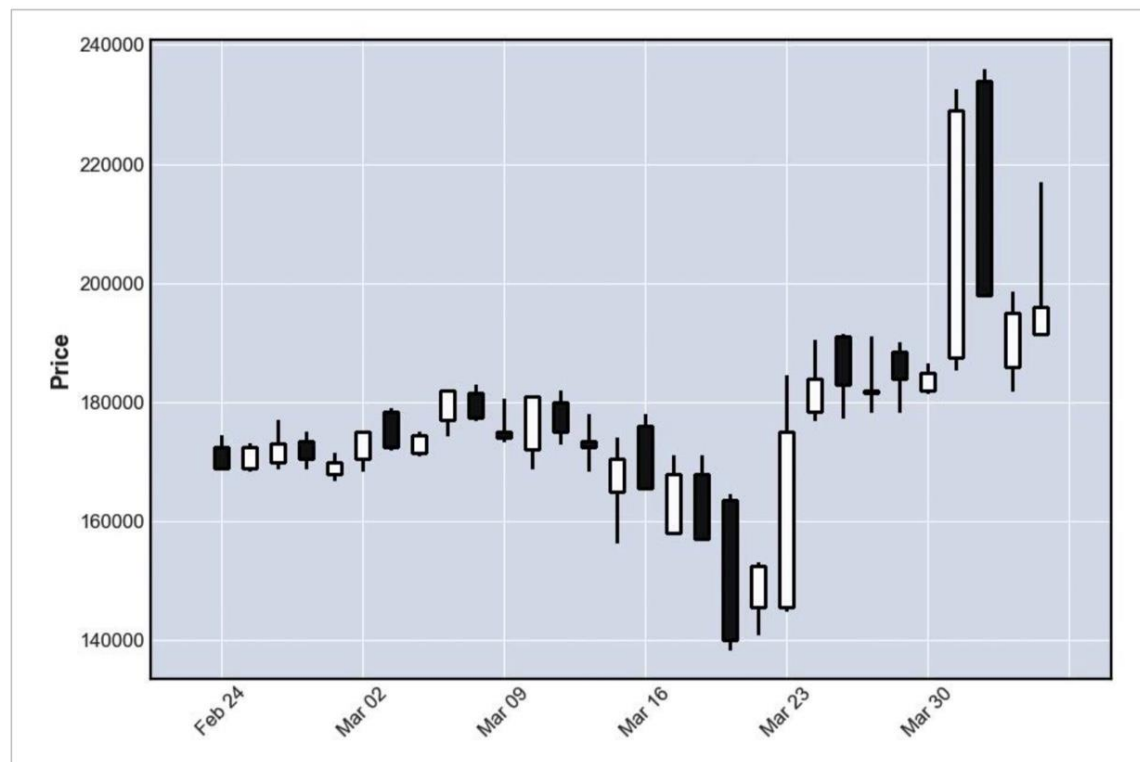
그림\_ 셀트리온 종가 차트



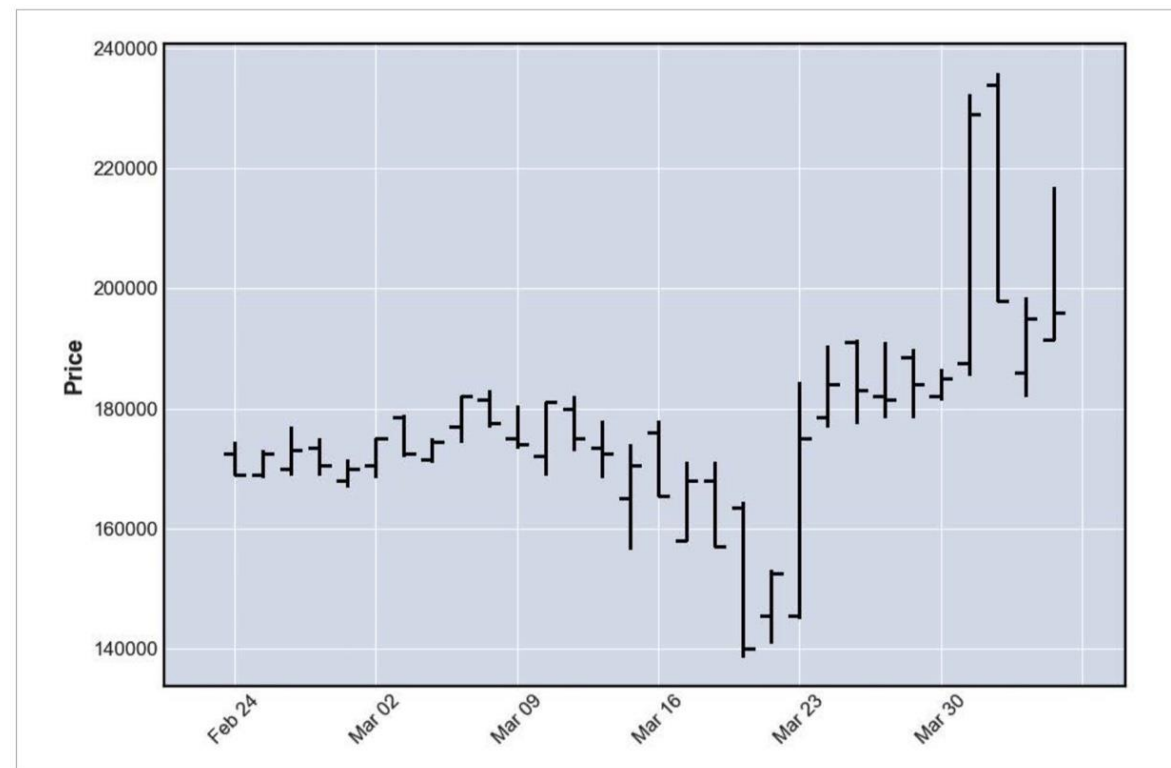
# Jupyter notebook 참고

- 신버전으로 캔들 차트 그리기

실행 결과 #1\_ 셀트리온 기본형 캔들 차트



실행 결과 #2\_ 셀트리온 기본형 OHLC 차트



# Jupyter notebook 참고

- 최종 캔들차트

실행 결과 #3\_ 셀트리온 캔들 차트

