

pandas

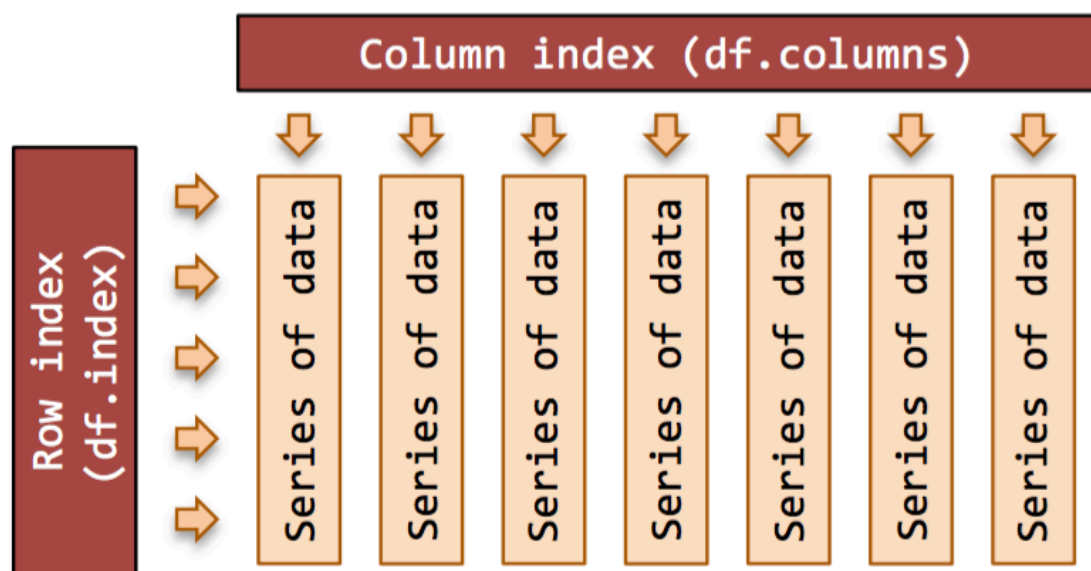
- pandas는 오픈소스이고, BSD 라이선스 라이브러리
- 고성능의 사용이 쉬운 데이터구조와 python 프로그래밍언어를 위한 데이터분석 도구
- **pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language**
- 2008년, AQR 투자 운용 회사(Capital Management)에 다니던 Wes McKinney가 개발 시작
- 데이터프레임을 제공하여 테이블 구조의 데이터 처리를 편리하게 해준다. R의 data.frame 구조체와 같은 용도로 사용된다.
- numpy의 배열에는 모두 숫자만 들어올 수 있으나, 데이터프레임에는 임의의 타입의 데이터를 담을 수 있다

import and set

```
In [1]: import numpy as np
import pandas as pd
from pandas import DataFrame, Series
```

Data Structure

DataFrame의 구성 - 자료화된 구조



```
In [2]: import pandas as pd
from IPython.display import display

# create a simple dataset of people
customer = {'Name': ["John", "Anna", "Peter", "Linda"],
            'Location': ["New York", "Paris", "Berlin", "London"],
            'Age': [24, 13, 53, 33]}
customer
```

```
Out[2]: {'Name': ['John', 'Anna', 'Peter', 'Linda'],
'Location': ['New York', 'Paris', 'Berlin', 'London'],
'Age': [24, 13, 53, 33]}
```

```
In [3]: a = pd.DataFrame(customer)
# IPython.display allows "pretty printing" of dataframes
# in the Jupyter notebook
a
```

```
Out[3]:
```

	Name	Location	Age
0	John	New York	24
1	Anna	Paris	13
2	Peter	Berlin	53
3	Linda	London	33

```
In [4]: # 조건에 맞는 데이터를 얻는 방법 예시로 나이가 30 이상인 경우만 출력한다
display(a[a.Age > 30])
```

	Name	Location	Age
2	Peter	Berlin	53
3	Linda	London	33

Series

생성

- Series는 일련의 객체를 담을 수 있는 1차원 배열 자료구조
- 색인(index)이라고 하는 배열의 데이터에 연관된 이름을 가지고 있음
- 가장 간단한 Series객체는 배열 데이터로부터 생성 할수 있음.

첫 번째 컬럼	두 번째 컬럼
Index Label	Series 객체의 Value

```
In [5]: from pandas import Series, DataFrame
```

```
In [6]: import pandas as pd
```

```
In [7]: pd.Series
```

```
Out[7]: pandas.core.series.Series
def __init__(data=None, index=None, dtype: Dtype | None=None, name=None, copy: bool | None=None, fastpath: bool | lib.NoDefault=lib.no_default) -> None

One-dimensional ndarray with axis labels (including time series).

Labels need not be unique but must be a hashable type. The object
supports both integer- and label-based indexing and provides a host of
methods for performing operations involving the index. Statistical
methods from ndarray have been overridden to automatically exclude
```

```
In [8]: a = Series([4,7,-5,3])
```

```
In [9]: a
```

```
Out[9]:
```

0	4
1	7
2	-5
3	3

dtype: int64

```
In [10]: a.values
```

```
Out[10]: array([ 4,  7, -5,  3])
```

```
In [11]: s = Series([10000, 2000, 30000, 40000], index=['Seattle', 'Seoul', 'San Hose', 'Beijing'])
s
```

```
Out[11]:
```

	0
Seattle	10000
Seoul	2000
San Hose	30000
Beijing	40000

dtype: int64

```
In [12]: s.index
```

```
Out[12]: Index(['Seattle', 'Seoul', 'San Hose', 'Beijing'], dtype='object')
```

```
In [13]: s.values
```

```
Out[13]: array([10000,  2000, 30000, 40000])
```

```
In [14]: s
```

```
Out[14]:
```

	0
Seattle	10000
Seoul	2000
San Hose	30000
Beijing	40000

dtype: int64

```
In [15]: #배열에서 값을 선택하거나 대입할 때는 색인을 이용해서 접근
s['Seattle']
```

```
Out[15]: np.int64(10000)
```

```
In [16]: s['Beijing']
```

```
Out[16]: np.int64(40000)
```

- 불리언 배열을 사용해서 값을 걸러내기

```
In [17]: s[s>10000]
```

```
Out[17]:
```

	0
San Hose	30000
Beijing	40000

dtype: int64

- 산술곱셈을 수행하거나 수학함수를 적용하는 등 Numpy배열연산을 수행해도 색인-값 연결은 유지됨

```
In [18]: s*2
```

```
Out[18]:
```

	0
Seattle	20000
Seoul	4000
San Hose	60000
Beijing	80000

dtype: int64

```
In [19]: 'Beijing' in s
```

```
Out[19]: True
```

```
In [20]: 'Japan' in s
```

```
Out[20]: False
```

2) indexing and slicing

```
In [21]: s['Seoul']
```

```
Out[21]: np.int64(2000)
```

```
In [22]: s[['Seattle', 'San Hose']]
```

```
Out[22]:
```

	0
Seattle	10000
San Hose	30000

dtype: int64

```
In [23]: s[['Seattle', 'San Hose']]
```

```
Out[23]:
```

	0
Seattle	10000
San Hose	30000

dtype: int64

```
In [24]: s['Seattle': 'San Hose']
```

```
Out[24]:
```

	0
Seattle	10000
Seoul	2000
San Hose	30000

dtype: int64

```
In [25]: sdata = {'Ohio': 35000, 'Texas': 71000, 'Oregon': 16000, 'Utah': 5000}
obj3 = pd.Series(sdata)
obj3
```

```
Out[25]:
```

	0
Ohio	35000
Texas	71000
Oregon	16000
Utah	5000

dtype: int64

```
In [26]: states = ['California', 'Ohio', 'Oregon', 'Texas']
obj4 = pd.Series(sdata, index=states)
obj4
```

```
Out[26]:
```

	0
California	NaN
Ohio	35000.0
Oregon	16000.0
Texas	71000.0

dtype: float64

```
In [27]: pd.isnull(obj4)
```

```
Out[27]:
```

	0
California	True
Ohio	False
Oregon	False
Texas	False

dtype: bool

```
In [28]: pd.notnull(obj4)
```

```
Out[28]:
```

	0
California	False
Ohio	True
Oregon	True
Texas	True

dtype: bool

```
In [29]: obj4.isnull()
```

```
Out[29]:
```

	0
California	True
Ohio	False
Oregon	False
Texas	False

dtype: bool

```
In [30]: obj3
```

```
Out[30]:
```

	0
Ohio	35000
Texas	71000
Oregon	16000
Utah	5000

dtype: int64

```
In [31]: print(obj4)
obj3 + obj4
```

California	NaN
Ohio	35000.0
Oregon	16000.0
Texas	71000.0

dtype: float64

```
Out[31]:
```

	0
California	NaN
Ohio	70000.0
Oregon	32000.0
Texas	142000.0
Utah	NaN

dtype: float64

3) Time Series

```
In [32]: import pandas as pd
```

```
In [33]: dates = pd.date_range('2016-05-01', '2016-05-07')
dates
```

```
Out[33]: DatetimeIndex(['2016-05-01', '2016-05-02', '2016-05-03', '2016-05-04',
                        '2016-05-05', '2016-05-06', '2016-05-07'],
                        dtype='datetime64[ns]', freq='D')
```

DataFrame

생성

```
In [34]: tmp1 = Series([80, 92, 82, 85, 97, 84, 78], index=dates)
```

```
In [35]: tmp1
```

```
Out[35]:
```

	0
2016-05-01	80
2016-05-02	92
2016-05-03	82
2016-05-04	85
2016-05-05	97
2016-05-06	84
2016-05-07	78

dtype: int64

```
In [36]: tmp2 = Series(np.random.randint(60, 100, size=7), index=dates)
```

In [37]:

```
tmp1
```

Out[37]:

	0
2016-05-01	80
2016-05-02	92
2016-05-03	82
2016-05-04	85
2016-05-05	97
2016-05-06	84
2016-05-07	78

dtype: int64

In [38]:

```
tmp2
```

Out[38]:

	0
2016-05-01	67
2016-05-02	78
2016-05-03	67
2016-05-04	64
2016-05-05	67
2016-05-06	85
2016-05-07	60

dtype: int64

In [39]:

```
exam = DataFrame({
    'Math': tmp1,
    'Philosophy': tmp2
})
exam
```

Out[39]:

	Math	Philosophy
2016-05-01	80	67
2016-05-02	92	78
2016-05-03	82	67
2016-05-04	85	64
2016-05-05	97	67
2016-05-06	84	85
2016-05-07	78	60

데이터 살펴보기

In [40]:

```
exam['Math']
```

```
Out[40]:
```

	Math
2016-05-01	80
2016-05-02	92
2016-05-03	82
2016-05-04	85
2016-05-05	97
2016-05-06	84
2016-05-07	78

dtype: int64

```
In [41]: exam.loc['2016-05-06']
```

```
Out[41]:
```

	2016-05-06
Math	84
Philosophy	85

dtype: int64

```
In [42]: exam.iloc[0]
```

```
Out[42]:
```

	2016-05-01
Math	80
Philosophy	67

dtype: int64

```
In [43]: exam[['Math', 'Philosophy']]
```

```
Out[43]:
```

	Math	Philosophy
2016-05-01	80	67
2016-05-02	92	78
2016-05-03	82	67
2016-05-04	85	64
2016-05-05	97	67
2016-05-06	84	85
2016-05-07	78	60

```
In [44]: exam['Math'][[1, 3]]
```

<ipython-input-44-335f50bf431e>:1: FutureWarning: Series.__getitem__ treating keys as positions is deprecated. In a future version, integer keys will always be treated as labels (consistent with DataFrame behavior). To access a value by position, use `ser.iloc[pos]`
 exam['Math'][[1, 3]]

```
Out[44]:
```

	Math
2016-05-02	92
2016-05-04	85

dtype: int64

```
In [45]: print('a')
```

a


```
In [46]: exam.Math
```

```
Out[46]:
```

	Math
2016-05-01	80
2016-05-02	92
2016-05-03	82
2016-05-04	85
2016-05-05	97
2016-05-06	84
2016-05-07	78

dtype: int64

```
In [47]: exam.columns
```

```
Out[47]: Index(['Math', 'Philosophy'], dtype='object')
```

```
In [48]: exam
```

```
Out[48]:
```

	Math	Philosophy
2016-05-01	80	67
2016-05-02	92	78
2016-05-03	82	67
2016-05-04	85	64
2016-05-05	97	67
2016-05-06	84	85
2016-05-07	78	60

연산

```
In [49]: a = exam.Math - exam.Philosophy
```

```
In [50]: a
```

```
Out[50]:
```

	0
2016-05-01	13
2016-05-02	14
2016-05-03	15
2016-05-04	21
2016-05-05	30
2016-05-06	-1
2016-05-07	18

dtype: int64

```
In [51]: exam['difference'] = a  
exam
```

```
Out[51]:
```

	Math	Philosophy	difference
2016-05-01	80	67	13
2016-05-02	92	78	14
2016-05-03	82	67	15
2016-05-04	85	64	21
2016-05-05	97	67	30
2016-05-06	84	85	-1
2016-05-07	78	60	18

```
In [52]: avg = np.mean(exam, axis=1) # axis: 1=row
avg
```

```
Out[52]:
```

	0
2016-05-01	53.333333
2016-05-02	61.333333
2016-05-03	54.666667
2016-05-04	56.666667
2016-05-05	64.666667
2016-05-06	56.000000
2016-05-07	52.000000

dtype: float64

컬럼 추가

```
In [53]: exam['avg'] = avg
exam
```

```
Out[53]:
```

	Math	Philosophy	difference	avg
2016-05-01	80	67	13	53.333333
2016-05-02	92	78	14	61.333333
2016-05-03	82	67	15	54.666667
2016-05-04	85	64	21	56.666667
2016-05-05	97	67	30	64.666667
2016-05-06	84	85	-1	56.000000
2016-05-07	78	60	18	52.000000

컬럼 이름 바꾸기

```
In [54]: # rename two of the columns by using the 'rename' method
exam.rename(columns={'Math': 'Mathmatics', 'avg': 'Average'}, inplace=True)
exam.columns
```

```
Out[54]: Index(['Mathmatics', 'Philosophy', 'difference', 'Average'], dtype='object')
```

```
In [55]: exam
```

```
Out[55]:
```

	Mathmatics	Philosophy	difference	Average
2016-05-01	80	67	13	53.333333
2016-05-02	92	78	14	61.333333
2016-05-03	82	67	15	54.666667
2016-05-04	85	64	21	56.666667
2016-05-05	97	67	30	64.666667
2016-05-06	84	85	-1	56.000000
2016-05-07	78	60	18	52.000000

컬럼 지우기

```
In [56]: # remove a single column (axis=1 refers to columns)
exam.drop('difference', axis=1, inplace=True)
exam.head()
```

```
Out[56]:
```

	Mathmatics	Philosophy	Average
2016-05-01	80	67	53.333333
2016-05-02	92	78	61.333333
2016-05-03	82	67	54.666667
2016-05-04	85	64	56.666667
2016-05-05	97	67	64.666667

데이터 찾기

```
In [57]: exam.loc['2016-05-01']
```

```
Out[57]:
```

2016-05-01
Mathmatics 80.000000
Philosophy 67.000000
Average 53.333333

dtype: float64

```
In [58]: exam.iloc[0] # location
```

```
Out[58]:
```

2016-05-01
Mathmatics 80.000000
Philosophy 67.000000
Average 53.333333

dtype: float64

```
In [59]: exam.index >= '2016-05-03'
```

```
Out[59]: array([False, False,  True,  True,  True,  True,  True])
```

```
In [60]: exam[
    (exam.index == '2016-05-03') | (exam.index == '2016-05-04')
]
```

```
Out[60]:
```

	Mathmatics	Philosophy	Average
2016-05-03	82	67	54.666667
2016-05-04	85	64	56.666667

Q. 2016-05-03부터 2016-05-05까지 데이터를 가져와보자.

```
In [61]: # Solution
```

```
In [64]: %time
```

CPU times: user 2 μ s, sys: 0 ns, total: 2 μ s
Wall time: 3.81 μ s

```
In [65]: exam.Mathmatics > 90
```

```
Out[65]:
```

	Mathmatics
2016-05-01	False
2016-05-02	True
2016-05-03	False
2016-05-04	False
2016-05-05	True
2016-05-06	False
2016-05-07	False

dtype: bool

```
In [66]: exam[exam.Mathmatics > 90]
```

```
Out[66]:
```

	Mathmatics	Philosophy	Average
2016-05-02	92	78	61.333333
2016-05-05	97	67	64.666667

```
In [67]: exam[exam.index < '2016-05-05']
```

```
Out[67]:
```

	Mathmatics	Philosophy	Average
2016-05-01	80	67	53.333333
2016-05-02	92	78	61.333333
2016-05-03	82	67	54.666667
2016-05-04	85	64	56.666667

```
In [68]: exam[(exam.index < '2016-05-05') & (exam.Mathmatics > 90)]
```

```
Out[68]:
```

	Mathmatics	Philosophy	Average
2016-05-02	92	78	61.333333

그래프로 그려보기

```
In [69]: %matplotlib inline
```

```
In [73]: import matplotlib.pyplot as plt
from matplotlib.pylab import rcParams
rcParams['font.family'] = 'AppleGothic' #Mac
#rcParams['font.family'] = 'NanumGothic' #Window용 폰트
```

```
In [75]: exam.plot()
```

```
Out[75]: <Axes: >
```

[illegible]

Q. 다른 형태의 그래프는 어떻게 그릴 수 있을까요?

Docstring

In []:

matplotlib

- 그래프용 라이브러리로 히스토그램, 산포도 등을 그리는데 사용된다
- 주피터 노트북에 결과 그래프가 나타나게 하려면 %matplotlib inline 매크로를 실행해야 한다
- (또는 matplotlib.pyplot.show 함수로 그림을 그려야 한다)

영화리뷰데이터 판다스로 필터링하기!

In [76]:

```
# IMDB 영화리뷰데이터셋을 인터넷에서 불러오기
movies = pd.read_csv('http://bit.ly/imdbratings')
movies.head()
```

Out[76]:

	star_rating	title	content_rating	genre	duration	actors_list
0	9.3	The Shawshank Redemption	R	Crime	142	[u'Tim Robbins', u'Morgan Freeman', u'Bob Gunt...
1	9.2	The Godfather	R	Crime	175	[u'Marlon Brando', u'Al Pacino', u'James Caan']
2	9.1	The Godfather: Part II	R	Crime	200	[u'Al Pacino', u'Robert De Niro', u'Robert Duv...
3	9.0	The Dark Knight	PG-13	Action	152	[u'Christian Bale', u'Heath Ledger', u'Aaron E...
4	8.9	Pulp Fiction	R	Crime	154	[u'John Travolta', u'Uma Thurman', u'Samuel L....

In [77]:

```
# 행과 열의 수 확인
movies.shape
```

Out[77]:

(979, 6)

우리의 목표: 영화상영시간이 200분 이상이 영화만 필터링해보자!

In [78]:

```
# 먼저, 영화상영시간이 200분 이상이면 true, 200분이하면 false라고 표시해주는 리스트를 생성해야함
booleans = []
for length in movies.duration:
    if length >= 200:
        booleans.append(True)
    else:
        booleans.append(False)
```

In [79]:

```
# 리스트에 있는 값의 수가 전체 영화수가 일치하는지 확인
len(booleans)
```

Out[79]:

979

In [80]:

```
# 1-5번영화먼저 확인
booleans[0:5]
```

Out[80]:

[False, False, True, False, False]

In [81]:

```
# 데이터프레임을 만들기전에 리스트를 Series로 변환
is_long = pd.Series(booleans)
is_long.head()
```

```
Out[81]: 0
0 False
1 False
2 True
3 False
4 False
```

dtype: bool

```
In [82]: # use bracket notation with the boolean Series to tell the DataFrame which rows to display
movies[is_long]
```

```
Out[82]:
```

	star_rating	title	content_rating	genre	duration	actors_list
2	9.1	The Godfather: Part II	R	Crime	200	[u'Al Pacino', u'Robert De Niro', u'Robert Duv...]
7	8.9	The Lord of the Rings: The Return of the King	PG-13	Adventure	201	[u'Elijah Wood', u'Viggo Mortensen', u'Ian McK...]
17	8.7	Seven Samurai	UNRATED	Drama	207	[u'Toshirxf4 Mifune', u'Takashi Shimura', u'K...]
78	8.4	Once Upon a Time in America	R	Crime	229	[u'Robert De Niro', u'James Woods', u'Elizabet...]
85	8.4	Lawrence of Arabia	PG	Adventure	216	[u"Peter O'Toole", u'Alec Guinness', u'Anthony...]
142	8.3	Lagaan: Once Upon a Time in India	PG	Adventure	224	[u'Aamir Khan', u'Gracy Singh', u'Rachel Shell...]
157	8.2	Gone with the Wind	G	Drama	238	[u'Clark Gable', u'Vivien Leigh', u'Thomas Mit...]
204	8.1	Ben-Hur	G	Adventure	212	[u'Charlton Heston', u'Jack Hawkins', u'Stephe...]
445	7.9	The Ten Commandments	APPROVED	Adventure	220	[u'Charlton Heston', u'Yul Brynner', u'Anne Ba...]
476	7.8	Hamlet	PG-13	Drama	242	[u'Kenneth Branagh', u'Julie Christie', u'Dere...]
630	7.7	Malcolm X	PG-13	Biography	202	[u'Denzel Washington', u'Angela Bassett', u'De...]
767	7.6	It's a Mad, Mad, Mad, Mad World	APPROVED	Action	205	[u'Spencer Tracy', u'Milton Berle', u'Ethel Me...]

```
In [83]: # 위과정을 아래와 같이 간편하게도 가능함
is_long = movies.duration >= 200
movies[is_long]

# is_long을 생성하지 않고 더 간단하게!
movies[movies.duration >= 200]
```


Out[83]:

	star_rating	title	content_rating	genre	duration	actors_list
2	9.1	The Godfather: Part II	R	Crime	200	[u'Al Pacino', u'Robert De Niro', u'Robert Duv...
7	8.9	The Lord of the Rings: The Return of the King	PG-13	Adventure	201	[u'Elijah Wood', u'Viggo Mortensen', u'Ian McK...
17	8.7	Seven Samurai	UNRATED	Drama	207	[u'Toshir�f4 Mifune', u'Takashi Shimura', u'K...
78	8.4	Once Upon a Time in America	R	Crime	229	[u'Robert De Niro', u'James Woods', u'Elizabet...
85	8.4	Lawrence of Arabia	PG	Adventure	216	[u"Peter O'Toole", u'Alec Guinness', u'Anthony...
142	8.3	Lagaan: Once Upon a Time in India	PG	Adventure	224	[u'Aamir Khan', u'Gracy Singh', u'Rachel Shell...
157	8.2	Gone with the Wind	G	Drama	238	[u'Clark Gable', u'Vivien Leigh', u'Thomas Mit...
204	8.1	Ben-Hur	G	Adventure	212	[u'Charlton Heston', u'Jack Hawkins', u'Stephe...
445	7.9	The Ten Commandments	APPROVED	Adventure	220	[u'Charlton Heston', u'Yul Brynner', u'Anne Ba...
476	7.8	Hamlet	PG-13	Drama	242	[u'Kenneth Branagh', u'Julie Christie', u'Dere...
630	7.7	Malcolm X	PG-13	Biography	202	[u'Denzel Washington', u'Angela Bassett', u'De...
767	7.6	It's a Mad, Mad, Mad, Mad World	APPROVED	Action	205	[u'Spencer Tracy', u'Milton Berle', u'Ethel Me...

In [84]:

```
# Seriese에서 장르만 선택하기
movies[movies.duration >= 200].genre

# loc내장함수를 이용하는 방법
movies.loc[movies.duration >= 200, 'genre']
```

Out[84]:

	genre
2	Crime
7	Adventure
17	Drama
78	Crime
85	Adventure
142	Adventure
157	Drama
204	Adventure
445	Adventure
476	Drama
630	Biography
767	Action

dtype: object

Rules for specifying **multiple filter criteria** in pandas:

- use `&` instead of `and`
- use `|` instead of `or`
- add **parentheses** around each condition to specify evaluation order

다음목표: 영화의 장르가 드라마이면서 상영시간이 200분이상인 영화를 필터링하자!

In [85]:

```
# CORRECT: use the '&' operator to specify that both conditions are required
movies[(movies.duration >=200) & (movies.genre == 'Drama')]
```

Out[85]:	star_rating	title	content_rating	genre	duration	actors_list
17	8.7	Seven Samurai	UNRATED	Drama	207	[u'Toshiro Mifune', u'Takashi Shimura', u'K...
157	8.2	Gone with the Wind	G	Drama	238	[u'Clark Gable', u'Vivien Leigh', u'Thomas Mit...
476	7.8	Hamlet	PG-13	Drama	242	[u'Kenneth Branagh', u'Julie Christie', u'Dere...

```
In [86]: # INCORRECT: using the '|' operator would have shown movies that are either long or dramas (or both)
movies[(movies.duration >=200) | (movies.genre == 'Drama')].head()
```

Out[86]:	star_rating	title	content_rating	genre	duration	actors_list
2	9.1	The Godfather: Part II	R	Crime	200	[u'Al Pacino', u'Robert De Niro', u'Robert Duv...
5	8.9	12 Angry Men	NOT RATED	Drama	96	[u'Henry Fonda', u'Lee J. Cobb', u'Martin Bals...
7	8.9	The Lord of the Rings: The Return of the King	PG-13	Adventure	201	[u'Elijah Wood', u'Viggo Mortensen', u'Ian McK...
9	8.9	Fight Club	R	Drama	139	[u'Brad Pitt', u'Edward Norton', u'Helena Bonh...
13	8.8	Forrest Gump	PG-13	Drama	142	[u'Tom Hanks', u'Robin Wright', u'Gary Sinise']

목표: 이번에는 장르가 'Crime' 또는 'Drama' 또는 'Action'인 영화리스트 보기!

```
In [87]: # use the '|' operator to specify that a row can match any of the three criteria
movies[(movies.genre == 'Crime') | (movies.genre == 'Drama') | (movies.genre == 'Action')].head(10)

# or equivalently, use the 'isin' method
movies[movies.genre.isin(['Crime', 'Drama', 'Action'])].head(10)
```

Out[87]:	star_rating	title	content_rating	genre	duration	actors_list
0	9.3	The Shawshank Redemption	R	Crime	142	[u'Tim Robbins', u'Morgan Freeman', u'Bob Gunt...
1	9.2	The Godfather	R	Crime	175	[u'Marlon Brando', u'Al Pacino', u'James Caan']
2	9.1	The Godfather: Part II	R	Crime	200	[u'Al Pacino', u'Robert De Niro', u'Robert Duv...
3	9.0	The Dark Knight	PG-13	Action	152	[u'Christian Bale', u'Heath Ledger', u'Aaron E...
4	8.9	Pulp Fiction	R	Crime	154	[u'John Travolta', u'Uma Thurman', u'Samuel L....
5	8.9	12 Angry Men	NOT RATED	Drama	96	[u'Henry Fonda', u'Lee J. Cobb', u'Martin Bals...
9	8.9	Fight Club	R	Drama	139	[u'Brad Pitt', u'Edward Norton', u'Helena Bonh...
11	8.8	Inception	PG-13	Action	148	[u'Leonardo DiCaprio', u'Joseph Gordon-Levitt'...
12	8.8	Star Wars: Episode V - The Empire Strikes Back	PG	Action	124	[u'Mark Hamill', u'Harrison Ford', u'Carrie Fi...
13	8.8	Forrest Gump	PG-13	Drama	142	[u'Tom Hanks', u'Robin Wright', u'Gary Sinise']

Documentation for `isin`

판다스로 그룹별 분석하기

```
In [88]: # 각나라별 음주현황데이터 불러오기
drinks = pd.read_csv('http://bit.ly/drinksbycountry')
drinks.head()
```

```
Out[88]:
```

	country	beer_servings	spirit_servings	wine_servings	total_litres_of_pure_alcohol	continent
0	Afghanistan	0	0	0	0.0	Asia
1	Albania	89	132	54	4.9	Europe
2	Algeria	25	0	14	0.7	Africa
3	Andorra	245	138	312	12.4	Europe
4	Angola	217	57	45	5.9	Africa

```
In [89]: # 전체 데이터의 맥주 평균 소비량
drinks.beer_servings.mean()
```

```
Out[89]: np.float64(106.16062176165804)
```

```
In [90]: # 아프리카 대륙의 평균 소비량
drinks[drinks.continent=='Africa'].beer_servings.mean()
```

```
Out[90]: np.float64(61.471698113207545)
```

```
In [91]: # 각 대륙별 평균 소비량
drinks.groupby('continent').beer_servings.mean()
```

```
Out[91]:
```

beer_servings	
continent	
Africa	61.471698
Asia	37.045455
Europe	193.777778
North America	145.434783
Oceania	89.687500
South America	175.083333

dtype: float64

```
In [92]: # 각 대륙별 최고 소비량
drinks.groupby('continent').beer_servings.max()
```

```
Out[92]:
```

beer_servings	
continent	
Africa	376
Asia	247
Europe	361
North America	285
Oceania	306
South America	333

dtype: int64

```
In [93]: # 다양한 통계치를 한꺼번에!
drinks.groupby('continent').beer_servings.agg(['count', 'mean', 'min', 'max'])
```

Out[93]:

	count	mean	min	max
continent				
Africa	53	61.471698	0	376
Asia	44	37.045455	0	247
Europe	45	193.777778	0	361
North America	23	145.434783	1	285
Oceania	16	89.687500	0	306
South America	12	175.083333	93	333

In [94]:

```
# 주피터 노트북에 그래프를 보이게 하려면 필수!  
%matplotlib inline
```

In [95]:

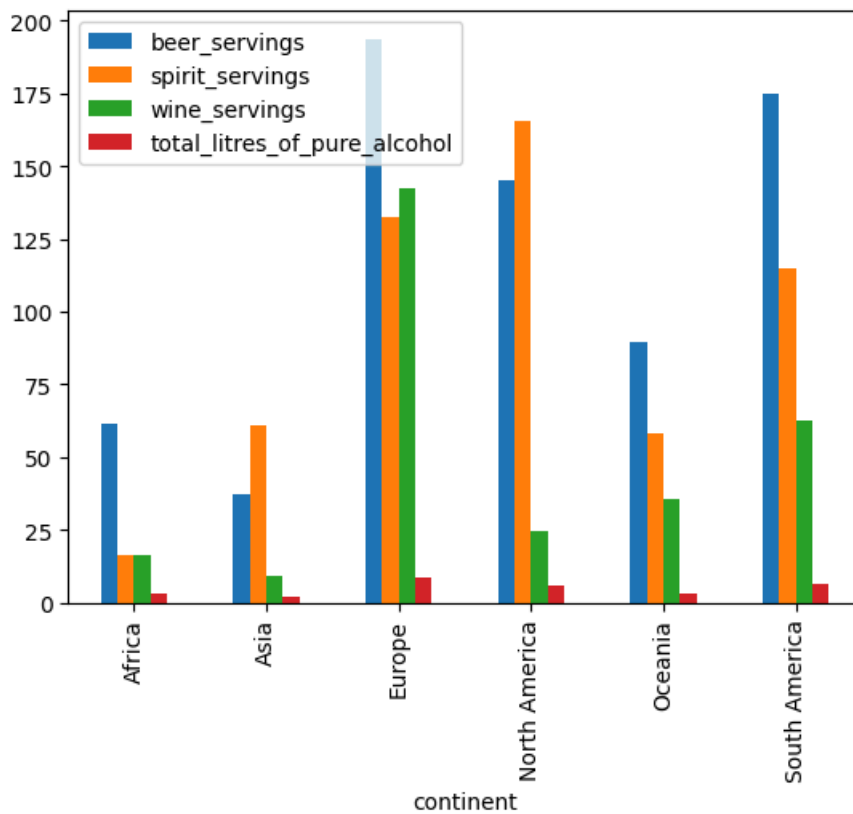
```
# 그래프를 그려보자!  
numeric_columns = drinks.select_dtypes(include='number').columns  
drinks.groupby('continent')[numeric_columns].mean().plot(kind='bar')
```

Out[95]:

```
<Axes: xlabel='continent'>
```

[illegible]

[illegible]



주식데이터 분석!

```
In [96]: # from 패키지.모듈 import 함수
from numpy.random import randint
```

```
In [97]: randint(1000, 2000, size=(7, 4))
```

```
Out[97]: array([[1183, 1164, 1757, 1445],
 [1732, 1323, 1714, 1195],
 [1180, 1524, 1356, 1169],
 [1975, 1175, 1106, 1767],
 [1441, 1955, 1597, 1479],
 [1654, 1132, 1872, 1662],
 [1013, 1960, 1505, 1163]])
```

```
In [98]: # pandas라는 패키지로 dummy data를 만드는 부분
import numpy as np
import pandas as pd

dates = pd.date_range('20180501', periods=7)
stocks = ['송원산업', '하림', '한라', '티엘아이']
values = np.random.randint(1000, 2000, size=(7, 4))

df = pd.DataFrame(values, index=dates, columns=stocks)
df
```

```
Out[98]:
```

	송원산업	하림	한라	티엘아이
2018-05-01	1733	1456	1158	1510
2018-05-02	1897	1630	1981	1043
2018-05-03	1000	1023	1347	1685
2018-05-04	1884	1761	1196	1967
2018-05-05	1412	1237	1188	1045
2018-05-06	1345	1291	1527	1555
2018-05-07	1145	1089	1482	1234

```
In [99]: # 여기서부터 작성해보세요.
df['티엘아이'][3]
```

```
<ipython-input-99-3dadef1a0e8>:2: FutureWarning: Series.__getitem__ treating keys as positions is deprecated. In a future version, integer keys will always be treated as labels (consistent with DataFrame behavior). To access a value by position, use `ser.iloc[pos]`  
df['티엘아이'][3]
```

Out[99]: np.int64(1967)

```
In [100]: df[df['티엘아이'] == 1523]
```

Out[100]: 송원산업 하림 한라 티엘아이

```
In [101]: %matplotlib inline
```

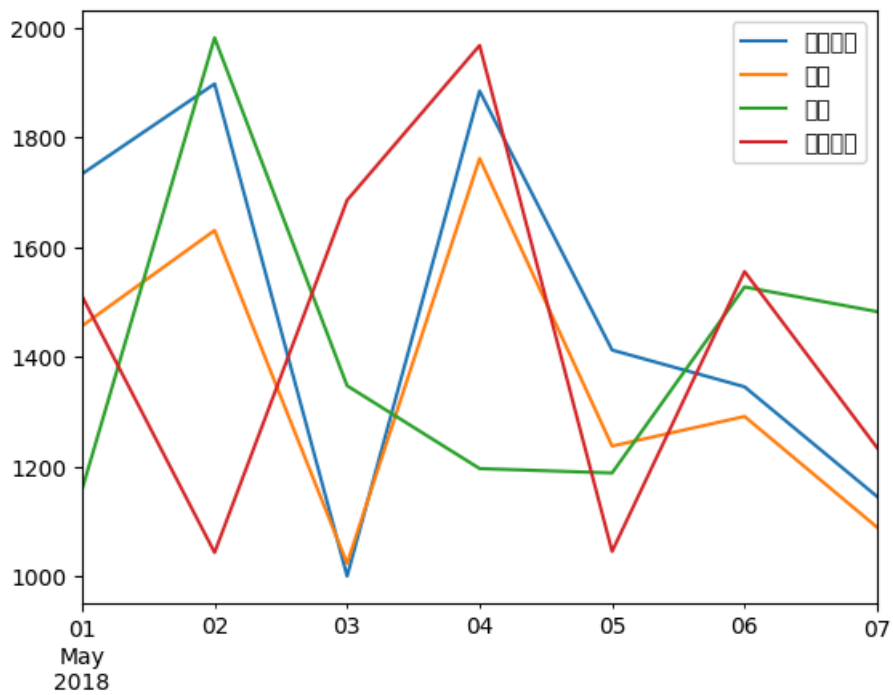
```
In [102]: df.plot()
```

Out[102]: <Axes: >

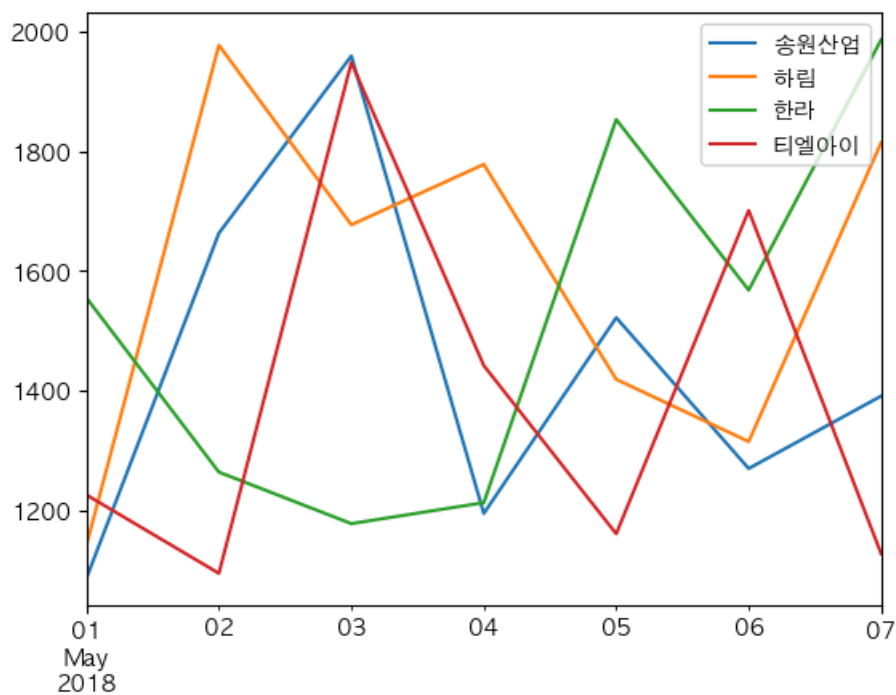
[illegible]

[illegible]

[illegible]



In []: `df.plot();`



Q 차트 사이즈가 너무 작아요! 어떻게 해볼까요?

In []: `# 괄호 안에 파라미터값을 넣어주면 간단하게 해결됩니다. 어떤 값을 넣어야할까요? 찾아봅시다.`

Q. 애플 주식의 거래량 증가값/증가율을 확인해보고 만약 40%이상 증가 했으면 시그널을 주는 프로그래밍을 해보자(for/if문 이용)

prerequisite

1. pandas
2. pandas-datareader
3. fix-yahoo-finance

In [103... `!pip install yfinance`

Requirement already satisfied: yfinance in /usr/local/lib/python3.11/dist-packages (0.2.55)
Requirement already satisfied: pandas>=1.3.0 in /usr/local/lib/python3.11/dist-packages (from yfinance) (2.2.2)
Requirement already satisfied: numpy>=1.16.5 in /usr/local/lib/python3.11/dist-packages (from yfinance) (2.0.2)
Requirement already satisfied: requests>=2.31 in /usr/local/lib/python3.11/dist-packages (from yfinance) (2.32.3)
Requirement already satisfied: multitasking>=0.0.7 in /usr/local/lib/python3.11/dist-packages (from yfinance) (0.0.11)
Requirement already satisfied: platformdirs>=2.0.0 in /usr/local/lib/python3.11/dist-packages (from yfinance) (4.3.7)
Requirement already satisfied: pytz>=2022.5 in /usr/local/lib/python3.11/dist-packages (from yfinance) (2025.2)
Requirement already satisfied: frozendict>=2.3.4 in /usr/local/lib/python3.11/dist-packages (from yfinance) (2.4.6)
Requirement already satisfied: peewee>=3.16.2 in /usr/local/lib/python3.11/dist-packages (from yfinance) (3.17.9)
Requirement already satisfied: beautifulsoup4>=4.11.1 in /usr/local/lib/python3.11/dist-packages (from yfinance) (4.13.3)
Requirement already satisfied: soupsieve>1.2 in /usr/local/lib/python3.11/dist-packages (from beautifulsoup4>=4.11.1->yfinance) (2.6)
Requirement already satisfied: typing-extensions>=4.0.0 in /usr/local/lib/python3.11/dist-packages (from beautifulsoup4>=4.11.1->yfinance) (4.13.1)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.11/dist-packages (from pandas>=1.3.0->yfinance) (2.8.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-packages (from pandas>=1.3.0->yfinance) (2025.2)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests>=2.31->yfinance) (3.4.1)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-packages (from requests>=2.31->yfinance) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests>=2.31->yfinance) (2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests>=2.31->yfinance) (2025.1.31)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.8.2->pandas>=1.3.0->yfinance) (1.17.0)

In [104... `import yfinance`

In [105... `pip install yfinance pandas_datareader`

Requirement already satisfied: yfinance in /usr/local/lib/python3.11/dist-packages (0.2.55)
 Requirement already satisfied: pandas_datareader in /usr/local/lib/python3.11/dist-packages (0.10.0)
 Requirement already satisfied: pandas>=1.3.0 in /usr/local/lib/python3.11/dist-packages (from yfinance) (2.2.2)
 Requirement already satisfied: numpy>=1.16.5 in /usr/local/lib/python3.11/dist-packages (from yfinance) (2.0.2)
 Requirement already satisfied: requests>=2.31 in /usr/local/lib/python3.11/dist-packages (from yfinance) (2.32.3)
 Requirement already satisfied: multitasking>=0.0.7 in /usr/local/lib/python3.11/dist-packages (from yfinance) (0.0.11)
 Requirement already satisfied: platformdirs>=2.0.0 in /usr/local/lib/python3.11/dist-packages (from yfinance) (4.3.7)
 Requirement already satisfied: pytz>=2022.5 in /usr/local/lib/python3.11/dist-packages (from yfinance) (2025.2)
 Requirement already satisfied: frozendict>=2.3.4 in /usr/local/lib/python3.11/dist-packages (from yfinance) (2.4.6)
 Requirement already satisfied: peewee>=3.16.2 in /usr/local/lib/python3.11/dist-packages (from yfinance) (3.17.9)
 Requirement already satisfied: beautifulsoup4>=4.11.1 in /usr/local/lib/python3.11/dist-packages (from yfinance) (4.13.3)
 Requirement already satisfied: lxml in /usr/local/lib/python3.11/dist-packages (from pandas_datareader) (5.3.2)
 Requirement already satisfied: soupsieve>1.2 in /usr/local/lib/python3.11/dist-packages (from beautifulsoup4>=4.11.1->yfinance) (2.6)
 Requirement already satisfied: typing-extensions>=4.0.0 in /usr/local/lib/python3.11/dist-packages (from beautifulsoup4>=4.11.1->yfinance) (4.13.1)
 Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.11/dist-packages (from pandas>=1.3.0->yfinance) (2.8.2)
 Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-packages (from pandas>=1.3.0->yfinance) (2025.2)
 Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests>=2.31->yfinance) (3.4.1)
 Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-packages (from requests>=2.31->yfinance) (3.10)
 Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests>=2.31->yfinance) (2.3.0)
 Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests>=2.31->yfinance) (2025.1.31)
 Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.8.2->pandas>=1.3.0->yfinance) (1.17.0)

In [133...

```
from pandas_datareader import data
import yfinance as yf

import datetime
```

Q. 먼저 설치해야할 패키지가 위에 있습니다. 각자 설치해보도록 합시다!

결과물

- 새로운 컬럼이 추가됨(diff): 전일 volume과의 증감량
- 새로운 컬럼이 추가됨(pct): 전일 volume과의 증감율
- 새로운 컬럼이 추가됨(volume_signal): 증감율이 40이상인 경우 'Over 40% UP' 출력

In [141...

```
IBM = yf.download('IBM', start='2020-01-01')
IBM
```

[*****100%*****] 1 of 1 completed

```
Out[141]:
```

	Price	Close	High	Low	Open	Volume
Ticker	IBM	IBM	IBM	IBM	IBM	IBM
Date						
2020-01-02	101.918869	102.295176	101.429671	101.602767	3293436	
2020-01-03	101.106041	101.497399	100.518997	100.526523	2482890	
2020-01-06	100.925407	101.030767	100.248050	100.413627	2537073	
2020-01-07	100.993149	101.572667	100.398592	100.616848	3232977	
2020-01-08	101.836067	102.250008	100.789935	101.233978	4545916	
...
2025-04-08	221.029999	233.050003	217.279999	232.559998	6850000	
2025-04-09	235.309998	236.300003	215.160004	217.119995	7302800	
2025-04-10	229.550003	232.570007	222.020004	231.000000	5656100	
2025-04-11	235.479996	237.580002	227.509995	229.720001	4324800	
2025-04-14	237.550003	241.769897	236.729996	239.770004	1359961	

1328 rows × 5 columns

```
In [142... diff = IBM['Volume'].diff()
```

```
In [143... IBM['Difference'] = diff
#print(type(IBM.Difference)) # series인지 확인
```

```
In [144... # yfinance 데이터 형식이 바뀌었는지 Ticker 가 들어가게 되면서
# MultiIndex가 되어서 Volume 값 불러올 때, 'IBM' 까지 같이 작성해야해요.
# 같이 작성해야 IBM.Difference 랑 IBM.Volume~ 이 전부 Series로 읽혀서 계산이 가능해집니다!

volume = IBM[('Volume', 'IBM')]
#print(type(volume)) # series인지 확인
```

```
In [145... pct = (IBM['Difference'] / volume) * 100
#print(type(pct)) #series인지 확인
```

```
In [146... IBM['Percent'] = pct.round(2)
```

```
In [147... IBM
```

```
Out[147]:
```

	Price	Close	High	Low	Open	Volume	Difference	Percent
Ticker	IBM	IBM	IBM	IBM	IBM	IBM		
Date								
2020-01-02	101.918869	102.295176	101.429671	101.602767	3293436		NaN	NaN
2020-01-03	101.106041	101.497399	100.518997	100.526523	2482890		-810546.0	-32.65
2020-01-06	100.925407	101.030767	100.248050	100.413627	2537073		54183.0	2.14
2020-01-07	100.993149	101.572667	100.398592	100.616848	3232977		695904.0	21.53
2020-01-08	101.836067	102.250008	100.789935	101.233978	4545916		1312939.0	28.88
...
2025-04-08	221.029999	233.050003	217.279999	232.559998	6850000		-947900.0	-13.84
2025-04-09	235.309998	236.300003	215.160004	217.119995	7302800		452800.0	6.20
2025-04-10	229.550003	232.570007	222.020004	231.000000	5656100		-1646700.0	-29.11
2025-04-11	235.479996	237.580002	227.509995	229.720001	4324800		-1331300.0	-30.78
2025-04-14	237.550003	241.769897	236.729996	239.770004	1359961		-2964839.0	-218.01

1328 rows × 7 columns

In [211...

```
import datetime

from pandas_datareader import data # module
import yfinance as yf
```

Q. (Jupyter Notebook 실습) 이 셀 아래에 새로운 셀을 추가해서 `fix_yahoo_finance` 라는 패키지를 왜 설치하는지 적어보세요! 마크다운으로!

In [212...

```
lge = yf.download('005930.KS', start='2018-07-01')
lge.head()
```

[*****100%*****] 1 of 1 completed

Out[212]:

	Price	Close	High	Low	Open	Volume
Ticker	005930.KS	005930.KS	005930.KS	005930.KS	005930.KS	005930.KS
Date						
2018-07-02	38217.359375	39559.791318	38175.408377	39014.428341	13112253	
2018-07-03	38720.777344	38972.483372	38385.169306	38385.169306	10959655	
2018-07-04	38804.671875	39475.887821	38636.867889	39182.230845	8776763	
2018-07-05	38552.964844	39056.376789	38259.307875	38678.817830	7039773	
2018-07-06	37672.003906	38469.073031	37462.248873	38175.415985	17843706	

In [213...

```
len(lge)
```

Out[213]: 1666

In [214...

```
start = datetime.datetime(2018, 1, 1)
```

In [215...

```
type(start)
```

Out[215]: datetime.datetime

In [216...

```
lge = yf.download("066570.KS", start=start) # 반환값은 pandas의 DataFrame 객체
lge.head()
```

[*****100%*****] 1 of 1 completed

Out[216]:

	Price	Close	High	Low	Open	Volume
Ticker	066570.KS	066570.KS	066570.KS	066570.KS	066570.KS	066570.KS
Date						
2018-01-02	102588.343750	103993.663527	98840.824344	99777.704195	948226	
2018-01-03	102588.343750	103525.223602	100714.584047	102588.343750	627983	
2018-01-04	99309.273438	103993.673128	98372.393499	103525.233159	977691	
2018-01-05	103993.664062	104462.103991	99309.264780	99309.264780	883832	
2018-01-08	98372.390625	106804.309821	96967.070759	106335.869866	2480029	

In [217...

```
lge['Volume'].head() # head(): pandas의 DataFrame 객체가 갖고 있는 기능
```


Out[217]: Ticker 066570.KS

Date	
2018-01-02	948226
2018-01-03	627983
2018-01-04	977691
2018-01-05	883832
2018-01-08	2480029

In [218... type(lge['Volume'])

Out[218]: pandas.core.frame.DataFrame

```
def __init__(data=None, index: Axes | None=None, columns: Axes | None=None, dtype: Dtype | None=None, copy: bool | None=None) -> None
```

Two-dimensional, size-mutable, potentially heterogeneous tabular data.

Data structure also contains labeled axes (rows and columns). Arithmetic operations align on both row and column labels. Can be thought of as a dict-like container for Series objects. The primary pandas data structure.

In [219... lge['Volume'].diff().head() # diff(): pandas의 DataFrame 객체가 갖고 있는 기능

Out[219]: Ticker 066570.KS

Date	
2018-01-02	NaN
2018-01-03	-320243.0
2018-01-04	349708.0
2018-01-05	-93859.0
2018-01-08	1596197.0

In [220... lge['diff'] = lge['Volume'].diff() # 새로운 컬럼 'diff'에 diff 반환값을 할당

In [221... lge.head()

Out[221]:

	Price	Close	High	Low	Open	Volume	diff
Ticker	066570.KS	066570.KS	066570.KS	066570.KS	066570.KS	066570.KS	
Date							
2018-01-02	102588.343750	103993.663527	98840.824344	99777.704195	948226	NaN	
2018-01-03	102588.343750	103525.223602	100714.584047	102588.343750	627983	-320243.0	
2018-01-04	99309.273438	103993.673128	98372.393499	103525.233159	977691	349708.0	
2018-01-05	103993.664062	104462.103991	99309.264780	99309.264780	883832	-93859.0	
2018-01-08	98372.390625	106804.309821	96967.070759	106335.869866	2480029	1596197.0	

In [229... # 확인용 나중에 지우기 $\pi\pi$
lge['diff'].isna().sum()
lge['Volume'].shift(1).isna().sum()
print(lge.columns)

```
MultiIndex([( 'Close', '066570.KS'),  
            ( 'High', '066570.KS'),  
            ( 'Low', '066570.KS'),  
            ( 'Open', '066570.KS'),  
            ('Volume', '066570.KS'),  
            ( 'diff', '')],  
           names=['Price', 'Ticker'])
```

pip install --user numexpr

In [230]...

```
# yfinance 데이터 형식이 바뀌었는지 Ticker 가 들어가게 되면서
# MultiIndex가 되어서 Volume 값 불러올 때, '066570.KS' 까지 같이 작성해야해요.
# 같이 작성해야 lge.diff 랑 lge.Volume~ 이 전부 Series로 읽혀서 계산이 가능해집니다!

pct = (lge['diff'] / lge[('Volume', '066570.KS')]) * 100
print(pct.head(10))
```

```
Date
2018-01-02      NaN
2018-01-03   -50.995489
2018-01-04    35.768765
2018-01-05   -10.619552
2018-01-08    64.362030
2018-01-09   -78.720348
2018-01-10   -52.655741
2018-01-11   -14.804683
2018-01-12     3.847602
2018-01-15   -18.410843
dtype: float64
```

In [231]...

pct

Out[231]:

0

Date	
2018-01-02	NaN
2018-01-03	-50.995489
2018-01-04	35.768765
2018-01-05	-10.619552
2018-01-08	64.362030
...	...
2025-04-08	-33.474432
2025-04-09	-17.806515
2025-04-10	37.551157
2025-04-11	-90.742702
2025-04-14	-inf

1787 rows × 1 columns

dtype: float64

In [232]...

lge['pct'] = pct

In [233]...

lge.head()

Out[233]:

	Price	Close	High	Low	Open	Volume	diff	pct
Ticker		066570.KS	066570.KS	066570.KS	066570.KS	066570.KS		
Date								
2018-01-02	102588.343750	103993.663527	98840.824344	99777.704195	948226	NaN	NaN	
2018-01-03	102588.343750	103525.223602	100714.584047	102588.343750	627983	-320243.0	-50.995489	
2018-01-04	99309.273438	103993.673128	98372.393499	103525.233159	977691	349708.0	35.768765	
2018-01-05	103993.664062	104462.103991	99309.264780	99309.264780	883832	-93859.0	-10.619552	
2018-01-08	98372.390625	106804.309821	96967.070759	106335.869866	2480029	1596197.0	64.362030	

소수점 이하 숫자가 너무 많습니다! 소수점 둘째자리까지 구해보겠습니다.

In [234]...

lge['pct'] = round(pct, 2)

In [235...

```
lge.tail()
```

Out[235]:

	Price	Close	High	Low	Open	Volume	diff	pct
Ticker	066570.KS	066570.KS	066570.KS	066570.KS	066570.KS	066570.KS		
Date								
2025-04-08	66900.0	69500.0	66800.0	69100.0	752667	-251951.0	-33.47	
2025-04-09	64700.0	66900.0	64100.0	65800.0	638901	-113766.0	-17.81	
2025-04-10	68700.0	70100.0	68000.0	69500.0	1023079	384178.0	37.55	
2025-04-11	67900.0	68100.0	66700.0	67200.0	536366	-486713.0	-90.74	
2025-04-14	69000.0	0.0	0.0	0.0	0	-536366.0	-inf	

Q. 지금까지 만든 lge 데이터프레임의 데이터를 잠깐 위에 몇 개만 확인해보세요. 위에서 이미 했었죠?

In [236...

```
# 여기에 작성해보세요.  
lge.head()
```

Out[236]:

	Price	Close	High	Low	Open	Volume	diff	pct
Ticker	066570.KS	066570.KS	066570.KS	066570.KS	066570.KS	066570.KS		
Date								
2018-01-02	102588.343750	103993.663527	98840.824344	99777.704195	948226	NaN	NaN	
2018-01-03	102588.343750	103525.223602	100714.584047	102588.343750	627983	-320243.0	-51.00	
2018-01-04	99309.273438	103993.673128	98372.393499	103525.233159	977691	349708.0	35.77	
2018-01-05	103993.664062	104462.103991	99309.264780	99309.264780	883832	-93859.0	-10.62	
2018-01-08	98372.390625	106804.309821	96967.070759	106335.869866	2480029	1596197.0	64.36	

! 지금까지 한 작업을 유식한 말로 Feature Engineering 이라고 합니다! 머신러닝할 때 많이 사용하는 기법이죠~

유익하게 생각되는 데이터를 있는 데이터 안에서 새롭게 만들어내는 방법을 말합니다!

In [237...

```
lge.head()
```

Out[237]:

	Price	Close	High	Low	Open	Volume	diff	pct
Ticker	066570.KS	066570.KS	066570.KS	066570.KS	066570.KS	066570.KS		
Date								
2018-01-02	102588.343750	103993.663527	98840.824344	99777.704195	948226	NaN	NaN	
2018-01-03	102588.343750	103525.223602	100714.584047	102588.343750	627983	-320243.0	-51.00	
2018-01-04	99309.273438	103993.673128	98372.393499	103525.233159	977691	349708.0	35.77	
2018-01-05	103993.664062	104462.103991	99309.264780	99309.264780	883832	-93859.0	-10.62	
2018-01-08	98372.390625	106804.309821	96967.070759	106335.869866	2480029	1596197.0	64.36	

In [238...

```
len(lge)
```

Out[238]:

1787

!? 이제 다 했지만 왜 for 반복문을 사용하는 것인가?!

여러 개의 데이터가 있는 위 데이터프레임에서 pct 값을 하나씩 하나씩 가져와서 40% 이상인지 아닌지 따져야 하는 작업일 땐... for 문이 필요합니다!

DataFrame의 여러 행에서 한 개 행씩 가져와서 pct의 값이 40이 넘는지 따져보고 넘으면 새로운 리스트에 추가해 보도록 합시다.

개발 방법

1. 새로운 컬럼에 넣을 새로운 데이터를 만드는 방법: 새로운 list를 만들어 list에 값을 하나씩 추가한다.
 - 새로운리스트 = []
 - 새로운리스트.append(넣을 값)
2. 만들어진 list를 새로운 컬럼에 넣는 방법: lge['새롭게 추가할 컬럼 이름'] = 새롭게 만든 list

```
In [239... volume_signal = [] # list
percentages = lge['pct'] # 증감율 데이터
```

```
In [240... percentages.head(10)
```

```
Out[240]:
```

	pct
Date	
2018-01-02	NaN
2018-01-03	-51.00
2018-01-04	35.77
2018-01-05	-10.62
2018-01-08	64.36
2018-01-09	-78.72
2018-01-10	-52.66
2018-01-11	-14.80
2018-01-12	3.85
2018-01-15	-18.41

dtype: float64

```
In [ ]: # 아래부터 작성해보세요.
```

```
In [241... lge.head(20)
```

Out[241]:

	Price	Close	High	Low	Open	Volume	diff	pct
Ticker	066570.KS	066570.KS	066570.KS	066570.KS	066570.KS	066570.KS		
Date								
2018-01-02	102588.343750	103993.663527	98840.824344	99777.704195	948226		NaN	NaN
2018-01-03	102588.343750	103525.223602	100714.584047	102588.343750	627983	-320243.0	-51.00	
2018-01-04	99309.273438	103993.673128	98372.393499	103525.233159	977691	349708.0	35.77	
2018-01-05	103993.664062	104462.103991	99309.264780	99309.264780	883832	-93859.0	-10.62	
2018-01-08	98372.390625	106804.309821	96967.070759	106335.869866	2480029	1596197.0	64.36	
2018-01-09	102119.898438	103056.778240	98372.379229	98372.379229	1387659	-1092370.0	-78.72	
2018-01-10	102119.898438	104462.097943	100714.578734	102119.898438	909012	-478647.0	-52.66	
2018-01-11	103525.218750	103993.658654	101651.459135	103056.778846	791790	-117222.0	-14.80	
2018-01-12	102588.343750	104462.103453	100246.144121	103525.223602	823474	31684.0	3.85	
2018-01-15	100246.140625	103056.780082	99309.260806	102588.340172	695438	-128036.0	-18.41	
2018-01-16	101183.015625	101651.455512	98372.376302	99777.695964	555065	-140373.0	-25.29	
2018-01-17	103525.218750	105867.418269	101183.019231	101651.459135	815085	260020.0	31.90	
2018-01-18	100714.593750	104930.553488	100714.593750	104930.553488	503773	-311312.0	-61.80	
2018-01-19	100714.593750	102588.353634	99777.713808	100246.153779	533503	29730.0	5.57	
2018-01-22	102119.898438	102119.898438	99777.698932	101651.458536	469543	-63960.0	-13.62	
2018-01-23	102588.343750	103056.783676	96967.064640	100714.584047	1166939	697396.0	59.76	
2018-01-24	96030.179688	101651.458498	95561.739787	100714.578697	1406328	239389.0	17.02	
2018-01-25	96967.078125	97903.958107	94624.878170	96030.198143	899675	-506653.0	-56.32	
2018-01-26	96498.625000	98840.824636	96030.185073	97903.944782	777285	-122390.0	-15.75	
2018-01-29	95561.750000	97435.509804	94624.870098	96498.629902	1031365	254080.0	24.64	

In [242...

```

# Solution
volume_signal = []

for i in IBM.Percent:
    if i > 40:
        volume_signal.append('UP')
    else:
        volume_signal.append('-')

```

In [243...

```

IBM['Signal'] = volume_signal
IBM.head()

```

Out[243]:

	Price	Close	High	Low	Open	Volume	Difference	Percent	Signal
Ticker	IBM	IBM	IBM	IBM	IBM	IBM			
Date									
2020-01-02	101.918869	102.295176	101.429671	101.602767	3293436		NaN	NaN	-
2020-01-03	101.106041	101.497399	100.518997	100.526523	2482890	-810546.0	-32.65		-
2020-01-06	100.925407	101.030767	100.248050	100.413627	2537073	54183.0	2.14		-
2020-01-07	100.993149	101.572667	100.398592	100.616848	3232977	695904.0	21.53		-
2020-01-08	101.836067	102.250008	100.789935	101.233978	4545916	1312939.0	28.88		-

In [244...

```

lge.head()

```

```
Out[244]:
```

	Price	Close	High	Low	Open	Volume	diff	pct
Ticker	066570.KS	066570.KS	066570.KS	066570.KS	066570.KS	066570.KS		
Date								
2018-01-02	102588.343750	103993.663527	98840.824344	99777.704195	948226		NaN	NaN
2018-01-03	102588.343750	103525.223602	100714.584047	102588.343750	627983	-320243.0	-51.00	
2018-01-04	99309.273438	103993.673128	98372.393499	103525.233159	977691	349708.0	35.77	
2018-01-05	103993.664062	104462.103991	99309.264780	99309.264780	883832	-93859.0	-10.62	
2018-01-08	98372.390625	106804.309821	96967.070759	106335.869866	2480029	1596197.0	64.36	

! apply 메소드를 사용하는 방법도 있습니다! 이것은 pandas가 익숙해지면 많이 사용하게 될거예요!

```
In [245...
def get_signal(pct):
    if pct >= 40:
        return 'Over'
    return '-'
```

```
In [246...
IBM['new'] = IBM['Percent'].apply(lambda x : get_signal(x))
```

```
In [247...
IBM.head()
```

```
Out[247]:
```

	Price	Close	High	Low	Open	Volume	Difference	Percent	Signal	new
Ticker	IBM	IBM	IBM	IBM	IBM	IBM				
Date										
2020-01-02	101.918869	102.295176	101.429671	101.602767	3293436		NaN	NaN	-	-
2020-01-03	101.106041	101.497399	100.518997	100.526523	2482890	-810546.0	-32.65		-	-
2020-01-06	100.925407	101.030767	100.248050	100.413627	2537073	54183.0	2.14		-	-
2020-01-07	100.993149	101.572667	100.398592	100.616848	3232977	695904.0	21.53		-	-
2020-01-08	101.836067	102.250008	100.789935	101.233978	4545916	1312939.0	28.88		-	-

! 해본 김에 간단한 그림도 그려볼까요!

```
In [248...
%autosave 1
```

Autosaving every 1 seconds

```
In [249...
%matplotlib inline
```

```
In [251...
IBM['Close'].plot(figsize=(15, 7))
plt.savefig('IBM.pdf')
```

[illegible]

Jupyter Notebook에서는 잘 모르는 우리를 위해 기능을 찾아볼 수 있도록 Shift+Tab이란 기능을 마련해주었습니다!!!!

```
In [252]: IBM['Close'].plot(figsize=(15, 7), title='IBM')
```

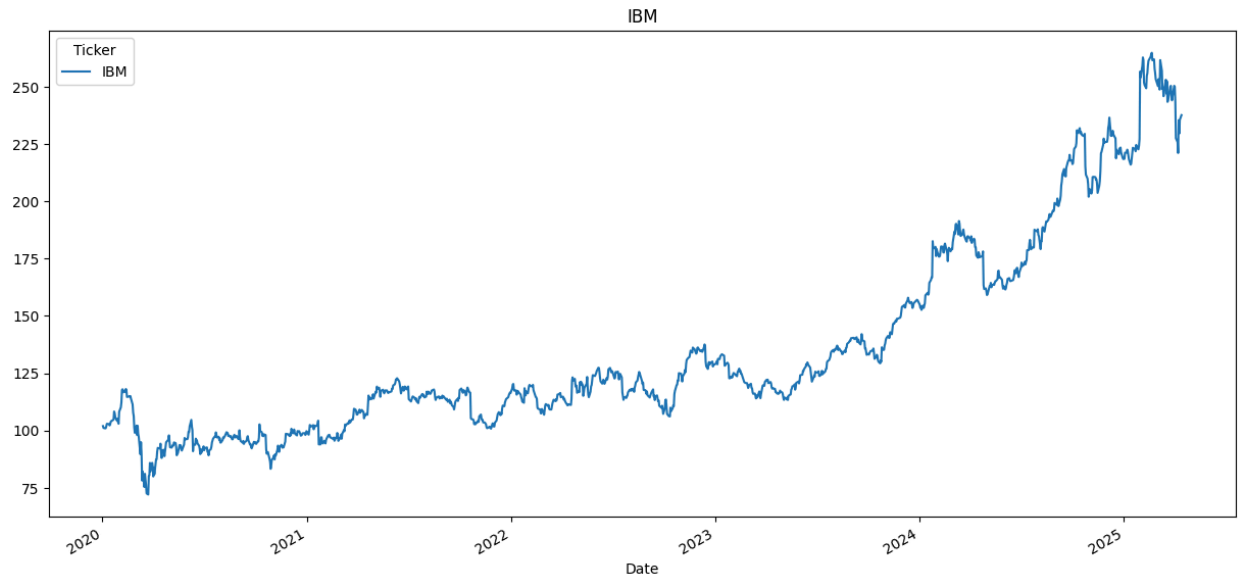
```
Out[252]: <Axes: title={'center': 'IBM'}, xlabel='Date'>
```

[illegible]

```

WARNING:matplotlib.font_manager:findfont: Font family 'AppleGothic' not found.
WARNING:matplotlib.font_manager:findfont: Font family 'AppleGothic' not found.
WARNING:matplotlib.font_manager:findfont: Font family 'AppleGothic' not found.
WARNING:matplotlib.font_manager:findfont: Font family 'AppleGothic' not found.
WARNING:matplotlib.font_manager:findfont: Font family 'AppleGothic' not found.
WARNING:matplotlib.font_manager:findfont: Font family 'AppleGothic' not found.
WARNING:matplotlib.font_manager:findfont: Font family 'AppleGothic' not found.
WARNING:matplotlib.font_manager:findfont: Font family 'AppleGothic' not found.
WARNING:matplotlib.font_manager:findfont: Font family 'AppleGothic' not found.
WARNING:matplotlib.font_manager:findfont: Font family 'AppleGothic' not found.
WARNING:matplotlib.font_manager:findfont: Font family 'AppleGothic' not found.
WARNING:matplotlib.font_manager:findfont: Font family 'AppleGothic' not found.
WARNING:matplotlib.font_manager:findfont: Font family 'AppleGothic' not found.
WARNING:matplotlib.font_manager:findfont: Font family 'AppleGothic' not found.
WARNING:matplotlib.font_manager:findfont: Font family 'AppleGothic' not found.
WARNING:matplotlib.font_manager:findfont: Font family 'AppleGothic' not found.
WARNING:matplotlib.font_manager:findfont: Font family 'AppleGothic' not found.
WARNING:matplotlib.font_manager:findfont: Font family 'AppleGothic' not found.

```



```
In [253... import matplotlib.pyplot as plt
```

```
In [254... plt.savefig('IBM.pdf')
```

<Figure size 640x480 with 0 Axes>

```
In [255... #에러가 나는 경우 위에 코딩셀을 추가하여 pd.core.common.is_list_like = pd.api.types.is_list_like 를 먼저 슬
from pandas_datareader import data
import yfinance as yf
```

Q. 각자 원하는 주식 코드(Quote)를 찾아서 위와 같은 그림을 그려보아요!

<http://finance.naver.com/>에서 찾아볼 수 있습니다.

```
In [ ]: #여기에 코딩을 작성하세요
```