2022.2학기 빅데이터 시각화

학번: 2004401

이름: 고유진

6주차: 파이썬 준비와 데이터 처리

시리즈 (Series)

더블클릭 또는 Enter 키를 눌러 수정

```
import numpy as np
import pandas as pd

sdata = [2000.0, 3000.0,4000.0, np.nan]
city = ['서울', '부산', '울산', '목포']
myseries2 = pd.Series(sdata, index=city)

myseries2.name = '호호호' # 객체 자체의 이름
myseries2.index.name = '크크크' # 색인의 이름
print(  myseries2 )
```

```
크크크
서울    2000.0
부산    3000.0
울산    4000.0
목포       NaN
Name: 호호호, dtype: float64
```

2.1.1 시리즈 생성 방법

```
myseries = pd.Series(range(0, 4))  # 연속된 숫자의 배열을 이용하여 생성합니다.
print(myseries)
```

```
0    0
1    1
2    2
3    3
dtype: int64
```

```
myseries = pd.Series([4, 5, 6])    # Python의 list 구조를 사용할 수 있습니다.
print(myseries)
```

```
0    4
1    5
2    6
dtype: int64
```

```
myseries = pd.Series([4, 5, 6], index=['a', 'b', 'c']) # 생성시 index를 이용하여 직접 색인을 지정할
print(myseries)
```

```
a    4
b    5
c    6
dtype: int64
```

```
sdata = {'서울' : 3000, '부산' : 2000}    # Python의 사전을 이용하여 생성합니다.
myseries = pd.Series( sdata )              # 사전의 key가 순서대로 색인으로 들어갑니다. 사전(키, 값
print(myseries)
```

```
서울    3000
부산    2000
dtype: int64
```

2.2. Series의 데이터 읽기와 쓰기

```
import pandas as pd

myindex = ['서울', '부산', '광주', '대구', '울산', '목포', '여수']
mylist = [50, 60, 40, 80, 70, 30, 20]
myseries = pd.Series(data=mylist, index=myindex)

myseries
```

```
서울    50
부산    60
광주    40
대구    80
울산    70
목포    30
여수    20
dtype: int64
```

```
print('₩n색인의 이름으로 값 읽기')
print(myseries[['대구']])
```

```
색인의 이름으로 값 읽기
대구    80
dtype: int64
```

```
print('₩n여러 개의 색인 이름으로 데이터 읽기')
print(myseries[['대구', '여수']])
```

```
여러 개의 색인 이름으로 데이터 읽기
대구    80
여수    20
dtype: int64
```

```
print('₩n정수를 이용한 데이터 읽기')
print(myseries[[2]])
```

```
정수를 이용한 데이터 읽기
광주    40
dtype: int64
```

```
print(myseries)
print('₩n0, 2, 4번째 데이터 읽기')
print(myseries[0:5:2]) # [시작:끝:간격]
```

```
서울    50
부산    60
광주    40
대구    80
울산    70
목포    30
여수    20
dtype: int64

0, 2, 4번째 데이터 읽기
서울    50
광주    40
울산    70
dtype: int64
```

```
myseries['대구']
```

```
80
```

```
myseries['대구':'목포']
```

```
대구    80
울산    70
목포    30
dtype: int64
```

```
print(myseries)
print('₩n2번째 항목의 값 변경')
myseries[2] = 22
print(myseries)
```

```
서울    50
부산    60
광주    40
대구    80
울산    70
```

```
목포    30
여수    20
dtype: int64

2번째 항목의 값 변경
서울    50
부산    60
광주    22
대구    80
울산    70
목포    30
여수    20
dtype: int64
```

```
print(myseries)
print('₩n1, 3, 5번째 데이터 읽기')
print(myseries[[1, 3, 5]])
```

```
서울    50
부산    60
광주    40
대구    80
울산    70
목포    30
여수    20
dtype: int64

1, 3, 5번째 데이터 읽기
부산    60
대구    80
목포    30
dtype: int64
```

```
print(myseries)
print('₩n슬라이싱 사용하기')
print(myseries[3:6])
```

```
서울    50
부산    60
광주    22
대구    80
울산    70
목포    30
여수    20
dtype: int64

슬라이싱 사용하기
대구    80
울산    70
목포    30
dtype: int64
```

```
print(myseries)
print('₩n2번째부터 4번째 까지 항목의 값 변경')
myseries[2:5] = 33
print(myseries)
```

```
서울    50
부산    60
광주    22
대구    80
울산    70
목포    30
여수    20
dtype: int64

2번째부터 4번째 까지 항목의 값 변경
서울    50
부산    60
광주    33
대구    33
울산    33
목포    30
여수    20
dtype: int64
```

```
print(myseries)
print('₩n서울과 대구만  55로 변경')
myseries[['서울', '대구']] = 55
print(myseries)
```

```
서울    50
부산    60
광주    33
대구    33
울산    33
목포    30
여수    20
dtype: int64

서울과 대구만  55로 변경
서울    55
부산    60
광주    33
대구    55
울산    33
목포    30
여수    20
dtype: int64
```

```
print(myseries)
print('₩n짝수 행만  77로 변경')
myseries[0::2] = 77
print(myseries)
```

```
서울    77
부산    60
광주    77
대구    55
울산    77
목포    30
여수    77
dtype: int64
```

```
짝수 행만  77로 변경
서울    77
부산    60
광주    77
대구    55
울산    77
목포    30
여수    77
dtype: int64
```

```
print('₩n시리즈 내용 확인')
print(myseries)
```

```
시리즈 내용 확인
서울    77
부산    60
광주    77
대구    55
울산    77
목포    30
여수    77
dtype: int64
```

```
print('₩nfinished')
```

```
finished
```

2.3 DataFrame

```
import pandas as pd

# 표를 만들기 위한 데이터 사전
sdata = {'city' : ['서울', '서울', '서울', '부산', '부산'],
         'year' : [2000, 2001, 2002, 2001, 2002],
         'pop' : [1.5, 1.7, 3.6, 2.4, 2.9 ]}

mycolumn = ['city', 'year', 'pop']    # 컬럼
myindex = ['one', 'two', 'three', 'four', 'five'] # row(색인)
myframe = pd.DataFrame( sdata, columns=mycolumn, index = myindex )
print( myframe )
```

```
      city  year  pop
one     서울  2000  1.5
two     서울  2001  1.7
three   서울  2002  3.6
four    부산  2001  2.4
five    부산  2002  2.9
```

2.4 DataFrame 데이터 읽기와 쓰기

```
import numpy as np
import pandas as pd
myindex = ['이순신', '김유신', '강감찬', '광해군', '연산군']
mycolumns = ['서울','부산','광주','목포', '경주']
mylist = list(10 * ondata for ondata in range(1, 26))
print(mylist)
```

```
[10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200,
```

```
myframe = pd.DataFrame(np.reshape(mylist, (5,5)), index=myindex, columns=mycolumns)
myframe
```

	서울	부산	광주	목포	경주
이순신	10	20	30	40	50
김유신	60	70	80	90	100
강감찬	110	120	130	140	150
광해군	160	170	180	190	200
연산군	210	220	230	240	250

```
result = myframe.iloc[1]
type(result)
result
```

```
서울     60
부산     70
광주     80
목포     90
경주    100
Name: 김유신, dtype: int64
```

```
result = myframe.iloc[1:3]
type(result)
result
```

	city	year	pop
two	서울	2001	1.7
three	서울	2002	3.6

```
print('₩n# 홀수행만 가져 오기')
result = myframe.iloc[1::2]
```

```
print(type(result))
print(result)
```

```
# 홀수행만 가져 오기
<class 'pandas.core.frame.DataFrame'>
     city  year  pop
two    서울  2001  1.7
four   부산  2001  2.4
```

```
result = myframe.iloc[0::2]
result
```

	서울	부산	광주	목포	경주
이순신	10	20	30	40	50
강감찬	110	120	130	140	150
연산군	210	220	230	240	250

```
result = myframe.loc['이순신']
type(result)
result
```

```
서울    10
부산    20
광주    30
목포    40
경주    50
Name: 이순신, dtype: int64
```

```
result = myframe.loc[['이순신']]
type(result)
result
```

	서울	부산	광주	목포	경주
이순신	10	20	30	40	50

```
result = myframe.loc[['이순신','강감찬']]
type(result)
result
```

	서울	부산	광주	목포	경주
이순신	10	20	30	40	50
강감찬	110	120	130	140	150

```
myframe.index
```

```
Index(['이순신', '김유신', '강감찬', '광해군', '연산군'], dtype='object')
```

```
mytarget = np.random.choice(myframe.index, 3)
mytarget
```

```
array(['김유신', '강감찬', '광해군'], dtype=object)
```

```
result = myframe.loc[mytarget]
result
```

	서울	부산	광주	목포	경주
김유신	60	70	80	90	100
강감찬	110	120	130	140	150
광해군	160	170	180	190	200

```
result = myframe.loc[['강감찬'], ['광주']]
result
```

	광주
강감찬	130

```
result = myframe.loc[['연산군', '강감찬'], ['광주', '목포']]
result
```

	광주	목포
연산군	230	240
강감찬	130	140

```
result = myframe.loc['김유신':'광해군','광주' :'목포']
result
```

	광주	목포
김유신	80	90
강감찬	130	140
광해군	180	190

```
result = myframe.loc['김유신':'광해군',['부산']]
result
```

	부산
김유신	70

```
result = myframe.loc[[False, True, True, False, True]]
result
```

	city	year	pop
two	서울	2001	1.7
three	서울	2002	3.6
five	부산	2002	2.9

```
print('₩nBoolean으로 데이터 처리하기')
result = myframe.loc[[False, True, True, False, True],[False, True, True, False, True]]
print(result)
```

```
Boolean으로 데이터 처리하기
      부산   광주   경주
김유신   70   80  100
강감찬  120  130  150
연산군  220  230  250
```

```
result = myframe.loc[myframe['부산']<=100]
result
```

	서울	부산	광주	목포	경주
이순신	10	20	30	40	50
김유신	60	70	80	90	100

```
result = myframe.loc[myframe['목포']<=140]
result
```

	서울	부산	광주	목포	경주
이순신	10	20	30	40	50
김유신	60	70	80	90	100
강감찬	110	120	130	140	150

```
cond1 = myframe['부산'] >= 70
cond2 = myframe['목포'] >= 140

type(cond1)
```

```
pandas.core.series.Series
```

```
cond1
```

```
이순신    False
김유신     True
강감찬     True
광해군     True
연산군     True
Name: 부산, dtype: bool
```

```
cond2
```

```
이순신    False
김유신    False
강감찬     True
광해군     True
연산군     True
Name: 목포, dtype: bool
```

```
df = pd.DataFrame([cond1, cond2])
print(df)
print('-' * 40)
```

```
      이순신    김유신    강감찬    광해군    연산군
부산  False   True  True  True  True
목포  False  False  True  True  True
----------------------------------------
```

```
df = pd.DataFrame([cond1, cond2])
df
```

	이순신	김유신	강감찬	광해군	연산군
부산	False	True	True	True	True
목포	False	False	True	True	True

```
df.all()
```

```
이순신    False
김유신    False
강감찬     True
광해군     True
연산군     True
dtype: bool
```

```
df.any()
```

```
이순신    False
김유신     True
강감찬     True
광해군     True
연산군     True
dtype: bool
```

```
result = myframe.loc[ df.all() ]
result
```

	서울	부산	광주	목포	경주
강감찬	110	120	130	140	150
광해군	160	170	180	190	200
연산군	210	220	230	240	250

```
result = myframe.loc[ df.any() ]
result
```

	서울	부산	광주	목포	경주
김유신	60	70	80	90	100
강감찬	110	120	130	140	150
광해군	160	170	180	190	200
연산군	210	220	230	240	250

```
result = myframe.loc[ lambda df : df['광주'] >= 130 ]
result
```

	서울	부산	광주	목포	경주
강감찬	110	120	130	140	150
광해군	160	170	180	190	200
연산군	210	220	230	240	250

```
myframe.loc[['이순신','강감찬'], ['부산']] = 30
myframe
```

	서울	부산	광주	목포	경주
이순신	10	30	30	40	50
김유신	60	70	80	90	100
강감찬	110	30	130	140	150
광해군	160	170	180	190	200
연산군	210	220	230	240	250

```
myframe.loc['김유신':'광해군',['경주']] = 80
myframe
```

	서울	부산	광주	목포	경주
이순신	10	30	30	40	50
김유신	60	70	80	90	80
강감찬	110	30	130	140	80
광해군	160	170	180	190	80
연산군	210	220	230	240	250

```
#연산군의 모든 실적을 50으로 변경하기
myframe.loc[['연산군'],:] = 50
myframe
```

	서울	부산	광주	목포	경주
이순신	10	30	30	40	50
김유신	60	70	80	90	80
강감찬	110	30	130	140	80
광해군	160	170	180	190	80
연산군	50	50	50	50	50

```
#모든 사람의 광주 컬럼을 60으로 변경하기
myframe.loc[:, ['광주']] = 60
myframe
```

	서울	부산	광주	목포	경주
이순신	10	20	60	40	50
김유신	60	70	60	90	100
강감찬	110	120	60	140	150
광해군	160	170	60	190	200
연산군	210	220	60	240	250

```
myframe2 = myframe
print('₩n# 광주의 모든 실적을 99으로 변경하기')
myframe2.loc[:, ['광주']] = 100
print(myframe2)
```

```
# 광주의 모든 실적을 99으로 변경하기
      서울   부산   광주   목포   경주
이순신   10   20  100   40   50
김유신   60   70  100   90  100
강감찬  110  120  100  140  150
광해군  160  170  100  190  200
연산군  210  220  100  240  250
```

```
#경주 실적이 150이하인 데이터를 모두 0으로 변경하기
myframe.loc[myframe['경주'] <= 150, ['경주', '광주']] = 0
myframe
```

	서울	부산	광주	목포	경주
이순신	10	30	0	40	0
김유신	60	70	0	90	0
강감찬	110	30	0	140	0
광해군	160	170	0	190	0
연산군	50	50	0	50	0

```
#데이터 프레임 사용하기
myframe
```

	서울	부산	광주	목포	경주
이순신	10	20	100	40	50
김유신	60	70	100	90	100
강감찬	110	120	100	140	150
광해군	160	170	100	190	200
연산군	210	220	100	240	250

```
print('₩nfinished')
```

```
finished
```

함수 적용과 매핑(apply 함수)

```
import pandas as pd
```

✓ 0초 오후 4:29에 완료됨