Pandas

Pandas

- 파이썬에서 사용하는 데이터분석 라이브러리
  - 행과 열로 이루어진 데이터 객체를 만들어 다룰 수 있게 되며 보다 안정적으로 대용량의 데이터들을 처리하는데 매우 편리한 도구
  - **import pandas as pd**
- **자료구조**
  - **Series**
    - 1차원 자료구조
  - **DataFrame**
    - 2차원 자료구조인 DataFrame는 행과 열이 있는 테이블 데이타

Pandas Series

```
#Series 만들기
import pandas as pd

ds = pd.Series(tempavg)
print(ds)
```

```
0      4.1
1      5.5
2      9.0
3     15.0
4     19.0
5     21.3
6     26.1
7     27.0
8     22.6
9     18.1
10    11.2
```

```
#기본 인덱스로 접근하기
print(ds[2:5])
```

```
2     9.0
3    15.0
4    19.0
dtype: float64
```

```
#인덱스 지정하기
ds = pd.Series(tempavg, index=day)
print(ds)
```

```
2017-01-01      4.1
2017-02-01      5.5
2017-03-01      9.0
2017-04-01     15.0
2017-05-01     19.0
2017-06-01     21.3
2017-07-01     26.1
2017-08-01     27.0
2017-09-01     22.6
2017-10-01     18.1
2017-11-01     11.2
```

```
#지정인덱스로 접근하기
print(ds['2017-03-01':'2017-05-01'])
```

```
2017-03-01      9.0
2017-04-01     15.0
2017-05-01     19.0
dtype: float64
```

Pandas DataFrame 데이터 가져오기

```
#데이터 프레임 만들기

data = {
    '일자' : day,
    '평균기온' : tempavg,
    '최고기온' : tempmax,
    '최저기온' : tempmin
}

df = pd.DataFrame(data)
df
```

	일자	최고기온	최저기온	평균기온
0	2017-01-01	15.1	-7.7	4.1
1	2017-02-01	17.7	-5.7	5.5
2	2017-03-01	18.1	-2.3	9.0
3	2017-04-01	23.2	5.0	15.0
4	2017-05-01	28.2	12.8	19.0
5	2017-06-01	30.4	15.4	21.3

열 가져오기

```
#1열가져오기
print(df['최저기온'])
```

```
0    -7.7
1    -5.7
```

```
#여러 열가져오기
print(df[['일자', '최저기온']])
```

```
          일자  최저기온
0  2017-01-01  -7.7
1  2017-02-01  -5.7
```

행 가져오기

```
#1행 가져오기
df.loc[4]
```

```
일자      2017-05-01
최고기온        28.2
최저기온        12.8
평균기온          19
Name: 4, dtype: object
```

```
#여러 행 가져오기
df.loc[:2]
```

	일자	최고기온	최저기온	평균기온
0	2017-01-01	15.1	-7.7	4.1
1	2017-02-01	17.7	-5.7	5.5
2	2017-03-01	18.1	-2.3	9.0

```
#여러 행 가져오기
df.loc[[2, 9, 27]]
```

	일자	최고기온	최저기온	평균기온
2	2017-03-01	18.1	-2.3	9.0
9	2017-10-01	28.7	7.3	18.1
27	2019-04-01	23.4	2.1	13.4

Pandas DataFrame 슬라이싱

- .loc[행인덱싱, 열인덱싱]
- .iloc[행순서번호, 열순서번호]

```
#데이터 프레임 슬라이싱
df.loc[0:2, '일자':'최고기온']
```

	일자	최고기온
0	2017-01-01	15.1
1	2017-02-01	17.7
2	2017-03-01	18.1

```
df.iloc[0:2, 0:2]
```

	일자	최고기온
0	2017-01-01	15.1
1	2017-02-01	17.7

Pandas 외부데이터 가져오기

- csv 읽기
  - df = pd.read_csv('파일명')
  - 한글 파일명을 사용할 경우 :
    df = pd.read_csv('파일명', engine='python')
- Csv 쓰기
  - df.to_csv('파일명')

```
import pandas as pd

#외부 데이터 가져오기
df = pd.read_csv('부산시기온.csv', engine='python')
df
```

	일시	평균	최고	최저
0	2017-01-01	4.1	15.1	-7.7
1	2017-02-01	5.5	17.7	-5.7
2	2017-03-01	9.0	18.1	-2.3
3	2017-04-01	15.0	23.2	5.0
4	2017-05-01	19.0	28.2	12.8
5	2017-06-01	21.3	30.4	15.4

```
#데이터 생성하기
subset = df.iloc[0:2, 0:2]
subset
```

	일시	평균
0	2017-01-01	4.1
1	2017-02-01	5.5

```
subset.to_csv('부분기온.csv')
```

Pandas 날짜 데이터

```
#날짜 데이터
#열별 데이터 타입 확인
df.dtypes

#열 타입 변경
df['일시'] = pd.to_datetime(df['일시'])

#년도 월 열 삽입
df['년도'] = df['일시'].dt.year
df['월'] = df['일시'].dt.month

df
```

날짜유형으로 변경

날짜유형에서 년,월 추출

cf)
숫자유형으로 변경
to_numeric()

	일시	평균	최고	최저	년도	월
0	2017-01-01	4.1	15.1	-7.7	2017	1
1	2017-02-01	5.5	17.7	-5.7	2017	2
2	2017-03-01	9.0	18.1	-2.3	2017	3
3	2017-04-01	15.0	23.2	5.0	2017	4
4	2017-05-01	19.0	28.2	12.8	2017	5
5	2017-06-01	21.3	30.4	15.4	2017	6

Pandas 기초 통계

```
#특정 열 합계 구하기

print('평균기온 합계 :' , round(df['평균'].sum(),1))
```

```
평균기온 합계 : 438.5
```

```
#열합계
colsum = df.sum(axis = 0)
print(colsum)
```

```
일시      2017-01-012017-02-012017-03-012017-04-012017-0...
평균                                                  438.5
최고                                                  734.4
최저                                                  157.7
dtype: object
```

```
#행합계
rowsum = df.sum(axis = 1)
print(rowsum)
```

```
0     11.5
1     17.5
2     24.8
3     43.2
```

Pandas 그룹화

```
#그룹으로 묶기
dfg = df.groupby('년도')[['평균','최고','최저']].mean()
dfg
```

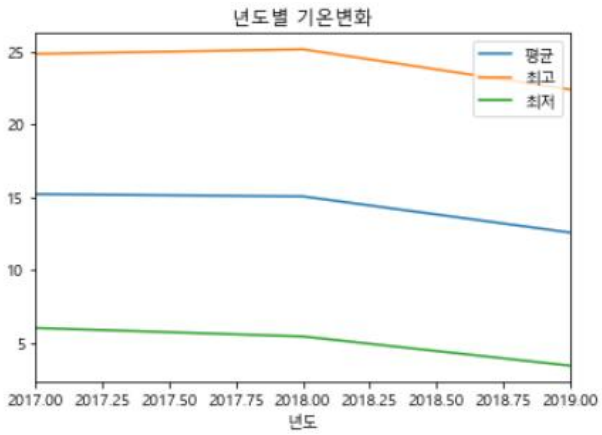
년도	평균	최고	최저
2017	15.208333	24.841667	6.008333
2018	15.050000	25.158333	5.425000
2019	12.566667	22.400000	3.416667

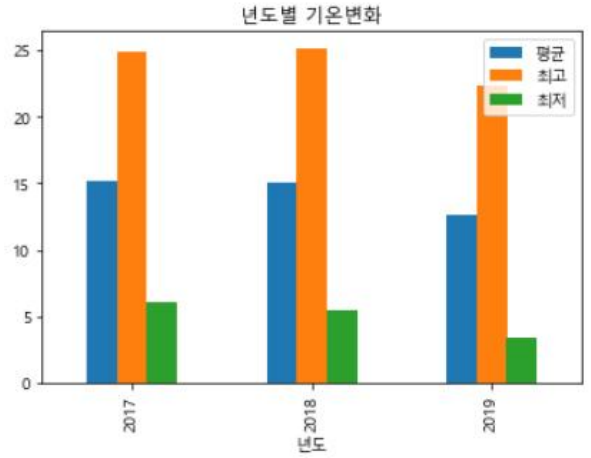
Pandas 시각화

```
#시각화
import matplotlib.pyplot as plt
#한글 폰트 사용
from matplotlib import font_manager, rc
font_name = font_manager.FontProperties(fname="c:/Windows/Fonts/malgun.ttf").get_name()
rc('font', family=font_name)

dfg.plot()
plt.title("년도별 기온변화")
plt.show()

dfg.plot(kind='bar')
plt.title("년도별 기온변화")
plt.show()
```





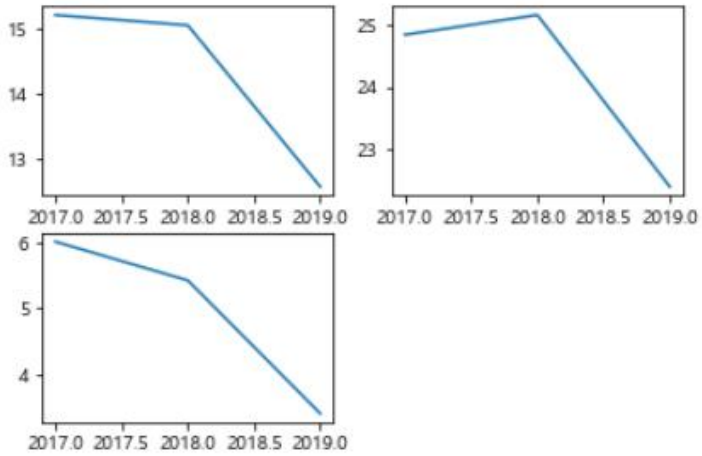
Pandas 시각화 subplot

```
#시각화
import matplotlib.pyplot as plt
#한글 폰트 사용
from matplotlib import font_manager, rc
font_name = font_manager.FontProperties(fname="c:/Windows/Fonts/malgun.ttf").get_name()
rc('font', family=font_name)

fig = plt.figure()
ax1 = fig.add_subplot(2,2,1)
ax2 = fig.add_subplot(2,2,2)
ax3 = fig.add_subplot(2,2,3)

ax1.plot(dfg['평균'])
ax2.plot(dfg['최고'])
ax3.plot(dfg['최저'])

fig
```



해결문제

- 부산시 2017년 ~2019년까지 기온 변화 시각화 자료를 이용하여 그래프를 그리시오.