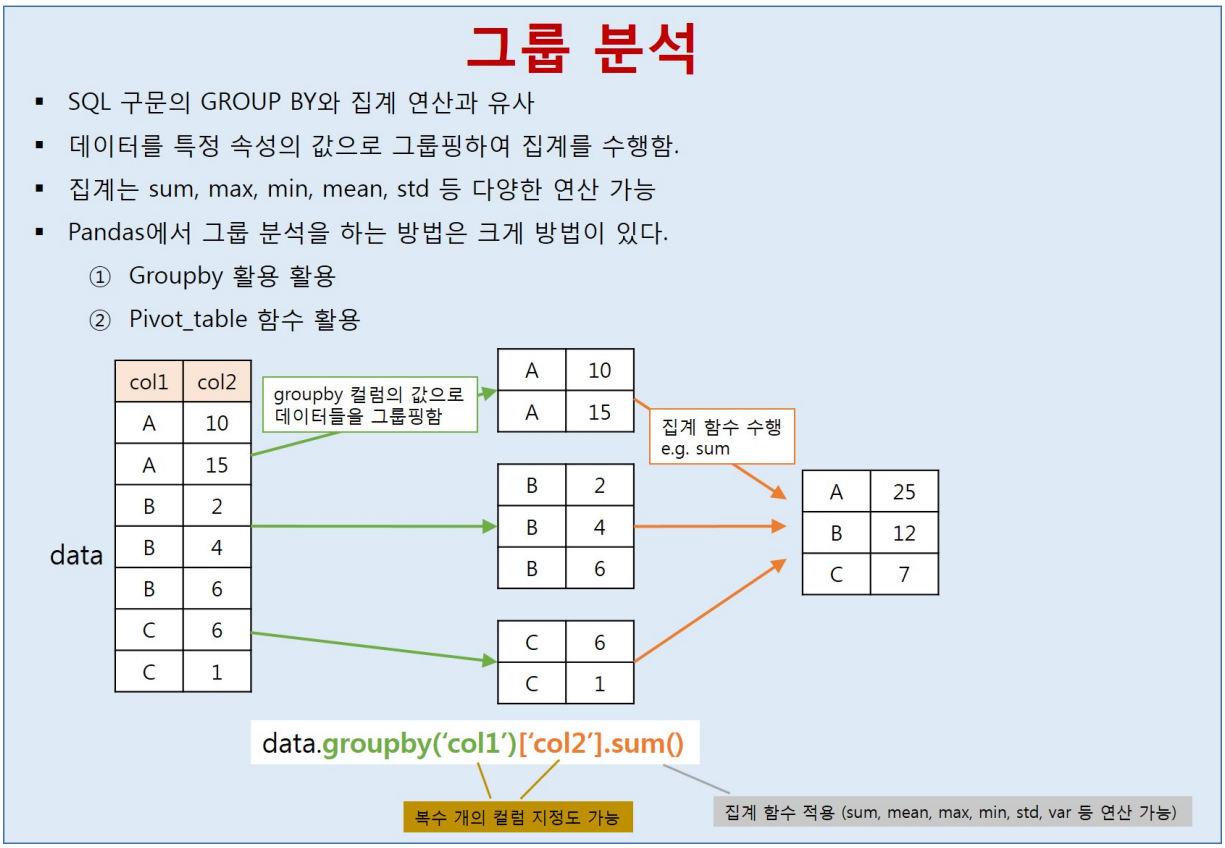
pandas, numpy module 불러오기

In [1]:
```
import pandas as pd
import numpy as np
```



In [2]:
```
# 실습 데이터 적재
data = pd.read_excel('data/인구수예제.xlsx')
data.head()
```

Out[2]:

	도시	자치구	연도	남자인구	여자인구	총인구
0	서울	강남구	2013	73	92	165
1	서울	강남구	2014	139	55	194
2	서울	강남구	2015	123	83	206
3	서울	강남구	2016	147	150	297
4	서울	강남구	2017	57	133	190

1. groupby

(dataframe).groupby('집단별로 묶고 싶은 기준 column')['이후에 연산을 하고 싶은 column'].함수

In [3]:
```
# 연도별 총인구의 합
# 집단별로 묶고 싶은 기준 : '연도'
# 연산을 하고 싶은 column : '총인구'
# 사용할 함수 : 'sum()'
data.groupby('연도')['총인구'].sum()
```

Out[3]:
```
연도
2013    2130
2014    2256
2015    1848
2016    2045
2017    2128
Name: 총인구, dtype: int64
```

In [4]:
```
pd.DataFrame(data.groupby('연도')['총인구'].sum())
```

Out[4]:

	총인구
연도	
2013	2130
2014	2256
2015	1848
2016	2045
2017	2128

In [5]:
```
# reset_index() : index의 값을 column으로 이동
pd.DataFrame(data.groupby('연도')['총인구'].sum()).reset_index()
```

Out[5]:

	연도	총인구
0	2013	2130
1	2014	2256
2	2015	1848
3	2016	2045
4	2017	2128

In [6]:
```
# 자치구별 총인구의 평균
data.groupby('자치구')['총인구'].mean()
```

Out[6]:
```
자치구
강남구      210.4
도봉구      207.0
동래구      207.2
동작구      207.2
서대문구     206.2
송파구      194.8
수영구      212.2
영등포구     238.2
종로구      171.2
해운대구     227.0
Name: 총인구, dtype: float64
```

2. pivot_table

(dataframe).pivot_table(index='', aggfunc='', values='')

index='집단별로 묶고 싶은 기준 column',

aggfunc=함수,

values='이후에 연산을 하고 싶은 column'

In [7]:
```
# 연도별 총인구의 합
data.pivot_table(index = '연도', aggfunc = 'sum', values = '총인구')
```

Out[7]:

	총인구
연도	
2013	2130
2014	2256
2015	1848
2016	2045
2017	2128

In [8]:
```
# 자치구별 총인구의 평균
data.pivot_table(index = '자치구', aggfunc = 'mean', values = '총인구')
```

Out[8]:

자치구	총인구
강남구	210.4
도봉구	207.0
동래구	207.2
동작구	207.2
서대문구	206.2
송파구	194.8
수영구	212.2
영등포구	238.2
종로구	171.2
해운대구	227.0

group, pivot_table 예제

1) 연도별 전체의 남자인구, 여자인구, 총인구 수 합 구하기

In [9]:
```
# groupby
data.groupby('연도')[['남자인구', '여자인구', '총인구']].sum()
```

Out[9]:

연도	남자인구	여자인구	총인구
2013	1099	1031	2130
2014	1196	1060	2256
2015	872	976	1848
2016	947	1098	2045
2017	1063	1065	2128

In [10]:
```
# pivot_table
data.pivot_table(index = '연도', aggfunc = 'sum', values = ['남자인구', '여자인구',
```

Out[10]:

연도	남자인구	여자인구	총인구
2013	1099	1031	2130
2014	1196	1060	2256
2015	872	976	1848
2016	947	1098	2045
2017	1063	1065	2128

2) 도시/자치구별 평균 총인구수

In [11]:
```
# groupby
data.groupby(['도시', '자치구'])['총인구'].mean()
```

Out[11]:
```
도시  자치구
부산  동래구      207.2
    수영구      212.2
    해운대구     227.0
서울  강남구      210.4
    도봉구      207.0
    동작구      207.2
    서대문구     206.2
    송파구      194.8
    영등포구     238.2
    종로구      171.2
Name: 총인구, dtype: float64
```

In [12]:
```
# pivot_table
data.pivot_table(index = ['도시', '자치구'], aggfunc = 'mean', values = ['총인구'])
```

Out[12]:

도시	자치구	총인구
부산	동래구	207.2
	수영구	212.2
	해운대구	227.0
서울	강남구	210.4
	도봉구	207.0
	동작구	207.2
	서대문구	206.2
	송파구	194.8
	영등포구	238.2
	종로구	171.2

3) 도시별, 연도별로 총인구수 합 출력

In [13]:
```
# groupby
pd.DataFrame(data.groupby(['도시', '연도'])['총인구'].sum()).unstack('도시')
```

Out[13]:

	총인구	
도시	부산	서울
연도		
2013	603	1527
2014	683	1573
2015	597	1251
2016	652	1393
2017	697	1431

In [14]:
```
# pivot_table
data.pivot_table(index = '연도', columns = '도시', aggfunc = 'sum', values = '총인
```

Out[14]:

도시	부산	서울
연도		
2013	603	1527
2014	683	1573
2015	597	1251
2016	652	1393
2017	697	1431

실습

1. 자치구별로 평균 총 인구수

In [15]:
```
# groupby 사용
data.groupby(['자치구'])['총인구'].mean()
```

Out[15]:
```
자치구
강남구      210.4
도봉구      207.0
동래구      207.2
동작구      207.2
서대문구     206.2
송파구      194.8
수영구      212.2
영등포구     238.2
종로구      171.2
해운대구     227.0
Name: 총인구, dtype: float64
```

In [16]:
```
# pivot_table 사용
data.pivot_table(index = '자치구', aggfunc = 'mean', values = '총인구')
```

Out[16]:

자치구	총인구
강남구	210.4
도봉구	207.0
동래구	207.2
동작구	207.2
서대문구	206.2
송파구	194.8
수영구	212.2
영등포구	238.2
종로구	171.2
해운대구	227.0

2. 도시/자치구별 평균 남자인구와 여자인구수 구하기

In [17]:
```
# groupby 사용
data.groupby(['도시', '자치구'])['남자인구', '여자인구'].mean()
```

Out[17]:

도시	자치구	남자인구	여자인구
부산	동래구	112.4	94.8
	수영구	100.4	111.8
	해운대구	124.0	103.0
서울	강남구	107.8	102.6
	도봉구	97.0	110.0
	동작구	90.8	116.4
	서대문구	97.6	108.6
	송파구	83.0	111.8
	영등포구	125.8	112.4
	종로구	96.6	74.6

In [18]:
```
# pivot_table 사용
data.pivot_table(index = ['도시', '자치구'], aggfunc = 'mean', values = ['남자인구',
```

Out[18]:

		남자인구	여자인구
도시	자치구		
부산	동래구	112.4	94.8
	수영구	100.4	111.8
	해운대구	124.0	103.0
서울	강남구	107.8	102.6
	도봉구	97.0	110.0
	동작구	90.8	116.4
	서대문구	97.6	108.6
	송파구	83.0	111.8
	영등포구	125.8	112.4
	종로구	96.6	74.6

3. 도시/자치구 별로 남자인구의 평균을 구한 후, 남자인구가 가장 많은 도시/자치구 5개를 출력하시오.

In [19]:
```
# groupby 사용
data.groupby(['도시', '자치구'])['남자인구'].mean().sort_values(ascending = False).h
```

Out[19]:
```
도시  자치구
서울  영등포구    125.8
부산  해운대구    124.0
    동래구     112.4
서울  강남구     107.8
부산  수영구     100.4
Name: 남자인구, dtype: float64
```

In [21]:
```
# pivot_table 사용
data.pivot_table(index = ['도시', '자치구'], aggfunc = 'mean', values = '남자인구').
```

Out[21]:

		남자인구
도시	자치구	
서울	영등포구	125.8
부산	해운대구	124.0
	동래구	112.4
서울	강남구	107.8
부산	수영구	100.4

함수를 여러개 적용하고 싶은 경우

In [4]:
```
# 연도, 도시별로 총인구의 평균과, 표준편차 구해보기
```

Out[4]:

	mean		std	
도시	부산	서울	부산	서울
연도				
2013	201.000000	218.142857	45.902070	33.982488
2014	227.666667	224.714286	16.563011	52.184654
2015	199.000000	178.714286	57.297469	31.731763
2016	217.333333	199.000000	49.541229	62.372537
2017	232.333333	204.428571	12.662280	32.315410

In [22]:
```
# pivot_table
# mean: 평균 / std: 표준편차

data.pivot_table(index = ['연도', '도시'], aggfunc = ['mean', 'std'],
                 values = '총인구')
```

Out[22]:

		mean	std
		총인구	총인구
연도	도시		
2013	부산	201.000000	45.902070
	서울	218.142857	33.982488
2014	부산	227.666667	16.563011
	서울	224.714286	52.184654
2015	부산	199.000000	57.297469
	서울	178.714286	31.731763
2016	부산	217.333333	49.541229
	서울	199.000000	62.372537
2017	부산	232.333333	12.662280
	서울	204.428571	32.315410

In [23]:
```
data.pivot_table(index = '연도', columns = '도시', aggfunc = ['mean', 'std'],
                 values = '총인구')
```

Out[23]:

	mean		std	
도시	부산	서울	부산	서울
연도				
2013	201.000000	218.142857	45.902070	33.982488
2014	227.666667	224.714286	16.563011	52.184654
2015	199.000000	178.714286	57.297469	31.731763
2016	217.333333	199.000000	49.541229	62.372537
2017	232.333333	204.428571	12.662280	32.315410

In [25]:
```
# groupby 함수에서는 agg() 함수를 적용하고 안에 list형식으로 넣어준다

data.groupby(['연도', '도시'])['총인구'].agg(['mean', 'std']).unstack(level = 1)
```

Out[25]:

	mean		std	
도시	부산	서울	부산	서울
연도				
2013	201.000000	218.142857	45.902070	33.982488
2014	227.666667	224.714286	16.563011	52.184654
2015	199.000000	178.714286	57.297469	31.731763
2016	217.333333	199.000000	49.541229	62.372537
2017	232.333333	204.428571	12.662280	32.315410

In [34]:
```
# 실습. 도시/자치구별 기준으로 남녀차이의 평균이 가장 많이 나는 도시/자치구 3개를 찾아보기
# Hint: abs() = 절대값을 구하는 함수

data['남녀차이'] = abs(data['남자인구'] - data['여자인구'])
```

In [35]:
```
data.groupby(['도시', '자치구'])['남녀차이'].mean().sort_values(ascending = False).h
```

Out[35]:
```
도시  자치구
서울  송파구    50.8
부산  수영구    45.8
서울  강남구    44.4
Name: 남녀차이, dtype: float64
```

In [33]:
```
temp = data.pivot_table(index = ['도시', '자치구'], aggfunc = 'mean', values = '남
temp.sort_values(by = '남녀차이', ascending = False).head(3)
```

Out[33]:

		남녀차이
도시	자치구	
서울	송파구	50.8
부산	수영구	45.8
서울	강남구	44.4