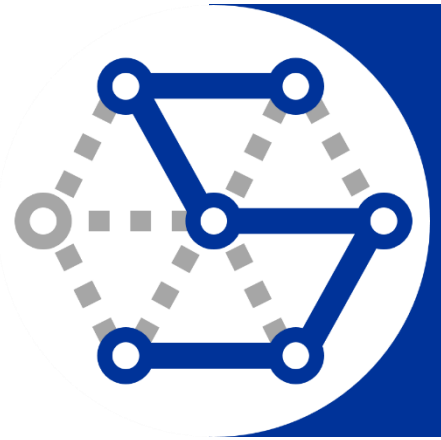
데이터 처리 프로그래밍

Data Processing Programming

12
데이터 처리 및 탐색

목차

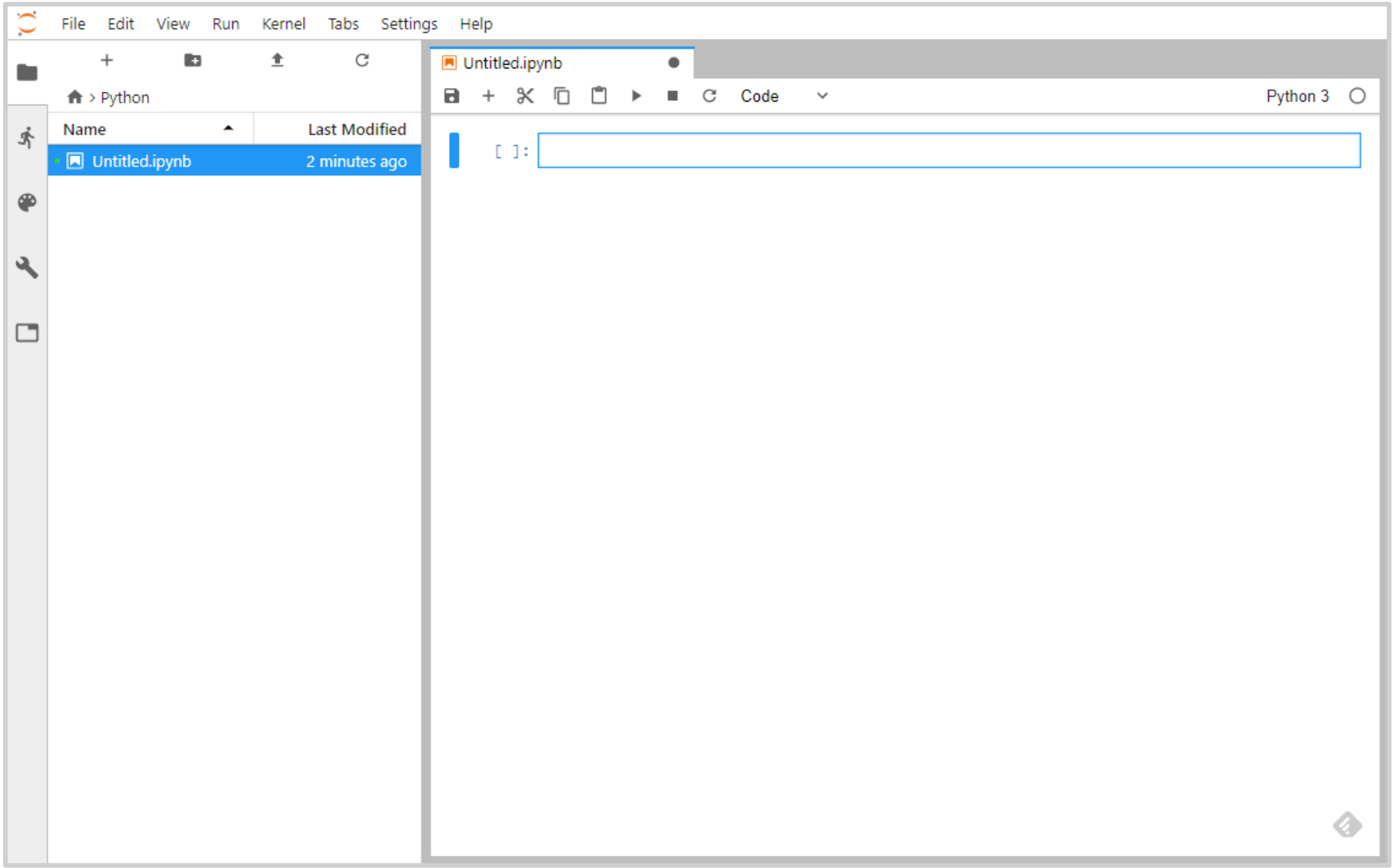


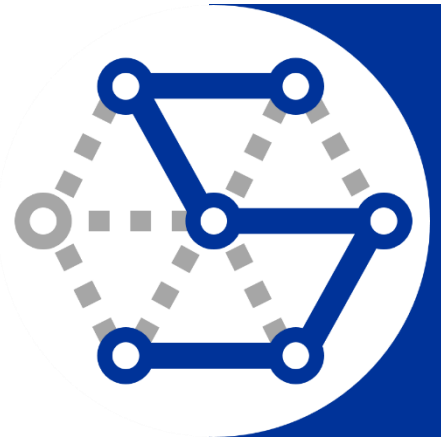
1. Jupyter 소개

Jupyter

- Jupyter Notebook은 오픈소스 웹 애플리케이션
- 라이브 코드, 등식, 시각화와 설명을 위한 텍스트 등을 포함하여 문서를 만들고 공유 가능
- 주로 데이터 전처리와 변형, 수치 시뮬레이션, 통계 모델링, 머신 러닝 등에 사용
- 다양한 프로그래밍 언어를 지원하고, 실시간으로 인터렉티브하게 데이터를 조작하고 시각화

JupyterLab


File Edit View Run Kernel Tabs Settings Help
Python
Name
Last Modified
Untitled.ipynb
2 minutes ago
Untitled.ipynb
Code
Python 3
[ ]:



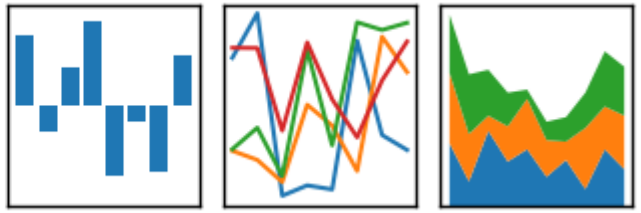
2. Pandas 소개

Pandas

- Pandas는 NumPy 패키지를 기반으로 만들어짐
- Pandas의 Series와 DataFrame 객체는 NumPy 배열 구조를 기반으로 함
- 인덱싱된 데이터의 1차원 배열 형태인 Series 객체 제공
- 행과 열 레이블이 부착된 다차원 배열로서 DataFrame이라는 효율적인 자료구조 제공
- DataFrame은 여러 가지 타입의 데이터를 가질 수 있고, 데이터 누락도 허용됨
- Pandas는 레이블이 붙은 데이터를 위한 편리한 스토리지 인터페이스를 제공
- 데이터베이스 프레임워크와 스프레드시트 프로그램 사용자에게 익숙한 강력한 데이터 연산을 구현

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



Pandas 이용

Pandas

Pandas 이용

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

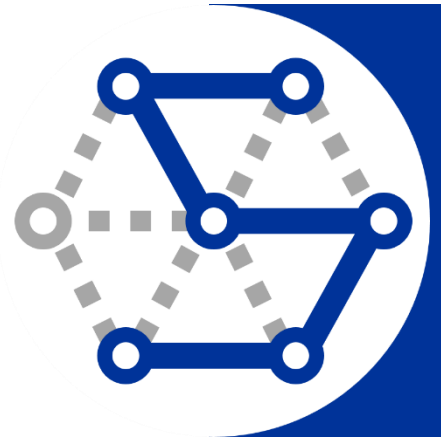
[2]:
```
np.__version__
```

[2]:
```
'1.15.4'
```

[3]:
```
pd.__version__
```

[3]:
```
'0.23.4'
```

- numpy와 pandas 패키지 버전



3. Pandas Series

Pandas Series 객체

Pandas Series 객체

```
[4]: series = pd.Series([0.1, 0.4, 0.2, 0.6, 0.5, 0.8, 0.7])

[5]: series

[5]: 0    0.1
     1    0.4
     2    0.2
     3    0.6
     4    0.5
     5    0.8
     6    0.7
     dtype: float64

[6]: series.values

[6]: array([0.1, 0.4, 0.2, 0.6, 0.5, 0.8, 0.7])

[7]: series.index

[7]: RangeIndex(start=0, stop=7, step=1)
```

```
[8]: series[0]

[8]: 0.1

[9]: series[2:4]

[9]: 2    0.2
     3    0.6
     dtype: float64

[10]: series[1:-3]

[10]: 1    0.4
      2    0.2
      3    0.6
      dtype: float64
```

Pandas Series 객체


```
[11]: series = pd.Series([0.1, 0.4, 0.2, 0.6, 0.5, 0.8, 0.7],
                         index=['a', 'b', 'c', 'd', 'e', 'f', 'g'])

[12]: series

[12]: a    0.1
      b    0.4
      c    0.2
      d    0.6
      e    0.5
      f    0.8
      g    0.7
      dtype: float64

[13]: series['c']

[13]: 0.2
```

```
[14]: series = pd.Series([0.1, 0.4, 0.2, 0.6, 0.5, 0.8, 0.7],
                         index=[1, 4, 2, 6, 5, 8, 7])

[15]: series

[15]: 1    0.1
      4    0.4
      2    0.2
      6    0.6
      5    0.5
      8    0.8
      7    0.7
      dtype: float64

[16]: series[5]

[16]: 0.5
```

Pandas Series 객체

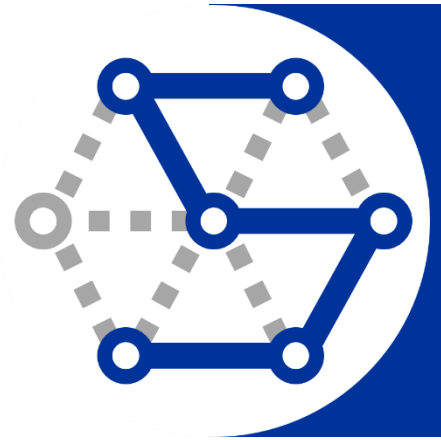
```
[17]: population = {'Seoul':9741381,
                    'Busan':3416918,
                    'Incheon':2925967,
                    'Daegu':2453041,
                    'Daejeon':1525849,
                    'Gwangju':1496172}

[18]: population_series = pd.Series(population)

[19]: population_series

[19]: Seoul      9741381
      Busan      3416918
      Incheon    2925967
      Daegu      2453041
      Daejeon    1525849
      Gwangju    1496172
      dtype: int64
```

```
[20]: population_series['Busan']

[20]: 3416918

[21]: population_series['Daegu':'Gwangju']

[21]: Daegu      2453041
      Daejeon    1525849
      Gwangju    1496172
      dtype: int64
```

Pandas Series 객체


```
[22]: pd.Series([1, 2, 3, 4, 5])

[22]: 0    1
      1    2
      2    3
      3    4
      4    5
      dtype: int64

[23]: pd.Series(5, index=['a', 'b', 'c'])

[23]: a    5
      b    5
      c    5
      dtype: int64
```

```
[24]: pd.Series({2:'b', 1:'a', 3:'c'})

[24]: 2    b
      1    a
      3    c
      dtype: object

[25]: pd.Series({2:'b', 1:'a', 3:'c'}, index=[2, 3])

[25]: 2    b
      3    c
      dtype: object
```



4. Pandas DataFrame

Pandas DataFrame 객체

[ ]:

Pandas DataFrame 객체

[26]:
```
population = {'Seoul':9741381,
              'Busan':3416918,
              'Incheon':2925967,
              'Daegu':2453041,
              'Daejeon':1525849,
              'Gwangju':1496172}
```

[27]:
```
population
```

[27]:
```
{'Seoul': 9741381,
 'Busan': 3416918,
 'Incheon': 2925967,
 'Daegu': 2453041,
 'Daejeon': 1525849,
 'Gwangju': 1496172}
```

[28]:
```
male = {'Seoul':4757642,
        'Busan':1680933,
        'Incheon':1472081,
        'Daegu':1218326,
        'Daejeon':765718,
        'Gwangju':745122}
```

[29]:
```
male
```

[29]:
```
{'Seoul': 4757642,
 'Busan': 1680933,
 'Incheon': 1472081,
 'Daegu': 1218326,
 'Daejeon': 765718,
 'Gwangju': 745122}
```

Pandas DataFrame 객체


```
[30]: female = { 'Seoul':4984229,
                 'Busan':1735985,
                 'Incheon':1453886,
                 'Daegu':1234715,
                 'Daejeon':760131,
                 'Gwangju':751050}

[31]: female

[31]: {'Seoul': 4984229,
       'Busan': 1735985,
       'Incheon': 1453886,
       'Daegu': 1234715,
       'Daejeon': 760131,
       'Gwangju': 751050}
```

```
[32]: korea = pd.DataFrame({'population':population,
                           'male':male,
                           'female':female})

[33]: korea
```

[33]:

	population	male	female
Busan	3416918	1680933	1735985
Daegu	2453041	1218326	1234715
Daejeon	1525849	765718	760131
Gwangju	1496172	745122	751050
Incheon	2925967	1472081	1453886
Seoul	9741381	4757642	4984229

Pandas DataFrame 객체


```
[34]: korea.index

[34]: Index(['Busan', 'Daegu', 'Daejeon', 'Gwangju', 'Incheon', 'Seoul'],
      dtype='object')

[35]: korea.columns

[35]: Index(['population', 'male', 'female'], dtype='object')

[36]: korea['female']

[36]: Busan      1735985
      Daegu      1234715
      Daejeon     760131
      Gwangju     751050
      Incheon    1453886
      Seoul      4984229
      Name: female, dtype: int64
```

```
[37]: pd.DataFrame([1, 2, 3, 4, 5], columns=['number'])
```

[37]:

	number
0	1
1	2
2	3
3	4
4	5

Pandas DataFrame 객체


```
[38]: l = [{'a':i, 'b':i*2, 'c':i/2}
           for i in range(5)]

[39]: l

[39]: [{'a': 0, 'b': 0, 'c': 0.0},
       {'a': 1, 'b': 2, 'c': 0.5},
       {'a': 2, 'b': 4, 'c': 1.0},
       {'a': 3, 'b': 6, 'c': 1.5},
       {'a': 4, 'b': 8, 'c': 2.0}]

[40]: pd.DataFrame(l)
```

[40]:

	a	b	c
0	0	0	0.0
1	1	2	0.5
2	2	4	1.0
3	3	6	1.5
4	4	8	2.0

```
[41]: pd.DataFrame([{'a':1, 'b':2},{'b':3, 'c':4},{'c':5, 'd':6}])
```

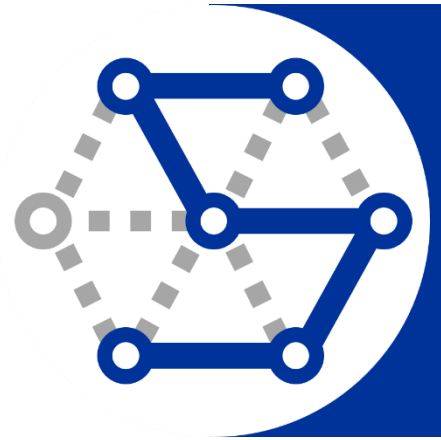
[41]:

	a	b	c	d
0	1.0	2.0	NaN	NaN
1	NaN	3.0	4.0	NaN
2	NaN	NaN	5.0	6.0

```
[42]: pd.DataFrame(np.random.rand(4, 3),
                   columns=['a', 'b', 'c'],
                   index=[1, 2, 3, 4])
```

[42]:

	a	b	c
1	0.572001	0.309018	0.231040
2	0.580856	0.196229	0.872685
3	0.779150	0.138881	0.497667
4	0.207357	0.480959	0.658109



5. Pandas Index

Pandas Index 객체

Pandas Index 객체

```
[43]: index = pd.Index([2, 4, 6, 8, 10])

[44]: index

[44]: Int64Index([2, 4, 6, 8, 10], dtype='int64')

[45]: index[1]

[45]: 4

[46]: index[1:2:2]

[46]: Int64Index([4], dtype='int64')

[47]: index[-1::]

[47]: Int64Index([10], dtype='int64')

[48]: index[::2]

[48]: Int64Index([2, 6, 10], dtype='int64')

[49]: print(index.size, index.shape, index.ndim, index.dtype)

      5 (5,) 1 int64
```

```
[50]: idx1 = pd.Index([1, 2, 3, 4, 5])

[51]: idx2 = pd.Index([2, 4, 6, 8, 10])

[52]: idx1 & idx2

[52]: Int64Index([2, 4], dtype='int64')

[53]: idx1 | idx2

[53]: Int64Index([1, 2, 3, 4, 5, 6, 8, 10], dtype='int64')

[54]: idx1 ^ idx2

[54]: Int64Index([1, 3, 5, 6, 8, 10], dtype='int64')
```

Pandas Index 객체

```
[55]: series = pd.Series([0.1, 0.2, 0.3, 0.4, 0.5],
                         index=['a', 'b', 'c', 'd', 'e'])

[56]: series

[56]: a    0.1
      b    0.2
      c    0.3
      d    0.4
      e    0.5
      dtype: float64

[57]: series['b']

[57]: 0.2

[58]: 'a' in series

[58]: True

[59]: series.keys()

[59]: Index(['a', 'b', 'c', 'd', 'e'], dtype='object')

[60]: list(series.items())

[60]: [('a', 0.1), ('b', 0.2), ('c', 0.3), ('d', 0.4), ('e', 0.5)]
```

```
[61]: series['f'] = 0.6

[62]: series

[62]: a    0.1
      b    0.2
      c    0.3
      d    0.4
      e    0.5
      f    0.6
      dtype: float64
```

Pandas Index 객체

```
[63]: series[2]

[63]: 0.3

[64]: series.loc['c']

[64]: 0.3

[65]: series.loc['b':'d']

[65]: b    0.2
      c    0.3
      d    0.4
      dtype: float64

[66]: series.iloc[1:4]

[66]: b    0.2
      c    0.3
      d    0.4
      dtype: float64
```

Pandas Index 객체


```
[67]: korea
```

[67]:

	population	male	female
Busan	3416918	1680933	1735985
Daegu	2453041	1218326	1234715
Daejeon	1525849	765718	760131
Gwangju	1496172	745122	751050
Incheon	2925967	1472081	1453886
Seoul	9741381	4757642	4984229

```
[68]: korea.population
```

```
[68]: Busan      3416918
      Daegu      2453041
      Daejeon    1525849
      Gwangju    1496172
      Incheon    2925967
      Seoul      9741381
      Name: population, dtype: int64
```

```
[69]: korea['gender_ratio'] = (korea['male'] * 100) / korea['female']
```

```
[70]: korea['gender_ratio']
```

```
[70]: Busan       96.828774
      Daegu       98.672649
      Daejeon    100.735005
      Gwangju     99.210705
      Incheon    101.251474
      Seoul       95.453921
      Name: gender_ratio, dtype: float64
```

```
[71]: korea.values
```

```
[71]: array([[3.41691800e+06, 1.68093300e+06, 1.73598500e+06, 9.68287744e+
      01],
             [2.45304100e+06, 1.21832600e+06, 1.23471500e+06, 9.86726492e+
      01],
             [1.52584900e+06, 7.65718000e+05, 7.60131000e+05, 1.00735005e+
      02],
             [1.49617200e+06, 7.45122000e+05, 7.51050000e+05, 9.92107050e+
      01],
             [2.92596700e+06, 1.47208100e+06, 1.45388600e+06, 1.01251474e+
      02],
             [9.74138100e+06, 4.75764200e+06, 4.98422900e+06, 9.54539208e+
      01]])
```

Pandas Index 객체

```
[72]: korea.T
```

[72]:

	Busan	Daegu	Daejeon	Gwangju	Incl
population	3.416918e+06	2.453041e+06	1.525849e+06	1.496172e+06	2.925967e
male	1.680933e+06	1.218326e+06	7.657180e+05	7.451220e+05	1.472081e
female	1.735985e+06	1.234715e+06	7.601310e+05	7.510500e+05	1.453886e
gender_ratio	9.682877e+01	9.867265e+01	1.007350e+02	9.921071e+01	1.012515e

```
[73]: korea.values[0]
```

```
[73]: array([3.41691800e+06, 1.68093300e+06, 1.73598500e+06, 9.68287744e+0
      1])
```

```
[74]: korea.iloc[:3, :2]
```

[74]:

	population	male
Busan	3416918	1680933
Daegu	2453041	1218326
Daejeon	1525849	765718

```
[75]: korea.loc[:'Incheon', :'male']
```

[75]:

	population	male
Busan	3416918	1680933
Daegu	2453041	1218326
Daejeon	1525849	765718
Gwangju	1496172	745122
Incheon	2925967	1472081

Pandas Index 객체

[76]:
```
korea.loc[korea.population > 2000000]
```

[76]:

	population	male	female	gender_ratio
Busan	3416918	1680933	1735985	96.828774
Daegu	2453041	1218326	1234715	98.672649
Incheon	2925967	1472081	1453886	101.251474
Seoul	9741381	4757642	4984229	95.453921

[77]:
```
korea.loc[korea.gender_ratio > 100]
```

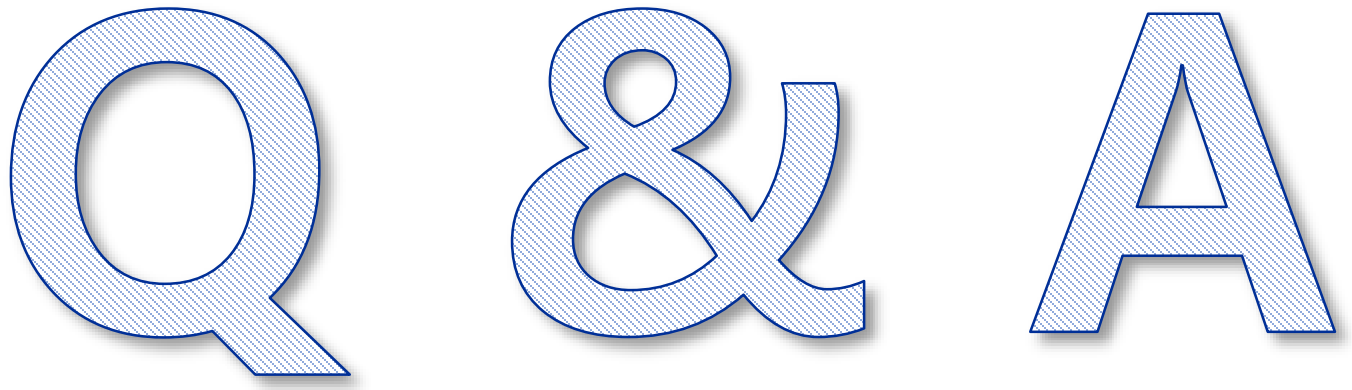
[77]:

	population	male	female	gender_ratio
Daejeon	1525849	765718	760131	100.735005
Incheon	2925967	1472081	1453886	101.251474

[78]:
```
korea[korea.gender_ratio < 100]
```

[78]:

	population	male	female	gender_ratio
Busan	3416918	1680933	1735985	96.828774
Daegu	2453041	1218326	1234715	98.672649
Gwangju	1496172	745122	751050	99.210705
Seoul	9741381	4757642	4984229	95.453921


Q & A