Statistical Modeling & Consulting

Data analysis

Importing data

1 Importing data

os.getcwd()

- 현재 작업 공간의 위치를 알려 줌.

os.chdir('경Th')

- 현재 작업 공간의 위치를 지정한 경로로 바꾸어 줌.

```
In [1]: import os

In [2]: os.getcwd()
Out[2]: 'C:\\Users\\IAN\\Documents\\Python Scripts'

In [3]: os.chdir('C:/Users/IAN')

In [4]: import pandas as pd
   ...: import numpy as np
   ...:
```

1 Importing data

pandas.read_csv('파일명과 확장자가 포함된 경로')

- 지정된 csv 파일을 작업 공간으로 불러 들임.

```
In [5]: mb = pd.read_csv("data/microbiome.csv")
   ...: mb
   ...:
Out[5]:
          Taxon  Patient  Tissue  Stool
0     Firmicutes        1     632    305
1     Firmicutes        2     136   4182
2     Firmicutes        3    1174    703
3     Firmicutes        4     408   3946
4     Firmicutes        5     831   8605
5     Firmicutes        6     693     50
6     Firmicutes        7     718    717
7     Firmicutes        8     173     33
8     Firmicutes        9     228     80
9     Firmicutes       10     162   3196
10    Firmicutes       11     372     32
```

1 Importing data

pandas.read_csv('파일명과 확장자가 포함된 경Th', header = None).head()

- 지정된 csv 파일을 작업 공간으로 불러 들임.
- Header = None: 첫 번째 행을 열의 이름으로 쓰지 않는다.
- 불러온 자료.head(): 첫 5번째 행만 불러온다.

```
In [6]: pd.read_csv("data/microbiome.csv", header=None).head()
Out[6]:
             0        1        2      3
0       Taxon  Patient  Tissue  Stool
1  Firmicutes        1     632    305
2  Firmicutes        2     136   4182
3  Firmicutes        3    1174    703
4  Firmicutes        4     408   3946
```

1 Importing data

pandas.read_csv('파일명과 확장자가 포함된 경Th', sep = ' , ')

- 지정된 csv 파일을 작업 공간으로 불러 들임.
- sep = ' , ': 구분자로 , 을 이용한다.

```
In [7]: mb = pd.read_table("data/microbiome.csv", sep=',')

In [8]: mb.head()
Out[8]:
         Taxon  Patient  Tissue  Stool
0   Firmicutes        1     632    305
1   Firmicutes        2     136   4182
2   Firmicutes        3    1174    703
3   Firmicutes        4     408   3946
4   Firmicutes        5     831   8605
```

1 Importing data

pandas.read_csv('파일명과 확장자가 포함된 경Th', index_col = ['Taxon', 'Patient'])

- 지정된 csv 파일을 작업 공간으로 불러 들임.
- Index_col = ['Taxon', 'Patient']: Taxon과 Patient라는 변수가 가지는 값들을 이용해 행의 Index를 만든다.

```
In [9]: mb = pd.read_csv("data/microbiome.csv", index_col=['Taxon','Patient'])

In [10]: mb.head()
Out[10]:
                    Tissue  Stool
Taxon      Patient
Firmicutes 1           632    305
           2           136   4182
           3          1174    703
           4           408   3946
           5           831   8605
```

1 Importing data

pandas.read_csv('파일명과 확장자가 포함된 경Th', skiprows = [3, 4, 6]).head()

- 지정된 csv 파일을 작업 공간으로 불러 들임.
- skiprows = [3, 4, 6]: 3번째, 4번째, 6번째 행을 빼고 작업 공간으로 자료를 불러 들인다.

```
In [11]: pd.read_csv("data/microbiome.csv", skiprows=[3,4,6]).head()
Out[11]:
         Taxon  Patient  Tissue  Stool
0  Firmicutes        1     632    305
1  Firmicutes        2     136   4182
2  Firmicutes        5     831   8605
3  Firmicutes        7     718    717
4  Firmicutes        8     173     33
```

1 Importing data

pandas.read_csv('파일명과 확장자가 포함된 경Th', nrows = 4)

- 지정된 csv 파일을 작업 공간으로 불러 들임.
- Nrows = 4: 4번째 행까지만 자료를 불러들인다.

```
In [12]: pd.read_csv("data/microbiome.csv", nrows=4)
Out[12]:
         Taxon  Patient  Tissue  Stool
0  Firmicutes        1     632    305
1  Firmicutes        2     136   4182
2  Firmicutes        3    1174    703
3  Firmicutes        4     408   3946
```

1 Importing data

pandas.read_csv('파일명과 확장자가 포함된 경Th', chunksize = 15)

- 지정된 csv 파일을 작업 공간으로 불러 들임.
- Chunksize = 15개의 행 단위로 자료를 나누어 저장한다.

```
In [22]: data_chunks = pd.read_csv("data/microbiome.csv", chunksize=15)
```

```
In [24]: list(data_chunks)
Out[24]:
[         Taxon  Patient  Tissue  Stool
 0   Firmicutes        1     632    305
 1   Firmicutes        2     136   4182
 2   Firmicutes        3    1174    703
 3   Firmicutes        4     408   3946
 4   Firmicutes        5     831   8605
 5   Firmicutes        6     693     50
 6   Firmicutes        7     718    717
 7   Firmicutes        8     173     33
 8   Firmicutes        9     228     80
 9   Firmicutes       10     162   3196
 10  Firmicutes       11     372     32
 11  Firmicutes       12    4255   4361
 12  Firmicutes       13     107   1667
 13  Firmicutes       14      96    223
 14  Firmicutes       15     281   2377,
              Taxon  Patient  Tissue  Stool
 0   Proteobacteria        1    1638   3886
 1   Proteobacteria        2    2469   1821
 2   Proteobacteria        3     839    661
 3   Proteobacteria        4    4414     18
 4   Proteobacteria        5   12044     83
 5   Proteobacteria        6    2310     12
 6   Proteobacteria        7    3053    547
 7   Proteobacteria        8     395   2174
```

- 자료를 list로 변환해 보면, 15개의 행 단위로 자료가 나누어 리스트의 원소에 저장되어 있음을 확인할 수 있다.
- 각 자료마다 특정 변수의 요약 통계량을 구할 수 있다.

- For 문을 이용하여 각 bacteria마다 tissue 값의 평균을 구한다.
- 각 평균값은 박테리아 이름을 키 값으로 가지는 사전의 원소로써 저장된다.

```
In [68]: data_chunks = pd.read_csv("data/microbiome.csv", chunksize=15)
    ...: data_chunks = list(data_chunks)
    ...:
    ...: mean_tissue = dict()
    ...: mean_tissue
    ...:
    ...: for chunk in data_chunks :
    ...:     mean_tissue[chunk.Taxon[0]] = chunk.Tissue.mean()
    ...:
    ...:
    ...: mean_tissue
Out[68]:
{'Actinobacteria': 449.06666666666666,
 'Bacteroidetes': 599.6666666666666,
 'Firmicutes': 684.4,
 'Other': 198.8,
 'Proteobacteria': 2943.0666666666666}
```

1 Importing data

{chunk.Taxon[0]:chunk.Tissue.mean() for chunk in data_chunks}

- 사전형 자료
- 각 자료의 Taxon 변수의 첫 번째 행의 값을 키 값으로 갖는다.
- 키 값에 대응되는 원소 값은 각 자료에서 Tissue 변수의 평균 값이다.

```
In [69]: data_chunks = pd.read_csv("data/microbiome.csv", chunksize=15)
    ...:
    ...: mean_tissue = {chunk.Taxon[0]:chunk.Tissue.mean() for chunk in data_chunks}
    ...:
    ...: mean_tissue
Out[69]:
{'Actinobacteria': 449.06666666666666,
 'Bacteroidetes': 599.6666666666666,
 'Firmicutes': 684.4,
 'Other': 198.8,
 'Proteobacteria': 2943.0666666666666}
```

1 Importing data

pandas.read_csv('파일명과 확장자가 포함된 경Th').head(20)

```
In [71]: pd.read_csv("data/microbiome_missing.csv").head(20)
    ...:
Out[71]:
            Taxon  Patient Tissue  Stool
0       Firmicutes        1    632    305
1       Firmicutes        2    136   4182
2       Firmicutes        3    NaN    703
3       Firmicutes        4    408   3946
4       Firmicutes        5    831   8605
5       Firmicutes        6    693     50
6       Firmicutes        7    718    717
7       Firmicutes        8    173     33
8       Firmicutes        9    228    NaN
9       Firmicutes       10    162   3196
10      Firmicutes       11    372 -99999
11      Firmicutes       12   4255   4361
12      Firmicutes       13    107   1667
13      Firmicutes       14      ?    223
14      Firmicutes       15    281   2377
15  Proteobacteria        1   1638   3886
16  Proteobacteria        2   2469   1821
17  Proteobacteria        3    839    661
18  Proteobacteria        4   4414     18
19  Proteobacteria        5  12044     83
```

- 불러온 자료.head(20): 첫 20번째 행만 불러온다.
- 결측치로 의심되는 값들이 존재한다.
- Na, 빈 칸만 결측치로 인식한다.

1 Importing data

Is.null(객체)

```
In [72]: pd.isnull(pd.read_csv("data/microbiome_missing.csv")).head(20)
Out[72]:
    Taxon Patient Tissue  Stool
0   False   False  False  False
1   False   False  False  False
2   False   False   True  False
3   False   False  False  False
4   False   False  False  False
5   False   False  False  False
6   False   False  False  False
7   False   False  False  False
8   False   False  False   True
9   False   False  False  False
10  False   False  False  False
11  False   False  False  False
12  False   False  False  False
13  False   False  False  False
14  False   False  False  False
15  False   False  False  False
16  False   False  False  False
17  False   False  False  False
18  False   False  False  False
19  False   False  False  False
```

- isnull 메써드는 결측 여부를 bool 값으로 반환해 준다.

1 Importing data

mb_file = pandas.ExcelFile('파일명과 확장자가 포함된 경Th')
mb1 = mb_file.parse("엑셀 sheet 이름", header=None)

- Pandas.ExcelFile('파일명과 확장자가 포함된 경로'): 지정된 경로의 엑셀 파일을 작업 공간으로 불러들인다.
- 불러들인 엑셀 파일.parse('sheet 이름'): 엑셀 파일 안의 sheet에 저장된 자료를 불러들인다.

```
In [74]: mb_file = pd.ExcelFile('data/microbiome/MID1.xls')
    ...: mb_file
Out[74]: <pandas.io.excel.ExcelFile at 0x9da05c0>

In [75]: mb1 = mb_file.parse("Sheet 1", header=None)
    ...: mb1.columns = ["Taxon", "Count"]
    ...: mb1.head()
Out[75]:
                                                Taxon  Count
0  Archaea "Crenarchaeota" Thermoprotei Desulfuro...      7
1  Archaea "Crenarchaeota" Thermoprotei Desulfuro...      2
2  Archaea "Crenarchaeota" Thermoprotei Sulfoloba...      3
3  Archaea "Crenarchaeota" Thermoprotei Thermopro...      3
4  Archaea "Euryarchaeota" "Methanomicrobia" Meth...      7
```

1 Importing data

mb2 = pd.read_excel(' 파일명과 확장자가 포함된 경Th', sheetname='Sheet 1', header=None)

- 다음과 같이 엑셀 파일 안의 sheet 1에 있는 자료를 한 번에 불러 올 수 있다.

```
In [79]: mb2 = pd.read_excel('data/microbiome/MID2.xls', sheetname='Sheet 1', header=None)
    ...: mb2.head()
Out[79]:
                                                   0   1
0  Archaea "Crenarchaeota" Thermoprotei Acidiloba...   2
1  Archaea "Crenarchaeota" Thermoprotei Acidiloba...  14
2  Archaea "Crenarchaeota" Thermoprotei Desulfuro...  23
3  Archaea "Crenarchaeota" Thermoprotei Desulfuro...   1
4  Archaea "Crenarchaeota" Thermoprotei Desulfuro...   2
```

Q & A