

Python Class 3

Excel 다루기

데이터 자료들은 구조화하기 위한 엑셀파일로 만들어져 있는 경우가 많다. 따라서 이번에는 엑셀파일을 다루어 보는 시간을 갖도록 하겠다.

먼저 아래와 같이 `pandas` 를 `import` 하도록 한다.

```
import pandas as pd
```

Pandas.read_csv()

Pandas에는 csv파일을 읽어서 Dataframe 객체에 테이블형 데이터를 리턴하는 기능을 제공하는 `read_csv` 가 있다. 아래와 같이 디렉토리경로/파일이름+확장자명까지 포함된 파일경로를 값으로 넘겨주고 head부분을 우선 출력해보도록 한다.

```
df = pd.read_csv('/~디렉토리경로/파일이름.csv')
print(df.head())
```

```
.dataframe thead th {
  text-align: left;
}

.dataframe tbody tr th {
  vertical-align: top;
}
```

	0	1	2	3	4	5	6	7	8
0	NaN	No Summer	01 !	02 !	03 !	Total	No Winter	01 !	02 !
1	Afghanistan (AFG)	13	0	0	2	2	0	0	0
2	Algeria (ALG)	12	5	2	8	15	3	0	0
3	Argentina (ARG)	23	18	24	28	70	18	0	0
4	Armenia (ARM)	5	1	2	9	12	6	0	0

읽어온 csv파일의 내용이 불필요한 인덱스와 컬럼을 포함하고 있으니 아래와 같이 코드를 수정해서 자료를 다시 구성하여 읽어들이도

록 한다. 읽은 자료는 `df` 라는 dataframe 객체에 할당하여 사용하도록 한다.

```
df = pd.read_csv('/디렉토리경로/파일이름.csv', index_col = 0, skiprows = 1)
print(df.head())
```

```
.dataframe thead th {
    text-align: left;
}

.dataframe tbody tr th {
    vertical-align: top;
}
```

	No Summer	01 !	02 !	03 !	Total	No Winter	01 !.1	02 !.1	03 !.1
Afghanistan (AFG)	13	0	0	2	2	0	0	0	0
Algeria (ALG)	12	5	2	8	15	3	0	0	0
Argentina (ARG)	23	18	24	28	70	18	0	0	0
Armenia (ARM)	5	1	2	9	12	6	0	0	0
Australasia (ANZ) [ANZ]	2	3	4	5	12	0	0	0	0

칼럼 데이터들 수정하기

```
df.columns
```

```
Index(['No Summer', '01 !', '02 !', '03 !', 'Total', 'No Winter', '01 !.1',
      '02 !.1', '03 !.1', 'Total.1', 'No Games', '01 !.2', '02 !.2', '03 !.2',
      'Combined total'],
      dtype='object')
```

```
for col in df.columns:
    if col[:2] == '01':
        df.rename(columns={col: 'Gold' + col[4:]}, inplace=True)
    if col[:2] == '02':
        df.rename(columns={col: 'Silver' + col[4:]}, inplace=True)
    if col[:2] == '03':
        df.rename(columns={col: 'Bronze' + col[4:]}, inplace=True)
    if col[:1] == 'No':
```

```
df.rename(columns={col: '#' + col[2:]}, inplace=True)
```

위 코드는 칼럼들 데이터 중에 '01', '02', '03', '№'를 각각 'Gold', 'Sivler', 'Bronze', '#'으로 수정하는 내용이다.

Dataframe 데이터 조회하기

아래 코드들은 **df** 라는 Dataframe 객체에 담긴 테이블형 데이터를 수정하고 조회하는 내용들이다. 아래와 같은 코드들을 통해 좀더 쉽게 데이터를 다룰 수 있게 된다.

```
df.count()
```

```
#Summer      147
Gold          147
Sivler        147
Bronze        147
Total         147
#Winter      147
Gold.1        147
Sivler.1      147
Bronze.1      147
Total.1       147
#Games        147
Gold.2        147
Sivler.2      147
Bronze.2      147
Combined total 147
dtype: int64
```

```
df.head()
```

```
.dataframe thead th {
  text-align: left;
}

.dataframe tbody tr th {
  vertical-align: top;
}
```

	№ Summer	01 !	02 !	03 !	Total	№ Winter	01 !.1	02 !.1	03 !.1
Afghanistan (AFG)	13	0	0	2	2	0	0	0	0
Algeria (ALG)	12	5	2	8	15	3	0	0	0
Argentina (ARG)	23	18	24	28	70	18	0	0	0

Armenia (ARM)	5	1	2	9	12	6	0	0	0
Australasia (ANZ) [ANZ]	2	3	4	5	12	0	0	0	0

```
df = df.drop('Total', axis=1)
```

```
df.head()
```

```
.dataframe thead th {
    text-align: left;
}

.dataframe tbody tr th {
    vertical-align: top;
}
```

	No Summer	01 !	02 !	03 !	No Winter	01 !.1	02 !.1	03 !.1	Total.1
Afghanistan (AFG)	13	0	0	2	0	0	0	0	0
Algeria (ALG)	12	5	2	8	3	0	0	0	0
Argentina (ARG)	23	18	24	28	18	0	0	0	0
Armenia (ARM)	5	1	2	9	6	0	0	0	0
Australasia (ANZ) [ANZ]	2	3	4	5	0	0	0	0	0

```
df.iloc[0]
```

```
#Summer      13
Gold          0
Sivler        0
Bronze        2
Total         2
#Winter       0
Gold.1        0
Sivler.1      0
Bronze.1      0
Total.1       0
#Games       13
Gold.2        0
```

```
Sivler.2      0
Bronze.2     2
Combined total 2
Name: Afghanistan (AFG), dtype: int64
```

```
df['Gold']
```

```
Afghanistan (AFG)      0
Algeria (ALG)          5
Argentina (ARG)       18
Armenia (ARM)          1
Australasia (ANZ) [ANZ] 3
Australia (AUS) [AUS] [Z] 139
Austria (AUT)         18
Azerbaijan (AZE)       6
Bahamas (BAH)          5
Bahrain (BRN)          0
Barbados (BAR) [BAR]   0
Belarus (BLR)         12
Belgium (BEL)         37
Bermuda (BER)          0
Bohemia (BOH) [BOH] [Z] 0
Botswana (BOT)         0
Brazil (BRA)          23
British West Indies (BWI) [BWI] 0
Bulgaria (BUL) [H]     51
Burundi (BDI)          1
Cameroon (CMR)         3
Canada (CAN)          59
Chile (CHI) [I]         2
China (CHN) [CHN]      201
Colombia (COL)         2
Costa Rica (CRC)        1
Ivory Coast (CIV) [CIV] 0
Croatia (CRO)          6
Cuba (CUB) [Z]        72
Cyprus (CYP)           0
...
Sri Lanka (SRI) [SRI]   0
Sudan (SUD)            0
Suriname (SUR) [E]      1
Sweden (SWE) [Z]       143
Switzerland (SUI)       47
Syria (SYR)            1
Chinese Taipei (TPE) [TPE] [TPE2] 2
Tajikistan (TJK)        0
Tanzania (TAN) [TAN]    0
Thailand (THA)          7
Togo (TOG)             0
Tonga (TGA)            0
Trinidad and Tobago (TRI) [TRI] 2
Tunisia (TUN)           3
Turkey (TUR)           39
Uganda (UGA)           2
Ukraine (UKR)          33
```

United Arab Emirates (UAE)	1
United States (USA) [P] [Q] [R] [Z]	976
Uruguay (URU)	2
Uzbekistan (UZB)	5
Venezuela (VEN)	2
Vietnam (VIE)	0
Virgin Islands (ISV)	0
Yugoslavia (YUG) [YUG]	26
Independent Olympic Participants (IOP) [IOP]	0
Zambia (ZAM) [ZAM]	0
Zimbabwe (ZIM) [ZIM]	3
Mixed team (ZZX) [ZZX]	8
Totals	4809

Name: Gold, Length: 147, dtype: int64

```
df['Gold']>0
```

Afghanistan (AFG)	False
Algeria (ALG)	True
Argentina (ARG)	True
Armenia (ARM)	True
Australasia (ANZ) [ANZ]	True
Australia (AUS) [AUS] [Z]	True
Austria (AUT)	True
Azerbaijan (AZE)	True
Bahamas (BAH)	True
Bahrain (BRN)	False
Barbados (BAR) [BAR]	False
Belarus (BLR)	True
Belgium (BEL)	True
Bermuda (BER)	False
Bohemia (BOH) [BOH] [Z]	False
Botswana (BOT)	False
Brazil (BRA)	True
British West Indies (BWI) [BWI]	False
Bulgaria (BUL) [H]	True
Burundi (BDI)	True
Cameroon (CMR)	True
Canada (CAN)	True
Chile (CHI) [I]	True
China (CHN) [CHN]	True
Colombia (COL)	True
Costa Rica (CRC)	True
Ivory Coast (CIV) [CIV]	False
Croatia (CRO)	True
Cuba (CUB) [Z]	True
Cyprus (CYP)	False
...	
Sri Lanka (SRI) [SRI]	False
Sudan (SUD)	False
Suriname (SUR) [E]	True
Sweden (SWE) [Z]	True
Switzerland (SUI)	True
Syria (SYR)	True
Chinese Taipei (TPE) [TPE] [TPE2]	True

Tajikistan (TJK)	False
Tanzania (TAN) [TAN]	False
Thailand (THA)	True
Togo (TOG)	False
Tonga (TGA)	False
Trinidad and Tobago (TRI) [TRI]	True
Tunisia (TUN)	True
Turkey (TUR)	True
Uganda (UGA)	True
Ukraine (UKR)	True
United Arab Emirates (UAE)	True
United States (USA) [P] [Q] [R] [Z]	True
Uruguay (URU)	True
Uzbekistan (UZB)	True
Venezuela (VEN)	True
Vietnam (VIE)	False
Virgin Islands (ISV)	False
Yugoslavia (YUG) [YUG]	True
Independent Olympic Participants (IOP) [IOP]	False
Zambia (ZAM) [ZAM]	False
Zimbabwe (ZIM) [ZIM]	True
Mixed team (ZZX) [ZZX]	True
Totals	True

Name: Gold, Length: 147, dtype: bool

```
df['Gold']+df['Sivler']+df['Bronze']
```

Afghanistan (AFG)	2
Algeria (ALG)	15
Argentina (ARG)	70
Armenia (ARM)	12
Australasia (ANZ) [ANZ]	12
Australia (AUS) [AUS] [Z]	468
Austria (AUT)	86
Azerbaijan (AZE)	26
Bahamas (BAH)	12
Bahrain (BRN)	1
Barbados (BAR) [BAR]	1
Belarus (BLR)	75
Belgium (BEL)	142
Bermuda (BER)	1
Bohemia (BOH) [BOH] [Z]	4
Botswana (BOT)	1
Brazil (BRA)	108
British West Indies (BWI) [BWI]	2
Bulgaria (BUL) [H]	214
Burundi (BDI)	1
Cameroon (CMR)	5
Canada (CAN)	279
Chile (CHI) [I]	13
China (CHN) [CHN]	473
Colombia (COL)	19
Costa Rica (CRC)	4
Ivory Coast (CIV) [CIV]	1
Croatia (CRO)	23

```

Cuba (CUB) [Z]                209
Cyprus (CYP)                   1
...
Sri Lanka (SRI) [SRI]         2
Sudan (SUD)                   1
Suriname (SUR) [E]           2
Sweden (SWE) [Z]             483
Switzerland (SUI)            185
Syria (SYR)                   3
Chinese Taipei (TPE) [TPE] [TPE2] 21
Tajikistan (TJK)              3
Tanzania (TAN) [TAN]          2
Thailand (THA)                24
Togo (TOG)                    1
Tonga (TGA)                   1
Trinidad and Tobago (TRI) [TRI] 18
Tunisia (TUN)                 10
Turkey (TUR)                  88
Uganda (UGA)                  7
Ukraine (UKR)                 115
United Arab Emirates (UAE)     1
United States (USA) [P] [Q] [R] [Z] 2399
Uruguay (URU)                 10
Uzbekistan (UZB)              20
Venezuela (VEN)              12
Vietnam (VIE)                 2
Virgin Islands (ISV)           1
Yugoslavia (YUG) [YUG]        83
Independent Olympic Participants (IOP) [IOP] 3
Zambia (ZAM) [ZAM]            2
Zimbabwe (ZIM) [ZIM]          8
Mixed team (ZZX) [ZZX]        17
Totals                        14714
Length: 147, dtype: int64

```

etc...

그 외에도 아래와 같은 코드가 있다. 간단히 설명하면 sdf에 df의 칼럼중 'Gold'인 칼럼 데이터가 15인 df의 칼럼을 할당한다. 그리고 sdf의 최대값과 갯수 등을 조회하는 코드이다.

```

sdf = df[(df['Gold'] == 15)]
sdf.max()
sdf.count()
sdf.argmax()

```