Excelファイルを読む

初めてのプログラミング

2020年度

只木進一（理工学部）

データサイエンスとPython

- データサイエンス
  - データ分析→客観的状況
  - 課題発見、施策立案
- Excelをpythonで利用
  - Excelファイルがたくさんだと？

Pandasパッケージ

- 表や列を扱う目的のパッケージ
  - 行や列のラベル指定が可能
- DataFrame
  - 表形式のデータ
- Series
  - 列形式のデータ
- 入出力

今日のサンプルファイル

- https://github.com/first-programming-saga/excelAndCSV

Excelファイルの読み込み

- pandas.read_excel()を利用してexcelを読む

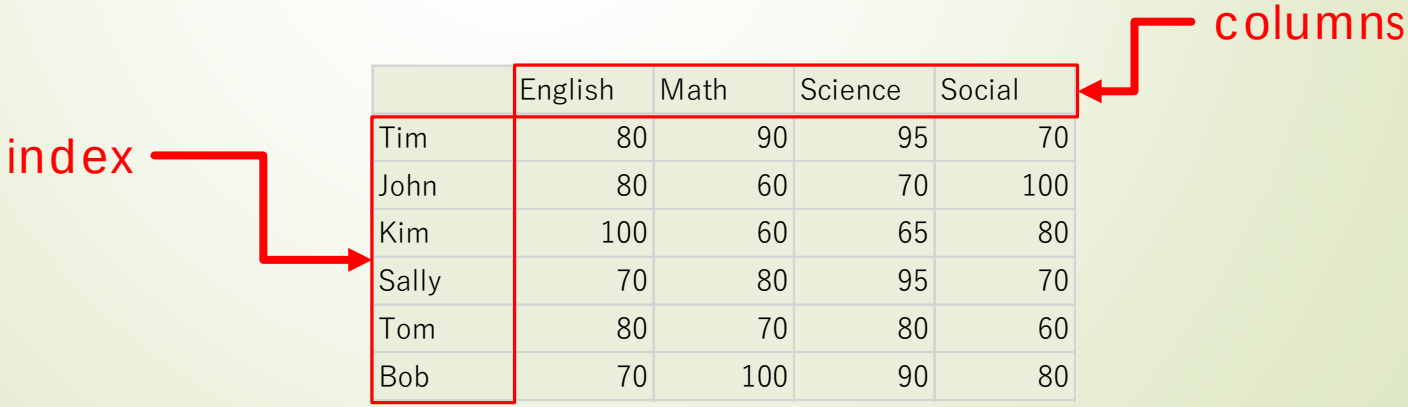
```
1. import pandas
2. with pandas.ExcelFile(filename) as f:
3.         data = pandas.read_excel(f)
```

- excelの内容はpandas.DataFrameクラス

excelAndCSV/howToUseDataFrame.ipynb

pandas.DataFrame

- 一行目をcolumnsとして識別する
  - DataFrame.columnsで参照
- 一列目をindexとして認識する
  - DataFrame.indexで参照



	English	Math	Science	Social
Tim	80	90	95	70
John	80	60	70	100
Kim	100	60	65	80
Sally	70	80	95	70
Tom	80	70	80	60
Bob	70	100	90	80

Columnの指定

- data[列名]：一列のデータ
  - pandas.Seriesクラス
- data[列名][行名]：指定位置のデータ

data['Math']

	English	Math	Science	Social
Tim	80	90	95	70
John	80	60	70	100
Kim	100	60	65	80
Sally	70	80	95	70
Tom	80	70	80	60
Bob	70	100	90	80

data['Social'] ['Sally']

Indexの指定

- data.loc[行名]：一行のデータ
  - pandas.Seriesクラス
- data.loc[行名][列名]：指定位置のデータ

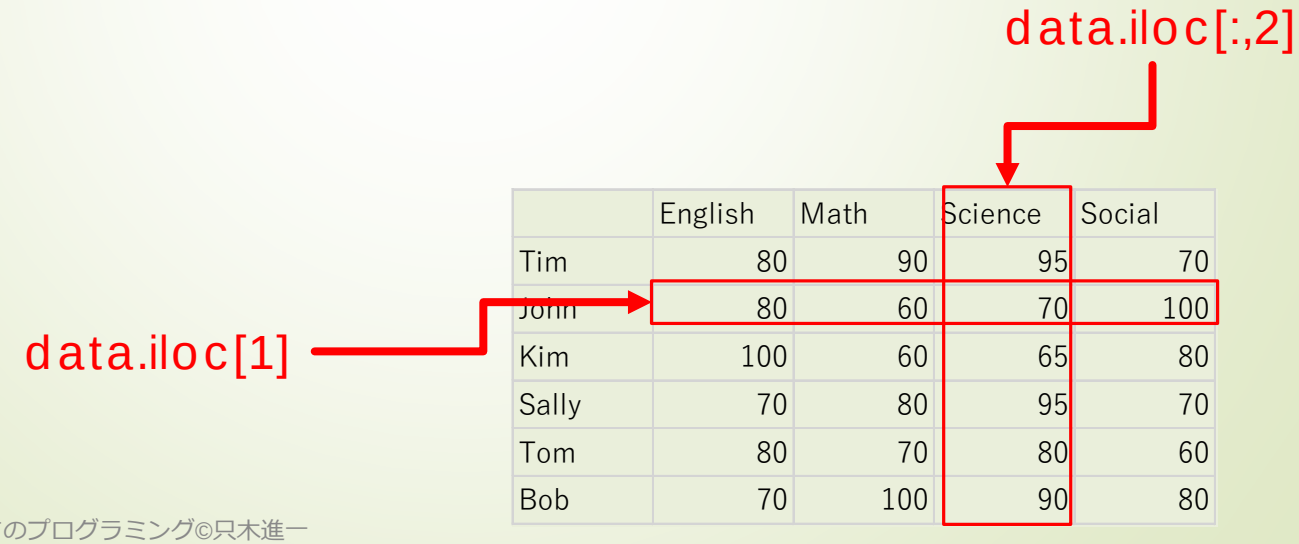
data.loc['John']

data.loc['Sally'] ['Math']

	English	Math	Science	Social
Tim	80	90	95	70
John	80	60	70	100
Kim	100	60	65	80
Sally	70	80	95	70
Tom	80	70	80	60
Bob	70	100	90	80

番号を使った指定

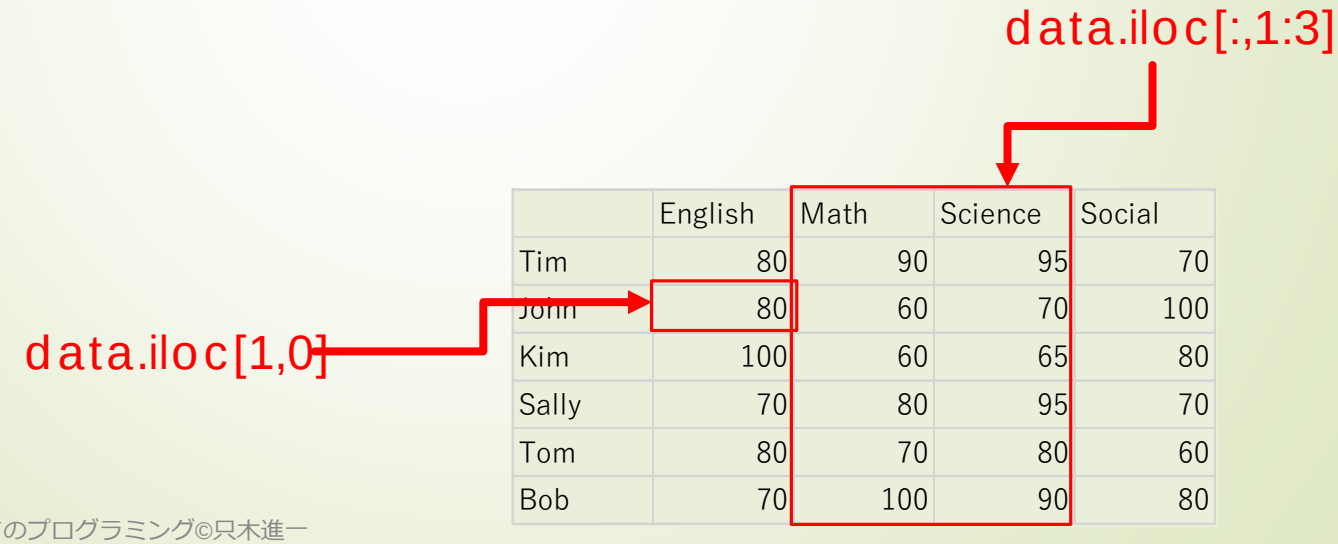
- data.iloc[整数]：一行のデータ
  - pandas.Seriesクラス
- data.iloc[:,整数]：一列



	English	Math	Science	Social
Tim	80	90	95	70
John	80	60	70	100
Kim	100	60	65	80
Sally	70	80	95	70
Tom	80	70	80	60
Bob	70	100	90	80

番号を使った指定

- data.iloc[整数,整数]：セル



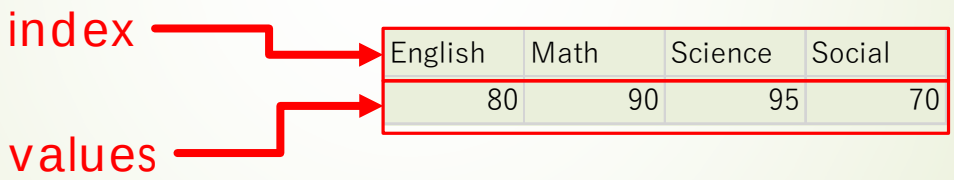
	English	Math	Science	Social
Tim	80	90	95	70
John	80	60	70	100
Kim	100	60	65	80
Sally	70	80	95	70
Tom	80	70	80	60
Bob	70	100	90	80

セルの値の変更

- セルを指定して代入
  - data[列名][行名] = 値
  - df.iloc[行番号][列番号] = 値

pandas.Series

- DataFrameから一行または一列取り出すとSeriesとなる
  - データの位置にindexが付く




```
for k in ser.index:
    v = ser[k]
    print(f'ser[{k}]:{v}')
```


excelAndCSV/howToUseDataFrame.ipynb

行を取り出す

In [4]:

```
1 print('Seriesの操作')
2 ser = data.loc['Tim']
3 for k in ser.index:
4     v = ser[k]
5     print(f'ser[{k}]:{v}')
```

```
Seriesの操作
ser[English]:80
ser[Math]:90
ser[Science]:95
ser[Social]:70
```

第一カラムにラベルを付ける

```
def useIndex(data):
    print('1列目にnameという名前を付ける')
    print('名前を付ける前のindex部のラベル')
    print(data.index.name)
    print('ラベルと付けた後')
    data.index.name='name'
    print(data)
    print()
```

	English	Math	Science	Social
name				
Tim	80	90	95	70
John	80	60	70	100
Kim	100	60	65	80
Sally	70	80	95	70
Tom	80	70	80	60

第一カラムにラベルを付ける

In [8]:

```
1 print("操作前")
2 print(data.index.name)
3 data.index.name = "name"
4 print("操作後")
5 print(data.index.name)
6 print(data)
```

```
操作前
None
操作後
name
```

name	English	Math	Science	Social
Tim	80	90	95	70
John	80	60	70	100
Kim	100	60	65	80
Sally	70	80	95	70
Tom	80	70	80	60
Bob	70	100	90	80

csvを読む

- csv (comma-separated values) ファイルはテキストファイル
- pandas.DataFrameとしても読むことができる

```
import pandas

with open(filename) as f:
    data = pandas.read_csv(f)
```

excelAndCSV/howToUseDataFrame3.ipynb

1カラムに名前がある場合

name	English	Math	Science	Social
Tim	80	90	95	70
John	80	60	70	100
Kim	100	60	65	80
Sally	70	80	95	70
Tom	80	70	80	60
Bob	70	100	90	80

- 番号のindexを付けられてしまう。

	name	English	Math	Science	Social
0	Tim	80	90	95	70
1	John	80	60	70	100
2	Kim	100	60	65	80
3	Sally	70	80	95	70
4	Tom	80	70	80	60
5	Bob	70	100	90	80

- set_index()で最初のカラムを指定
  - data.set_index(‘name’,inplace=True)