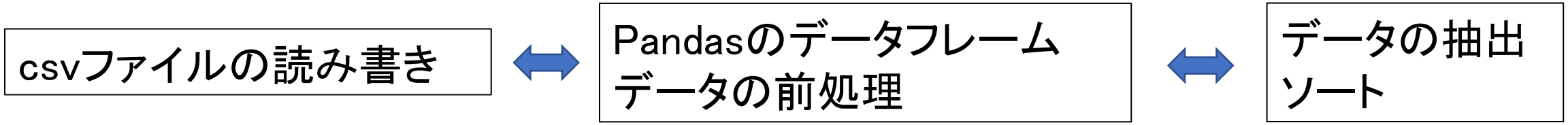
pandas

データの集計と解析のライブラリ



① pandasのインストール

Windows　Powershell等で＞の後にpy -m pip install pandasを入力

②Jupyter notebookを起動

```python
import numpy as np
arr1=['A']*5 + ['B']*5 +['C']*5+['D']*5+['E']*5+['F']*5+['G']*5+['H']*5+['I']*5+['J']*5
arr2=['国語','数学','英語','理科','社会']*10
arr3=[np.random.randint(0,100) for i in range(50)]
print(arr1)
print(arr2)
print(arr3)

import pandas as pd
data=list(zip(arr1,arr2,arr3))
df=pd.DataFrame(data=data,columns=['名前','教科','点数'])
df.info()
print()
df[:5]
```

リストデータを用意

データをまとめる

data=list(zip(arr1,arr2,arr3))

DataFrameを作成

変数＝DataFrame(data=ﾃﾞｰﾀ、columns=列名)

最初の5行だけ表示

```
['A', 'A', 'A', 'A', 'A', 'B', 'B', 'B', 'B', 'B', 'C', 'C', 'C', 'C', 'C', 'D', 'D', 'D', 'D', 'D', 'E', 'E',
'E', 'E', 'E', 'F', 'F', 'F', 'F', 'F', 'G', 'G', 'G', 'G', 'G', 'H', 'H', 'H', 'H', 'H', 'I', 'I', 'I', 'I',
'I', 'J', 'J', 'J', 'J', 'J']
['国語', '数学', '英語', '理科', '社会', '国語', '数学', '英語', '理科', '社会', '国語', '数学', '英語', '理科',
'社会', '国語', '数学', '英語', '理科', '社会', '国語', '数学', '英語', '理科', '社会', '国語', '数学', '英語',
'理科', '社会', '国語', '数学', '英語', '理科', '社会', '国語', '数学', '英語', '理科', '社会', '国語', '数学',
'英語', '理科', '社会', '国語', '数学', '英語', '理科', '社会']
[8, 47, 63, 32, 62, 24, 20, 96, 53, 63, 53, 4, 44, 38, 71, 45, 61, 5, 97, 86, 44, 79, 84, 96, 99, 18, 59, 43, 7
6, 85, 90, 8, 77, 26, 0, 59, 72, 98, 97, 51, 28, 32, 21, 6, 67, 11, 18, 91, 94, 79]
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50 entries, 0 to 49
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   名前      50 non-null     object
 1   教科      50 non-null     object
 2   点数      50 non-null     int64
dtypes: int64(1), object(2)
memory usage: 1.3+ KB
```

DataFrameの情報出力　df.info

	名前	教科	点数
0	A	国語	8
1	A	数学	47
2	A	英語	63
3	A	理科	32
4	A	社会	62

Print(df[:5])とするとテーブルにならずテキスト表示
テーブルにするには最後にdfを実行

```
df['点数'][:10]
```

```
0    77
1    73
2    15
3    43
4    98
5    87
6    35
7    35
8     5
9    32
Name: 点数, dtype: int64
```

```
df['総合']=[np.random.randint(0,100)for i in range(50)]
df[:5]
```

	名前	教科	点数	総合
0	A	国語	77	82
1	A	数学	73	81
2	A	英語	15	85
3	A	理科	43	52
4	A	社会	98	11

列の追加

列の削除

del df[‘文字’]

行の追加

変数＝[DataFrame].append([DataFrame])

```
new_data=[('K','国語',95,89),('K','数学',87,71),('K','英語',69,93),('K','理科',73,52),('K','社会',54,60)]
new_df=pd.DataFrame(data=new_data,columns=['氏名','教科','中間','期末'])
df2=df.append(new_df,ignore_index=True)
df2[45:]
```

←元のインデックスを無視

	名前	教科	点数	総合
0	A	国語	77	82
1	A	数学	73	81
2	A	英語	15	85
3	A	理科	43	52
4	A	社会	98	11

	氏名	教科	中間	期末
45	J	国語	62	92
46	J	数学	94	53
47	J	英語	46	94
48	J	理科	24	3
49	J	社会	97	84
50	K	国語	95	89
51	K	数学	87	71
52	K	英語	69	93
53	K	理科	73	52
54	K	社会	54	60

行に追加データ

loc[row+数字]=[追加データ]

```
(rows,cols)=df2.shape
df2.loc[rows]=['L','国語',15,20]
df2.loc[rows+1]=['L','数学',25,30]
df2.loc[rows+2]=['L','英語',35,40]
df2.loc[rows+3]=['L','理科',45,50]
df2.loc[rows+4]=['L','社会',55,60]
df2.tail(10)
```

最後のデータ10行分表示

頭の10行は、df2.head(10)

	氏名	教科	中間	期末
50	K	国語	95	89
51	K	数学	87	71
52	K	英語	69	93
53	K	理科	73	52
54	K	社会	54	60
55	L	国語	15	20
56	L	数学	25	30
57	L	英語	35	40
58	L	理科	45	50
59	L	社会	55	60

```
df2.index=list('abcdefghijklmnopqrstuvwxyzABCDEFGHIJKLMNOPQRSTUVWXYZ12345678')
df2.head(10)
```

	氏名	教科	中間	期末
0	A	国語	51	88
1	A	数学	43	11
2	A	英語	47	15
3	A	理科	64	1
4	A	社会	69	51
5	B	国語	91	73
6	B	数学	7	23
7	B	英語	30	54
8	B	理科	93	96
9	B	社会	27	51

	氏名	教科	中間	期末
a	A	国語	51	88
b	A	数学	43	11
c	A	英語	47	15
d	A	理科	64	1
e	A	社会	69	51
f	B	国語	91	73
g	B	数学	7	23
h	B	英語	30	54
i	B	理科	93	96
j	B	社会	27	51

```
df2.iloc[23:28]
```

23
24
25
26
27
28

```
df2.loc['x':'B']
```

	氏名	教科	中間	期末
x	E	理科	20	29
y	E	社会	38	20
z	F	国語	23	15
A	F	数学	79	33
B	F	英語	53	56

locで指定した後で「. T」を使う

	氏名	教科	中間	期末
x	E	理科	20	29
y	E	社会	38	20
z	F	国語	23	15
A	F	数学	79	33
B	F	英語	53	56

```
df2.loc['x':'B'].T
```

	x	y	z	A	B
氏名	E	E	F	F	F
教科	理科	社会	国語	数学	英語
中間	20	38	23	79	53
期末	29	20	15	33	56

```
df2.T.loc['x':'B']
```

表示されない

```
df2.T.iloc[0:5]
```

	a	b	c	d	e	f	g	h	i	j	...	Y	Z	1	2	3	4	5	6	7	8
氏名	A	A	A	A	A	B	B	B	B	B	...	K	K	K	K	K	L	L	L	L	L
教科	国語	数学	英語	理科	社会	国語	数学	英語	理科	社会	...	国語	数学	英語	理科	社会	国語	数学	英語	理科	社会
中間	51	43	47	64	69	91	7	30	93	27	...	95	87	69	73	54	15	25	35	45	55
期末	88	11	15	1	51	73	23	54	96	51	...	89	71	93	52	60	20	30	40	50	60

データの取得　[DataFrame]['列名'],unique()

```
print(df2['氏名'].unique())
print()
print(df2['教科'].unique())
```

```
['A' 'B' 'C' 'D' 'E' 'F' 'G' 'H' 'I' 'J' 'K' 'L']

['国語' '数学' '英語' '理科' '社会']
```

```
print(df2['中間'][:5])
print()
print((df2['中間']*2)[:5])
print()
print((df2['中間']//10)[:5])
```

	氏名	教科	中間	期末
a	A	国語	43	44
b	A	数学	90	55
c	A	英語	58	39
d	A	理科	12	75
e	A	社会	0	59
f	B	国語	22	6
g	B	数学	6	87
h	B	英語	48	57
i	B	理科	18	81
j	B	社会	95	75

```
a    43
b    90
c    58
d    12
e     0
Name: 中間, dtype: int64
```

2倍

```
a    86
b   180
c   116
d    24
e     0
Name: 中間, dtype: int64
```

$\frac{1}{10}$

```
a    4
b    9
c    5
d    1
e    0
Name: 中間, dtype: int64
```

統計

```
print('合計:%s' % df2['中間'].sum())
print('平均:%s' % df2['中間'].mean())
print('中央:%s' % df2['中間'].median())
print('最小:%s' % df2['中間'].min())
print('最大:%s' % df2['中間'].max())
```

```
合計:2776
平均:46.266666666666666
中央:44.0
最小:0
最大:97
```

合計　[Series].sum
平均　[Series].mean
中央　[Series].median
最小　[Series].min
最大　[Series].max

並べ替え（基のデータは変わらない）

```
df2.sort_values(['期末'],ascending=False)[:10]
```

変数=[DataFrame].sort_values([列],ascending=真偽値)

True　→　昇順
False　→　降順

falseのように小文字では動かない

	氏名	教科	中間	期末
m	C	英語	2	97
1	K	英語	69	93
G	G	英語	65	92
x	E	理科	54	92
r	D	英語	40	91
o	C	社会	74	89
Y	K	国語	95	89
q	D	数学	78	87
g	B	数学	6	87
l	G	社会	8	84

グループ

‘中間’、‘期末’と並べると、中間を優先して
並べ、同じ値がある場合は期末を考慮します

```
srt=df2.sort_values(['中間','期末'],ascending=False)
grp=srt.groupby('教科')
grp.first()
```

first　最初のデータ表示
head
last　最後のデータ表示
tail

変数=[DataFrame].groupby(列名)

教科	氏名	中間	期末
国語	K	95	89
数学	C	97	16
理科	D	85	21
社会	B	95	75
英語	J	96	31

```
grp.mean()
```

教科	中間	期末
国語	48.166667	37.250000
数学	49.416667	53.916667
理科	42.166667	44.750000
社会	43.083333	57.833333
英語	48.500000	61.166667

合計　[Groupby].sum ()
平均　[Groupby].mean ()
中央　[Groupby].median ()
最小　[Groupby].min ()
最大　[Groupby].max ()
分散　[Groupby].var ()
標準偏差　[Groupby].std ()
データ数　[Groupby].count()

各々の氏名の点数を合計し、中間の点数順に並べる

```
grp2=srt.groupby('氏名')
grp2.sum().sort_values(['中間','期末'],ascending=False)
```

氏名	中間	期末
K	378	365
D	366	229
J	287	217
E	259	276
F	231	118
G	223	283
C	208	321
A	203	272
B	189	306
L	175	200
I	173	279
H	84	193

教科	氏名	中間	期末
国語	K	95	89
数学	C	97	16
理科	D	85	21
社会	B	95	75
英語	J	96	31

```
grp.agg(['sum','mean','median','min','max'])
```

	中間					期末				
	sum	mean	median	min	max	sum	mean	median	min	max
教科										
国語	578	48.166667	41.0	15	95	447	37.250000	39.5	2	89
数学	593	49.416667	41.0	6	97	647	53.916667	54.5	16	87
理科	506	42.166667	42.0	2	85	537	44.750000	44.5	8	92
社会	517	43.083333	43.5	0	95	694	57.833333	60.0	5	89
英語	582	48.500000	48.0	2	96	734	61.166667	61.0	7	97

特定のグループを取り出す

```
grp2=srt.groupby('氏名')
grp2.get_group('A').sort_values(['中間'],ascending=False)
```

変数=[Groupby].get_group(ｸﾞﾙｰﾌﾟ名)

	氏名	教科	中間	期末
b	A	数学	90	55
c	A	英語	58	39
a	A	国語	43	44
d	A	理科	12	75
e	A	社会	0	59

中間、期末いずれも80以上

```
df2.query('中間 >= 80 & 期末 >=80')
```

	氏名	教科	中間	期末
Y	K	国語	95	89

中間と期末の合計が170以上

```
df2.query('中間+期末 >=170')
```

	氏名	教科	中間	期末
j	B	社会	95	75
Y	K	国語	95	89

ピボットテーブル

```
df2.pivot_table(values='中間',index=['氏名'],columns=['教科'])
```

変数=[DataFrame].pivot_table(values=[列],index=[列],columns=[列])

教科	国語	数学	理科	社会	英語
氏名					
A	43	90	12	0	58
B	22	6	18	95	48
C	21	97	14	74	2
D	68	78	85	95	40
E	56	32	54	69	48
F	94	10	2	33	92
G	91	47	12	8	65
H	17	22	39	2	4
I	17	35	72	24	25
J	39	64	80	8	96
K	95	87	73	54	69
L	15	25	45	55	35