7. 集約・フィルタ・変換のためのグループ分け（P146）

In [47]:
```
import pandas as pd
import pathlib
import os
import numpy as np
```

- **データを独立なグループに分割して、各グループごとに計算をする**
- 「分割、適用、結合」と呼ぶ
- グループの分け方（グループ分けカラム）
  1. `df.groupby(['list', 'of', 'grouping', 'columns'])`
  2. `df.groupby('single_column')`
- 戻り値：groupbyオブジェクト

レシピ53 集約の定義（P147）

もっとも応用が利く書き方はコレ、他の用法は「★（補足）集約パターン」を参照

`df.groupby ( グループ分けカラム ) .agg { 集約カラム： 集約関数 }`

1. **グループ分けカラム**：データを独立なグループに分割して、各グループごとに計算をする単位
2. **集約カラム**：そのカラムを使って値が集約される
3. **集約関数**：集約をどのように行うか
   - sum, min, max, mean, count（件数、NaN除く）, size（件数、NaN含む）, variance, std,etc

- このレシピでは、航空便データセットを調べて、1つのグループ分けカラム、1つの集約カラム、1つの集約関数、といった簡単な集約を行い
- 各航空会社の平均遅延を求める

（1）航空便のデータセットを読み込み、グループ分けカラム（AIRLINE）、集約カラム（ARR_DELAY）、集約関数（mean）で処理する

In [48]:
```
flights = pd.read_csv('flights.csv')
flights.head()
```

Out[48]:

	MONTH	DAY	WEEKDAY	AIRLINE	ORG_AIR	DEST_AIR	SCHED_DEP	DEP_DELAY	AIR_TIME	DIST	SCHED_ARR	ARR_DELAY	DIVERTED	CANCELLED
0	1	1	4	WN	LAX	SLC	1625	58.0	94.0	590	1905	65.0	0	0
1	1	1	4	UA	DEN	IAD	823	7.0	154.0	1452	1333	-13.0	0	0
2	1	1	4	MQ	DFW	VPS	1305	36.0	85.0	641	1453	35.0	0	0
3	1	1	4	AA	DFW	DCA	1555	7.0	126.0	1192	1935	-7.0	0	0
4	1	1	4	WN	LAX	MCI	1720	48.0	166.0	1363	2225	39.0	0	0

（2）groupbyメソッドにグループ分けカラムを渡し、aggメソッドに集約カラムと集約関数を**dict**で渡す

- mean：平均

In [49]:
```
flights.groupby('AIRLINE').agg({'ARR_DELAY':'mean'}).head()
#             グループ分けカラム                    集約カラム    集約関数
```

Out[49]:

	ARR_DELAY
AIRLINE	
AA	5.542661
AS	-0.833333
B6	8.692593
DL	0.339691
EV	7.034580

（3）（2）の別のやり方

In [50]:
```
flights.groupby('AIRLINE')['ARR_DELAY'].agg('mean').head()
#             グループ分けカラム     集約カラム      集約関数
```

Out[50]:
```
AIRLINE
AA    5.542661
AS   -0.833333
B6    8.692593
DL    0.339691
EV    7.034580
Name: ARR_DELAY, dtype: float64
```

（4）（3）のやり方だと、aggにいろんな関数を渡すことができる...?

In [51]:
```
flights.groupby('AIRLINE')['ARR_DELAY'].agg(np.mean).head()
```

Out[51]:
```
AIRLINE
AA    5.542661
AS   -0.833333
B6    8.692593
DL    0.339691
EV    7.034580
Name: ARR_DELAY, dtype: float64
```

（5）（3）の場合、aggメソッドを省略しても可能

- 集約関数が1つのときだけ使える

In [52]:
```
flights.groupby('AIRLINE')['ARR_DELAY'].mean().head()
```

Out[52]:
```
AIRLINE
AA    5.542661
AS   -0.833333
B6    8.692593
DL    0.339691
EV    7.034580
Name: ARR_DELAY, dtype: float64
```
