レシピ49 別のDataFrameからカラムを追加（P132）

- 別のDataFrameやSeriesから新たなカラムを追加する場合
- まずインデックスをアラインメントしてから新たなカラムが作られる

- このレシピでは、従業員データセットにその従業員の部門の最高給与のカラムを追加する

（1）employeeデータセットから必要なカラムを選ぶ

In [2]:
```
import pandas as pd
employee = pd.read_csv('employee.csv')
dept_sal = employee[['DEPARTMENT', 'BASE_SALARY']]
dept_sal.head()
```

Out[2]:

	DEPARTMENT	BASE_SALARY
0	Municipal Courts Department	121862.0
1	Library	26125.0
2	Houston Police Department-HPD	45279.0
3	Houston Fire Department (HFD)	63166.0
4	General Services Department	56347.0

In [25]:
```
len(dept_sal)
```

Out[25]:
```
2000
```

（2）より小さなDataFrameである部門内で給与をソートする

- 部門別：DEPARTMENTを昇順、給与を降順（大きい順）にソート

In [8]:
```
dept_sal = dept_sal.sort_values(['DEPARTMENT', 'BASE_SALARY'],
                                ascending=[True, False])
dept_sal
```

Out[8]:

	DEPARTMENT	BASE_SALARY
1494	Admn. & Regulatory Affairs	140416.0
237	Admn. & Regulatory Affairs	130416.0
1679	Admn. & Regulatory Affairs	103776.0
988	Admn. & Regulatory Affairs	72741.0
693	Admn. & Regulatory Affairs	66825.0
...	...	...
1140	Solid Waste Management	30410.0
1243	Solid Waste Management	30410.0
387	Solid Waste Management	28829.0
57	Solid Waste Management	27622.0
536	Solid Waste Management	26499.0

2000 rows × 2 columns

（3）

`drop_duplicates` メソッド

- `df.drop_duplicates(keep=first)`
- デフォルトでは引数 `keep='first'` のため、先頭以外の重複を除外する（参考）

- 各部門の先頭行を残す

- 部門ごとに給与を高い順に並び替えているため、先頭だけを残す

In [10]:
```
max_dept_sal = dept_sal.drop_duplicates(subset='DEPARTMENT')
#                                        列 DEPARTMENTに対し、先頭以外の重複行を削除する
max_dept_sal.head()
```

Out[10]:

	DEPARTMENT	BASE_SALARY
1494	Admn. & Regulatory Affairs	140416.0
149	City Controller's Office	64251.0
236	City Council	100000.0
647	Convention and Entertainment	38397.0
1500	Dept of Neighborhoods (DON)	89221.0

（4）DEPARTMENTカラムをそれぞれのDataFrameのインデックスにする

In [11]:
```
max_dept_sal = max_dept_sal.set_index('DEPARTMENT')
employee = employee.set_index(('DEPARTMENT'))
```

In [31]:
```
# cf
print(f'max_dept_sal: {len(max_dept_sal)}')
print(f'employee: {len(employee)}')
```

```
max_dept_sal: 24
employee: 2000
```

In [30]:
```
# cf.
max_dept_sal['BASE_SALARY']
```

Out[30]:
```
DEPARTMENT
Admn. & Regulatory Affairs        140416.0
City Controller's Office           64251.0
City Council                      100000.0
Convention and Entertainment       38397.0
Dept of Neighborhoods (DON)        89221.0
Finance                            96272.0
Fleet Management Department       125884.0
General Services Department        89194.0
Health & Human Services           180416.0
Housing and Community Devp.        98536.0
Houston Airport System (HAS)      186192.0
Houston Emergency Center (HEC)     84456.0
Houston Fire Department (HFD)     210588.0
Houston Information Tech Svcs     102019.0
Houston Police Department-HPD     199596.0
Human Resources Dept.             110547.0
Legal Department                  275000.0
Library                           107763.0
Mayor's Office                    120750.0
Municipal Courts Department       121862.0
Parks & Recreation                 85055.0
Planning & Development             68762.0
Public Works & Engineering-PWE    178331.0
Solid Waste Management            110005.0
Name: BASE_SALARY, dtype: float64
```

（5）インデックスをセットしたので、employee DataFrameに新たなカラムを追加する

このあたりから理解不足▼

- 左辺のDataFrameのemployeeが、右辺のDataFrameのmax_dept_salのインデックスと
- **1対1対応でアラインメントするからうまくいく** → インデックスにセットしている点がポイントか？？？
- もしmax_dept_salで、インデックスの部門に重複があれば、処理は失敗となる（→解説参照）

In [33]:
```
employee['MAX_DEPT_SALARY'] = max_dept_sal['BASE_SALARY']
# 件数2000                      件数：24  だけど、インデックスが1対1だから埋まるらしい。。。。。

# 見にくいので3列目に移動
move_col = employee.pop('MAX_DEPT_SALARY')
employee.insert(3, 'MAX_DEPT_SALARY', move_col)

employee
```

Out[33]:

	UNIQUE_ID	POSITION_TITLE	BASE_SALARY	MAX_DEPT_SALARY	RACE	EMPLOYMENT_TYPE	GENDER	EMPLOYMENT_STATUS	HIRE_DATE	JOB_DATE	MAX_SALARY2
DEPARTMENT											
Municipal Courts Department	0	ASSISTANT DIRECTOR (EX LVL)	121862.0	121862.0	Hispanic/Latino	Full Time	Female	Active	2006-06-12	2012-10-13	NaN
Library	1	LIBRARY ASSISTANT	26125.0	107763.0	Hispanic/Latino	Full Time	Female	Active	2000-07-19	2010-09-18	NaN
Houston Police Department-HPD	2	POLICE OFFICER	45279.0	199596.0	White	Full Time	Male	Active	2015-02-03	2015-02-03	NaN
Houston Fire Department (HFD)	3	ENGINEER/OPERATOR	63166.0	210588.0	White	Full Time	Male	Active	1982-02-08	1991-05-25	NaN
General Services Department	4	ELECTRICIAN	56347.0	89194.0	White	Full Time	Male	Active	1989-06-19	1994-10-22	NaN
...	...	...	...	...	...	...	...	...	...	...	...
Houston Police Department-HPD	1995	POLICE OFFICER	43443.0	199596.0	White	Full Time	Male	Active	2014-06-09	2015-06-09	NaN
Houston Fire Department (HFD)	1996	COMMUNICATIONS CAPTAIN	66523.0	210588.0	Black or African American	Full Time	Male	Active	2003-09-02	2013-10-06	NaN
Houston Police Department-HPD	1997	POLICE OFFICER	43443.0	199596.0	White	Full Time	Male	Active	2014-10-13	2015-10-13	NaN
Houston Police Department-HPD	1998	POLICE OFFICER	55461.0	199596.0	Asian/Pacific Islander	Full Time	Male	Active	2009-01-20	2011-07-02	NaN
Houston Fire Department (HFD)	1999	FIRE FIGHTER	51194.0	210588.0	Hispanic/Latino	Full Time	Male	Active	2009-01-12	2010-07-12	NaN

2000 rows × 11 columns

In [32]:
```
# cf.
len(employee)
```

Out[32]:
```
2000
```

（6）BASE_SALARYがMAX_DEPT_SALARYより大きな行があるか `query` でチェックする→なし

In [18]:
```
employee.query('BASE_SALARY > MAX_DEPT_SALARY')
```

Out[18]:

	UNIQUE_ID	POSITION_TITLE	BASE_SALARY	MAX_DEPT_SALARY	RACE	EMPLOYMENT_TYPE	GENDER	EMPLOYMENT_STATUS	HIRE_DATE	JOB_DATE
DEPARTMENT										

解説

- 右辺のDataFrameのインデックス値に重複があると何が起こるか？
- DataFrameのsampleメソッドで10行を非復元無作為抽出する

In [21]:
```
import numpy as np
np.random.seed(1234)
random_salary = dept_sal.sample(n=10).set_index('DEPARTMENT')
random_salary
```

Out[21]:

	BASE_SALARY
DEPARTMENT	
Public Works & Engineering-PWE	50586.0
Houston Police Department-HPD	66614.0
Houston Police Department-HPD	66614.0
Housing and Community Devp.	78853.0
Houston Police Department-HPD	66614.0
Parks & Recreation	NaN
Public Works & Engineering-PWE	37211.0
Public Works & Engineering-PWE	54683.0
Human Resources Dept.	58474.0
Health & Human Services	47050.0

- インデックスの部門にいくつか重複が出ている
- 新たなカラムを追加しようとすると、重複が出てエラーが発生する
- employee DataFrameの少なくとも1つのインデックスラベルがrandom_salaryの複数の
- インデックスラベルとジョインしようとした

In [ ]:
```
employee['RANDOM_SALARY'] = random_salary['BASE_SALARY']
```

```
ValueError: cannot reindex from a duplicate axis
```

補足

- 等式の左辺の全インデックスが合致する必要はないが、1つは合致しないといけない
- 左辺のDataFrameのインデックスのどれもアラインメントする相手がないと、結果は欠損値になる
- そうなる例を示す
- max_dept_sal Seriesの先頭3行だけを用いて新たなカラムを追加してみる

In [24]:
```
employee['MAX_SALARY2'] = max_dept_sal['BASE_SALARY'].head(3)
employee.MAX_SALARY2.value_counts()
```

Out[24]:
```
140416.0    29
100000.0    11
64251.0      5
Name: MAX_SALARY2, dtype: int64
```

- 演算は成功したが、3部門の給与しか値が埋められない
- max_dept_sal Seriesの先頭3行以外の部門は欠損値になる
