

レシピ39 SQLのWHERE節の翻訳（P102）

2022/02/18

- SQLを使ってデータを選択、フィルタリング、集約、ソートするにはSELECT文を使う
- SQLのSELECT文では、WHERE節がデータのフィルタリングによく使われる
- pandasにはデータベースにコネクしてSQL文を送る機能がある

- このレシピでは、従業員データセットの部分集合を選択するSQLクエリに等しいpandasコードを書く
- 警察署が消防署で働いている基本給が8万～12万ドルの女性従業員を抽出してみる

In [] :

```
# <<SQLの場合>>

# SELECT
#     UNIQUE_ID,
#     DEPARTMENT,
#     GENDER,
#     BASE_SALARY
# FROM
#     EMPLOYEE
# WHERE
#     DEPARTMENT IN ( 'Houston Plice Department-HPD',
#                     'Houston Fire Department (HFD)' ) AND
#     GENDER = 'Female' AND
#     BASE_SALARY BETWEEN 80000 AND 120000;
```

(1) employeeデータセットを読み込む

In [2] :

```
import pandas as pd
employee = pd.read_csv('employee.csv')
employee.head()
```

Out[2] :

	UNIQUE_ID	POSITION_TITLE	DEPARTMENT	BASE_SALARY	RACE	EMPLOYMENT_TYPE	GENDER	EMPLOYMENT_STATUS	HIRE_DATE	JOB_DATE
0	0	ASSISTANT DIRECTOR (EX LVL)	Municipal Courts Department	121862.0	Hispanic/Latino	Full Time	Female	Active	2006-06-12	2012-10-13
1	1	LIBRARY ASSISTANT	Library	26125.0	Hispanic/Latino	Full Time	Female	Active	2000-07-19	2010-09-18
2	2	POLICE OFFICER	Houston Police Department-HPD	45279.0	White	Full Time	Male	Active	2015-02-03	2015-02-03
3	3	ENGINEER/OPERATOR	Houston Fire Department (HFD)	63166.0	White	Full Time	Male	Active	1982-02-08	1991-05-25
4	4	ELECTRICIAN	General Services Department	56347.0	White	Full Time	Male	Active	1989-06-19	1994-10-22

(2) データセットをフィルタリングする前に、対象カラムを自分の目で調べてフィルタリングする正確な値がどうなっているかを確認する

value_counts()

- ある列の要素ごとの数
- 多い順（降順）でから並ぶ

In [5] :

```
# 列「DEPARTMENT」
employee.DEPARTMENT.value_counts()
```

Out[5] :

```
Houston Police Department-HPD    638
Houston Fire Department (HFD)     384
Public Works & Engineering-PWE   343
Health & Human Services           110
Houston Airport System (HAS)      106
Parks & Recreation                 74
Solid Waste Management            43
Library                           36
Fleet Management Department       36
Admn. & Regulatory Affairs         29
Municipal Courts Department       28
Human Resources Dept.             24
Houston Emergency Center (HEC)    23
Housing and Community Devp.       22
General Services Department        22
Dept of Neighborhoods (DON)       17
Legal Department                  17
City Council                      11
Finance                           10
Houston Information Tech Svcs      9
Planning & Development             7
City Controller's Office           5
Mayor's Office                    5
Convention and Entertainment       1
Name: DEPARTMENT, dtype: int64
```

In [7] :

```
# 列「GENDER」
employee.GENDER.value_counts()
```

Out[7] :

```
Male      1397
Female     603
Name: GENDER, dtype: int64
```

In [16] :

```
# 列「BASE_SALARY」はdescribeで。intにするとわかりやすい
print(employee.BASE_SALARY.describe(),'\n')
print(employee.BASE_SALARY.describe().astype(int))
```

```
count      1886.000000
mean       55767.931601
std        21693.706679
min        24960.000000
25%        40170.000000
50%        54461.000000
75%        66614.000000
max        275000.000000
Name: BASE_SALARY, dtype: float64

count      1886
mean       55767
std        21693
min        24960
25%        40170
50%        54461
75%        66614
max        275000
Name: BASE_SALARY, dtype: int64
```

In [4] :

```
employee.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 10 columns):
#   Column              Non-Null Count  Dtype
---  ---
0   UNIQUE_ID           2000 non-null  int64
1   POSITION_TITLE        2000 non-null  object
2   DEPARTMENT           2000 non-null  object
3   BASE_SALARY          1886 non-null  float64
4   RACE                 1965 non-null  object
5   EMPLOYMENT_TYPE      2000 non-null  object
6   GENDER               2000 non-null  object
7   EMPLOYMENT_STATUS    2000 non-null  object
8   HIRE_DATE            2000 non-null  object
9   JOB_DATE             1997 non-null  object
dtypes: float64(1), int64(1), object(8)
memory usage: 156.4+ KB
```

(3) フィルタリング基準を作成する。isinメソッドを使って、多くの値の一つと等しいかチェックする

isin()

- ここでは、booleanインデックス法やQueryを使うよりシンプルだが..

In [20] :

```
depts = ['Houston Police Department-HPD', 'Houston Fire Department (HFD)']

criteria_dept = employee.DEPARTMENT.isin(depts) # Boolean
criteria_gender = employee.GENDER == 'Female'
criteria_sal = (employee.BASE_SALARY >= 80000) & (employee.BASE_SALARY <= 120000)
```

between(A,B)

- 上のcriteria_salは以下のような書き方もできる

In [27] :

```
criteria_sal = employee.BASE_SALARY.between(80000, 120000)
```

(4) すべての条件を組み合わせる

- たしかにこんな風にまとめたほうがわかりやすいな...

In [23] :

```
criteria_final = (criteria_dept & criteria_gender & criteria_sal)
```

(5) Booleanインデックス法を使って、最終的なフィルタリング基準に合致する行だけを選ぶ

In [25] :

```
select_columns = ['UNIQUE_ID', 'DEPARTMENT', 'GENDER', 'BASE_SALARY']
employee.loc[criteria_final, select_columns]
```

Out[25] :

	UNIQUE_ID	DEPARTMENT	GENDER	BASE_SALARY
61	61	Houston Fire Department (HFD)	Female	96668.0
136	136	Houston Police Department-HPD	Female	81239.0
367	367	Houston Police Department-HPD	Female	86534.0
474	474	Houston Police Department-HPD	Female	91181.0
513	513	Houston Police Department-HPD	Female	81239.0
564	564	Houston Police Department-HPD	Female	81239.0
595	595	Houston Police Department-HPD	Female	99953.0
937	937	Houston Police Department-HPD	Female	81239.0
954	954	Houston Police Department-HPD	Female	91181.0
1118	1118	Houston Police Department-HPD	Female	104455.0
1519	1519	Houston Police Department-HPD	Female	91181.0
1565	1565	Houston Police Department-HPD	Female	81239.0
1622	1622	Houston Police Department-HPD	Female	91181.0

(補足)

- 「もっとも人気の多い上記5位部門を除く」場合
- ずばらしい...

In [28] :

```
# value_countsで列「組織」の要素ごとの数が出る
top_5_depts = employee.DEPARTMENT.value_counts().index[:5]
top_5_depts
```

Out[28] :

```
Index(['Houston Police Department-HPD', 'Houston Fire Department (HFD)',
      'Public Works & Engineering-PWE', 'Health & Human Services',
      'Houston Airport System (HAS)'],
      dtype='object')
```

In [29] :

```
criteria = ~employee.DEPARTMENT.isin(top_5_depts)
employee[criteria]
```

Out[29] :

	UNIQUE_ID	POSITION_TITLE	DEPARTMENT	BASE_SALARY	RACE	EMPLOYMENT_TYPE	GENDER	EMPLOYMENT_STATUS	HIRE_DATE	JOB_DATE
0	0	ASSISTANT DIRECTOR (EX LVL)	Municipal Courts Department	121862.0	Hispanic/Latino	Full Time	Female	Active	2006-06-12	2012-10-13
1	1	LIBRARY ASSISTANT	Library	26125.0	Hispanic/Latino	Full Time	Female	Active	2000-07-19	2010-09-18
4	4	ELECTRICIAN	General Services Department	56347.0	White	Full Time	Male	Active	1989-06-19	1994-10-22
18	18	MAINTENANCE MECHANIC III	General Services Department	40581.0	Hispanic/Latino	Full Time	Male	Active	2008-12-29	2008-12-29
32	32	SENIOR ACCOUNTANT	Finance	46963.0	Black or African American	Full Time	Male	Active	1991-02-11	2016-02-13
...	...	...	...	...	...	...	...	...	...	...
1976	1976	SENIOR SIDeloader OPERATOR	Solid Waste Management	34466.0	Hispanic/Latino	Full Time	Female	Active	2015-07-20	2016-01-30
1983	1983	ADMINISTRATIVE SPECIALIST	Admn. & Regulatory Affairs	55172.0	Black or African American	Full Time	Female	Active	2006-10-16	2006-10-16
1985	1985	TRUCK DRIVER	Fleet Management Department	30222.0	Black or African American	Full Time	Male	Active	2013-06-10	2015-08-01
1988	1988	SENIOR ASSISTANT CITY ATTORNEY II	Legal Department	124115.0	White	Full Time	Male	Active	2013-01-23	2013-03-02
1990	1990	BUILDING MAINTENANCE SUPERVISOR	Parks & Recreation	30347.0	Black or African American	Full Time	Female	Active	1995-10-14	2010-03-20

419 rows × 10 columns

あまりSQLは関係しなかった。

In [] :