

Data Analysis Tools with Pandas 2 - SF Salaries Exercise

แบบฝึกหัดนี้เป็นแบบฝึกหัดทดสอบทักษะการใช้งาน library pandas ด้วย [SF Salaries Dataset](#) จากเว็บไซต์ Kaggle ให้ทำตามคำสั่ง ต่อไปนี้

Import pandas as pd.

```
In [ ]: import pandas as pd
```

ให้นำเข้าข้อมูลจากไฟล์ Salaries.csv มาในรูปของ dataframe โดยตั้งชื่อตัวแปรว่า sal

```
In [ ]: sal = pd.read_csv('Salaries.csv')
```

Check the head of the DataFrame.

```
In [ ]: sal.head()
```

```
Out[ ]:   Id  EmployeeName  JobTitle  BasePay  OvertimePay  OtherPay  Benefits  Total
```

0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.00	400184.25	NaN	56759
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	NaN	53890
2	3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.60	NaN	33527
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.00	56120.71	198306.90	NaN	33234
4	5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT, (FIRE DEPARTMENT)	134401.60	9737.00	182234.59	NaN	32637

ใช้คำสั่ง .info() method to ในการดูภาพรวมของข้อมูลทั้งหมด

```
In [ ]: sal.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 148654 entries, 0 to 148653
Data columns (total 13 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Id                    148654 non-null  int64
1   EmployeeName          148654 non-null  object
2   JobTitle              148654 non-null  object
3   BasePay               148045 non-null  float64
4   OvertimePay           148650 non-null  float64
5   OtherPay              148650 non-null  float64
6   Benefits              112491 non-null  float64
7   TotalPay              148654 non-null  float64
8   TotalPayBenefits      148654 non-null  float64
9   Year                  148654 non-null  int64
10  Notes                  0 non-null       float64
11  Agency                148654 non-null  object
12  Status                 0 non-null       float64
dtypes: float64(8), int64(2), object(3)
memory usage: 14.7+ MB
```

ลบคอลัมน์ Notes และ Status ออก

```
In [ ]: sal.drop(['Notes'],axis=1,inplace=True)
sal.drop(['Status'],axis=1,inplace=True)
```

```
In [ ]: sal
```

Out[]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Bene
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.00	400184.25	↑
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	↑
2	3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.60	↑
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.00	56120.71	198306.90	↑
4	5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT, (FIRE DEPARTMENT)	134401.60	9737.00	182234.59	↑
...	
148649	148650	Roy I Tillery	Custodian	0.00	0.00	0.00	
148650	148651	Not provided	Not provided	NaN	NaN	NaN	↑
148651	148652	Not provided	Not provided	NaN	NaN	NaN	↑
148652	148653	Not provided	Not provided	NaN	NaN	NaN	↑
148653	148654	Joe Lopez	Counselor, Log Cabin Ranch	0.00	0.00	-618.13	

148654 rows × 11 columns



หาค่าเฉลี่ยของ Benefits ใน sal

In []: `sal['Benefits'].mean()`

Out[]: 25007.893150829852

ใน sal แทน Benefits ที่เป็น null ด้วย 0

```
In [ ]: sal['Benefits'].fillna(value=0, inplace=True)
```

```
In [ ]: sal.head()
```

```
Out[ ]:
```

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	Tota
0	1	NATHANIEL FORD	GENERAL MANAGER- METROPOLITAN TRANSIT AUTHORITY	167411.18	0.00	400184.25	0.0	56759
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	0.0	53890
2	3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.60	0.0	33527
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.00	56120.71	198306.90	0.0	33234
4	5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT, (FIRE DEPARTMENT)	134401.60	9737.00	182234.59	0.0	32637

หาค่าเฉลี่ยของ Benefits ใน sal อีกครั้ง

```
In [ ]: sal['Benefits'].mean()
```

```
Out[ ]: 18924.23283887417
```

จงเพิ่มคอลัมน์ Year(TH) ใน sal ให้เป็นเลขปี พศ

```
In [ ]: sal['Years(TH)'] = sal['Year'] + 543
```

```
In [ ]: sal
```

Out[]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Bene
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.00	400184.25	
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	
2	3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.60	
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.00	56120.71	198306.90	
4	5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT, (FIRE DEPARTMENT)	134401.60	9737.00	182234.59	
...
148649	148650	Roy I Tillery	Custodian	0.00	0.00	0.00	
148650	148651	Not provided	Not provided	NaN	NaN	NaN	
148651	148652	Not provided	Not provided	NaN	NaN	NaN	
148652	148653	Not provided	Not provided	NaN	NaN	NaN	
148653	148654	Joe Lopez	Counselor, Log Cabin Ranch	0.00	0.00	-618.13	

148654 rows × 12 columns



จงเพิ่มคอลัมน์ Level มีค่าเป็น L เมื่อ TotalPayBenefits น้อยกว่า 1 แสน และเป็น H เมื่อมากกว่าเท่ากับ 1 แสน

In []:

```
def LH(x):
    if x < 100000 :
        return 'L'
    else :
        return 'H'
```

```
In [ ]: sal['Level'] = sal['TotalPayBenefits'].apply(LH)
```

```
In [ ]: sal
```

```
Out[ ]:
```

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Bene
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.00	400184.25	
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	
2	3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.60	
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.00	56120.71	198306.90	
4	5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT, (FIRE DEPARTMENT)	134401.60	9737.00	182234.59	
...
148649	148650	Roy I Tillery	Custodian	0.00	0.00	0.00	
148650	148651	Not provided	Not provided	NaN	NaN	NaN	
148651	148652	Not provided	Not provided	NaN	NaN	NaN	
148652	148653	Not provided	Not provided	NaN	NaN	NaN	
148653	148654	Joe Lopez	Counselor, Log Cabin Ranch	0.00	0.00	-618.13	

148654 rows × 13 columns



เซต Id ให้เป็น index

```
In [ ]: sal.set_index('Id', inplace=True)
```

In []: sal

Out[]:

	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	To
--	--------------	----------	---------	-------------	----------	----------	----

		GENERAL MANAGER- METROPOLITAN TRANSIT AUTHORITY					
1	NATHANIEL FORD		167411.18	0.00	400184.25	0.0	567
2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	0.0	538
3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.60	0.0	335
4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.00	56120.71	198306.90	0.0	332
5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT, (FIRE DEPARTMENT)	134401.60	9737.00	182234.59	0.0	326
...
148650	Roy I Tillery	Custodian	0.00	0.00	0.00	0.0	
148651	Not provided	Not provided	NaN	NaN	NaN	0.0	
148652	Not provided	Not provided	NaN	NaN	NaN	0.0	
148653	Not provided	Not provided	NaN	NaN	NaN	0.0	
148654	Joe Lopez	Counselor, Log Cabin Ranch	0.00	0.00	-618.13	0.0	-

148654 rows × 12 columns



เปลี่ยนชื่อคอลัมน์ Year เป็น Year(Eng)

In []: sal.rename(columns={'Year': 'Year(ENG)'}, inplace=True)

In []: sal

Out []:

	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	To
Id							
1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.00	400184.25	0.0	567
2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	0.0	538
3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.60	0.0	335
4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.00	56120.71	198306.90	0.0	332
5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT, (FIRE DEPARTMENT)	134401.60	9737.00	182234.59	0.0	326
...
148650	Roy I Tillery	Custodian	0.00	0.00	0.00	0.0	
148651	Not provided	Not provided	NaN	NaN	NaN	0.0	
148652	Not provided	Not provided	NaN	NaN	NaN	0.0	
148653	Not provided	Not provided	NaN	NaN	NaN	0.0	
148654	Joe Lopez	Counselor, Log Cabin Ranch	0.00	0.00	-618.13	0.0	-

148654 rows × 12 columns



เพิ่มคนชื่อ David Copperfield ทำงานเป็น Magician คอลัมอื่นๆเป็น null

```
In [ ]: sal.loc[len(sal)+1] = pd.Series(['David Copperfield', 'Magician'], index = ['Employee', 'JobTitle'])
```

```
In [ ]: sal
```


Out[]:

	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	To
Id							
1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.00	400184.25	0.0	567
2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	0.0	538
3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.60	0.0	335
4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.00	56120.71	198306.90	0.0	332
5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT, (FIRE DEPARTMENT)	134401.60	9737.00	182234.59	0.0	326
...
148651	Not provided	Not provided	NaN	NaN	NaN	0.0	
148652	Not provided	Not provided	NaN	NaN	NaN	0.0	
148653	Not provided	Not provided	NaN	NaN	NaN	0.0	
148654	Joe Lopez	Counselor, Log Cabin Ranch	0.00	0.00	-618.13	0.0	-
148655	David Copperfield	Magician	NaN	NaN	NaN	NaN	

148655 rows × 12 columns



สร้าง Dataframe ที่ EmployeeName มีนาย A , B และ C ซึ่งมี BasePay เป็น 10000 แล้วนำไปรวมกับ sal

In []:

```
sal2 = pd.DataFrame({ 'EmployeeName':['A','B','C'],
                      'BasePay':[10000]},
                    index=[0,1,2])
```

```
In [ ]: sal = pd.concat([sal,sal2])
sal
```

```
Out[ ]:
```

	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	To
1	NATHANIEL FORD	GENERAL MANAGER- METROPOLITAN TRANSIT AUTHORITY	167411.18	0.00	400184.25	0.0	567
2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	0.0	538
3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.60	0.0	335
4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.00	56120.71	198306.90	0.0	332
5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT, (FIRE DEPARTMENT)	134401.60	9737.00	182234.59	0.0	326
...
148654	Joe Lopez	Counselor, Log Cabin Ranch	0.00	0.00	-618.13	0.0	-
148655	David Copperfield	Magician	NaN	NaN	NaN	NaN	NaN
0	A	NaN	10000.00	NaN	NaN	NaN	NaN
1	B	NaN	10000.00	NaN	NaN	NaN	NaN
2	C	NaN	10000.00	NaN	NaN	NaN	NaN

148658 rows × 12 columns



สร้างตาราง salB ซึ่งเก็บเฉพาะของคนที่ไม่ได้ BasePay

```
In [ ]: salB = sal[sal['BasePay'].isnull()]
```

```
In [ ]: salB.head()
```

Out[]:

	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay
81392	Kevin P Cashman	Deputy Chief 3	NaN	0.0	149934.11	0.00	149934.11
84507	Demetrya Mullens	Licensed Vocational Nurse	NaN	0.0	110485.41	20779.00	110485.41
84961	Michael M Horan	Park Patrol Officer	NaN	0.0	120000.00	8841.48	120000.00
90526	Thomas Tang	Police Officer 3	NaN	0.0	106079.31	0.00	106079.31
90787	Michael C Hill	Deputy Sheriff	NaN	0.0	81299.02	23877.53	81299.02

In []: `salB.info()`

```

<class 'pandas.core.frame.DataFrame'>
Index: 610 entries, 81392 to 148655
Data columns (total 12 columns):
#   Column                Non-Null Count  Dtype
---  -
0   EmployeeName          610 non-null    object
1   JobTitle              610 non-null    object
2   BasePay               0 non-null      float64
3   OvertimePay           605 non-null    float64
4   OtherPay              605 non-null    float64
5   Benefits              609 non-null    float64
6   TotalPay              609 non-null    float64
7   TotalPayBenefits      609 non-null    float64
8   Year(ENG)             609 non-null    float64
9   Agency                609 non-null    object
10  Years(TH)             609 non-null    float64
11  Level                 609 non-null    object
dtypes: float64(8), object(4)
memory usage: 62.0+ KB

```

----- ภาพนามยปัญญา ปัญญาที่เกิดจากการลงมือทำ! -----