Data Analysis Tools with Pandas 2 - SF Salaries Exercise

แบบฝึกหัดนี้เป็นแบบฝึกหัดทดสอบทักษะการใช้งาน library pandas ด้วย SF Salaries Dataset (https://www.kaggle.com/kaggle/sf-salaries) จากเว็ปไซต์ Kaggle ให้ทำตามคำสั่ง ต่อไปนี้

Import pandas as pd.

In [135]:
```
import pandas as pd
```

ให้นำเข้าข้อมูลจากไฟล์ Salaries.csv มาในรูปของ dataframe โดยตังชื่อตัวแปรว่า sal

In [136]:
```
sal = pd.read_csv('Salaries.csv')
```

Check the head of the DataFrame.

In [137]:
```
sal.head()
```

Out[137]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPayBe
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.00	400184.25	NaN	567595.43	5675
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	NaN	538909.28	5389
2	3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.60	NaN	335279.91	3352
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.00	56120.71	198306.90	NaN	332343.61	3323
4	5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT, (FIRE DEPARTMENT)	134401.60	9737.00	182234.59	NaN	326373.19	3263

ใช้คำสั่ง .info() method to ในการดูภาพรวมของข้อมูลทั้งหมด

In [138]:
```
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 [139]:
```
sal.drop('Notes',axis=1,inplace = True)
sal.drop('Status',axis=1,inplace = True)
```

In [140]:
```
sal
```

1	2	GARY JIMENEZ	(POLICE DEPARTMENT)	155966.02	245131.88	137811.38	NaN	538909.28
2	3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.60	NaN	335279.91
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.00	56120.71	198306.90	NaN	332343.61
4	5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT,(FIRE DEPARTMENT)	134401.60	9737.00	182234.59	NaN	326373.19
...	...	...	...	...	...	...	...	...
148649	148650	Roy I Tillery	Custodian	0.00	0.00	0.00	0.0	0.00
148650	148651	Not provided	Not provided	NaN	NaN	NaN	NaN	0.00

หาค่าเฉลี่ยนของ Benefits ใน sal

In [141]:
```
sal["Benefits"].mean()
```

Out[141]:
```
25007.893150829852
```

ใน sal แทน Benefits ที่เป็น null ด้วย 0

In [142]:
```
sal['Benefits'].fillna(0 , inplace=True)
```

In [143]:
```
sal.head()
```

Out[143]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPayI
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.00	400184.25	0.0	567595.43	56
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	0.0	538909.28	53
2	3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.60	0.0	335279.91	33
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.00	56120.71	198306.90	0.0	332343.61	33
			DEPUTY CHIEF						

หาค่าเฉลี่ยนของ Benefits ใน sal อีกครั้ง

In [144]:
```
sal["Benefits"].mean()
```

Out[144]:
```
18924.23283887417
```

จงเพิ่มคอลัมน์ Year(TH) ใน sal ให้เป็นเลขปี พศ

In [145]:
```
sal["Year(TH)"] = sal["Year"]+543
```

In [146]:
```
sal
```

		FORD	TRANSIT AUTHORITY					
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	0.0	538909.28
2	3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.60	0.0	335279.91
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.00	56120.71	198306.90	0.0	332343.61
4	5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT, (FIRE DEPARTMENT)	134401.60	9737.00	182234.59	0.0	326373.19
...	...	...	...	...	...	...	...	...

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

In [147]:
```
sal["Level"] = sal["TotalPayBenefits"].apply(levels)
```

In [148]:
```
def levels(input):
    if(input < 100000 ) :
        return "L"
    else :
        return "H"
```

In [149]:
```
sal
```

Out[149]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.00	400184.25	0.0	567595.43
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	0.0	538909.28
2	3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.60	0.0	335279.91
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.00	56120.71	198306.90	0.0	332343.61
			DEPUTY CHIEF					

เซ็ต Id ให้เป็น index

In [150]:
```
sal.set_index('Id',inplace = True)
```

In [151]:
```
sal
```

Out[151]:

	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPa
Id								
1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.00	400184.25	0.0	567595.43	
2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	0.0	538909.28	
3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.60	0.0	335279.91	
4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.00	56120.71	198306.90	0.0	332343.61	

เปลี่ยนชื่อคอลัมน์ Year เป็น Year(Eng)

In [152]:
```
sal.rename(columns = {'Year':'Year(Eng)'},inplace = True)
```

In [153]:
```
sal
```

Out[153]:

	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPayB
Id								
1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.00	400184.25	0.0	567595.43	567
2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	0.0	538909.28	538
3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.60	0.0	335279.91	335
4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.00	56120.71	198306.90	0.0	332343.61	332
5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT,(FIRE DEPARTMENT)	134401.60	9737.00	182234.59	0.0	326373.19	326
...	...	...	...	...	...	...	...	
148650	Roy I Tillery	Custodian	0.00	0.00	0.00	0.0	0.00	
148651	Not provided	Not provided	NaN	NaN	NaN	0.0	0.00	
148652	Not provided	Not provided	NaN	NaN	NaN	0.0	0.00	
148653	Not provided	Not provided	NaN	NaN	NaN	0.0	0.00	
148654	Joe Lopez	Counselor, Log Cabin Ranch	0.00	0.00	-618.13	0.0	-618.13	

148654 rows × 12 columns

เพิ่มคนชื่อ David Copperfield ทำงานเป็น Magician คอลัมน์อื่นๆเป็น null

In [154]:
```
sal.loc[len(sal.index)+1,["EmployeeName","JobTitle"]] = ["David Copperfield","Magician"
```

In [155]:
```
sal
```

Out[155]:

	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPayB
Id								
1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.00	400184.25	0.0	567595.43	567
2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	0.0	538909.28	538
3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.60	0.0	335279.91	335
4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.00	56120.71	198306.90	0.0	332343.61	332
5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT,(FIRE DEPARTMENT)	134401.60	9737.00	182234.59	0.0	326373.19	326
...	...	...	...	...	...	...	...	
148651	Not provided	Not provided	NaN	NaN	NaN	0.0	0.00	
148652	Not provided	Not provided	NaN	NaN	NaN	0.0	0.00	
148653	Not provided	Not provided	NaN	NaN	NaN	0.0	0.00	
148654	Joe Lopez	Counselor, Log Cabin Ranch	0.00	0.00	-618.13	0.0	-618.13	
148655	David Copperfield	Magician	NaN	NaN	NaN	NaN	NaN	

148655 rows × 12 columns

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

In [156]:
```
data = [['A',10000],['B',10000],['C',10000]]
df = pd.DataFrame(data,columns=['EmployeeName','BasePay'])
df
```

Out[156]:

	EmployeeName	BasePay
0	A	10000
1	B	10000
2	C	10000

In [157]:

```
sal = pd.concat([sal,df])
sal
```

Out[157]:

	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPa
1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.00	400184.25	0.0	567595.43	
2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	0.0	538909.28	
3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.60	0.0	335279.91	
4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.00	56120.71	198306.90	0.0	332343.61	
		DEPUTY CHIEF						

สร้างตาราง salB ซึ่งเก็บเฉพาะของคนที่ไม่มี BasePay

In [158]:

```
salB = pd.DataFrame(sal[sal["BasePay"].isnull()])
salB.head()
```

Out[158]:

	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPayBenefits
81392	Kevin P Cashman	Deputy Chief 3	NaN	0.0	149934.11	0.00	149934.11	149934.11
84507	Demetrya Mullens	Licensed Vocational Nurse	NaN	0.0	110485.41	20779.00	110485.41	131264.41
84961	Michael M Horan	Park Patrol Officer	NaN	0.0	120000.00	8841.48	120000.00	128841.48
90526	Thomas Tang	Police Officer 3	NaN	0.0	106079.31	0.00	106079.31	106079.31
90787	Michael C Hill	Deputy Sheriff	NaN	0.0	81299.02	23877.53	81299.02	105176.55

```
In [161]: 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  Year(TH)          609 non-null    float64
 11  Level             609 non-null    object 
dtypes: float64(8), object(4)
memory usage: 62.0+ KB
```

----- ภาวนามยปัญญา ปัญญาที่เกิดจากการลงมือทำ! -----