

Data Analysis Tools with Pandas - SF Salaries Exercise

แบบฝึกหัดนี้เป็นแบบฝึกหัดทดสอบทักษะการใช้งาน library pandas ด้วย [SF Salaries Dataset](https://www.kaggle.com/kaggle/sf-salaries) (<https://www.kaggle.com/kaggle/sf-salaries>) จากเว็บไซต์ Kaggle ให้ทำตามคำสั่ง ต่อไปนี้

Import pandas as pd.

In [1]:

```
import pandas as pd
```

ให้นำเข้าข้อมูลจากไฟล์ Salaries.csv มาในรูปของ dataframe โดยตั้งชื่อตัวแปรว่า sal

In [2]:

```
sal = pd.read_csv('Salaries.csv')
```

Check the head of the DataFrame.

In [3]:

```
sal.head()
```

Out[3]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.00	400184.25	NaN	567595.43
1	2	GARY JIMENEZ	CAPTAIN III (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

ใช้คำสั่ง .info() method to ในการดูภาพรวมของข้อมูลทั้งหมด

In [4]:

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
```

ให้หาค่า average ของ BasePay ?

In [5]:

sal['BasePay'].mean()

Out[5]:

66325.44884050643

OvertimePay สูงที่สุด ใน dataset เท่ากับเท่าไร?

In [6]:

sal['OvertimePay'].max()

Out[6]:

245131.88

JOSEPH DRISCOLL ทำงานอะไร (jobTitle)?

Note: Use all caps, otherwise you may get an answer that doesn't match up (there is also a lowercase Joseph Driscoll).

In [7]:

sal[sal['EmployeeName'] == 'JOSEPH DRISCOLL']['JobTitle']

Out[7]:

```
24   CAPTAIN, FIRE SUPPRESSION
Name: JobTitle, dtype: object
```

JOSEPH DRISCOLL ได้เงินไปทั้งหมดเท่าไร (รวมทั้ง benefits)?

In [8]:

```
sal[sal['EmployeeName'] == 'JOSEPH DRISCOLL']['JobTitle']
```

Out[8]:

```
24    CAPTAIN, FIRE SUPPRESSION
Name: JobTitle, dtype: object
```

ใครคือคนที่ได้รับเงินมากที่สุด (รวมทั้ง benefits)?

In [9]:

```
sal[sal['TotalPayBenefits'].max() == sal['TotalPayBenefits']]
```

Out[9]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.0	400184.25	NaN	567595.43

ใครคือคนที่ได้รับเงินน้อยที่สุด (รวมทั้ง benefits)?

Do you notice something strange about how much he or she is paid?

In [10]:

```
sal[sal['TotalPayBenefits'].min() == sal['TotalPayBenefits']]
```

Out[10]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay
148653	148654	Joe Lopez	Counselor, Log Cabin Ranch	0.0	0.0	-618.13	0.0	-618.13

จงหาค่า average (mean) ของ BasePay ของ employees ทั้งหมดในแต่ละปี (2011-2014)

In [11]:

```
sal.groupby('Year').mean()['BasePay']
```

Out[11]:

```
Year
2011    63595.956517
2012    65436.406857
2013    69630.030216
2014    66564.421924
Name: BasePay, dtype: float64
```

มีชื่อตำแหน่งงานต่างๆ (unique job) อยู่กี่ชื่อ?

In [12]:

```
len(sal['JobTitle'].unique())
```

Out[12]:

```
2159
```

top 5 ตำแหน่งเป็นที่ต้องการในที่ต่างๆ มีอะไรบ้าง ?

In [13]:

```
sal['JobTitle'].value_counts().head()
```

Out[13]:

```
Transit Operator          7036
Special Nurse             4389
Registered Nurse          3736
Public Svc Aide-Public Works  2518
Police Officer 3          2421
Name: JobTitle, dtype: int64
```

มีจำนวนที่ตำแหน่งที่ต้องการเพียง 2 คน ในปี 2013? (e.g. Job Titles with only one occurrence in 2013?)

In [14]:

```
# sal['Year'].apply(lambda x: x == 2013)

df1 = sal[sal['Year'] == 2013]['JobTitle'].value_counts() == 2
df2 = sal[sal['Year'] == 2013]['JobTitle'].value_counts()
df2[df1].count()
#ตอบ 69
```

Out[14]:

```
69
```

มีคนที่คนที่มีความว่า Chief อยู่ในชื่อตำแหน่ง job title ของเค้า (This is pretty tricky)

In [15]:

```
sal[sal['JobTitle'].apply(lambda x: 'chief' in x.lower())]['JobTitle'].count()
```

Out[15]:

627

Bonus: Is there a correlation between length of the Job Title string and Salary?

In []:

In []:

----- ถาวนามยปัญญา ปัญญาที่เกิดจากการลงมือทำ! -----