SF Salaries Exercise

For this exercise, you will be using the SF Salaries Dataset from Kaggle!. You can also used the SF Salaries Dataset on D2L to answer the questions. Please download the "Salaries.csv" file from kaggle or D2L and complete the below exercises. Please submit the completed jupyter notebook file to D2L.

Exercise 1

Import pandas as pd.

In [1]:

```
import pandas as pd
```

Read Salaries.csv as a dataframe called sal.

In [2]:

```
sal=pd.read_csv('..\\Data\\Salaries.csv',low_memory=False)
```

Check the head of the DataFrame.

In [3]:

```
sal.head()
```

Out[3]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	Tota
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.0	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	16452.6	NaN	335279.91	
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.0	56120.71	198306.9	NaN	332343.61	
4	5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT,(FIRE DEPARTMENT)	134401.6	9737.0	182234.59	NaN	326373.19	

Exercise 2 - Use the .info() method to find out how many entries there are.

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           148049 non-null  object 
 4   OvertimePay       148654 non-null  object 
 5   OtherPay          148654 non-null  object 
 6   Benefits          112495 non-null  object 
 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            38119 non-null   object 
dtypes: float64(3), int64(2), object(8)
memory usage: 14.7+ MB
```

Exercise 3 - What is the average BasePay ?

In [5]:

```
import numpy as np
sal=sal.replace('Not Provided',np.nan)
sal.BasePay=sal.BasePay.astype('float')
sal.BasePay.mean()
```

Out[5]:

```
66325.44884050643
```

Exercise 4 - What is the highest amount of OvertimePay in the dataset ?

In [6]:

```
sal.OvertimePay=sal.OvertimePay.astype('float')
sal.OvertimePay.max()
```

Out[6]:

```
245131.88
```

Exercise 5 - What is the job title of JOSEPH DRISCOLL ? Note: Use all caps, otherwise you may get an answer that doesn't match up (there is also a lowercase Joseph Driscoll).

In [7]:

```
tuple(sal.loc[sal.EmployeeName=='JOSEPH DRISCOLL','JobTitle'])[0]
```

Out[7]:

```
'CAPTAIN, FIRE SUPPRESSION'
```

Exercise 6 - How much does JOSEPH DRISCOLL make (including benefits)?

In [8]:

```
tuple(sal.loc[sal.EmployeeName=='JOSEPH DRISCOLL','TotalPayBenefits'])[0]
```

Out[8]:

```
270324.91
```

Exercise 7 - What is the name of highest paid person (including benefits)?

In [9]:

```
tuple(sal.loc[sal.TotalPayBenefits==sal.TotalPayBenefits.max(),'EmployeeName'])[0]
```

Out[9]:

```
'NATHANIEL FORD'
```

Exercise 8 - What is the name of lowest paid person (including benefits)?

In [10]:

```
tuple(sal.loc[sal.TotalPayBenefits==sal.TotalPayBenefits.min(),'EmployeeName'])[0]
```

Out[10]:

```
'Joe Lopez'
```

Exercise 9 - What was the average (mean) BasePay of all employees per year? (2011-2014)?

In [11]:

```
sal.groupby(by='Year')['BasePay'].mean().round(2)
```

Out[11]:

```
Year
2011    63595.96
2012    65436.41
2013    69630.03
2014    66564.42
Name: BasePay, dtype: float64
```

Exercise 10 - How many unique job titles are there?

In [12]:

```
len(sal.JobTitle.unique())
```

Out[12]:

```
2159
```

Exercise 11 - What are the top 5 most common jobs?

In [13]:

```
sal.JobTitle.value_counts().sort_values(ascending=False).head(5)
```

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
```

Exercise 12 - How many Job Titles were represented by only one person in 2013? (e.g. Job Titles with only one occurence in 2013?)

In [14]:

```
series=sal[sal.Year==2013].groupby('JobTitle')['Id'].count()
len(series[series==1])
```

Out[14]:

```
202
```

Exercise 13: Is there a correlation between length of the Job Title string and Salary?

In [15]:

```
# is the Salary the total pay?
df=pd.DataFrame({'JobTitle_len':[len(x) for x in sal.JobTitle],'TotalPayBenefits':sal.
df.corr()
```

Out[15]:

	JobTitle_len	TotalPayBenefits
JobTitle_len	1.000000	-0.036878
TotalPayBenefits	-0.036878	1.000000

A linear relation between Job Title and Total Pay Benefits are almost inexistent