SF Salaries Exercise

This is the practice file for Pandas library of Python. I am using San Francisco Data Set from Kaggle.com https://www.kaggle.com/kaggle/sf-salaries (https://www.kaggle.com/kaggle/sf-salaries)

I am practicing these questions from Udemy.com, Course Name: Python for Data Science and Machine Learning Bootcamp, Tutor: Jose Portilla

Import pandas as pd.

In [3]:
```
import pandas as pd
```

Read Salaries.csv as a dataframe called sal.

In [4]:
```
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	Tota
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.00	400184.25	NaN	5675
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	NaN	5389
2	3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.60	NaN	3352
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.00	56120.71	198306.90	NaN	3323
4	5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT, (FIRE DEPARTMENT)	134401.60	9737.00	182234.59	NaN	3263

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):
Id                  148654 non-null int64
EmployeeName        148654 non-null object
JobTitle            148654 non-null object
BasePay             148045 non-null float64
OvertimePay         148650 non-null float64
OtherPay            148650 non-null float64
Benefits            112491 non-null float64
TotalPay            148654 non-null float64
TotalPayBenefits    148654 non-null float64
Year                148654 non-null int64
Notes               0 non-null float64
Agency              148654 non-null object
Status              0 non-null float64
dtypes: float64(8), int64(2), object(3)
memory usage: 14.7+ MB
```

What is the average BasePay ?

In [8]:
```
sal['BasePay'].mean()
```

Out[8]:
```
66325.44884050643
```

In [9]:
```
sal['BasePay'].count()
```

Out[9]:
```
148045
```

Manual Calculation for average

In [11]:
```
(sal['BasePay'].sum())/sal['BasePay'].count()
```

Out[11]:
```
66325.448840506433
```

What is the highest amount of OvertimePay in the dataset ?

In [11]:
```
# manual way to search as a WHERE clause.
sal['OvertimePay'].idxmax()
sal.iloc[1]
# SELECT * FROM SF_SALARIES
# WHERE overtimepay = max
```

Out[11]:
```
Id                                            2
EmployeeName                        GARY JIMENEZ
JobTitle            CAPTAIN III (POLICE DEPARTMENT)
BasePay                                    155966
OvertimePay                                245132
OtherPay                                   137811
Benefits                                      NaN
TotalPay                                   538909
TotalPayBenefits                           538909
Year                                         2011
Notes                                         NaN
Agency                              San Francisco
Status                                        NaN
Name: 1, dtype: object
```

Returning a row with a specified WHERE condition (Similar to SQL is written in this way)

In [7]:
```
sal[sal['OvertimePay'] == sal['OvertimePay'].max()]
```

Out[7]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalP
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	NaN	53890

Finding the people with a total pay of more than $100,000 in San Francisco

In [34]:
```
sal[sal['TotalPay']>100000].sort_values('BasePay', ascending = False)
```

Out[34]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Be
72925	72926	Gregory P Suhr	Chief of Police	319275.01	0.00	20007.06	86
110532	110533	Amy P Hart	Asst Med Examiner	318835.49	10712.95	60563.54	89
72929	72930	Robert L Shaw	Dep Dir for Investments, Ret	315572.01	0.00	0.00	82
72926	72927	Joanne M Hayes-White	Chief, Fire Department	313686.01	0.00	23236.00	85
72931	72932	Harlan L Kelly-Jr	Executive Contract Employee	313312.52	0.00	0.00	82

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 [54]:
```
sal[sal['EmployeeName']=='JOSEPH DRISCOLL'][['EmployeeName','JobTitle','TotalPayB
```

Out[54]:

	EmployeeName	JobTitle	TotalPayBenefits
24	JOSEPH DRISCOLL	CAPTAIN, FIRE SUPPRESSION	270324.91

Comparing the Joseph's Total Pay (Found in previous command line) with others of same profile

In [53]:
```
sal[sal['JobTitle']=='CAPTAIN, FIRE SUPPRESSION']['TotalPayBenefits'].mean()
```

Out[53]:
```
179758.84239436616
```

How much does JOSEPH DRISCOLL make (including benefits)?

In [52]:
```
sal[sal['EmployeeName']=='JOSEPH DRISCOLL']['TotalPayBenefits']
```

Out[52]:
```
24    270324.91
Name: TotalPayBenefits, dtype: float64
```

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

In [50]:
```
sal[sal['TotalPayBenefits']== sal['TotalPayBenefits'].max()]
```

Out[50]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	Tota
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.0	400184.25	NaN	5675

What is the name of lowest paid person (including benefits)? Do you notice something strange about how much he or she is paid?

In [7]:
```
sal[sal['BasePay']==sal['BasePay'].min()]
```

Out[7]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	Totall
72832	72833	Irwin Sidharta	Junior Clerk	-166.01	249.02	0.0	6.56	83.01

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

In [37]:
```
sal.groupby(by='Year').max()['BasePay']
```

Out[37]:
```
Year
2011    294580.02
2012    302578.00
2013    319275.01
2014    318835.49
Name: BasePay, dtype: float64
```

Finding the most common job title in San Francisco

In [23]:
```
sal.groupby('JobTitle').count()['Id'].sort_values(ascending = False).head()
```

Out[23]:
```
JobTitle
Transit Operator                7036
Special Nurse                   4389
Registered Nurse                3736
Public Svc Aide-Public Works    2518
Police Officer 3                2421
Name: Id, dtype: int64
```

How many unique job titles are there?

In [14]:
```
sal['JobTitle'].nunique()
```

Out[14]:
```
2159
```

What are the top 5 most common jobs?

In [17]:
```
sal.groupby(by='JobTitle').count()['Id'].sort_values(ascending = False).head(5)
# easy way to do is:
# sal['JobTitle'].value_counts().....done below
```

Out[17]:
```
JobTitle
Transit Operator                7036
Special Nurse                   4389
Registered Nurse                3736
Public Svc Aide-Public Works    2518
Police Officer 3                2421
Name: Id, dtype: int64
```

In [19]:
```
sal['JobTitle'].value_counts().head(5)
```

Out[19]:
```
Transit Operator                7036
Special Nurse                   4389
Registered Nurse                3736
Public Svc Aide-Public Works    2518
Police Officer 3                2421
Name: JobTitle, dtype: int64
```

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

In [65]:
```
sal[(sal['Year']==2013) & (sal['JobTitle'].value_counts==1)] #This is not the cor
```

Out[65]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	Total

In [25]:
```
sal[sal['Year']== 2013]['JobTitle'].value_counts()==1
# sal[sal['TotalPayBenefits']== sal['TotalPayBenefits'].max()]
# sal[sal['JobTitle'] == sal['JobTitle].value_counts == 1]
```

Out[25]:
```
Transit Operator                  False
Special Nurse                     False
Registered Nurse                  False
Public Svc Aide-Public Works      False
Custodian                         False
Firefighter                       False
Police Officer 3                  False
Patient Care Assistant            False
Recreation Leader                 False
Deputy Sheriff                    False
Police Officer                    False
Public Service Trainee            False
Police Officer 2                  False
Attorney (Civil/Criminal)         False
Sergeant 3                        False
Porter                            False
Eligibility Worker                False
General Laborer                   False
Gardener                          False
EMT/Paramedic/Firefighter         False
Senior Clerk                      False
Parking Control Officer           False
Library Page                      False
Senior Eligibility Worker         False
Senior Clerk Typist               False
Licensed Vocational Nurse         False
Clerk                             False
Stationary Engineer               False
Nurse Practitioner                False
PS Aide to Prof                   False
                                  ...
Track Maint Supt, Muni Railway     True
Auto Body & Fender Wrk Sprv 1      True
Gen Mgr, Public Trnsp Dept         True
Sr Employee Asst Counselor         True
Adm, SFGH Medical Center           True
Administrative Analyst I           True
Asphalt Plant Supervisor 1         True
Orthopedic Technician 1            True
Sprv Adult Prob Ofc (SFERS)        True
Assistant Law Librarian            True
Field Svcs Asst Supv               True
Media Production Specialist        True
Special Assistant 16               True
Assoc Musm Cnsrvt, AAM             True
Chief Nursery Specialist           True
Assistant Director, Probate        True
Research Psychologist              True
Signal and Systems Engineer        True
Pr Investigator, Tax Collector     True
Employment & Training Spec 6       True
Payroll Supervisor                 True
Public Safety Comm Tech            True
```

```
Arborist Technician Supv II        True
IS Operator-Journey                True
Drug Court Coordinator             True
Transit Paint Shop Sprv1           True
Real Estate Devt. Mgr, SFMTA       True
Ex Asst to the Controller          True
Statistician                       True
Legislation Clerk                  True
Name: JobTitle, dtype: bool
```

In [55]:
```
# sal[sal['JobTitle'].value_counts == 1]
sum(sal[sal['Year']==2013]['JobTitle'].value_counts() == 1)
```

Out[55]: 202

How many people have the word Chief in their job title? And return a table of results for them

In [12]:
```
def chief_string(title):
    if 'chief' in title.lower():
        return True
    else:
        return False
```

In [26]:
```
#Counting
sal[sal['JobTitle'].apply(lambda x: chief_string(x))]['Id'].count()
```

Out[26]: 627

In [27]:
```
# Extracting information with people have job title 'Chief'.
sal[sal['JobTitle'].apply(lambda x: chief_string(x))][['Id','EmployeeName','Total
```

Out[27]:

	Id	EmployeeName	TotalPay	JobTitle
4	5	PATRICK GARDNER	326373.19	DEPUTY CHIEF OF DEPARTMENT,(FIRE DEPARTMENT)
5	6	DAVID SULLIVAN	316285.74	ASSISTANT DEPUTY CHIEF II
6	7	ALSON LEE	315981.05	BATTALION CHIEF, (FIRE DEPARTMENT)
8	9	MICHAEL MORRIS	303427.55	BATTALION CHIEF, (FIRE DEPARTMENT)
9	10	JOANNE HAYES-WHITE	302377.73	CHIEF OF DEPARTMENT, (FIRE DEPARTMENT)

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

In [76]:
```
sal['length_sal'] = sal['JobTitle'].apply(len)
```

In [77]:
```
sal[['length_sal','TotalPayBenefits']].corr()
```

Out[77]:

	length_sal	TotalPayBenefits
length_sal	1.000000	-0.036878
TotalPayBenefits	-0.036878	1.000000