python project in data set salaries

first import the pandas

In [1]:
```
import pandas as pd
```

In [2]:
```
data = pd.read_csv('D:\\Edubriage assements submet\\Salaries.csv')
```

```
C:\Users\ASUS\anaconda3\lib\site-packages\IPython\core\interactiveshell.py:3165: DtypeWarning: Columns (3,4,5,6,12) have mixed types.Specify dtype option on i
mport or set low_memory=False.
  has_raised = await self.run_ast_nodes(code_ast.body, cell_name,
```

In [3]:
```
data
```

Out[3]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPayBenefits	Year	Notes	Agency	Status
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.0	400184.25	NaN	567595.43	567595.43	2011	NaN	San Francisco	NaN
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	NaN	538909.28	538909.28	2011	NaN	San Francisco	NaN
2	3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.6	NaN	335279.91	335279.91	2011	NaN	San Francisco	NaN
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.0	56120.71	198306.9	NaN	332343.61	332343.61	2011	NaN	San Francisco	NaN
4	5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT,(FIRE DEPARTMENT)	134401.6	9737.0	182234.59	NaN	326373.19	326373.19	2011	NaN	San Francisco	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...
148649	148650	Roy I Tillery	Custodian	0.00	0.00	0.00	0.00	0.00	0.00	2014	NaN	San Francisco	PT
148650	148651	Not provided	Not provided	Not Provided	Not Provided	Not Provided	Not Provided	0.00	0.00	2014	NaN	San Francisco	NaN
148651	148652	Not provided	Not provided	Not Provided	Not Provided	Not Provided	Not Provided	0.00	0.00	2014	NaN	San Francisco	NaN
148652	148653	Not provided	Not provided	Not Provided	Not Provided	Not Provided	Not Provided	0.00	0.00	2014	NaN	San Francisco	NaN
148653	148654	Joe Lopez	Counselor, Log Cabin Ranch	0.00	0.00	-618.13	0.00	-618.13	-618.13	2014	NaN	San Francisco	PT

148654 rows × 13 columns

1. display Top 10 rows of the datasets

In [4]:
```
data.head(10)
```

Out[4]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPayBenefits	Year	Notes	Agency	Status
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.0	400184.25	NaN	567595.43	567595.43	2011	NaN	San Francisco	NaN
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	NaN	538909.28	538909.28	2011	NaN	San Francisco	NaN
2	3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.6	NaN	335279.91	335279.91	2011	NaN	San Francisco	NaN
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.0	56120.71	198306.9	NaN	332343.61	332343.61	2011	NaN	San Francisco	NaN
4	5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT,(FIRE DEPARTMENT)	134401.6	9737.0	182234.59	NaN	326373.19	326373.19	2011	NaN	San Francisco	NaN
5	6	DAVID SULLIVAN	ASSISTANT DEPUTY CHIEF II	118602.0	8601.0	189082.74	NaN	316285.74	316285.74	2011	NaN	San Francisco	NaN
6	7	ALSON LEE	BATTALION CHIEF, (FIRE DEPARTMENT)	92492.01	89062.9	134426.14	NaN	315981.05	315981.05	2011	NaN	San Francisco	NaN
7	8	DAVID KUSHNER	DEPUTY DIRECTOR OF INVESTMENTS	256576.96	0.0	51322.5	NaN	307899.46	307899.46	2011	NaN	San Francisco	NaN
8	9	MICHAEL MORRIS	BATTALION CHIEF, (FIRE DEPARTMENT)	176932.64	86362.68	40132.23	NaN	303427.55	303427.55	2011	NaN	San Francisco	NaN
9	10	JOANNE HAYES-WHITE	CHIEF OF DEPARTMENT, (FIRE DEPARTMENT)	285262.0	0.0	17115.73	NaN	302377.73	302377.73	2011	NaN	San Francisco	NaN

2.check last 10 Rows of the data sets

In [5]:
```
data.tail(10)
```

Out[5]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPayBenefits	Year	Notes	Agency	Status
148644	148645	Randy D Winn	Stationary Eng, Sewage Plant	0.00	0.00	0.00	0.00	0.00	0.00	2014	NaN	San Francisco	PT
148645	148646	Carolyn A Wilson	Human Services Technician	0.00	0.00	0.00	0.00	0.00	0.00	2014	NaN	San Francisco	PT
148646	148647	Not provided	Not provided	Not Provided	Not Provided	Not Provided	Not Provided	0.00	0.00	2014	NaN	San Francisco	NaN
148647	148648	Joann Anderson	Communications Dispatcher 2	0.00	0.00	0.00	0.00	0.00	0.00	2014	NaN	San Francisco	PT
148648	148649	Leon Walker	Custodian	0.00	0.00	0.00	0.00	0.00	0.00	2014	NaN	San Francisco	PT
148649	148650	Roy I Tillery	Custodian	0.00	0.00	0.00	0.00	0.00	0.00	2014	NaN	San Francisco	PT
148650	148651	Not provided	Not provided	Not Provided	Not Provided	Not Provided	Not Provided	0.00	0.00	2014	NaN	San Francisco	NaN
148651	148652	Not provided	Not provided	Not Provided	Not Provided	Not Provided	Not Provided	0.00	0.00	2014	NaN	San Francisco	NaN
148652	148653	Not provided	Not provided	Not Provided	Not Provided	Not Provided	Not Provided	0.00	0.00	2014	NaN	San Francisco	NaN
148653	148654	Joe Lopez	Counselor, Log Cabin Ranch	0.00	0.00	-618.13	0.00	-618.13	-618.13	2014	NaN	San Francisco	PT

3. Find shape of out dateset(no of rows and no of columns)

In [6]:
```
data.shape
```

Out[6]:
```
(148654, 13)
```

In [7]:
```
print("number of rows",data.shape[0])
print("number of colums ",data.shape[1])
```

```
number of rows 148654
number of colums  13
```

4.Getting information about Data set Total no of Row ,Total no of columns

In [8]:
```
data.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
```

5.Check the null values in the datasets

In [9]:
```
data.isnull().sum()
```

Out[9]:
```
Id                       0
EmployeeName             0
JobTitle                 0
BasePay                605
OvertimePay              0
OtherPay                 0
Benefits             36159
TotalPay                 0
TotalPayBenefits         0
Year                     0
Notes               148654
Agency                   0
Status              110535
dtype: int64
```

6.Drop Id,notes,Agencs and status columns

In [10]:
```
data.columns
```

Out[10]:
```
Index(['Id', 'EmployeeName', 'JobTitle', 'BasePay', 'OvertimePay', 'OtherPay',
       'Benefits', 'TotalPay', 'TotalPayBenefits', 'Year', 'Notes', 'Agency',
       'Status'],
      dtype='object')
```

In [11]:
```
data = data.drop(['Id','Notes','Agency','Status'],axis=1)
```

In [12]:
```
data.head(1)
```

Out[12]:

	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPayBenefits	Year
0	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.0	400184.25	NaN	567595.43	567595.43	2011

7.Get Overall statistics About the data frame

In [13]:
```
data.describe(include='all')
```

Out[13]:

	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPayBenefits	Year
count	148654	148654	148049.0	148654.0	148654.0	112495.0	148654.000000	148654.000000	148654.000000
unique	110811	2159	109900.0	66555.0	84968.0	99635.0	NaN	NaN	NaN
top	Kevin Lee	Transit Operator	0.0	0.0	0.0	0.0	NaN	NaN	NaN
freq	13	7036	875.0	66103.0	35218.0	1053.0	NaN	NaN	NaN
mean	NaN	NaN	NaN	NaN	NaN	NaN	74768.321972	93692.554811	2012.522643
std	NaN	NaN	NaN	NaN	NaN	NaN	50517.005274	62793.533483	1.117538
min	NaN	NaN	NaN	NaN	NaN	NaN	-618.130000	-618.130000	2011.000000
25%	NaN	NaN	NaN	NaN	NaN	NaN	36168.995000	44065.650000	2012.000000
50%	NaN	NaN	NaN	NaN	NaN	NaN	71426.610000	92404.090000	2013.000000
75%	NaN	NaN	NaN	NaN	NaN	NaN	105839.135000	132876.450000	2014.000000
max	NaN	NaN	NaN	NaN	NaN	NaN	567595.430000	567595.430000	2014.000000

8.Find occurrence of the Employee name (Top 5)

In [57]:
```
data.columns
```

Out[57]:
```
Index(['EmployeeName', 'JobTitle', 'BasePay', 'OvertimePay', 'OtherPay',
       'Benefits', 'TotalPay', 'TotalPayBenefits', 'Year'],
      dtype='object')
```

In [14]:
```
data['EmployeeName'].value_counts().head(
)
```

Out[14]:
```
Kevin Lee       13
William Wong    11
Steven Lee      11
Richard Lee     11
Stanley Lee      9
Name: EmployeeName, dtype: int64
```

9.Find The Number of Unique job Titles

In [15]:
```
data.columns
```

Out[15]:
```
Index(['EmployeeName', 'JobTitle', 'BasePay', 'OvertimePay', 'OtherPay',
       'Benefits', 'TotalPay', 'TotalPayBenefits', 'Year'],
      dtype='object')
```

In [16]:
```
data['JobTitle'].nunique()
```

Out[16]:
```
2159
```

10.Total Number of Job Titles contian captain

In [17]:
```
data.columns
```

Out[17]:
```
Index(['EmployeeName', 'JobTitle', 'BasePay', 'OvertimePay', 'OtherPay',
       'Benefits', 'TotalPay', 'TotalPayBenefits', 'Year'],
      dtype='object')
```

In [69]:
```
len(data[data['JobTitle'].str.contains('CAPTAIN',case=False)])
```

Out[69]:
```
552
```

In [18]:
```
data[data['JobTitle'].str.contains('CAPTAIN',case=False)].count()
```

Out[18]:
```
EmployeeName        552
JobTitle            552
BasePay             551
OvertimePay         552
OtherPay            552
Benefits            411
TotalPay            552
TotalPayBenefits    552
Year                552
dtype: int64
```

11.Display All The EmployeeNames From Fire Departments

In [19]:
```
data.columns
```

Out[19]:
```
Index(['EmployeeName', 'JobTitle', 'BasePay', 'OvertimePay', 'OtherPay',
       'Benefits', 'TotalPay', 'TotalPayBenefits', 'Year'],
      dtype='object')
```

In [20]:
```
data[data['JobTitle'].str.contains('fire',case=False)]['EmployeeName']
```

Out[20]:
```
4             PATRICK GARDNER
6                   ALSON LEE
8              MICHAEL MORRIS
9          JOANNE HAYES-WHITE
10              ARTHUR KENNEY
                 ...
145956      Kenneth C Farris
147556         Edward A Dunn
148021        Kari A Johnson
148209          Sheryl K Lee
148554       Lawrence F Gatt
Name: EmployeeName, Length: 5879, dtype: object
```

12.Find Minimum and maximum and Average Base pay

In [21]:
```
data.columns
```

Out[21]:
```
Index(['EmployeeName', 'JobTitle', 'BasePay', 'OvertimePay', 'OtherPay',
       'Benefits', 'TotalPay', 'TotalPayBenefits', 'Year'],
      dtype='object')
```

In [22]:
```
data['BasePay'].describe()
```

Out[22]:
```
count     148049.0
unique    109900.0
top            0.0
freq         875.0
Name: BasePay, dtype: float64
```

13 Replace Not provided in Employee name column to nan

In [23]:
```
data.columns
```

Out[23]:
```
Index(['EmployeeName', 'JobTitle', 'BasePay', 'OvertimePay', 'OtherPay',
       'Benefits', 'TotalPay', 'TotalPayBenefits', 'Year'],
      dtype='object')
```

In [96]:
```
data['EmployeeName']
```

Out[96]:
```
0             NATHANIEL FORD
1               GARY JIMENEZ
2             ALBERT PARDINI
3          CHRISTOPHER CHONG
4            PATRICK GARDNER
                 ...
148649        Roy I Tillery
148650         Not provided
148651         Not provided
148652         Not provided
148653            Joe Lopez
Name: EmployeeName, Length: 148654, dtype: object
```

In [24]:
```
import numpy as np
data['EmployeeName'].replace('Not provided',np.nan)
```

Out[24]:
```
0             NATHANIEL FORD
1               GARY JIMENEZ
2             ALBERT PARDINI
3          CHRISTOPHER CHONG
4            PATRICK GARDNER
                 ...
148649        Roy I Tillery
148650                  NaN
148651                  NaN
148652                  NaN
148653            Joe Lopez
Name: EmployeeName, Length: 148654, dtype: object
```

14.Find job Titles of ALBERT PARDINI

In [25]:
```
data.columns
```

Out[25]:
```
Index(['EmployeeName', 'JobTitle', 'BasePay', 'OvertimePay', 'OtherPay',
       'Benefits', 'TotalPay', 'TotalPayBenefits', 'Year'],
      dtype='object')
```

In [26]:
```
data.head()
data[data['EmployeeName']=='ALBERT PARDINI']['JobTitle']
```

Out[26]:
```
2    CAPTAIN III (POLICE DEPARTMENT)
Name: JobTitle, dtype: object
```

15.how much ALBERT PARDINI make(include benefits)

In [27]:
```
data.columns
```

Out[27]:
```
Index(['EmployeeName', 'JobTitle', 'BasePay', 'OvertimePay', 'OtherPay',
       'Benefits', 'TotalPay', 'TotalPayBenefits', 'Year'],
      dtype='object')
```

In [28]:
```
data[data['EmployeeName']=='ALBERT PARDINI']
```

Out[28]:

	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPayBenefits	Year
2	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.6	NaN	335279.91	335279.91	2011

In [29]:
```
data[data['EmployeeName']=='ALBERT PARDINI']['TotalPayBenefits']
```

Out[29]:
```
2    335279.91
Name: TotalPayBenefits, dtype: float64
```

In [ ]: