In [1]:
```
import numpy as np
import pandas as pd
#%matplotlib notebook
%matplotlib inline
```

import the dataset into a dataframe

In [2]:
```
df = pd.read_csv('Salaries.csv')
df
```

Out[2]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPayBenefits	Year	Notes	Agenc
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.00	400184.25	NaN	567595.43	567595.43	2011	NaN	Sa Francis
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	NaN	538909.28	538909.28	2011	NaN	Sa Francis
2	3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.60	NaN	335279.91	335279.91	2011	NaN	Sa Francis
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.00	56120.71	198306.90	NaN	332343.61	332343.61	2011	NaN	Sa Francis
4	5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT, (FIRE DEPARTMENT)	134401.60	9737.00	182234.59	NaN	326373.19	326373.19	2011	NaN	Sa Francis
...	...	...	...	...	...	...	...	...	...	...	...	
148649	148650	Roy I Tillery	Custodian	0.00	0.00	0.00	0.0	0.00	0.00	2014	NaN	Sa Francis
148650	148651	Not provided	Not provided	NaN	NaN	NaN	NaN	0.00	0.00	2014	NaN	Sa Francis
148651	148652	Not provided	Not provided	NaN	NaN	NaN	NaN	0.00	0.00	2014	NaN	Sa Francis
148652	148653	Not provided	Not provided	NaN	NaN	NaN	NaN	0.00	0.00	2014	NaN	Sa Francis
148653	148654	Joe Lopez	Counselor, Log Cabin Ranch	0.00	0.00	-618.13	0.0	-618.13	-618.13	2014	NaN	Sa Francis

148654 rows × 13 columns

display the column names

In [3]:
```
df.columns
```

Out[3]:
```
Index(['Id', 'EmployeeName', 'JobTitle', 'BasePay', 'OvertimePay', 'OtherPay',
       'Benefits', 'TotalPay', 'TotalPayBenefits', 'Year', 'Notes', 'Agency',
       'Status'],
      dtype='object')
```

display the number of rows and cols

In [4]:
```
df.shape
```

Out[4]:
```
(148654, 13)
```

display the dataframe info (types of data in columns and not null values etc.)

In [5]:
```
df.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
```

display stats of the dataframe like count, mean, std, max, 25% etc.....

In [6]:
```
df.describe(include='all')
```

Out[6]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPayBenefits	1
count	148654.000000	148654	148654	148045.000000	148650.000000	148650.000000	112491.000000	148654.000000	148654.000000	
unique	NaN	110811	2159	NaN	NaN	NaN	NaN	NaN	NaN	
top	NaN	Kevin Lee	Transit Operator	NaN	NaN	NaN	NaN	NaN	NaN	
freq	NaN	13	7036	NaN	NaN	NaN	NaN	NaN	NaN	
mean	74327.500000	NaN	NaN	66325.448841	5066.059886	3648.767297	25007.893151	74768.321972	93692.554811	
std	42912.857795	NaN	NaN	42764.635495	11454.380559	8056.601866	15402.215858	50517.005274	62793.533483	
min	1.000000	NaN	NaN	-166.010000	-0.010000	-7058.590000	-33.890000	-618.130000	-618.130000	
25%	37164.250000	NaN	NaN	33588.200000	0.000000	0.000000	11535.395000	36168.995000	44065.650000	
50%	74327.500000	NaN	NaN	65007.450000	0.000000	811.270000	28628.620000	71426.610000	92404.090000	
75%	111490.750000	NaN	NaN	94691.050000	4658.175000	4236.065000	35566.855000	105839.135000	132876.450000	
max	148654.000000	NaN	NaN	319275.010000	245131.880000	400184.250000	96570.660000	567595.430000	567595.430000	

In [7]:
```
#OR

df.describe()
```

Out[7]:

	Id	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPayBenefits	Year	Notes	Statu
count	148654.000000	148045.000000	148650.000000	148650.000000	112491.000000	148654.000000	148654.000000	148654.000000	0.0	0.(
mean	74327.500000	66325.448841	5066.059886	3648.767297	25007.893151	74768.321972	93692.554811	2012.522643	NaN	Nal
std	42912.857795	42764.635495	11454.380559	8056.601866	15402.215858	50517.005274	62793.533483	1.117538	NaN	Nal
min	1.000000	-166.010000	-0.010000	-7058.590000	-33.890000	-618.130000	-618.130000	2011.000000	NaN	Nal
25%	37164.250000	33588.200000	0.000000	0.000000	11535.395000	36168.995000	44065.650000	2012.000000	NaN	Nal
50%	74327.500000	65007.450000	0.000000	811.270000	28628.620000	71426.610000	92404.090000	2013.000000	NaN	Nal
75%	111490.750000	94691.050000	4658.175000	4236.065000	35566.855000	105839.135000	132876.450000	2014.000000	NaN	Nal
max	148654.000000	319275.010000	245131.880000	400184.250000	96570.660000	567595.430000	567595.430000	2014.000000	NaN	Nal

display null values per column

In [8]:
```
df.isnull().sum()
```

Out[8]:
```
Id                     0
EmployeeName           0
JobTitle               0
BasePay              609
OvertimePay            4
OtherPay               4
Benefits           36163
TotalPay               0
TotalPayBenefits       0
Year                   0
Notes             148654
Agency                 0
Status            148654
dtype: int64
```

remove columns will all values as NaN

In [9]:
```
df.dropna(axis = 1, how = 'all', inplace = True)
df
```

Out[9]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPayBenefits	Year	Agency
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.00	400184.25	NaN	567595.43	567595.43	2011	San Francisco
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	NaN	538909.28	538909.28	2011	San Francisco
2	3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.60	NaN	335279.91	335279.91	2011	San Francisco
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.00	56120.71	198306.90	NaN	332343.61	332343.61	2011	San Francisco
4	5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT,(FIRE DEPARTMENT)	134401.60	9737.00	182234.59	NaN	326373.19	326373.19	2011	San Francisco
...	...	...	...	...	...	...	...	...	...	...	...
148649	148650	Roy I Tillery	Custodian	0.00	0.00	0.00	0.0	0.00	0.00	2014	San Francisco
148650	148651	Not provided	Not provided	NaN	NaN	NaN	NaN	0.00	0.00	2014	San Francisco
148651	148652	Not provided	Not provided	NaN	NaN	NaN	NaN	0.00	0.00	2014	San Francisco
148652	148653	Not provided	Not provided	NaN	NaN	NaN	NaN	0.00	0.00	2014	San Francisco
148653	148654	Joe Lopez	Counselor, Log Cabin Ranch	0.00	0.00	-618.13	0.0	-618.13	-618.13	2014	San Francisco

148654 rows × 11 columns

display number of unique values in each column

In [10]:
```
for x in df.columns:
    print(x, df[x].nunique())
```

```
Id 148654
EmployeeName 110811
JobTitle 2159
BasePay 109489
OvertimePay 65998
OtherPay 83225
Benefits 98465
TotalPay 138486
TotalPayBenefits 142098
Year 4
Agency 1
```

mean of total pay of all people based on year

In [11]:
```
df.groupby('Year')['TotalPay'].mean()
```

Out[11]:
```
Year
2011    71744.103871
2012    74113.262265
2013    77611.443142
2014    75463.918140
Name: TotalPay, dtype: float64
```

how many people have 0 overtime pay

In [12]:
```
len(df[df['OvertimePay']==0])
```

Out[12]:
```
77321
```

max, min, mean, median and other stats of TotalPay of people having 0 OvertimePay

In [13]:
```
df[df['OvertimePay']==0]['TotalPay'].describe()
```

Out[13]:
```
count     77321.000000
mean      60229.348901
std       49307.912350
min        -618.130000
25%       13290.450000
50%       58158.590000
75%       91115.090000
max      567595.430000
Name: TotalPay, dtype: float64
```

find Id of that person with max TotalPay you got in previous question

In [14]:
```
df[df['TotalPay']==567595.430000].index
```

Out[14]:
```
Int64Index([0], dtype='int64')
```

name of employee with total pay benefits = 87619.78

In [15]:
```
len(df[df['TotalPayBenefits']==87619.78])
```

Out[15]:
```
1
```

how many people have BasePay > 150000 and OvertimePay > 100000

In [16]:
```
len(df[(df['BasePay']>150000) & (df['OvertimePay']>100000)])
```

Out[16]:
```
12
```

which job title generally has highest average TotalPayBenefits

In [17]:
```
df[df['TotalPayBenefits']==max(df['TotalPayBenefits'])]['JobTitle']
```

Out[17]:
```
0    GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY
Name: JobTitle, dtype: object
```

How many employees are POLICE

In [18]:
```
# .str.contains()
len(df[df['JobTitle'].str.contains('POLICE')])
```

Out[18]:
```
2512
```