Load Necessary libraries

In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

Read Salaries.csv as a dataframe called sal.

In [8]:

```
#pd.read_csv('Salaries.csv')

# read a file from desktop/final
#pd.read_csv('C:\\Users\\hakim\\OneDrive\\Desktop\\final\\Salaries.csv')
sal = pd.read_csv(r'C:\Users\hakim\OneDrive\Desktop\final\Salaries.csv')
```

Check the head of the DataFrame.

In [9]:

```
# head will five 5 elements
sal.head()
```

Out[9]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPayBenefits	Year	Notes	Agency	Status
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.00	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	16452.60	NaN	335279.91	335279.91	2011	NaN	San Francisco	NaN
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.00	56120.71	198306.90	NaN	332343.61	332343.61	2011	NaN	San Francisco	NaN

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPayBenefits	Year	Notes	Agency	Status
4	5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT,(FIRE DEPARTMENT)	134401.60	9737.00	182234.59	NaN	326373.19	326373.19	2011	NaN	San Francisco	NaN

Check nan/missing values

In [11]:

```
sum([True,False,True])
```

Out[11]:

```
2
```

In [13]:

```
#sal.isna().sum()
sal.isnull().sum()
```

Out[13]:

```
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
```

total records , rows and columns

In [14]:

```
sal.shape
```

Out[14]:

```
(148654, 13)
```

check feature names

In [15]:

```
sal.columns
```

Out[15]:

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

drop Notes and Status columns bcz they have 100% missing values

In [18]:

```
sal.drop(columns=['Notes','Status'],inplace=True)
```

In [19]:

```
#check new shape
sal.shape
```

Out[19]:

```
(148654, 11)
```

In [20]:

```
#check columns
sal.columns
```

Out[20]:

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

fill BasePay NaN by using some technique

In [32]:

```
#sal.BasePay --> Series
#sal['BasePay'] --> Series
#sal[['BasePay']] --> df

# fill NaN by 0
#sal.BasePay.fillna(0)

# fill NaN by -1
#sal.BasePay.fillna(-1)

# fill NaN by mean() of BasePay column
#sal.BasePay.fillna(round(sal.BasePay.mean(),2))
```

```
mn = round(sal.BasePay.mean(),2)
sal.BasePay.fillna(mn)
# we can use inplace to store the changes permnt.
```

Out[32]:

```
0         167411.18
1         155966.02
2         212739.13
3          77916.00
4         134401.60
             ...
148649         0.00
148650     66325.45
148651     66325.45
148652     66325.45
148653         0.00
Name: BasePay, Length: 148654, dtype: float64
```

In [ ]: