# Importing Libraries

In [2]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

# Loading the dataset

In [3]:

```
sal_df = pd.read_csv("F://Big Data//Poc//Salaries.csv//Salaries.csv")
sal_df.head()
```

```
C:\Users\Pritesh\Anaconda3\lib\site-packages\IPython\core\interactiveshel
l.py:2698: DtypeWarning: Columns (3,4,5,6,12) have mixed types. Specify dt
ype option on import or set low_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
```

Out[3]:

| | Id | EmployeeName | JobTitle | BasePay | OvertimePay | OtherPay | Benefits | |
|---|---|---|---|---|---|---|---|---|
| **0** | 1 | NATHANIEL FORD | GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY | 167411 | 0 | 400184 | NaN | 5 |
| **1** | 2 | GARY JIMENEZ | CAPTAIN III (POLICE DEPARTMENT) | 155966 | 245132 | 137811 | NaN | 5 |
| **2** | 3 | ALBERT PARDINI | CAPTAIN III (POLICE DEPARTMENT) | 212739 | 106088 | 16452.6 | NaN | 3 |
| **3** | 4 | CHRISTOPHER CHONG | WIRE ROPE CABLE MAINTENANCE MECHANIC | 77916 | 56120.7 | 198307 | NaN | 3 |
| **4** | 5 | PATRICK GARDNER | DEPUTY CHIEF OF DEPARTMENT, (FIRE DEPARTMENT) | 134402 | 9737 | 182235 | NaN | 3 |

# Info about Dataset

In [4]:

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

# What are the descriptive statistics of the dataset

In [6]:

```
sal_df.describe()
```

Out[6]:

| | Id | TotalPay | TotalPayBenefits | Year | Notes |
|---|---|---|---|---|---|
| **count** | 148654.000000 | 148654.000000 | 148654.000000 | 148654.000000 | 0.0 |
| **mean** | 74327.500000 | 74768.321972 | 93692.554811 | 2012.522643 | NaN |
| **std** | 42912.857795 | 50517.005274 | 62793.533483 | 1.117538 | NaN |
| **min** | 1.000000 | -618.130000 | -618.130000 | 2011.000000 | NaN |
| **25%** | 37164.250000 | 36168.995000 | 44065.650000 | 2012.000000 | NaN |
| **50%** | 74327.500000 | 71426.610000 | 92404.090000 | 2013.000000 | NaN |
| **75%** | 111490.750000 | 105839.135000 | 132876.450000 | 2014.000000 | NaN |
| **max** | 148654.000000 | 567595.430000 | 567595.430000 | 2014.000000 | NaN |

# Find the Unique values in the dataset

In [7]:

```
sal_df.nunique()
```

Out[7]:

```
Id                148654
EmployeeName      110811
JobTitle            2159
BasePay           109900
OvertimePay        66555
OtherPay           84968
Benefits           99635
TotalPay          138486
TotalPayBenefits  142098
Year                   4
Notes                  0
Agency                 1
Status                 2
dtype: int64
```

# How many unique job titles are there?

In [14]:

```
sal_df['JobTitle'].nunique()
```

Out[14]:

```
2159
```

# What are the top 10 most common jobs?

In [15]:

```
sal_df['JobTitle'].value_counts().head(10)
```

Out[15]:

```
Transit Operator                7036
Special Nurse                   4389
Registered Nurse                3736
Public Svc Aide-Public Works    2518
Police Officer 3                2421
Custodian                       2418
TRANSIT OPERATOR                2388
Firefighter                     2359
Recreation Leader               1971
Patient Care Assistant          1945
Name: JobTitle, dtype: int64
```

# What is the job title of Nathaniel Ford and CHRISTOPHER CHONG?

In [34]:

```
sal_df[sal_df['EmployeeName']=='NATHANIEL FORD']['JobTitle']
```

Out[34]:

```
0    GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY
Name: JobTitle, dtype: object
```

In [29]:

```
sal_df[sal_df['EmployeeName']=='CHRISTOPHER CHONG']['JobTitle']
```

Out[29]:

```
3    WIRE ROPE CABLE MAINTENANCE MECHANIC
Name: JobTitle, dtype: object
```

# How much does CHRISTOPHER CHONG make (including benefits)?

In [36]:

```
sal_df[sal_df['EmployeeName']=='CHRISTOPHER CHONG']['TotalPayBenefits']
```

Out[36]:

```
3    332343.61
Name: TotalPayBenefits, dtype: float64
```

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

In [37]:

```
sal_df[sal_df['TotalPayBenefits']== sal_df['TotalPayBenefits'].max()]
```

Out[37]:

| | Id | EmployeeName | JobTitle | BasePay | OvertimePay | OtherPay | Benefits | |
|---|---|---|---|---|---|---|---|---|
| **0** | 1 | NATHANIEL FORD | GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY | 167411 | 0 | 400184 | NaN | 5 |

# What is the name of lowest paid person?

In [40]:

```
sal_df[sal_df['TotalPayBenefits']== sal_df['TotalPayBenefits'].min()]
```

Out[40]:

| | Id | EmployeeName | JobTitle | BasePay | OvertimePay | OtherPay | Benef |
|---|---|---|---|---|---|---|---|
| **148653** | 148654 | Joe Lopez | Counselor, Log Cabin Ranch | 0 | 0 | -618.13 | 0 |

# How many Job Titles were represented by only one person in 2011 to 2014?

In [46]:

```
sum(sal_df[sal_df['Year']==2011]['JobTitle'].value_counts() == 1)
```

Out[46]:

200

In [48]:

```
sum(sal_df[sal_df['Year']==2012]['JobTitle'].value_counts() == 1)
```

Out[48]:

190

In [49]:

```
sum(sal_df[sal_df['Year']==2013]['JobTitle'].value_counts() == 1)
```

Out[49]:

202

In [50]:

```
sum(sal_df[sal_df['Year']==2014]['JobTitle'].value_counts() == 1)
```

Out[50]:

175

# Maximum Salary received from 2011 to 2014?

In [65]:

```
sal_df.groupby('Year').max()
```

Out[65]:

| | Id | EmployeeName | JobTitle | TotalPay | TotalPayBenefits | Notes | Agenc |
|---|---|---|---|---|---|---|---|
| **Year** | | | | | | | |
| **2011** | 36159 | ZURI JONES | ZOO CURATOR | 567595.43 | 567595.43 | NaN | San Francis |
| **2012** | 72925 | Zuri Jones | Youth Comm Advisor | 362844.66 | 407274.78 | NaN | San Francis |
| **2013** | 110531 | Zuri Jones | Youth Comm Advisor | 347102.32 | 425815.28 | NaN | San Francis |
| **2014** | 148654 | Zuri Jones | Youth Comm Advisor | 471952.64 | 510732.68 | NaN | San Francis |

# What is the Mean TotalPay By Year?

In [66]:

```
sal_df[['Year', 'TotalPay']].groupby('Year').mean()
```

Out[66]:

| | TotalPay |
|---|---|
| **Year** | |
| **2011** | 71744.103871 |
| **2012** | 74113.262265 |
| **2013** | 77611.443142 |
| **2014** | 75463.918140 |

# VISUALIZATION OF DATA

# PLotting the Top 10 common jobs

In [72]:

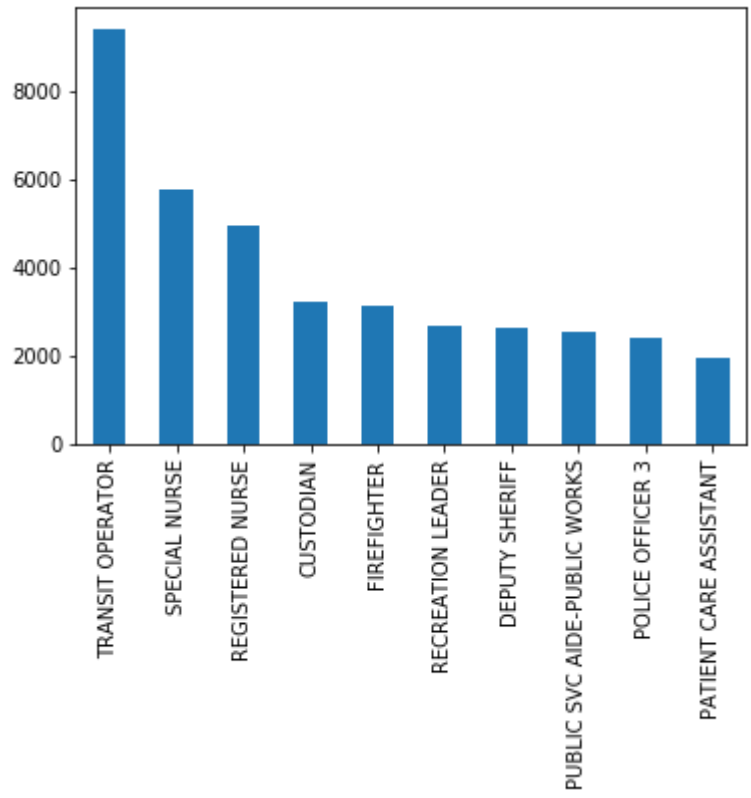
```
job_title_counts = sal_df['JobTitle'].value_counts()[:10]
print(job_title_counts)

#plot bar graph
job_title_counts.plot(kind = 'bar')
```

```
TRANSIT OPERATOR                9424
SPECIAL NURSE                   5791
REGISTERED NURSE                4955
CUSTODIAN                       3214
FIREFIGHTER                     3153
RECREATION LEADER               2663
DEPUTY SHERIFF                  2618
PUBLIC SVC AIDE-PUBLIC WORKS    2518
POLICE OFFICER 3                2421
PATIENT CARE ASSISTANT          1945
Name: JobTitle, dtype: int64
```

Out[72]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1aa0384e7b8>
```

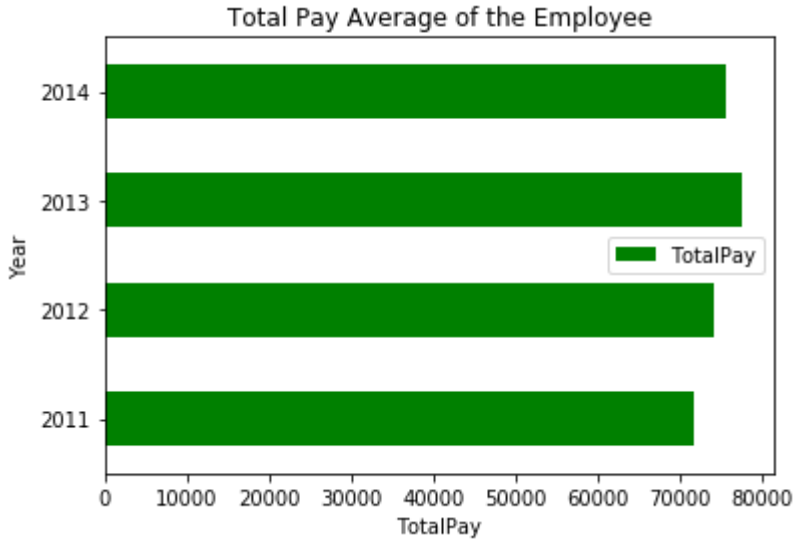



# Total Pay Average of the Employee by Year

In [95]:

```
df1=sal_df[['Year', 'TotalPay']].groupby('Year').mean()
print(df1)
plt1 = df1.plot(kind='barh',color='g');
plt.xlabel('TotalPay')
plt.ylabel('Year')
plt.title('Total Pay Average of the Employee')
print(plt1)
```

```
          TotalPay
Year
2011  71744.103871
2012  74113.262265
2013  77611.443142
2014  75463.918140
AxesSubplot(0.125,0.125;0.775x0.755)
```




# Total Pay of the Emplyee by job title

In [99]:

```
df2=sal_df[['TotalPayBenefits', 'JobTitle']].groupby('JobTitle').mean()
print(df2)
df3 = df2.ix[1:25]
plt2 = df3.plot(kind='bar',color='r');
plt.xlabel('TotalPay')
plt.ylabel('Year')
plt.title('Total Pay Average of the Employee')
print(plt2)
```

| JobTitle | TotalPayBenefits |
|---|---|
| ACCOUNT CLERK | 58212.534872 |
| ACCOUNTANT | 47429.268000 |
| ACCOUNTANT I | 88122.188750 |
| ACCOUNTANT II | 95086.024027 |
| ACCOUNTANT III | 107741.412158 |
| ACCOUNTANT INTERN | 48726.873796 |
| ACCOUNTANT IV | 124236.643275 |
| ACPO,JUVP, JUV PROB (SFERS) | 80266.370000 |
| ACUPUNCTURIST | 97055.530000 |
| ADM, SFGH MEDICAL CENTER | 347079.706667 |
| ADMIN ANALYST 3 | 94863.188333 |
| ADMIN HEARING EXAMINER | 66641.102258 |
| ADMINISTRATIVE ANALYST | 89983.720434 |
| ADMINISTRATIVE ANALYST I | 33384.780000 |
| ADMINISTRATIVE ANALYST II | 58915.042000 |
| ADMINISTRATIVE ANALYST III | 92698.515000 |
| ADMINISTRATIVE ENGINEER | 153759.660000 |
| ADMINISTRATIVE HEARING SUP | 131790.260000 |
| ADMINISTRATIVE SERVICES MANAGER | 77015.580000 |
| ADMINISTRATIVE SERVICES MGR | 125826.777273 |
| ADMINISTRATOR, DPH | 331564.035000 |
| ADMINISTRATOR, SFGH MEDICAL CENTER | 257124.440000 |
| ADMISSION ATTENDANT | 31550.802744 |
| AFFIRMATIVE ACTION SPECIALIST | 68213.983333 |
| AGRICULTURAL INSPECTOR | 81456.723750 |
| AIRPORT ASSISTANT DEPUTY DIRECTOR, BUSINESS ADMINI | 1927.500000 |
| AIRPORT ASSISTANT DEPUTY DIRECTOR, OPERATIONS | 15420.000000 |
| AIRPORT COMMUNICATIONS DISP | 109911.220235 |
| AIRPORT COMMUNICATIONS OFFICER | 122251.730000 |
| AIRPORT COMMUNICATIONS OPERATOR | 81214.809630 |
| ... | ... |
| WATER QUALITY TECH III | 98219.567083 |
| WATER QUALITY TECHNICIAN | 94591.214687 |
| WATER QUALITY TECHNICIAN I/II | 62157.373438 |
| WATER QUALITY TECHNICIAN III | 55968.794444 |
| WATER QUALITYTECH I/II | 91938.448696 |
| WATER SERVICE INSPECTOR | 113935.699727 |
| WATERSHED FORESTER | 140865.673333 |
| WATERSHED FORESTER MANAGER | 92877.545000 |
| WATERSHED KEEPER | 87672.161429 |
| WATERSHED KEEPER SUPERVISOR | 102931.722727 |
| WATERSHED WORKER (SEASONAL) | 9662.175690 |
| WELDER | 108096.224167 |
| WELFARE FRAUD INVESTIGATOR | 102839.558182 |
| WHARFINGER 1 | 71288.443333 |
| WHARFINGER 2 | 104724.598000 |
| WHARFINGER I | 65259.480000 |
| WHARFINGER II | 57837.853333 |
| WINDOW CLEANER | 85210.670882 |
| WINDOW CLEANER SUPERVISOR | 98453.107500 |
| WIRE ROPE CABLE MAINT MECHANIC | 138837.434333 |
| WIRE ROPE CABLE MAINT SPRV | 242118.323333 |
| WIRE ROPE CABLE MAINTENANCE MECHANIC | 145073.492500 |
| WIRE ROPE CABLE MAINTENANCE SUPERVISOR | 199628.970000 |
| WORKER'S COMP SUPERVISOR 1 | 96125.531429 |
| WORKER'S COMPENSATION ADJUSTER | 95269.564526 |
| WORKER'S COMPENSATION SUPERVISOR I | 91020.726000 |
| X-RAY LABORATORY AIDE | 66051.311190 |
| YOUTH COMM ADVISOR | 60118.550000 |

```
YOUTH COMMISSION ADVISOR, BOARD OF SUPERVISORS          53632.870000
ZOO CURATOR                                             66686.560000

[1637 rows x 1 columns]
AxesSubplot(0.125,0.125;0.775x0.755)

C:\Users\Pritesh\Anaconda3\lib\site-packages\ipykernel_launcher.py:3: Depr
ecationWarning:
.ix is deprecated. Please use
.loc for label based indexing or
.iloc for positional indexing

See the documentation here:
http://pandas.pydata.org/pandas-docs/stable/indexing.html#ix-indexer-is-de
precated
  This is separate from the ipykernel package so we can avoid doing import
s until
```