salary-analysis

November 16, 2024

[1]:

```python
import pandas as pd
```

[2]:

```python
sal_data=pd.read_csv('E:\Sql querry\Salaries.csv')
sal_data
```

[2]:

```
            Id        EmployeeName  \
0            1      NATHANIEL FORD
1            2        GARY JIMENEZ
2            3      ALBERT PARDINI
3            4   CHRISTOPHER CHONG
4            5     PATRICK GARDNER
…        …                 …
148649  148650       Roy I Tillery
148650  148651        Not provided
148651  148652        Not provided
148652  148653        Not provided
148653  148654           Joe Lopez

                                               JobTitle    BasePay  \
0       GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY  167411.18
1                      CAPTAIN III (POLICE DEPARTMENT)  155966.02
2                      CAPTAIN III (POLICE DEPARTMENT)  212739.13
3                 WIRE ROPE CABLE MAINTENANCE MECHANIC   77916.00
4         DEPUTY CHIEF OF DEPARTMENT,(FIRE DEPARTMENT)  134401.60
…                                                 …        …
148649                                       Custodian       0.00
148650                                    Not provided        NaN
148651                                    Not provided        NaN
148652                                    Not provided        NaN
148653                      Counselor, Log Cabin Ranch       0.00

        OvertimePay   OtherPay  Benefits   TotalPay  TotalPayBenefits  Year   \
0              0.00  400184.25       NaN  567595.43         567595.43  2011
1         245131.88  137811.38       NaN  538909.28         538909.28  2011
2         106088.18   16452.60       NaN  335279.91         335279.91  2011
3          56120.71  198306.90       NaN  332343.61         332343.61  2011
4           9737.00  182234.59       NaN  326373.19         326373.19  2011
```

```
…             …        …       …        …               …   …
148649         0.00       0.00       0.0        0.00              0.00  2014
148650          NaN        NaN       NaN        0.00              0.00  2014
148651          NaN        NaN       NaN        0.00              0.00  2014
148652          NaN        NaN       NaN        0.00              0.00  2014
148653         0.00    -618.13       0.0     -618.13           -618.13  2014

        Notes         Agency  Status
0         NaN  San Francisco     NaN
1         NaN  San Francisco     NaN
2         NaN  San Francisco     NaN
3         NaN  San Francisco     NaN
4         NaN  San Francisco     NaN
…       …            …     …
148649    NaN  San Francisco     NaN
148650    NaN  San Francisco     NaN
148651    NaN  San Francisco     NaN
148652    NaN  San Francisco     NaN
148653    NaN  San Francisco     NaN

[148654 rows x 13 columns]
```

[3]:
```
# 1 - Diaplay 10 rows of dataset
```

[4]:
```
sal_data.head(10)
```

[4]:
```
   Id        EmployeeName                                        JobTitle  \
0   1      NATHANIEL FORD  GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY
1   2        GARY JIMENEZ                 CAPTAIN III (POLICE DEPARTMENT)
2   3      ALBERT PARDINI                 CAPTAIN III (POLICE DEPARTMENT)
3   4   CHRISTOPHER CHONG            WIRE ROPE CABLE MAINTENANCE MECHANIC
4   5     PATRICK GARDNER    DEPUTY CHIEF OF DEPARTMENT,(FIRE DEPARTMENT)
5   6      DAVID SULLIVAN                       ASSISTANT DEPUTY CHIEF II
6   7           ALSON LEE              BATTALION CHIEF, (FIRE DEPARTMENT)
7   8       DAVID KUSHNER                  DEPUTY DIRECTOR OF INVESTMENTS
8   9      MICHAEL MORRIS              BATTALION CHIEF, (FIRE DEPARTMENT)
9  10  JOANNE HAYES-WHITE          CHIEF OF DEPARTMENT, (FIRE DEPARTMENT)

     BasePay  OvertimePay   OtherPay  Benefits   TotalPay  TotalPayBenefits  \
0  167411.18         0.00  400184.25       NaN  567595.43         567595.43
1  155966.02    245131.88  137811.38       NaN  538909.28         538909.28
2  212739.13    106088.18   16452.60       NaN  335279.91         335279.91
3   77916.00     56120.71  198306.90       NaN  332343.61         332343.61
4  134401.60      9737.00  182234.59       NaN  326373.19         326373.19
5  118602.00      8601.00  189082.74       NaN  316285.74         316285.74
6   92492.01     89062.90  134426.14       NaN  315981.05         315981.05
7  256576.96         0.00   51322.50       NaN  307899.46         307899.46
```

```
8  176932.64     86362.68   40132.23       NaN  303427.55         303427.55
9  285262.00         0.00   17115.73       NaN  302377.73         302377.73

   Year  Notes         Agency  Status
0  2011    NaN  San Francisco     NaN
1  2011    NaN  San Francisco     NaN
2  2011    NaN  San Francisco     NaN
3  2011    NaN  San Francisco     NaN
4  2011    NaN  San Francisco     NaN
5  2011    NaN  San Francisco     NaN
6  2011    NaN  San Francisco     NaN
7  2011    NaN  San Francisco     NaN
8  2011    NaN  San Francisco     NaN
9  2011    NaN  San Francisco     NaN
```

[5]:
```
# 2 - Display last 10 rows of dataset
```

[6]:
```
sal_data.tail(10)
```

[6]:
```
            Id      EmployeeName                      JobTitle  BasePay  \
148644  148645      Randy D Winn  Stationary Eng, Sewage Plant      0.0
148645  148646  Carolyn A Wilson     Human Services Technician      0.0
148646  148647      Not provided                  Not provided      NaN
148647  148648    Joann Anderson   Communications Dispatcher 2      0.0
148648  148649       Leon Walker                     Custodian      0.0
148649  148650     Roy I Tillery                     Custodian      0.0
148650  148651      Not provided                  Not provided      NaN
148651  148652      Not provided                  Not provided      NaN
148652  148653      Not provided                  Not provided      NaN
148653  148654         Joe Lopez    Counselor, Log Cabin Ranch      0.0

        OvertimePay  OtherPay  Benefits  TotalPay  TotalPayBenefits  Year  \
148644          0.0      0.00       0.0      0.00              0.00  2014
148645          0.0      0.00       0.0      0.00              0.00  2014
148646          NaN       NaN       NaN      0.00              0.00  2014
148647          0.0      0.00       0.0      0.00              0.00  2014
148648          0.0      0.00       0.0      0.00              0.00  2014
148649          0.0      0.00       0.0      0.00              0.00  2014
148650          NaN       NaN       NaN      0.00              0.00  2014
148651          NaN       NaN       NaN      0.00              0.00  2014
148652          NaN       NaN       NaN      0.00              0.00  2014
148653          0.0   -618.13       0.0   -618.13           -618.13  2014

        Notes         Agency  Status
148644    NaN  San Francisco     NaN
148645    NaN  San Francisco     NaN
148646    NaN  San Francisco     NaN
```

```
148647    NaN  San Francisco     NaN
148648    NaN  San Francisco     NaN
148649    NaN  San Francisco     NaN
148650    NaN  San Francisco     NaN
148651    NaN  San Francisco     NaN
148652    NaN  San Francisco     NaN
148653    NaN  San Francisco     NaN
```

[7]:

```
# 3 - find shape of dataset with number of rows and columns
```

[8]:

```
sal_data.shape
```

[8]:

```
(148654, 13)
```

[9]:

```
print("no of rows :",sal_data.shape[0])
print("no of columns :",sal_data.shape[1])
```

```
no of rows : 148654
no of columns : 13
```

[10]:

```
# 4 - get information of data
```

[11]:

```
sal_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           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
```

[12]:

```
# check null values in dataset
```

[13]:
```
sal_data.isnull()
```

[13]:
```
           Id  EmployeeName  JobTitle  BasePay  OvertimePay  OtherPay  \
0       False         False     False    False        False     False
1       False         False     False    False        False     False
2       False         False     False    False        False     False
3       False         False     False    False        False     False
4       False         False     False    False        False     False
…       …           …       …      …          …       …
148649  False         False     False    False        False     False
148650  False         False     False     True         True      True
148651  False         False     False     True         True      True
148652  False         False     False     True         True      True
148653  False         False     False    False        False     False

        Benefits  TotalPay  TotalPayBenefits   Year  Notes  Agency  Status
0           True     False             False  False   True   False    True
1           True     False             False  False   True   False    True
2           True     False             False  False   True   False    True
3           True     False             False  False   True   False    True
4           True     False             False  False   True   False    True
…          …       …                …    …    …     …     …
148649     False     False             False  False   True   False    True
148650      True     False             False  False   True   False    True
148651      True     False             False  False   True   False    True
148652      True     False             False  False   True   False    True
148653     False     False             False  False   True   False    True

[148654 rows x 13 columns]
```

[14]:
```
sal_data.isnull().sum()
```

[14]:
```
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
```

```
[15]: # Drop  Notes,Agency and Status.
```

```
[16]: sal_data.columns
```

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

```
[17]: sal_data=sal_data.drop(['Notes','Agency','Status'],axis=1)
```

```
[18]: sal_data.columns
```

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

```
[19]: # Get overall statistic of dataset
```

```
[20]: sal_data.describe()
```

```
[20]:                   Id        BasePay    OvertimePay       OtherPay  \
      count  148654.000000  148045.000000  148650.000000  148650.000000
      mean    74327.500000   66325.448840    5066.059886    3648.767297
      std     42912.857795   42764.635495   11454.380559    8056.601866
      min         1.000000    -166.010000      -0.010000   -7058.590000
      25%     37164.250000   33588.200000       0.000000       0.000000
      50%     74327.500000   65007.450000       0.000000     811.270000
      75%    111490.750000   94691.050000    4658.175000    4236.065000
      max    148654.000000  319275.010000  245131.880000  400184.250000

                  Benefits       TotalPay  TotalPayBenefits           Year
      count  112491.000000  148654.000000     148654.000000  148654.000000
      mean    25007.893151   74768.321972      93692.554811    2012.522643
      std     15402.215858   50517.005274      62793.533483       1.117538
      min       -33.890000    -618.130000       -618.130000    2011.000000
      25%     11535.395000   36168.995000      44065.650000    2012.000000
      50%     28628.620000   71426.610000      92404.090000    2013.000000
      75%     35566.855000  105839.135000     132876.450000    2014.000000
      max     96570.660000  567595.430000     567595.430000    2014.000000
```

```
[21]: sal_data.describe(include='all')
```

```
[21]:                    Id EmployeeName          JobTitle        BasePay  \
      count   148654.000000       148654            148654  148045.000000
      unique            NaN       110811              2159            NaN
      top               NaN    Kevin Lee  Transit Operator            NaN
```

```
freq              NaN           13              7036            NaN
mean     74327.500000          NaN               NaN   66325.448840
std      42912.857795          NaN               NaN   42764.635495
min          1.000000          NaN               NaN    -166.010000
25%      37164.250000          NaN               NaN   33588.200000
50%      74327.500000          NaN               NaN   65007.450000
75%     111490.750000          NaN               NaN   94691.050000
max     148654.000000          NaN               NaN  319275.010000

          OvertimePay       OtherPay       Benefits       TotalPay  \
count   148650.000000  148650.000000  112491.000000  148654.000000
unique            NaN            NaN            NaN            NaN
top               NaN            NaN            NaN            NaN
freq              NaN            NaN            NaN            NaN
mean      5066.059886    3648.767297   25007.893151   74768.321972
std      11454.380559    8056.601866   15402.215858   50517.005274
min         -0.010000   -7058.590000     -33.890000    -618.130000
25%          0.000000       0.000000   11535.395000   36168.995000
50%          0.000000     811.270000   28628.620000   71426.610000
75%       4658.175000    4236.065000   35566.855000  105839.135000
max     245131.880000  400184.250000   96570.660000  567595.430000

        TotalPayBenefits           Year
count      148654.000000  148654.000000
unique               NaN            NaN
top                  NaN            NaN
freq                 NaN            NaN
mean        93692.554811    2012.522643
std         62793.533483       1.117538
min          -618.130000    2011.000000
25%         44065.650000    2012.000000
50%         92404.090000    2013.000000
75%        132876.450000    2014.000000
max        567595.430000    2014.000000
```

[22]:
```
# find all occurance of all employe name of top 5
```

[23]:
```
sal_data['EmployeeName'].value_counts().head()
```

[23]:
```
EmployeeName
Kevin Lee        13
Richard Lee      11
Steven Lee       11
William Wong     11
Stanley Lee       9
Name: count, dtype: int64
```

[24]:

```
# Find number of unique job title
```

[25]:

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

[25]:

```
2159
```

[26]:

```
# Display total no of job titles contains captain
```

[27]:

```
sal_data[sal_data['JobTitle'].str.contains('captain',case=False)].count()␣
 ↪['JobTitle']
```

[27]:

```
552
```

[28]:

```
#Display allemployee name from fire department
```

[29]:

```
sal_data.columns
```

[29]:

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

[30]:

```
sal_data[sal_data['JobTitle'].str.contains('fire',case=False)]['EmployeeName']
```

[30]:

```
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
```

[31]:

```
# Find minimum ,maximum,avg of basepay
```

[32]:

```
sal_data.columns
```

[32]:

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

[33]:

```
sal_data['BasePay'].max()
```

[33]:

```
319275.01
```

```
[34]: sal_data['BasePay'].min()
```

```
[34]: -166.01
```

```
[35]: sal_data['BasePay'].mean()
```

```
[35]: 66325.4488404877
```

```
[36]: sal_data['BasePay'].describe()
```

```
[36]: count    148045.000000
      mean      66325.448840
      std       42764.635495
      min        -166.010000
      25%       33588.200000
      50%       65007.450000
      75%       94691.050000
      max      319275.010000
      Name: BasePay, dtype: float64
```

```
[37]: # 13- Display 'not provided' to NaN from 'Employee Name' column
```

```
[38]: import numpy as np
```

```
[39]: sal_data.columns
```

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

```
[40]: sal_data['EmployeeName']
```

```
[40]: 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
```

```
[41]: sal_data['EmployeeName'].replace('Not provided',np.nan)
```

```
[41]: 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
```

```
[42]: sal_data['EmployeeName']=sal_data['EmployeeName'].replace('Not provided',np.nan)
```

```
[43]: sal_data['EmployeeName']
```

```
[43]: 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
```

```
[44]: # 14- Drop the Rows Having 5 missing values
```

```
[45]: sal_data.drop(sal_data[sal_data.isnull().sum(axis=1)==5].
       ↪index,axis=0,inplace=True)
```

```
[46]: sal_data.isnull().sum(axis=1)
```

```
[46]: 0         1
      1         1
      2         1
      3         1
      4         1
               ..
      148645    0
      148647    0
      148648    0
      148649    0
```

```
148653    0
Length: 148650, dtype: int64
```

```
[47]: # 15- Find job title of Albert pardini
```

```
[48]: sal_data[sal_data['EmployeeName']=='ALBERT PARDINI']['JobTitle']
```

```
[48]: 2    CAPTAIN III (POLICE DEPARTMENT)
      Name: JobTitle, dtype: object
```

```
[49]: # 16- How much ALBERT PARDINI make( include Benefits)?
```

```
[50]: sal_data.columns
```

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

```
[51]: sal_data[sal_data['EmployeeName']=='ALBERT PARDINI']['TotalPayBenefits']
```

```
[51]: 2    335279.91
      Name: TotalPayBenefits, dtype: float64
```

```
[52]: # 17- Display the name of the person having highest Basepay?
```

```
[53]: sal_data[sal_data['BasePay'].max()==sal_data['BasePay']]['EmployeeName']
```

```
[53]: 72925    Gregory P Suhr
      Name: EmployeeName, dtype: object
```

```
[54]: # 18 - Display rop 5 common job
```

```
[55]: sal_data['JobTitle'].value_counts().head()
```

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

```
[56]: # 19- Find avg Basepay of emp having job title Accountant
```

```
[57]: sal_data[sal_data['JobTitle']=='ACCOUNTANT']['BasePay'].mean()
```

```
[57]: 46643.172
```

[ ]:

```
```