Data_Analysis_SF_Salaries

February 21, 2020

1 SF _ Salaries Data Analysis

[1]:
```
import numpy as np
import pandas as pd
```

[ ]:
```
sal = pd.read_csv('Salaries.csv')
```

[18]:
```
sal
```

[18]:

```
            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
1                      CAPTAIN III (POLICE DEPARTMENT)        155966
2                      CAPTAIN III (POLICE DEPARTMENT)        212739
3                  WIRE ROPE CABLE MAINTENANCE MECHANIC         77916
4           DEPUTY CHIEF OF DEPARTMENT,(FIRE DEPARTMENT)        134402
…                                                    …             …
148649                                       Custodian          0.00
148650                                   Not provided  Not Provided
148651                                   Not provided  Not Provided
148652                                   Not provided  Not Provided
148653                      Counselor, Log Cabin Ranch          0.00

         OvertimePay      OtherPay      Benefits   TotalPay  TotalPayBenefits  \
0                  0        400184           NaN  567595.43         567595.43
```

```
1             245132        137811           NaN  538909.28         538909.28
2             106088       16452.6           NaN  335279.91         335279.91
3            56120.7        198307           NaN  332343.61         332343.61
4               9737        182235           NaN  326373.19         326373.19
…                …             …             …         …                 …
148649          0.00          0.00          0.00       0.00              0.00
148650  Not Provided  Not Provided  Not Provided       0.00              0.00
148651  Not Provided  Not Provided  Not Provided       0.00              0.00
148652  Not Provided  Not Provided  Not Provided       0.00              0.00
148653          0.00       -618.13          0.00    -618.13           -618.13

        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
…       …     …             …     …
148649  2014    NaN  San Francisco     PT
148650  2014    NaN  San Francisco    NaN
148651  2014    NaN  San Francisco    NaN
148652  2014    NaN  San Francisco    NaN
148653  2014    NaN  San Francisco     PT

[148654 rows x 13 columns]
```

[19]:
```
sal.head()
```

[19]:
```
   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)

  BasePay OvertimePay OtherPay Benefits   TotalPay  TotalPayBenefits  Year  \
0  167411           0   400184      NaN  567595.43         567595.43  2011
1  155966      245132   137811      NaN  538909.28         538909.28  2011
2  212739      106088  16452.6      NaN  335279.91         335279.91  2011
3   77916     56120.7   198307      NaN  332343.61         332343.61  2011
4  134402        9737   182235      NaN  326373.19         326373.19  2011

   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
```

[76]:
```
sal.tail(10)
```

[76]:

	Id	EmployeeName	JobTitle	BasePay	\
148644	148645	Randy D Winn	Stationary Eng, Sewage Plant	0.00	
148645	148646	Carolyn A Wilson	Human Services Technician	0.00	
148646	148647	Not provided	Not provided	Not Provided	
148647	148648	Joann Anderson	Communications Dispatcher 2	0.00	
148648	148649	Leon Walker	Custodian	0.00	
148649	148650	Roy I Tillery	Custodian	0.00	
148650	148651	Not provided	Not provided	Not Provided	
148651	148652	Not provided	Not provided	Not Provided	
148652	148653	Not provided	Not provided	Not Provided	
148653	148654	Joe Lopez	Counselor, Log Cabin Ranch	0.00	

	OvertimePay	OtherPay	Benefits	TotalPay	TotalPayBenefits	\
148644	0.00	0.00	0.00	0.00	0.00	
148645	0.00	0.00	0.00	0.00	0.00	
148646	Not Provided	Not Provided	Not Provided	0.00	0.00	
148647	0.00	0.00	0.00	0.00	0.00	
148648	0.00	0.00	0.00	0.00	0.00	
148649	0.00	0.00	0.00	0.00	0.00	
148650	Not Provided	Not Provided	Not Provided	0.00	0.00	
148651	Not Provided	Not Provided	Not Provided	0.00	0.00	
148652	Not Provided	Not Provided	Not Provided	0.00	0.00	
148653	0.00	-618.13	0.00	-618.13	-618.13	

	Year	Notes	Agency	Status	title_len
148644	2014	NaN	San Francisco	PT	28
148645	2014	NaN	San Francisco	PT	25
148646	2014	NaN	San Francisco	NaN	12
148647	2014	NaN	San Francisco	PT	27
148648	2014	NaN	San Francisco	PT	9
148649	2014	NaN	San Francisco	PT	9
148650	2014	NaN	San Francisco	NaN	12
148651	2014	NaN	San Francisco	NaN	12
148652	2014	NaN	San Francisco	NaN	12
148653	2014	NaN	San Francisco	PT	26

[6]:
```
sal.info() # no. of column and no. of entry rows
```

```
<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
```

[20]:
```
sal['BasePay']
```

[20]:
```
0               167411
1               155966
2               212739
3                77916
4               134402
               ...
148649            0.00
148650    Not Provided
148651    Not Provided
148652    Not Provided
148653            0.00
Name: BasePay, Length: 148654, dtype: object
```

[33]:
```
sal['TotalPay'].max()
```

[33]:
```
567595.43
```

[44]:
```
sal[sal['EmployeeName'] == 'ALBERT PARDINI']
```

[44]:
```
    Id    EmployeeName                          JobTitle BasePay OvertimePay  \
2   3  ALBERT PARDINI  CAPTAIN III (POLICE DEPARTMENT)  212739      106088

  OtherPay Benefits   TotalPay  TotalPayBenefits  Year  Notes         Agency  \
2  16452.6      NaN  335279.91         335279.91  2011    NaN  San Francisco

  Status
2    NaN
```

[46]:
```
sal['TotalPayBenefits'] == sal['TotalPayBenefits'].max()
```

[46]:
```
0          True
1         False
2         False
3         False
4         False
          …
148649    False
148650    False
148651    False
148652    False
148653    False
Name: TotalPayBenefits, Length: 148654, dtype: bool
```

[47]:
```
sal.loc[sal['TotalPayBenefits'].idxmax()]

# sal.iloc[sal['TotalPayBenefits'].argmax()]
```

[47]:
```
Id                                                          1
EmployeeName                                     NATHANIEL FORD
JobTitle            GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY
BasePay                                                167411
OvertimePay                                                 0
OtherPay                                               400184
Benefits                                                  NaN
TotalPay                                               567595
TotalPayBenefits                                       567595
Year                                                     2011
Notes                                                     NaN
Agency                                           San Francisco
Status                                                    NaN
Name: 0, dtype: object
```

[48]:
```
sal.iloc[sal['TotalPayBenefits'].argmin()]
# sal[sal['TotalPayBenefits'] == sal['TotalPayBenefits'].min()]
#lowest paid person including benefits
```

[48]:
```
Id                                  148654
EmployeeName                     Joe Lopez
JobTitle            Counselor, Log Cabin Ranch
BasePay                               0.00
OvertimePay                           0.00
OtherPay                           -618.13
Benefits                              0.00
TotalPay                           -618.13
TotalPayBenefits                   -618.13
Year                                  2014
Notes                                  NaN
```

```
Agency                            San Francisco
Status                                       PT
Name: 148653, dtype: object
```

[55]:
```
sal.groupby('Year').mean()
# average mean all employees per year
```

[55]:
```
             Id      TotalPay  TotalPayBenefits  Notes
Year
2011    18080.0  71744.103871      71744.103871    NaN
2012    54542.5  74113.262265     100553.229232    NaN
2013    91728.5  77611.443142     101440.519714    NaN
2014   129593.0  75463.918140     100250.918884    NaN
```

[57]:
```
sal['JobTitle'].nunique() #unique job title
```

[57]:
```
2159
```

[58]:
```
sal['JobTitle'].value_counts().head(5) # top 5 most common jobs
```

[58]:
```
Transit Operator                7036
Special Nurse                   4389
Registered Nurse                3736
Public Svc Aide-Public Works    2518
Police Officer 3                2421
Name: JobTitle, dtype: int64
```

[61]:
```
sum(sal[sal['Year']==2013]['JobTitle'].value_counts() ==1)
#How many job titles were represented by only one person in 2013
```

[61]:
```
202
```

[62]:
```
def chief_string(title): #upper and lower case that's why used functions
    if 'chief' in title.lower().split():
        return True
    else:
        return False
```

[ ]:
```

```

[67]:
```
sal['JobTitle'].iloc[0]
```

[67]:
```
'GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY'
```

[72]:
```
sal['title_len']= sal['JobTitle'].apply(len)
```

[73]:
```
sal[['JobTitle','title_len']]
```

[73]:

	JobTitle	title_len
0	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	46
1	CAPTAIN III (POLICE DEPARTMENT)	31
2	CAPTAIN III (POLICE DEPARTMENT)	31
3	WIRE ROPE CABLE MAINTENANCE MECHANIC	36
4	DEPUTY CHIEF OF DEPARTMENT,(FIRE DEPARTMENT)	44
…	…	…
148649	Custodian	9
148650	Not provided	12
148651	Not provided	12
148652	Not provided	12
148653	Counselor, Log Cabin Ranch	26

[148654 rows x 2 columns]

[74]:
```
sal[['JobTitle','title_len']].corr() #correlation
```

[74]:

	title_len
title_len	1.0

[75]:
```
sal[['TotalPayBenefits','title_len']].corr() # no-correlation
```

[75]:

	TotalPayBenefits	title_len
TotalPayBenefits	1.000000	-0.036878
title_len	-0.036878	1.000000

[ ]:
```

```