SF Salaries Analysis

Explore San Francisco city employee salary data.

Datasets

I will be using the SF Salaries Dataset

About the Datasets

One way to understand how a city government works is by looking at who it employs and how its employees are compensated. This data contains the names, job title, and compensation for San Francisco city employees on an annual basis from 2011 to 2014.

Import pandas as pd

In [2]:
```python
import numpy as np
import pandas as pd
```

Read Salaries.csv as a dataframe called sal.

In [3]:
```python
sal = pd.read_csv("Salaries.csv")
sal
```

Out[3]:

	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
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
...	...	...	...	...	...	...	...	...	...	...	...	...	...
148649	148650	Roy I Tillery	Custodian	0.00	0.00	0.00	0.0	0.00	0.00	2014	NaN	San Francisco	NaN
148650	148651	Not provided	Not provided	NaN	NaN	NaN	NaN	0.00	0.00	2014	NaN	San Francisco	NaN
148651	148652	Not provided	Not provided	NaN	NaN	NaN	NaN	0.00	0.00	2014	NaN	San Francisco	NaN
148652	148653	Not provided	Not provided	NaN	NaN	NaN	NaN	0.00	0.00	2014	NaN	San Francisco	NaN
148653	148654	Joe Lopez	Counselor, Log Cabin Ranch	0.00	0.00	-618.13	0.0	-618.13	-618.13	2014	NaN	San Francisco	NaN

148654 rows × 13 columns

Check the head of the DataFrame.

In [4]:
```python
sal.head()
```

Out[4]:

	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
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

Use the .info() method to find out how many entries there are.

In [5]:
```python
sal.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
```

1. What is the average BasePay ?

In [6]:
```python
sal["BasePay"].mean()
```

Out[6]:
```
66325.44884050643
```

> The average BasePay is **$66325.45**

2. What is the highest amount of OvertimePay in the dataset?

In [7]:
```python
sal["OvertimePay"].max()
```

Out[7]:
```
245131.88
```

> The highest amount of OvertimePay is **$245131.88**

3. What is the job title of JOSEPH DRISCOLL ? Note: Use all caps, otherwise you may get an answer that doesn't match up (there is also a lowercase Joseph Driscoll).

In [8]:
```python
sal[sal["EmployeeName"] == "JOSEPH DRISCOLL"]["JobTitle"]
```

Out[8]:
```
24    CAPTAIN, FIRE SUPPRESSION
Name: JobTitle, dtype: object
```

> JOSEPH DRISCOLL is **a CAPTAIN, FIRE SUPPRESSION**

4. How much does JOSEPH DRISCOLL make (including benefits)?

In [9]:
```python
sal[sal["EmployeeName"] == "JOSEPH DRISCOLL"]["TotalPayBenefits"]
```

Out[9]:
```
24    270324.91
Name: TotalPayBenefits, dtype: float64
```

> JOSEPH DRISCOLL make **$270324.91**

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

In [10]:
```python
sal[sal["TotalPayBenefits"] == sal["TotalPayBenefits"].max()]
```

Out[10]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPayBenefits	Year	Notes	Agency	Status
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.0	400184.25	NaN	567595.43	567595.43	2011	NaN	San Francisco	NaN

In [11]:
```python
sal.loc[sal['TotalPayBenefits'].idxmax()]
```

Out[11]:
```
Id                                                            1
EmployeeName                                     NATHANIEL FORD
JobTitle            GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY
BasePay                                               167411.18
OvertimePay                                                 0.0
OtherPay                                              400184.25
Benefits                                                    NaN
TotalPay                                              567595.43
TotalPayBenefits                                      567595.43
Year                                                       2011
Notes                                                       NaN
Agency                                            San Francisco
Status                                                      NaN
Name: 0, dtype: object
```

In [12]:
```python
sal.iloc[sal['TotalPayBenefits'].argmax()]
```

Out[12]:
```
Id                                                            1
EmployeeName                                     NATHANIEL FORD
JobTitle            GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY
BasePay                                               167411.18
OvertimePay                                                 0.0
OtherPay                                              400184.25
Benefits                                                    NaN
TotalPay                                              567595.43
TotalPayBenefits                                      567595.43
Year                                                       2011
Notes                                                       NaN
Agency                                            San Francisco
Status                                                      NaN
Name: 0, dtype: object
```

> The person who has the highest pay in San Francisco is **NATHANIEL FORD** with **$567595.43**

6. What is the name of lowest paid person (including benefits)? Do you notice something strange about how much he or she is paid?

In [13]:
```python
sal[sal["TotalPayBenefits"] == sal["TotalPayBenefits"].min()]
```

Out[13]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPayBenefits	Year	Notes	Agency	Status
148653	148654	Joe Lopez	Counselor, Log Cabin Ranch	0.0	0.0	-618.13	0.0	-618.13	-618.13	2014	NaN	San Francisco	NaN

> The person who has the lowest pay in San Francisco is **Joe Lopez** with **$-618.13**

7. What was the average (mean) BasePay of all employees per year? (2011-2014)?

In [14]:
```python
sal.head()
```

Out[14]:

	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
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

In [15]:
```python
sal["Year"].unique()
```

Out[15]:
```
array([2011, 2012, 2013, 2014], dtype=int64)
```

In [16]:
```python
sal[["BasePay", "Year"]].groupby("Year").mean()
```

Out[16]:

Year	BasePay
2011	63595.956517
2012	65436.406857
2013	69630.030216
2014	66564.421924

8. How many unique job titles are there?

In [39]:
```python
sal["JobTitle"].nunique()
```

Out[39]:
```
2159
```

> There are **2159** distinct job titles in San Francisco.

9. What are the top 5 most common jobs?

In [53]:
```python
sal["JobTitle"].value_counts().head(5)
```

Out[53]:
```
Transit Operator                7036
Special Nurse                   4389
Registered Nurse                3736
Public Svc Aide-Public Works    2518
Police Officer 3                2421
Name: JobTitle, dtype: int64
```

In [20]:
```python
sal["JobTitle"].value_counts().sort_values(ascending= False).head(10)
```

Out[20]:
```
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
```

> **The 5 Common Jobs in San Francisco**
>
> 1. Transit Operator 7036
> 2. Special Nurse 4389
> 3. Registered Nurse 3736
> 4. Public Svc Aide-Public Works 2518
> 5. Police Officer 3 2421

10. How many Job Titles were represented by only one person in 2013? (e.g. Job Titles with only one occurence in 2013?)

In [23]:
```python
sum(sal[sal["Year"] == 2013]["JobTitle"].value_counts() == 1)
```

Out[23]:
```
202
```

> In 2013, there were **202** job titles represented by only one person in San Francisco.

11. How many people have the word Chief in their job title? (This is pretty tricky)

In [40]:
```python
sum(sal["JobTitle"].str.contains("Chief", case=False))
```

Out[40]:
```
627
```

In [39]:
```python
sal[sal["JobTitle"].str.contains("Chief", case=False)]
```

Out[39]:

	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
5	6	DAVID SULLIVAN	ASSISTANT DEPUTY CHIEF II	118602.00	8601.00	189082.74	NaN	316285.74	316285.74	2011	NaN	San Francisco	NaN
6	7	ALSON LEE	BATTALION CHIEF, (FIRE DEPARTMENT)	92492.01	89062.90	134426.14	NaN	315981.05	315981.05	2011	NaN	San Francisco	NaN
8	9	MICHAEL MORRIS	BATTALION CHIEF, (FIRE DEPARTMENT)	176932.64	86362.68	40132.23	NaN	303427.55	303427.55	2011	NaN	San Francisco	NaN
9	10	JOANNE HAYES-WHITE	CHIEF OF DEPARTMENT, (FIRE DEPARTMENT)	285262.00	0.00	17115.73	NaN	302377.73	302377.73	2011	NaN	San Francisco	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...
138011	138012	Sharon C Jackson	Chief Deputy Adlt Probation Of	56660.18	0.00	0.00	566.61	56660.18	57226.79	2014	NaN	San Francisco	NaN
139620	139621	Anita S Lee	Chief Clerk	38598.86	0.00	0.00	386.00	38598.86	38984.86	2014	NaN	San Francisco	NaN
139709	139710	George D Vaughan	Chief Nursery Specialist	27319.47	483.90	0.00	10334.67	27803.37	38138.04	2014	NaN	San Francisco	NaN
140080	140081	Patrick W Gillespie	Chief Preparator	24091.00	0.00	0.00	10176.77	24091.00	34267.77	2014	NaN	San Francisco	NaN
140156	140157	Christopher L Ochoa	Chief Stationary Engineer	24392.00	418.71	0.00	8918.44	24810.71	33729.15	2014	NaN	San Francisco	NaN

627 rows × 13 columns

In [41]:
```python
len(sal[sal["JobTitle"].str.contains("Chief", case=False)])
```

Out[41]:
```
627
```

In [42]:
```python
def chief_string(title):
    if 'chief' in title.lower():
        return True
    else:
        return False
```

In [43]:
```python
sum(sal['JobTitle'].apply(lambda x: chief_string(x)))
```

Out[43]:
```
627
```

> The number of individuals with the word 'Chief' in their job title is **627** in San Francisco.

12. Is there a correlation between length of the Job Title string and Salary?

In [44]:
```python
sal['title_len'] = sal['JobTitle'].apply(len)
```

In [45]:
```python
sal.head()
```

Out[45]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPayBenefits	Year	Notes	Agency	Status	title_len
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	46
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	NaN	538909.28	538909.28	2011	NaN	San Francisco	NaN	31
2	3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.60	NaN	335279.91	335279.91	2011	NaN	San Francisco	NaN	31
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	36
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	44

In [46]:
```python
sal[["title_len", "TotalPayBenefits"]].corr()
```

Out[46]:

	title_len	TotalPayBenefits
title_len	1.000000	-0.036878
TotalPayBenefits	-0.036878	1.000000

> The correlation coefficient value is negative and tends to 0, so it can be concluded that **there is no correlation between Length of JobTitle String and Salary**

Great Job!