

21 June 2024


Import pandas

```
import pandas as pd
```

Read the data from Salaries.csv and store it in a dataframe

```
df=pd.read_csv("/content/Salaries (1).csv")
```


df



	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPayBenefits	Year	Notes
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.00	400184.25	NaN	567595.43	567595.43	2011	NaN
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	NaN	538909.28	538909.28	2011	NaN
2	3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.60	NaN	335279.91	335279.91	2011	NaN
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.00	56120.71	198306.90	NaN	332343.61	332343.61	2011	NaN
4	5	PATRICK	DEPUTY CHIEF OF DEPARTMENT	134401.60	9737.00	182234.59	NaN	326373.19	326373.19	2011	NaN

Check if the dataframe is properly read or not using the head function


```
df.head()
```



	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.00	400184.25	NaN	567595.43
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	NaN	538909.28

What columns exist in this dataframe?

```
df.columns
```



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

How many rows does this dataframe have?


```
len(df.index)
```



```
40409
```

Display the information about the dataframe using the info function. Which of these columns have missing values in them?

```
df.info()
```



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40409 entries, 0 to 40408
```

```
Data columns (total 13 columns):
#   Column              Non-Null Count  Dtype
---  -
0   Id                   40409 non-null  int64
1   EmployeeName         40409 non-null  object
2   JobTitle             40409 non-null  object
3   BasePay              40408 non-null  float64
4   OvertimePay          40408 non-null  float64
5   OtherPay             40408 non-null  float64
6   Benefits             4249 non-null   float64
7   TotalPay             40408 non-null  float64
8   TotalPayBenefits     40408 non-null  float64
9   Year                 40408 non-null  float64
10  Notes                0 non-null     float64
11  Agency              40408 non-null  object
12  Status              0 non-null     float64
dtypes: float64(9), int64(1), object(3)
memory usage: 4.0+ MB
```

**What is the total BasePay?**

```
df.BasePay.sum()
```

```
2872874878.9
```

**What is the highest amount of overtime pay?**

```
df.OvertimePay.max()
```

```
245131.88
```

**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).**

```
df[df['EmployeeName']=='JOSEPH DRISCOLL']['JobTitle']
```

```
24    CAPTAIN, FIRE SUPPRESSION
Name: JobTitle, dtype: object
```

**How much does JOSEPH DRISCOLL make (including benefits)?**

```
df[df['EmployeeName']=='JOSEPH DRISCOLL']['TotalPayBenefits']
```

```
24    270324.91
Name: TotalPayBenefits, dtype: float64
```

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

```
df.loc[df['TotalPayBenefits'].idxmax()]
```

```
1
Id
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
```

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

```
df['BasePay'].mean()
```

```
66325.4488404877
```

**Replace all the missing values in the Benefits column with 0**

```
df.fillna("0")
```



	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Ben
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.0	400184.25	
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	
2	3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.6	
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.0	56120.71	198306.9	
4	5	PATRICK	DEPUTY CHIEF OF DEPARTMENT	134401.6	9737.0	182234.59	

**How many unique job titles exist in the dataframe?**

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



2159

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

```
df.loc[df['TotalPayBenefits'].idxmin()]['EmployeeName']
```



'Joe Lopez'

**What are the top 5 most common jobs?**

```
df['JobTitle'].value_counts().head(5)
```



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

**How many people have the word Chief in their job title?**

Hint: Use lambda expression here

```
def chief_string(title):
    if 'chief' in title.lower():
        return True
    else:
        return False

sum(df['JobTitle'].apply(lambda x: chief_string(x)))
```



627

