```
import pandas as pd
import numpy as np
```

```
df = pd.read_csv('Salaries.csv')
```

```
/usr/local/lib/python3.8/dist-packages/IPython/core/interactiveshell.py:3326: DtypeWarning: Columns (3,4,5,6,12) have mixed types.S
  exec(code_obj, self.user_global_ns, self.user_ns)
```

```
df.head(10)
```

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	Tota
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.0	400184.25	NaN	56759
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	NaN	53890
2	3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.6	NaN	33527
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.0	56120.71	198306.9	NaN	33234
4	5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT, (FIRE DEPARTMENT)	134401.6	9737.0	182234.59	NaN	32637
5	6	DAVID SULLIVAN	ASSISTANT DEPUTY CHIEF II	118602.0	8601.0	189082.74	NaN	31628
6	7	ALSON LEE	BATTALION CHIEF, (FIRE	92492.01	89062.9	134426.14	NaN	31598

```
df.tail(10)
```

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	Tot
148644	148645	Randy D Winn	Stationary Eng, Sewage Plant	0.00	0.00	0.00	0.00	0.00	
148645	148646	Carolyn A Wilson	Human Services Technician	0.00	0.00	0.00	0.00	0.00	
148646	148647	Not provided	Not provided	Not Provided	Not Provided	Not Provided	Not Provided	0.00	
148647	148648	Joann Anderson	Communications Dispatcher 2	0.00	0.00	0.00	0.00	0.00	
148648	148649	Leon Walker	Custodian	0.00	0.00	0.00	0.00	0.00	
148649	148650	Roy I Tillery	Custodian	0.00	0.00	0.00	0.00	0.00	
148650	148651	Not provided	Not provided	Not Provided	Not Provided	Not Provided	Not Provided	0.00	

```
df.shape
```

```
(148654, 13)
```

```
df.describe()
```

	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

```
df.isnull().sum()
```

```
Id                      0
EmployeeName            0
JobTitle                0
BasePay               605
OvertimePay             0
OtherPay                0
Benefits            36159
TotalPay                0
TotalPayBenefits        0
Year                    0
Notes              148654
Agency                  0
Status             110535
dtype: int64
```

```
df = df.drop(['Id', 'Notes', 'Status'], axis = 1)
```

```
df.columns
```

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

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

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

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

```
2159
```

```
len(df[df['JobTitle'].str.contains('CAPTAIN', case = False)])
```

```
552
```

```
df.columns
```

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

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

```
4              PATRICK GARDNER
6                    ALSON LEE
8                MICHAEL MORRIS
9            JOANNE HAYES-WHITE
10                ARTHUR KENNEY
                    ...
32623              JAMES BARDEN
36162       Joanne Hayes-White
72926     Joanne M Hayes-White
102303          Robert E Evans
110535    Joanne M Hayes-White
Name: EmployeeName, Length: 226, dtype: object
```

```
df.columns
```

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

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

```
count     148049.0
unique    109900.0
top            0.0
freq         875.0
Name: BasePay, dtype: float64
```

```
df.shape
```

```
(148654, 10)
```

```
modified_dataframe = df['BasePay'].replace('Not Provided', 0)
```

```
df['find_mean'] = modified_dataframe
```

```
df.head(5)
```

	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPayBenef
0	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.0	400184.25	NaN	567595.43	56759
1	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	NaN	538909.28	53890
2	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.6	NaN	335279.91	33527
3	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.0	56120.71	198306.9	NaN	332343.61	33234

```
df['find_mean'].describe()
```

```
count     148049.0
unique    109899.0
top            0.0
freq         879.0
Name: find_mean, dtype: float64
```

```
#replacing a set of values of a column using a custom value
```

```
df['EmployeeName'].replace('Not provided', np.nan)
```

```
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
```

```
#drop the rows having five missing values
```

```
df[df.isnull().sum(axis = 1) == 5]
```

	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPayBenefits	Year

```
df.columns
```

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

```
df[df['EmployeeName'] == 'Albert Pardini']['JobTitle']
```

```
36519    Captain 3
Name: JobTitle, dtype: object
```

```
df[df['EmployeeName'].str.contains('albert pardini', case = False)]['JobTitle']
```

```
2        CAPTAIN III (POLICE DEPARTMENT)
36519                          Captain 3
Name: JobTitle, dtype: object
```

```
(df['BasePay'] == 'Not Provided').sum()
```

```
4
```

```
df['BasePay'].replace('Not Provided', np.nan)
```

```
0         167411.18
1         155966.02
2         212739.13
3           77916.0
4          134401.6
            ...
148649         0.00
148650          NaN
148651          NaN
148652          NaN
148653         0.00
Name: BasePay, Length: 148654, dtype: object
```

```
df[(df['BasePay'] == 'Not Provided')]
```

	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPayBenefits
148646	Not provided	Not provided	Not Provided	Not Provided	Not Provided	Not Provided	0.0	0.0
148650	Not provided	Not provided	Not Provided	Not Provided	Not Provided	Not Provided	0.0	0.0
148651	Not provided	Not provided	Not Provided	Not Provided	Not Provided	Not Provided	0.0	0.0

```
x = df['BasePay'].str.contains('not provided', case = False)
```

```
x.dtype
```

```
dtype('O')
```

```
x.replace('Not provided', np.nan)
```

```
0          NaN
1          NaN
2          NaN
3          NaN
4          NaN
           ...
148649    False
148650     True
148651     True
148652     True
148653    False
Name: BasePay, Length: 148654, dtype: object
```

✓ 0s completed at 11:09 AM