In [6]:

```python
import pandas as pd
```

In [7]:

```python
# read dataset
df = pd.read_csv('employees.csv')
```

In [3]:

```python
df
```

Out[3]:

	First Name	Gender	Start Date	Last Login Time	Salary	Bonus %	Senior Management	Team
0	Douglas	Male	8/6/1993	12:42 PM	97308	6.945	True	Marketing
1	Thomas	Male	3/31/1996	6:53 AM	61933	4.170	True	NaN
2	Maria	Female	4/23/1993	11:17 AM	130590	11.858	False	Finance
3	Jerry	Male	3/4/2005	1:00 PM	138705	9.340	True	Finance
4	Larry	Male	1/24/1998	4:47 PM	101004	1.389	True	Client Services
...	...	...	...	...	...	...	...	...
995	Henry	NaN	11/23/2014	6:09 AM	132483	16.655	False	Distribution
996	Phillip	Male	1/31/1984	6:30 AM	42392	19.675	False	Finance
997	Russell	Male	5/20/2013	12:39 PM	96914	1.421	False	Product
998	Larry	Male	4/20/2013	4:45 PM	60500	11.985	False	Business Development
999	Albert	Male	5/15/2012	6:24 PM	129949	10.169	True	Sales

1000 rows × 8 columns

In [8]:

```python
df.head()
```

Out[8]:

	First Name	Gender	Start Date	Last Login Time	Salary	Bonus %	Senior Management	Team
0	Douglas	Male	8/6/1993	12:42 PM	97308	6.945	True	Marketing
1	Thomas	Male	3/31/1996	6:53 AM	61933	4.170	True	NaN
2	Maria	Female	4/23/1993	11:17 AM	130590	11.858	False	Finance
3	Jerry	Male	3/4/2005	1:00 PM	138705	9.340	True	Finance
4	Larry	Male	1/24/1998	4:47 PM	101004	1.389	True	Client Services

In [5]:

```
df.describe()
```

Out[5]:

	Salary	Bonus %
count	1000.000000	1000.000000
mean	90662.181000	10.207555
std	32923.693342	5.528481
min	35013.000000	1.015000
25%	62613.000000	5.401750
50%	90428.000000	9.838500
75%	118740.250000	14.838000
max	149908.000000	19.944000

In [9]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 8 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   First Name         933 non-null    object 
 1   Gender             855 non-null    object 
 2   Start Date         1000 non-null   object 
 3   Last Login Time    1000 non-null   object 
 4   Salary             1000 non-null   int64  
 5   Bonus %            1000 non-null   float64
 6   Senior Management  933 non-null    object 
 7   Team               957 non-null    object 
dtypes: float64(1), int64(1), object(6)
memory usage: 62.6+ KB
```

In [7]:

```
df.columns
```

Out[7]:

```
Index(['First Name', 'Gender', 'Start Date', 'Last Login Time', 'Salary',
       'Bonus %', 'Senior Management', 'Team'],
      dtype='object')
```

In [8]:

```
# count rows which include null
df.isna().sum()
```

Out[8]:

```
First Name            67
Gender               145
Start Date             0
Last Login Time        0
Salary                 0
Bonus %                0
Senior Management     67
Team                  43
dtype: int64
```

In [11]:

```
# count rows which not include null
cleaned = df.dropna()
```

In [12]:
```
cleaned
```

Out[12]:

	First Name	Gender	Start Date	Last Login Time	Salary	Bonus %	Senior Management	Team
0	Douglas	Male	8/6/1993	12:42 PM	97308	6.945	True	Marketing
2	Maria	Female	4/23/1993	11:17 AM	130590	11.858	False	Finance
3	Jerry	Male	3/4/2005	1:00 PM	138705	9.340	True	Finance
4	Larry	Male	1/24/1998	4:47 PM	101004	1.389	True	Client Services
5	Dennis	Male	4/18/1987	1:35 AM	115163	10.125	False	Legal
...	...	...	...	...	...	...	...	...
994	George	Male	6/21/2013	5:47 PM	98874	4.479	True	Marketing
996	Phillip	Male	1/31/1984	6:30 AM	42392	19.675	False	Finance
997	Russell	Male	5/20/2013	12:39 PM	96914	1.421	False	Product
998	Larry	Male	4/20/2013	4:45 PM	60500	11.985	False	Business Development
999	Albert	Male	5/15/2012	6:24 PM	129949	10.169	True	Sales

764 rows × 8 columns

In [13]:
```
len(cleaned)
```

Out[13]:
```
764
```

In [14]:
```
pd.isna(df["Gender"])
```

Out[14]:
```
0       False
1       False
2       False
3       False
4       False
        ...  
995      True
996     False
997     False
998     False
999     False
Name: Gender, Length: 1000, dtype: bool
```

In [15]:
```
# filling missing values
mostFrequentGender = df['Gender'].mode()[0]
df["Gender"].fillna(mostFrequentGender, inplace=True)
```

In [16]:

```
df["Gender"]
```

Out[16]:

```
0        Male
1        Male
2      Female
3        Male
4        Male
        ...  
995    Female
996      Male
997      Male
998      Male
999      Male
Name: Gender, Length: 1000, dtype: object
```

In [17]:

```
# sum of salary
df['Salary'].sum()
```

Out[17]:

```
90662181
```

In [18]:

```
# delete "Last Login Time" column
df = df.drop("Last Login Time", axis='columns')
```

In [19]:

```
df
```

Out[19]:

	First Name	Gender	Start Date	Salary	Bonus %	Senior Management	Team
0	Douglas	Male	8/6/1993	97308	6.945	True	Marketing
1	Thomas	Male	3/31/1996	61933	4.170	True	NaN
2	Maria	Female	4/23/1993	130590	11.858	False	Finance
3	Jerry	Male	3/4/2005	138705	9.340	True	Finance
4	Larry	Male	1/24/1998	101004	1.389	True	Client Services
...	...	...	...	...	...	...	...
995	Henry	Female	11/23/2014	132483	16.655	False	Distribution
996	Phillip	Male	1/31/1984	42392	19.675	False	Finance
997	Russell	Male	5/20/2013	96914	1.421	False	Product
998	Larry	Male	4/20/2013	60500	11.985	False	Business Development
999	Albert	Male	5/15/2012	129949	10.169	True	Sales

1000 rows × 7 columns

In [22]:

```
# one hot encoding for "Gender" column
oneHotEncodedData = pd.get_dummies(df, columns = ['Gender'])
oneHotEncodedData
```

Out[22]:

	First Name	Start Date	Salary	Bonus %	Senior Management	Team	Gender_Female	Gender
0	Douglas	8/6/1993	97308	6.945	True	Marketing	False	
1	Thomas	3/31/1996	61933	4.170	True	NaN	False	
2	Maria	4/23/1993	130590	11.858	False	Finance	True	
3	Jerry	3/4/2005	138705	9.340	True	Finance	False	
4	Larry	1/24/1998	101004	1.389	True	Client Services	False	
...	...	...	...	...	...	...	...	
995	Henry	11/23/2014	132483	16.655	False	Distribution	True	
996	Phillip	1/31/1984	42392	19.675	False	Finance	False	
997	Russell	5/20/2013	96914	1.421	False	Product	False	
998	Larry	4/20/2013	60500	11.985	False	Business Development	False	
999	Albert	5/15/2012	129949	10.169	True	Sales	False	

1000 rows × 8 columns

In [ ]: