ML Project on Employee Dataset

```
import pandas as pd
import numpy as np

df=pd.read_csv('Employee_Salary_Dataset.csv')
print(df)
```

```
    ID  Experience_Years  Age  Gender    Salary
0    1                 5   28  Female    250000
1    2                 1   21    Male     50000
2    3                 3   23  Female    170000
3    4                 2   22    Male     25000
4    5                 1   17    Male     10000
5    6                25   62    Male   5001000
6    7                19   54  Female    800000
7    8                 2   21  Female      9000
8    9                10   36  Female     61500
9   10                15   54  Female    650000
10  11                 4   26  Female    250000
11  12                 6   29    Male   1400000
12  13                14   39    Male   6000050
13  14                11   40    Male    220100
14  15                 2   23    Male      7500
15  16                 4   27  Female     87000
16  17                10   34  Female    930000
17  18                15   54  Female   7900000
18  19                 2   21    Male     15000
19  20                10   36    Male    330000
20  21                15   54    Male   6570000
21  22                 4   26    Male     25000
22  23                 5   29    Male   6845000
23  24                 1   21  Female      6000
24  25                 4   23  Female      8900
25  26                 3   22  Female     20000
26  27                 1   18    Male      3000
27  28                27   62  Female  10000000
28  29                19   54  Female   5000000
29  30                 2   21  Female      6100
30  31                10   34    Male     80000
31  32                15   54    Male    900000
32  33                20   55  Female   1540000
33  34                19   53  Female   9300000
34  35                16   49    Male   7600000
```

```
df.head()
```

```
   ID  Experience_Years  Age  Gender  Salary
0   1                 5   28  Female  250000
1   2                 1   21    Male   50000
```

```
2   3                 3   23  Female  170000
3   4                 2   22    Male   25000
4   5                 1   17    Male   10000
```

```
print(df.columns)
print(df.shape)
```

```
Index(['ID', 'Experience_Years', 'Age', 'Gender', 'Salary'],
dtype='object')
(35, 5)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 35 entries, 0 to 34
Data columns (total 5 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   ID                35 non-null     int64
 1   Experience_Years  35 non-null     int64
 2   Age               35 non-null     int64
 3   Gender            35 non-null     object
 4   Salary            35 non-null     int64
dtypes: int64(4), object(1)
memory usage: 1.5+ KB
```

```
#percentage of null valules
(df.isnull().sum()/len(df))*100
```

```
ID                  0.0
Experience_Years    0.0
Age                 0.0
Gender              0.0
Salary              0.0
dtype: float64
```

```
#removes duplicates
df=df[~df.duplicated()]
df.shape
```

```
(35, 5)
```

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

```
ID                     int64
Experience_Years       int64
Age                    int64
Gender                object
Salary                 int64
dtype: object
```

```
#mode
df['Age'].mode()

0    54
Name: Age, dtype: int64

#mean
df['Salary'].fillna(df['Salary'].mean())
df['SalaryExperience']=df['Experience_Years'].fillna(df['Experience_Ye
ars'].mean())
print(df.isnull().sum())

ID                    0
Experience_Years      0
Age                   0
Gender                0
Salary                0
SalaryExperience      0
dtype: int64

df.shape

(35, 6)

df.head()
# and condition
df1=df[(df['Salary']>20000) & (df['Salary']>80000)]
df1.head()

   ID  Experience_Years  Age  Gender   Salary  SalaryExperience
0   1                 5   28  Female   250000                 5
2   3                 3   23  Female   170000                 3
5   6                25   62    Male  5001000                25
6   7                19   54  Female   800000                19
9  10                15   54  Female   650000                15

df.shape

(35, 6)

df2=df[np.logical_and(df['Salary']>20000,df['Salary']<90000)]
df2.shape

(6, 6)

df.shape

(35, 6)

df.query('Salary>20000 and Salary <80000')
df
```

```
    ID  Experience_Years  Age  Gender    Salary  SalaryExperience
0    1                 5   28  Female    250000                 5
1    2                 1   21    Male     50000                 1
2    3                 3   23  Female    170000                 3
3    4                 2   22    Male     25000                 2
4    5                 1   17    Male     10000                 1
5    6                25   62    Male   5001000                25
6    7                19   54  Female    800000                19
7    8                 2   21  Female      9000                 2
8    9                10   36  Female     61500                10
9   10                15   54  Female    650000                15
10  11                 4   26  Female    250000                 4
11  12                 6   29    Male   1400000                 6
12  13                14   39    Male   6000050                14
13  14                11   40    Male    220100                11
14  15                 2   23    Male      7500                 2
15  16                 4   27  Female     87000                 4
16  17                10   34  Female    930000                10
17  18                15   54  Female   7900000                15
18  19                 2   21    Male     15000                 2
19  20                10   36    Male    330000                10
20  21                15   54    Male   6570000                15
21  22                 4   26    Male     25000                 4
22  23                 5   29    Male   6845000                 5
23  24                 1   21  Female      6000                 1
24  25                 4   23  Female      8900                 4
25  26                 3   22  Female     20000                 3
26  27                 1   18    Male      3000                 1
27  28                27   62  Female  10000000                27
28  29                19   54  Female   5000000                19
29  30                 2   21  Female      6100                 2
30  31                10   34    Male     80000                10
31  32                15   54    Male    900000                15
32  33                20   55  Female   1540000                20
33  34                19   53  Female   9300000                19
34  35                16   49    Male   7600000                16
```

```
df=pd.read_csv('Employee_Salary_Dataset.csv')
def function_name(x):
    if x=='Male':
        x='m'
    elif x=='Female':
        x='f'
    else:
        x='other'
    return x
df["Gender1"]=df['Gender'].apply(function_name)
pd.set_option('display.max_rows',None)
pd.set_option('display.max_columns',None)
print(df.head())
```

```
   ID  Experience_Years  Age  Gender  Salary Gender1
0   1                 5   28  Female  250000       f
1   2                 1   21    Male   50000       m
2   3                 3   23  Female  170000       f
3   4                 2   22    Male   25000       m
4   5                 1   17    Male   10000       m
```

```
#bouns
df['bouns']=df['Salary']*10/100
df.head()
```

```
   ID  Experience_Years  Age  Gender  Salary Gender1    bouns
0   1                 5   28  Female  250000       f  25000.0
1   2                 1   21    Male   50000       m   5000.0
2   3                 3   23  Female  170000       f  17000.0
3   4                 2   22    Male   25000       m   2500.0
4   5                 1   17    Male   10000       m   1000.0
```

```
df['next_year_salary']=df['bouns']+df['Salary']
df.head()
```

```
   ID  Experience_Years  Age  Gender  Salary Gender1    bouns  \
0   1                 5   28  Female  250000       f  25000.0
1   2                 1   21    Male   50000       m   5000.0
2   3                 3   23  Female  170000       f  17000.0
3   4                 2   22    Male   25000       m   2500.0
4   5                 1   17    Male   10000       m   1000.0

   next_year_salary
0          275000.0
1           55000.0
2          187000.0
3           27500.0
4           11000.0
```

```
df.dtypes
```

```
ID                     int64
Experience_Years       int64
Age                    int64
Gender                object
Salary                 int64
Gender1               object
bouns                float64
next_year_salary     float64
dtype: object
```

```
df.head()
```

```
   ID  Experience_Years  Age  Gender  Salary Gender1    bouns  \
0   1                 5   28  Female  250000       f  25000.0
```

```
1   2                 1   21    Male   50000       m    5000.0
2   3                 3   23  Female  170000       f   17000.0
3   4                 2   22    Male   25000       m    2500.0
4   5                 1   17    Male   10000       m    1000.0

   next_year_salary
0          275000.0
1           55000.0
2          187000.0
3           27500.0
4           11000.0
```

```
#revers columns names
print(df.columns[::-1])
```

```
Index(['next_year_salary', 'bouns', 'Gender1', 'Salary', 'Gender',
'Age',
       'Experience_Years', 'ID'],
      dtype='object')
```

```
df.columns
```

```
Index(['ID', 'Experience_Years', 'Age', 'Gender', 'Salary', 'Gender1',
'bouns',
       'next_year_salary'],
      dtype='object')
```

```
#row reverse
df[::-1]
df.head()
```

```
   ID  Experience_Years  Age  Gender  Salary Gender1     bouns  \
0   1                 5   28  Female  250000       f   25000.0
1   2                 1   21    Male   50000       m    5000.0
2   3                 3   23  Female  170000       f   17000.0
3   4                 2   22    Male   25000       m    2500.0
4   5                 1   17    Male   10000       m    1000.0

   next_year_salary
0          275000.0
1           55000.0
2          187000.0
3           27500.0
4           11000.0
```

```
df=pd.pivot_table(df,index='ID',columns='Gender',aggfunc='count')
df.head()
```

```
          Age      Experience_Years      Gender1      Salary
bouns       \
Gender Female Male           Female Male  Female Male Female Male
```

```
Female Male
ID

1          1.0  NaN              1.0  NaN     1.0  NaN     1.0  NaN
1.0  NaN
2          NaN  1.0              NaN  1.0     NaN  1.0     NaN  1.0
NaN  1.0
3          1.0  NaN              1.0  NaN     1.0  NaN     1.0  NaN
1.0  NaN
4          NaN  1.0              NaN  1.0     NaN  1.0     NaN  1.0
NaN  1.0
5          NaN  1.0              NaN  1.0     NaN  1.0     NaN  1.0
NaN  1.0
```

```
       next_year_salary
Gender           Female Male
ID
1                   1.0  NaN
2                   NaN  1.0
3                   1.0  NaN
4                   NaN  1.0
5                   NaN  1.0
```

```
df=pd.read_csv('Employee_Salary_Dataset.csv')
df.columns
```

```
Index(['ID', 'Experience_Years', 'Age', 'Gender', 'Salary'],
dtype='object')
```

```
df1=df.groupby('Experience_Years')['Salary'].sum()
print(df1)
```

```
Experience_Years
1        69000
2        62600
3       190000
4       370900
5      7095000
6      1400000
10     1401500
11      220100
14     6000050
15    16020000
16     7600000
19    15100000
20     1540000
25     5001000
27    10000000
Name: Salary, dtype: int64
```

```
df1=df.groupby('Experience_Years')
['Salary'].agg(['mean','sum','count'])
df1
```

```
                          mean       sum  count
Experience_Years
1                 1.725000e+04     69000      4
2                 1.252000e+04     62600      5
3                 9.500000e+04    190000      2
4                 9.272500e+04    370900      4
5                 3.547500e+06   7095000      2
6                 1.400000e+06   1400000      1
10                3.503750e+05   1401500      4
11                2.201000e+05    220100      1
14                6.000050e+06   6000050      1
15                4.005000e+06  16020000      4
16                7.600000e+06   7600000      1
19                5.033333e+06  15100000      3
20                1.540000e+06   1540000      1
25                5.001000e+06   5001000      1
27                1.000000e+07  10000000      1
```