Problem 2(Explore Python)

- Import pandas and read data

In [1]:

```
import pandas as pd
import numpy as np
Column_name = ['Annual income',
               'Sex',
               'Marital Status',
               'Age',
               'Education',
               'Occupation',
               'How long have you lived in the SF/O/SJ area',
               'Dual Incomes',
               'Persons in your household',
               'Persons in household under 18',
               'Householder Status',
               'Type of Home',
               'Ethnic Classification',
               'language spoken most often']
data = pd.DataFrame(pd.read_table('income1.data',header=None, delim_whitespa
data.head()
```

Out[1]:

	Annual income	Sex	Marital Status	Age	Education	Occupation	How long have you lived in the SF/O/SJ area	Dual Incomes	Persons in your household	Persons in household under 18	H
0	9	2	1.0	5	4.0	5.0	5.0	3	3.0	0	
1	9	1	1.0	5	5.0	5.0	5.0	3	5.0	2	
2	9	2	1.0	3	5.0	1.0	5.0	2	3.0	1	
3	1	2	5.0	1	2.0	6.0	5.0	1	4.0	2	
4	1	2	5.0	1	2.0	6.0	3.0	1	4.0	2	

- a)Indicate the number of lines in the file.

In [2]:

```
len(data)
```

Out[2]:

```
8993
```

- b) Indicate the number of lines in the file after eliminating those lines that have fields characterized by unavailable (NA) data.

```
In [3]: data = data.dropna(axis=0,how='any')
        len(data)
```

```
Out[3]: 6876
```

- c) Indicate the most common education level (the fifth column corresponds to education level).

```
In [4]: most_common_edu_lev = data['Education'].replace([1,2,3,4,5,6], ['Grade 8 or
                                                                 'Grades 9 to
                                                                 'Graduated h
                                                                 '1 to 3 year
                                                                 'College gra
                                                                 'Grad Study'
        most_common_edu_lev = most_common_edu_lev.value_counts()
        print(most_common_edu_lev)
```

```
1 to 3 years of college    2407
Graduated high school      1479
College graduate           1207
Grad Study                  820
Grades 9 to 11              787
Grade 8 or less             176
Name: Education, dtype: int64
```

Answer: The most common education level is 4 which represent 1 to 3 years of college level.

- d) Indicate the level of income for households with some graduate school.
  - Graduate study - Education = 6
  - Annual income of household = 'Annual income'

```
In [5]: Income_education = data.loc[data["Education"] == 6.0,
                                      ['Annual income',
                                       'Education']].sort_values(['Annual income'],
                                                                  ascending=False)
        Income_education['Annual income'] = Income_education['Annual income'].replac

        Income_education
        Level_income_graduate = Income_education['Annual income'].value_counts()
        print(Level_income_graduate)
```

```
$75,000 or more       245
$50,000 to $74,999    189
$40,000 to $49,999    121
$30,000 to $39,999     95
$25,000 to $29,999     40
$20,000 to $24,999     40
$15,000 to $19,999     33
Less than $10,000      29
$10,000 to $14,999     28
Name: Annual income, dtype: int64
```

#Problem 2 End