Assignment - 1

Import the necessary libraries

In [5]:
```
import pandas as pd
```

Import the dataset from this(https://raw.githubusercontent.com/justmarkham/DAT8/master/

Use sep= "|" while reading the data

In [6]:
```
url = 'https://raw.githubusercontent.com/justmarkham/DAT8/master/data/u.user'
```

Assign it to a variable called users and use the 'user_id' as index

In [7]:
```
user=pd.read_csv(url,sep="|",index_col="user_id")
```

See the first 10 and last 10 entries

In [8]:
```
user.head(10)
```

Out[8]:

user_id	age	gender	occupation	zip_code
1	24	M	technician	85711
2	53	F	other	94043
3	23	M	writer	32067
4	24	M	technician	43537
5	33	F	other	15213
6	42	M	executive	98101
7	57	M	administrator	91344
8	36	M	administrator	05201
9	29	M	student	01002
10	53	M	lawyer	90703

In [9]:
```
user.tail(10)
```

Out[9]:

user_id	age	gender	occupation	zip_code
934	61	M	engineer	22902
935	42	M	doctor	66221

user_id	age	gender	occupation	zip_code
936	24	M	other	32789
937	48	M	educator	98072
938	38	F	technician	55038
939	26	F	student	33319
940	32	M	administrator	02215
941	20	M	student	97229
942	48	F	librarian	78209
943	22	M	student	77841

What is the number of observations in the dataset?

In [11]:
```
user1=pd.read_csv(url,sep="|")
user1["user_id"].astype("object").nunique()
```

Out[11]:
```
943
```

What is the number of columns in the dataset?

In [12]:
```
user.columns.nunique()
```

Out[12]:
```
4
```

Print the name of all the columns.

In [13]:
```
user.columns
```

Out[13]:
```
Index(['age', 'gender', 'occupation', 'zip_code'], dtype='object')
```

How is the dataset indexed?

In [14]:
```
user.index
```

Out[14]:
```
Int64Index([  1,   2,   3,   4,   5,   6,   7,   8,   9,  10,
            ...
            934, 935, 936, 937, 938, 939, 940, 941, 942, 943],
           dtype='int64', name='user_id', length=943)
```

What is the data type of each column?

In [27]:
```
df = pd.read_csv(url)
df.dtypes
```

Out[27]:
```
user_id|age|gender|occupation|zip_code    object
dtype: object
```

Print only the occupation column

In [15]:
```
user["occupation"]
```

Out[15]:
```
user_id
1         technician
2              other
3             writer
4         technician
5              other
           ...
939          student
940    administrator
941          student
942        librarian
943          student
Name: occupation, Length: 943, dtype: object
```

How many different occupations are in this dataset?

In [16]:
```
user["occupation"].nunique()
```

Out[16]:
```
21
```

What is the most frequent occupation?

In [17]:
```
user["occupation"].mode()
```

Out[17]:
```
0    student
dtype: object
```

DataFrame Info.

In [18]:
```
user.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 943 entries, 1 to 943
Data columns (total 4 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   age         943 non-null    int64 
 1   gender      943 non-null    object
 2   occupation  943 non-null    object
 3   zip_code    943 non-null    object
dtypes: int64(1), object(3)
memory usage: 36.8+ KB
```

Describe all the columns

In [19]:
```
user.describe(include="all")
```

Out[19]:

	age	gender	occupation	zip_code
count	943.000000	943	943	943
unique	NaN	2	21	795
top	NaN	M	student	55414
freq	NaN	670	196	9

	age	gender	occupation	zip_code
mean	34.051962	NaN	NaN	NaN
std	12.192740	NaN	NaN	NaN
min	7.000000	NaN	NaN	NaN
25%	25.000000	NaN	NaN	NaN
50%	31.000000	NaN	NaN	NaN
75%	43.000000	NaN	NaN	NaN
max	73.000000	NaN	NaN	NaN

Summarize only the occupation column

In [20]:
```
user["occupation"].value_counts()
```

Out[20]:
```
student          196
other            105
educator          95
administrator     79
engineer          67
programmer        66
librarian         51
writer            45
executive         32
scientist         31
artist            28
technician        27
marketing         26
entertainment     18
healthcare        16
retired           14
lawyer            12
salesman          12
none               9
homemaker          7
doctor             7
Name: occupation, dtype: int64
```

What is the mean age of users?

In [21]:
```
user["age"].mean()
```

Out[21]:
```
34.05196182396607
```

What is the age with least occurrence?

In [22]:
```
user["age"].min()
```

Out[22]:
```
7
```