In [21]:

```
#Import the necessary libraries
import pandas as pd
```

In [22]:

```
#Import the dataset from this(https://raw.githubusercontent.com/justmarkham/DAT8/master/data/u.user).
#Use sep= "|" while reading the data

url = 'https://raw.githubusercontent.com/justmarkham/DAT8/master/data/u.user'
pd.read_csv(url,sep='|')
```

Out[22]:

	user_id	age	gender	occupation	zip_code
0	1	24	M	technician	85711
1	2	53	F	other	94043
2	3	23	M	writer	32067
3	4	24	M	technician	43537
4	5	33	F	other	15213
...	...	...	...	...	...
938	939	26	F	student	33319
939	940	32	M	administrator	02215
940	941	20	M	student	97229
941	942	48	F	librarian	78209
942	943	22	M	student	77841

943 rows × 5 columns

In [36]:

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

users=pd.read_csv(url,sep='|')
users=users.set_index('user_id')
print(users)
```

```
         age gender     occupation zip_code
user_id
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
...      ...    ...            ...      ...
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

[943 rows x 4 columns]
```

In [37]:

```
#See the first 10 and last 10 entries

print("------------First 10 entries--------------")
print(users.head(10))
print("------------Last 10 entries--------------")
print(users.tail(10))
```

```
------------First 10 entries--------------
         age gender     occupation zip_code
user_id
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
------------Last 10 entries--------------
         age gender     occupation zip_code
user_id
934       61      M       engineer    22902
935       42      M         doctor    66221
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
```

In [39]:

```
#What is the number of observations in the dataset?

row=users.shape[0]
print("Number of observations is ",row)
```

```
Number of observations is  943
```

In [41]:

```
#What is the number of columns in the dataset?

col=users.shape[1]
print("Number of columns is ",col)
```

```
Number of columns is  4
```

In [42]:

```
#Print the name of all the columns.

print(users.columns)
```

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

In [43]:

```
#How is the dataset indexed?

print(users.index)
```

```
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)
```

In [44]:

```
#What is the data type of each column?

DataType = users.dtypes
print('Data type of each column:')
print(DataType)
```

```
Data type of each column:
age            int64
gender        object
occupation    object
zip_code      object
dtype: object
```

In [47]:

```
#Print only the occupation column

users['occupation']
```

Out[47]:

```
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
```

In [49]:

```
#How many different occupations are in this dataset?

users['occupation'].nunique()
```

Out[49]:

```
21
```

In [52]:

```
#What is the most frequent occupation?

users['occupation'].value_counts().idxmax()
```

Out[52]:

```
'student'
```

In [53]:

```
#DataFrame Info.

users.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
```

In [54]:

```
#Describe all the columns

users.describe()
```

Out[54]:

	age
count	943.000000
mean	34.051962
std	12.192740
min	7.000000
25%	25.000000
50%	31.000000
75%	43.000000
max	73.000000

In [62]:

```
#Summarize only the occupation column

users['occupation'].value_counts()
```

Out[62]:

```
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
```

In [57]:

```
#What is the mean age of users?

users['age'].mean()
```

Out[57]:

```
34.05196182396607
```

In [58]:

```
#What is the age with least occurrence?

users['age'].value_counts().idxmin()
```

Out[58]:

```
7
```

In [ ]:

In [ ]: