Assignment - 1

Import the necessary libraries

In [44]:
```
import pandas as pd
import numpy as np

import matplotlib
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline

import warnings
warnings.filterwarnings("ignore")
from IPython.display import Image
```

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

Use sep= "|" while reading the data

In [45]:
```
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 [65]:
```
users = pd.read_csv(url,sep="|",index_col="user_id")
```

In [66]:
```
users
```

Out[66]:

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
...	...	...	...	...
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 × 4 columns

See the first 10 and last 10 entries

In [5]:
```
first = df.head(10)
print(first)
```

```
   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
5                 6|42|M|executive|98101
6              7|57|M|administrator|91344
7              8|36|M|administrator|05201
8                   9|29|M|student|01002
9                  10|53|M|lawyer|90703
```

In [6]:
```
last = df.tail(10)
print(last)
```

```
     user_id|age|gender|occupation|zip_code
933                934|61|M|engineer|22902
934                 935|42|M|doctor|66221
935                  936|24|M|other|32789
936               937|48|M|educator|98072
937              938|38|F|technician|55038
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
```

What is the number of observations in the dataset?

In [9]:
```
total_rows = len(df.axes[0])
print(total_rows)
```

```
943
```

What is the number of columns in the dataset?

In [14]:
```
total_column = len(df.axes[1])
print(total_column)
```

```
1
```

Print the name of all the columns.

In [16]:
```
print(df.columns)
```

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

How is the dataset indexed?

In [62]:
```
users.index
```

Out[62]:
```
RangeIndex(start=0, stop=943, step=1)
```

What is the data type of each column?

In [67]:
```
users.dtypes
```

Out[67]:
```
age             int64
gender         object
occupation     object
zip_code       object
dtype: object
```

Print only the occupation column

In [68]:
```
users['occupation']
```

Out[68]:
```
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 [70]:
```
users["occupation"].value_counts().count()
```

Out[70]:
```
21
```

What is the most frequent occupation?

In [75]:
```
users["occupation"].value_counts().sort_values(ascending=False).head()
```

Out[75]:
```
student          196
other            105
educator          95
administrator     79
engineer          67
Name: occupation, dtype: int64
```

DataFrame Info.

In [81]:
```
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
```

Describe all the columns

In [78]:
```
users.describe(include="all")
```

Out[78]:

	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
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 [76]:
```
users.occupation.describe()
```

Out[76]:
```
count          943
unique          21
top        student
freq           196
Name: occupation, dtype: object
```

What is the mean age of users?

In [79]:
```
users["age"].mean()
```

Out[79]:
```
34.05196182396607
```

What is the age with least occurrence?

In [80]:
```
users["age"].value_counts().tail()
```

Out[80]:
```
11    1
10    1
73    1
66    1
7     1
Name: age, dtype: int64
```

In [ ]:
```

```