Ex3 - Getting and Knowing your Data

This time we are going to pull data directly from the internet. Special thanks to: https://github.com/justmarkham for sharing the dataset and materials.

Step 1. Import the necessary libraries

import pandas as pd

Step 2. Import the dataset from this address.

Step 3. Assign it to a variable called users and use the 'user_id' as index

user_id = pd.read_csv('https://raw.githubusercontent.com/thieu1995/csv-files/main/data/pandas/u.user_i, sep='|', index_col='user_id')

→		age	gender	occupation	zip_code
	user_id				
	1	24	М	technician	85711
	2	53	F	other	94043
	3	23	M	writer	32067
	4	24	М	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	М	student	77841
	943 rows ×	4 col	umns		

∨ Step 4. See the first 25 entries

user_id.head(25)

	age	gender	occupation	zip_code
user_id				
1	24	М	technician	85711
2	53	F	other	94043
3	23	M	writer	32067
4	24	М	technician	43537
5	33	F	other	15213
6	42	М	executive	98101
7	57	М	administrator	91344
8	36	М	administrator	05201
9	29	М	student	01002
10	53	М	lawyer	90703
11	39	F	other	30329
12	28	F	other	06405
13	47	М	educator	29206
14	45	М	scientist	55106
15	49	F	educator	97301
16	21	M	entertainment	10309
17	30	М	programmer	06355
18	35	F	other	37212
19	40	М	librarian	02138
20	42	F	homemaker	95660
21	26	М	writer	30068
22	25	М	writer	40206
23	30	F	artist	48197
24	21	F	artist	94533
25	აი	N.4	ondinoor	EE107
	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24	user_id 1 24 2 53 3 23 4 24 5 33 6 42 7 57 8 36 9 29 10 53 11 39 12 28 13 47 14 45 15 49 16 21 17 30 18 35 19 40 20 42 21 26 22 25 23 30 24 21	user_id 1	user_id 1 24 M technician 2 53 F other 3 23 M writer 4 24 M technician 5 33 F other 6 42 M executive 7 57 M administrator 8 36 M administrator 9 29 M student 10 53 M lawyer 11 39 F other 12 28 F other 13 47 M educator 14 45 M educator 16 21 M entertainment 17 30 M programmer 18 35 F other 19 40 M librarian 20 42 F homemaker

✓ Step 5. See the last 10 entries

user_id.tail(10) ₹ age gender occupation zip_code user_id 934 22902 61 engineer 935 42 Μ doctor 66221 32789 936 24 M other 937 48 Μ educator 98072 938 55038 38 technician 939 26 student 33319 32 02215 940 M administrator 941 20 Μ student 97229 942 librarian 78209

Step 6. What is the number of observations in the dataset?

user_id.shape[0]

→ 943

Step 7. What is the number of columns in the dataset?

user_id.shape[1]

→ 4

Step 8. Print the name of all the columns.

user_id.columns

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

▼ Step 9. How is the dataset indexed?

Step 10. What is the data type of each column?

user_id.dtypes



Step 11. Print only the occupation column

print(user_id['occupation'])

```
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
```

▼ Step 12. How many different occupations are in this dataset?

user_id['occupation'].nunique()

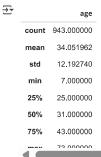
Step 13. What is the most frequent occupation?

user_id['occupation'].value_counts().idxmax()



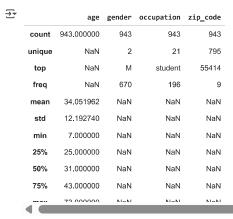
Step 14. Summarize the DataFrame.

user_id.describe()



 ✓ Step 15. Summarize all the columns

user_id.describe(include='all')



Step 16. Summarize only the occupation column

user_id['occupation'].describe()



print(user_id['age'].mean())

→ 34.05196182396607

print(user_id['age'].value_counts().idxmin())

→ 7