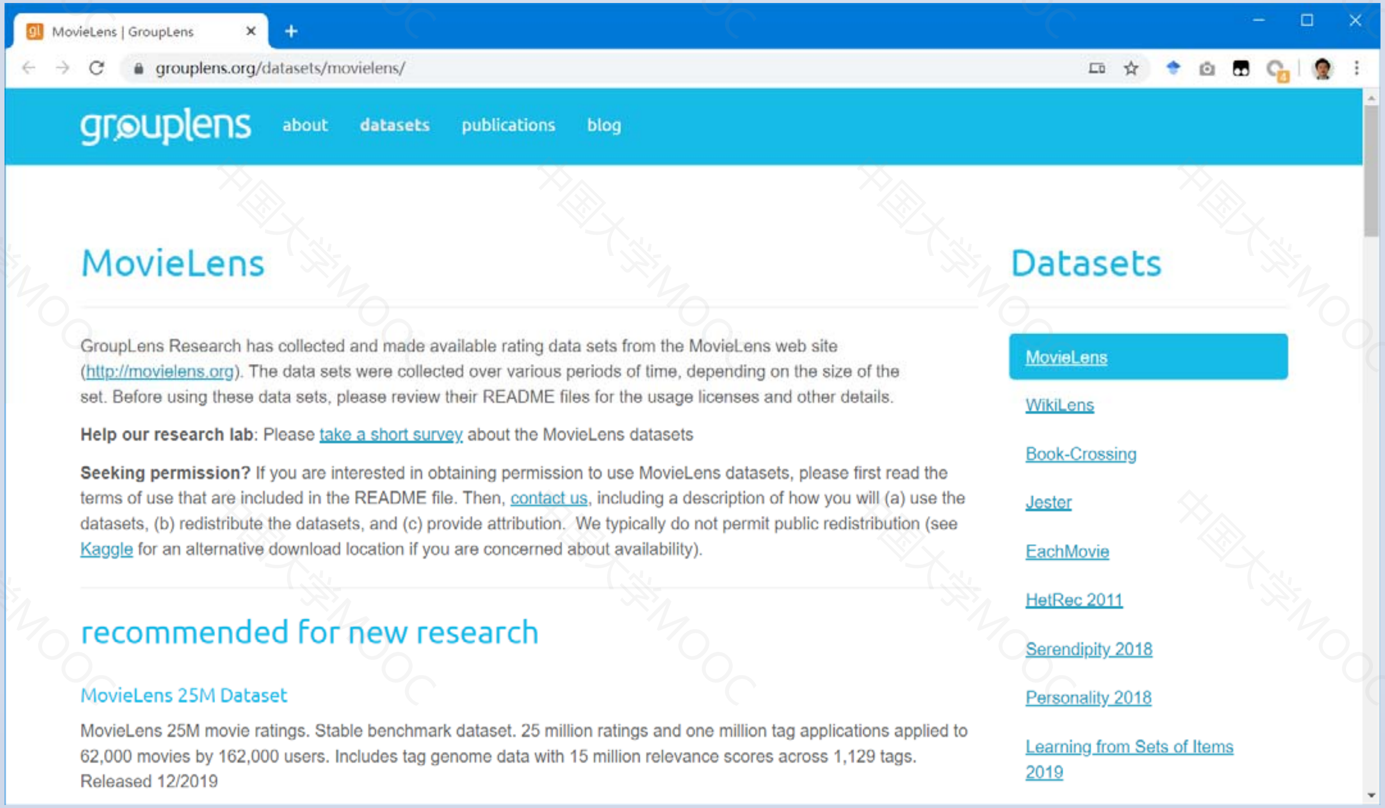
Python大数据分析

案例1：电影评分数据集的分析1

https://grouplens.org/datasets/movielens/



MovieLens

GroupLens Research has collected and made available rating data sets from the MovieLens web site (http://movielens.org). The data sets were collected over various periods of time, depending on the size of the set. Before using these data sets, please review their README files for the usage licenses and other details.

Help our research lab: Please take a short survey about the MovieLens datasets

Seeking permission? If you are interested in obtaining permission to use MovieLens datasets, please first read the terms of use that are included in the README file. Then, contact us, including a description of how you will (a) use the datasets, (b) redistribute the datasets, and (c) provide attribution. We typically do not permit public redistribution (see Kaggle for an alternative download location if you are concerned about availability).

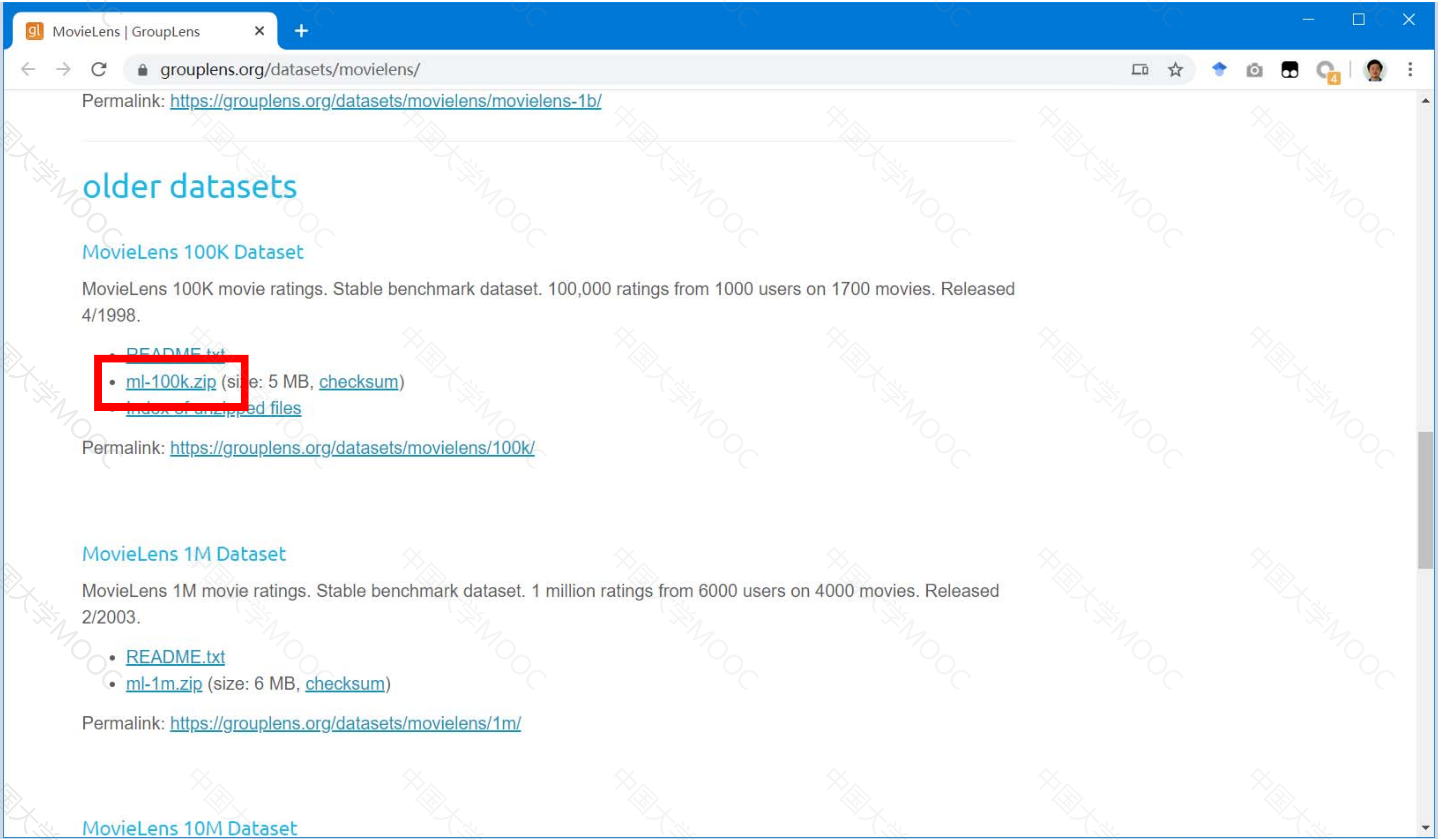
recommended for new research

MovieLens 25M Dataset

MovieLens 25M movie ratings. Stable benchmark dataset. 25 million ratings and one million tag applications applied to 62,000 movies by 162,000 users. Includes tag genome data with 15 million relevance scores across 1,129 tags. Released 12/2019

Datasets

- MovieLens
- WikiLens
- Book-Crossing
- Jester
- EachMovie
- HetRec 2011
- Serendipity 2018
- Personality 2018
- Learning from Sets of Items 2019



Permalink: https://grouplens.org/datasets/movielens/movielens-1b/

older datasets

MovieLens 100K Dataset

MovieLens 100K movie ratings. Stable benchmark dataset. 100,000 ratings from 1000 users on 1700 movies. Released 4/1998.

- README.txt
- ml-100k.zip (size: 5 MB, checksum)
- Index of unzipped files

Permalink: https://grouplens.org/datasets/movielens/100k/

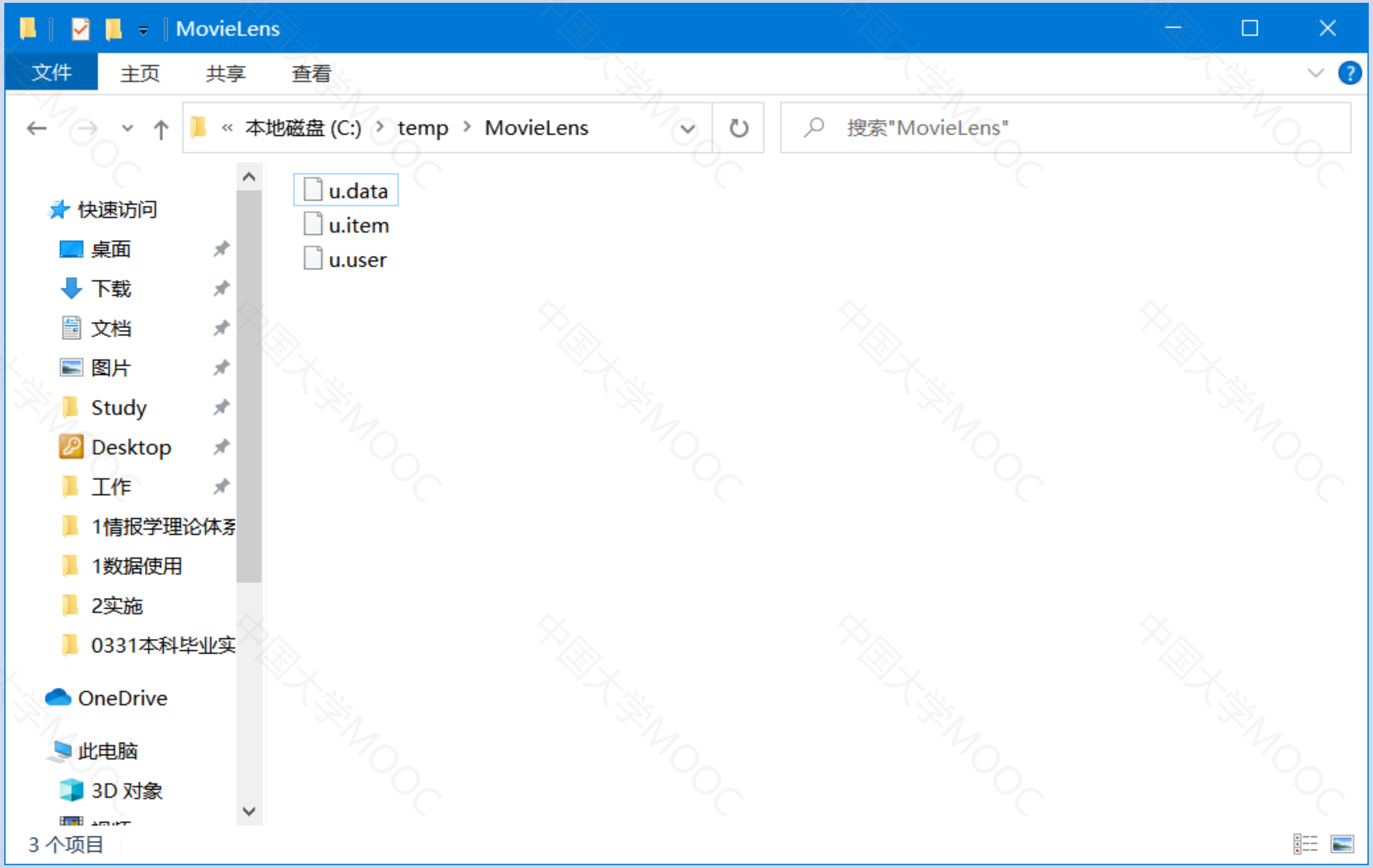
MovieLens 1M Dataset

MovieLens 1M movie ratings. Stable benchmark dataset. 1 million ratings from 6000 users on 4000 movies. Released 2/2003.

- README.txt
- ml-1m.zip (size: 6 MB, checksum)

Permalink: https://grouplens.org/datasets/movielens/1m/

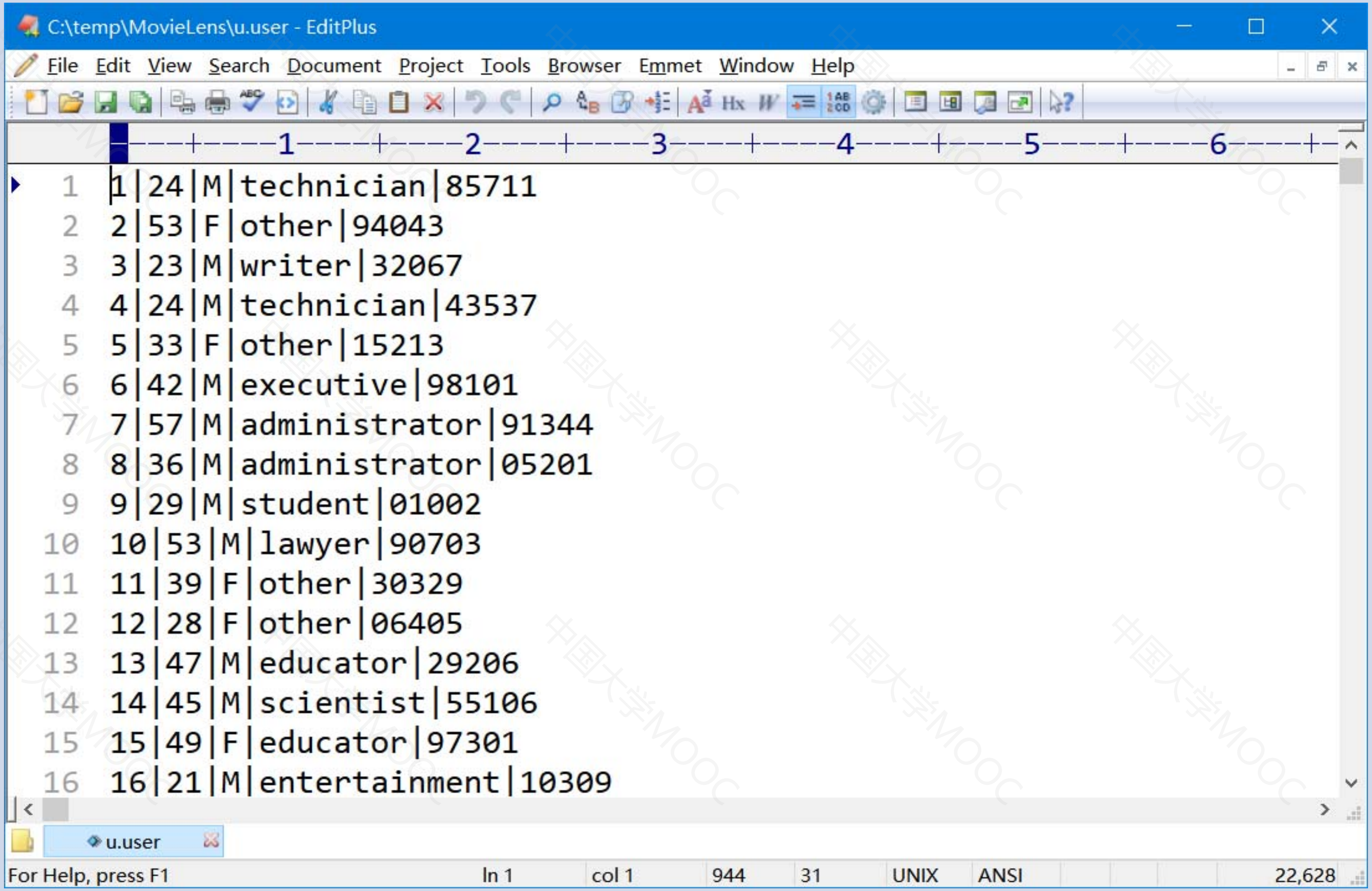
MovieLens 10M Dataset



u.data

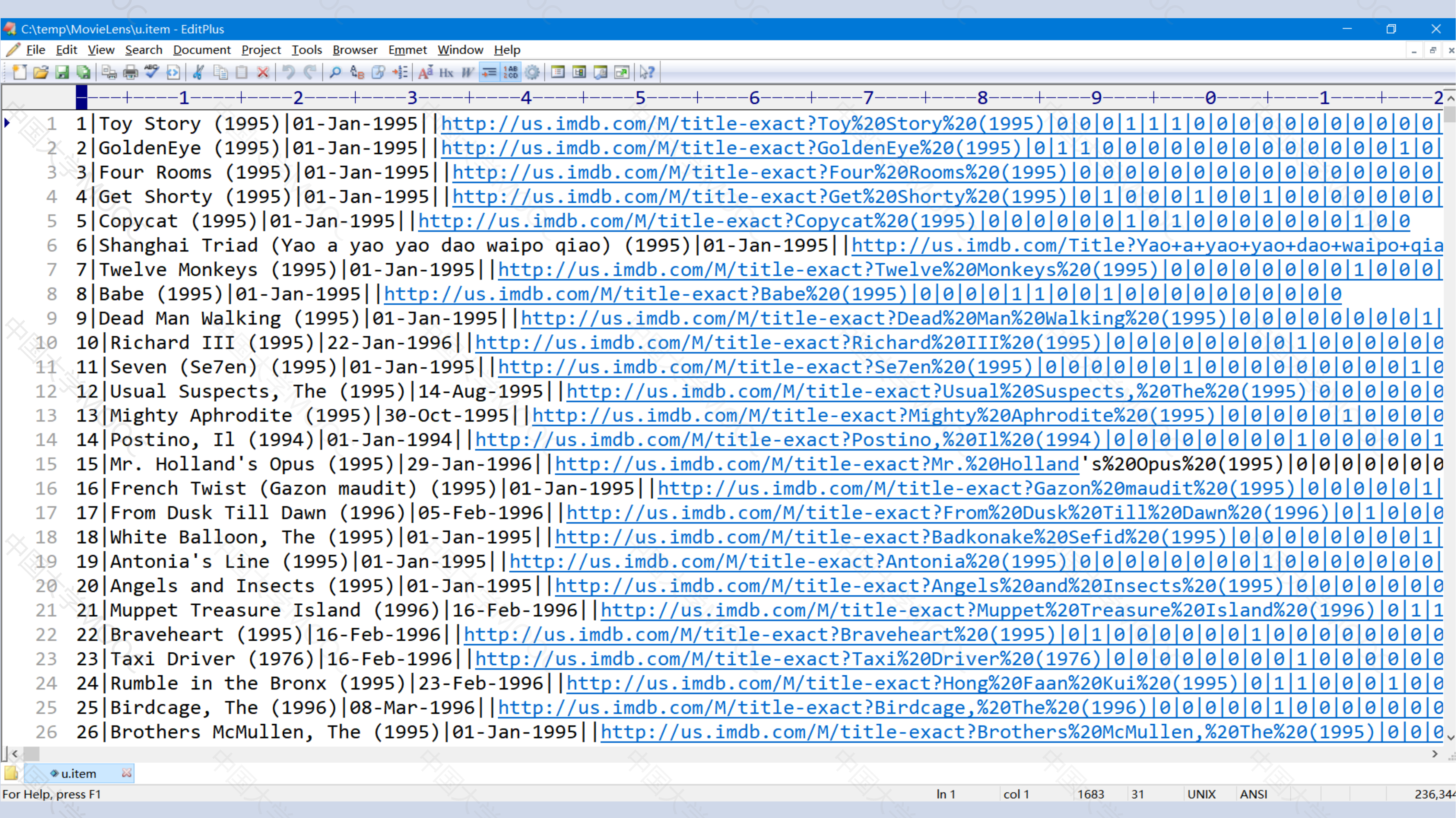
u.item

u.user



C:\temp\MovieLens\u.user - EditPlus
File Edit View Search Document Project Tools Browser Emmet Window Help
u.user
For Help, press F1
ln 1
col 1
944
31
UNIX
ANSI
22,628

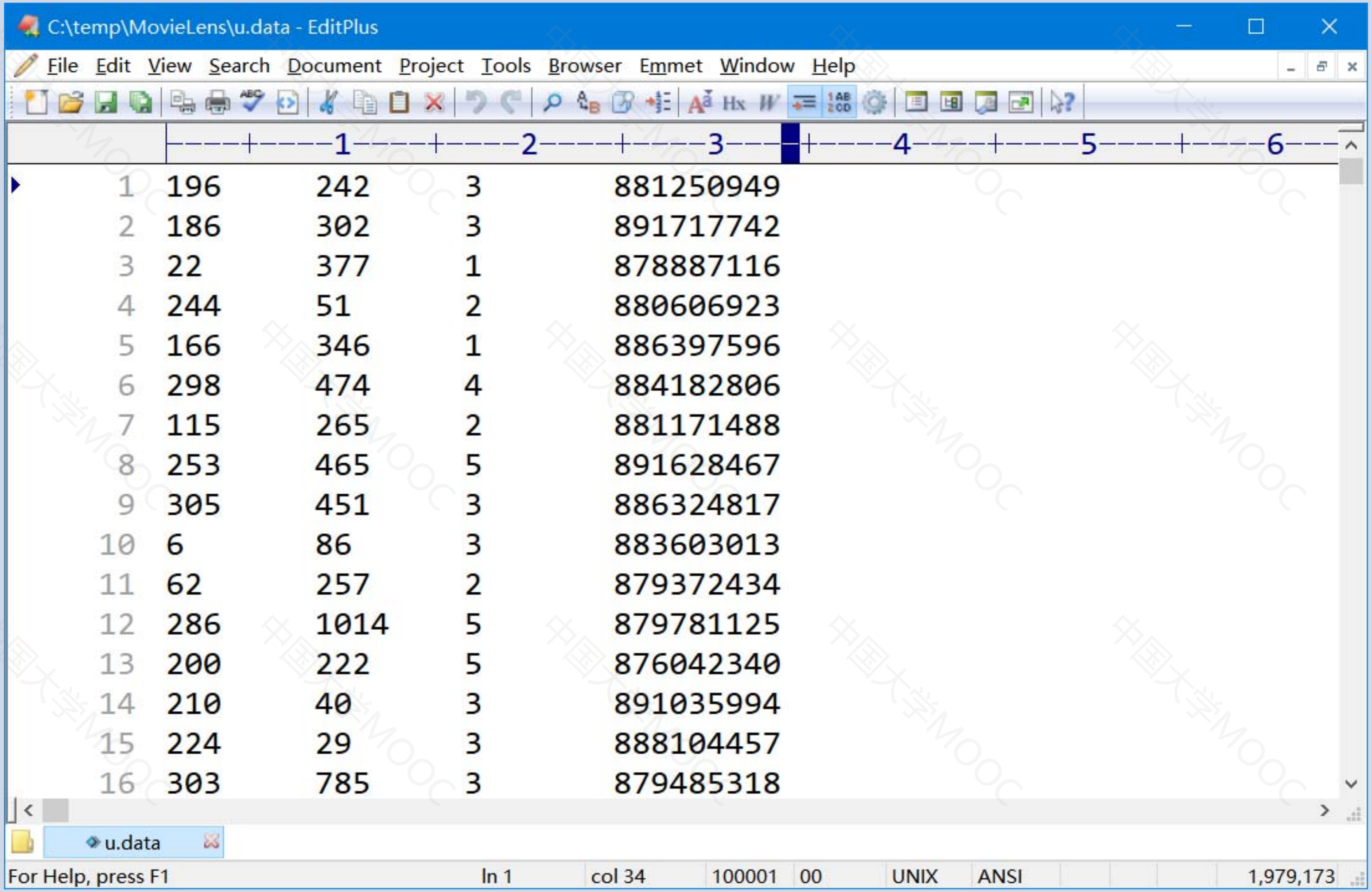
```
1|24|M|technician|85711
2|53|F|other|94043
3|23|M|writer|32067
4|24|M|technician|43537
5|33|F|other|15213
6|42|M|executive|98101
7|57|M|administrator|91344
8|36|M|administrator|05201
9|29|M|student|01002
10|53|M|lawyer|90703
11|39|F|other|30329
12|28|F|other|06405
13|47|M|educator|29206
14|45|M|scientist|55106
15|49|F|educator|97301
16|21|M|entertainment|10309
```



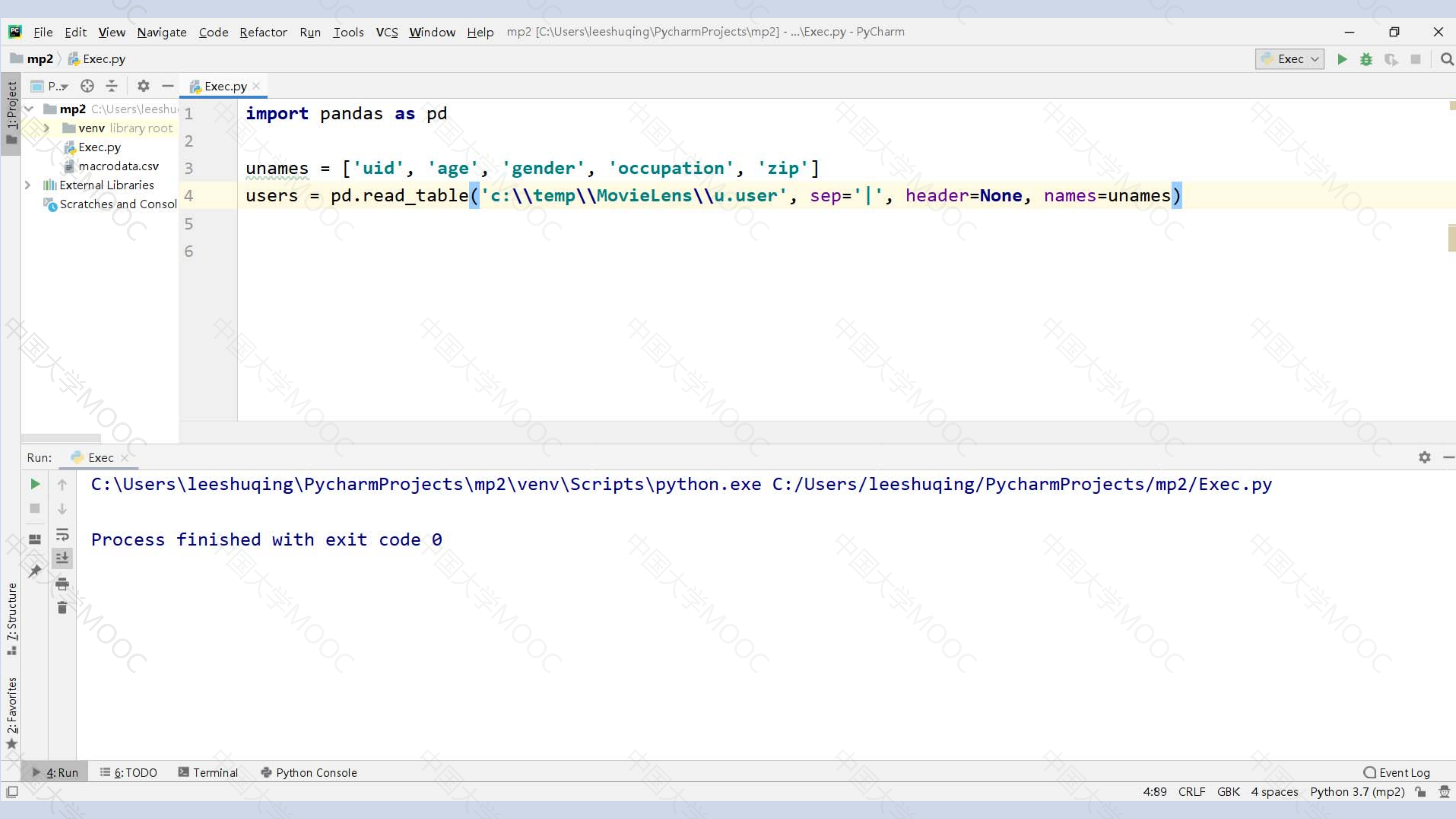
C:\temp\MovieLens\u.item - EditPlus
File Edit View Search Document Project Tools Browser Emmet Window Help
u.item
For Help, press F1

```
1|Toy Story (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Toy%20Story%20(1995)|0|0|0|1|1|1|0|0|0|0|0|0|0|0|0|0|0|
2|GoldenEye (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?GoldenEye%20(1995)|0|1|1|0|0|0|0|0|0|0|0|0|0|0|0|1|0|
3|Four Rooms (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Four%20Rooms%20(1995)|0|0|0|0|0|0|0|0|0|0|0|0|0|0|0|0|
4|Get Shorty (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Get%20Shorty%20(1995)|0|1|0|0|0|1|0|0|1|0|0|0|0|0|0|0|
5|Copycat (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Copycat%20(1995)|0|0|0|0|0|0|1|0|1|0|0|0|0|0|0|0|1|0|0
6|Shanghai Triad (Yao a yao yao dao waipo qiao) (1995)|01-Jan-1995||http://us.imdb.com/Title?Yao+a+yao+yao+dao+waipo+qia
7|Twelve Monkeys (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Twelve%20Monkeys%20(1995)|0|0|0|0|0|0|0|0|1|0|0|0|
8|Babe (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Babe%20(1995)|0|0|0|0|1|1|0|0|1|0|0|0|0|0|0|0|0|0|0
9|Dead Man Walking (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Dead%20Man%20Walking%20(1995)|0|0|0|0|0|0|0|0|1|
10|Richard III (1995)|22-Jan-1996||http://us.imdb.com/M/title-exact?Richard%20III%20(1995)|0|0|0|0|0|0|0|0|1|0|0|0|0|0|0
11|Seven (Se7en) (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Se7en%20(1995)|0|0|0|0|0|0|1|0|0|0|0|0|0|0|0|0|1|0
12|Usual Suspects, The (1995)|14-Aug-1995||http://us.imdb.com/M/title-exact?Usual%20Suspects,%20The%20(1995)|0|0|0|0|0|0
13|Mighty Aphrodite (1995)|30-Oct-1995||http://us.imdb.com/M/title-exact?Mighty%20Aphrodite%20(1995)|0|0|0|0|0|1|0|0|0|0
14|Postino, Il (1994)|01-Jan-1994||http://us.imdb.com/M/title-exact?Postino,%20Il%20(1994)|0|0|0|0|0|0|0|0|1|0|0|0|0|0|1
15|Mr. Holland's Opus (1995)|29-Jan-1996||http://us.imdb.com/M/title-exact?Mr.%20Holland's%20Opus%20(1995)|0|0|0|0|0|0|0
16|French Twist (Gazon maudit) (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Gazon%20maudit%20(1995)|0|0|0|0|0|1|
17|From Dusk Till Dawn (1996)|05-Feb-1996||http://us.imdb.com/M/title-exact?From%20Dusk%20Till%20Dawn%20(1996)|0|1|0|0|0
18|White Balloon, The (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Badkonake%20Sefid%20(1995)|0|0|0|0|0|0|0|0|1|
19|Antonia's Line (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Antonia%20(1995)|0|0|0|0|0|0|0|0|1|0|0|0|0|0|0|0|
20|Angels and Insects (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Angels%20and%20Insects%20(1995)|0|0|0|0|0|0|0
21|Muppet Treasure Island (1996)|16-Feb-1996||http://us.imdb.com/M/title-exact?Muppet%20Treasure%20Island%20(1996)|0|1|1
22|Braveheart (1995)|16-Feb-1996||http://us.imdb.com/M/title-exact?Braveheart%20(1995)|0|1|0|0|0|0|0|0|1|0|0|0|0|0|0|0|0
23|Taxi Driver (1976)|16-Feb-1996||http://us.imdb.com/M/title-exact?Taxi%20Driver%20(1976)|0|0|0|0|0|0|0|0|1|0|0|0|0|0|0
24|Rumble in the Bronx (1995)|23-Feb-1996||http://us.imdb.com/M/title-exact?Hong%20Faan%20Kui%20(1995)|0|1|1|0|0|0|1|0|0
25|Birdcage, The (1996)|08-Mar-1996||http://us.imdb.com/M/title-exact?Birdcage,%20The%20(1996)|0|0|0|0|0|1|0|0|0|0|0|0|0
26|Brothers McMullen, The (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Brothers%20McMullen,%20The%20(1995)|0|0|0
```

In 1 col 1 1683 31 UNIX ANSI 236,344

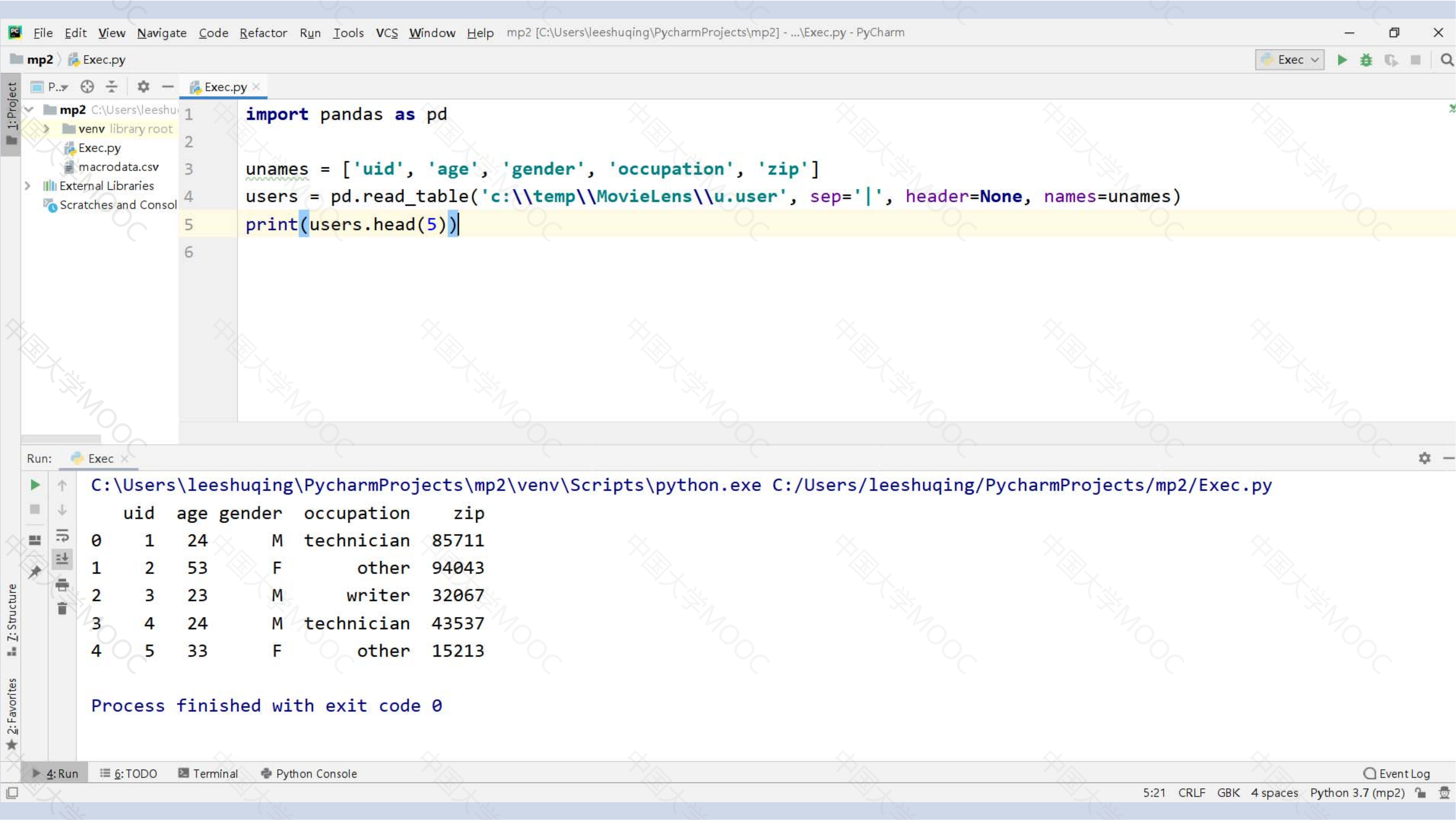

C:\temp\MovieLens\u.data - EditPlus
File Edit View Search Document Project Tools Browser Emmet Window Help
u.data
For Help, press F1
ln 1
col 34
100001
00
UNIX
ANSI
1,979,173

```
196     242     3       881250949
186     302     3       891717742
22      377     1       878887116
244     51      2       880606923
166     346     1       886397596
298     474     4       884182806
115     265     2       881171488
253     465     5       891628467
305     451     3       886324817
6       86      3       883603013
62      257     2       879372434
286     1014    5       879781125
200     222     5       876042340
210     40      3       891035994
224     29      3       888104457
303     785     3       879485318
```


File Edit View Navigate Code Refactor Run Tools VCS Window Help mp2 [C:\Users\leeshuqing\PycharmProjects\mp2] - ...\Exec.py - PyCharm
mp2
Exec.py
Exec
mp2 C:\Users\leeshu
venv library root
Exec.py
macrodata.csv
External Libraries
Scratches and Consol
Run:
Exec
4: Run
6: TODO
Terminal
Python Console
Event Log
4:89 CRLF GBK 4 spaces Python 3.7 (mp2)

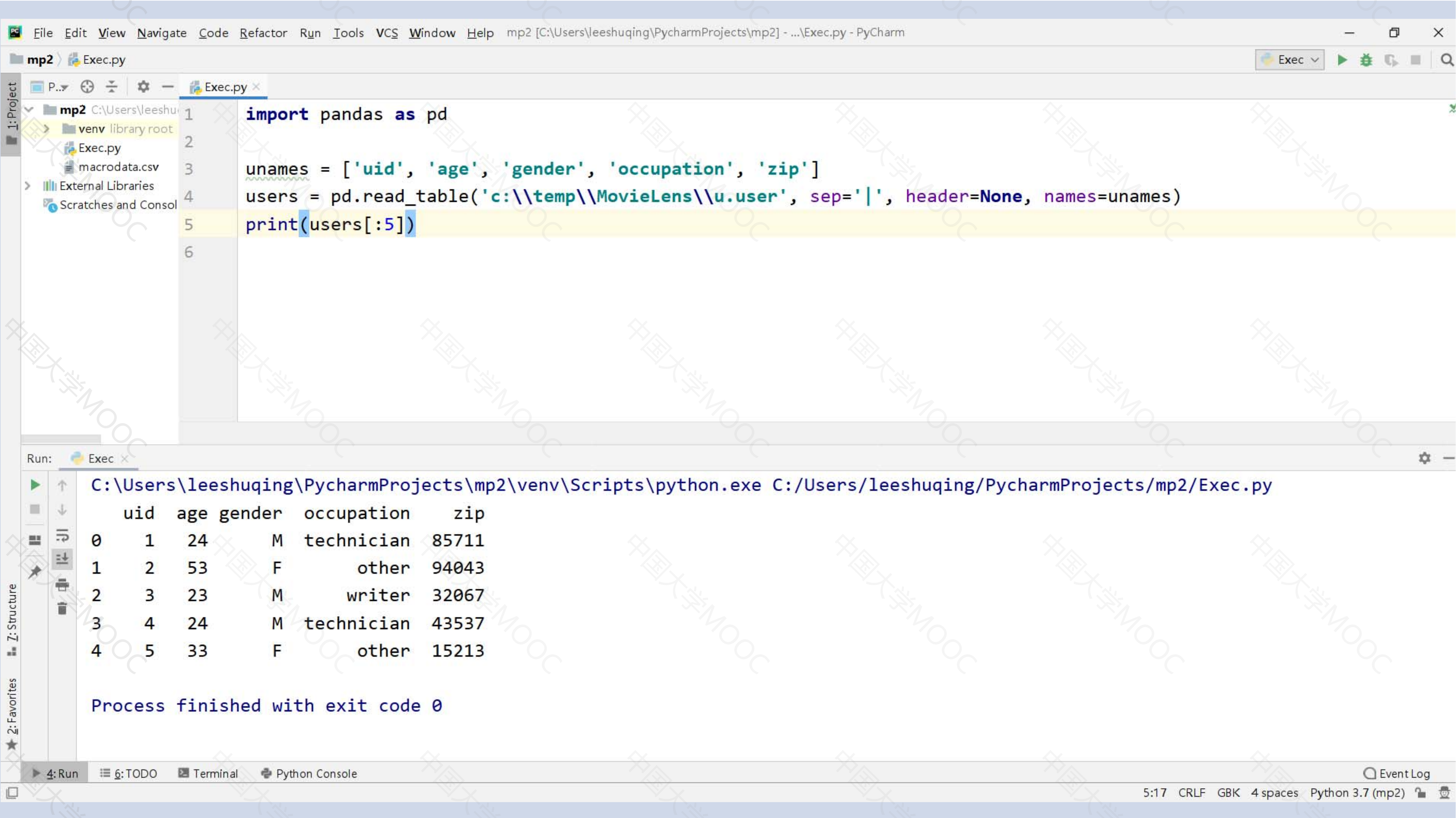
```
import pandas as pd

unames = ['uid', 'age', 'gender', 'occupation', 'zip']
users = pd.read_table('c:\\temp\\MovieLens\\u.user', sep='|', header=None, names=unames)
```

```
C:\Users\leeshuqing\PycharmProjects\mp2\venv\Scripts\python.exe C:/Users/leeshuqing/PycharmProjects/mp2/Exec.py

Process finished with exit code 0
```



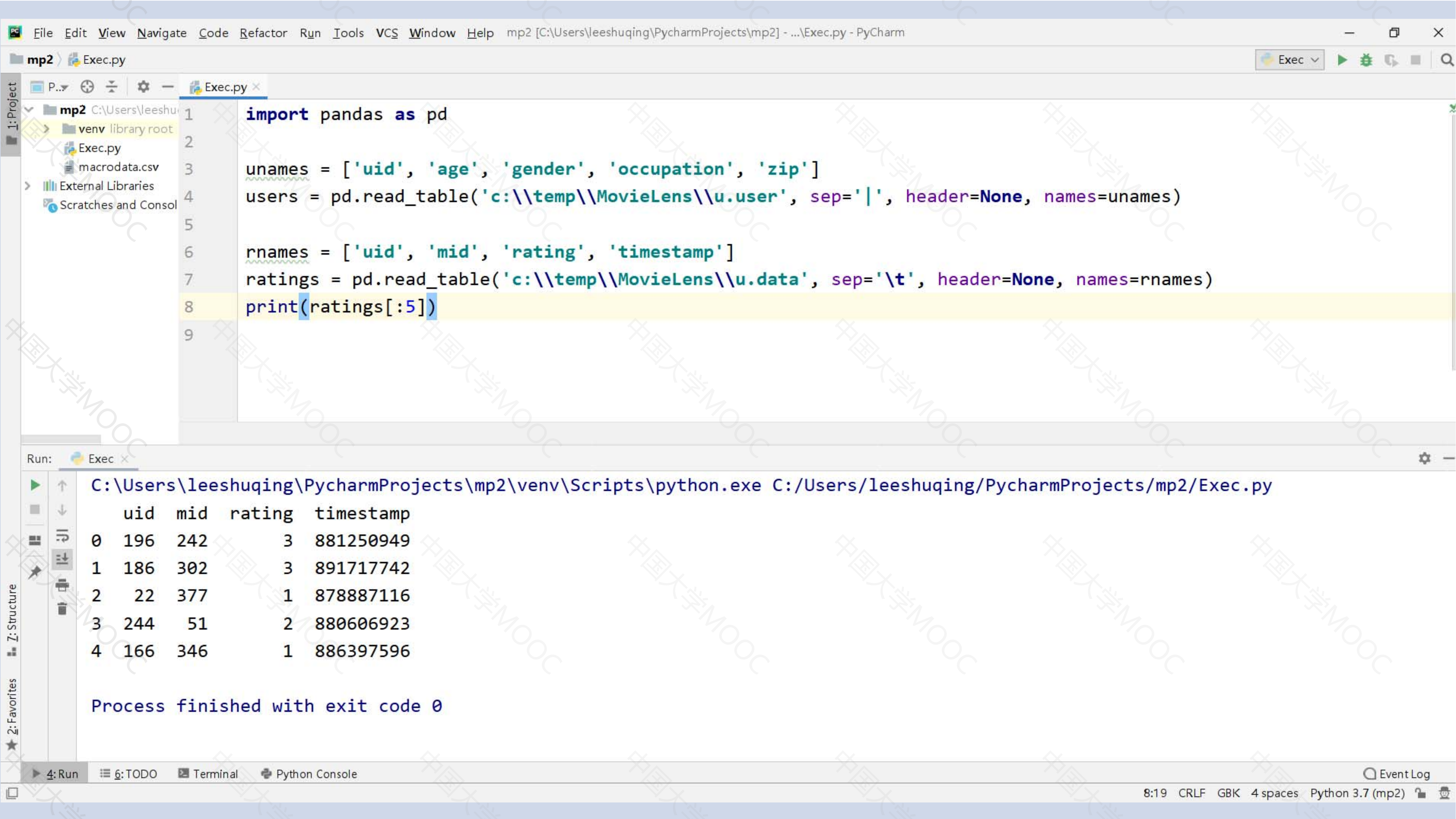
```python
import pandas as pd

unames = ['uid', 'age', 'gender', 'occupation', 'zip']
users = pd.read_table('c:\\temp\\MovieLens\\u.user', sep='|', header=None, names=unames)
print(users.head(5))
```

```
C:\Users\leeshuqing\PycharmProjects\mp2\venv\Scripts\python.exe C:/Users/leeshuqing/PycharmProjects/mp2/Exec.py
   uid  age gender  occupation    zip
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

Process finished with exit code 0
```



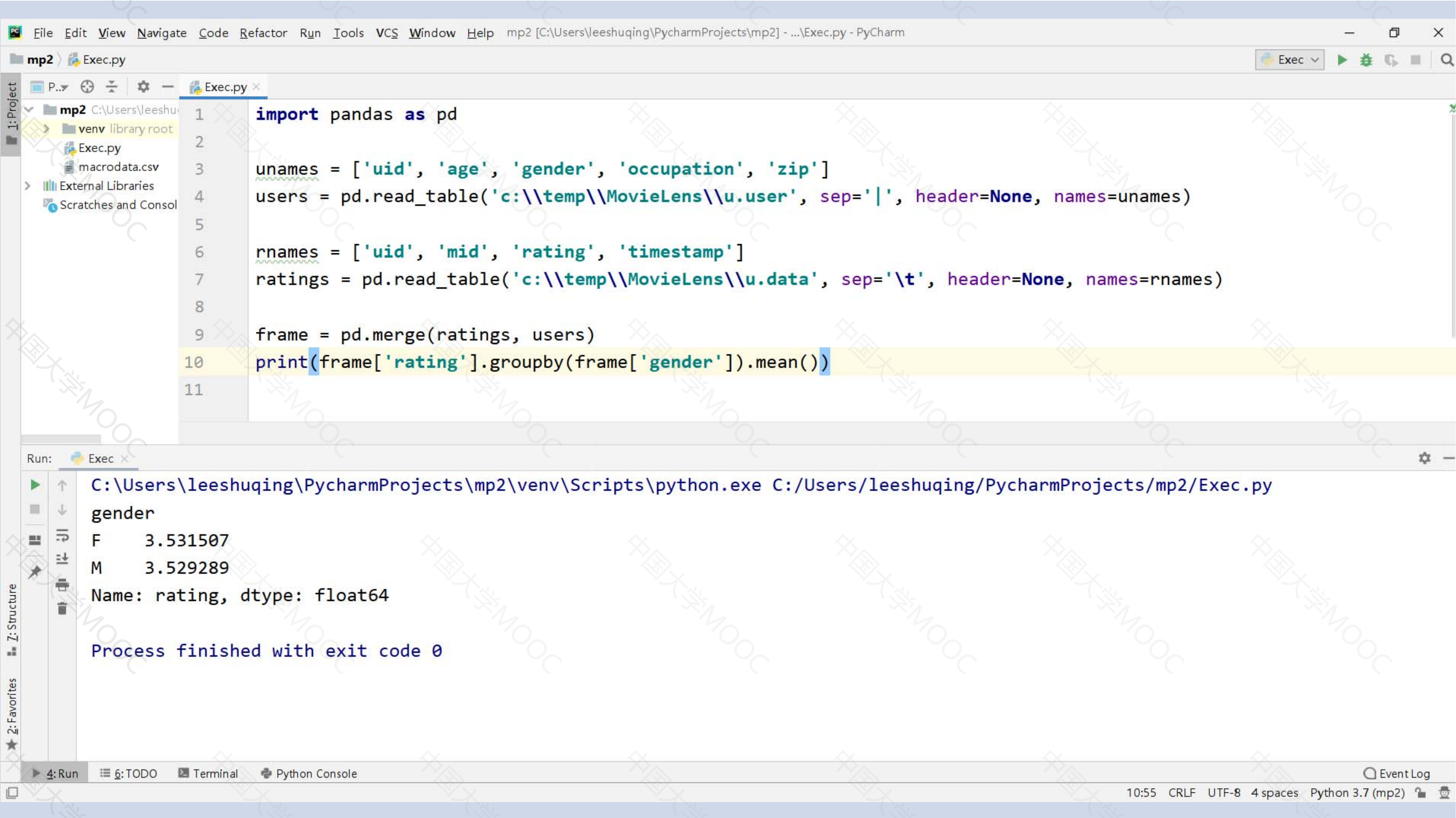
```python
import pandas as pd

unames = ['uid', 'age', 'gender', 'occupation', 'zip']
users = pd.read_table('c:\\temp\\MovieLens\\u.user', sep='|', header=None, names=unames)
print(users[:5])
```

```
C:\Users\leeshuqing\PycharmProjects\mp2\venv\Scripts\python.exe C:/Users/leeshuqing/PycharmProjects/mp2/Exec.py
   uid  age gender  occupation    zip
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

Process finished with exit code 0
```


File Edit View Navigate Code Refactor Run Tools VCS Window Help mp2 [C:\Users\leeshuqing\PycharmProjects\mp2] - ...\Exec.py - PyCharm
mp2 Exec.py
Exec
mp2 C:\Users\leeshu
venv library root
Exec.py
macrodata.csv
External Libraries
Scratches and Consol
Run: Exec
4: Run 6: TODO Terminal Python Console Event Log
8:19 CRLF GBK 4 spaces Python 3.7 (mp2)

```
import pandas as pd

unames = ['uid', 'age', 'gender', 'occupation', 'zip']
users = pd.read_table('c:\\temp\\MovieLens\\u.user', sep='|', header=None, names=unames)

rnames = ['uid', 'mid', 'rating', 'timestamp']
ratings = pd.read_table('c:\\temp\\MovieLens\\u.data', sep='\t', header=None, names=rnames)
print(ratings[:5])
```

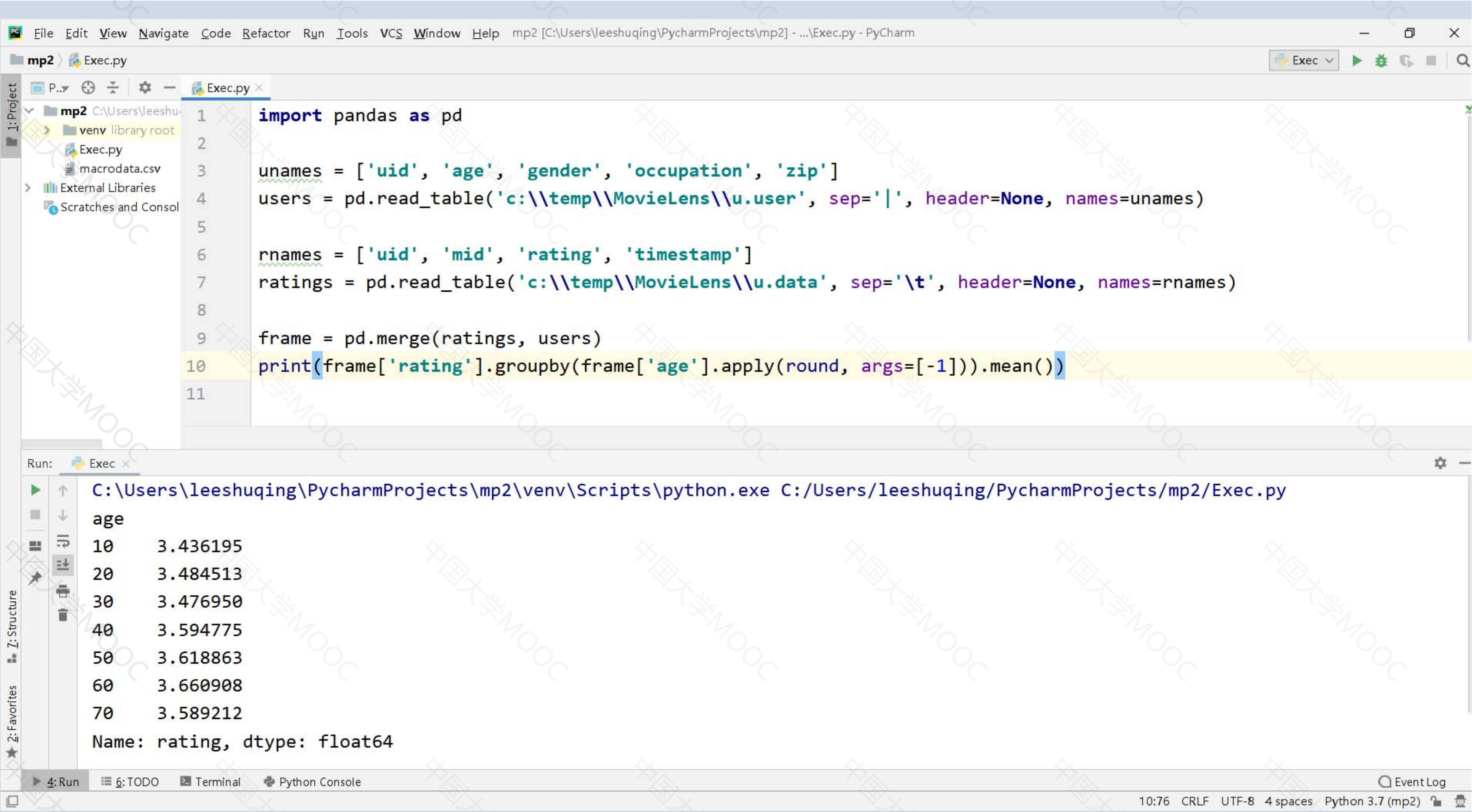
```
C:\Users\leeshuqing\PycharmProjects\mp2\venv\Scripts\python.exe C:/Users/leeshuqing/PycharmProjects/mp2/Exec.py
   uid  mid  rating  timestamp
0  196  242       3  881250949
1  186  302       3  891717742
2   22  377       1  878887116
3  244   51       2  880606923
4  166  346       1  886397596

Process finished with exit code 0
```



File Edit View Navigate Code Refactor Run Tools VCS Window Help mp2 [C:\Users\leeshuqing\PycharmProjects\mp2] - ...\Exec.py - PyCharm
mp2 Exec.py
Exec
mp2 C:\Users\leeshu
venv library root
Exec.py
macrodata.csv
External Libraries
Scratches and Consol
Run: Exec
4: Run 6: TODO Terminal Python Console
Event Log
10:55 CRLF UTF-8 4 spaces Python 3.7 (mp2)

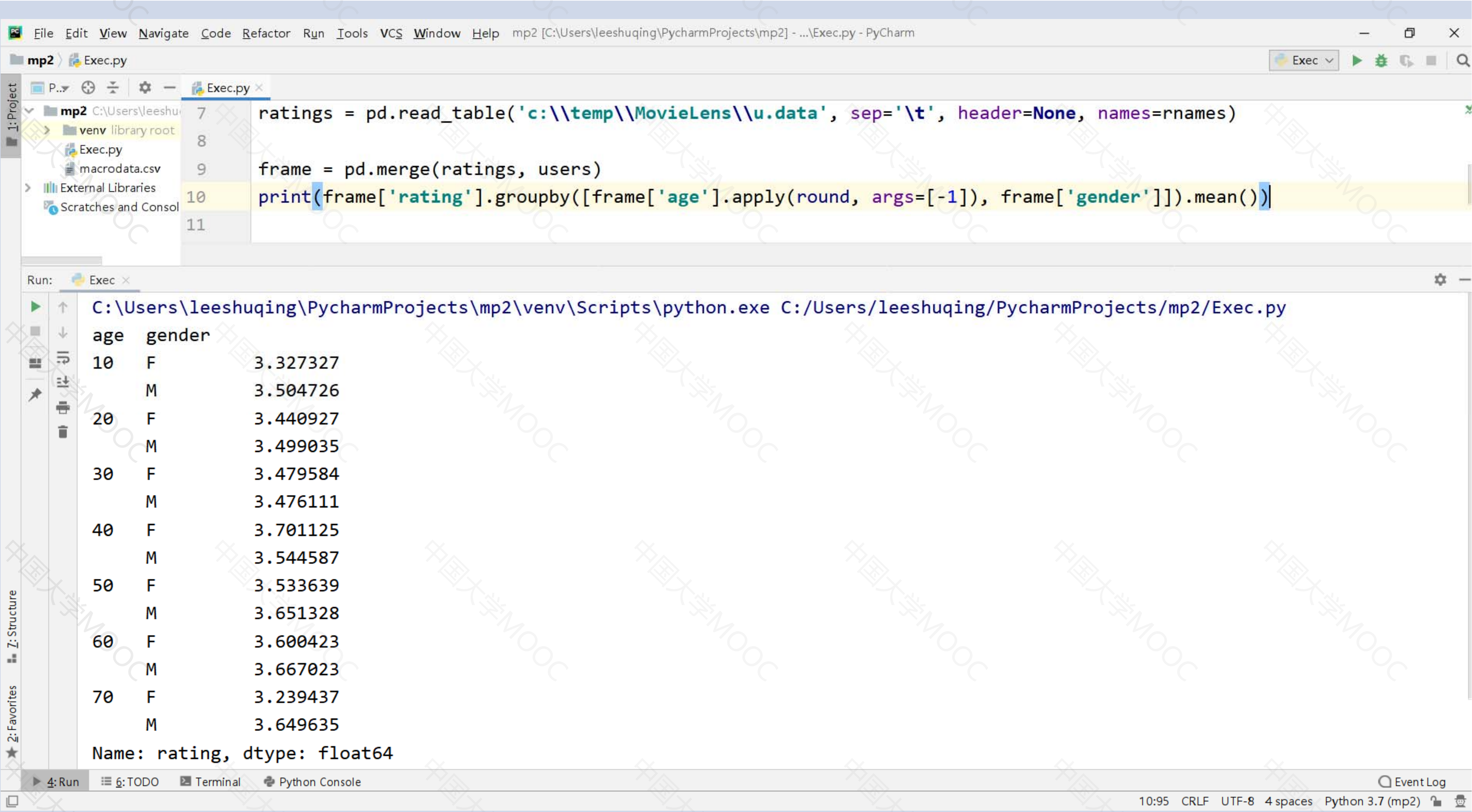
```python
import pandas as pd

unames = ['uid', 'age', 'gender', 'occupation', 'zip']
users = pd.read_table('c:\\temp\\MovieLens\\u.user', sep='|', header=None, names=unames)

rnames = ['uid', 'mid', 'rating', 'timestamp']
ratings = pd.read_table('c:\\temp\\MovieLens\\u.data', sep='\t', header=None, names=rnames)

frame = pd.merge(ratings, users)
print(frame['rating'].groupby(frame['gender']).mean())
```

```
C:\Users\leeshuqing\PycharmProjects\mp2\venv\Scripts\python.exe C:/Users/leeshuqing/PycharmProjects/mp2/Exec.py
gender
F    3.531507
M    3.529289
Name: rating, dtype: float64

Process finished with exit code 0
```



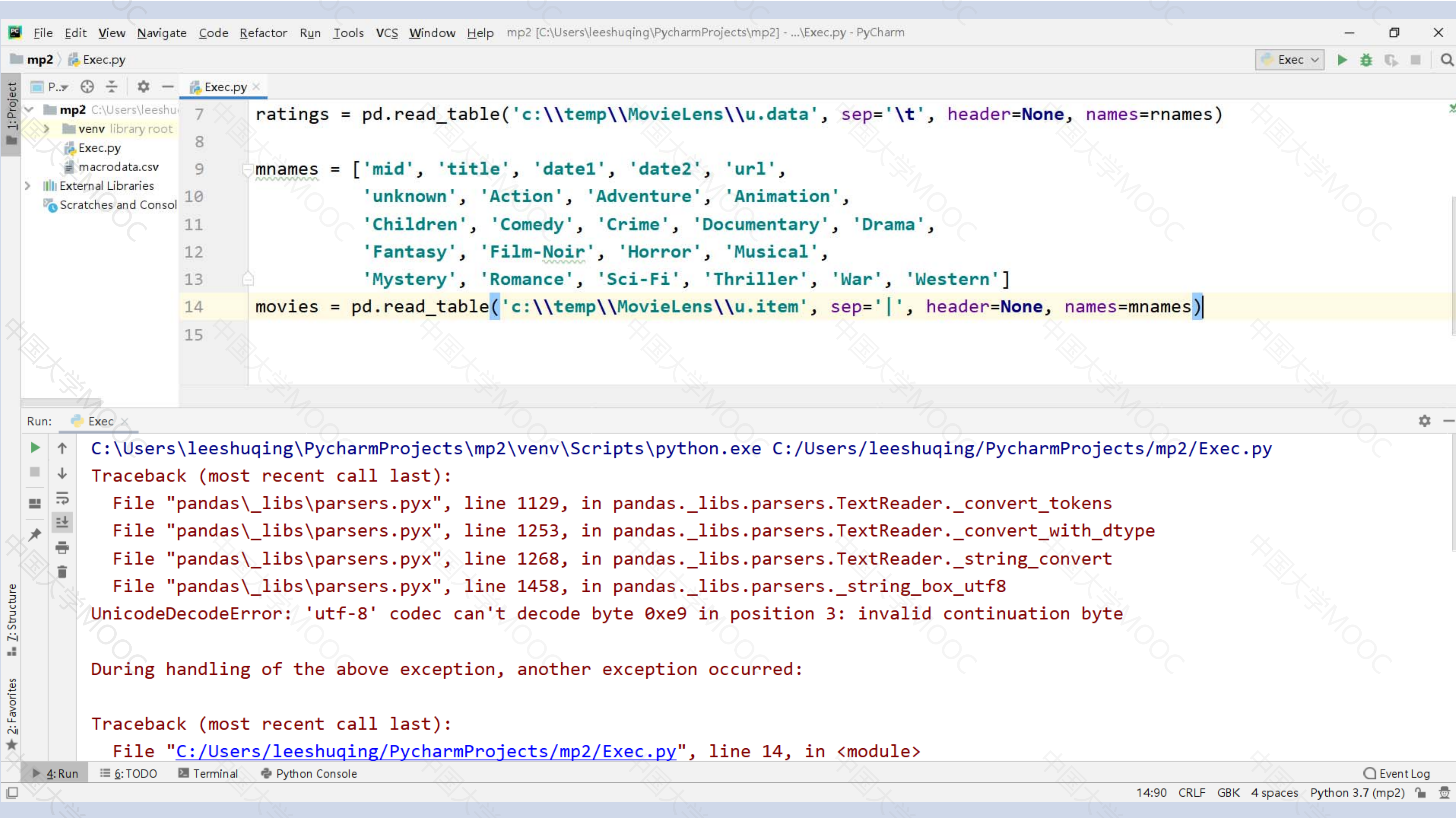
```python
import pandas as pd

unames = ['uid', 'age', 'gender', 'occupation', 'zip']
users = pd.read_table('c:\\temp\\MovieLens\\u.user', sep='|', header=None, names=unames)

rnames = ['uid', 'mid', 'rating', 'timestamp']
ratings = pd.read_table('c:\\temp\\MovieLens\\u.data', sep='\t', header=None, names=rnames)

frame = pd.merge(ratings, users)
print(frame['rating'].groupby(frame['age'].apply(round, args=[-1])).mean())
```

```
C:\Users\leeshuqing\PycharmProjects\mp2\venv\Scripts\python.exe C:/Users/leeshuqing/PycharmProjects/mp2/Exec.py
age
10    3.436195
20    3.484513
30    3.476950
40    3.594775
50    3.618863
60    3.660908
70    3.589212
Name: rating, dtype: float64
```



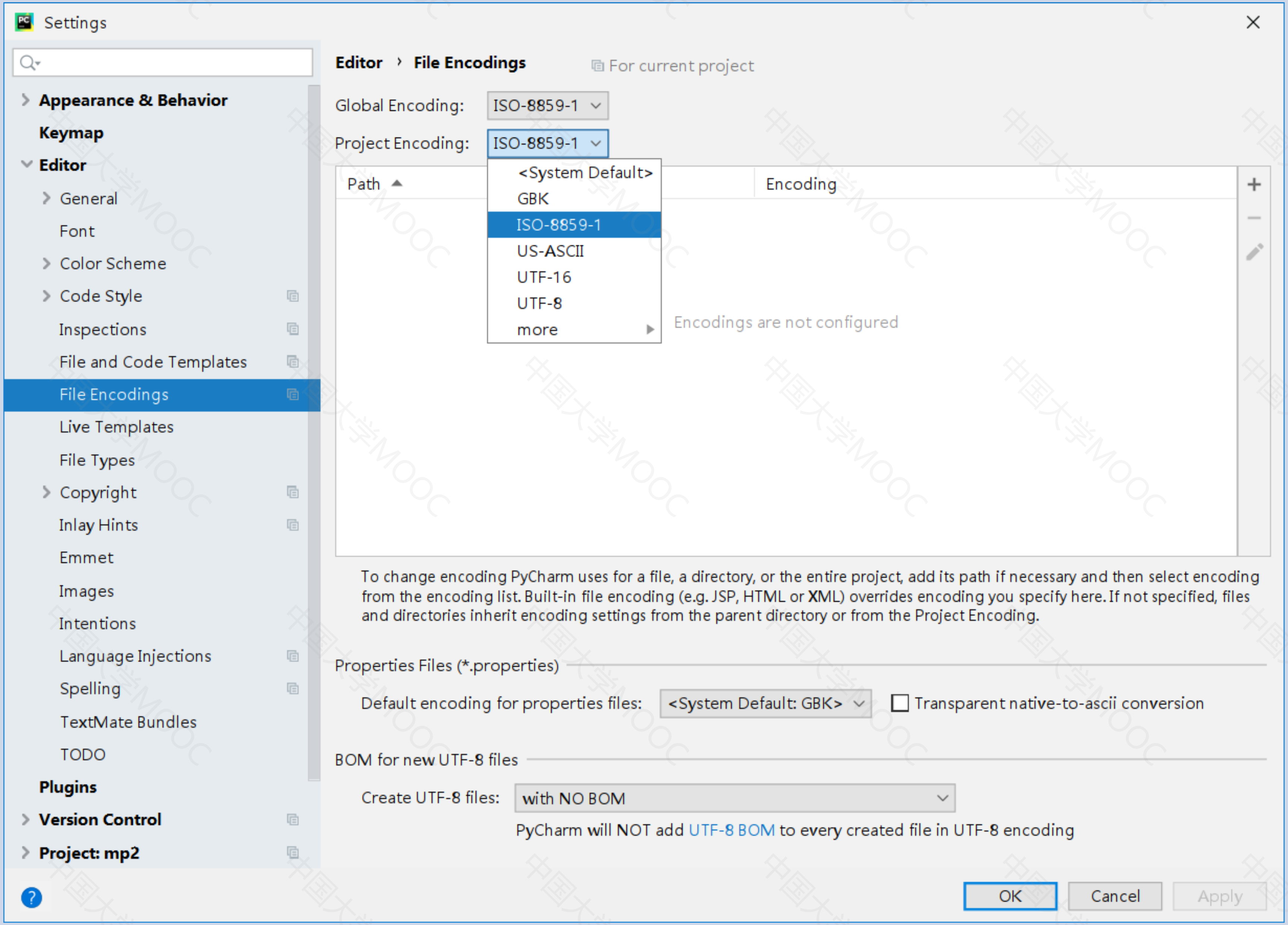
```python
ratings = pd.read_table('c:\\temp\\MovieLens\\u.data', sep='\t', header=None, names=rnames)

frame = pd.merge(ratings, users)
print(frame['rating'].groupby([frame['age'].apply(round, args=[-1]), frame['gender']]).mean())

```

```
C:\Users\leeshuqing\PycharmProjects\mp2\venv\Scripts\python.exe C:/Users/leeshuqing/PycharmProjects/mp2/Exec.py
age  gender
10   F         3.327327
     M         3.504726
20   F         3.440927
     M         3.499035
30   F         3.479584
     M         3.476111
40   F         3.701125
     M         3.544587
50   F         3.533639
     M         3.651328
60   F         3.600423
     M         3.667023
70   F         3.239437
     M         3.649635
Name: rating, dtype: float64
```



```
ratings = pd.read_table('c:\\temp\\MovieLens\\u.data', sep='\t', header=None, names=rnames)

mnames = ['mid', 'title', 'date1', 'date2', 'url',
          'unknown', 'Action', 'Adventure', 'Animation',
          'Children', 'Comedy', 'Crime', 'Documentary', 'Drama',
          'Fantasy', 'Film-Noir', 'Horror', 'Musical',
          'Mystery', 'Romance', 'Sci-Fi', 'Thriller', 'War', 'Western']
movies = pd.read_table('c:\\temp\\MovieLens\\u.item', sep='|', header=None, names=mnames)

```

```
C:\Users\leeshuqing\PycharmProjects\mp2\venv\Scripts\python.exe C:/Users/leeshuqing/PycharmProjects/mp2/Exec.py
Traceback (most recent call last):
  File "pandas\_libs\parsers.pyx", line 1129, in pandas._libs.parsers.TextReader._convert_tokens
  File "pandas\_libs\parsers.pyx", line 1253, in pandas._libs.parsers.TextReader._convert_with_dtype
  File "pandas\_libs\parsers.pyx", line 1268, in pandas._libs.parsers.TextReader._string_convert
  File "pandas\_libs\parsers.pyx", line 1458, in pandas._libs.parsers._string_box_utf8
UnicodeDecodeError: 'utf-8' codec can't decode byte 0xe9 in position 3: invalid continuation byte

During handling of the above exception, another exception occurred:

Traceback (most recent call last):
  File "C:/Users/leeshuqing/PycharmProjects/mp2/Exec.py", line 14, in <module>
```



Global Encoding: ISO-8859-1

Project Encoding: ISO-8859-1

To change encoding PyCharm uses for a file, a directory, or the entire project, add its path if necessary and then select encoding from the encoding list. Built-in file encoding (e.g. JSP, HTML or XML) overrides encoding you specify here. If not specified, files and directories inherit encoding settings from the parent directory or from the Project Encoding.

Properties Files (*.properties)

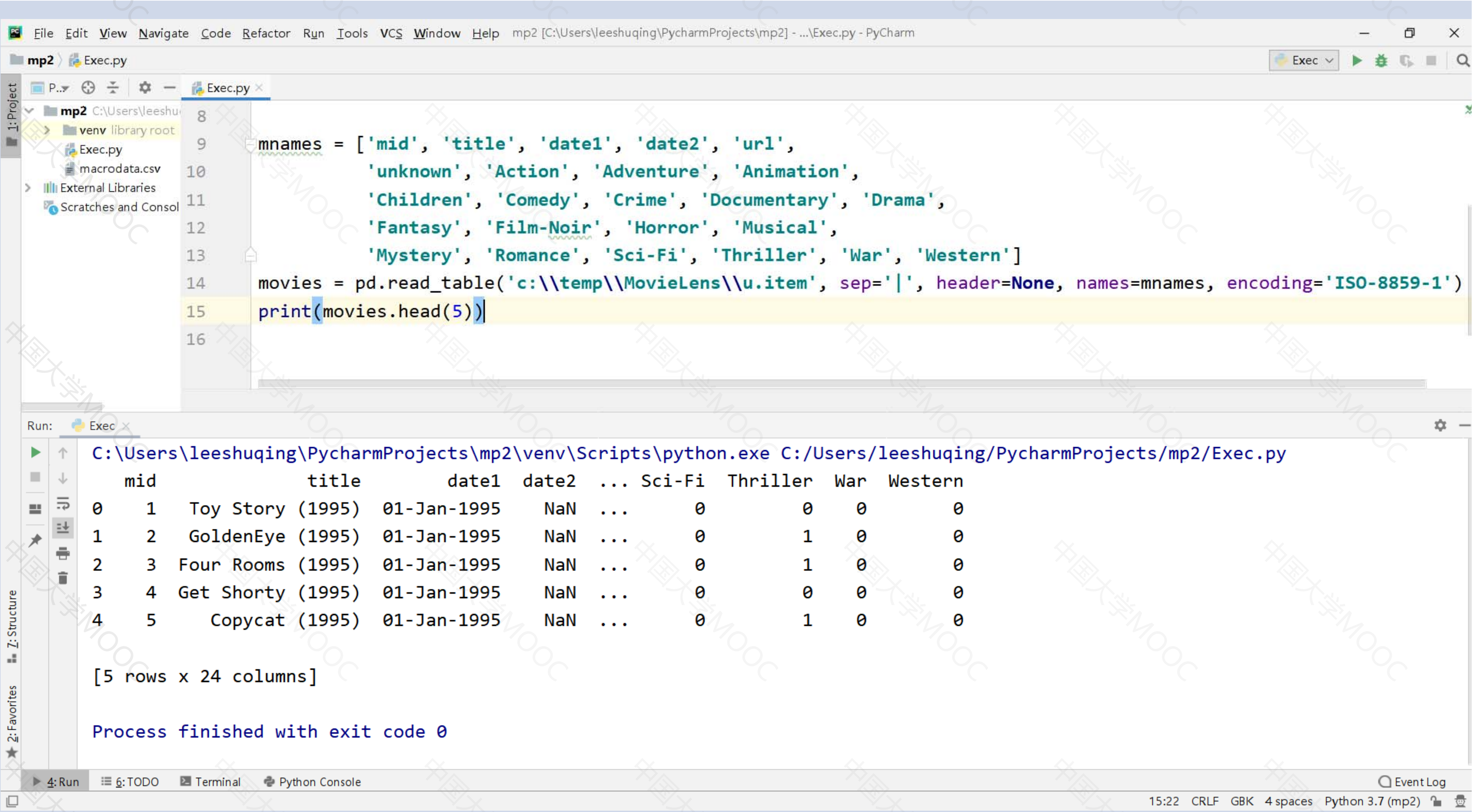
Default encoding for properties files: <System Default: GBK>

□ Transparent native-to-ascii conversion

BOM for new UTF-8 files

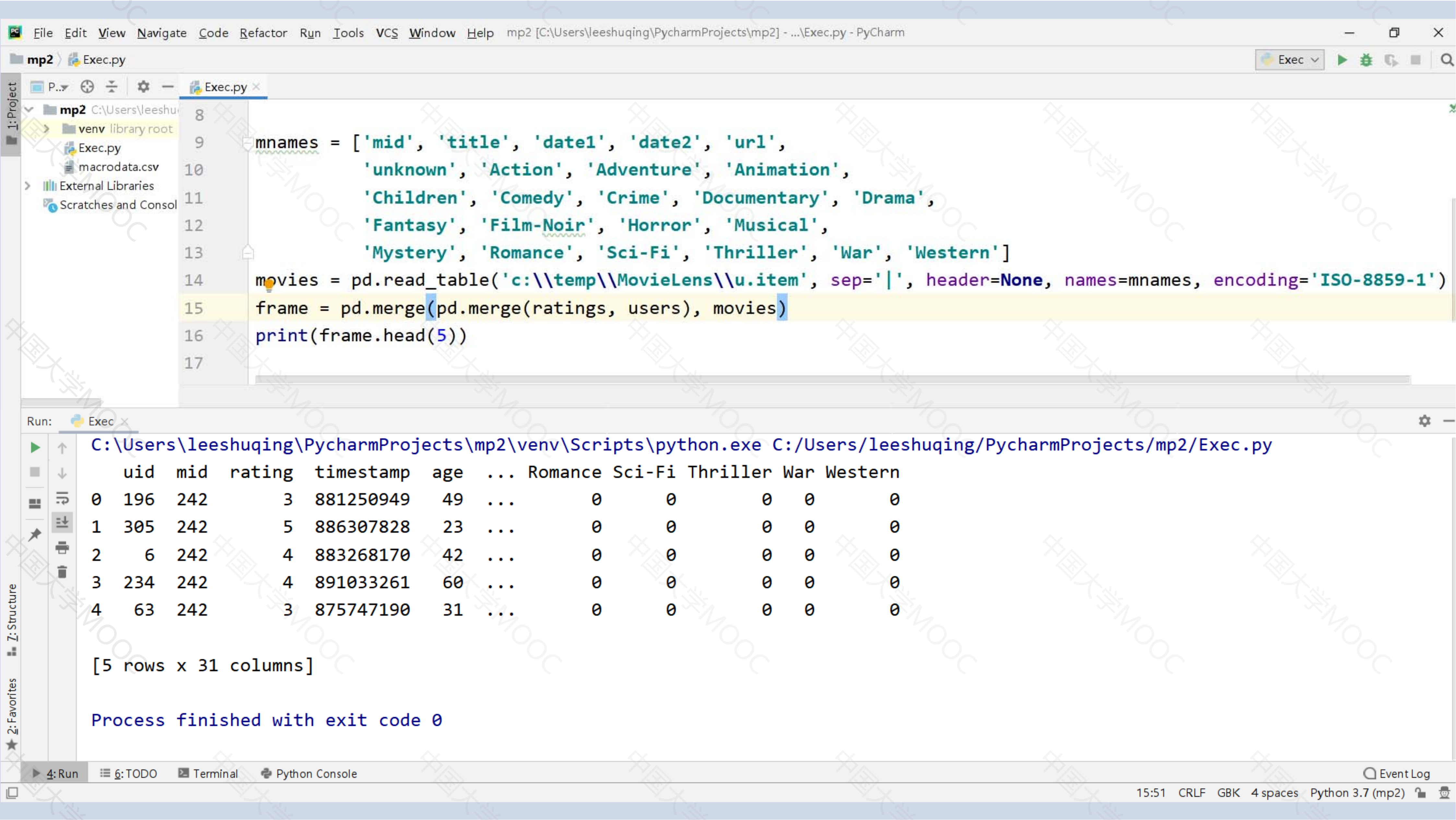
Create UTF-8 files: with NO BOM

PyCharm will NOT add UTF-8 BOM to every created file in UTF-8 encoding



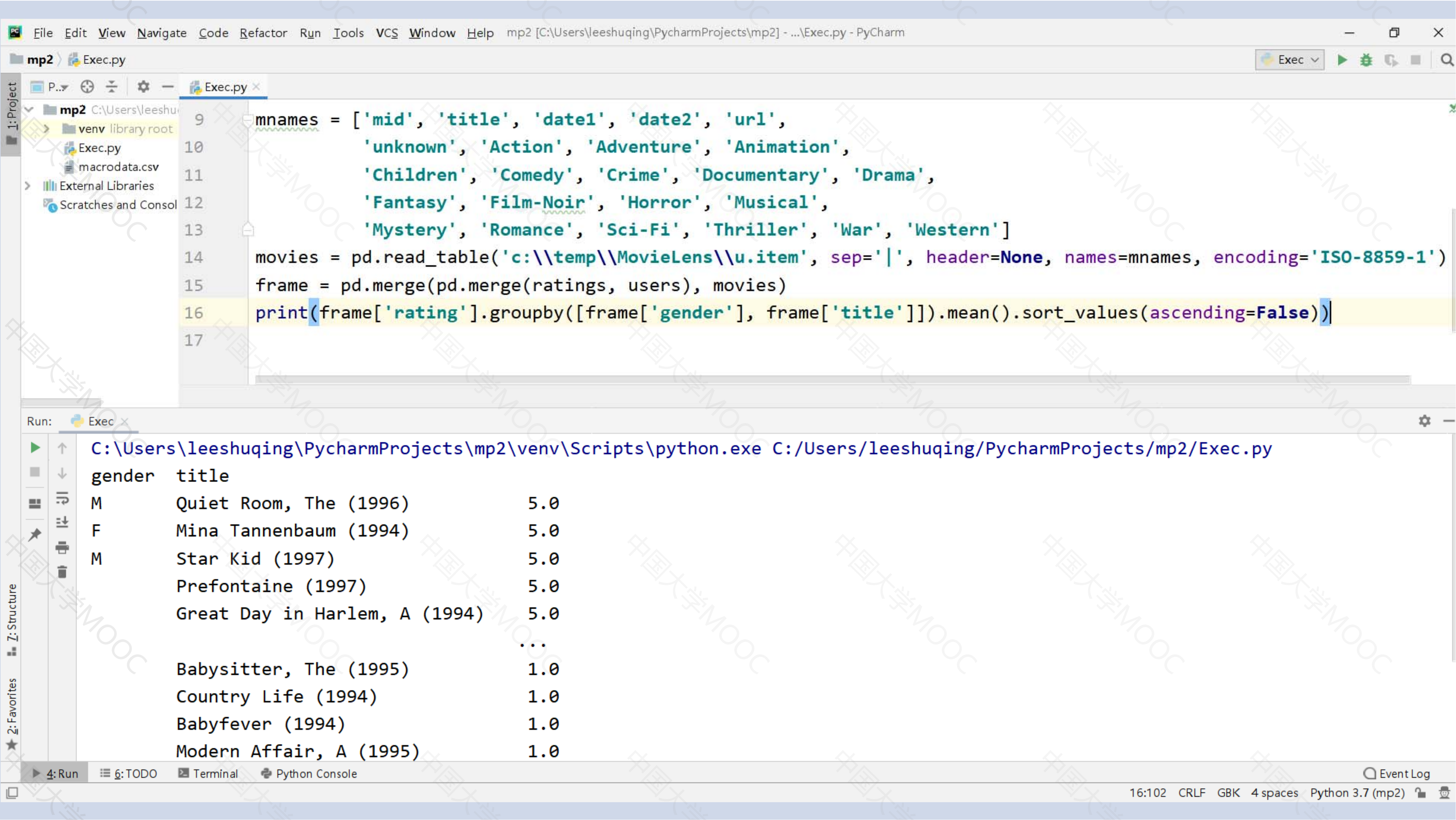
```python

mnames = ['mid', 'title', 'date1', 'date2', 'url',
          'unknown', 'Action', 'Adventure', 'Animation',
          'Children', 'Comedy', 'Crime', 'Documentary', 'Drama',
          'Fantasy', 'Film-Noir', 'Horror', 'Musical',
          'Mystery', 'Romance', 'Sci-Fi', 'Thriller', 'War', 'Western']
movies = pd.read_table('c:\\temp\\MovieLens\\u.item', sep='|', header=None, names=mnames, encoding='ISO-8859-1')
print(movies.head(5))

```

```
C:\Users\leeshuqing\PycharmProjects\mp2\venv\Scripts\python.exe C:/Users/leeshuqing/PycharmProjects/mp2/Exec.py
   mid              title        date1  date2  ... Sci-Fi  Thriller  War  Western
0    1   Toy Story (1995)  01-Jan-1995    NaN  ...      0         0    0        0
1    2   GoldenEye (1995)  01-Jan-1995    NaN  ...      0         1    0        0
2    3  Four Rooms (1995)  01-Jan-1995    NaN  ...      0         1    0        0
3    4  Get Shorty (1995)  01-Jan-1995    NaN  ...      0         0    0        0
4    5     Copycat (1995)  01-Jan-1995    NaN  ...      0         1    0        0

[5 rows x 24 columns]

Process finished with exit code 0
```



```python
mnames = ['mid', 'title', 'date1', 'date2', 'url',
          'unknown', 'Action', 'Adventure', 'Animation',
          'Children', 'Comedy', 'Crime', 'Documentary', 'Drama',
          'Fantasy', 'Film-Noir', 'Horror', 'Musical',
          'Mystery', 'Romance', 'Sci-Fi', 'Thriller', 'War', 'Western']
movies = pd.read_table('c:\\temp\\MovieLens\\u.item', sep='|', header=None, names=mnames, encoding='ISO-8859-1')
frame = pd.merge(pd.merge(ratings, users), movies)
print(frame.head(5))
```

```
C:\Users\leeshuqing\PycharmProjects\mp2\venv\Scripts\python.exe C:/Users/leeshuqing/PycharmProjects/mp2/Exec.py
   uid  mid  rating  timestamp  age  ... Romance  Sci-Fi  Thriller  War  Western
0  196  242       3  881250949   49  ...       0       0         0    0        0
1  305  242       5  886307828   23  ...       0       0         0    0        0
2    6  242       4  883268170   42  ...       0       0         0    0        0
3  234  242       4  891033261   60  ...       0       0         0    0        0
4   63  242       3  875747190   31  ...       0       0         0    0        0

[5 rows x 31 columns]

Process finished with exit code 0
```



```python
mnames = ['mid', 'title', 'date1', 'date2', 'url',
          'unknown', 'Action', 'Adventure', 'Animation',
          'Children', 'Comedy', 'Crime', 'Documentary', 'Drama',
          'Fantasy', 'Film-Noir', 'Horror', 'Musical',
          'Mystery', 'Romance', 'Sci-Fi', 'Thriller', 'War', 'Western']
movies = pd.read_table('c:\\temp\\MovieLens\\u.item', sep='|', header=None, names=mnames, encoding='ISO-8859-1')
frame = pd.merge(pd.merge(ratings, users), movies)
print(frame['rating'].groupby([frame['gender'], frame['title']]).mean().sort_values(ascending=False))
```

```
C:\Users\leeshuqing\PycharmProjects\mp2\venv\Scripts\python.exe C:/Users/leeshuqing/PycharmProjects/mp2/Exec.py
gender  title
M       Quiet Room, The (1996)          5.0
F       Mina Tannenbaum (1994)          5.0
M       Star Kid (1997)                 5.0
        Prefontaine (1997)              5.0
        Great Day in Harlem, A (1994)   5.0
                                       ...
        Babysitter, The (1995)          1.0
        Country Life (1994)             1.0
        Babyfever (1994)                1.0
        Modern Affair, A (1995)         1.0
```

一次不学多，下次再学