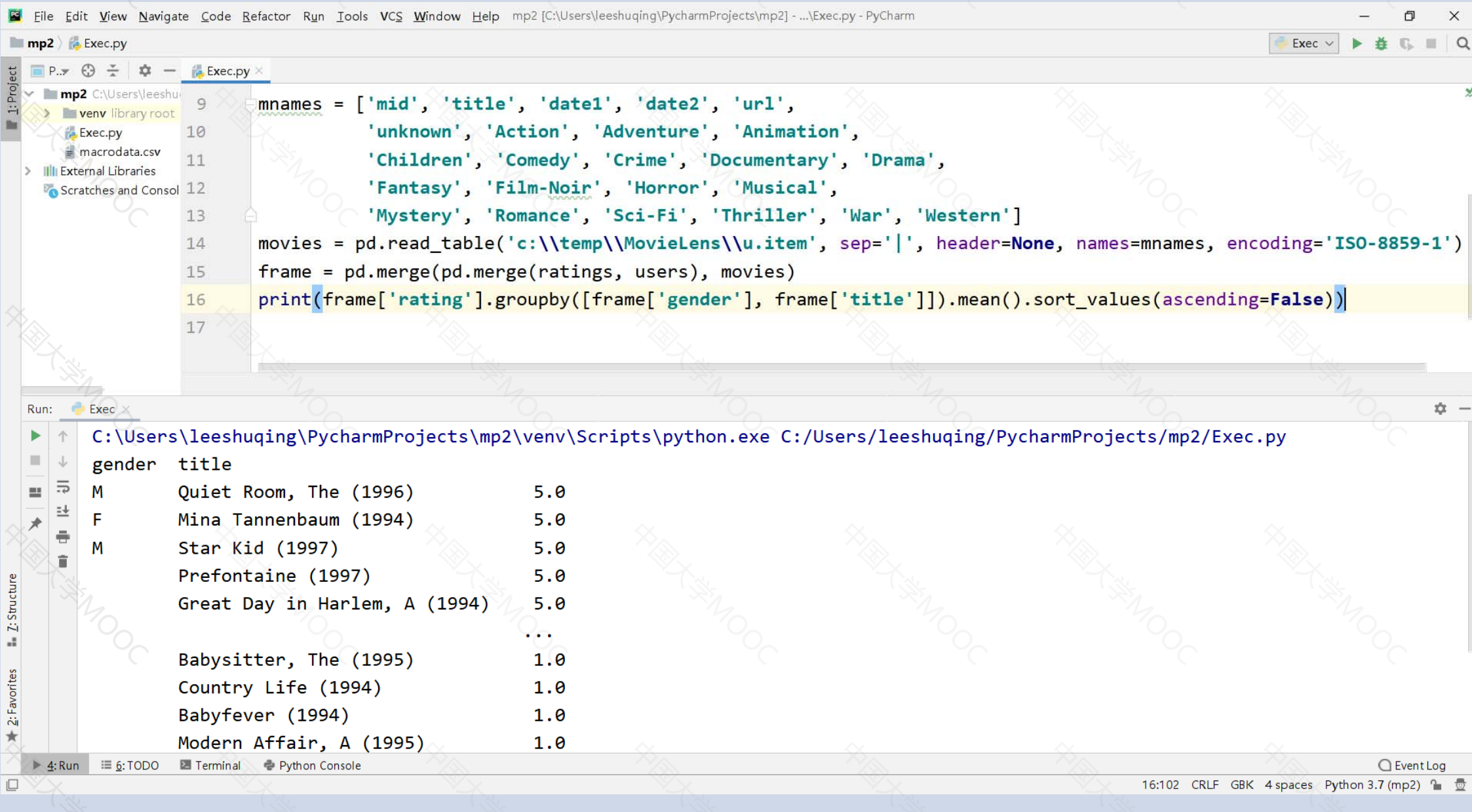
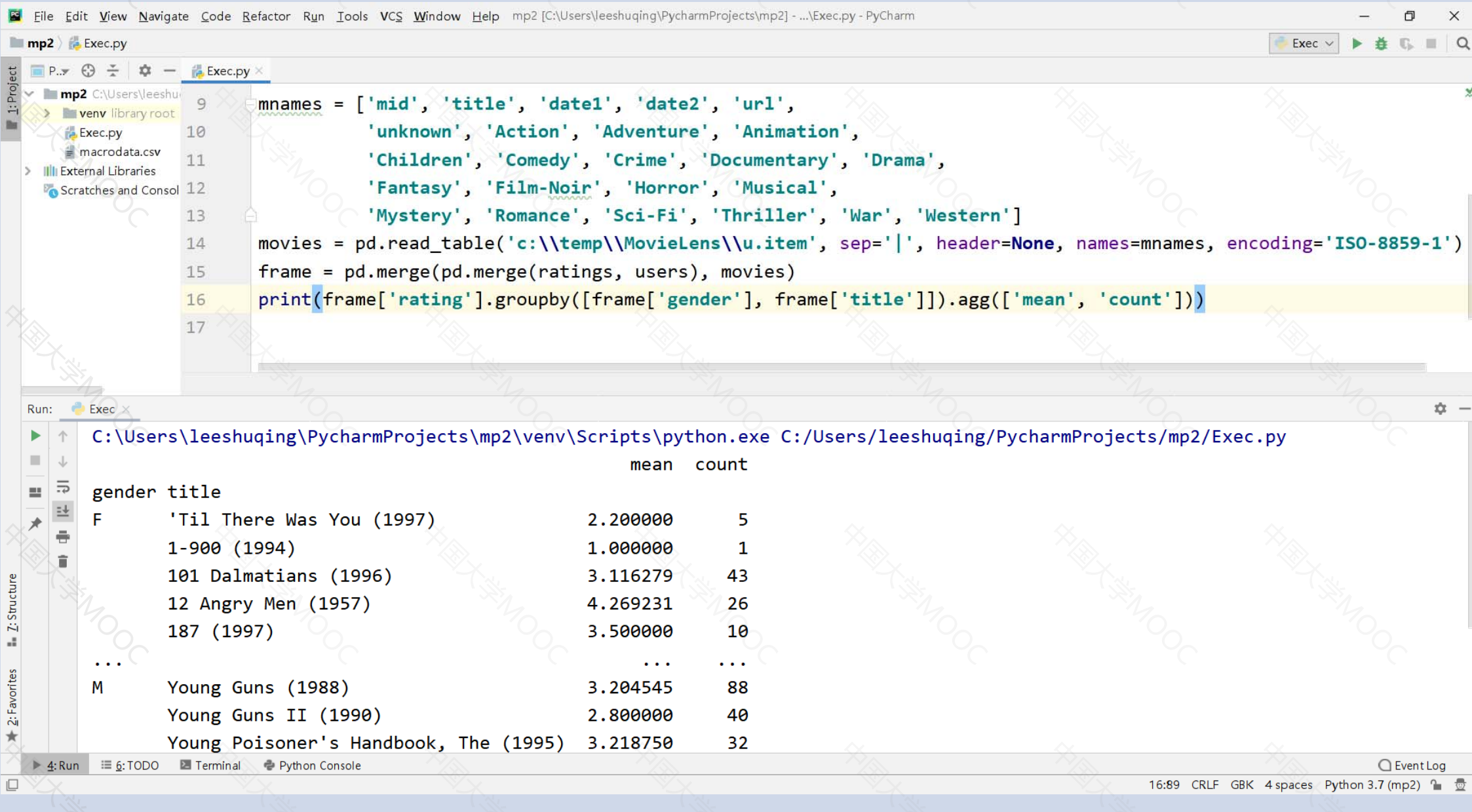
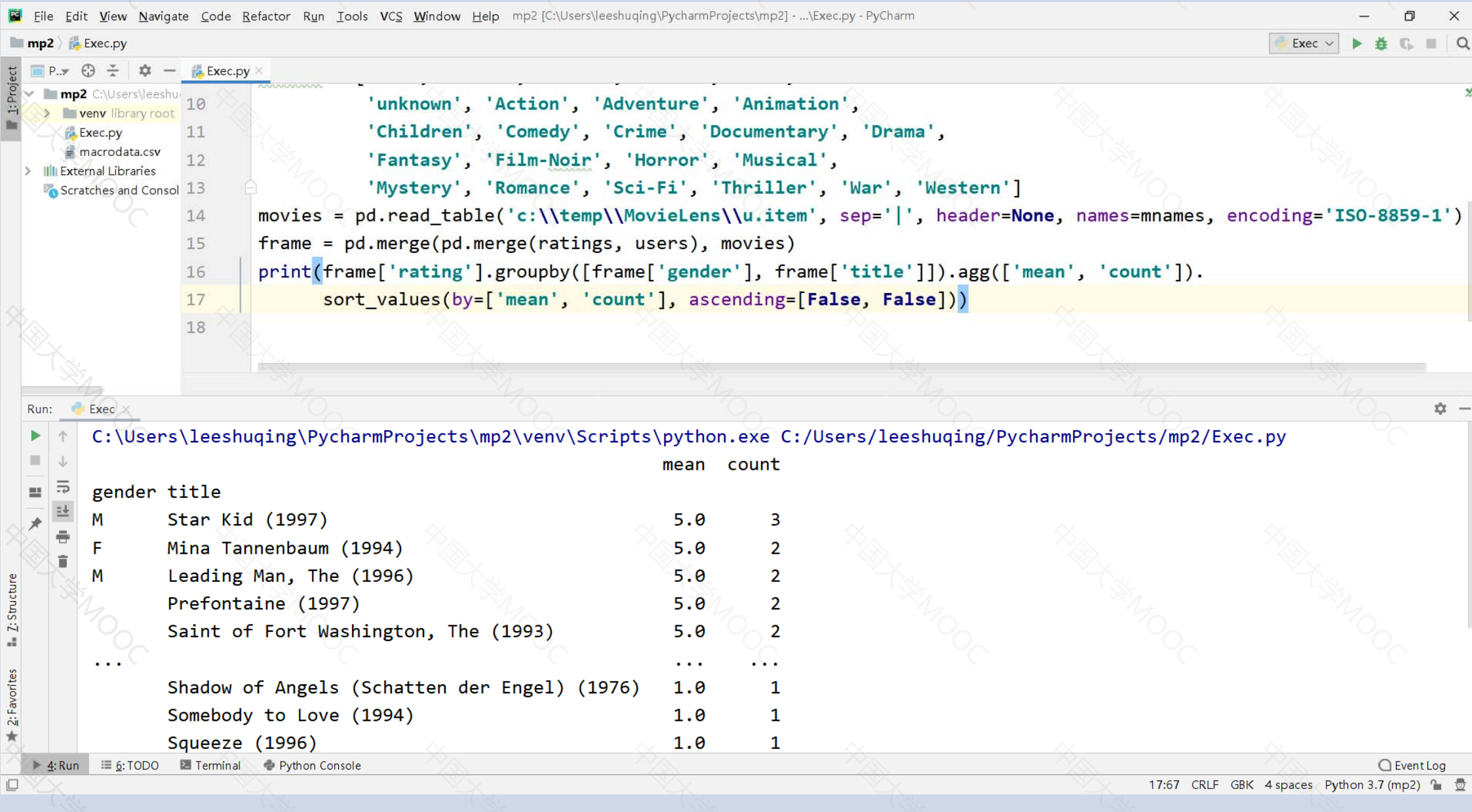
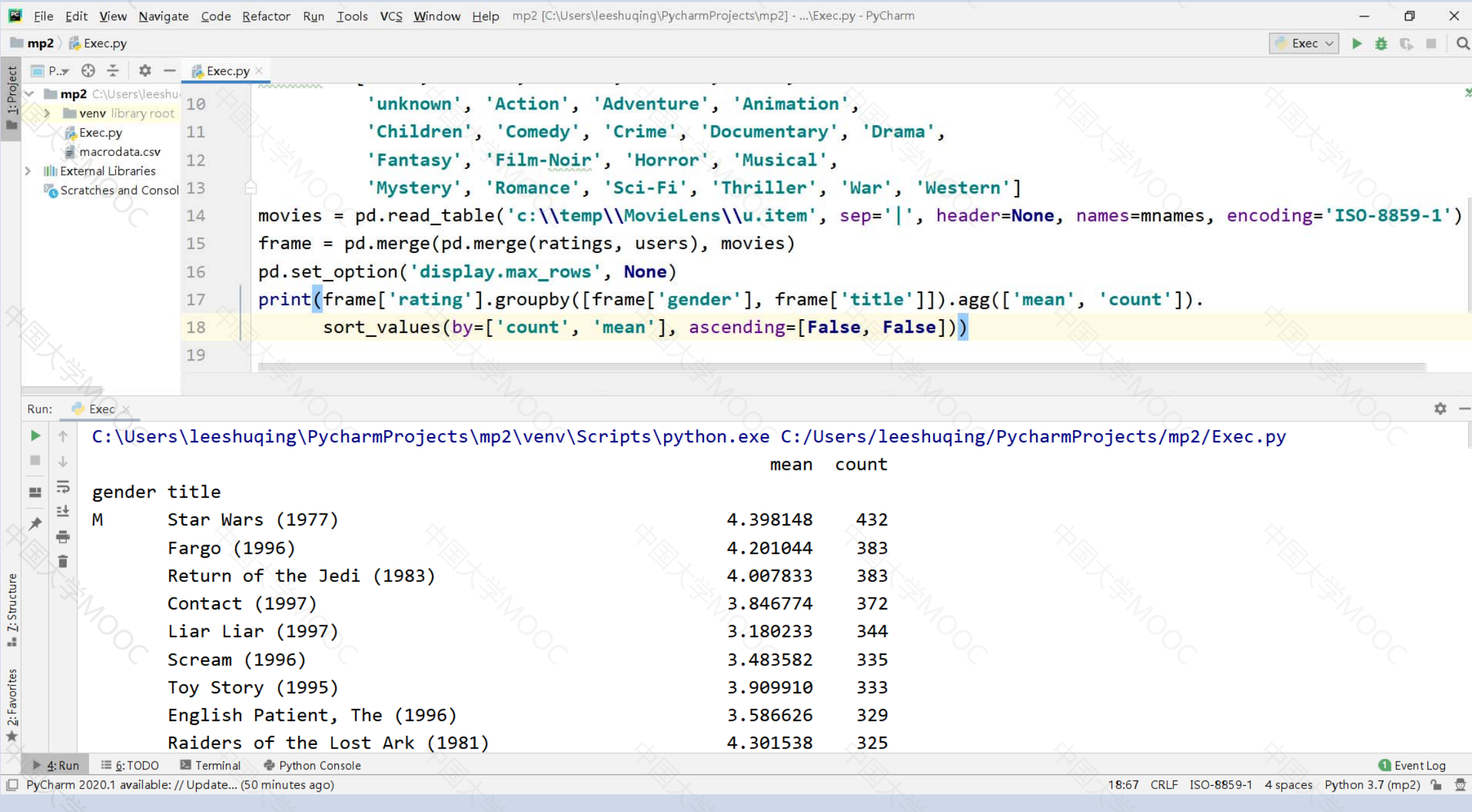
Python大数据分析

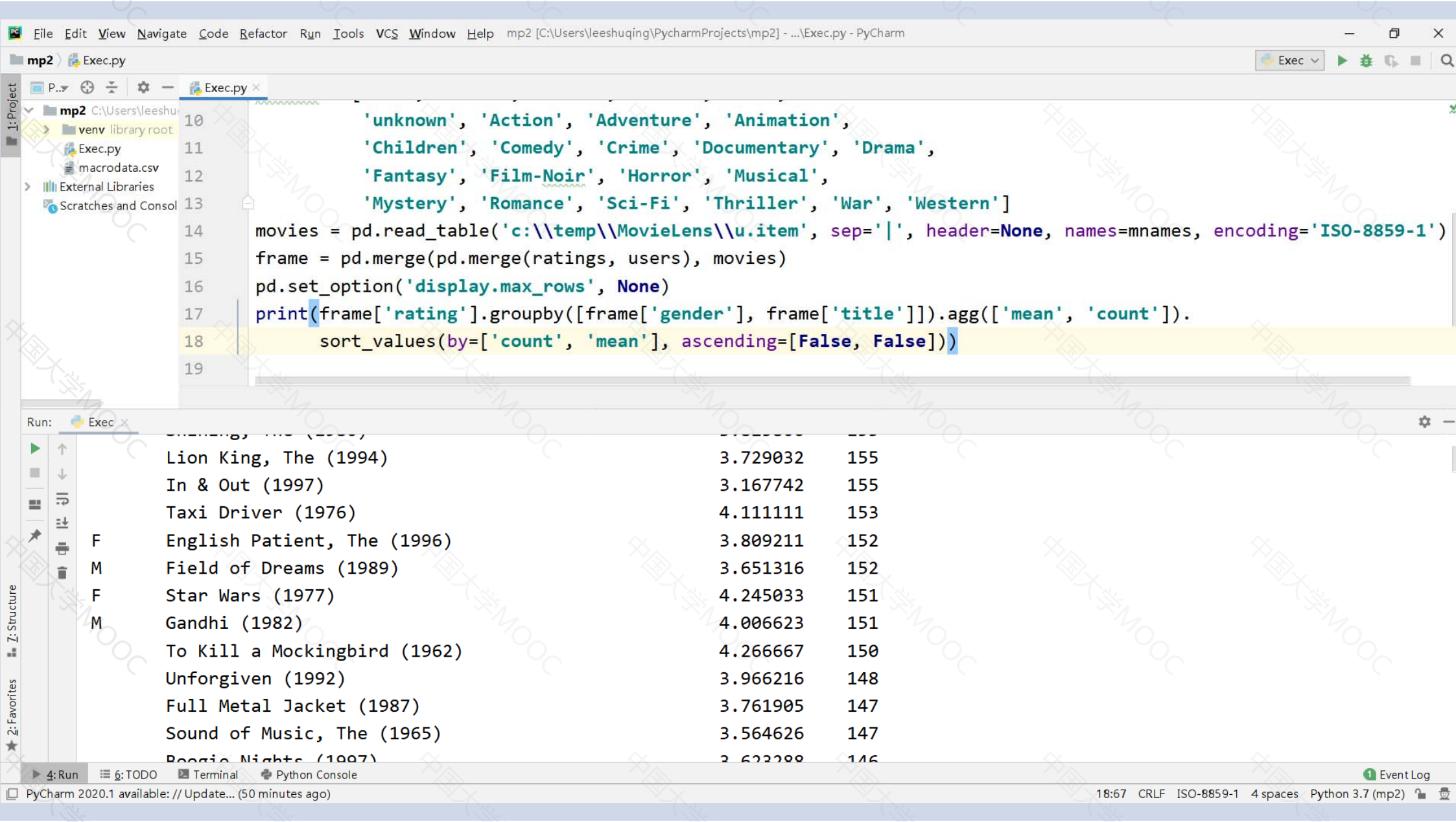
案例2：电影评分数据集的分析2

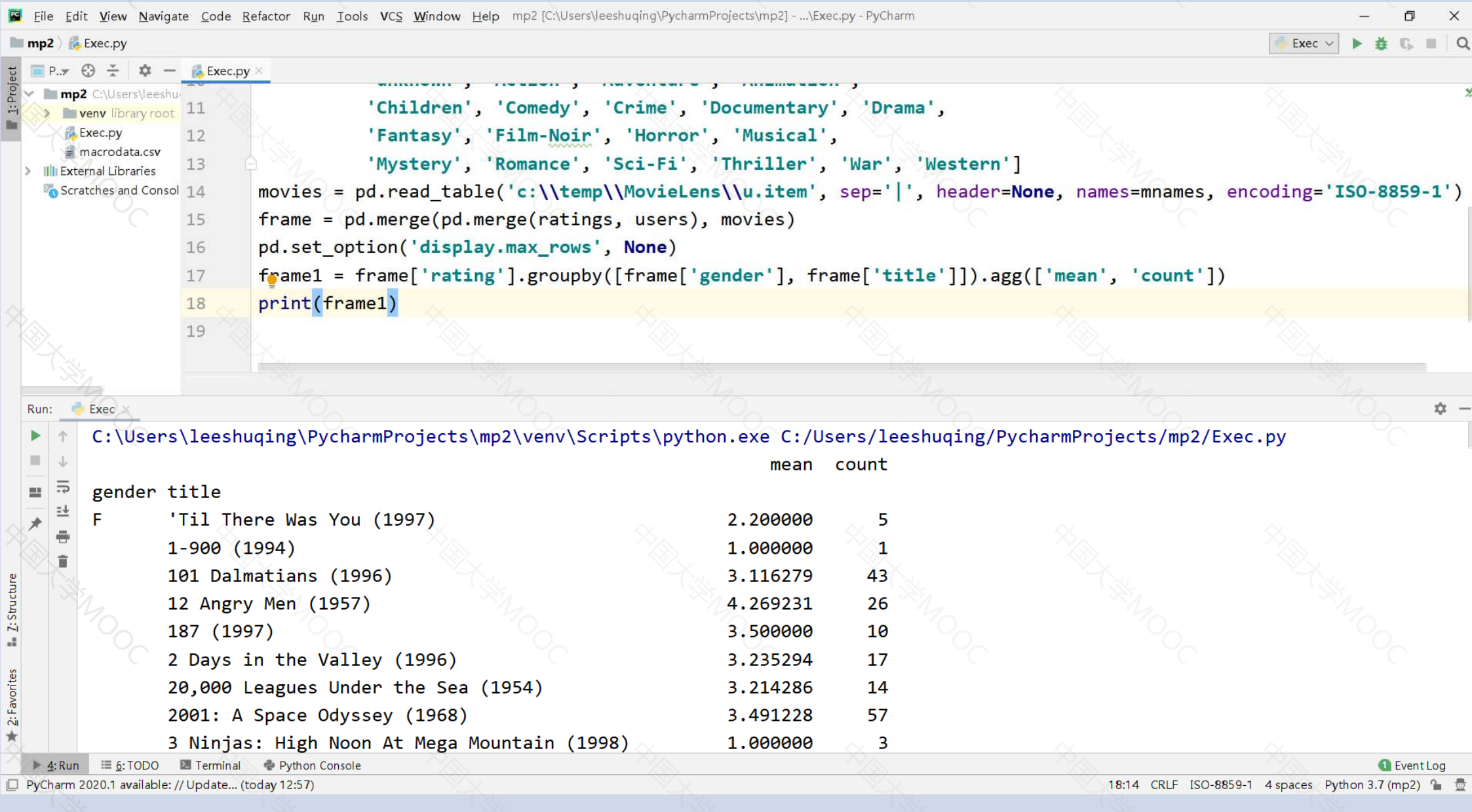

File Edit View Navigate Code Refactor Run Tools VCS Window Help mp2 [C:\Users\leeshuqing\PycharmProjects\mp2] - ...\Exec.py - PyCharm
mp2 Exec.py
Exec
Exec.py
mp2 C:\Users\leeshu
venv library root
Exec.py
macrodata.csv
External Libraries
Scratches and Consol
9 mnames = ['mid', 'title', 'date1', 'date2', 'url',
10 'unknown', 'Action', 'Adventure', 'Animation',
11 'Children', 'Comedy', 'Crime', 'Documentary', 'Drama',
12 'Fantasy', 'Film-Noir', 'Horror', 'Musical',
13 'Mystery', 'Romance', 'Sci-Fi', 'Thriller', 'War', 'Western']
14 movies = pd.read_table('c:\\temp\\MovieLens\\u.item', sep='|', header=None, names=mnames, encoding='ISO-8859-1')
15 frame = pd.merge(pd.merge(ratings, users), movies)
16 print(frame['rating'].groupby([frame['gender'], frame['title']]).mean().sort_values(ascending=False))
17
Run: Exec
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
4: Run
6: TODO
Terminal
Python Console
Event Log
16:102 CRLF GBK 4 spaces Python 3.7 (mp2)

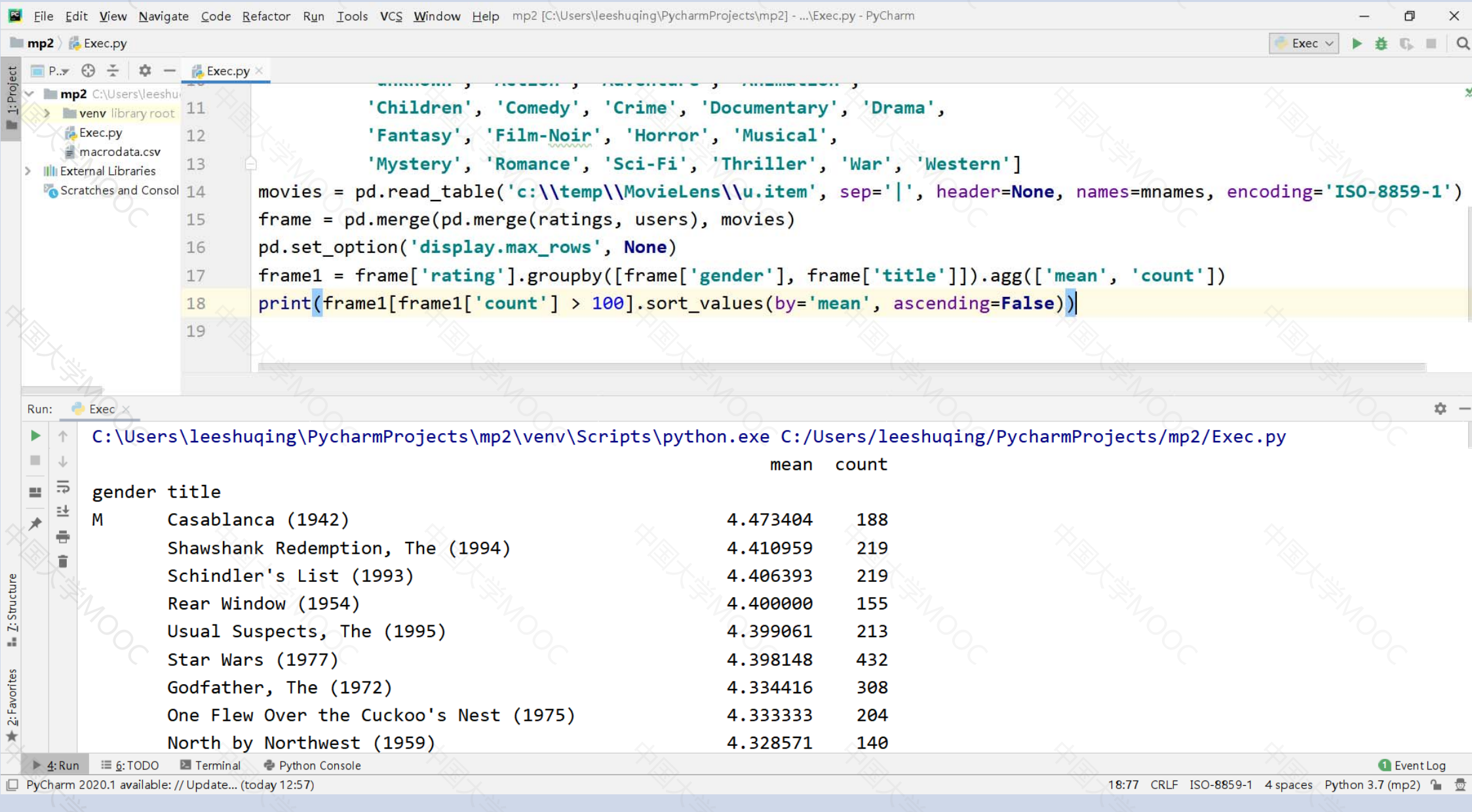

File Edit View Navigate Code Refactor Run Tools VCS Window Help mp2 [C:\Users\leeshuqing\PycharmProjects\mp2] - ...\Exec.py - PyCharm
mp2 Exec.py
Exec
Exec.py
mp2 C:\Users\leeshu
venv library root
Exec.py
macrodata.csv
External Libraries
Scratches and Consol
9 mnames = ['mid', 'title', 'date1', 'date2', 'url',
10 'unknown', 'Action', 'Adventure', 'Animation',
11 'Children', 'Comedy', 'Crime', 'Documentary', 'Drama',
12 'Fantasy', 'Film-Noir', 'Horror', 'Musical',
13 'Mystery', 'Romance', 'Sci-Fi', 'Thriller', 'War', 'Western']
14 movies = pd.read_table('c:\\temp\\MovieLens\\u.item', sep='|', header=None, names=mnames, encoding='ISO-8859-1')
15 frame = pd.merge(pd.merge(ratings, users), movies)
16 print(frame['rating'].groupby([frame['gender'], frame['title']]).agg(['mean', 'count']))
17
Run: Exec
C:\Users\leeshuqing\PycharmProjects\mp2\venv\Scripts\python.exe C:/Users/leeshuqing/PycharmProjects/mp2/Exec.py
mean count
gender title
F 'Til There Was You (1997) 2.200000 5
1-900 (1994) 1.000000 1
101 Dalmatians (1996) 3.116279 43
12 Angry Men (1957) 4.269231 26
187 (1997) 3.500000 10
... ...
M Young Guns (1988) 3.204545 88
Young Guns II (1990) 2.800000 40
Young Poisoner's Handbook, The (1995) 3.218750 32
4: Run
6: TODO
Terminal
Python Console
Event Log
16:89 CRLF GBK 4 spaces Python 3.7 (mp2)


File Edit View Navigate Code Refactor Run Tools VCS Window Help mp2 [C:\Users\leeshuqing\PycharmProjects\mp2] - ...\Exec.py - PyCharm
mp2 Exec.py
Exec
Exec.py
mp2 C:\Users\leeshu
venv library root
Exec.py
macrodata.csv
External Libraries
Scratches and Consol
'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']]).agg(['mean', 'count']).
sort_values(by=['mean', 'count'], ascending=[False, False]))
Run: Exec
C:\Users\leeshuqing\PycharmProjects\mp2\venv\Scripts\python.exe C:/Users/leeshuqing/PycharmProjects/mp2/Exec.py
mean count
gender title
M Star Kid (1997) 5.0 3
F Mina Tannenbaum (1994) 5.0 2
M Leading Man, The (1996) 5.0 2
Prefontaine (1997) 5.0 2
Saint of Fort Washington, The (1993) 5.0 2
... ...
Shadow of Angels (Schatten der Engel) (1976) 1.0 1
Somebody to Love (1994) 1.0 1
Squeeze (1996) 1.0 1
4: Run
6: TODO
Terminal
Python Console
Event Log
17:67 CRLF GBK 4 spaces Python 3.7 (mp2)

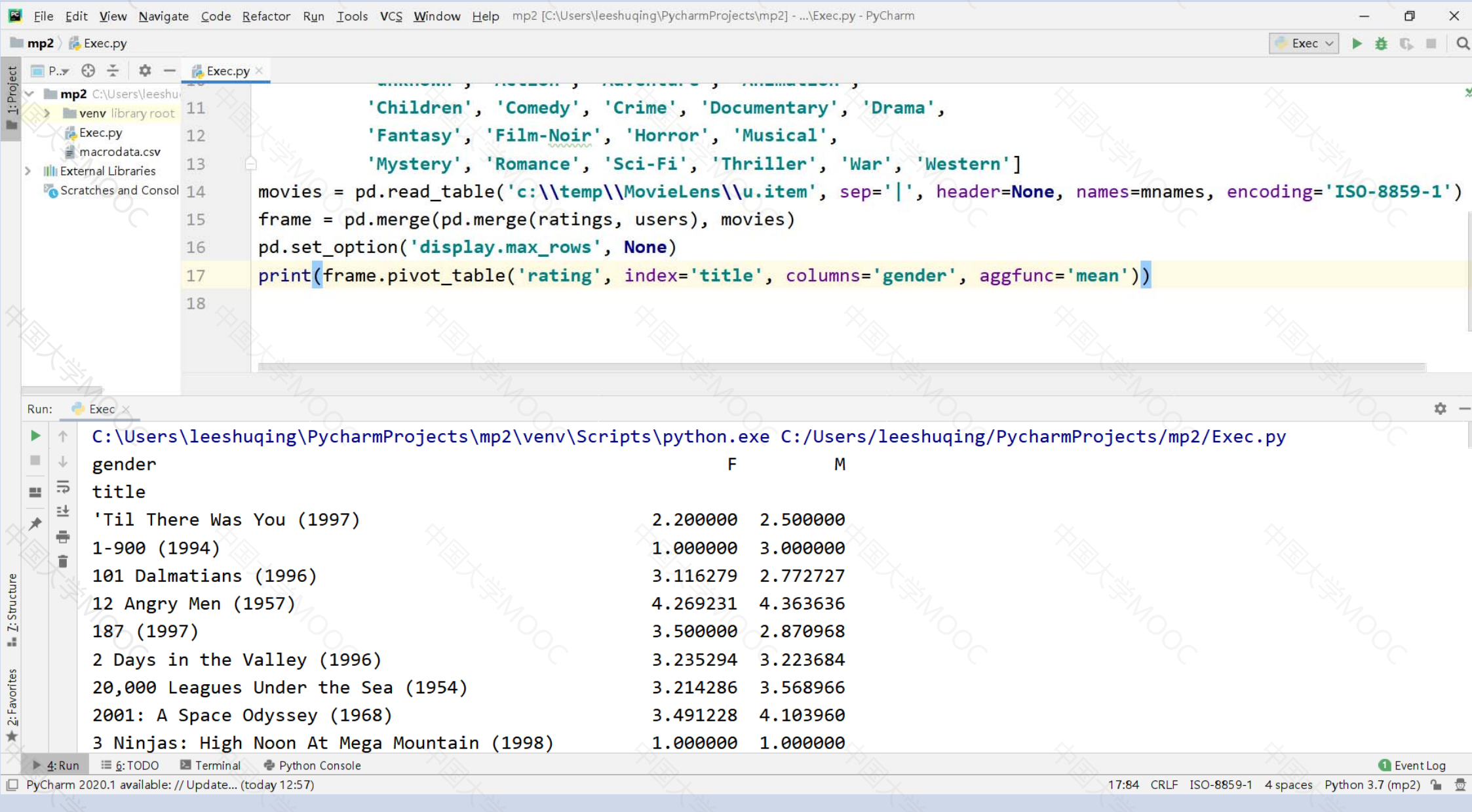

File Edit View Navigate Code Refactor Run Tools VCS Window Help mp2 [C:\Users\leeshuqing\PycharmProjects\mp2] - ...\Exec.py - PyCharm
mp2 Exec.py
Exec
Exec.py
mp2 C:\Users\leeshu
venv library root
Exec.py
macrodata.csv
External Libraries
Scratches and Consol
'unknown', 'Action', 'Adventure', 'Animation',
'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)
pd.set_option('display.max_rows', None)
print(frame['rating'].groupby([frame['gender'], frame['title']]).agg(['mean', 'count']).
sort_values(by=['count', 'mean'], ascending=[False, False]))
Run: Exec
C:\Users\leeshuqing\PycharmProjects\mp2\venv\Scripts\python.exe C:/Users/leeshuqing/PycharmProjects/mp2/Exec.py
mean count
gender title
M Star Wars (1977) 4.398148 432
Fargo (1996) 4.201044 383
Return of the Jedi (1983) 4.007833 383
Contact (1997) 3.846774 372
Liar Liar (1997) 3.180233 344
Scream (1996) 3.483582 335
Toy Story (1995) 3.909910 333
English Patient, The (1996) 3.586626 329
Raiders of the Lost Ark (1981) 4.301538 325
4: Run
6: TODO
Terminal
Python Console
Event Log
PyCharm 2020.1 available: // Update... (50 minutes ago)
18:67 CRLF ISO-8859-1 4 spaces Python 3.7 (mp2)

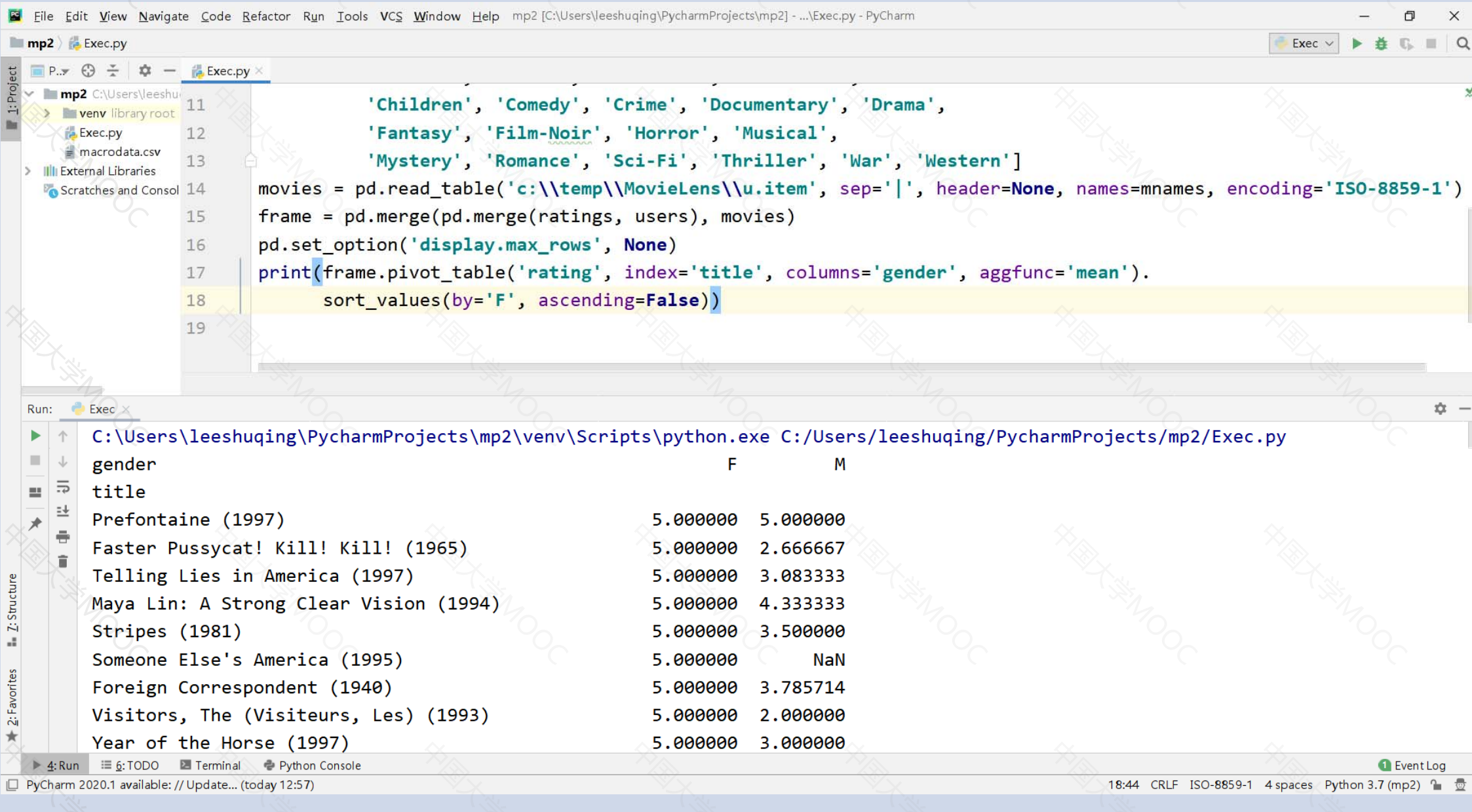

File Edit View Navigate Code Refactor Run Tools VCS Window Help mp2 [C:\Users\leeshuqing\PycharmProjects\mp2] - ...\Exec.py - PyCharm
mp2 Exec.py
Exec
Exec.py
mp2 C:\Users\leeshu
venv library root
Exec.py
macrodata.csv
External Libraries
Scratches and Consol
'unknown', 'Action', 'Adventure', 'Animation',
'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)
pd.set_option('display.max_rows', None)
print(frame['rating'].groupby([frame['gender'], frame['title']]).agg(['mean', 'count']).
sort_values(by=['count', 'mean'], ascending=[False, False]))
Run: Exec
Lion King, The (1994) 3.729032 155
In & Out (1997) 3.167742 155
Taxi Driver (1976) 4.111111 153
F English Patient, The (1996) 3.809211 152
M Field of Dreams (1989) 3.651316 152
F Star Wars (1977) 4.245033 151
M Gandhi (1982) 4.006623 151
To Kill a Mockingbird (1962) 4.266667 150
Unforgiven (1992) 3.966216 148
Full Metal Jacket (1987) 3.761905 147
Sound of Music, The (1965) 3.564626 147
4: Run
6: TODO
Terminal
Python Console
Event Log
PyCharm 2020.1 available: // Update... (50 minutes ago)
18:67 CRLF ISO-8859-1 4 spaces Python 3.7 (mp2)

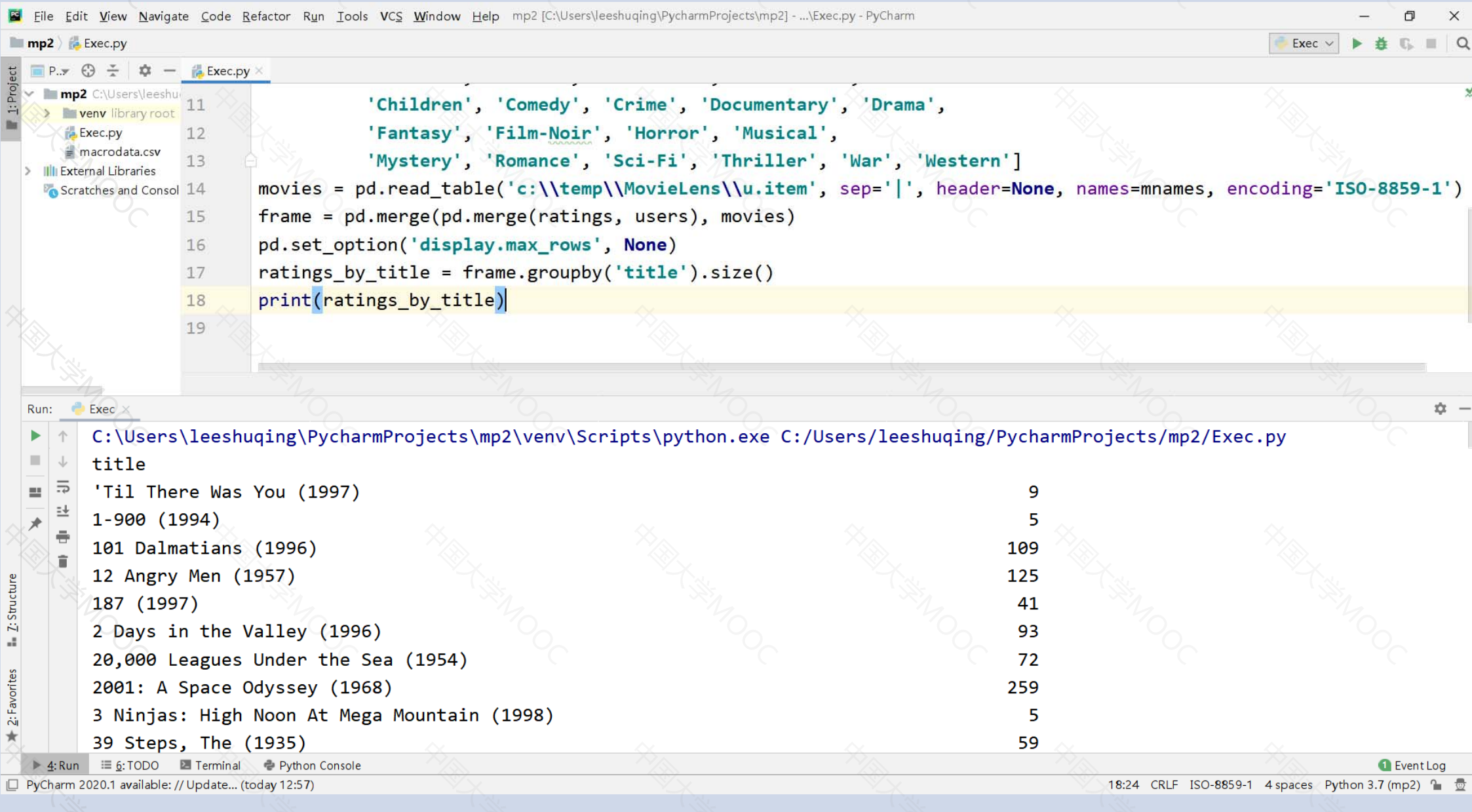

File Edit View Navigate Code Refactor Run Tools VCS Window Help mp2 [C:\Users\leeshuqing\PycharmProjects\mp2] - ...\Exec.py - PyCharm
mp2 Exec.py
Exec
Exec.py
mp2 C:\Users\leeshu
venv library root
Exec.py
macrodata.csv
External Libraries
Scratches and Consol
'Children', 'Comedy', 'Crime', 'Documentary', '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)
pd.set_option('display.max_rows', None)
frame1 = frame['rating'].groupby([frame['gender'], frame['title']]).agg(['mean', 'count'])
print(frame1)
Run: Exec
C:\Users\leeshuqing\PycharmProjects\mp2\venv\Scripts\python.exe C:/Users/leeshuqing/PycharmProjects/mp2/Exec.py
mean count
gender title
F 'Til There Was You (1997) 2.200000 5
1-900 (1994) 1.000000 1
101 Dalmatians (1996) 3.116279 43
12 Angry Men (1957) 4.269231 26
187 (1997) 3.500000 10
2 Days in the Valley (1996) 3.235294 17
20,000 Leagues Under the Sea (1954) 3.214286 14
2001: A Space Odyssey (1968) 3.491228 57
3 Ninjas: High Noon At Mega Mountain (1998) 1.000000 3
4: Run
6: TODO
Terminal
Python Console
Event Log
PyCharm 2020.1 available: // Update... (today 12:57)
18:14 CRLF ISO-8859-1 4 spaces Python 3.7 (mp2)

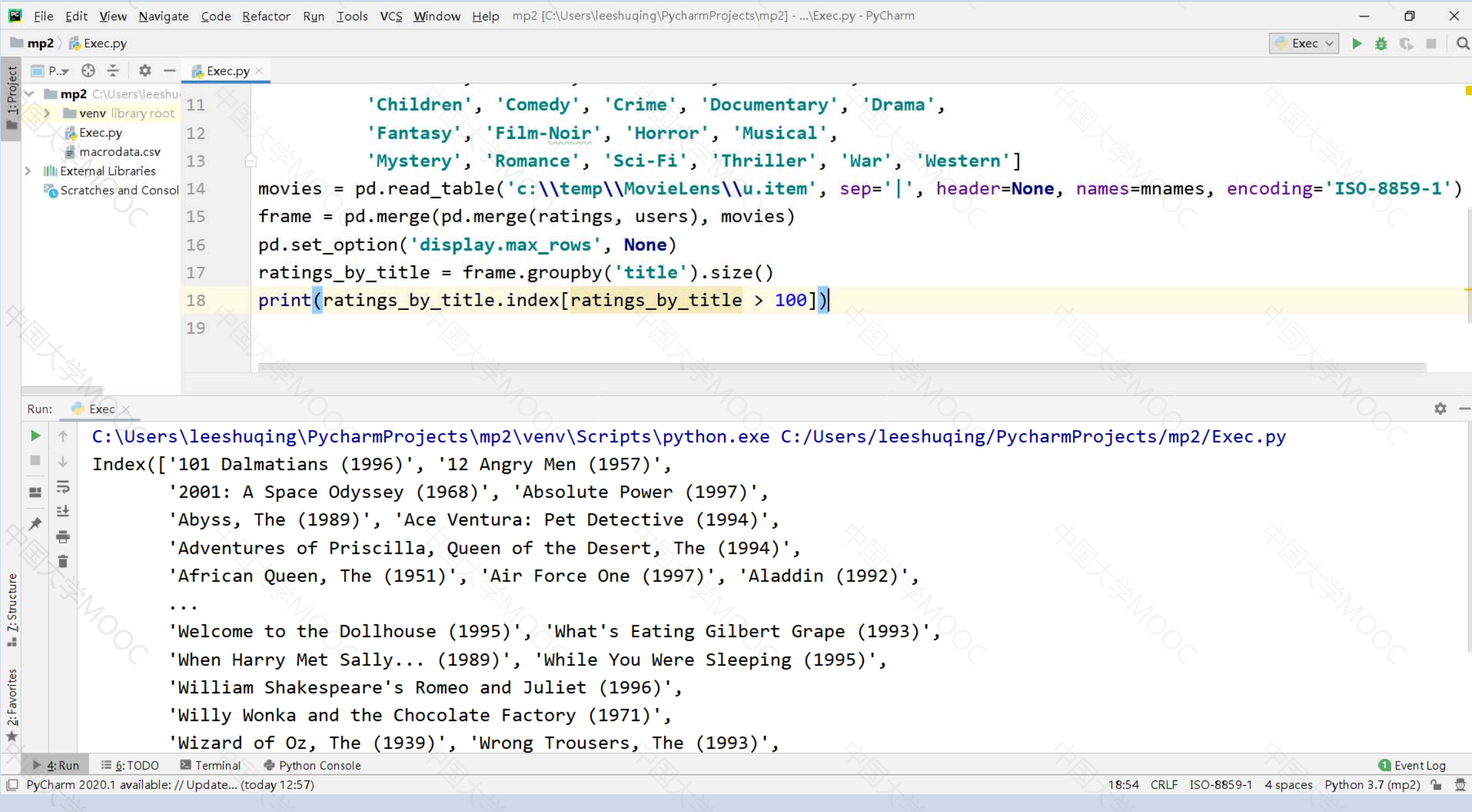

File Edit View Navigate Code Refactor Run Tools VCS Window Help mp2 [C:\Users\leeshuqing\PycharmProjects\mp2] - ...\Exec.py - PyCharm
mp2 Exec.py
Exec
Exec.py
mp2 C:\Users\leeshu
venv library root
Exec.py
macrodata.csv
External Libraries
Scratches and Consol
'Children', 'Comedy', 'Crime', 'Documentary', '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)
pd.set_option('display.max_rows', None)
frame1 = frame['rating'].groupby([frame['gender'], frame['title']]).agg(['mean', 'count'])
print(frame1[frame1['count'] > 100].sort_values(by='mean', ascending=False))
Run: Exec
C:\Users\leeshuqing\PycharmProjects\mp2\venv\Scripts\python.exe C:/Users/leeshuqing/PycharmProjects/mp2/Exec.py
mean count
gender title
M Casablanca (1942) 4.473404 188
Shawshank Redemption, The (1994) 4.410959 219
Schindler's List (1993) 4.406393 219
Rear Window (1954) 4.400000 155
Usual Suspects, The (1995) 4.399061 213
Star Wars (1977) 4.398148 432
Godfather, The (1972) 4.334416 308
One Flew Over the Cuckoo's Nest (1975) 4.333333 204
North by Northwest (1959) 4.328571 140
4: Run
6: TODO
Terminal
Python Console
Event Log
PyCharm 2020.1 available: // Update... (today 12:57)
18:77 CRLF ISO-8859-1 4 spaces Python 3.7 (mp2)

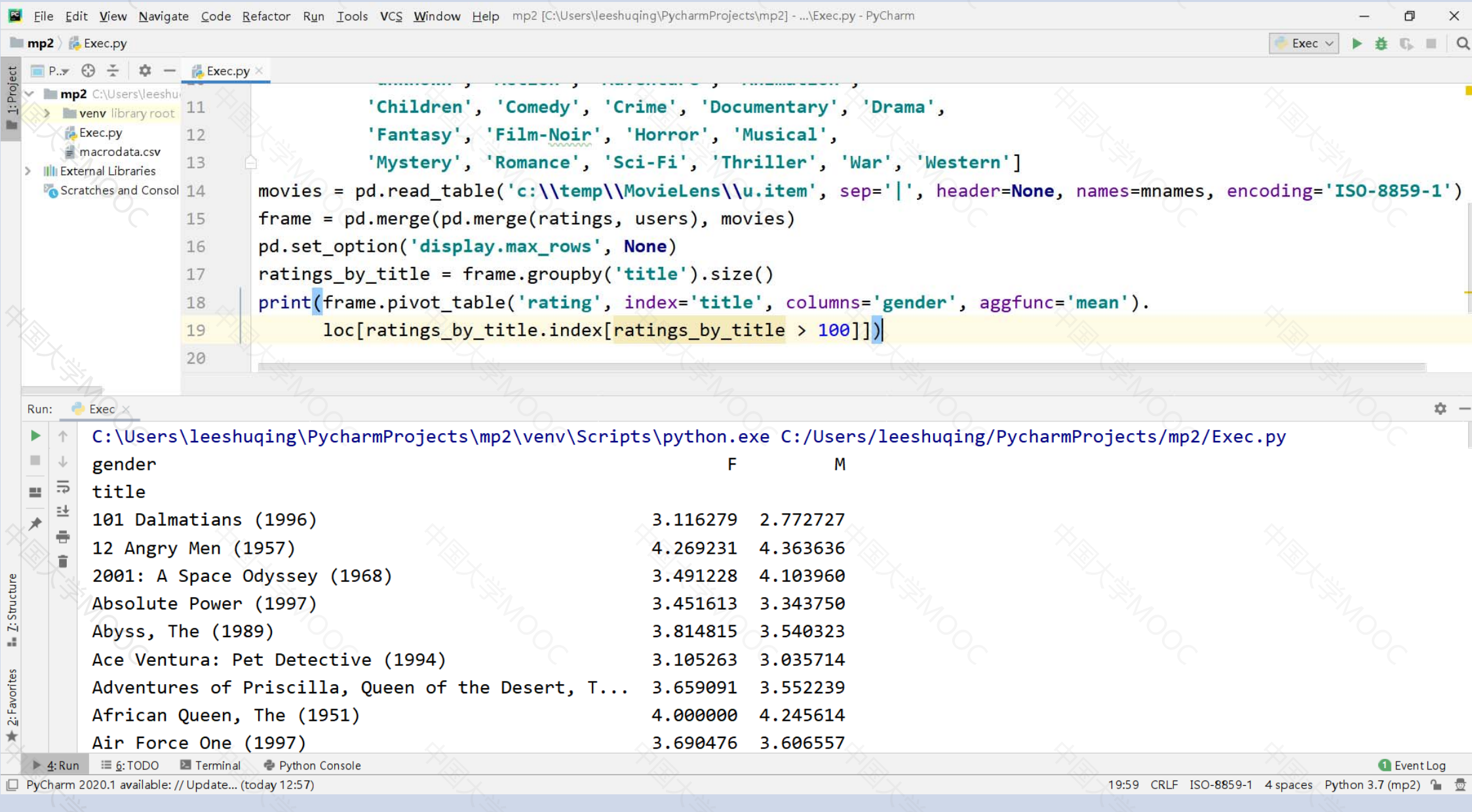

File Edit View Navigate Code Refactor Run Tools VCS Window Help mp2 [C:\Users\leeshuqing\PycharmProjects\mp2] - ...\Exec.py - PyCharm
mp2 Exec.py
Exec
Exec.py
mp2 C:\Users\leeshu
venv library root
Exec.py
macrodata.csv
External Libraries
Scratches and Consol
'Children', 'Comedy', 'Crime', 'Documentary', '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)
pd.set_option('display.max_rows', None)
frame1 = frame['rating'].groupby([frame['gender'], frame['title']]).agg(['mean', 'count'])
print(frame1[frame1['count'] > 100].sort_values(by='mean', ascending=False))
Run: Exec
Silence of the Lambs, The (1991) 4.279310 290
Empire Strikes Back, The (1980) 4.278986 276
F Titanic (1997) 4.278846 104
M To Kill a Mockingbird (1962) 4.266667 150
Manchurian Candidate, The (1962) 4.250000 108
Godfather: Part II, The (1974) 4.248521 169
Dr. Strangelove or: How I Learned to Stop Worry... 4.247059 170
African Queen, The (1951) 4.245614 114
F Star Wars (1977) 4.245033 151
M Titanic (1997) 4.231707 246
Boot, Das (1981) 4.229814 161
Good Will Hunting (1997) 4.223022 139
4: Run
6: TODO
Terminal
Python Console
Event Log
PyCharm 2020.1 available: // Update... (today 12:57)
18:77 CRLF ISO-8859-1 4 spaces Python 3.7 (mp2)

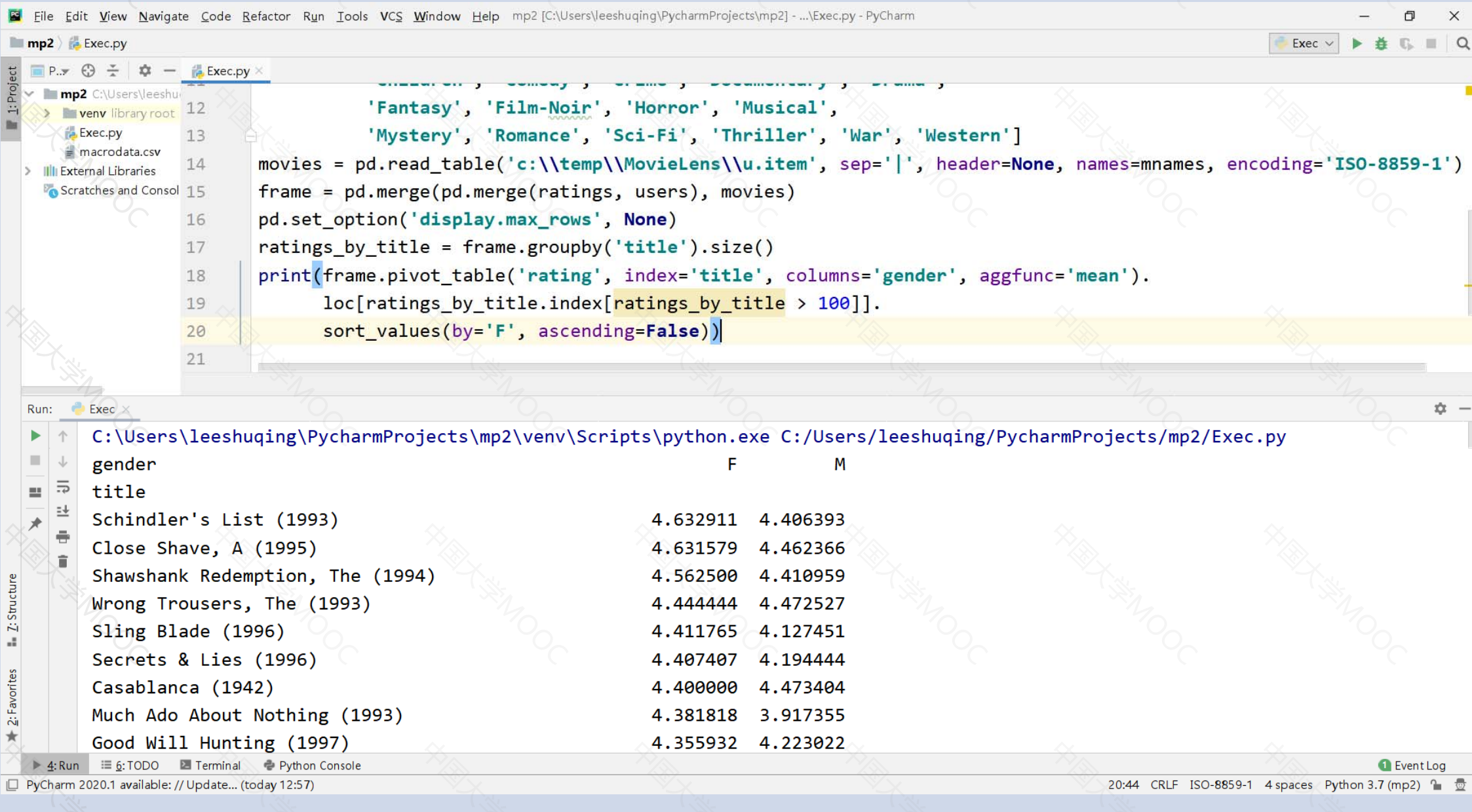

File Edit View Navigate Code Refactor Run Tools VCS Window Help mp2 [C:\Users\leeshuqing\PycharmProjects\mp2] - ...\Exec.py - PyCharm
mp2 Exec.py
Exec
Exec.py
mp2 C:\Users\leeshu
venv library root
Exec.py
macrodata.csv
External Libraries
Scratches and Consol
'Children', 'Comedy', 'Crime', 'Documentary', '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)
pd.set_option('display.max_rows', None)
print(frame.pivot_table('rating', index='title', columns='gender', aggfunc='mean'))
Run: Exec
C:\Users\leeshuqing\PycharmProjects\mp2\venv\Scripts\python.exe C:/Users/leeshuqing/PycharmProjects/mp2/Exec.py
gender F M
title
'Til There Was You (1997) 2.200000 2.500000
1-900 (1994) 1.000000 3.000000
101 Dalmatians (1996) 3.116279 2.772727
12 Angry Men (1957) 4.269231 4.363636
187 (1997) 3.500000 2.870968
2 Days in the Valley (1996) 3.235294 3.223684
20,000 Leagues Under the Sea (1954) 3.214286 3.568966
2001: A Space Odyssey (1968) 3.491228 4.103960
3 Ninjas: High Noon At Mega Mountain (1998) 1.000000 1.000000
4: Run
6: TODO
Terminal
Python Console
Event Log
PyCharm 2020.1 available: // Update... (today 12:57)
17:84 CRLF ISO-8859-1 4 spaces Python 3.7 (mp2)

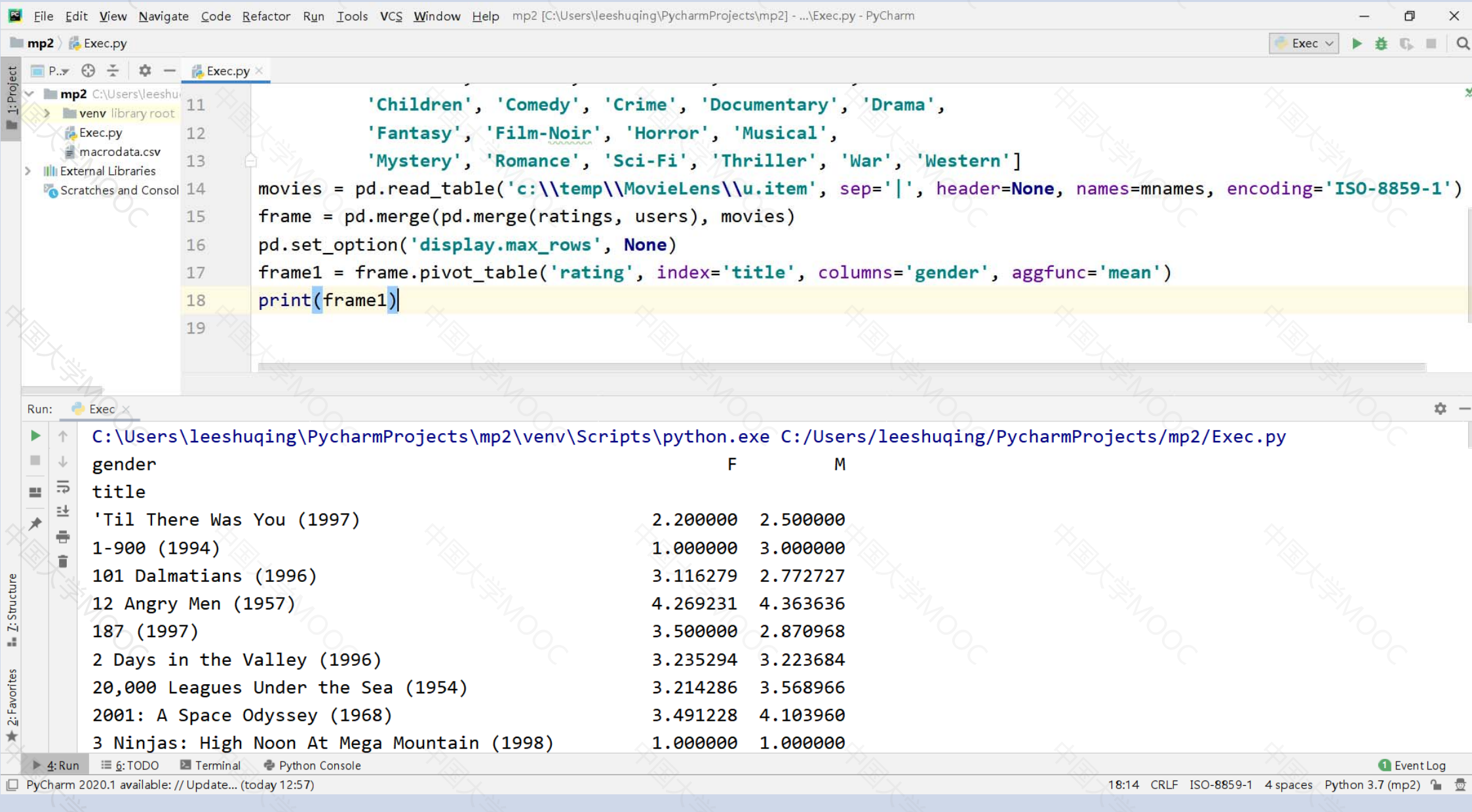

File Edit View Navigate Code Refactor Run Tools VCS Window Help mp2 [C:\Users\leeshuqing\PycharmProjects\mp2] - ...\Exec.py - PyCharm
mp2 Exec.py
Exec
Exec.py
mp2 C:\Users\leeshu
venv library root
Exec.py
macrodata.csv
External Libraries
Scratches and Consol
'Children', 'Comedy', 'Crime', 'Documentary', '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)
pd.set_option('display.max_rows', None)
print(frame.pivot_table('rating', index='title', columns='gender', aggfunc='mean').
sort_values(by='F', ascending=False))
Run: Exec
C:\Users\leeshuqing\PycharmProjects\mp2\venv\Scripts\python.exe C:/Users/leeshuqing/PycharmProjects/mp2/Exec.py
gender F M
title
Prefontaine (1997) 5.000000 5.000000
Faster Pussycat! Kill! Kill! (1965) 5.000000 2.666667
Telling Lies in America (1997) 5.000000 3.083333
Maya Lin: A Strong Clear Vision (1994) 5.000000 4.333333
Stripes (1981) 5.000000 3.500000
Someone Else's America (1995) 5.000000 NaN
Foreign Correspondent (1940) 5.000000 3.785714
Visitors, The (Visiteurs, Les) (1993) 5.000000 2.000000
Year of the Horse (1997) 5.000000 3.000000
4: Run
6: TODO
Terminal
Python Console
Event Log
PyCharm 2020.1 available: // Update... (today 12:57)
18:44 CRLF ISO-8859-1 4 spaces Python 3.7 (mp2)

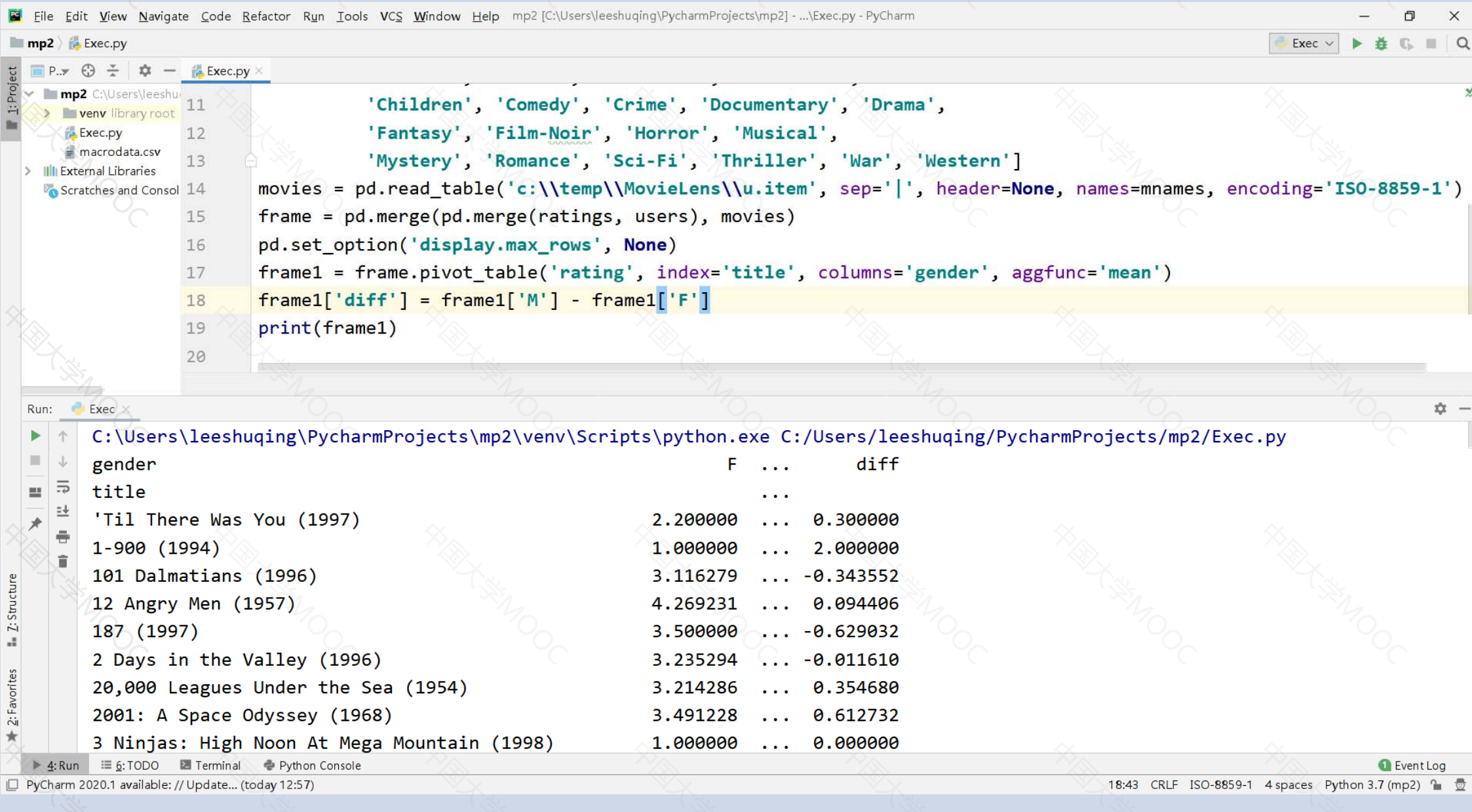

File Edit View Navigate Code Refactor Run Tools VCS Window Help mp2 [C:\Users\leeshuqing\PycharmProjects\mp2] - ...\Exec.py - PyCharm
mp2 Exec.py
Exec
Exec.py
mp2 C:\Users\leeshu
venv library root
Exec.py
macrodata.csv
External Libraries
Scratches and Consol
'Children', 'Comedy', 'Crime', 'Documentary', '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)
pd.set_option('display.max_rows', None)
ratings_by_title = frame.groupby('title').size()
print(ratings_by_title)
Run: Exec
C:\Users\leeshuqing\PycharmProjects\mp2\venv\Scripts\python.exe C:/Users/leeshuqing/PycharmProjects/mp2/Exec.py
title
'Til There Was You (1997) 9
1-900 (1994) 5
101 Dalmatians (1996) 109
12 Angry Men (1957) 125
187 (1997) 41
2 Days in the Valley (1996) 93
20,000 Leagues Under the Sea (1954) 72
2001: A Space Odyssey (1968) 259
3 Ninjas: High Noon At Mega Mountain (1998) 5
39 Steps, The (1935) 59
4: Run
6: TODO
Terminal
Python Console
Event Log
PyCharm 2020.1 available: // Update... (today 12:57)
18:24 CRLF ISO-8859-1 4 spaces Python 3.7 (mp2)


File Edit View Navigate Code Refactor Run Tools VCS Window Help mp2 [C:\Users\leeshuqing\PycharmProjects\mp2] - ...\Exec.py - PyCharm
mp2 Exec.py
Exec
Exec.py
mp2 C:\Users\leeshu
venv library root
Exec.py
macrodata.csv
External Libraries
Scratches and Consol
'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)
pd.set_option('display.max_rows', None)
ratings_by_title = frame.groupby('title').size()
print(ratings_by_title.index[ratings_by_title > 100])
Run: Exec
C:\Users\leeshuqing\PycharmProjects\mp2\venv\Scripts\python.exe C:/Users/leeshuqing/PycharmProjects/mp2/Exec.py
Index(['101 Dalmatians (1996)', '12 Angry Men (1957)',
'2001: A Space Odyssey (1968)', 'Absolute Power (1997)',
'Abyss, The (1989)', 'Ace Ventura: Pet Detective (1994)',
'Adventures of Priscilla, Queen of the Desert, The (1994)',
'African Queen, The (1951)', 'Air Force One (1997)', 'Aladdin (1992)',
...
'Welcome to the Dollhouse (1995)', 'What's Eating Gilbert Grape (1993)',
'When Harry Met Sally... (1989)', 'While You Were Sleeping (1995)',
'William Shakespeare's Romeo and Juliet (1996)',
'Willy Wonka and the Chocolate Factory (1971)',
'Wizard of Oz, The (1939)', 'Wrong Trousers, The (1993)',
4: Run
6: TODO
Terminal
Python Console
Event Log
PyCharm 2020.1 available: // Update... (today 12:57)
18:54 CRLF ISO-8859-1 4 spaces Python 3.7 (mp2)


File Edit View Navigate Code Refactor Run Tools VCS Window Help mp2 [C:\Users\leeshuqing\PycharmProjects\mp2] - ...\Exec.py - PyCharm
mp2 Exec.py
Exec
Exec.py
mp2 C:\Users\leeshu
venv library root
Exec.py
macrodata.csv
External Libraries
Scratches and Consol
'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)
pd.set_option('display.max_rows', None)
ratings_by_title = frame.groupby('title').size()
print(frame.pivot_table('rating', index='title', columns='gender', aggfunc='mean').
loc[ratings_by_title.index[ratings_by_title > 100]])
Run: Exec
C:\Users\leeshuqing\PycharmProjects\mp2\venv\Scripts\python.exe C:/Users/leeshuqing/PycharmProjects/mp2/Exec.py
gender F M
title
101 Dalmatians (1996) 3.116279 2.772727
12 Angry Men (1957) 4.269231 4.363636
2001: A Space Odyssey (1968) 3.491228 4.103960
Absolute Power (1997) 3.451613 3.343750
Abyss, The (1989) 3.814815 3.540323
Ace Ventura: Pet Detective (1994) 3.105263 3.035714
Adventures of Priscilla, Queen of the Desert, T... 3.659091 3.552239
African Queen, The (1951) 4.000000 4.245614
Air Force One (1997) 3.690476 3.606557
4: Run
6: TODO
Terminal
Python Console
Event Log
PyCharm 2020.1 available: // Update... (today 12:57)
19:59 CRLF ISO-8859-1 4 spaces Python 3.7 (mp2)

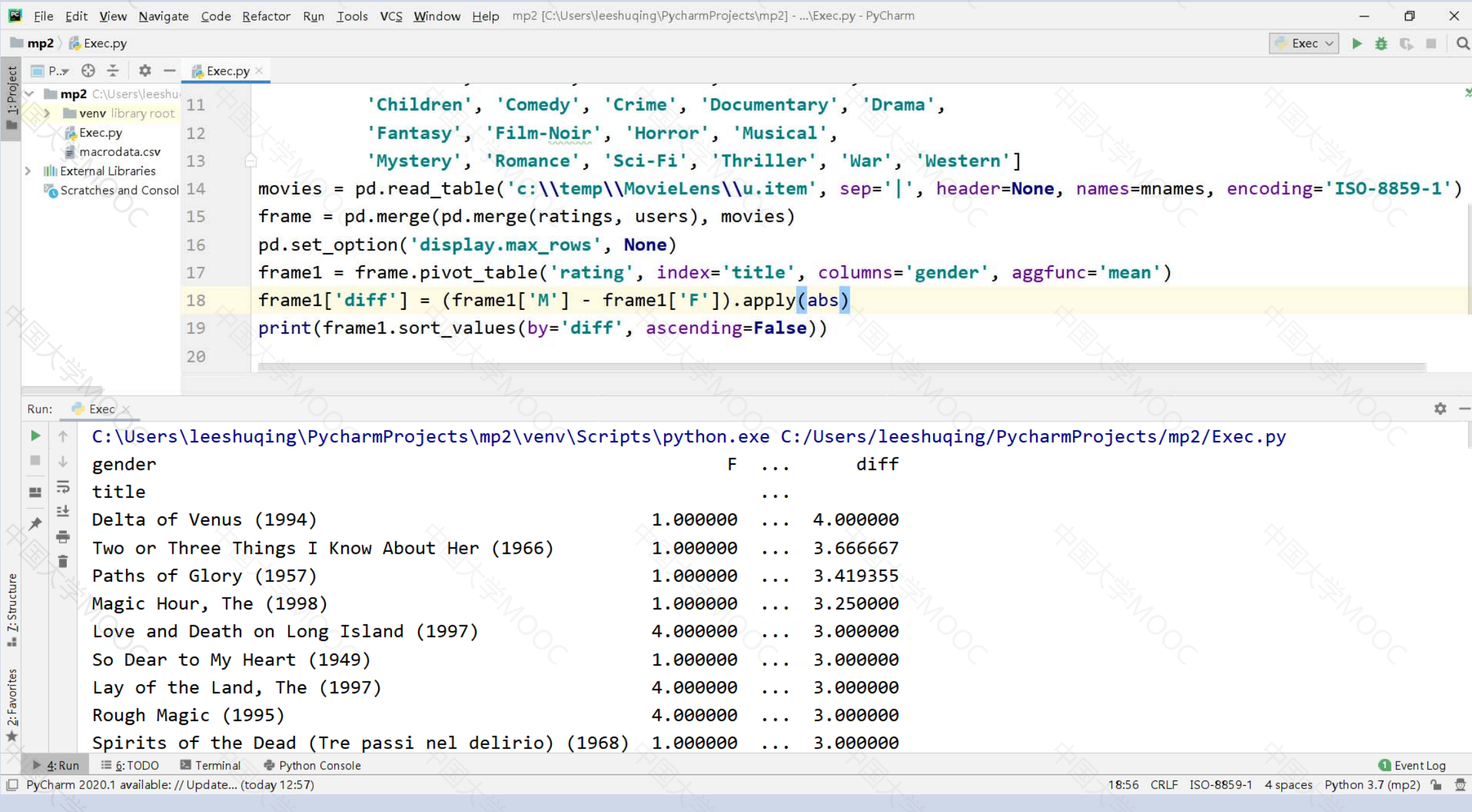

File Edit View Navigate Code Refactor Run Tools VCS Window Help mp2 [C:\Users\leeshuqing\PycharmProjects\mp2] - ...\Exec.py - PyCharm
mp2 Exec.py
Exec
Exec.py
mp2 C:\Users\leeshu
venv library root
Exec.py
macrodata.csv
External Libraries
Scratches and Consol
'Fantasy', 'Film-Noir', 'Horror', 'Musical',
'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)
pd.set_option('display.max_rows', None)
ratings_by_title = frame.groupby('title').size()
print(frame.pivot_table('rating', index='title', columns='gender', aggfunc='mean').
loc[ratings_by_title.index[ratings_by_title > 100]].
sort_values(by='F', ascending=False))
Run: Exec
C:\Users\leeshuqing\PycharmProjects\mp2\venv\Scripts\python.exe C:/Users/leeshuqing/PycharmProjects/mp2/Exec.py
gender F M
title
Schindler's List (1993) 4.632911 4.406393
Close Shave, A (1995) 4.631579 4.462366
Shawshank Redemption, The (1994) 4.562500 4.410959
Wrong Trousers, The (1993) 4.444444 4.472527
Sling Blade (1996) 4.411765 4.127451
Secrets & Lies (1996) 4.407407 4.194444
Casablanca (1942) 4.400000 4.473404
Much Ado About Nothing (1993) 4.381818 3.917355
Good Will Hunting (1997) 4.355932 4.223022
4: Run
6: TODO
Terminal
Python Console
Event Log
PyCharm 2020.1 available: // Update... (today 12:57)
20:44 CRLF ISO-8859-1 4 spaces Python 3.7 (mp2)

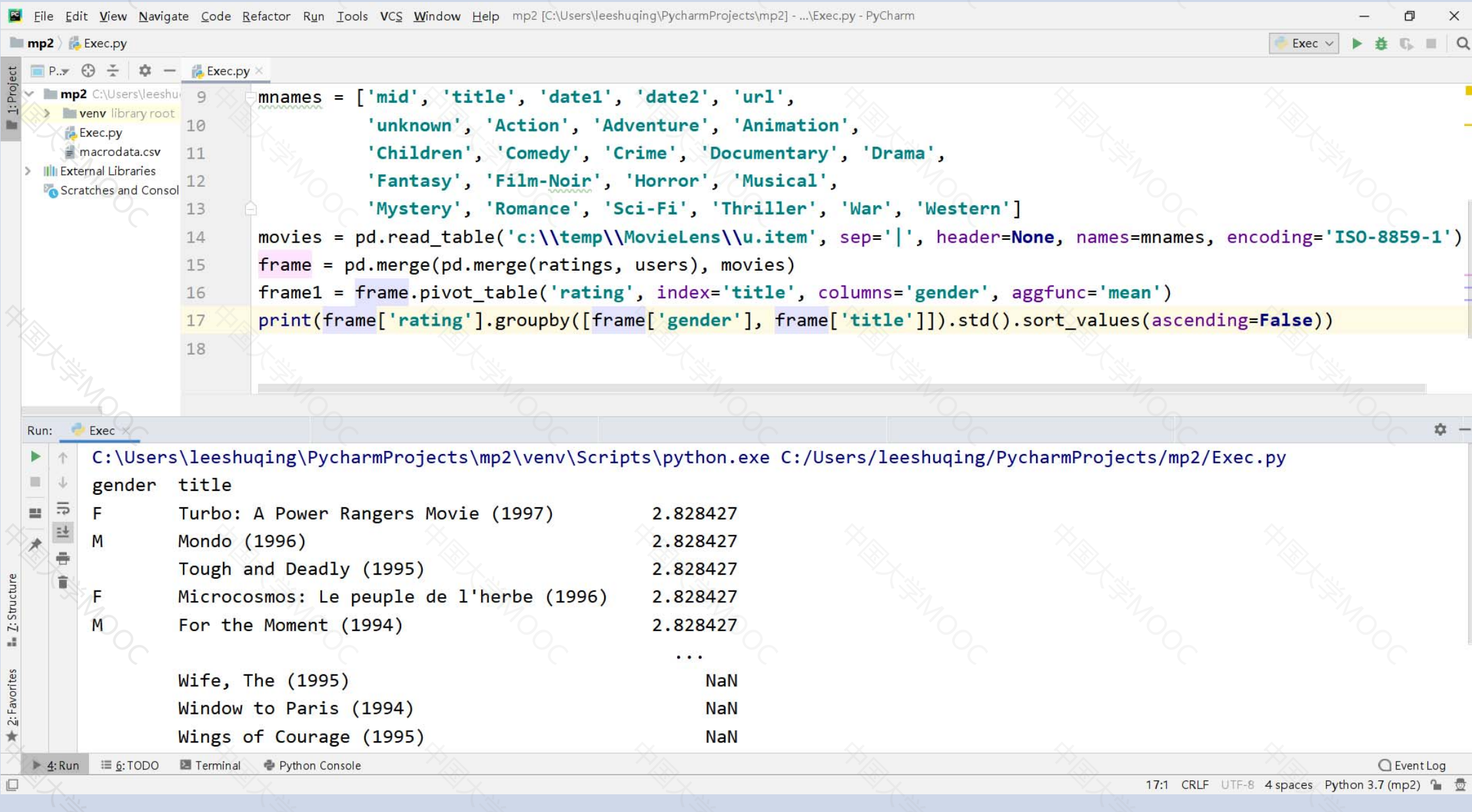

File Edit View Navigate Code Refactor Run Tools VCS Window Help mp2 [C:\Users\leeshuqing\PycharmProjects\mp2] - ...\Exec.py - PyCharm
mp2 Exec.py
Exec
Exec.py
mp2 C:\Users\leeshu
venv library root
Exec.py
macrodata.csv
External Libraries
Scratches and Consol
'Children', 'Comedy', 'Crime', 'Documentary', '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)
pd.set_option('display.max_rows', None)
frame1 = frame.pivot_table('rating', index='title', columns='gender', aggfunc='mean')
print(frame1)
Run: Exec
C:\Users\leeshuqing\PycharmProjects\mp2\venv\Scripts\python.exe C:/Users/leeshuqing/PycharmProjects/mp2/Exec.py
gender F M
title
'Til There Was You (1997) 2.200000 2.500000
1-900 (1994) 1.000000 3.000000
101 Dalmatians (1996) 3.116279 2.772727
12 Angry Men (1957) 4.269231 4.363636
187 (1997) 3.500000 2.870968
2 Days in the Valley (1996) 3.235294 3.223684
20,000 Leagues Under the Sea (1954) 3.214286 3.568966
2001: A Space Odyssey (1968) 3.491228 4.103960
3 Ninjas: High Noon At Mega Mountain (1998) 1.000000 1.000000
4: Run
6: TODO
Terminal
Python Console
Event Log
PyCharm 2020.1 available: // Update... (today 12:57)
18:14 CRLF ISO-8859-1 4 spaces Python 3.7 (mp2)


File Edit View Navigate Code Refactor Run Tools VCS Window Help mp2 [C:\Users\leeshuqing\PycharmProjects\mp2] - ...\Exec.py - PyCharm
mp2 Exec.py
Exec
Exec.py
mp2 C:\Users\leeshu
venv library root
Exec.py
macrodata.csv
External Libraries
Scratches and Consol
'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)
pd.set_option('display.max_rows', None)
frame1 = frame.pivot_table('rating', index='title', columns='gender', aggfunc='mean')
frame1['diff'] = frame1['M'] - frame1['F']
print(frame1)
Run: Exec
C:\Users\leeshuqing\PycharmProjects\mp2\venv\Scripts\python.exe C:/Users/leeshuqing/PycharmProjects/mp2/Exec.py
gender F ... diff
title ...
'Til There Was You (1997) 2.200000 ... 0.300000
1-900 (1994) 1.000000 ... 2.000000
101 Dalmatians (1996) 3.116279 ... -0.343552
12 Angry Men (1957) 4.269231 ... 0.094406
187 (1997) 3.500000 ... -0.629032
2 Days in the Valley (1996) 3.235294 ... -0.011610
20,000 Leagues Under the Sea (1954) 3.214286 ... 0.354680
2001: A Space Odyssey (1968) 3.491228 ... 0.612732
3 Ninjas: High Noon At Mega Mountain (1998) 1.000000 ... 0.000000
4: Run
6: TODO
Terminal
Python Console
Event Log
PyCharm 2020.1 available: // Update... (today 12:57)
18:43 CRLF ISO-8859-1 4 spaces Python 3.7 (mp2)


File Edit View Navigate Code Refactor Run Tools VCS Window Help mp2 [C:\Users\leeshuqing\PycharmProjects\mp2] - ...\Exec.py - PyCharm
mp2 Exec.py
Exec
Exec.py
mp2 C:\Users\leeshu
venv library root
Exec.py
macrodata.csv
External Libraries
Scratches and Consol
'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)
pd.set_option('display.max_rows', None)
frame1 = frame.pivot_table('rating', index='title', columns='gender', aggfunc='mean')
frame1['diff'] = frame1['M'] - frame1['F']
print(frame1.sort_values(by='diff', ascending=False))
Run: Exec
C:\Users\leeshuqing\PycharmProjects\mp2\venv\Scripts\python.exe C:/Users/leeshuqing/PycharmProjects/mp2/Exec.py
gender F ... diff
title ...
Delta of Venus (1994) 1.000000 ... 4.000000
Two or Three Things I Know About Her (1966) 1.000000 ... 3.666667
Paths of Glory (1957) 1.000000 ... 3.419355
Magic Hour, The (1998) 1.000000 ... 3.250000
So Dear to My Heart (1949) 1.000000 ... 3.000000
Spirits of the Dead (Tre passi nel delirio) (1968) 1.000000 ... 3.000000
Little City (1998) 2.000000 ... 3.000000
Aparajito (1956) 1.000000 ... 3.000000
Killer (Bulletproof Heart) (1994) 1.000000 ... 3.000000
4: Run
6: TODO
Terminal
Python Console
Event Log
PyCharm 2020.1 available: // Update... (today 12:57)
19:54 CRLF ISO-8859-1 4 spaces Python 3.7 (mp2)


File Edit View Navigate Code Refactor Run Tools VCS Window Help mp2 [C:\Users\leeshuqing\PycharmProjects\mp2] - ...\Exec.py - PyCharm
mp2 Exec.py
Exec
Exec.py
mp2 C:\Users\leeshu
venv library root
Exec.py
macrodata.csv
External Libraries
Scratches and Consol
'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)
pd.set_option('display.max_rows', None)
frame1 = frame.pivot_table('rating', index='title', columns='gender', aggfunc='mean')
frame1['diff'] = frame1['M'] - frame1['F']
print(frame1.sort_values(by='diff', ascending=True))
Run: Exec
C:\Users\leeshuqing\PycharmProjects\mp2\venv\Scripts\python.exe C:/Users/leeshuqing/PycharmProjects/mp2/Exec.py
gender F ... diff
title ...
Loch Ness (1995) 4.000000 ... -3.000000
Love and Death on Long Island (1997) 4.000000 ... -3.000000
Visitors, The (Visiteurs, Les) (1993) 5.000000 ... -3.000000
Lay of the Land, The (1997) 4.000000 ... -3.000000
Rough Magic (1995) 4.000000 ... -3.000000
Faster Pussycat! Kill! Kill! (1965) 5.000000 ... -2.333333
Chairman of the Board (1998) 4.000000 ... -2.285714
Two Much (1996) 4.000000 ... -2.200000
Simple Wish, A (1997) 3.000000 ... -2.000000
4: Run
6: TODO
Terminal
Python Console
Event Log
PyCharm 2020.1 available: // Update... (today 12:57)
19:53 CRLF ISO-8859-1 4 spaces Python 3.7 (mp2)


File Edit View Navigate Code Refactor Run Tools VCS Window Help mp2 [C:\Users\leeshuqing\PycharmProjects\mp2] - ...\Exec.py - PyCharm
mp2 Exec.py
Exec
Exec.py
mp2 C:\Users\leeshu
venv library root
Exec.py
macrodata.csv
External Libraries
Scratches and Consol
'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)
pd.set_option('display.max_rows', None)
frame1 = frame.pivot_table('rating', index='title', columns='gender', aggfunc='mean')
frame1['diff'] = (frame1['M'] - frame1['F']).apply(abs)
print(frame1.sort_values(by='diff', ascending=False))
Run: Exec
C:\Users\leeshuqing\PycharmProjects\mp2\venv\Scripts\python.exe C:/Users/leeshuqing/PycharmProjects/mp2/Exec.py
gender F ... diff
title ...
Delta of Venus (1994) 1.000000 ... 4.000000
Two or Three Things I Know About Her (1966) 1.000000 ... 3.666667
Paths of Glory (1957) 1.000000 ... 3.419355
Magic Hour, The (1998) 1.000000 ... 3.250000
Love and Death on Long Island (1997) 4.000000 ... 3.000000
So Dear to My Heart (1949) 1.000000 ... 3.000000
Lay of the Land, The (1997) 4.000000 ... 3.000000
Rough Magic (1995) 4.000000 ... 3.000000
Spirits of the Dead (Tre passi nel delirio) (1968) 1.000000 ... 3.000000
4: Run
6: TODO
Terminal
Python Console
Event Log
PyCharm 2020.1 available: // Update... (today 12:57)
18:56 CRLF ISO-8859-1 4 spaces Python 3.7 (mp2)


File Edit View Navigate Code Refactor Run Tools VCS Window Help mp2 [C:\Users\leeshuqing\PycharmProjects\mp2] - ...\Exec.py - PyCharm
mp2 Exec.py
Exec
Exec.py
mp2 C:\Users\leeshu
venv library root
Exec.py
macrodata.csv
External Libraries
Scratches and Consol
9 mnames = ['mid', 'title', 'date1', 'date2', 'url',
10 'unknown', 'Action', 'Adventure', 'Animation',
11 'Children', 'Comedy', 'Crime', 'Documentary', 'Drama',
12 'Fantasy', 'Film-Noir', 'Horror', 'Musical',
13 'Mystery', 'Romance', 'Sci-Fi', 'Thriller', 'War', 'Western']
14 movies = pd.read_table('c:\\temp\\MovieLens\\u.item', sep='|', header=None, names=mnames, encoding='ISO-8859-1')
15 frame = pd.merge(pd.merge(ratings, users), movies)
16 frame1 = frame.pivot_table('rating', index='title', columns='gender', aggfunc='mean')
17 print(frame['rating'].groupby([frame['gender'], frame['title']]).std().sort_values(ascending=False))
18
Run: Exec
C:\Users\leeshuqing\PycharmProjects\mp2\venv\Scripts\python.exe C:/Users/leeshuqing/PycharmProjects/mp2/Exec.py
gender title
F Turbo: A Power Rangers Movie (1997) 2.828427
M Mondo (1996) 2.828427
Tough and Deadly (1995) 2.828427
F Microcosmos: Le peuple de l'herbe (1996) 2.828427
M For the Moment (1994) 2.828427
...
Wife, The (1995) NaN
Window to Paris (1994) NaN
Wings of Courage (1995) NaN
4: Run
6: TODO
Terminal
Python Console
Event Log
17:1 CRLF UTF-8 4 spaces Python 3.7 (mp2)

一次不学多，下次再学