《用 Python 玩转数据》综合项目 1

一、程序功能

基于 MovieLens 100k 数据集中男性女性对电影的评分来判断男性还是女性电影评分的差异性更大。

二、数据来源

数据集下载：

http://files.grouplens.org/datasets/movielens/ml-100k.zip

数据含义：

u.data 表示 100k 条评分记录，每一列的数值含义是：

user id | item id | rating | timestamp

u.user 表示用户的信息，每一列的数值含义是：

user id | age | gender | occupation | zip code

u.item 文件表示电影的相关信息，每一列的数值含义是：

movie/item id | movie title | release date | video release date |IMDb URL | unknown | Action | Adventure | Animation | Children's | Comedy | Crime | Documentary | Drama | Fantasy |Film-Noir | Horror | Musical | Mystery | Romance | Sci-Fi |Thriller | War | Western |

三、分析和代码

基于本数据集可以进行很多分析，例如简单的可基于男生和女生评分均值统计男女各自最喜爱的 10 部电影，结果如下：

```
>>> mean_ratings[:10]
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
39 Steps, The (1935)	4.000000	4.060000

而要判断男性还是女性电影评分的差异性大小则可以利用标准差，标准差越大表示评分离散程度大，即差异性大，反之表示数据越聚集，差异性小。

程序:

```
# API 文档请参考 http://pandas.pydata.org/pandas-docs/stable/

import pandas as pd
import numpy as np

#  读入数据

unames = ['user id', 'age', 'gender', 'occupation', 'zip code']
users = pd.read_table('ml-100k/u.user', sep = '\|', names = unames, engine='python')
rnames = ['user id', 'item id', 'rating', 'timestamp']
ratings = pd.read_table('ml-100k/u.data', sep='\t', names = rnames, engine='python')

#  选择需要的数据列，提高效率

users_df = pd.DataFrame()
users_df['user id'] = users['user id']
users_df['gender'] = users['gender']
ratings_df = pd.DataFrame()
ratings_df['user id'] = ratings['user id']
ratings_df['rating'] = ratings['rating']

#  将数据合并

rating_df = pd.merge(users_df, ratings_df)
gender_table = pd.pivot_table(rating_df, index = ['gender', 'user id'], values = 'rating')
#  利用 pandas 中的数据透视表 pivot_table()函数对数据进行聚合，gender_table 中的数据形式为:
# gender    user id
# F            2                3.709677
#               5                2.874286
# …
# M            898              3.500000
#               899              3.525926
# …
gender_df = pd.DataFrame(gender_table)
```

```
# 分男女
Female_df = gender_df.query("gender == ['F']")
Male_df = gender_df.query("gender == ['M']")
# 按性别计算评分的标准差
Female_std = np.std(Female_df)
Male_std = np.std(Male_df)
print ('Gender', '\nF\t%.6f' % Female_std, '\nM\t%.6f' % Male_std)
```

程序执行结果：

```
Gender
F     0.480358
M    0.429755
```

结论：女生的电影评分差异性更大