德阳科贸职业学院
Deyang Vocational College of Technology and Trade

德阳科贸职业学院

《Python 数据采集与分析专周实训》指导书

二级学院（部）：信息工程学院

适 用 对 象：2021 级大数据技术专业

执 笔 人：唐雪

审 核 人：张玮

制 定 时 间：2023 年 2 月

德阳科贸职业学院教务处制

二〇二三年 二 月

目录

德阳科贸职业学院
Deyang Vocational College of Technology and Trade

电影数据的处理与可视化

目标

1. 掌握 Python 的语法规则和基本的程序设计方法，能够使用 Python 实现对文档的读取、合并功能；

2. 掌握使用 Python 操作文档、处理数据的方法，能够开发简单的逻辑功能模块，实现对数据的处理功能；

3. 掌握数据可视化开发的方法，能够开发简单数据可视化功能。

软硬件环境

硬件：PC 电脑一台

配置：win7 或 win10 系统，内存大于 4G，硬盘 250G 及以上；

软件环境：Pycharm、Python3.7。

实训任务

本案例参考参考 Gred Reda 的分析

http://grouplens.org/datasets/movielens/

对电影评分数据集进行数据处理与可视化，设计要求如下：

一、需求分析

下载公开数据集《电影数据》，进行数据分析。

1. 理解数据，数据如下：

u.data
u.item
u.user

使用上述电影信息，理解并分析数据表字段，表与表之间的关系。

文档中数据如下：

	A	B	C	D
1	196	242	3	881250949
2	186	302	3	891717742
3	22	377	1	878887116
4	244	51	2	880606923
5	166	346	1	886397596
6	298	474	4	884182806
7	115	265	2	881171488
8	253	465	5	891628467
9	305	451	3	886324817
10	6	86	3	883603013
11	62	257	2	879372434
12	286	1014	5	879781125
13	200	222	5	876042340
14	210	40	3	891035994
15	224	29	3	888104457
16	303	785	3	879485318
17	122	387	5	879270459
18	194	274	2	879539794
19	291	1042	4	874834944
20	234	1184	2	892079237
21	119	392	4	886176814
22	167	486	4	892738452
23	299	144	4	877881320
24	291	118	2	874833878
25	308	1	4	887736532
26	95	546	2	879196566

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
17|30|M|programmer|06355
18|35|F|other|37212
19|40|M|librarian|02138
20|42|F|homemaker|95660
21|26|M|writer|30068
22|25|M|writer|40206
23|30|F|artist|48197
24|21|F|artist|94533
25|39|M|engineer|55107

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|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|0|1|0|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|1|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|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+qiao+(1995)|0|0
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|0|1|0|0|0|0|0|0|0|1|0|0|
8|Babe (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Babe%20(1995)|0|0|0|0|0|1|1|0|0|1|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|0|0|1|0|0|0|0|0|0|0|
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|0|1|0|0|0|0|0|0|0|0|1|0
11|Seven (Se7en) (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Se7en%20(1995)|0|0|0|0|0|0|0|1|0|0|0|0|0|0|0|0|0|1|0|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|0|1|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|0|1|0|0|0|0|0|0|0|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|0|1|0|0|0|0
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|0|0|1|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|0|1|0|0|0|0|0|0|0|
17|From Dusk Till Dawn (1996)|05-Feb-1996||http://us.imdb.com/M/title-exact?From%20Dusk%20Till%20Dawn%20(1996)|0|0|1|0|0|0|0|1|1|0|0|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|0|1|0|0|0|0|0|0|0|0|
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|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|0|0|1|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|0|0|1|0|0|0
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|1|0

2.将数据表读取并设置表头：

	user_id	age	sex	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

	user_id	movie_id	rating	unix-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

	movie_id	...	unix-imdb_url
0	1	...	http://us.imdb.com/M/title-exact?Toy%20Story%2...
1	2	...	http://us.imdb.com/M/title-exact?GoldenEye%20(...
2	3	...	http://us.imdb.com/M/title-exact?Four%20Rooms%...
3	4	...	http://us.imdb.com/M/title-exact?Get%20Shorty%...
4	5	...	http://us.imdb.com/M/title-exact?Copycat%20(1995)

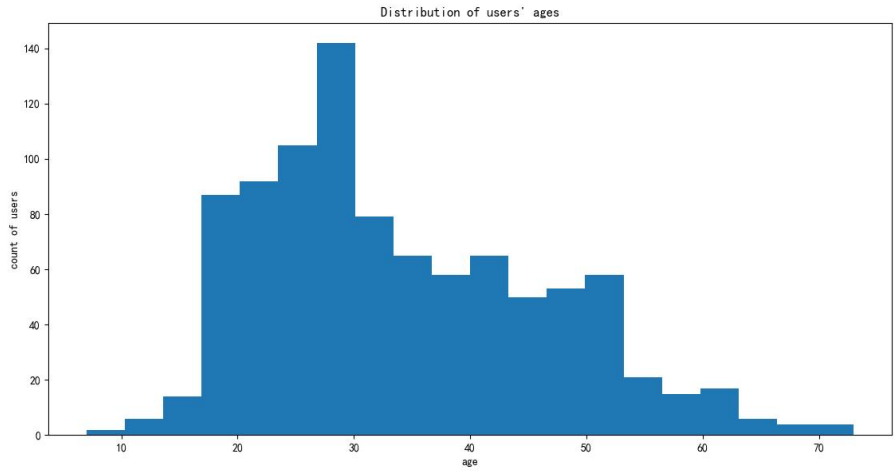
3.合并上述三张数据表

	movie_id	title	release_date	...	sex	occupation	zip_code
0	1	Toy Story (1995)	01-Jan-1995	...	M	retired	95076
1	4	Get Shorty (1995)	01-Jan-1995	...	M	retired	95076
2	5	Copycat (1995)	01-Jan-1995	...	M	retired	95076
3	7	Twelve Monkeys (1995)	01-Jan-1995	...	M	retired	95076
4	8	Babe (1995)	01-Jan-1995	...	M	retired	95076

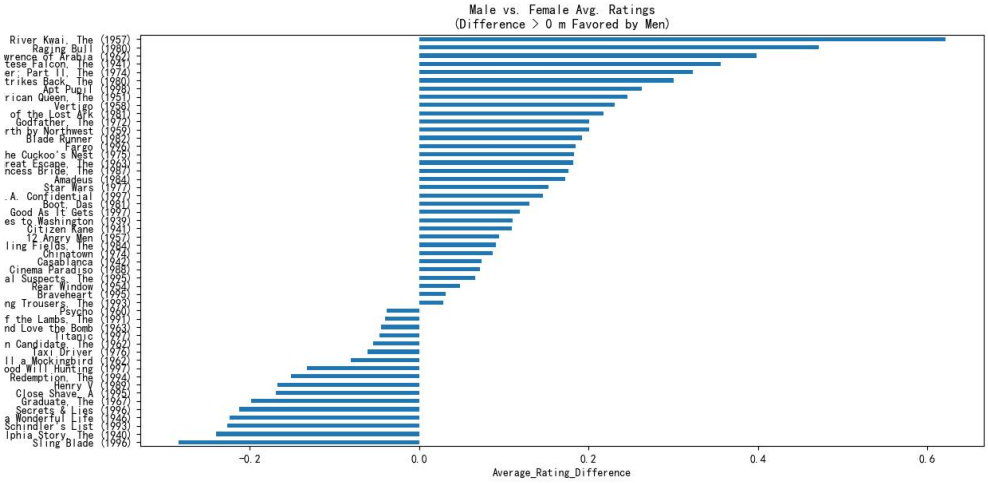
4.按照四个数据分析指标进行电影数据评分的分析任务

title	
Star Wars (1977)	583
Contact (1997)	509
Fargo (1996)	508
Return of the Jedi (1983)	507
Liar Liar (1997)	485
English Patient, The (1996)	481
Scream (1996)	478
Toy Story (1995)	452
Air Force One (1997)	431
Independence Day (ID4) (1996)	429
Raiders of the Lost Ark (1981)	420
Godfather, The (1972)	413
Pulp Fiction (1994)	394
Twelve Monkeys (1995)	392
Silence of the Lambs, The (1991)	390
Jerry Maguire (1996)	384
Chasing Amy (1997)	379

5.根据上一步中统计的数据，开发可视化功能，使用柱状图的方式展示评分与年龄的关系，如下图：



6.根据上一步中统计的数据，开发可视化功能，使用柱状图的方式展示评分与性别的关系，如下图：



二、实训步骤

1. 分析功能需求，设计功能模块；

2. 创建工程，安装工程使用的库；

3. 分析理解数据和业务指标；

4. 将数据进行预处理和特征提取；

5. 将得到的数据进行可视化展示