2024年甘肃省职业院校技能大赛
中职学生组电子与信息大类大数据应用与服务赛项
样题三

模块一 数据库系统运维

环境说明：

编号	主机名	类型	用户	密码
1	database	MySQL数据库	root	123456
2	desktop1	桌面1	/	/
3	desktop2	桌面2	/	/
4	desktop3	桌面3	/	/

补充说明：

①mysql服务器地址 database:3306

②desktop1、desktop2、desktop3完全一致，各位选手可以各使用其中一台桌面主机进行操作

③可以在desktop1/desktop2/desktop3主机上通过如下命令连接到MySQL数据库：mysql -h database -p123456

④也可以直接切换到database主机上操作MySQL数据库

模块一涉及到的数据库数据表信息如下：

数据库-数据表

数据库	数据表	备注
MovieDB	movies	电影表
	ratings	评分表
	users	用户表

movies表

表名	列名	数据类型	备注
movies	movie_id	int	电影ID
	title	varchar	电影标题
	genres	varchar	类型，类型是管道分离的，可从以下类型中选择： Action 行动 Adventure 冒险 Animation 动画

			Children's 儿童 Comedy 喜剧 Crime 犯罪 Documentary 纪录片 Drama 戏剧 Fantasy 幻想 Film-Noir 胶片噪声 Horror 恐怖 Musical 音乐剧 Mystery 神秘 Romance 浪漫 Sci-Fi 科幻 Thriller 惊悚片 War 战争 Western 西部

ratings表

表名	列名	数据类型	备注
ratings	id	int	主键ID
	user_id	int	用户ID，范围在1到6040之间
	movie_id	int	电影ID，范围在1到3952之间
	rating	int	评分，评分等级为1-5
	timestamp	varchar	时间戳

users表

表名	列名	数据类型	备注
users	user_id	int	用户ID
	gender	varchar	性别，用“M”表示男性，用“F”表示女性
	age	int	年龄，可从以下范围中选择： 1:“Under 18” 18:“18-24” 25:“25-34” 35:“35-44” 45:“45-49” 50:“50-55” 56:“56+”

	occupation	int	职业，可从以下选项中选择： 0:“其他”或未指定 1:“学术/教育家” 2:“艺术家” 3:“文书/行政人员” 4:“大学/研究生” 5:“客户服务” 6:“医生/医疗保健” 7:“执行/管理” 8:“农民” 9:“家庭主妇” 10:“K-12学生” 11:“律师” 12:“程序员” 13:“退休” 14:“销售/营销” 15:“科学家” 16:“个体经营者” 17:“技术员/工程师” 18:“商人/工匠” 19:“失业” 20:“作家”
	zip_code	varchar	邮编

基本要求：

1、本模块为技能实操，满分25分。

2、禁止携带参考资料入场。

任务一：数据库系统之用户与权限管理

【任务要求】

本环节需要使用MySQL数据库系统完成关于用户管理与权限管理的操作。

【任务需求背景】

MySQL是一个多用户数据库，具有功能强大的访问控制系统，可以为不同用

户指定不同权限。root用户是超级管理员，拥有所有权限，包括创建用户、删除用户和修改用户密码等管理权限。

为了实际项目的需要，可以定义不同的用户角色，并为不同的角色赋予不同的操作权限。当用户访问数据库时，需要先验证该用户是否为合法用户，再约束该用户只能在被赋予的权限范围内操作。

【具体任务】

1、为本地主机数据库创建一个名为competitor的用户，密码为cpttor123，将完整命令及结果截图粘贴到对应答题报告中；

2、查看用户，确认有刚才创建的competitor用户，将完整命令及结果截图粘贴到对应答题报告中；

3、将用户名competitor修改为competitor01，将完整命令及结果截图粘贴到对应答题报告中；

4、使用新用户competitor01登录MySQL数据库，将完整命令及结果截图粘贴到对应答题报告中；

5、授予用户competitor01对MovieDB数据库中所有表的所有权限，将完整命令及结果截图粘贴到对应答题报告中；

6、使用新用户competitor01登录MySQL数据库，然后查看数据库，将完整命令及结果截图粘贴到对应答题报告中；

7、撤销用户competitor01对MovieDB数据库中所有表的所有权限，将完整命令及结果截图粘贴到对应答题报告中；

8、删除competitor01的用户，将完整命令及结果截图粘贴到对应答题报告中；

任务二：数据库系统之数据表管理

【任务要求】

本环节需要使用MySQL数据库系统完成关于电影信息的建库、建表、数据的导入、数据表的管理等操作。

【任务需求背景】

在今天的数字娱乐时代，电影产业扮演着至关重要的角色，为观众提供了无尽的娱乐选择。了解观众对电影的评分和喜好是制作和推荐电影的关键因素之一。因此，我们决定建立一个电影评分信息管理系统，以更好地了解和分析电影评分数据，提供更精准的电影推荐服务，并深入了解市场趋势和用户口味。

【具体任务】

1、在MySQL数据库的MovieDB库中，创建一个名为users的数据表，包含的字段见上面的数据表说明，指定user_id字段为主键，该字段非空，数据库引擎为InnoDB,默认字符集为utf8。将完整命令及运行结果截图粘贴到对应答题报告中；

2、查看刚才创建的users表结构，将完整命令及结果截图粘贴到对应答题报告中；

3、执行database主机/usr/local/src目录下的users.sql文件，将数据导入到刚才创建的users表中，将完整命令及结果截图粘贴到对应答题报告中；

4、使用SQL命令查看users表中前15条数据（查询结果只显示前15条数据），将完整命令及结果截图粘贴到对应答题报告中；

5、使用SQL命令查看users表中第1001至第1010条数据（查询结果只显示第1001至第1010条数据），将完整命令及结果截图粘贴到对应答题报告中；

6、使用SQL命令复制users表的表结构到new_users表中，将完整命令及结果截图粘贴到对应答题报告中；

7、使用SQL命令复制users表的表结构及表中第666至第888条数据到new_users_new表中，将完整命令及结果截图粘贴到对应答题报告中；

8、使用SQL命令修改new_users_new表中occupation列的列名为work，将完整命令及结果截图粘贴到对应答题报告中；

9、使用SQL命令修改new_users_new表中work字段的类型和长度为varchar(255)，将完整命令及结果截图粘贴到对应答题报告中；

10、使用SQL命令删除new_users_new表中的zip_code字段，将完整命令及结果截图粘贴到对应答题报告中；

11、使用SQL命令给new_users_new表增加一个字段address（代表家庭地址），字段类型应符合实际意义，将完整命令及结果截图粘贴到对应答题报告中；

12、使用SQL命令删除new_users表和new_users_new表，将完整命令及结果截图粘贴到对应答题报告中；

任务三：数据库系统之数据管理

【任务要求】

本环节需要使用SQL语句对数据表的数据进行查询和统计。

【任务需求背景】

SQL作为一种全球通用的语言，任何人都可以学习使用。虽然看起来很复杂，除开特定数据库系统专用的SQL命令，其它基本上不需要任何事先的知识，而且命令通常比较少。SQL能够快速的查询和统计大量数据，发现数据的趋势和数据之间的关系。SQL是一种与数据库打交道的标准语言，熟练地使用SQL可以确保每个使用数据库的人都会使用相同的命令，使得开发人员更容易创建与多个数据库一起工作的应用程序。

【具体任务】

1、使用SQL语句查询users表中职业为程序员的女性用户。将完整SQL语句和运行结果的后5条数据以及总数据行数截图粘贴到对应答题报告中；

2、使用SQL语句查询users表中年龄大于等于18岁且小于45岁的用户。将完整SQL语句和运行结果的后5条数据以及总数据行数截图粘贴到对应答题报告中；

3、使用SQL语句查询movies表中电影类型包含冒险和恐怖的电影。将完整SQL语句和运行结果以及总数据行数截图粘贴到对应答题报告中；

4、使用SQL语句查询被user_id为100的用户评分过的电影，输出用户id、电影id、电影标题、评分、时间戳。将完整SQL语句和运行结果的后5条数据以及总数据行数截图粘贴到对应答题报告中；

5、使用SQL语句查询users表中user_id的最大值和最小值。将完整SQL语句及运行结果截图粘贴到对应答题报告中；

6、使用SQL语句统计ratings表中每个用户所评分电影的平均分，输出用户id及他评论电影的平均分。将完整SQL语句和运行结果的后5条数据以及总数据行数截图粘贴到对应答题报告中；

7、使用SQL语句统计ratings表中movie_id大于等于2500且小于等于2510的电影的最高评分、最低评分、和平均评分，输出格式需包含movie_id。将完整SQL语句及运行结果截图粘贴到对应答题报告中；

8、使用SQL语句查询邮编为55117的用户们对标题为Toy Story (1995)的电影的评分，输出用户id、用户年龄、电影id、电影标题、评分、评分时间戳。将完整SQL语句及运行结果截图粘贴到对应答题报告中；

9、使用SQL语句统计用户各职业对电影的平均评分，输出职业和平均分。将完整SQL语句及运行结果截图粘贴到对应答题报告中；

10、使用SQL语句统计哪个年龄段参与电影评分的次数最多，输出年龄和评分次数。将完整SQL语句及运行结果截图粘贴到对应答题报告中。

11、使用SQL语句给users表中插入一条数据，数据的具体信息如下：用户ID为6041、性别为女性、年龄为25-34岁、职业为医生、邮编为11106。将完整SQL语句及运行结果截图粘贴到对应答题报告中。

12、使用SQL语句批量给movies表中插入两条数据，数据的具体信息如下：电影ID为3953、电影标题为Titanic、电影类型为浪漫；电影ID为3954、电影标题为Under the Light、电影类型为犯罪和神秘。将完整SQL语句及运行结果截图粘贴到对应答题报告中。

13、使用SQL语句修改users表中用户ID为6041的用户信息，将邮编改为02460。将完整SQL语句及运行结果截图粘贴到对应答题报告中。

14、使用SQL语句删除movies表中电影ID为3954的数据。将完整SQL语句及运行结果截图粘贴到对应答题报告中。

模块二 数据采集与处理

基本要求：

1、本模块为技能实操，满分30分。

2、禁止携带参考资料入场。

任务一：电影数据采集

【任务要求】

本任务是使用Python开发网络爬虫程序爬取电影数据，并将爬取的数据进行持久化存储。

请在“Desktop/大数据应用与服务竞赛/模块二/任务一 电影数据采集/FilmCrawl”项目中的“crawl_year”模块中编写代码，该模块用于从“慧影网”中爬取不同年代的电影数据。

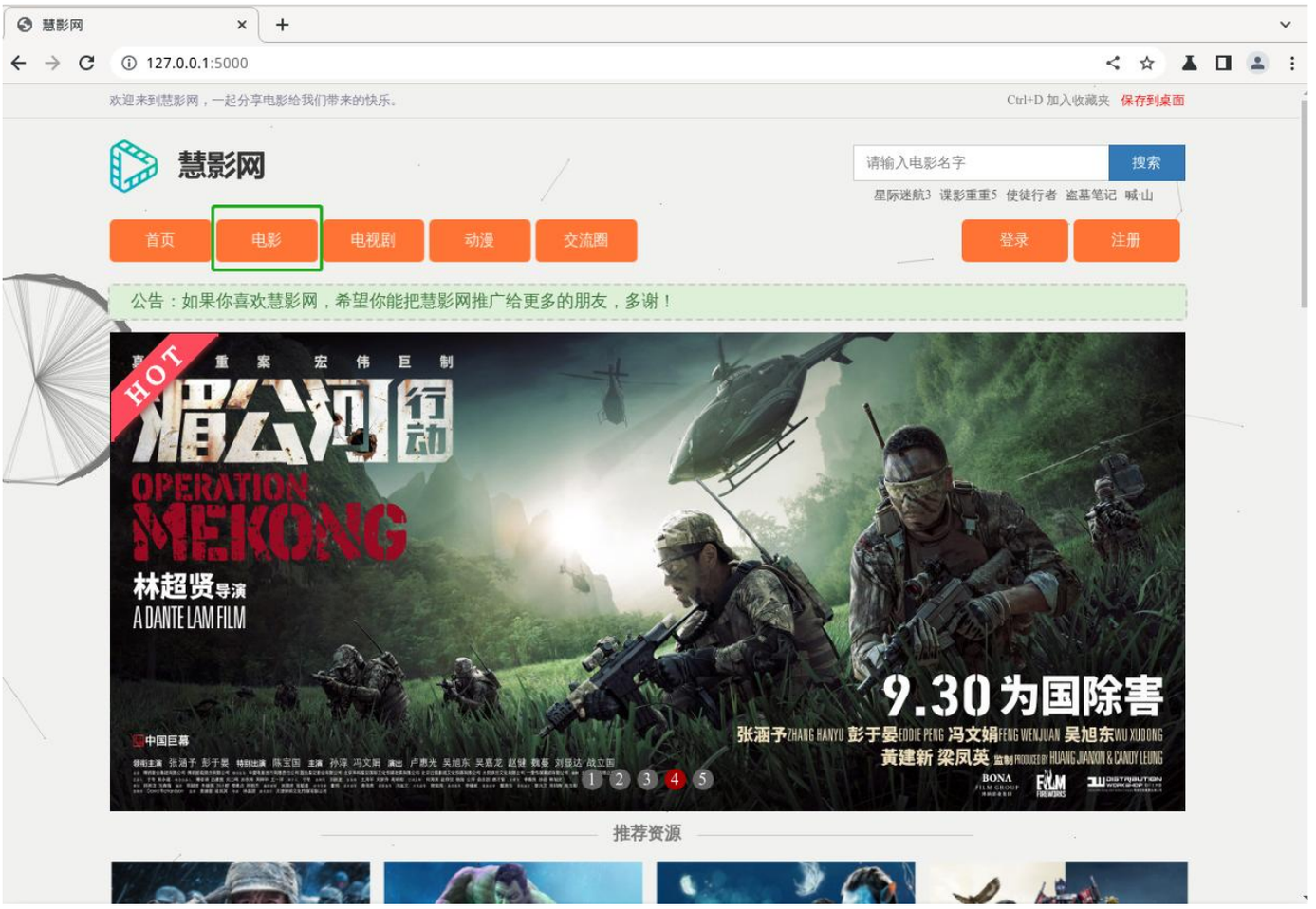
【任务需求背景】

当今时代，数据为王，掌握了多少数据、数据中蕴含有多少价值往往决定了我们能够获取多少收益。数据可以有不同的来源，一些互联网巨头拥有很多用户，掌握了大量的用户数据；数据也可以通过一些渠道进行购买。但对于规模不那么大的中小型企业来说，没有太多的用户，也就没有足够的用户数据，也没有足够的资金去购买数据。那么，他们应该怎样获取数据呢？--爬虫或许是一种答案，通过爬虫去互联网上抓取需要的数据，是一种低成本的数据获取方式。

本任务就是使用网络爬虫技术对数据进行采集，从“慧影网”中抓取电影数据，并将采集到的数据进行持久化存储。

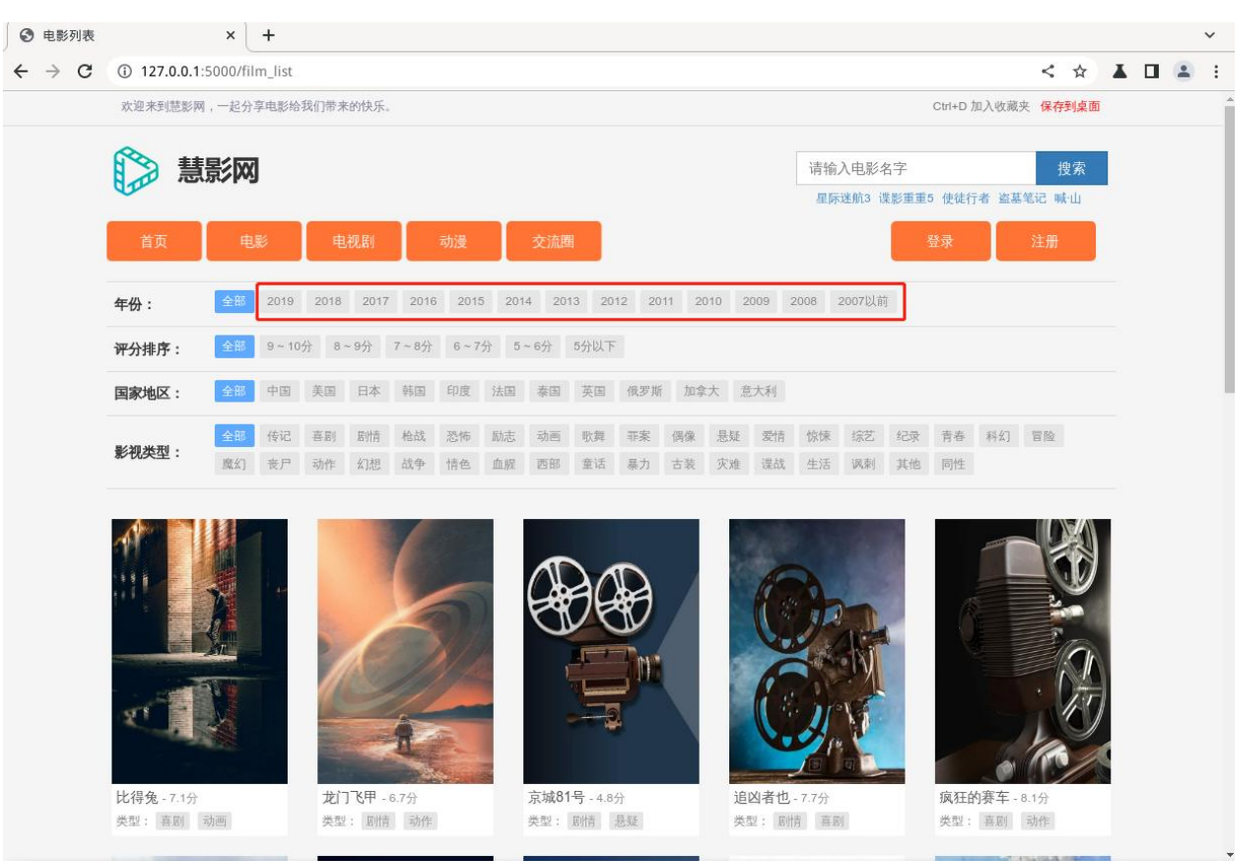
【具体任务】

1、使用谷歌浏览器访问“慧影网首页”，网站访问地址为【http://127.0.0.1:5000】，该网站顶部有首页、电影、电视剧、动漫、交流圈几个按钮，其中“电影”按钮可跳转到电影页，网站首页效果图如下：



慧影网首页

2、点击“电影”按钮可跳转到电影页面，如下图所示：

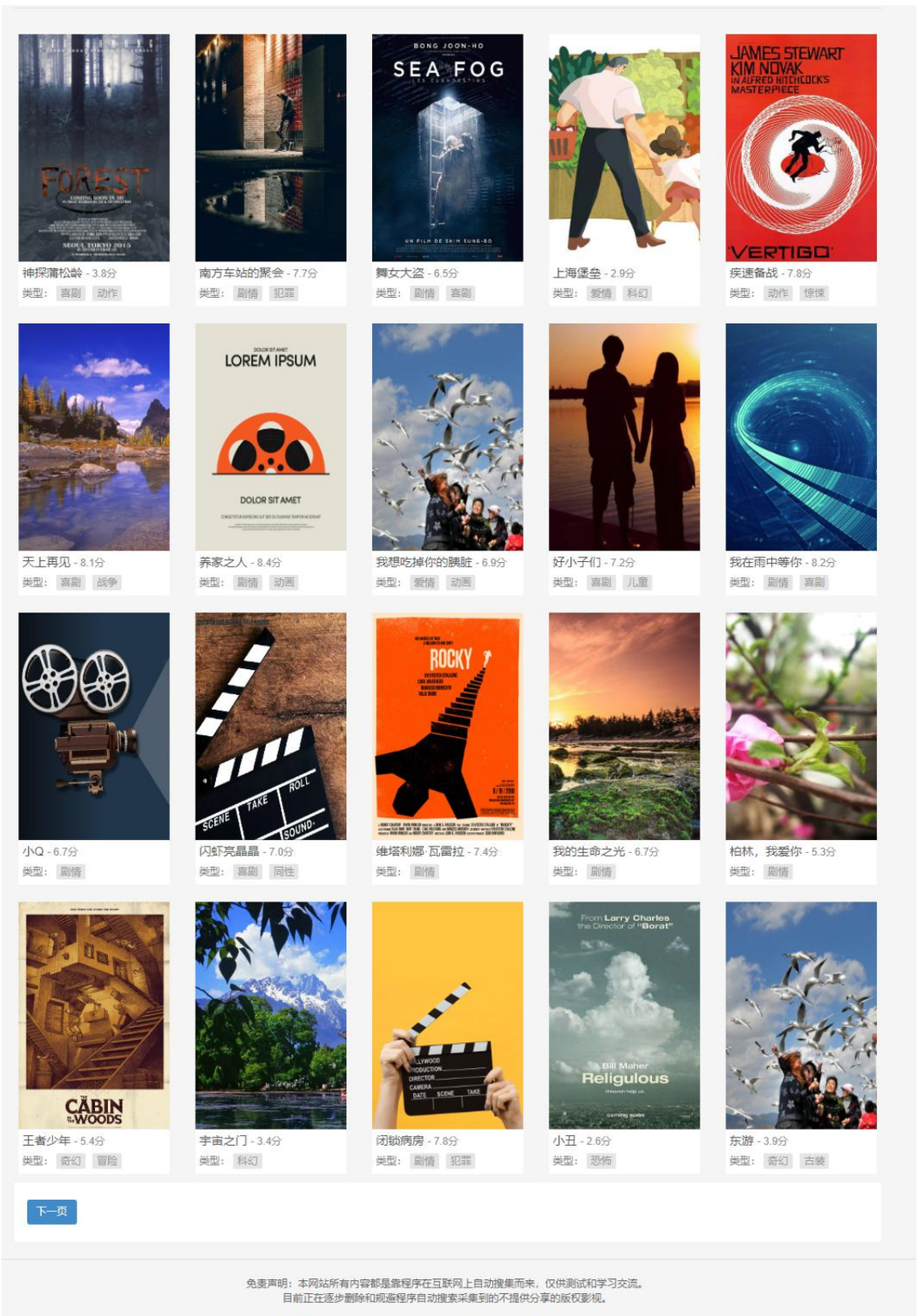


电影页面

该页面可以将电影按年代进行分类，页面上部有2019、2018、2017等年代按

钮。

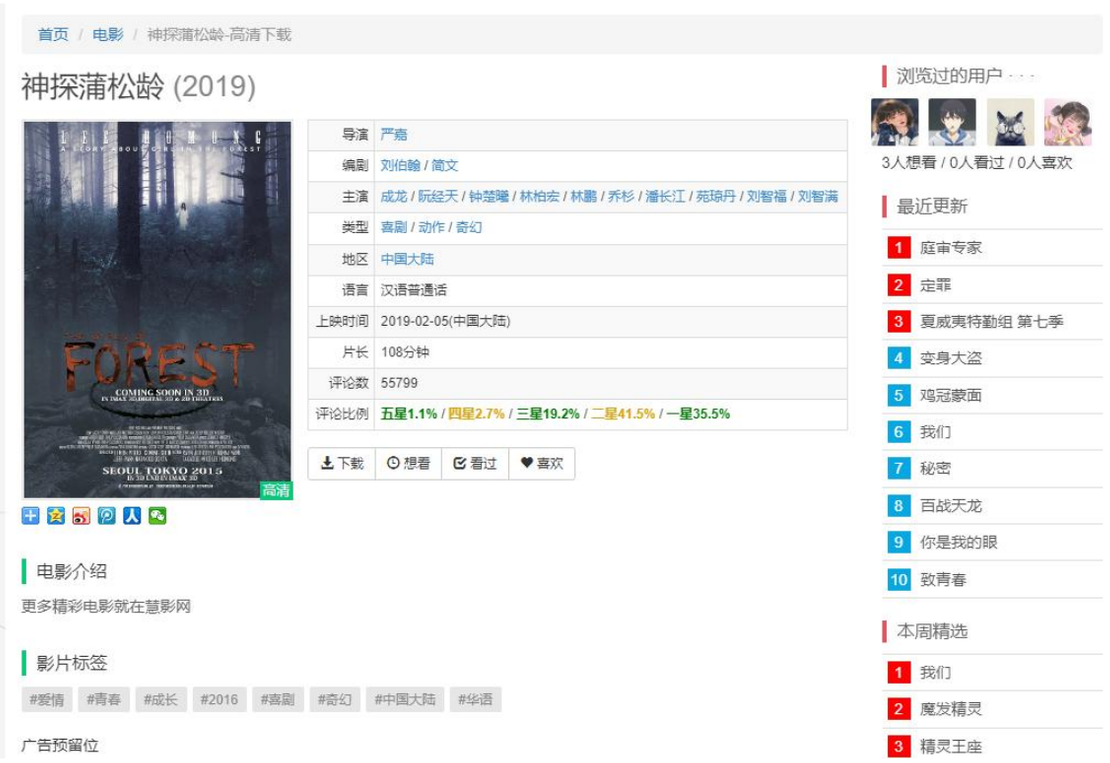
3、点击年代按钮，页面上便展示的是相应年代的电影，这里以2019为例，点击2019以后，页面上便展示2019年的电影。默认每页有20部电影，点击页面底部的“下一页”按钮可以进行翻页：



年代电影页面

4、点击具体的某部电影，可跳转到电影详情页面，页面上包含每部电影的

详细内容，如下图所示：



电影详情页

5、现需要按年代分类爬取电影数据，每个年代类别的数据均需要全部爬取，然后将不同年代的数据保存于不同的CSV文件中，CSV文件存储到FilmCrawl目录下的【data】目录中，若目录不存在，需要自行创建目录。数据文件名见如下“电影数据文件名说明表”，每个文件中需要爬取数据的列名见如下“抓取数据字段说明表”，注意，保存到文件中时，必须按表中字段的列举顺序进行保存。

电影数据文件名说明表

数据文件名	说明
年代名.csv （如“2019.csv”、“2007以前.csv”）	电影数据CSV文件一共包含2019至2007以前共计13个文件

抓取数据字段说明表

字段名	说明
movie_id	自己生成，从1开始，依次递增1
movie_name	电影名
year	电影年代（注意和数据文件名可能会有区别）
director	导演
editor	编剧，多个编剧以‘/’分隔

actor	演员，多个演员以‘/’分隔
movie_type	电影类型，多种类型以‘/’分隔
region	地区，多个地区以‘/’分隔
language	语言
on_time	上映时间
duration	时长
score	评分
comments	评论数
five_star	五星比例
four_star	四星比例
three_star	三星比例
two_star	二星比例
one_star	一星比例

保存到CSV文件中的内容示例如下：

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	
1	movie_	movie_nam	year	director	editor	actor	movie_typ	region	language	on_time	duration	score	comment	five_star	four_star	three_sta	two_star	one
2	1	无极	2005	陈凯歌	陈凯歌/张	张东健/张	剧情/动作	中国大陆	汉语普通	2005-12-1121分钟		5.3分	151997	五星8.0%	四星13.8%	三星32.1%	二星27.8%	一星

CSV文件中的内容示例

6、爬虫任务完成后，需要将相应的结果截图粘贴到答题报告中。（截图具体要求见答题报告）

任务二：电影数据处理

【任务要求】

现有一份电影数据集，具有多个字段，各字段信息如下表所示：

电影数据集字段说明表

数据字段	字段说明
movie_name	电影名
director	导演
editor	编辑
actor	演员
movie_type	电影类型
score	电影评分
comments	评论数
five_star	五星评论比例

four_star	四星评论比例
three_star	三星评论比例
two_star	两星评论比例
one_star	一星评论比例
language	语言
region	地区
on_time	上映时间
duration	时长

本任务需要使用pandas库对数据进行清洗及处理，包括数据类型处理、数据计算、缺失数据处理、数据分组聚合、数据排序、数据存储等。

请在“Desktop/大数据应用与服务竞赛/模块二/任务二 电影数据处理/FilmDataProcess”项目中编写代码，子任务1在“data_process.py”中编写代码，子任务2在“data_calc.py”中编写代码，数据集位于项目的“data”目录中，名为“movie.csv”。

【任务需求背景】

随着数字化时代的到来，数据已成为企业的一项不可或缺的资源和财富。然而，我们从互联网中得到的数据集往往是一些非结构化数据，其中通常会存在脏数据。所谓数据清洗，指的是对数据进行一些处理，以确保其准确性、完整性和可用性。在数据分析过程中，数据清洗是一项非常重要的任务，因为清洗数据可以减少错误率，提高数据的质量，使企业更好地利用数据资源进行数据分析及数据挖掘。本任务需要使用Python语言根据要求对数据进行清洗及处理，并将处理后的数据集进行存储。

【具体任务】

1、电影数据清洗，具体要求如下：

(1) 评分和评分星级处理：评分和评分星级数据类型不便于后续进行计算，经过必要的处理将其处理为数字型，其中score处理为浮点型，四舍五入保留1位浮点数，评分星级也处理为浮点型，四舍五入保留3位浮点数。其中如果某列有缺失，需要根据已有的评分和评分星级进行计算

得到。

(2) 上映时间处理：时间统一处理为“年/月/日”的样式，例如“2009/01/01”，如果包含多个日期，只保留第一个日期，如果日期中不包含年，则默认处理为2019年，如果日期不包含月，则默认处理为01月，如果不包含日，则默认处理为01日；

(3) 时长处理：时长处理为整型，如果有秒数的，可忽略不计，时长有缺失的，要使用电影上映同年的其它电影的平均时长来填充，如果没有和其同年上映的电影，则使用其它所有年份电影上映平均时长的平均值来填充（例如现有年份2020、2021、2022，没有2020年上映的电影，则用2021年上映电影的平均时长和2022年上映电影的平均时长取平均值去填充）；

(4) 处理完成之后，将其保存到“FilmDataProcess”项目下的“deal_data”目录中，命名为“movie_deal.csv”，注意不要改变原来列数据的顺序，列名需要保留，然后按要求截图并粘贴到答题报告相应位置。

2、数据计算，具体要求如下：

(1) 本任务需要读取子任务1处理得到的“movie_deal.csv”文件；

(2) 计算演出电影最多的20名演员，并按降序排序，计算完成后，将其保存到“FilmDataProcess”项目下的“deal_data”目录中，命名为“count.csv”，保存的字段包括actor(演员名)、count(参演电影数)、score(演员参演电影平均评分)；

(3) 计算参演电影大于等于10部的平均评分最高的前20名演员，并按降序排序，计算完成后，将其保存到“FilmDataProcess”项目下的“deal_data”目录中，命名为“score.csv”，保存的字段包括actor(演员名)、count(参演电影数)、score(演员参演电影平均评分)；

(4) 计算最佳的20部电影，最佳电影的定义为：评论数大于等于5000，首先看评分高的，如果评分一致再看评论星级，优先考虑五星比例高的，再考虑四星，再考虑三星，再考虑二星，再考虑一星；计算完成后，将其保存到“FilmDataProcess”项目下的“deal_data”目录中，命名为

“best_movie.csv”。保存的字段包括movie_name(电影名),score(评分),five_star(五星比例),four_star(四星比例),three_star(三星比例),two_star(二星比例),one_star(一星比例);

(5) 处理完成后，按要求截图并粘贴到答题报告对应的位置。

任务三：电影数据标注

【任务要求】

本任务是使用WPS对给定的电影数据进行标注，并进行持久化存储。

原始数据保存在“Desktop/大数据应用与服务竞赛/模块二/任务三 电影数据标注/FilmDataAnnotation”项目中，使用WPS工具打开电影数据完成此任务，电影数据集具有多个字段，各字段信息如下表所示：

电影数据集字段说明表

数据字段	字段说明
id	标识号
title	电影名
overview	电影摘要
release_date	首映日期
runtime	电影时长
tagline	电影标语

【任务需求背景】

数据标注是人工智能产业的基础，是机器感知现实世界的起点。随着AI行业的蓬勃发展，对数据的需求呈井喷式增长，从某种程度上来说，没有经过标注的数据就是无用数据。数据标注的越精准、对算法模型训练的效果就越好。大部分算法在拥有足够多普通标注数据的情况下，能够将准确率提升到95%，但从95%再提升到99%甚至99.9%，就需要大量高质量的标注数据。

【具体任务】

1、使用WPS打开Initial_movie.csv。

2、在末尾新增一列数据为“movie_type”，若电影时长超过120分钟，则打标签为“1”，否则标记为“0”。标记完成后保存到当前目录，并将标记后的数据截

图粘贴到答题报告对应位置。

模块三：大数据应用开发

基本要求：

1、本模块为技能实操，满分45分。

2、禁止携带参考资料入场。

任务一：基于 Tableau 进行数据分析与可视化

【任务要求】

本环节需要使用数据可视化工具Tableau，基于亚马逊股票数据进行分析、可视化展示；股价数据存储在Windows桌面“大数据应用与服务竞赛”目录下的“T_amzn.csv”中，数据表中记录2013年至2022年的股票历史信息，包含日期、收盘价、开盘价、最高价、最低价、交易量六列指标，其中交易量的“M”表示单位：百万。

数据字段表

表名	字段	字段说明
T_amzn.csv	日期	/
	收盘价	当日收盘价
	开盘价	当日开盘价
	高	当日最高价
	低	当日最低价
	交易量	当日交易量，单位百万

【任务需求背景】

Tableau是一款强大的可视化BI工具，可以非常便捷的进行数据连接、数据分析与可视化，交互式的界面帮助用户探索、分析数据。使用Tableau工具分析亚马逊股价数据。

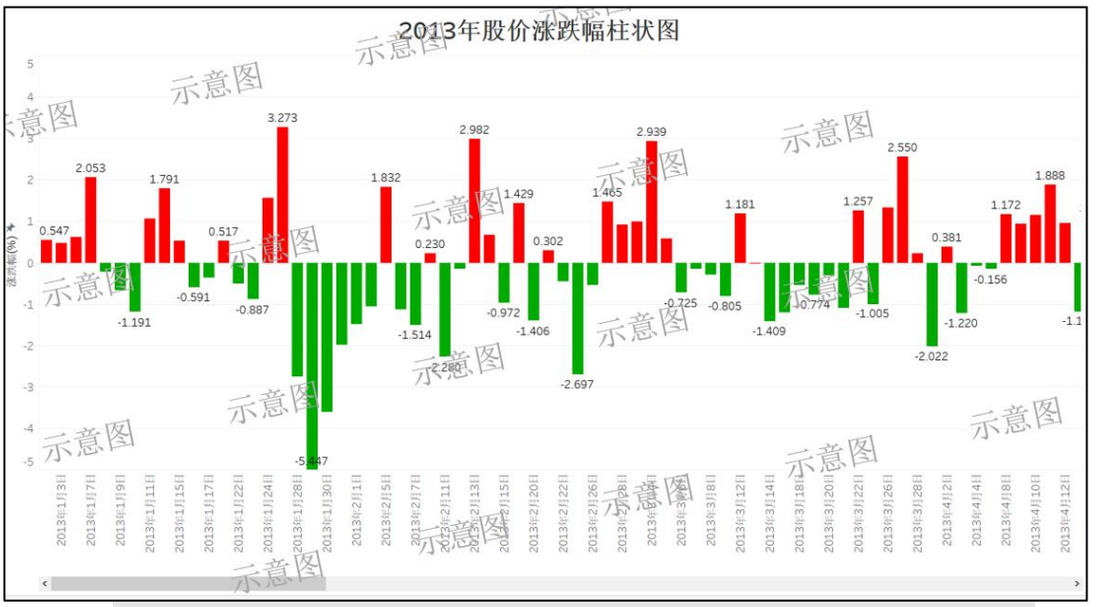
【具体任务】

（1）**绘制柱状图展示2022年股价的涨跌幅**，具体要求如下：

- 柱状图标题设置为“2022 年股价涨跌幅柱状图”，字号 20、加粗、居中显示。
- 横轴显示 2022 年的所有数据日期信息，日期格式显示为【年/月/日】，不显示横轴标签；
- 纵轴显示涨跌幅数值，纵轴标签设置为“涨跌幅（%）”。

- 柱状图设置显示数据标签，数据标签水平显示在柱子顶部。
- 设置柱状图的颜色，涨跌幅为正时显示为红色，涨跌幅为负时显示为绿色。

其中涨跌幅计算公式为：

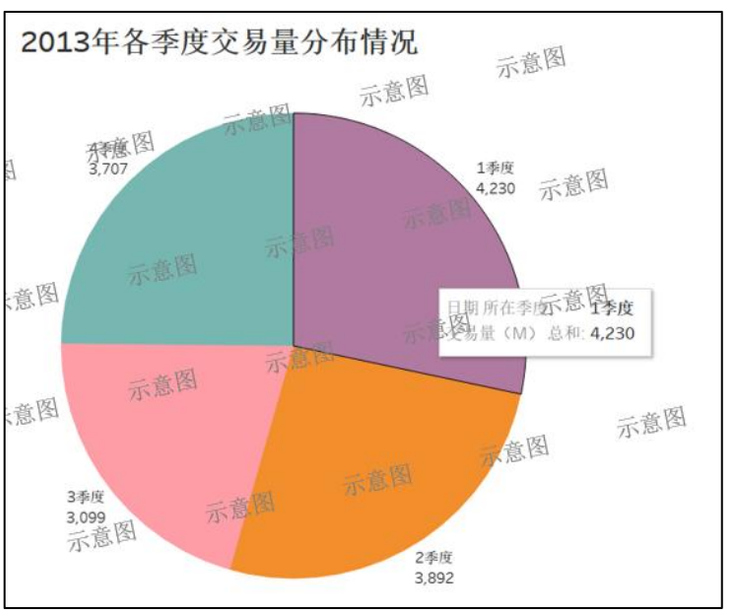


$$\text{涨跌幅} = (\text{当日收盘价} - \text{当日开盘价})/\text{当日开盘价} * 100$$

涨跌幅示意图

（2）绘制**2022年各季度交易量分布饼图**，具体要求如下：

- 饼图标题设置为【2022 年各季度交易量分布情况】，字号 18、居左显示。
- 在饼图外设置显示饼图的标签及对应的交易量总和。
- 各扇形区域设置显示不同的颜色，其他参数可自行调整。



饼图示意图

（3）将绘制完成后的图表进行截图，粘贴到竞赛平台答题报告上对应位置。

任务二：基于 Excel 进行数据分析与可视化

【任务要求】

本环节需要使用办公软件Excel工具，对亚马逊股票信息进行分析与可视化。

【任务需求背景】

股票作为一个情绪市场指标、企业的融资工具，向上连接了公司经营情况，向下是各类衍生投资产品为我们带来收益，如基金、可转债等。亚马逊作为多元化零售行业，为客户提供一系列产品和服务。使用Excel工具对其历史股票信息进行分析与可视化，掌握使用Excel进行数据分析应用。

【具体任务】

股价数据存储在Windows桌面“大数据应用与服务竞赛”目录下的“E_amzn.csv”中，数据表中记录2013年至2022年的股票历史信息，包含日期、收盘价、开盘价、最高价、最低价、交易量六列指标，其中交易量的“M”表示单位：百万。使用Excel打开“E_amzn.csv”文件，对数据进行分析与可视化，具体要求如下：

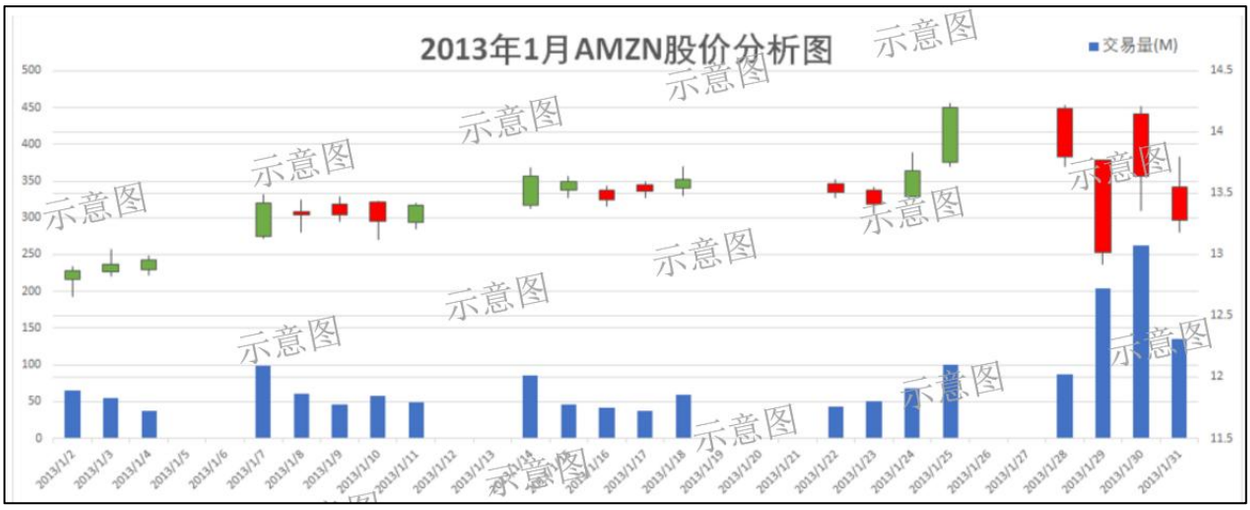
（1）将csv数据表读取为Excel数据表，并分析每个数据字段类型，使字段能进行统计、计算等。

（2）数据中每日的股票数据应该只有一份，数据中包含重复数据，请过滤掉重复日期的数据，并对数据根据日期升序进行排序。

（3）对数据进行统计分析，找出2013~2022年中交易总量最高的季度，并使用该季度中交易量最高的月数据绘制股价图【成交量-开盘-盘高-盘低-收盘图】，图表基本设置要求如下：

- 设置图表标题为【X 年 X 月 AMAZ 股价分析图】，标题居中显示（注意：日期对应交易总量最高年月）。
- 横坐标显示为日期轴，日期显示格式为【年/月/日】，倾斜显示完整数据。
- 绘制的图表为双坐标，合理设置坐标轴取值范围，使交易量显示在箱线图下方，实现图表不重叠显示。
- 图例保留“交易量（M）”。

- 默认的股票图是黑白色调的，将箱线图颜色改为红绿色调（红色表示上涨、绿色表示下跌）。



股价示意图

（4）分析2013年~2020年期间每年的最高收盘价，使用该数据绘制柱状图，具体要求如下：

- 标题设置为“2013 年~2020 年股票收盘价分析”，居中显示。
- 横坐标显示年份数据、纵坐标数据显示收盘价，并设置纵坐标标题为收盘价。
- 显示数据标签，其中收盘价超出 100 的值标红显示。

柱状图示意图

（5）将绘制完成后的图表进行截图，粘贴到竞赛平台答题报告上对应位置。

任务三：基于 Python 的美职篮球员数据分析

【任务要求】

现有一份关于NBA球员的数据集，字段说明如下表：

球员数据集说明表

列名	字段说明	列名	字段说明
Rk	序号	DRB	防守篮板
PLAYER	球员	TRB	总篮板
POSITION	球员位置	AST	助攻
AGE	年龄	STL	抢断
MP	出场时间	BLK	盖帽
FG	命中次数	TOV	失误
FGA	出手次数	PF	犯规
FG%	命中率	POINTS	得分
3P	三分球命中数	TEAM	球队
3PA	三分球出手数	GP	场次
3P%	三分球命中率	MPG	出场时间
2P	两分命中数	ORPM	进攻正负值
2PA	两分出手数	DRPM	防守正负值
2P%	两分命中率	RPM	正负值
eFG%	真实命中率	WINS_RPM	赢球正负值
FT	罚球命中数	PIE	球员贡献值
FTA	罚球次数	PACE	每48分钟回合数
FT%	罚球命中率	W	赢球场次
ORB	进攻篮板	SALARY_MILLIONS	薪水

本任务需要使用Numpy、Pandas、Matplotlib、Seaborn等库按要求对数据进行处理及分析，然后将结果进行可视化。

请在“Desktop/大数据应用与服务竞赛/模块三/任务三 美职篮球员数据分析/NBA”项目下的“NBA_data_analysis”模块中编写代码实现功能，数据集在项目的“data”目录中。

【任务需求背景】

NBA是美国运动员薪酬最高的几大联赛之一，在2023-2024赛季，斯蒂芬库里

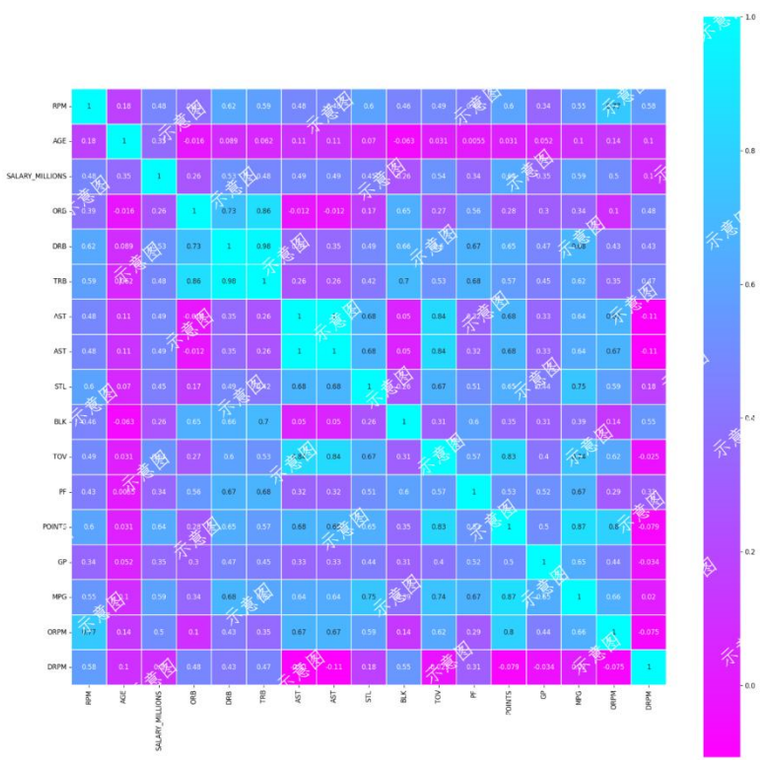
获得了超过5191万美元的薪资，勒布朗詹姆斯获得了超过4447万美元的薪资，由于每支球队都有工资上限，作为球队管理人员如何更有效的确定球员薪水（避免溢价合同）是十分重要的。本任务就是对NBA球员数据集做相关分析来了解球员和球队相关指标，进而帮助球队做出有益的决策。

【具体任务】

1、请使用以下表格中的字段对数据进行相关性分析，并使用seaborn绘制出热力图，要求使热力图的每个单元格为正方形，在每个热力图单元格中写入数据值，颜色映射为“cool_r”。绘制完成后将热力图粘贴到答题报告对应位置。

RPM, AGE, SALARY_MILLIONS, ORB, DRB, TRB, AST, STL, BLK, TOV, PF, POINTS, GP, MPG, ORPM, DRPM

相关性分析字段



热力图示意图

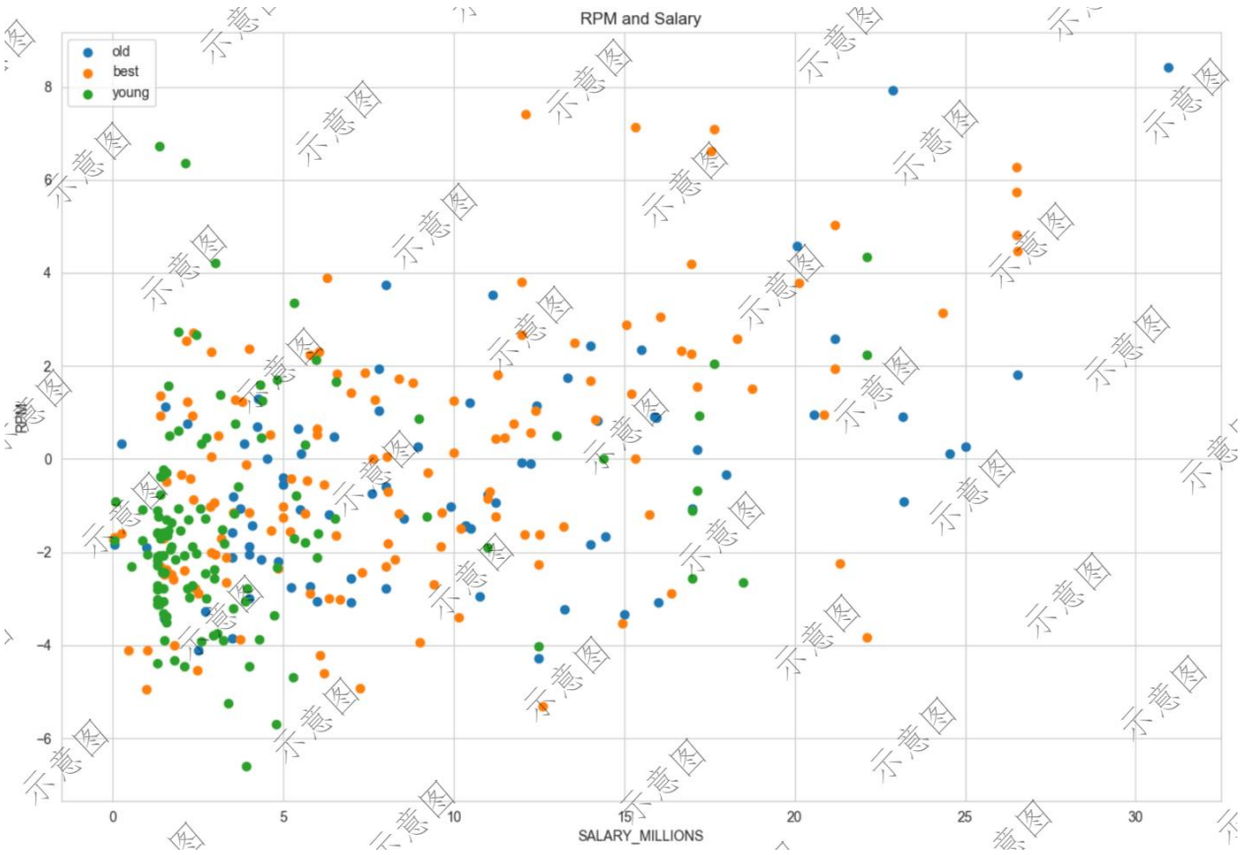
2、按照球员的效率值进行排名，取出效率值最高的前五名球员，并输出球员的["PLAYER", "RPM", "AGE"]三个特征值，将输出结果截图粘贴到答题报告对应位置。

3、按照下表要求将NBA球员年龄根据岁数划分为三个类别并放在名为“age_cut”的字段中，然后基于年龄段对球员的薪水和效率值进行分析，并使用Matplotlib绘制出散点图，画布大小为（20,15），dpi为120，主题使用“whitegrid”；标题为“RPM and Salary”，x轴标签为“SALARY_MILLIONS”，

y轴标签为“RPM”；在左上角添加图例，绘制完成后将散点图粘贴到答题报告对应位置。

年龄分类表

年龄范围	类别
AGE <= 24	young
AGE >= 30	old
30 < AGE < 24	best



散点图示意图

任务四：基于 Python 的亚马逊股票数据分析

【任务要求】

现有一份亚马逊股票数据集，字段说明如下表：

股票数据集说明表

列名	字段说明
Date	日期
Open	开盘价
High	最高价

Low	最低价
Close	收盘价
Adj Close	调整后的收盘价
Volume	成交量

本任务需要使用Numpy、Pandas、Matplotlib、Seaborn等库按要求对数据进行处理及分析，然后将结果进行可视化。

请在“Desktop/大数据应用与服务竞赛/模块三/任务四 亚马逊股票数据分析/stock”项目下的“amazon”模块中编写代码实现功能，数据集在项目的“data”目录中。

绘图过程中如有使用中文字体的地方，请统一使用“SimSun”字体。

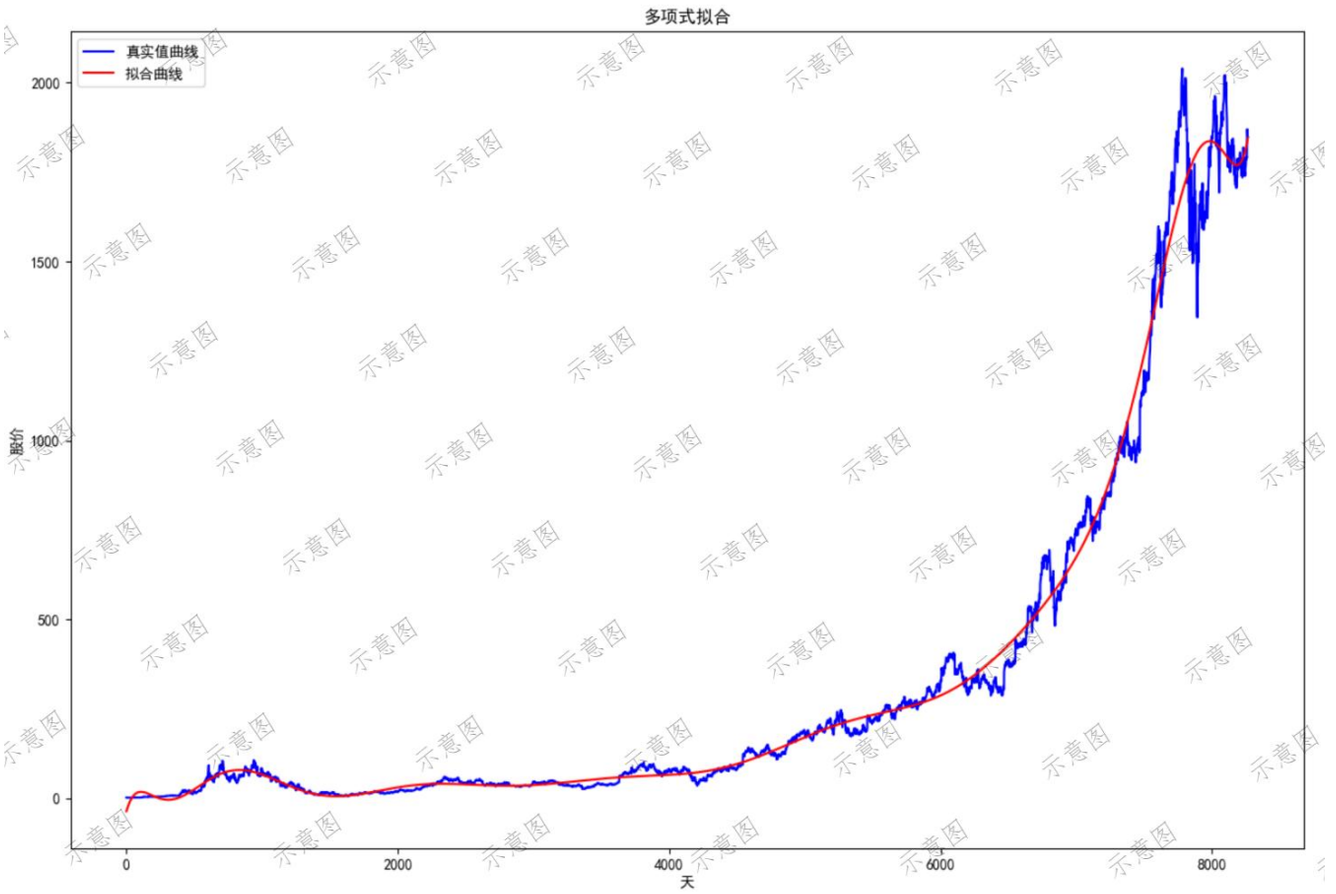
【任务需求背景】

股票数据分析是对股票交易数据进行分析、归纳、反映和挖掘等过程的总称。它可以帮助分析人员从繁杂的股票数据中获取有价值的信息，推测市场变化趋势，并作出相应的决策。在当前股票交易市场，股票数据分析已经成为一种必备的工具，它可以帮助投资者规避风险、获取收益。除此之外，通过股票数据分析，可以发掘公司经营、市场变化等有关信息，推断股票入市和出市的最佳时机。

随着互联网的普及和金融领域的不断创新，股票交易已经变得越来越普遍，也越来越复杂。在这样一个背景下，股票数据分析的作用越来越凸显。本任务将对亚马逊股票数据分析作出深入剖析，以期从中挖掘出一些有用的信息。

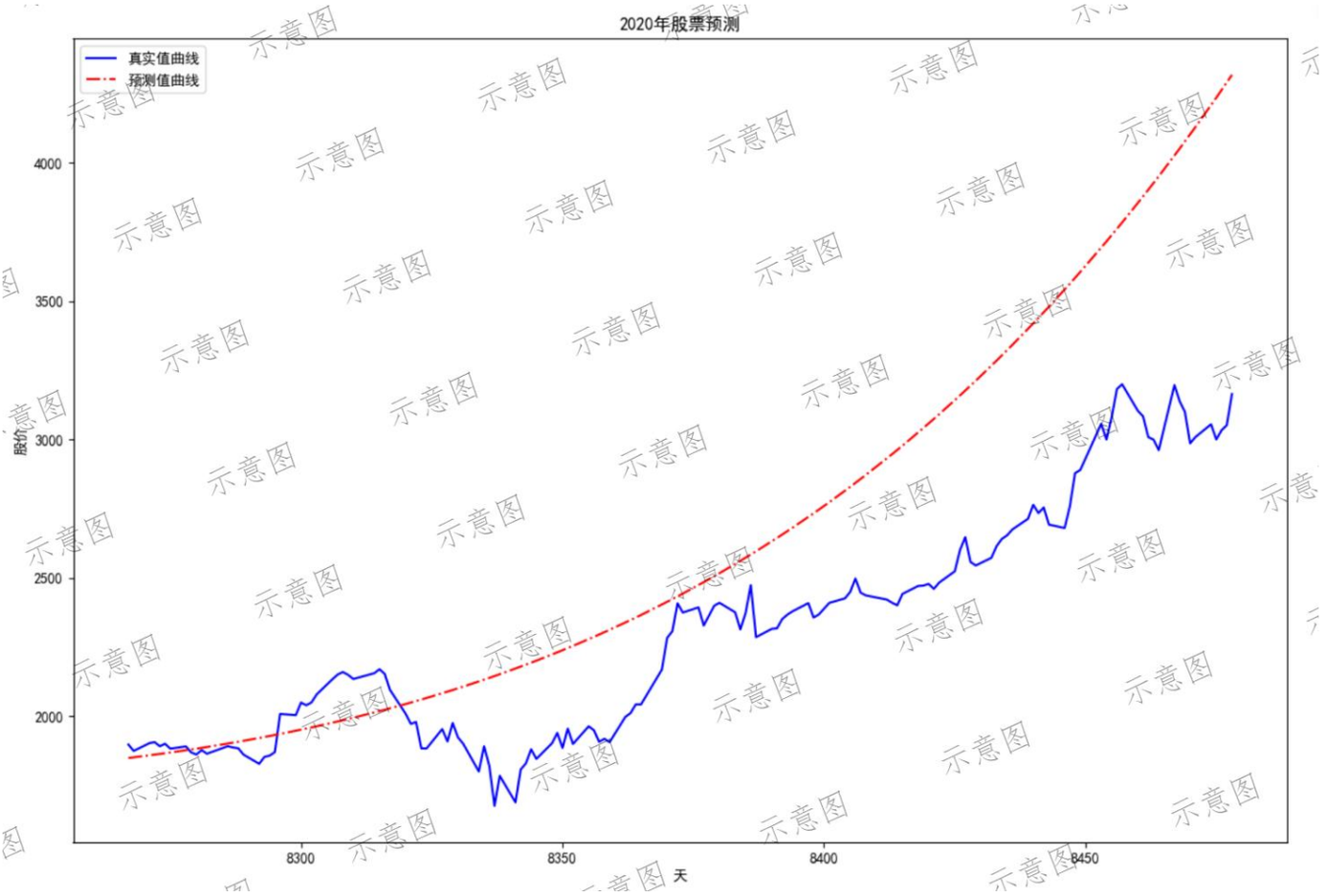
【具体任务】

1、给数据添加“day”一列用来表示当天与1997-05-15相差的天数，然后使用2020年以前的样本进行多项式拟合，并将该部分的股价曲线以及拟合后的曲线使用Matplotlib绘制出来，画布大小为（15，10）；x轴标签为“天”，y轴标签为“股价”，标题为“多项式拟合”；然后添加图例，图例标签分别为“真实值曲线”和“拟合曲线”；真实值曲线使用蓝色，拟合曲线使用红色。绘制完成后将图粘贴到答题报告对应位置。



多项式拟合示意图

2、使用拟合的多项式预测2020年的股价，并将2020年的真实股票价格和预测值通过Matplotlib绘制折线图进行可视化。画布大小为（15，10）；x轴标签为“天”，y轴标签为“股价”，标题为“2020年股票预测”；然后添加图例，图例标签分别为“真实值曲线”和“预测值曲线”；真实值曲线使用蓝色，预测值曲线使用红色并将其线型修改为点划线（-.）。绘制完成后将图粘贴到答题报告对应位置。



股票预测示意图

任务五：职业素养

【任务要求】

参赛选手操作规范、遵守考场纪律、收纳整理干净整洁、安全意识良好、文明竞赛。