



计算机科学与技术学院

《企业实训-大数据》

项目设计报告

学 号 17210122

姓 名 林栩

企业教师 王大瑞

助课教师 邹淑雪

撰写日期 2025/02/27

一、实训项目名称：

基于Hadoop和Spark大数据分析的电影历史与观众偏好研究

二、项目概述

2.1 项目设计目标

本项目的设计目标是通过大数据技术，深入挖掘电影历史数据与观众行为数据，揭示电影产业的发展规律与观众偏好的变化趋势，为电影行业的学术研究、市场决策以及文化传播提供数据支持和理论依据。具体目标包括以下几个方面：

a. 构建高效的数据处理流程

利用 Hadoop 的分布式存储系统 HDFS 存储海量电影数据，并通过 Spark 实现高效的数据清洗、整理与分析，构建一个稳定、可扩展的大数据处理框架，为后续分析提供高质量的数据基础。

b. 分析电影历史发展趋势

通过对电影历史数据的深度挖掘，研究电影产业在技术、题材、风格等方面演变规律，揭示不同时期电影发展的特点及其背后的社会文化因素。

c. 探索观众偏好变化

基于观众评分、评论、观影行为等数据，分析观众对电影类型、导演、演员等的偏好变化，研究不同地区、年龄段、文化背景下的观众需求差异，为电影制作和市场营销提供参考。

d. 实现数据可视化与直观展示

利用 Python 中的数据处理与可视化库（如 pandas、matplotlib、seaborn 等），将分析结果以图表、热力图、趋势图等形式直观呈现，便于研究者与决策者快速理解数据背后的规律。

e. 支持电影产业的实践应用

通过数据分析结果，为电影制作方提供内容创作建议，为发行方提供市场策略支持，同时为学术研究提供数据驱动的理论依据，推动电影产业的可持续发展。

f. 验证大数据技术的应用价值

通过本项目，验证 Hadoop、Spark 等大数据技术在文化研究与市场分析中的实际应用效果，探索其在更广泛领域的潜力与价值。

2.2 项目应用场景

1. 电影制作与创作：电影制作方可以根据研究结果，选择受观众欢迎的电影类型和导演，调整电影时长，优化影片结构，从而提高电影质量，增加观众满意度。

2. 市场营销与宣传：电影公司和流媒体平台可以利用观众的评分偏好数据进行精准的市场营销，制定符合观众需求的宣传策略，提升电影的曝光度和观众参与度。

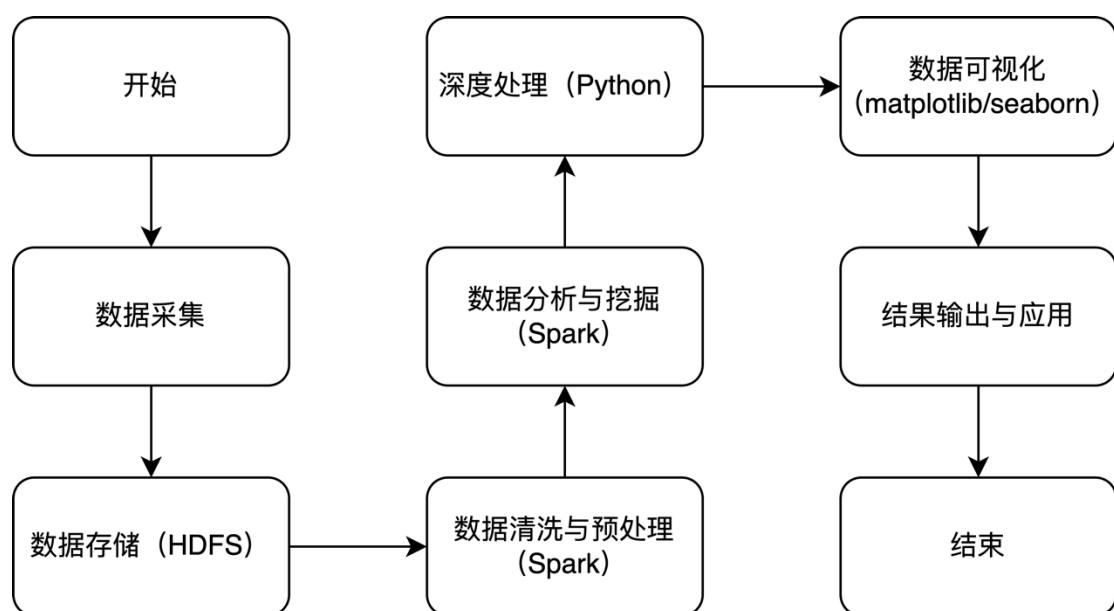
3. 观众行为分析：本项目为电影行业提供了关于观众评分行为的深入洞察，能够帮助分析不同观众群体的偏好，为定制化观影推荐系统提供数据支持。

4. 电影评估与决策支持：影视投资方和电影节组织可以参考评分分析结果，评估电影项目的潜力，做出更为科学和精准的投资决策，确保更高的投资回报率。

三、项目设计

3.1 大数据框架设计

1、数据流框架



2、业务流框架

a. 数据采集与存储

数据来源：从公开的电影数据库（豆瓣）以及观众评分、评论平台获取电影历史数据与观众行为数据。

数据存储：将采集到的原始数据存储于 Hadoop 分布式文件系统（HDFS）中，利用其高容错性和高吞吐量特性，确保数据的安全性与可扩展性。

b. 数据清洗与预处理

数据清洗：使用 Spark 对原始数据进行清洗，处理缺失值、重复值、异常值等问题，确保数据质量。

数据预处理：对清洗后的数据进行格式转换、字段提取、数据归一化等操作，为后续分析做好准备。

c. 数据分析与挖掘

统计分析: 利用 Spark 对电影历史数据进行统计分析, 计算电影均分、评分等指标的分布与趋势, 并设计一些新的指标来量化结果。

观众偏好分析: 基于观众行为数据, 分析观众对电影类型、导演等的偏好, 研究不同群体的需求差异。

d. 深度处理与可视化

深度处理: 使用 Python 中的 numpy、pandas 等库对 Spark 处理后的结果进行进一步分析, 提取关键特征与指标。

可视化展示: 通过 matplotlib、seaborn 等可视化工具, 将分析结果以图表、热力图、趋势图等形式呈现, 便于直观理解与决策支持。

e. 结果输出与应用

报告生成: 将分析结果整理为研究报告, 包含电影历史发展趋势、观众偏好分析、市场建议等内容。

实践应用: 为电影制作方、发行方提供数据驱动的建议, 支持内容创作与市场策略制定。

3、大数据框架

本项目的大数据框架基于分布式计算与存储技术, 结合多种大数据组件, 构建了一个高效、可扩展的数据处理与分析平台。以下是项目中用到的主要大数据组件:

a. Hadoop 分布式文件系统 (HDFS)

b. Apache Spark

c. Python 数据处理与可视化库:

- o numpy
- o pandas
- o matplotlib
- o seaborn
- o collections
- o scipy
- o math 与 ast

3.2 项目表结构设计

a. movies.csv

字段名	类型	说明
Movie_id	string	电影id
Name	string	电影名称
Alias	string	别名
Actors	string	主演
Cover	string	封面地址
Director	string	导演
DOUBAN_Score	int	豆瓣评分

DOUBAN_Votes	int	豆瓣评分人数
Genres	string	类型
IMDB_id	string	IMDB电影id
Languages	string	语言
Mins	int	时长
Official_Site	string	地址
Regions	string	国家/地区
Release_Date	date	上映时间
Slug	string	加密的url
Storyline	string	电影描述
Tags	string	标签
Year	date	发行年份
Actor_ids	string	演员与个人id的关系
Director_ids	string	导演与个人id的关系

b.ratings.csv		
字段名	类型	说明
Rating_id	string	评价id
User_MD5	string	用户id
Movie_id	string	评价电影id
Rating	int	评分
Rating_Time	date	评价时间

四、项目实现

4.1 大数据环境配置

1、资源环境配置

服务器	暂无
操作系统	Windows10
大数据平台	Hadoop-2.7.2
服务和组件	Spark-2.1.0 numpy-1.26.4 pandas-1.5.1 matplotlib-3.8.2 seaborn-0.13.2 scipy-1.15.2 collections math ast

2、生产集群

在本项目中，生产集群的设计主要考虑到大数据处理和计算任务的高效执行。集群资源规划应注重以下几个方面：

a. 计算资源规划：

节点配置: 根据数据分析的计算需求，选择高性能计算节点进行任务处理。每个节点配备至少 16 核 CPU、64GB 内存，并且采用高效的 GPU 进行深度学习模型训练和大规模数据处理。集群总节点数根据具体的数据规模可进行灵活扩展。

负载均衡: 采用分布式负载均衡策略，确保任务在计算节点之间均匀分配，避免计算资源浪费，提升整体系统性能。

b. 存储资源规划:

分布式存储: 使用分布式存储系统（如 Hadoop HDFS、Ceph 或分布式对象存储）来处理海量数据。数据应按类别和重要性分层存储，采用热冷数据分离策略，提高存储效率和成本效益。

数据备份与冗余: 考虑到数据安全性，集群配置多副本数据备份，并定期进行数据迁移和灾备演练。

c. 网络资源规划:

高速网络连接: 集群节点之间通过高带宽低延迟的网络进行连接，使用 10GbE 或更高规格的网络接口，确保数据传输不成为瓶颈。

网络隔离与安全: 采用虚拟化技术和网络隔离策略，以确保集群中不同任务的安全性，防止数据泄露或计算资源受到不必要的干扰。

d. 任务调度与管理:

集群调度系统: 使用如 Kubernetes 或 YARN 等资源管理和任务调度系统，自动化调度数据处理任务和计算任务，保证集群资源的最大化利用，并支持动态扩容和容错处理。

监控与优化: 集成系统监控工具（如 Prometheus、Grafana）进行实时性能监控，基于监控数据自动优化集群资源分配，减少任务执行时间，提高处理效率。

e. 弹性扩展与容错机制:

横向扩展: 集群应支持按需横向扩展，随着数据量增长或计算需求增加，能够快速增加计算节点和存储节点，确保项目可持续增长。

容错机制: 在集群内实现故障自动恢复和容错机制，确保单节点故障不会影响整体系统的稳定性和可用性。

4.2 项目数据说明

见4.2节

4.3 大数据环境搭建

a. Hadoop环境检查

1. 启动hadoop，用jsp指令检查进程完整性
2. 访问hdfs，创建目录，上传文件

b. spark环境检查

3.启动spark shell进行环境检查

c. idea环境配置

- 4.配置jdk和scala
- 5.本机配置hadoop
- 6.配置maven
- 7.在idea中安装Big Data Tools插件
- 8.在idea项目中创建java的Maven项目，选择jdk1.8，配置pom.xml
- 9.进入src/main目录中，修改java名称为scala
- 10.scala文件夹下建包，开始编程

4.4 业务流实现

S1.电影评分历史趋势分析

S1.1 涉及字段:

`movies_Year, movies_DOUBAN_Score`

S1.2 分析目的与方法:

了解评分的最直观的方式就是观察变化趋势，通过这Year字段可以将每一部电影归到其属于的年份集合中，并带电影的评分属性，针对年份集合内的所有电影计算均分，即可得到不同年份的电影均分水平，随后利用python进行可视化呈现数据。

S1.3 数据分析可视化结果:

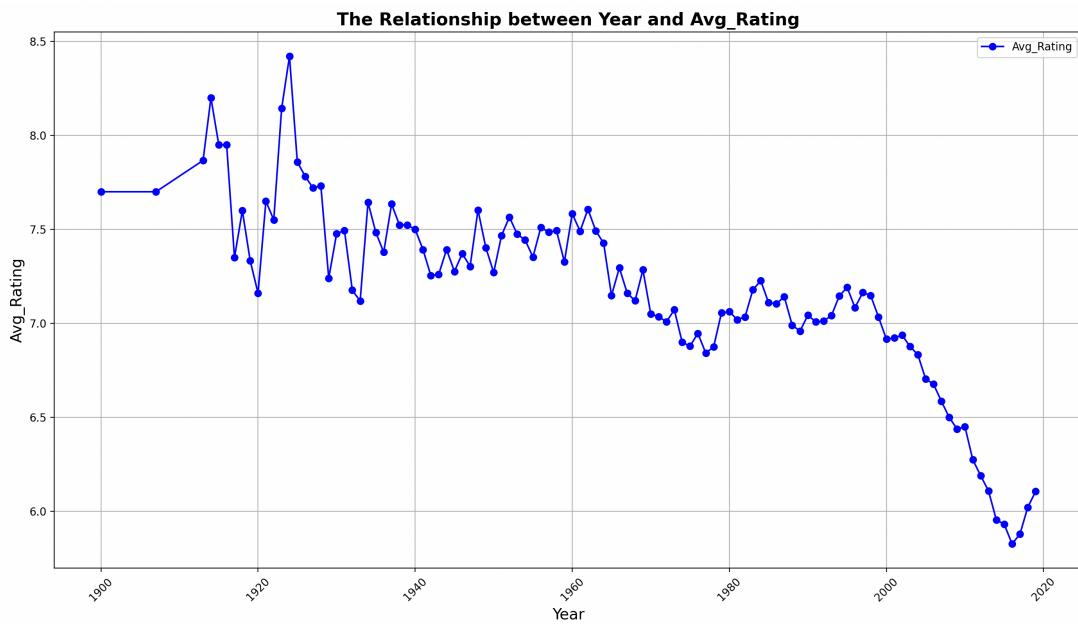


Fig1.1 年度-均分折线图

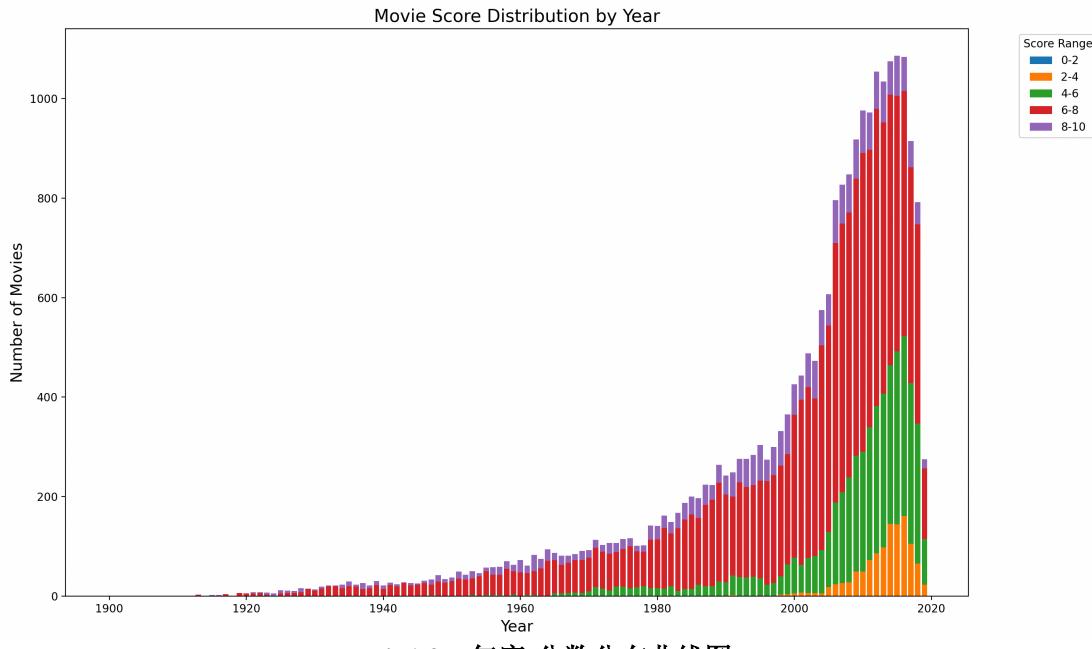


Fig1.2 年度-分数分布曲线图

S1.4 结论：

总体来看，电影的平均评分呈现出逐步波动并下降的趋势，特别是进入21世纪后，评分出现下降。而电影评分的分布情况则显示，随着电影产业的发展，电影数量的激增以及观众评分标准的提高，使得高评分电影逐渐增多，但也伴随着低评分电影的增加。这一现象反映了电影质量的两极分化。

S2. 电影评分时间分布

S2.1 涉及字段：

ratings_userId, ratings_Rating, ratings_Time

S2.2 分析目的与方法：

对豆瓣这个网站而言，上面呈现的电影评分并非是完全由参与评分的用户的意志决定的，豆瓣有一个属于自己的算法来过滤掉一些异常的用户值之后再经过一套换算算法得出最终评分，因此，为了更好地理解电影产业的演变、观众行为以及市场趋势，我们需要使用ratings中用户的原始评价来进行分析。

通过userId唯一标识每个客户，客户的评价时间包含年、月、日、时、分、秒，这一字段便可提炼三种信息：用户评价时的年份、季节、时段（白天或是黑夜），通过这一信息丰富的字段，我们便可以很好地进行分析。

针对不同维度，我们采取了不同的分析尺度，以保证分析的多样化。

S2.3.1.1 数据分析可视化结果1：

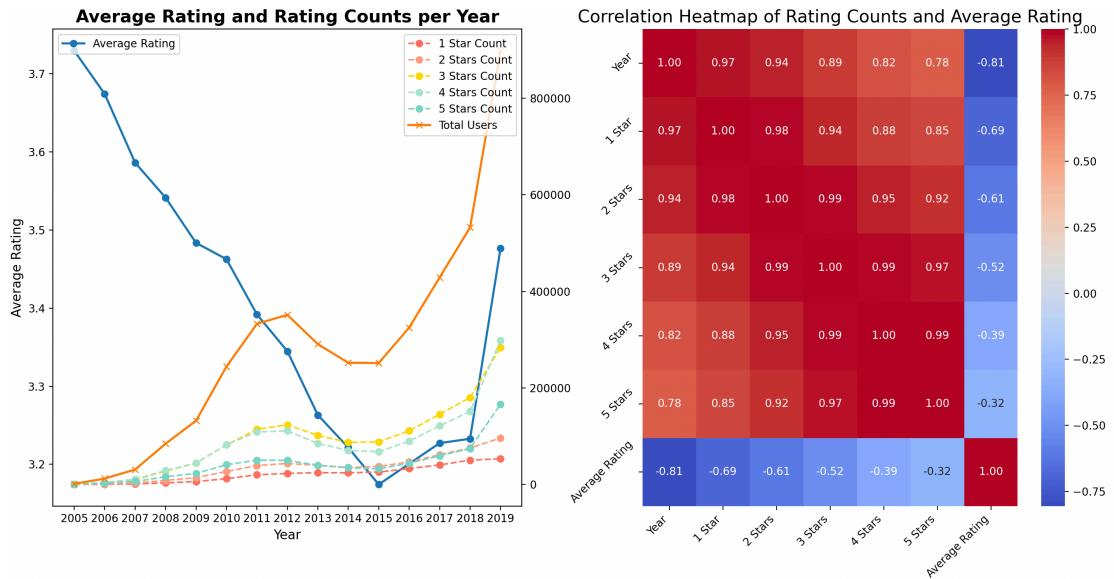


Fig2.1 年份评分曲线图/年份-均分-评价数量相关系数热力图

S2.3.1.2 结论1:

从图表中可以观察到，近年来，电影评分呈现明显的波动，特别是从2000年开始，评分数量逐年增长。年份与均分以及各评价数量之间存在强相关性，总体上，高评分（4星和5星）电影数量不断增加，而低评分电影（1星和2星）数量相对较少。随着观众对电影质量的要求逐渐提高，评分趋向集中于高分电影。

在2010年之后，随着豆瓣等平台的用户激增，评分人数的大幅增长，电影评分分布趋于极端，即更多的电影获得极高或极低的评分，这可能与电影的商业化和娱乐化程度提高以及观众口味多样化有关。

S2.3.2.1 数据分析可视化结果2:

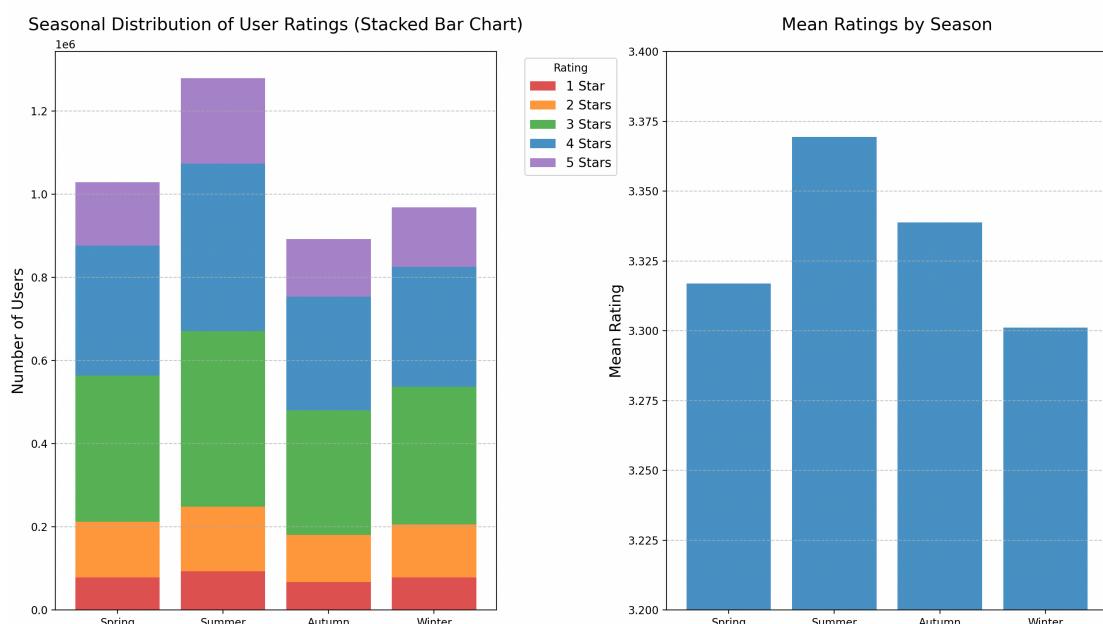


Fig2.2 季节评分柱状堆叠图/季节评分均分柱状图

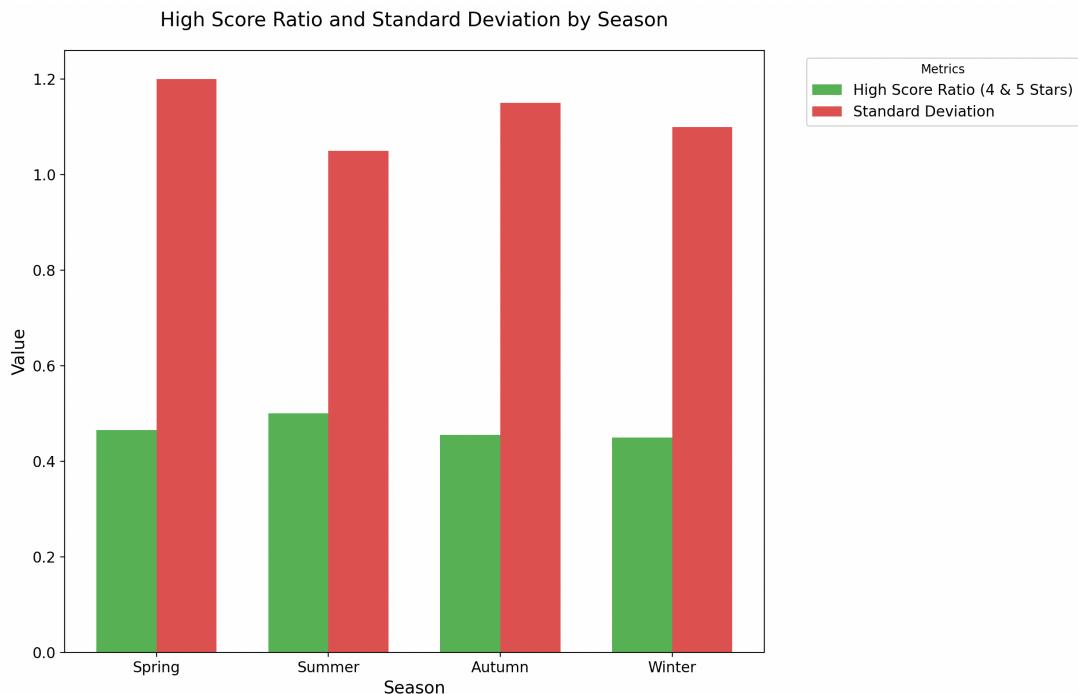


Fig2.3 季节高评分率与标准差双柱状图

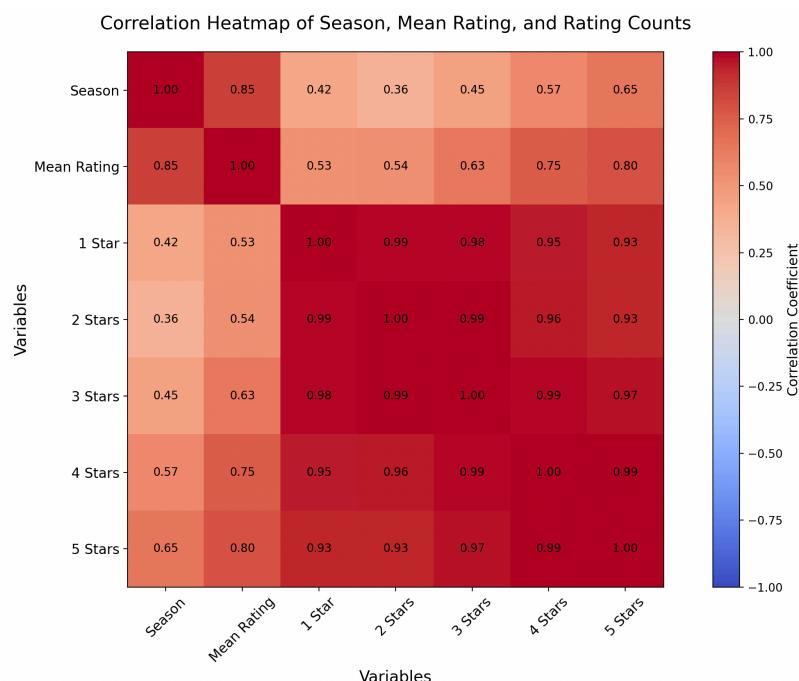


Fig2.4 季节-均分-评分数量相关系数热力图

S2.3.2.2 结论2:

从季节性分析中，夏季的评分数量最多，且高评分电影的比例也较大。这可能与假期观影量的增加有关。春季和秋季的评分相对较少，而冬季的评分则略高于秋季。这种季节性的波动反映了观众在不同季节的观影习惯和电影需求变化，特别是假期对电影观看行为的影响。高评分电影在各个季节中都占据较大比例，尤其是夏季，表明在观众有更多空闲时间时，他们倾向于观看和评价更高质量的电影。

S2.3.3.1 数据分析可视化结果3:

Correlation Heatmap of Hour and Rating Distributions

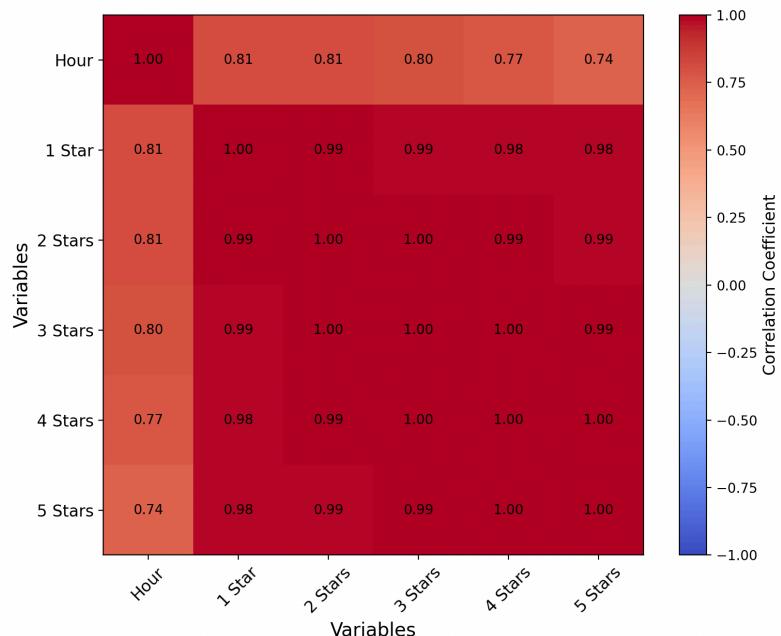


Fig2.5 时段-评价人数相关系数热力图

Hourly Distribution of User Ratings (Stacked Bar Chart)

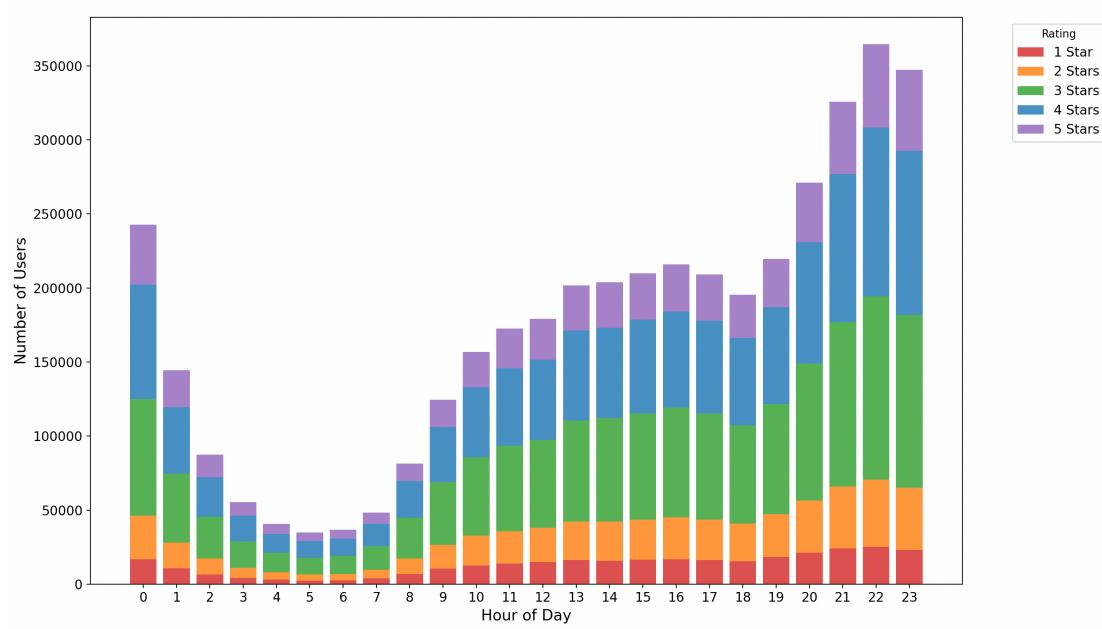


Fig2.6 小时评分分布堆叠柱状图

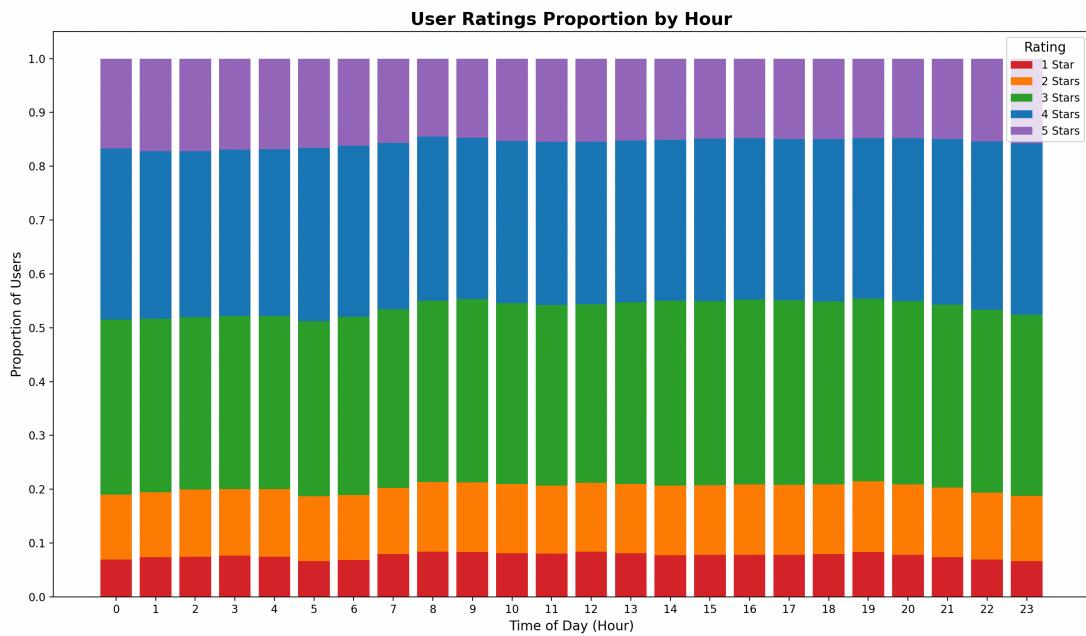


Fig2.7 小时评分占比堆叠柱状图

S2.3.3.2 结论3:

从小时分布图来看，评分的数量在一天的不同时间段内呈现出不同的趋势。深夜（00:00-03:00）时段的评分人数显著减少，而在晚上（18:00-22:00）期间，评分的数量达到峰值。尤其是在晚上的时段，5星电影的评分占比较高，说明此时段的观众可能偏向较高的评分。与此相对，凌晨和早晨时段的评分则更多集中在低评分（1星和2星）上，可能与观众的观看状态和心理状态有关。

这种时段分布趋势反映了观众观看电影的时段偏好以及不同时间段内观众的情感状态差异。

S2.3.4 总结:

根据相关性热力图，评分时间和评分之间有较强的相关性。比如，深夜的低评分与白天的高评分之间呈现出较强的负相关性，表明观众在不同时间段的情感状态差异较大。评分的季节性变化也与平均评分之间有着较高的相关性，尤其是夏季和冬季的评分高峰，反映了观众在这些季节对电影的偏好和评分差异。

电影的评分分布与观众的情感和心理状态密切相关，不同时间段内的评分差异反映了观众在日常生活中的观影习惯。高评分电影在夏季和晚间时段尤为集中，反映了假期和空闲时间对电影观看行为的影响。评分数量的急剧增加和电影评分的集中化趋势，表明了电影产业逐渐向娱乐化、商业化发展，观众对电影质量的需求不断提升。

S3.评分-区域关系分析

S3.1 涉及字段:

`movies_DOUBAN_Score, movies_region`

S3.2 分析目的与方法:

对电影的评分有了一个总体趋势的把握之后，我们就可以针对电影的单个特征来进行一些分析。首先我们想先了解一下不同国家不同地区的电影拍摄质量是否存在显著差异，因此我们提取了所有电影的评分和地区进行分析，为了量化不同国家的电影质量，我们采用如下公式来计算不同国家的电影质量：

$$\text{综合评分} = 0.5 \times \text{均分} + 0.3 \times \log(\text{电影数量}) + 0.2 \times (10 - \text{标准差}) + \text{高分电影比例得分}$$

其中：

- 高分电影比例得分的计算规则：

$$\text{高分电影比例得分} = \begin{cases} 0.2 \times \text{高分电影比例} \times 100 & \text{如果电影数量} \geq 10 \\ 0 & \text{如果电影数量} < 10 \end{cases}$$

这样设计的思路在于给均分一个比较重的权重，同时也不忽略其他因素，对于一个国家来说，其电影拍摄的质量的稳定性也是重要的因素，因而引入了标准差相关的项，电影数量则体现国家的电影工业能力，且其跨度较大，从1到几千兼而有之，因此采用对数降低跨度，为了防止部分国家只拍摄一部电影就拿了高分，导致高分率为100，因此在电影数量上对高分电影比例得分作限制。

S3.3 数据分析可视化结果:

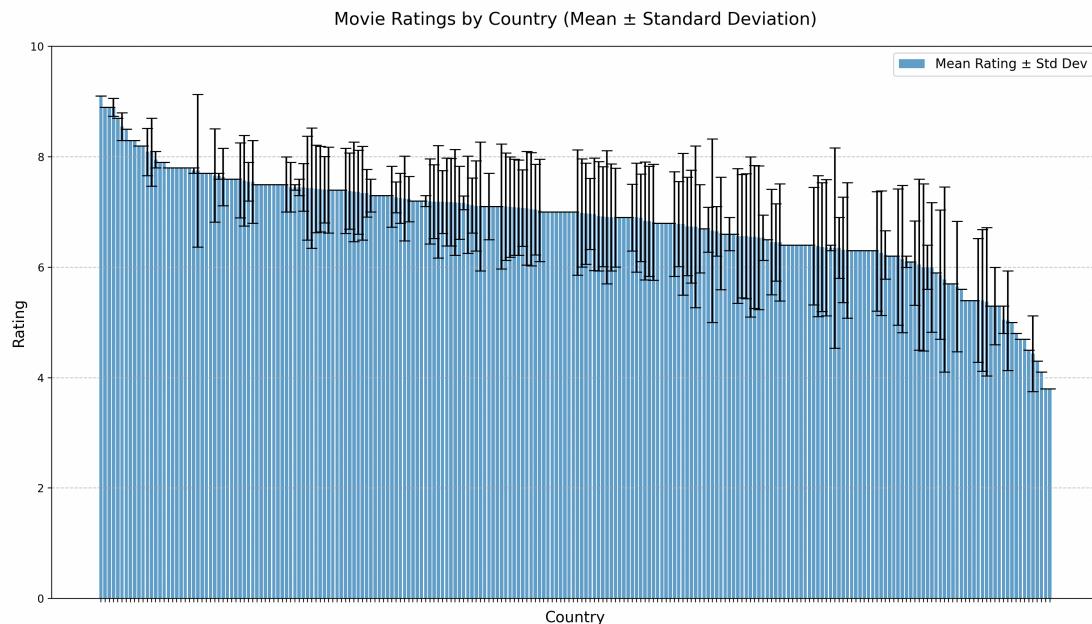


Fig3.1 不同国家的“原始均分-标准差”柱状图

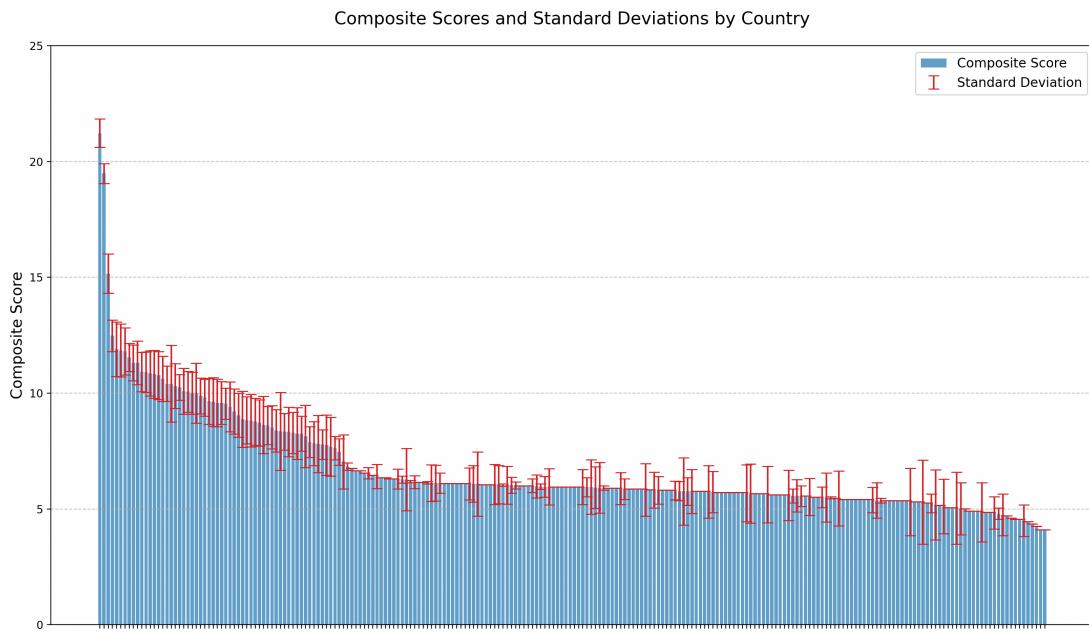


Fig3.2 不同国家的“电影质量评分-标准差”柱状图

国家	均分	标准差	电影数量	高分率	复合评分
苏联	8.08	0.61	38	71.05%	21.23
捷克斯洛伐克	8.09	0.43	19	63.16%	19.49
南斯拉夫	7.66	0.85	21	42.86%	15.16
伊朗	7.58	0.68	76	27.63%	12.48
英国	7.10	1.17	2231	21.34%	11.90

Table3.1 排名前五的国家的电影质量量化表

S3.4 结论：

根据提供的数据分析结果，可以得出如下结论：

高分电影的区域集中性：一些地区（如苏联、捷克斯洛伐克等）具有较高的电影评分和较高的高分电影占比，这些地区的电影质量普遍较高，能够吸引更多观众。

电影产业质量参差不齐：电影评分标准差较大的地区（如印度）显示出电影产业的质量波动较大，需要进一步提高制作水平和质量控制。

综合评分反映电影产业发展水平：综合得分较高的地区（如苏联、捷克斯洛伐克）代表了较为成熟且受欢迎的电影产业，而得分较低的地区则可能存在技术、创意和市场等方面的风险。

整体而言，S3分析为全球电影市场的区域性发展提供了可操作的指导，并有助于电影产业在全球范围内的文化交流与合作。

S4. 评分-类型关系分析

S4.1 涉及字段：

`movies_Movies_id, movies_DOUBAN_Score, movies_Genre,`
`ratings_Movie_id, ratings_Rating`

S4.2 分析目的与方法:

接下来我们想研究不同类型的电影是否对评分有影响，这一研究为理解电影的观众偏好提供了重要视角。通过考察各类电影的均分和评分分布，研究可以帮助电影制作方、平台运营者和市场营销人员更好地把握观众的偏好，调整电影的制作方向和市场推广策略。此外，这一分析也揭示了不同电影类型的受欢迎程度和观众评价的差异性，从而为电影产业的发展和观众需求的精准满足提供了数据支持。

针对不同类型的电影，统计其均分以及不同评价的数量，进行可视化分析，并结合排名

S4.3 数据可视化分析结果:

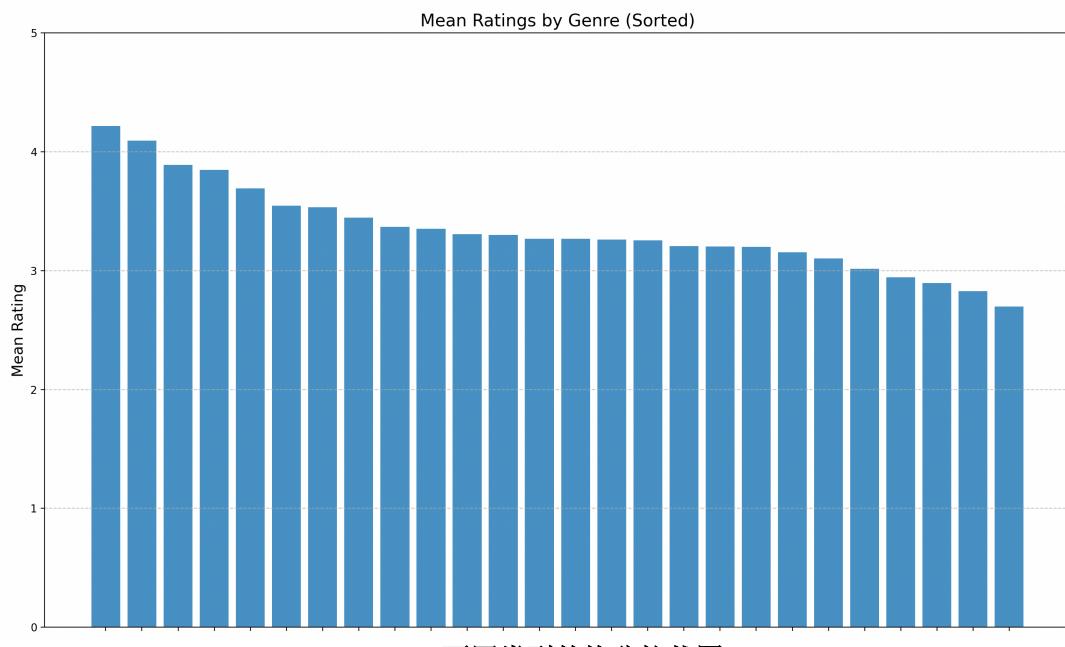


Fig4.1 不同类型的均分柱状图

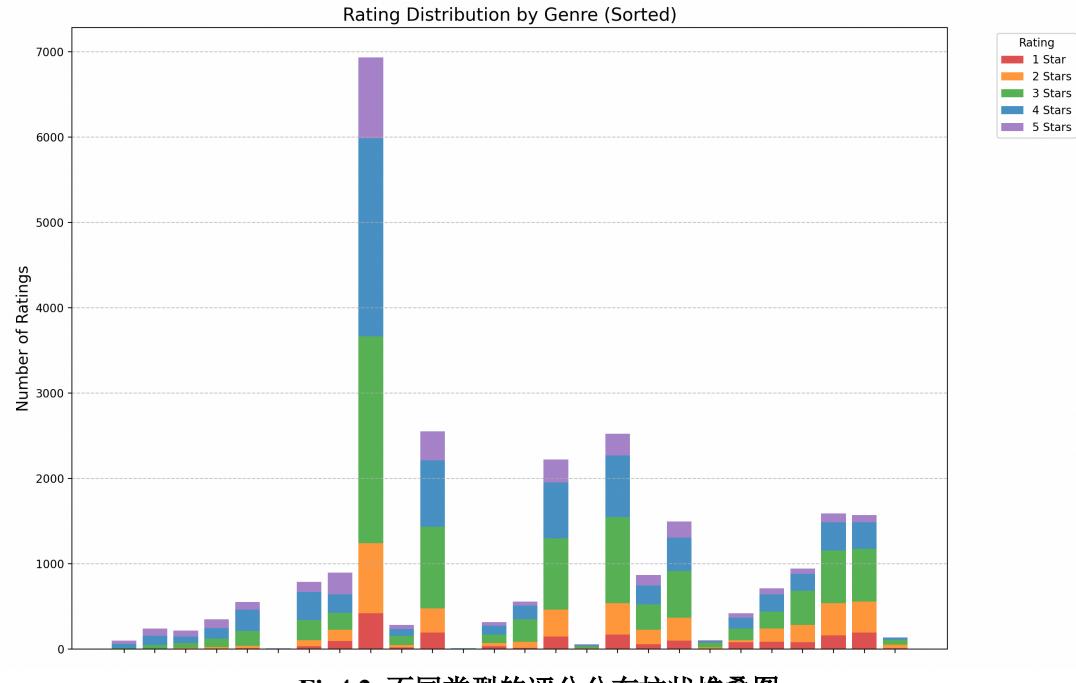


Fig4.2 不同类型的评分分布柱状堆叠图

S4.4 结论：

电影类型与评分之间的关系明显。在所有类型中，**古装**和**武侠**电影的均分最高，分别为4.22和4.10，显示这些类型的电影通常能够获得较高的观众评价，可能与其特殊的文化背景和受众基础相关。相比之下，**灾难**和**惊悚**电影的均分最低，分别为2.70和2.83，这表明这些类型的电影可能较为极端，观众的评价分布存在较大差异。

从评分数量来看，**剧情**和**喜剧**电影的评分数最多，表明这两类电影是观众最常观看和评价的类型，这也反映了它们在全球电影市场中的主流地位。其他类型如**爱情**、**动作**和**科幻**电影，虽然评分数量较大，但整体评分较低，可能与这些类型电影的创作质量、观众期望和市场竞争程度有关。

总体来说，不同电影类型的评分和分布情况揭示了观众在不同类型电影中的偏好差异，为电影创作、市场推广和内容分发提供了重要的决策依据。

S5. 评分-时长关系分析

S5.1 涉及字段：

`movies_Mins`, `movies_DOUBAN_Score`

S5.2 分析目的与方法：

接下来我们想研究不同时长的电影是否对评分有影响，这一研究能够揭示观众对电影时长的偏好，以及时长是否对观众的评分产生影响。这对电影制作方和发行方在决策时长、调整影片结构以及优化观众体验具有重要的实际指导意义。

针对不同时长的电影，统计其均分以及不同评价的数量，进行可视化分析，并结合排名，值得注意的是，由于电影时长跨度过大（短则几分钟，长则10h，而大部分电影时长1-5h），为了防止可视化分析时长尾数据对图片美观度的影响，因此在时间上采用对数尺度

S5.3 数据分析可视化结果：

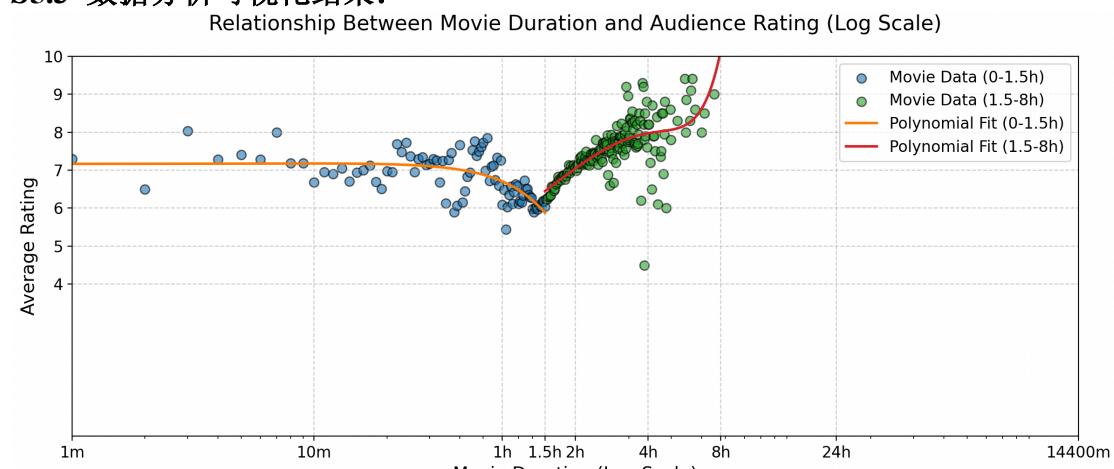


Fig5.1 时长-均分散点图

S5.4 结论：

根据图表中的数据，电影时长与观众评分之间呈现出非线性关系。对于较短时长的电影（0-1.5小时），平均评分趋于稳定，大部分电影的评分较为集中

在7分左右，这些电影通常较为精炼，符合观众的偏好。对于时长较长的电影（1.5小时至8小时），评分出现较大的波动，尤其是在1.5到4小时区间，部分电影获得了较高的评分，而一些电影的评分则较低。这表明观众可能对时长过长的电影产生疲劳感或期望过高，从而影响其评价。

总体来看，较短时长的电影倾向于获得更高的评分和一致的观众评价，而较长时长的电影则面临更大的观众分歧，制片方需要在时长控制和影片质量之间找到合适的平衡点。

S6.评分-导演关系分析

S6.1 涉及字段：

`movies_Director`, `movies_DOUBAN_Score`

S6.2 分析目的与方法：

接下来我们想研究不同导演执导是否对评分有影响，研究揭示了不同导演执导的电影是否对评分产生影响，能够为电影产业中的导演选拔、电影制作和市场推广提供参考。通过比较导演的平均评分，研究有助于识别哪些导演能够持续吸引高评分观众，从而为电影制作方提供关于导演选择的有力依据。这也有助于理解导演在电影历史发展中的作用以及如何通过导演的风格影响观众偏好。

通过统计不同导演的执导电影的均分，来进行可视化分析

S6.3 数据可视化分析结果：

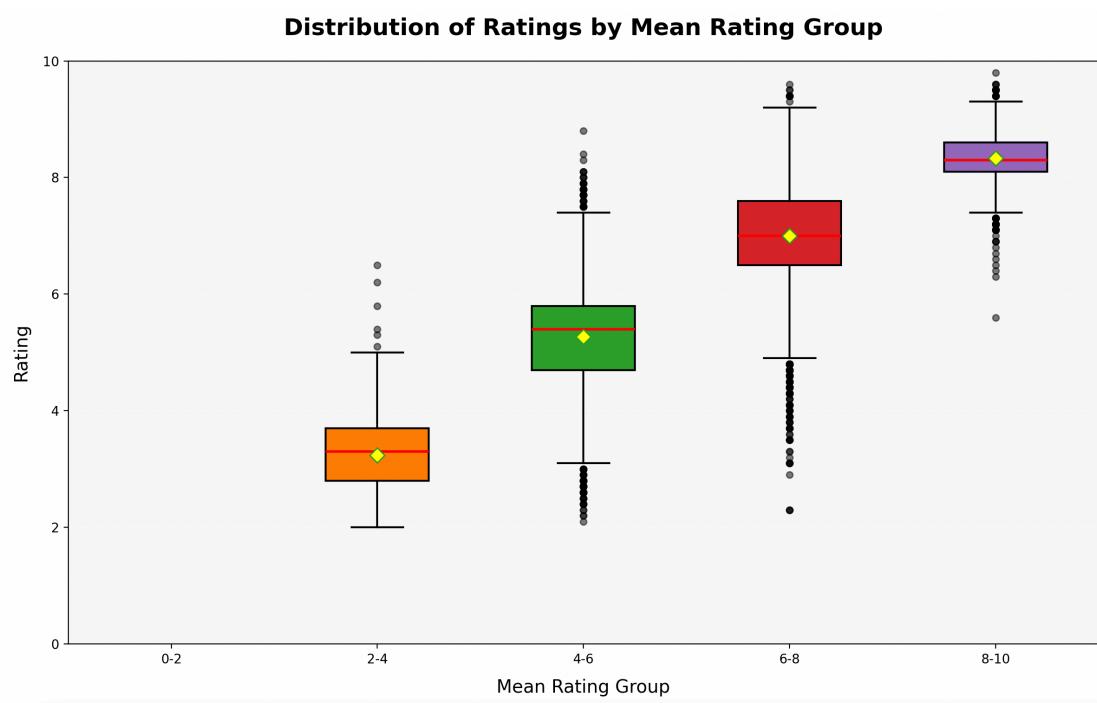


Fig6.1 不同导演的均分分布情况

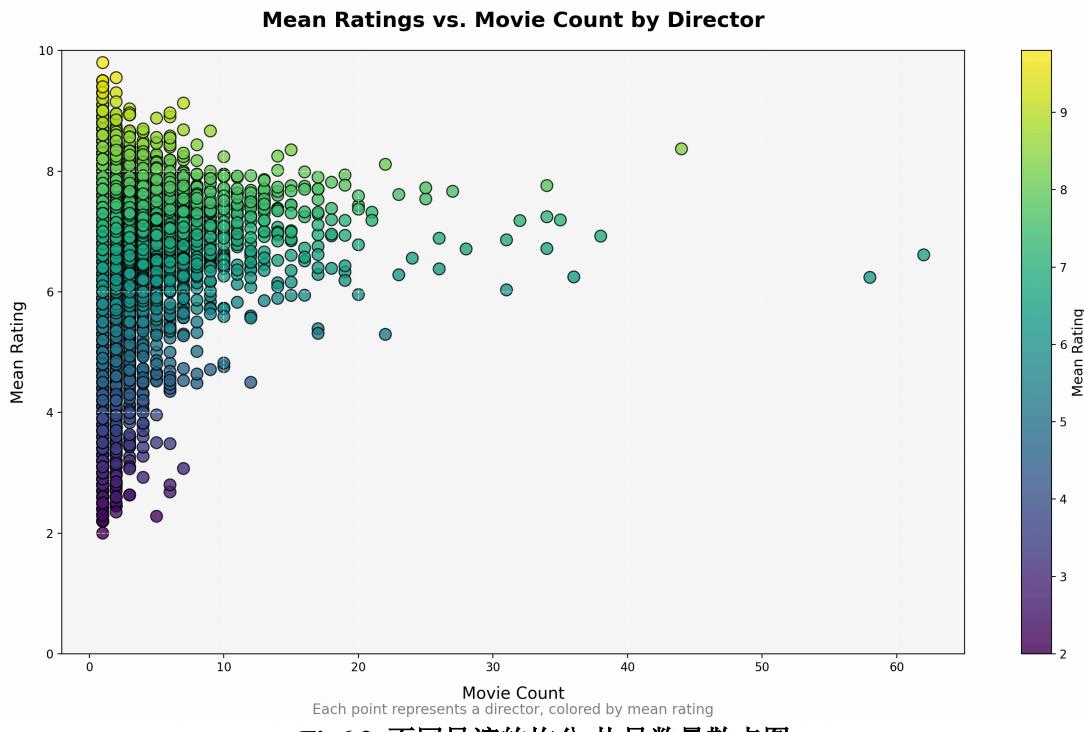


Fig6.2 不同导演的均分-执导数量散点图

S6.4 结论：

根据提供的图表数据，可以观察到导演执导的电影评分存在显著差异。对于那些执导电影数目较多的导演（如中等评分组和高评分组），其电影通常获得较高的评分，尤其是在**6-8分**的区间。这表明这些导演在作品质量和观众评价上积累了较高的信任度。而在评分较低的导演群体中（如**2-4分**评分组），执导的电影评分波动较大，显示出这些导演的作品可能质量参差不齐，未能稳定吸引观众。

另外，从导演作品数量和评分的关系来看，导演数量较多的群体中，不一定所有导演的评分都处于高端，表明频繁执导可能导致作品质量的起伏。总体来看，拥有高评分的导演通常在市场中占有较高的影响力和观众基础，电影制片方可以通过合作这些导演来提高影片的市场表现和观众满意度。

S7. 用户打分习惯分析

S7.1 涉及字段：

ratings_userId, rating_Movie_id, movies_Rating

S7.2 分析目的与方法：

在分门别类地研究了电影的不同特征对评分的影响后，我们发现电影的诸多特征都会左右观众的判断，因此我们再次聚焦于观众，分析观众们的打分习惯，研究揭示了观众在评分过程中可能存在的偏好和规律。这一分析有助于理解观众如何在不同的评分区间内进行评价，从而为电影制作方和平台运营方提供有价值的参考，帮助他们更好地把握观众的评分趋势，调整电影内容和观众互动方式，以提升观影体验和满意度。

我们统计了所有用户给出1星-5星评价的数量，研究这些数量的分布情况。

S7.3 数据可视化分析结果：

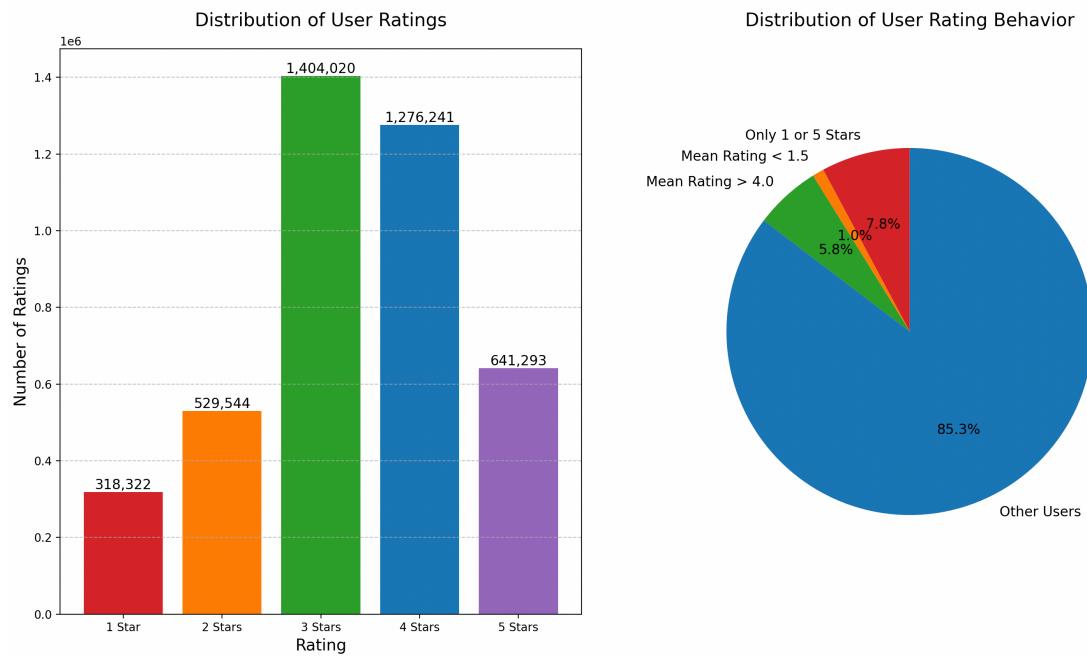


Fig7.1 用户评分分布柱状图/用户评分分布饼图
Distribution of User Rating Standard Deviations

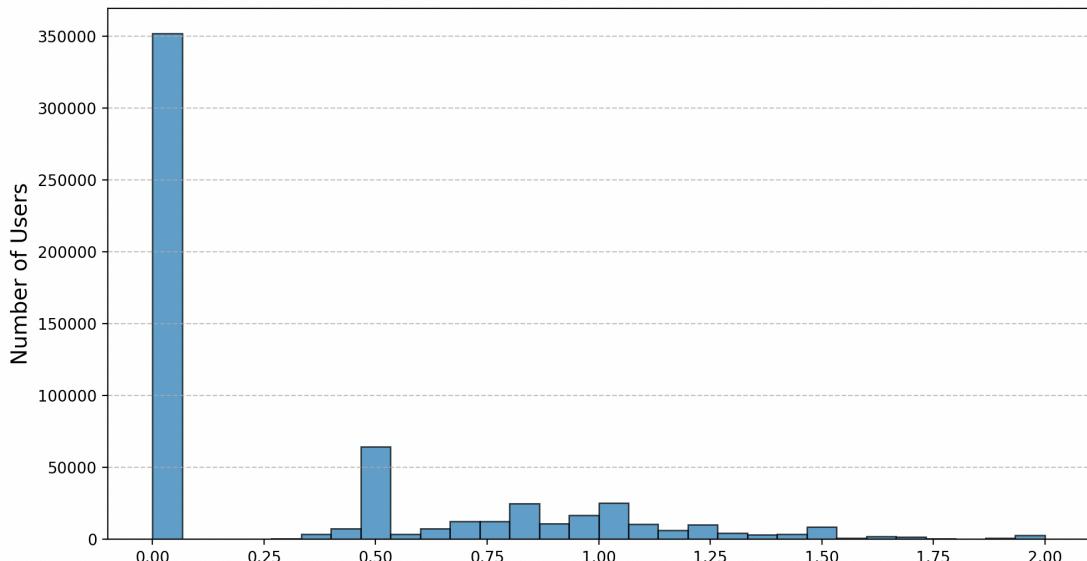


Fig7.2 用户评分标准差分布柱状图

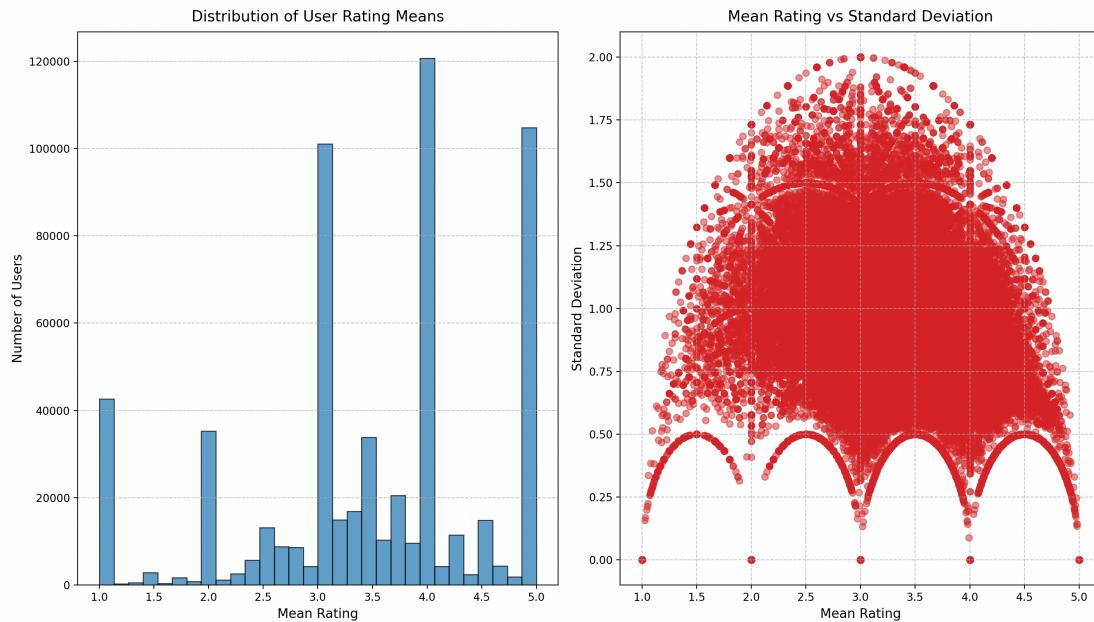


Fig7.3 用户均分分布柱状图/用户均分-评分标准差散点图

S7.4 结论：

根据图表数据，用户评分的分布呈现出明显的集中趋势。3分和4分的评分占据了最大的比例，显示出大多数观众的评分偏向中等和较高的区间。而极端评分（如1分和5分）则较少，且大部分观众的评分标准趋于一致，表明大多数观众倾向于在较为中立的评分范围内做出评价。

在评分标准差方面，大部分用户的评分标准差非常小，说明大部分观众的评分比较稳定，不容易受到个别情感波动或电影质量的过度影响。也有一小部分用户的评分波动较大，表现出更为极端的评价倾向。

这些数据反映了观众评分的集中化特征，说明绝大多数用户在评分时较为保守，偏向中等评分，而极端评分的观众群体相对较少。这一分析为电影行业在观众群体分析、观影体验优化以及电影评价体系的设计提供了重要依据。

S8. 冷门佳作分析

S8.1 涉及字段：

`movies_Name, movies_DOUBAN_Score, movies_DOUBAN_Votes`

S8.2 分析目的与方法：

经过上面的研究，我们已经对电影评分发展的历史趋势、电影的特征对评分的影响、观众打分偏好有了较为深入的了解，此时，我们想去探究一下是否存在那些质量上乘，却鲜为人知的电影，这个研究帮助理解观众偏好与电影评价之间的关系，以及探索可能影响电影流行度的因素。

我们针对一部电影的均分以及评分人数，设计了一个指标，称为冷门指数，冷门指数的计算方法如下：

$$\text{冷门指数} = \left(\frac{\text{均分}}{\text{最大均分}} \right) \times \left(1 - \frac{\log(\text{评分人数})}{\log(\text{最大评分人数})} \right) \times \left(1 - \frac{\text{评分人数}}{\text{评分人数阈值}} \right) \times 100$$

这个乘式的第一项是对均分进行归一化，将均分归结到0-1范围，衡量了评分高低，乘式的第二项衡量冷门程度，用对数平滑了跨度过大的评分人数，乘式的第三项设计了惩罚因子，防止评分人数过少缺乏客观性，当评分人数大于评分阈值时，第三项乘式取1

S8.3 数据可视化分析结果：

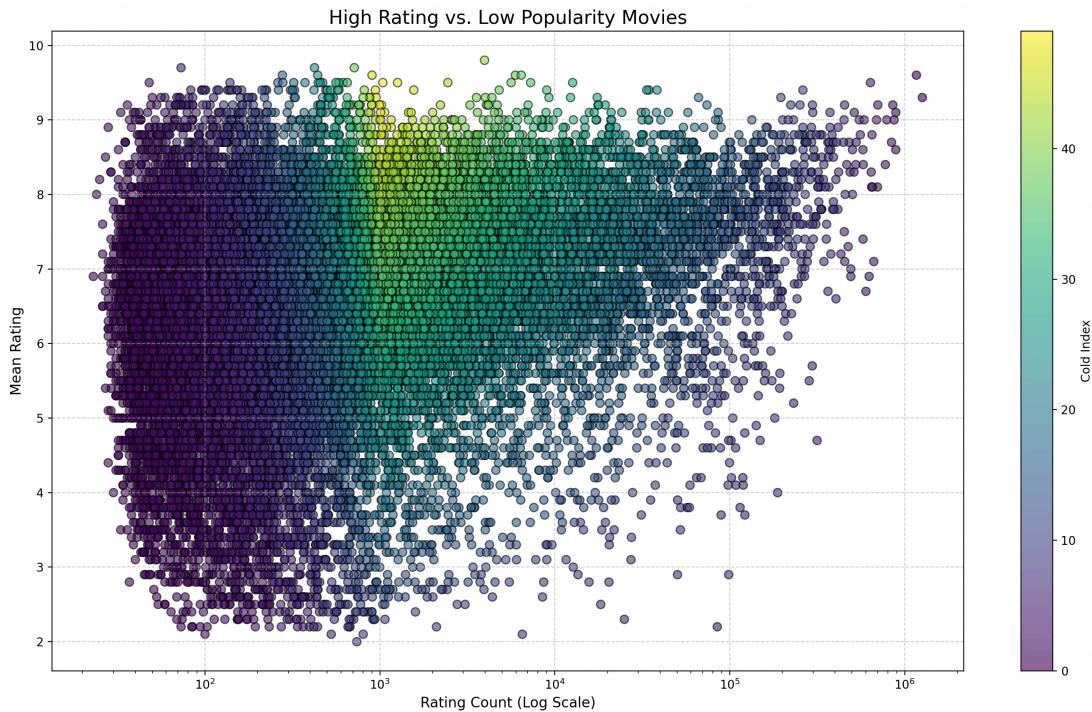


Fig8.1 高分冷门程度散点图

S8.4 结论：

根据数据分析结果图，高分冷门电影的存在表明，尽管某些电影在少数观众中获得了极高的评价，但由于各种原因（如宣传不足、题材小众等），它们未能获得广泛的关注。这提示我们在电影评价和推荐系统中，除了考虑大众喜好外，还应关注那些可能被忽视的优质作品，以更全面地反映电影的艺术价值和多样性。

五、实训总结

5.1 大数据环境配置总结

环境配置过程整体顺利，只遇到过一个问题，就是尝试在idea中使用python进行可视化时，错误地将原来的SDK从Maven切换成Python，导致scala程序无法运行，解决方式是将SDK切换回Maven，且在解释器栏位中增加python解释器部分

5.2 项目设计与实现总结

5.3 实训收获

通过本次基于 Hadoop 和 Spark 的电影历史发展与观众偏好研究项目，我在技术能力、项目实践等方面获得了显著的提升。以下是本次实训的主要收获：

1. 技术能力的提升

我深入理解了 Hadoop HDFS 的分布式存储原理和 Spark 的分布式计算机机制，能够熟练运用这些技术处理海量数据。掌握了数据清洗、转换、统计分析和模式识别的基本方法，能够使用 Spark 和 Python 库（如 pandas、numpy）高效处理复杂数据。学会了使用 matplotlib、seaborn 等工具将数据分析结果以直观的图表形式呈现，提升了数据表达能力。

2. 项目实践经验的积累

从数据采集、存储、处理到分析与可视化，参与了项目的全流程，积累了大数据项目的实战经验。在项目中遇到了数据缺失等问题，通过查阅文档、调试代码和优化配置，成功解决了这些问题。学会了从业务需求出发，设计合理的技术方案，并通过数据分析为电影产业提供有价值的洞察。

3. 行业认知的深化

通过对电影历史数据与观众偏好的分析，深入了解了电影产业的发展规律与市场需求，我认识到大数据技术在文化研究与市场分析中的巨大潜力，为未来在相关领域的职业发展奠定了基础。