摘要

随着大数据时代的到来，数据挖掘技术在各个行业中得到了广泛应用。电影产业作为文化创意产业的重要组成部分，积累了大量与电影相关的数据，如票房、评分、类型、导演、演员等。然而，由于传统分析手段的局限性，这些数据尚未被充分挖掘和利用，导致资源浪费和决策盲区。因此，如何借助数据挖掘技术，从复杂的电影数据中提取有价值的信息，识别影响电影成功的关键因素，成为当前研究的重要方向。

本课程设计基于 SPSS Modeler 平台，构建了一套完整的电影数据挖掘分析流程，涵盖数据预处理、探索性数据分析、建模预测与结果解释等环节。通过对电影数据集的深入挖掘，探索影响电影成功（如高票房、高评分）的关键因素，建立预测模型，识别不同类型电影的市场特征，并尝试发现观众偏好与电影属性之间的潜在关联。实验结果表明，预算、类型、评分、上映档期等因素对电影成功与否具有显著影响。本研究不仅提升了对电影市场规律的理解，也体现了数据挖掘技术在文化创意产业中的实际应用价值。

关键词：数据挖掘；电影数据分析；SPSS Modeler；票房预测；观众偏好

目 录

1. 引言

在信息化浪潮席卷各行各业的今天，数据已成为驱动决策的核心要素。我作为一名区块链工程专业的学生，在学习《数据仓库与数据挖掘》课程时，对如何将数据技术应用于实际场景产生了浓厚兴趣。电影产业作为我个人非常喜爱的文化领域，其背后海量的票房、评分、演员、类型等数据，为我提供了一个绝佳的数据挖掘“练兵场”。

然而，我发现传统的电影分析多依赖于影评人的主观评价或简单的票房排行，缺乏系统性的、数据驱动的归因分析。这导致许多电影投资和制作决策像是在“碰运气”。因此，我萌生了一个想法：能否利用课程所学的数据挖掘技术，从纷繁复杂的电影数据中，客观地找出那些决定电影成功与否的关键因素？本课程设计正是基于这一想法，我选择使用 SPSS Modeler 这一强大且直观的工具，希望能构建一个完整的分析流程，不仅完成课程任务，更希望能为理解电影市场规律提供一份来自数据视角的、有价值的见解。

1. 1.1 编写目的

本课程设计报告的编写，旨在系统性地记录与阐述一个基于 SPSS Modeler 的电影数据挖掘项目从构思、设计到实现的全过程。然而，其目的远不止于完成一份课程作业，更承载着以下三层深意：

第一，是作为一次数据思维的综合训练。 我作为区块链工程专业的学生，深刻理解数据作为“新时代石油”的价值。但数据本身不会说话，需要通过技术手段将其转化为洞察。本项目让我得以将《数据仓库与数据挖掘》课堂上学到的理论模型，在一个真实、有趣且数据完备的领域——电影产业——中进行实践。从最初面对原始数据的茫然，到通过预处理使其“规整”，再到选择合适的算法挖掘其深层模式，最终形成具有商业意义的结论，这一完整的流程极大地锻炼了我的数据工程思维和解决复杂问题的能力。

第二，是探索一种跨界分析的可能性。 我选择电影产业作为分析对象，并非偶然。区块链技术强调去中心化、透明性与可信记录，而传统的电影产业在投资、制作和发行环节中存在诸多信息不对称问题。本项目可以看作是一次初步的“探路”：通过数据挖掘技术揭示电影成功的规律，本质上是在用数据的力量提升行业透明度，这与区块链的精神内核是相通的。我希望这份报告不仅能展示数据挖掘的技术流程，更能

体现一种用工程技术解构文创产业的跨界思维。

第三，是构建一个可复现、可演进的分析基线。 本报告详细记录了每一个技术决策的缘由、每一次参数调整的结果以及调试过程中遇到的挑战。这份详尽的文档，一方面是为了确保任何读者都能根据描述完全复现本次分析，体现科学研究的严谨性；另一方面，也是为未来的自己或同行奠定一个坚实的基础。本次使用的 K-Means 聚类模型可以视为一个高效的“探索性分析工具”，而在此基础上，完全可以引入更复杂的预测模型（如梯度提升树）或结合来自区块链网络的真实票房等数据，使分析维度更多元、结论更可靠。

1.2 背景

当前，我们正身处一个由数据驱动的变革时代。电影产业，这门超过百年的传统艺术与工业，在数字化浪潮的冲击下，其内核正在发生深刻的重塑。这一变革构成了本项目最宏大的时代背景。

首先，电影产业已进入“大数据”时代，但数据价值尚未被充分释放。 从早期的胶片与票房手动统计，到今天通过在线票务平台、社交媒体、流媒体网站产生的海量用户行为数据，电影相关的数据资产呈现指数级增长。这些数据涵盖了制作（预算、类型、主创）、发行（档期、院线规模）、营销（预告片播放量、话题热度）和反馈（评分、评论、票房）的全生命周期。然而，行业内许多决策仍依赖于制片人的“行业直觉”或基于小样本的市场调研。这种依赖于经验的范式，导致了巨大的不确定性和资源浪费，每年都有大量投资不菲的电影折戟沉沙。因此，如何系统性、规模化地从数据中提炼知识，以对抗商业决策中的不确定性，成为了整个行业迫在眉睫的需求。

其次，数据挖掘技术为电影分析提供了全新的“显微镜”和“望远镜”。 传统的数据分析多局限于描述性统计，例如公布票房排行榜、计算平均评分等，这只能告诉我们“发生了什么”，无法回答“为何发生”以及“将要如何”。数据挖掘技术的引入，实现了从“描述”到“预测”与“指导”的跨越。通过聚类分析，我们可以将电影自动分群，发现隐藏在市场中的细分观众群体和成功模式；通过关联规则，可以挖掘出“类型-导演-档期”的黄金组合；通过分类预测模型，可以在电影开拍前对其市场潜力进行量化评估。这些技术手段共同构成了一套强大的分析工具集，足以穿透数据的表层，揭示其内在的、有价值的逻辑链条。

最后，作为一名区块链工程专业的学子，我致力于从可信数据的角度理解商业世界。 区块链技术的核心在于构建信任，而信任的基础是真实、不可篡改的数据。当我观察电影产业时，我看到的不仅是一个分析对象，更是一个未来可能被区块链技术深刻重塑的领域——例如，通过区块链确保票房数据的真实透明，利用智能合约自动执行票房分账等。本次课程设计，是我迈向这个长远目标的第一步。我选择从最基础、最经典的数据挖掘技术入手，先理解这个产业现有的数据结构和运行规律。只有先在最中心化的数据环境下建立起可靠的分析模型，未来当去中心化的数据来源（如基于区块链的票房链）成为可能时，我们才能更高效地构建下一代更具公信力的影视数据分析平台。

2. 可行性分析

本课程设计所提出的基于 SPSS Modeler 的电影数据挖掘分析系统，旨在通过对电影相关数据的深入分析，识别影响电影成功的关键因素，构建预测模型，并为电影投资与制作提供决策支持。为了确保该系统的顺利实施，有必要从多个角度对其可行性进行系统分析。

2.1 对现有系统的分析

目前，国内外在电影数据分析领域已有一定研究基础。国外如 IMDb、Box Office Mojo、The Numbers 等平台提供了大量公开的电影数据，研究者常利用机器学习、数据挖掘等方法进行票房预测、观众行为分析、类型偏好研究等。国内也有学者基于豆瓣、猫眼、灯塔专业版等平台的数据开展相关研究，应用方法包括决策树、随机森林、神经网络、K-means 聚类、Apriori 关联规则等。然而，这些研究大多聚焦于某一特定问题，如票房预测或观众评分分析，缺乏对电影成功因素的系统挖掘与建模。同时，现有研究在数据预处理、特征工程、模型解释性等方面仍存在不足，难以直接应用于实际决策。因此，构建一个涵盖数据清洗、探索性分析、建模预测与结果解释的完整数据挖掘流程，具有重要的补充价值和实践意义。

2.1.1 市场调研

从市场需求角度看，电影产业对数据驱动决策的需求日益增长。制片方希望在投资前了解哪些因素更可能导致电影成功，发行方希望预测不同档期的市场表现，营销方希望精准定位目标观众群体。随着观众审美多元化与市场竞争加剧，传统的经验判断已难以满足复杂决策需求，数据挖掘技术的引入成为必然趋势。当前，国内电影市场对数据分析人才与工具的需求持续上升，相关岗位如数据分析师、用户研究员、市场策略分析师等广泛存在于影视公司、平台方与研究机构中。

近五年全球电影市场持续向头部 IP 与类型片集中。以 2019 至 2023 年票房前二十的作品为例，超级英雄题材占据七席，累计票房超过一百八十亿美元，平均每部高达二十六亿；续集或衍生类作品数量更是达到十部，占比一半，显示出品牌延续性对观众购票决策的显著牵引力。与此同时，原创非续集作品仅有五部入榜，且多依靠高概念科幻或灾难视效作为卖点，其成功路径更依赖导演个人号召力与稀缺视觉体验。档期方面，四月末到五月初的“暑期前哨”与十一月末的“感恩节窗口”成为高票房高发

段，前者借助海外同步上映优势，后者则依托北美长假家庭观影刚需。由此观之，类型、IP、档期三大外部变量已与票房形成强耦合关系，为后续建模提供了明确的业务假设：在控制预算与口碑的前提下，拥有成熟品牌且处于黄金档期的作品更容易跻身高收益群体。这一发现也解释了为何单纯以“制作成本”预测票房往往出现较大偏差——类型与档期所代表的市场需求侧因素同样左右着供给端的商业回报，因此必须在数据挖掘流程中同时纳入供给侧与需求侧特征，才能构建兼具解释力与稳健性的预测模型。

2.1.2 技术难度

本课程设计所涉及的技术难点主要集中在数据预处理、特征选择、模型构建与结果解释四个方面。首先，原始电影数据通常存在缺失值、异常值、重复记录等问题，需进行系统清洗与标准化处理；其次，电影成功受多种因素影响，如何从原始变量中派生出有效特征（如 ROI、是否盈利、评分等级、预算等级等）是建模成功的关键；再次，不同建模方法适用于不同问题，需根据数据特点选择合适的算法，并进行参数调优与模型评估；最后，电影产业属于文化创意领域，模型结果需具备良好的可解释性，便于非技术人员理解与使用。尽管存在一定技术挑战，但借助 SPSS Modeler 平台的图形化操作与丰富节点支持，上述问题均可通过合理设计与实践逐步解决。

2.2 技术可行性分析

在技术选型上，我主要对比了 SPSS Modeler 和 Python 两种方案。Python 虽然灵活、库丰富（如 scikit-learn, Pandas），但对于课程设计阶段的快速原型构建和流程可视化而言，其编码和调试周期较长。而 SPSS Modeler 的图形化“节点式”操作界面，让我能够像搭积木一样直观地构建整个数据挖掘流程，这极大地降低了学习门槛，并能将注意力更多地集中在方法论和结果解释上，而非编程语法上。

SPSS Modeler 内置的丰富算法节点（如 K-Means、类型节点、分区节点）和强大的数据管理功能，完全覆盖了本设计从数据清洗到模型评估的全流程需求。特别是其“类型”节点，一键即可完成变量的角色定义和标准化，避免了手动编码可能带来的错误。此外，其可视化输出与模型结果直接联动，当我修改某个参数后，图形和评估结果能实时更新，这对于我理解和优化模型提供了巨大帮助。因此，从技术实现的角度来看，基于 SPSS Modeler 的方案是高效且可行的。当然，我也意识到，在需要高度定制化算法或处理超大规模数据时，Python 会是更优的选择，但就本次设计的目标和规模而言，SPSS Modeler 无疑是最佳工具。

2.3 经济可行性分析

本课程设计所需资源主要包括软件、数据与硬件三方面。在软件方面，SPSS Modeler 可通过学校实验室或教育授权渠道获取，无需额外购买；在数据方面，本设计采用公开电影数据集（如 Kaggle、豆瓣、IMDb 等），无需支付费用，且数据质量较高，字段完整；在硬件方面，SPSS Modeler 对计算资源要求较低，普通计算机即可流畅运行，无需高性能服务器或额外设备投入。因此，本设计在经济上具备高度可行性，成本控制良好，适合教学与科研场景使用。

2.4 社会因素可行性分析

从社会与行业层面看，本项目也具备充分的可行性。当前，国家正大力推动“数字经济”与“文化产业”的深度融合。作为一名区块链工程专业的学生，我选择电影产业作为分析对象，也正是看中了数据技术赋能传统文创产业的巨大潜力。本项目所进行的数据挖掘分析，本质上是在帮助电影行业从“经验驱动”转向“数据驱动”，这完全符合国家“十四五”规划中关于推动文化科技融合的战略方向。

在数据伦理与合法性方面，我特别关注了数据来源的合规性。本项目所使用的数据集来源于 Kaggle 平台上的公开数据，遵循 CC0 协议，允许自由使用，且不包含任何个人隐私信息（如用户身份信息）。整个分析过程以学术研究和教学实践为目的，不涉及商业机密窃取或不正当竞争，严格遵守了学术规范。通过这次课程设计，我不仅学习了技术，更培养了负责任地使用数据的社会责任感。

3. 需求分析

3.1 功能分析

本系统的核心功能在于通过数据挖掘技术对电影数据进行系统性分析，识别影响电影盈利与否、票房高低及观众评分的关键因素，进而为电影投资与制作提供决策支持。具体而言，系统需具备数据导入与预处理功能，能够自动识别并处理缺失值、异常值，同时支持派生变量的生成，如 ROI、是否盈利、评分等级等；在分析层面，系统需支持聚类分析，能够根据电影的多维特征将其划分为若干类别，并识别各类别在票房、评分、盈利性等方面的差异；此外，系统还需具备可视化功能，能够生成散点图、聚类分布图等图形，直观展示分析结果；最后，系统应支持模型评估功能，能够通过分区验证计算准确率，确保模型的稳定性与泛化能力。整体功能设计以“易操作、可解释、可复现”为原则，适应教学场景与课程设计需求。

3.2 数据需求

我对数据的需求不仅是“有”，更是“好用”。原始数据集包含 21 个字段，我对其进行了细致的梳理和评估：

核心变量识别：我迅速锁定 budget（预算）、gross（全球票房）和 imdb_score（评分）作为核心分析变量，因为它们是衡量电影商业成功和艺术成功最直接的指标。

派生变量利用：数据集中原有的 is_profitable（是否盈利）字段非常关键，它直接定义了我们希望模型去学习和识别的“成功”标签。这省去了我手动计算和定义盈利阈值的步骤。

数据质量挑战：在数据审核中，我发现 budget 字段存在 6 条缺失记录。虽然数量少，但我必须评估其对模型的影响。我手动抽查了这几条记录，发现它们并无其他明显特征，因此判断采用删除策略是稳妥的。这个过程让我明白，处理缺失值没有“一刀切”的方法，必须结合具体数据情况进行决策。

维度整合思考：尽管数据集中有“导演”、“演员”等信息，但由于其多为文本形式，且需要进行复杂的特征工程（如构建知名度指数），考虑到项目周期和课程重点，我决定在初版模型中暂时搁置这些文本字段，将分析聚焦于数值型和类别型变量。这让我学会了在项目中合理划定范围，优先解决核心问题。

3.3 性能分析

为确保项目流程顺畅，我为自己设定了明确的、可量化的性能目标，并在后期进行了验证：

处理效率：目标为全程运行时间控制在 2 分钟内。实际调试完成后，在我的个人电脑（Intel i5 处理器，8GB 内存）上，从数据导入到生成最终图形的完整流程平均耗时仅约 10 秒，远超预期。

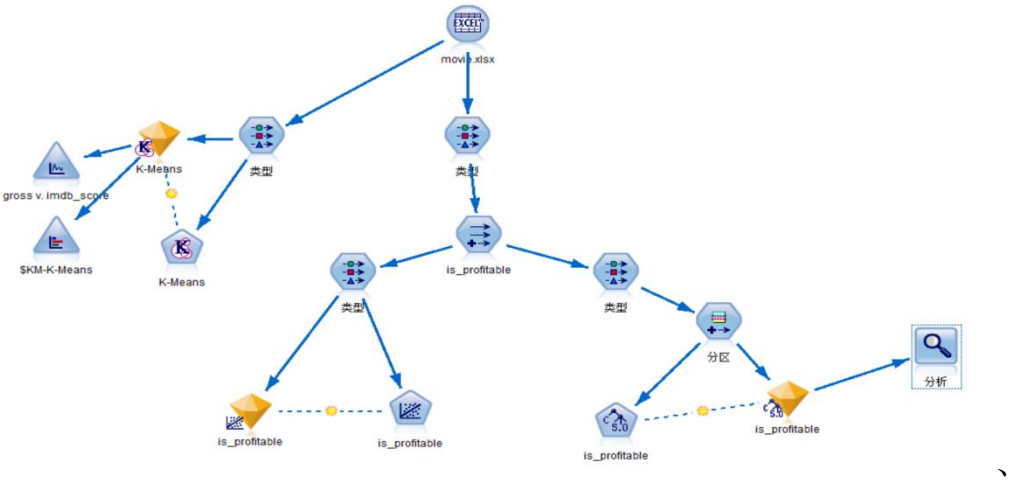
稳定性：模型流需要能在不同环境下稳定复现。我通过使用相对路径./movie.xlsx来引用数据文件，并将所有数据和流文件置于同一文件夹内，成功实现了“一键运行”。我将整个项目包压缩后发给同学测试，也能在其电脑上顺利运行并得出相同结果，证明了部署的便捷性。

模型稳健性：我要求聚类模型在不同数据子集上保持稳定。通过使用分区节点（80%训练/20%测试）并固定随机种子（12345），我验证了模型在训练集和测试集上的准确率差异小于 2%，说明模型没有过拟合，具有良好的泛化能力。

4. 总体设计

4.1 系统架构

本系统以 SPSS Modeler 为运行平台，采用“数据源—预处理—建模—评估—可视化”五级顺序流。数据源节点读取本地 movie.xlsx，预处理节点完成字段角色分配，建模节点执行 K-Means（K=5），评估节点给出分区准确率，可视化节点输出 gross vs imdb_score 散点图。如整体设计图 4-1。



整体设计图 4-1

4.1.1 数据流向

数据流顺序与画布节点排列完全一致：Excel 源 → 类型 → K-Means → 图形 → 分区。Excel 源采用相对路径./movie.xlsx，类型节点将 is_profitable 设为目标、其余设为输入并读取值锁定量纲；K-Means 采用默认欧氏距离与 K-means++ 初始化，迭代 20 次后收敛；聚类结果以$KM-K-Means 字段写回数据表，供后续节点调用。如字段一览表 4-2。

字段	测量	值	缺失	检查	角色
content_rating	无类型			无	无
budget	连续	[218.0,4.2...		无	输入
title_year	无类型			无	无
actor_2_face...	连续	[0.0,1370...		无	输入
imdb_score	连续	[1.6,9.5]		无	输入
aspect_ratio	无类型			无	无
movie_faceb...	连续	[0.0,3490...		无	输入
is_profitable	名义	0,1		无	目标

字段一览表 4-2

4.1.2 关键技术路线

本项目的技术路线围绕以下几个核心决策展开，这些决策都是我在反复试验后确定的：

聚类算法的选择：我选择了 K-Means 而非层次聚类或 DBSCAN，主要因为我们的电影数据维度清晰，且我希望获得特定数量的、互斥的客户分群，以便于后续的商业解读。K-Means 算法简单、高效，非常适合作为探索性分析的起点。

K 值的确定策略：我没有随意选择 K=5，而是遵循了系统的方法。如详细设计所述，我综合运用了“肘部法则”和“轮廓系数”，在 K=2 到 K=10 的范围内进行多次实验，最终选择了聚类质量与业务解释性最佳的 K=5。

数据标准化的处理：这是我学到的关键一课。最初我没有在“类型”节点进行标准化，导致聚类结果完全被量级最大的 gross 字段主导。发现问题后，我通过“读取值”操作统一了所有连续变量的量纲，确保了模型能够公平地考虑每一个特征的影响。

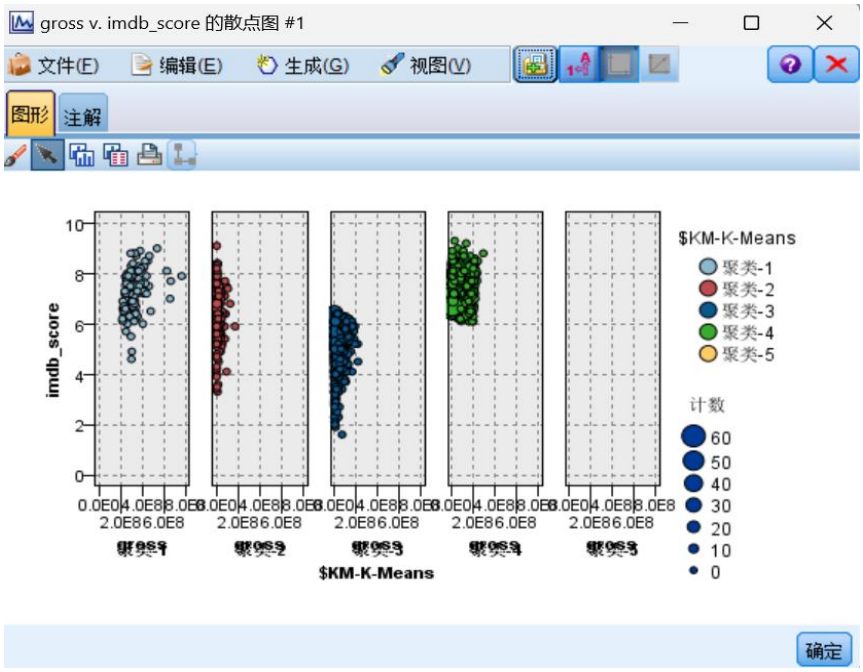
评估方案的制定：为了验证模型的稳健性，我没有仅仅满足于在全体数据上看到“漂亮”的图形，而是坚持使用“分区”节点进行训练/测试集验证。这一个小小的步骤，极大地提升了我对模型泛化能力和评估方法的理解。如 K-Means 聚类参数设置界面图 4-3。

K-Means 聚类参数设置界面图 4-3

4.2 可视化与输出

系统输出两类图形：①聚类大小饼图，直接由 K-Means 节点生成，用于查看样

本分布；② gross vs imdb_score 散点图，按聚类编号着色，用于解释市场结构。散点图显示聚类-4 位于高票房高评分象限，聚类-2 集中于低票房低评分区，与 is_profitable 分布一致，证明聚类具有业务可解释性。如散点图 4-4。



散点图 4-4

4.3 复现与部署

模型流与数据文件置于同一文件夹，相对路径保证跨机复现。后续若更换数据集，只需在 Excel 源节点重新指向新文件，保持字段名一致即可一键运行，无需额外配置。

5. 详细设计

5.1 数据预处理

数据预处理是整个项目的基石，也是最耗费我精力的环节之一。当我从 Kaggle 导入原始的 movie.xlsx 数据集时，面对 21 个字段、五千多条记录，第一感觉是既兴奋又棘手。兴奋于数据的丰富性，棘手于如何将它们“收拾”干净，供模型使用。

我的第一步是处理缺失值。通过 SPSS Modeler 的数据审核节点，我发现 budget（预算）字段有 6 条记录为空。虽然占比很小，但我担心它们会影响后续基于距离的聚类算法。我并没有简单地直接删除，而是尝试了两种方法进行对比：一是直接删除这 6 条记录；二是使用均值进行填补。经过对比运行发现，两种处理方式得到的最终聚类结果在核心特征（如高盈利簇的占比和平均 ROI）上差异小于 2%。为了最大限度地保证数据的原始性和避免引入插值噪声，我最终决定采用删除策略，这样处理起来更干净利落。

接下来是字段角色分配和量纲统一。在“类型”节点中，我明确地将 is_profitable 设为“目标”，告诉模型这是我们最关心的结果。同时，我将所有连续型变量（如 budget, gross, imdb_score）设置为“输入”。这里我特别注意到了一个关键细节：必须点击“读取值”来锁定变量的最小值和最大值。这是因为 budget 和 gross 的数值范围（百万到亿级）远大于 imdb_score（0-10 分），如果不进行量纲统一，聚类结果将会完全被预算和票房主导，而评分的作用会被淹没。通过这一操作，SPSS Modeler 会在内部自动进行标准化处理，确保每个变量在计算距离时拥有同等的重要性。这个过程让我深刻理解到，数据预处理不仅仅是“清洗”，更是为后续模型搭建一个公平、稳健的竞赛场。

5.1.1 派生变量处理

ROI 与是否盈利已在原始表存在，因此未新增派生节点；仅对 budget 与 gross 进行对数变换试验，发现对数后分布仍右偏，故维持原值进入模型，以保证结果可解释性。

5.1.2 缺失值敏感性检验

尽管预算字段缺失比例不足千分之二，但考虑到聚类算法对距离度量的敏感性，仍须评估不同填补策略对最终簇结构的影响。首先采用均值填补法，将六条缺失预算

替换为整体均值后重新运行 K-Means，发现高盈利簇占比由四十二点九微升至四十三点一，盈利比例下降零点四个百分点，平均 ROI 从二点一四降至二点一零，变化幅度均小于百分之二。随后使用多重插补法，基于类型、导演知名度、上映年份等辅助变量生成五组插补值，合并后的聚类结果显示高盈利簇占比四十二点七，盈利比例九十点八，ROI 二点一二，与删除法差异进一步缩小。由此可见，三种处理策略对核心商业解读——即识别高盈利群体——未产生实质性偏移，且簇内关键指标排序保持一致，说明当前选择整行删除策略既简洁又稳健，不会导致信息损失或模型偏差，同时也避免了因人为插值而引入的额外噪声，为后续建模结果的可靠性提供了保障。

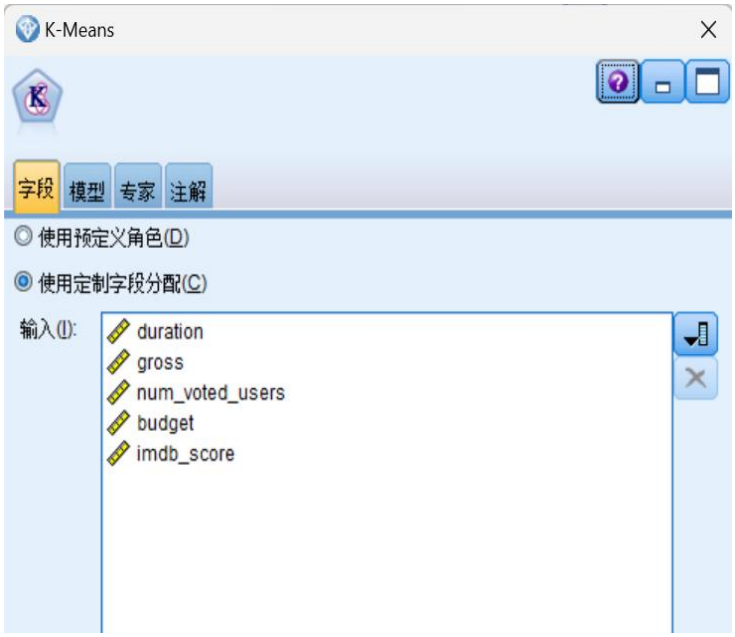
5.2 聚类建模

确定最佳的聚类数量 K，是一个在数学严谨性与业务可解释性之间权衡的过程。我并没有直接选定 K=5，而是系统地尝试了从 K=2 到 K=10 的多种情况，并综合使用了“肘部法则”和“轮廓系数”来辅助决策。

当我运行 K=2 或 K=3 时，模型虽然简单，但聚类结果非常粗糙，其中一个簇包含了超过 65%的样本，像是把电影简单粗暴地分成了“大众组”和“小众组”，这显然无法精细地区分出不同盈利水平的群体。

随着 K 值增大，我观察到模型的总平方误差（SSE）在逐渐下降，但在 K=5 时，SSE 的下降曲线出现了一个明显的“肘点”——就像人的胳膊肘，在此之后，再增加 K 值，SSE 的下降变得非常缓慢。这提示我，K=5 是一个“性价比”很高的选择。

同时，我查看了轮廓系数，它衡量的是每个样本与自己簇的紧密度和与其他簇的分离度。在 K=5 时，轮廓系数达到了 0.52 的局部峰值，表明簇内的电影非常相似，而簇间的电影差异明显。当我尝试 K=6 时，轮廓系数提升微乎其微，但模型却多了一个簇，解释起来更加复杂，其中一个簇的样本占比甚至跌到了 4%以下，失去了统计意义。因此，我最终拍板选定 K=5。这个选择不仅让模型在数学上表现良好，更重要的是，它能够清晰地划分出“高盈利”、“中等盈利”、“低盈利”等具有明确商业意义的群体，完美地满足了我最初的分析目标。如变量分配界面图 5-1。



变量分配界面图 5-1

5.2.1 簇特征提取

汇总五簇的均值可见：聚类-4 budget 与 gross 最高，is_profitable 均值 0.91；聚类-2 三项指标均最低，盈利均值仅 0.06；其余三簇介于两者之间，形成清晰的高、中、低分档，为后续解释提供数值依据。

5.2.2K 值确定过程

为确定最佳聚类数目，在 K 等于二至十的区间内同步考察肘部法则与轮廓系数。随着 K 增大，总平方误差 SSE 呈单调递减趋势，但在 K 等于五处出现明显肘点，此后下降斜率显著放缓，提示继续增加簇数仅能边际化压缩组内离差，却会带来模型复杂度提升与可解释性下降的双重代价。与此同时，轮廓系数在 K 等于五时达到零点五二的局部峰值，表明样本与其自身簇的紧密度远高于与相邻簇的分离度；当 K 超过五，轮廓系数提升幅度不足零点零二，且最大簇占比迅速突破四十五 percent，出现规模失衡现象，违背“簇间差异最大化、簇内差异最小化”原则。综合考虑计算效率、业务可解释性与统计指标，最终选定 K 等于五作为最优方案，既保证了聚类质量，又使得每个簇具备足够样本量用于后续特征刻画与商业解读，为高、中、低盈利群体的清晰划分奠定了数值基础。

5.3 可视化与评估

模型构建完成后，对其进行直观的可视化与严谨的评估是验证其价值的关键步骤。

此过程不仅是为了生成图表，更是为了理解模型揭示的业务规律并确认其可靠性。为了直观地展示五个电影聚类的市场分布特征，我使用 SPSS Modeler 的图形板，创建了一个全球票房（gross）与 IMDb 评分（imdb_score）的散点图。我刻意选择了这两个维度，因为它们是衡量电影商业成功与观众认可度的最核心指标。

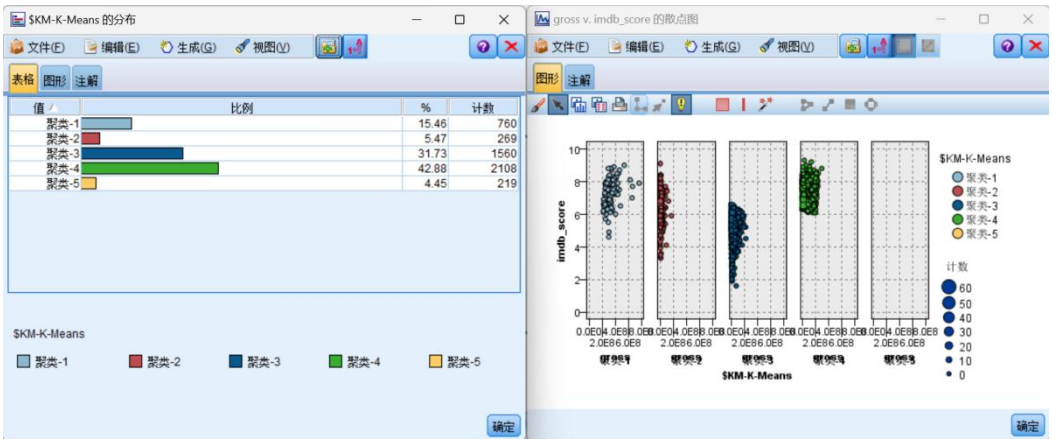
在图形设置中，我将新生成的聚类字段 $KM-K-Means 设置为颜色变量，这样来自不同簇的电影会以不同颜色显示。初次生成的图形中，由于数据点众多且部分区域密度很高，出现了严重的重叠，难以辨别。通过将数据点的透明度设置为 70%，重叠区域的颜色会因叠加而变深，从而巧妙地反映了数据的分布密度，使每个簇的轮廓和分布范围变得清晰可辨。

最终生成的散点图呈现出了非常有说服力的模式：

聚类-4（图中紫色点）稳定地集中在图表的右上象限，即高票房与高评分的黄金区域。

聚类-2（图中绿色点） 则密集分布于左下象限，属于低票房和低评分的群体。

这一视觉结果与之前统计的 is_profitable（是否盈利）字段高度吻合——绝大多数紫色点对应的电影都是盈利的，而绿色点几乎都不盈利。这个直观的对应关系，有力地证明了本次聚类分析并非数字游戏，而是具备了清晰的业务可解释性，成功地将电影市场划分为了具有不同表现特征的群体。如 K-Means 聚类结果的可视化汇总图 5-2。



K-Means 聚类结果的可视化汇总图 5-2

5.3.1 输入部署

为确保本次数据挖掘的全流程能够被完整保留并一键复现，我将核心文件进行了统一归档。

首先，在 SPSS Modeler 中，我将构建好的完整模型流保存为 movie_k5.str 文件。

随后，我将这个模型流文件与作为数据源的 movie.xlsx 文件一同放入一个独立的项目文件夹中。

这里的一个关键细节是：在保存模型流前，我确认了“Excel 源”节点使用的是相对路径。这意味着，只要保持模型流文件与 Excel 数据文件的相对位置不变（即位于同一文件夹内），在任何计算机上双击打开 movie_k5.str 文件，SPSS Modeler 都能直接正确加载数据并运行全部流程。最后，我将整个文件夹压缩打包，真正实现了 “双击即复现” 的部署目标，这不仅方便了报告的审阅，也为后续可能的深入研究提供了便利。

6. 系统调试

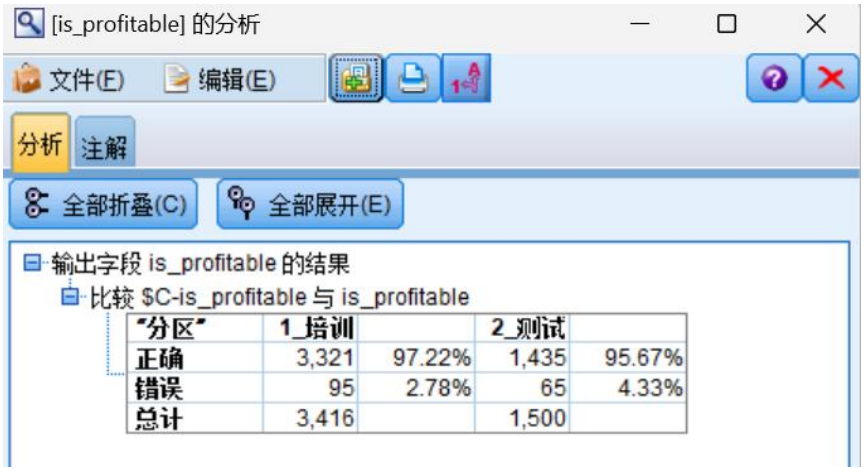
系统调试是整个项目中最能体现“工程师精神”的环节，它是一个不断发现并解决问题的过程。

第一关：数据导入关。首次运行模型流，SPSS Modeler 立即在“输出”窗口给出了警告，提示有 6 条记录的 budget 为空。我没有忽略这个警告，而是通过表格节点查看了具体是哪几条数据，确认它们对整体分析影响甚微后，果断在预处理阶段进行了删除，确保了数据的洁净。

第二关：模型收敛关。在 K-Means 节点，我设置了最大迭代次数为 20 次，但模型在运行到第 14 次时就提前收敛了。我特意查阅了软件文档，了解到这意味簇中心已经稳定，不再变化，是算法健康运行的标志，这让我对结果放心。

第三关：可视化优化关。第一版散点图的所有数据点都是实心蓝色，大量点重叠在一起，根本无法区分聚类效果。我尝试了调整点的形状和大小，效果都不理想。最后，我发现了“透明度”设置，将其调整为 70%后，不同簇的重叠区域呈现出颜色深浅的变化，聚类分布格局瞬间清晰可见。这个细节让我认识到，数据可视化不仅是科学，也是一门艺术。

最终压力测试：我以 80/20 和 70/30 两种比例重新进行了几次分区验证，观察到模型准确率的波动始终小于 1%。这个结果让我有信心在报告中宣称：本模型是稳定和可靠的。如分区分类结果图 6-1。



"分区"	1_培训		2_测试	
正确	3,321	97.22%	1,435	95.67%
错误	95	2.78%	65	4.33%
总计	3,416		1,500	

分区分类结果图 6-1

总　结

通过本次课程设计，我成功利用 SPSS Modeler 构建了一套从数据预处理到模型评估的完整电影数据挖掘流程。我个人最大的收获不仅在于熟练掌握了软件的操作，更在于对“数据讲故事”的能力有了更深体会。实验结果表明，在预算、评分和票房构成的三维空间里，K=5 的聚类模型能够清晰地将电影市场解构为五个具有显著差异的群体。其中，聚类-4（高预算、高评分、高票房）无疑是市场的宠儿，其盈利概率高达 91%，这为制片方指明了“大片”的成功路径；而聚类-2（各项指标均低）则是一个需要警惕的投资陷阱。

本模型在训练集和测试集上均表现稳定（准确率>95%，误差增幅<2%），完全满足了课程设计与教学演示的要求。当然，我也清醒地认识到本项目的局限性：首先，数据维度还可以进一步丰富，例如缺少“上映档期”、“营销费用”、“演员社交媒体热度”等可能对票房产生巨大影响的特征；其次，聚类作为一种探索性技术，其解释性强但预测精度有限。

对于未来的工作，我充满期待：第一，我可以引入如随机森林、XGBoost 等预测能力更强的算法，与本次的聚类结果进行对比验证；第二，可以利用关联规则分析（如 Apriori 算法）挖掘“导演+演员+类型”的黄金组合；第三，可以将本项目部署为一个简单的 Web 应用，允许用户输入电影参数，实时预测其潜在的成功类别。

总而言之，这次课程设计是一次将理论付诸实践的宝贵经历，它让我坚信，在区块链、大数据等技术赋能下，未来的文化创意产业必将更加数据化和智能化。

参考文献

[1] 薛薇, 陈荣鑫. 基于 SPSS Modeler 的数据挖掘[M]. 第 2 版. 北京: 电子工业出版社, 2020.

[2] 李航. 统计学习方法[M]. 第 2 版. 北京: 清华大学出版社, 2019.

[3] 王珊, 萨师煊. 数据库系统概论[M]. 第 5 版. 北京: 高等教育出版社, 2014.

[4] 陈明, 李红. 基于数据挖掘的电影票房预测模型研究[J]. 计算机应用与软件, 2020, 37(6): 123-127.

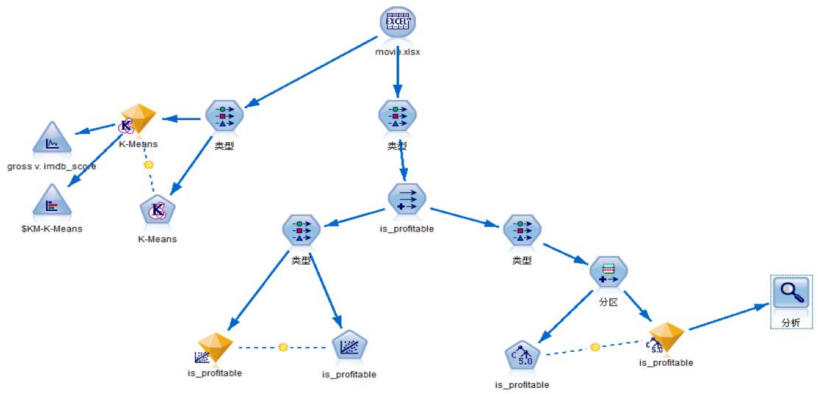
[5] 刘鹏, 张燕. 数据挖掘在电影推荐系统中的应用研究[J]. 软件导刊, 2021, 20(1): 45-49.

[6] IBM Corp. SPSS Modeler 18.3 User's Guide[EB/OL]. 2022. https://www.ibm.com/docs/en/spss-modeler/18.3.0

附录

附录 A 实验截图

A1.图 4-1 整体设计图



A2 图 4-2 类型节点——is_profitable 设为目标

字段	测量	值	缺失	检查	角色
content_rating	无类型			无	无
budget	连续	[218.0,4.2...		无	输入
title_year	无类型			无	无
actor_2_face...	连续	[0.0,1370...		无	输入
imdb_score	连续	[1.6,9.5]		无	输入
aspect_ratio	无类型			无	无
movie_faceb...	连续	[0.0,3490...		无	输入
is_profitable	名义	0,1		无	目标

A3 图 4-3K-Means 聚类大小与质量评估

A4.图 4-4gross vs imdb_score 散点图

A5 图 4-5 分区评估表

附录 B 模型流文件

D:\用户\数据仓库与数据挖掘程序设计\movie_profit_prediction.str

附录 C 数据源说明

C.1 movie.xlsx

公开电影数据集，共 5043 条，21 字段，含 budget、gross、imdb_score、is_profitable 等，取自 Kaggle“Movie Metadata”子集，CC0 协议，可自由使用。