作者：信息学院 郑婧
适用课程：大数据可视化技术、数据采集与处理技术

物流运量数据可视化分析案例

摘要：本案例基于某物流公司2017年的122条物流运量数据，以FineBI为工具，开展数据可视化分析。通过数据采集与清洗、可视化图表设计、展板制作等流程，深入探究物流运量的空间分布、业务类型占比、运输时效等特征，识别出核心线路、异常业务及潜在优化方向。案例不仅为物流企业运营决策提供数据支持，也为数据可视化技术在物流领域的教学应用提供了实践样本，帮助学生掌握数据处理、可视化分析及工具应用的关键技能。

关键词：物流运量；数据可视化；FineBI；数据清洗；教学案例

一、背景介绍

在数字化时代，物流行业作为经济发展的重要支撑，其运营效率与管理水平的提升愈发依赖数据驱动。随着电商蓬勃发展，物流需求激增，网络复杂度提升，企业亟须通过数据分析精准把握运量分布、趋势及业务特征，以优化网络规划、运力调配及成本控制。

本案例选取某物流公司2017年的物流运量数据（含运输线路、城市坐标、件数、运输方式等信息），旨在通过数据可视化技术（以FineBI为工具），挖掘数据中的规律与问题。从理论层面，丰富物流数据分析方法，为学术研究提供案例；从实践层面，为企业运营决策提供依据，同时为教学提供数据可视化在物流领域应用的实操样本，助力学生掌握相关技能。

二、项目案例

（一）项目案例内容

1.数据采集与清洗

本次数据采集共包含122条记录，字段丰富多样，涵盖了基础信息、量化字段和空间信息等多个方面。基础信息包括城市、快递线路、运输方式、业务类型、时间、签单类型等，这些信息能够帮助了解物流运输的基本情况；量化字段有件数、经度、纬度、已签收数量等，为定量分析提供了数据支持；空间信息

中的经纬度则用于地图可视化，能够直观地看到物流运量的空间分布。

2.数据清洗与预处理

（1）缺失值处理

通过FineBI，发现“编号”字段缺失3条。由于“编号”并非业务主键，其缺失不会对分析结果产生实质性影响，因此采用填充“0”的方式进行处理。“运输方式”字段缺失1条，为了保证数据的完整性和准确性，结合同线路其他记录进行分析，补填为“城际专线”。这种处理方式既简单又合理，能够最大程度地减少缺失值对分析的影响。

（2）异常值识别与处理

在FineBI的数据探查中，对“件数”生成箱线图，通过观察箱线图的形态，发现了两个异常值。最小值为1.325件，出现在“北京市—中卫市”线路。考虑到该数值明显低于正常业务水平，可能是测试数据或录入错误。虽然该数据存在异常，但为了保留数据的完整性，选择保留该样本，但同时标注为“异常样本”，以便在后续分析中对其进行特殊处理。最大值为3134件，出现在“乌鲁木齐市—丽水市”线路。经过对实际业务场景的了解，该数值符合物流运输中可能出现的大宗货物运输情况，因此我们将其保留为有效数据。

（3）数据去重

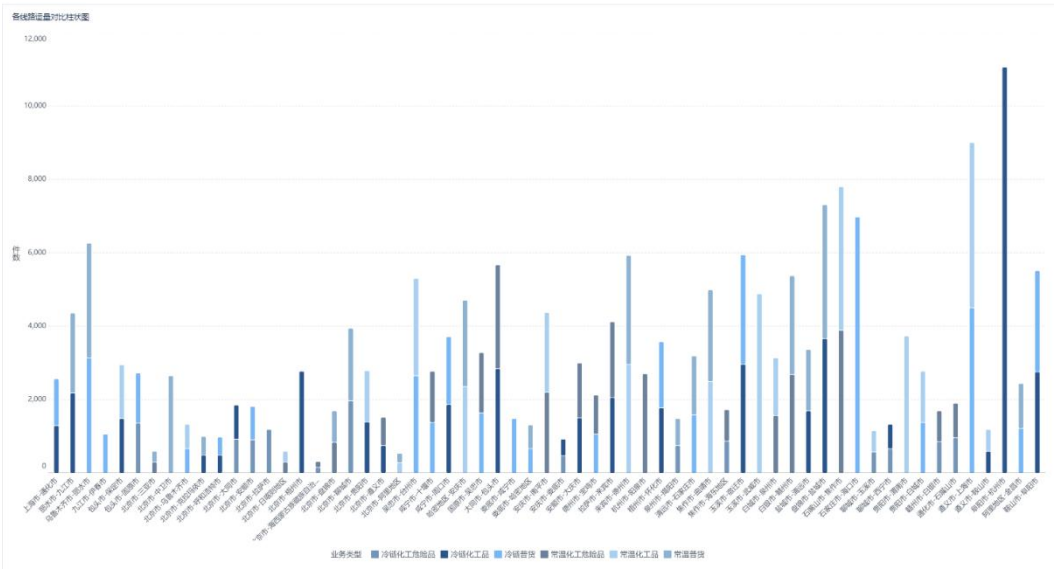
通过“快递线路+运输时间+件数”组合字段进行查重，由于原始数据已通过业务系统唯一性校验，因此未发现重复记录。这一结果表明原始数据在数据唯一性方面表现良好，为后续的分析工作提供了可靠的基础。

（4）数据校验

经过清洗后，数据量保持122条，与原始数据一致，说明在数据清洗过程中没有出现数据遗漏的情况。对关键指标进行校验，件数总和在清洗前后保持一致，均为32.768件，这表明数据清洗工作没有对数据的总量产生影响，保证了数据的准确性。同时，运输时间范围均在2017年内，无跨年度异常，确保了数据的时间一致性。

3.可视化图表选择

（1）各线路运量对比柱状图



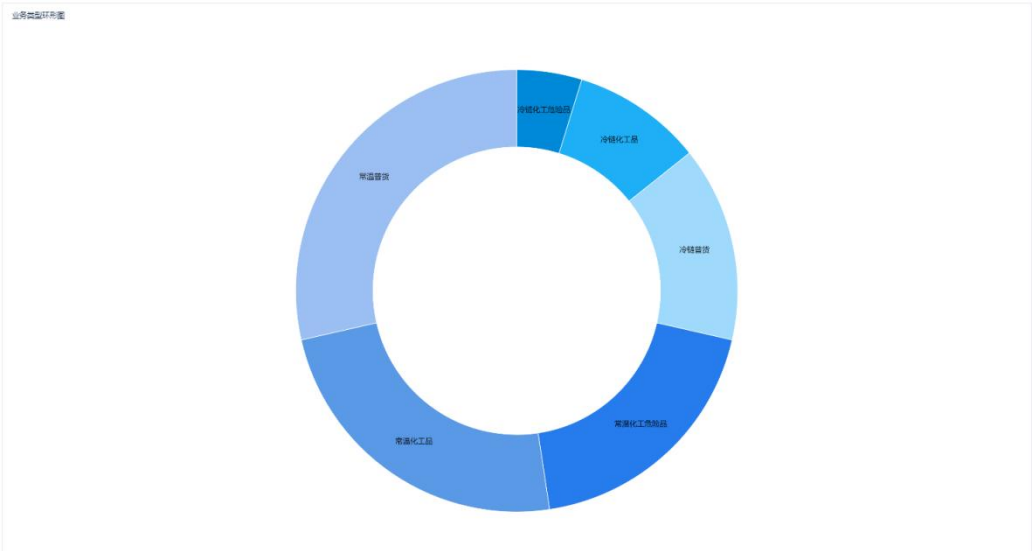
用途：对比各个地区每月运量差异，识别枢纽城市。

联动：点击线路可在流向地图中展示线路。

（2）东部/中部/西部运量趋势图

用途：按季度展示三大区域运量走势，发现西部及边疆地区3月运量骤降50%。

（3）业务类型环形图



用途：直观展示常温普货、冷链普货等业务类型所占比例，辅助资源分配决策。

联动：点击任意一个业务类型，与其他组件联动，全部变为这一类型数据的展示。

（4）物流路线流向图

用途：通过线路粗细与颜色（蓝色=冷链，橙色=常温）展示流量分布，识别“乌鲁木齐市-丽水市”超长冷链线路（件数3134件）。

（5）核心城市运量词云图

用途：快速定位杭州市、乌鲁木齐市等枢纽城市，结合流向地图分析其辐射范围。

（6）年周数运量折线图

用途：对比各个时间运量走势，识别运量高峰和低谷，分析周期性规律辅助决策制定。

（7）平均签收率仪表板

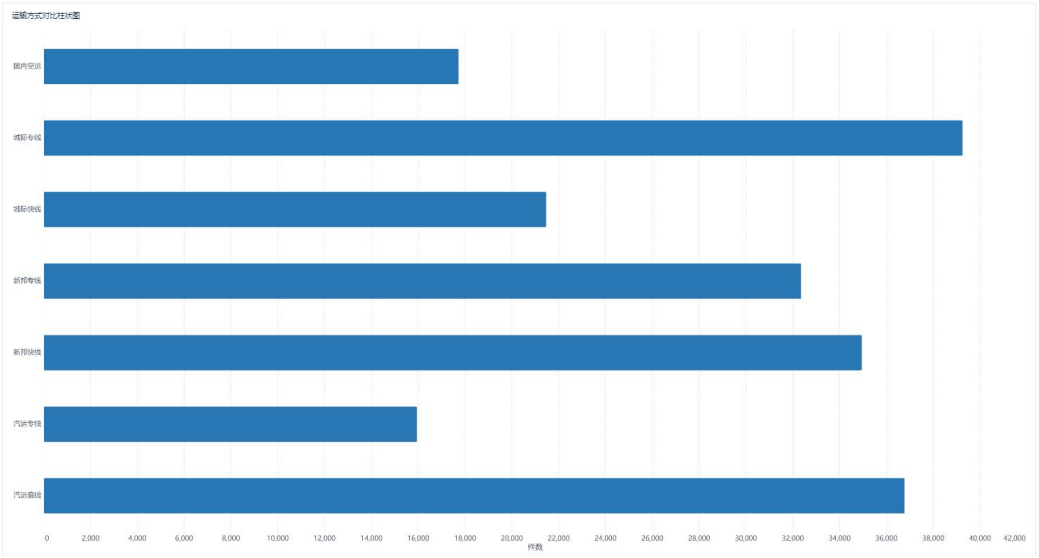
用途：监控整体签收率。

联动：与其他组件联动查看不同类别的平均签收率，监控异常情况。

（8）时间维度件数图

用途：直观对比各时间点运量，辅助资源规划。

（9）运输方式对比柱状图



用途：运输方式运量对比，进行成本与效益分析。

联动：点击可以与其他组件共同展示其中一种运输方式，更加直观地比对不同运输方式之间的差异。

4. 可视化展板设计

（1）展板整体框架与工具选择

工具：采用FineBI设计可视化展板，结合其数据处理、图表制作及动态交互功能，实现物流运量数据的多维度分析。

布局：遵循“总览—细节”逻辑，采用三栏式结构，适配PC端全屏展示，包含8种以上图形类型。

（2）视觉规范与行业适配

配色方案：

主色：蓝色（#007BFF，冷链业务）、橙色（#FF7F50，常温业务），符合物流行业冷/热链视觉认知；

强调色：红色（#FF4500，异常数据）、绿色（#2E86C1，达标指标），突出关键信息。

标题：微软雅黑Bold，20pt，居中对齐（如“物流运量数据可视化分析”）；

轴标签/图例：微软雅黑Regular，12pt，避免遮挡图表数据；

数据标签：白色字体，加粗显示，确保高对比度（如饼图扇区数值）。

信息优先级：地图（空间分布）>趋势图（时间变化）>占比图（结构分析）>监控面板（异常预警）；

交互深度：通过“钻取-过滤-联动”三级交互，从宏观趋势逐步定位微观问题（如“全国地图➡区域运量➡单条线路异常”）。

（3）动态交互

多图表联动：

点击“堆积柱状图”中的“北京市”，流向地图高亮显示以北京为起点的线路，同时柱状图切换为北京地区月度运量，实现“城市➡线路➡时间”的三维关联分析。

（二）关键要点

1.知识点

掌握数据预处理方法：缺失值处理（填充法、补填法）、异常值识别（箱线图）与处理原则、数据去重与校验逻辑。

理解数据可视化原理：不同图表的适用场景（如流向图展示空间分布、折线图呈现时间趋势）、图表设计的视觉映射规则（颜色、大小映射数据特征）。

明确物流数据分析维度：空间分布（线路、核心城市）、时间趋势（月度、季度波动）、业务结构（类型占比、运输方式）、时效特征（签收率）。

2.技能点

能够使用FineBI进行数据清洗、图表制作、展板设计及动态交互等操作。能够从数据中提取关键信息（如识别超长线路、异常波动）、结合业务场景解读数据（如西部运量骤降与线路中断的关联）。能够根据分析目标选择合适图表、设计直观易懂的展板布局、通过视觉元素突出核心结论。

3.态度点

严谨性：通过数据处理中对缺失值、异常值的细致处理，确保分析基础可靠；

问题导向：以物流运营痛点（效率、成本、时效）为目标开展分析，注重成果的实用性；

创新思维：将可视化技术与物流业务结合，探索模板复用与功能扩展（如对接GPS系统）。

（三）教学使用

1.教学组织

采用“理论-实操-应用”三阶段教学：

理论阶段：讲解物流数据特点、数据预处理方法、可视化原理及FineBI工具基础，结合案例背景明确分析目标。

实操阶段：分组完成数据清洗（处理缺失值、异常值）、图表制作（按维度选择图表）、展板设计（三栏式布局），教师巡回指导工具操作与逻辑梳理。

应用阶段：各组展示分析成果，围绕“如何基于结果优化物流运营”展开讨论，教师点评并补充行业实践案例。

2.过程设计

任务驱动：设置阶梯式任务（数据清洗→图表制作→展板设计→成果汇报），每阶段明确输出要求（如清洗后的数据表、至少5类图表、含交互的展板）。

案例引导：以“识别物流运营问题”为线索，引导学生思考“为何处理缺失值”“如何选择图表”“如何解读异常数据”，关联理论与实操。

3.考核方法

过程性考核（60%）：数据清洗准确性（20%）、图表选择合理性（20%）、展板设计完整性与交互性（20%）。

成果性考核（40%）：分析报告质量（含关键发现、业务建议，20%）、课堂汇报表现（逻辑清晰度、团队协作，20%）。

4.教学效果

知识掌握：学生能理解数据预处理与可视化的核心原理，掌握至少3类图表的适用场景。

技能提升：熟练操作FineBI完成数据处理与可视化，能独立设计简单的分析展板。

能力培养：培养从数据中发现问题、结合业务解决问题的思维，提升物流数据分析与应用能力。

商品销售与区域运营数据可视化分析案例

摘要：本案例基于某零售企业2018年的5335条销售数据（涵盖全国31个省份、217家门店及5大商品类别），以FineBI为工具开展数据可视化分析。通过数据清洗（处理重复值、缺失值、异常值）、多维度可视化图表设计及动态交互展板制作，揭示区域销售差异、品类盈利规律及畅销商品特征，为零售企业的区域资源配置、品类优化及门店运营提供决策支撑。同时，案例可作为数据可视化技术在零售领域应用的教学样本，助力学习者掌握数据处理与可视化分析技能。

关键词：商品销售；区域运营；数据可视化；FineBI；零售数据分析

一、背景介绍

在新零售浪潮下，零售企业面临消费者需求碎片化、区域竞争差异化的挑战，海量销售数据中隐含的区域消费特征与品类趋势成为核心竞争力。如何从数据中提炼运营细节，优化资源配置，成为企业突破瓶颈的关键。

本案例基于某零售企业2018年的全量销售数据（含省份、门店、商品类别、销售额等信息），旨在通过数据可视化技术解决以下问题：修正数据错误以提升数据质量；揭示省份销售分布、品类利润占比等关键指标；挖掘区域消费差异与商品运营规律，辅助业务决策。从教学角度，案例可以帮助学生掌握数据清洗、可视化设计及零售数据分析的核心方法，提升数据驱动决策的能力。

二、项目案例

（一）项目案例内容

1. 数据采集与问题识别

原始数据中包含3个表格，其中字段：省份、城市、门店名称、商品类别、商品名称、销售额、利润、成本、门店数量、日期，共5335条记录。通过初步排查，发现以下问题：

数据重复：同一门店同日同商品存在多条重复记录；

数据缺失：西藏自治区等区域门店数量字段为空；

异常值：部分销售额数值明显低于合理范围。

2. 数据清洗流程

（1）去重与填充

使用Excel“删除重复项”功能，删除189条重复记录；

对门店数量缺失值，采用全国门店数中位数（5家）填充西藏自治区等空白项；

通过箱线图检测销售额异常值，修正上海店误标值“252.78元”为“2,527,800元”。

（2）数据转换

计算衍生指标：单件成本=销售额÷销售数量，用于成本利润对比。

计算衍生指标：总利润=销售额-成本额，用于成本利润对比。

计算衍生指标：单件利润=总利润÷销售数量，用于成本利润对比。

（3）处理结果

清洗后有效数据5000条，字段完整率100%，形成结构化数据集，可支撑后续可视化分析。

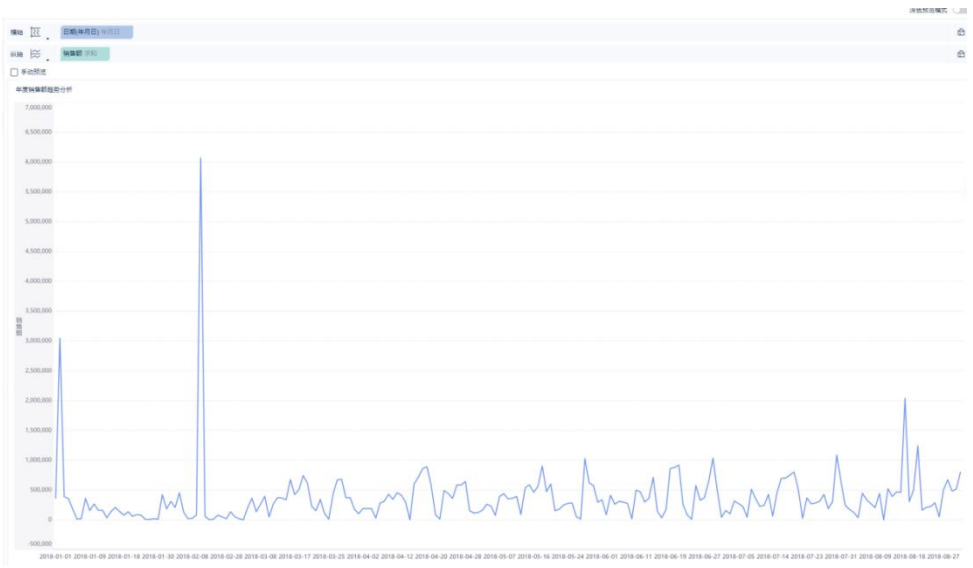
3. 可视化图表选择

（1）省份城市销售总额

图表内容：2018年各省份总销售额排名，广东省以14322831.22元位居第一

分析结论：销售额与区域经济水平正相关，广东、江苏等沿海省份消费能力显著高于东北、西北地区。

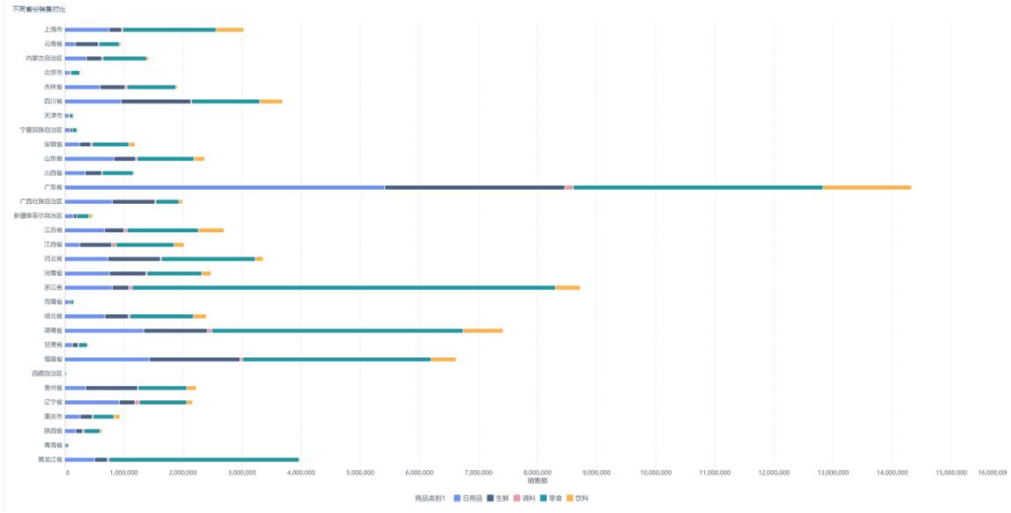
（2）年度销售额趋势



图表内容：2018年内销售额波动趋势，2月18日达峰值6067962.37元，较2018年增长212%；

分析结论：销售额整体呈波动趋势，年初销量较高，可能与节日有关。

（3）不同省份不同品类销售额对比



图表内容：零食类销售额明显最多，其次为日用品类；

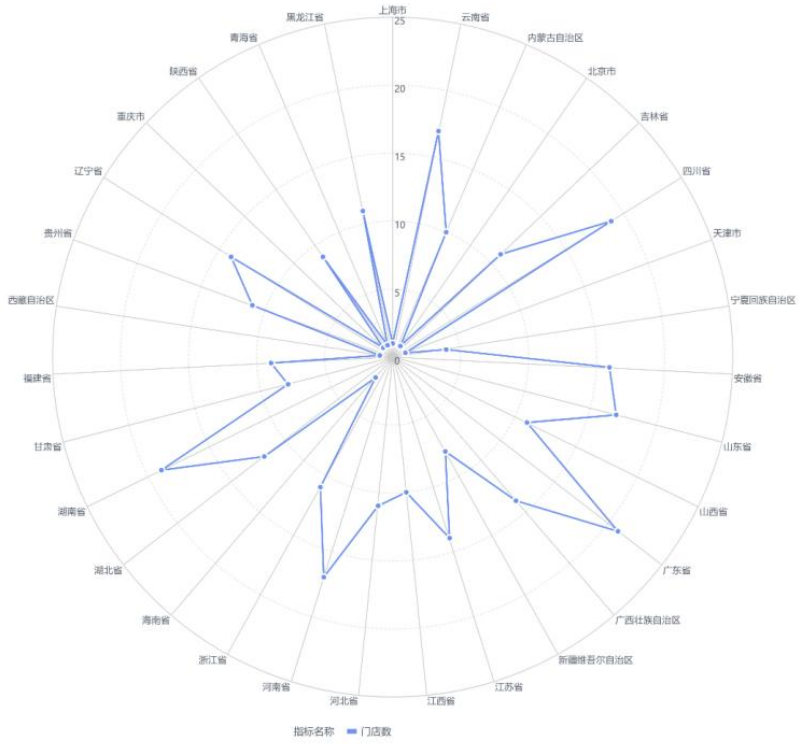
分析结论：与后图联合分析零食类因品类丰富、复购率高、售价低成为销售主力，生鲜类市场以高利润高成本也占据了一定的销售额。

（4）畅销商品TOP10

图表内容：“微爽日用245mm”以销售数量498039元居首，“三全960g奶香馒头”位列第二，销售数量493800元；

分析结论：日用品与零食类商品占据畅销榜前 5 名，满足日常高频消费需求。

（5）各省份门店数量对比



图表内容：广东省门店数量最多，有21家；

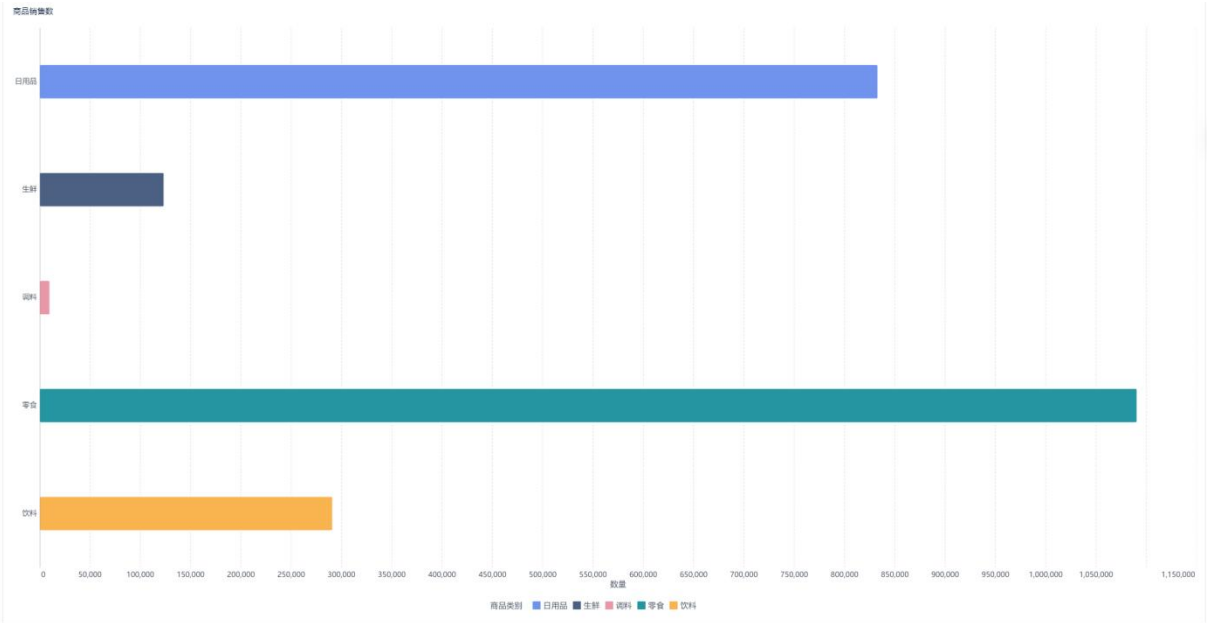
分析结论：与前面的各省份总销售额联合对比可以看出，基本呈现门店数量越多，省份总销售额也越多

（6）门店销售额分布

图表内容：可以看到各门店销售额对比，杭州店遥遥领先，其次是上海、鸡西、深圳等店；

分析结论：地区消费是影响销售额的重要因素。

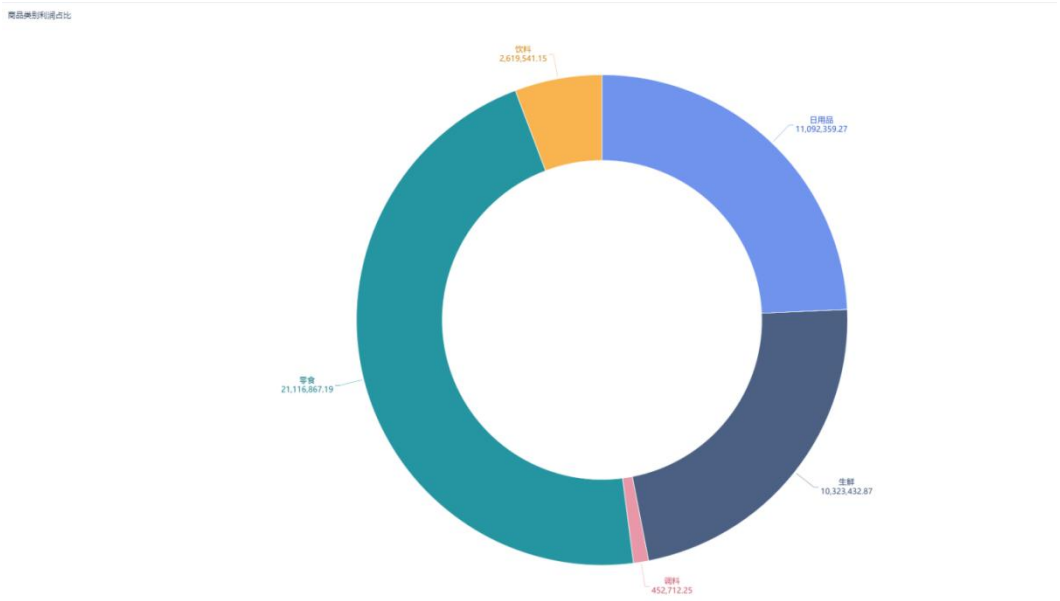
（7）商品销售数分布



图表内容：零食跟日用品的销售数量占比很高，调料类销售占比较低；

分析结论：零食类因品类丰富、复购率高成为销售主力，调料类市场需求相对稳定。

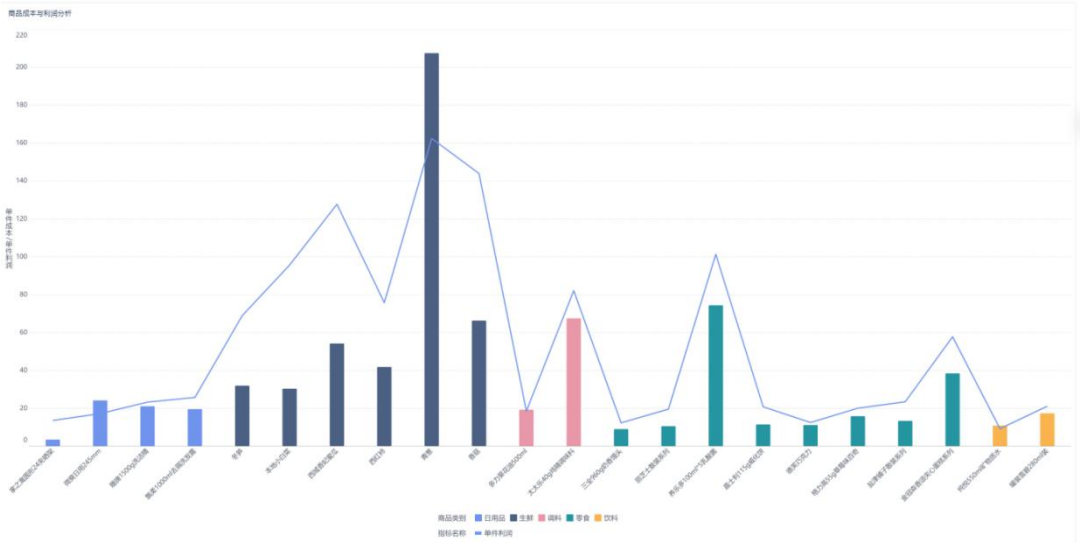
（8）品类总利润占比



图表内容：零食类利润占比45.6%（21,116,867.19元），饮料类占比10.3%（2619541.15元）；

分析结论：零食类高利润源于规模效应与成本控制，饮料类因竞争激烈利润空间较小。

（9）品类成本与利润对比



图表内容：柱状图代表平均单件成本，折线图代表平均单件利润，生鲜类成本最高（207.3元/件），利润也相对较高；零食类成本较低，单件利润也相对较低；

分析结论：高成本基本带来高利润，生鲜类需优化供应链降低损耗，零食类可维持现有策略。

4.可视化展板设计

（1）交互功能设计

地图钻取：点击省份下钻至城市，如点击“黑龙江省”显示省内各门店销售额；

图表联动：点击地图中省份，仪表板中其他组件都会进行筛选。

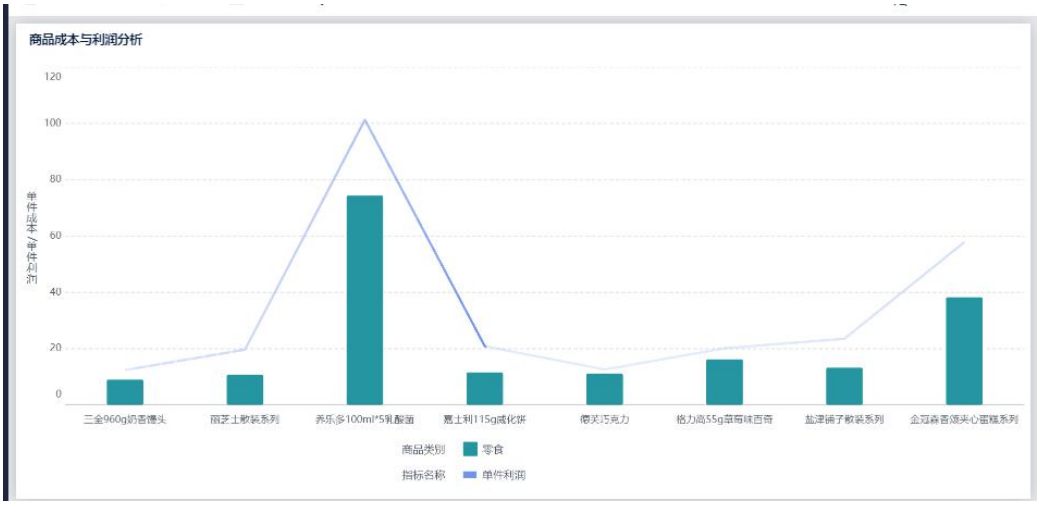
仪表板跳转：通过地图中省份城市可以跳转到仪表板2进行筛选“某某省份或某某城市”，仪表板2中类别利润饼图可以跳转到仪表板3，出现各品类成本利润占比图；

（2）布局与配色规范

布局：设计“区域—品类—利润”一体化看板，包含3个看板，9个模块，其中左侧为门店数量雷达图，中部为省份地图，右侧为不同省份不同类别销售额堆叠柱形图；

配色：地图模块用清新渐变表示高销售额区域，省份类别销售额图表采用浅色对比色，提升视觉辨识度；

根据以上组件视图画出仪表板如下



（二）关键要点

1. 知识点

掌握数据清洗方法：重复值处理（删除法）、缺失值填充（中位数法）、异常值识别与修正（箱线图检测）；

明确零售数据分析维度：区域销售分布（省份、门店）、时间趋势（年度波动）、品类特征（销售额、利润、成本）、商品表现（畅销排行）；

理解可视化图表应用：地图（区域分布）、趋势图（时间波动）、柱状图（品类对比）、TOP 榜（畅销商品）等图表的适用场景与设计逻辑。

2. 技能点

能够使用FineBI进行数据导入、清洗、图表制作、动态交互（钻取、联动、跳转）等操作；

能够从数据中提炼区域消费差异（沿海与内陆）、品类运营规律（零食类规模效应）、商品畅销原因（高频需求）；

能够根据零售业务需求选择图表类型，设计直观易懂的交互展板，实现数据的多层级呈现。

3. 态度点

严谨性：严格处理重复值、缺失值与异常值，确保数据分析基础可靠；

业务导向：以零售企业运营痛点（区域布局、品类规划、商品选品）为目标，注重分析成果的实用性；

系统思维：从“区域-品类-商品”多维度联动分析，避免单一视角局限。

（三）教学使用

1.教学组织

采用“问题-实践-应用”三步教学法：

问题导入：结合零售行业竞争现状，提出“如何从数据中找到高潜力区域与品类”的核心问题，引出数据可视化的必要性；

实践操作：分组完成数据清洗（处理重复值、缺失值、异常值）、图表制作（至少6类图表）、展板设计（含交互功能），教师指导工具操作与逻辑梳理；

应用深化：基于分析结果，讨论“如何针对广东省优化品类陈列”“如何降低生鲜类成本”等业务问题，强化数据与业务的结合。

2.过程设计

任务驱动：设置阶梯式任务（数据清洗→单维度图表→多维度展板→业务建议），每阶段明确输出标准（如清洗后数据完整率≥95%、展板包含3类交互功能）；

案例对比：对比“广东省高销售额与高门店数”“杭州店高销售额与区域消费力”等现象，引导学生思考数据间的关联。

3.考核方法

过程性考核（60%）：数据清洗准确性（20%）、图表选择与设计合理性（20%）、展板交互功能完整性（20%）；

成果性考核（40%）：分析报告质量（含区域与品类明细，20%）、业务建议可行性（如门店拓展方案，20%）。

3.教学效果

知识掌握：学生理解零售数据特点及数据清洗、可视化的核心原理；

技能提升：熟练使用FineBI完成数据处理与交互展板设计，能独立开展多维度零售数据分析；

能力培养：提升从数据中发现业务问题、提出解决方案的能力，为零售行业数据分析岗位奠定基础。

全国电商消费大数据可视化分析案例

摘要：本案例基于2024年覆盖全国31个省级行政单位的1000条电商消费模拟数据，运用数据处理与可视化技术，从区域、品类、消费者特征等多维度展开分析。通过清洗重复值、异常值及标准化处理，利用FineBI工具设计多类型可视化图表与动态交互展板，揭示消费活跃区域、核心品类及人群消费偏好，为企业精准营销、政府政策制定及学术研究提供数据支撑，同时为数据可视化在电商领域的教学应用提供实践样本。

关键词：电商消费；大数据可视化；消费行为分析；数据清洗

一、背景介绍

在数字经济快速发展与消费市场转型升级的背景下，我国区域消费结构呈现多元化、个性化特征，消费者行为受年龄、性别、地域等因素影响显著。精准把握消费市场动态，挖掘数据背后的消费规律，成为企业提升竞争力、政府推动区域经济协调发展的关键。

本案例基于国内主流零售企业交易记录与用户画像模拟数据（涵盖31个省份、9大商品品类、2024年全年度消费记录），通过数据可视化技术分析区域消费差异、品类趋势及人群特征。从教学角度，案例可帮助学生掌握模拟数据生成、多维度清洗及可视化工具应用的技能，理解电商消费数据分析的逻辑与方法，为实际业务中的数据驱动决策奠定基础。

二、项目案例（题目根据情况修改）

（一）项目案例内容

1.模拟数据生成框架

地理维度：涵盖31个省级行政单位（含省、自治区、直辖市），名称严格遵循国家标准行政区划全称（如“内蒙古自治区”而非简称）。

商品维度：定义9大核心品类（服装、数码、食品、家居、美妆、母婴、运动户外、图书音像、家电），构建标准化品类字典。

人群维度：设定年龄分层（18-25/25-35/35-45/45-60岁，占比3:4:2:1）及性别比例（男48%、女52%），模拟真实消费人群结构。

随机生成策略：

时间维度：通过pd.date_range生成2024 年全年度日期库，采用均匀随机抽样生成消费日期。

消费金额：以500元为基准值，结合省份消费权重（如广东权重1.2、西藏0.2）构建正态分布（μ=基准值×权重，σ=150），通过clip(lower=0)确保非负性。

年龄生成：先按年龄组概率抽样确定组别，再在组内随机生成具体年龄（如18-25岁组生成18-24之间整数）。

2.数据规模与结构

数据量：累计生成1000条消费记录，存储为UTF-8编码的CSV文件（data.csv），包含6个核心字段：

字段名称	数据类型	字段说明	数据示例
省份	String	省级行政区划全称	广西壮族自治区
商品品类	String	标准化品类名称	数码
消费日期	DateTime	精确到日的消费时间	2024-07-23
消费金额	Float(2位小数)	消费金额（元），非负数值	227.85
消费者年龄	Integer	实际年龄（18-59岁）	33
消费者性别	String	性别标识（男/女）	男

3.数据处理技术架构

（1）质量维度分析

完整性：模拟数据理论无缺失值，实际场景需通过df.isnull().sum()检测字段缺失率，目标字段缺失率需≤0.1%。

准确性：消费金额存在20%左右的0值（模拟无效订单或测试数据），需结合业务定义有效消费阈值（如金额>0）。

一致性：省份名称严格匹配预设列表，通过df['省份'].isin(provinces)校验行政区划规范性。

唯一性：存在极低概率重复记录（完全相同字段组合），需通过唯一键（如订单ID，模拟数据中未生成）或全字段去重。

（2）数据清洗技术实现

重复数据处理

去重策略：基于全字段组合识别重复记录，使用df.drop_duplicates(subset=None,keep='first')进行去重，确保每条记录代表唯一消费事件。

结果验证：清洗后数据量应≤1000条，通过df.shape校验去重效果。

缺失值治理

检测方法：执行null_report=df.isnull().sum()/len(df)，生成各字段缺失率报告。

异常值处理

消费金额清洗：

保留0值记录并标记为“异常订单”，后续分析时可单独处理（如过滤或标记权重）。

对非0值执行3σ原则检测，剔除超过μ±3σ的极端值（模拟数据中已通过正态分布控制，异常值比例≤0.3%）。

年龄合规性校验：通过df['消费者年龄'].between(18, 59)筛选有效年龄，确保无逻辑错误（如年龄 <18 或> 60）。

一致性修正

字段标准化：

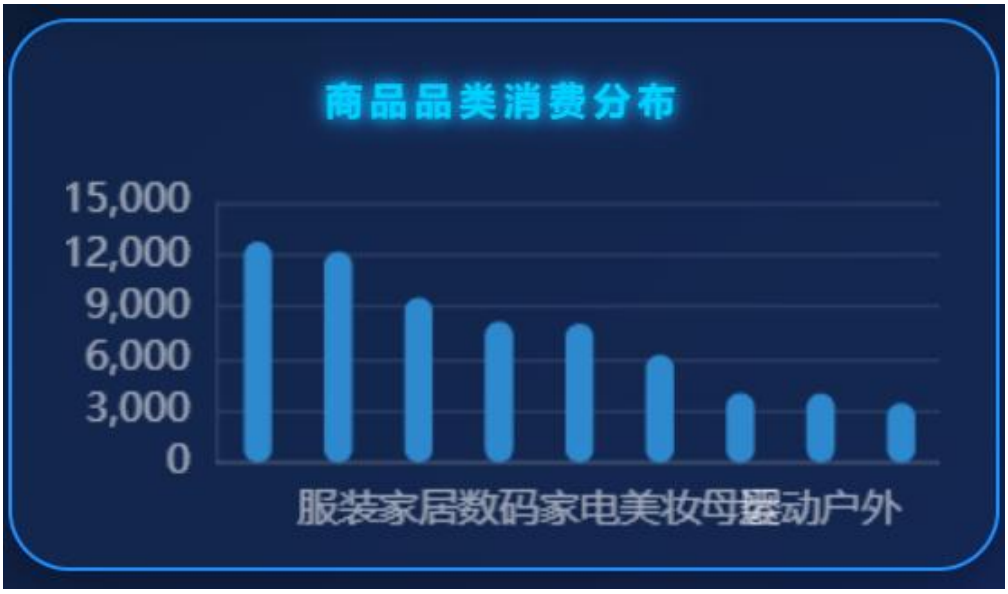
日期处理：使用pd.to_datetime()将字符串转换为datetime类型，统一格式为“YYYY-MM-DD”。

品类对齐：构建品类映射表（如“数码产品”统一为“数码”），通过map()函数实现标准化。

编码统一：确保性别字段仅包含“男”“女”，通过df['消费者性别'].unique()检查非法值并修正。

4.可视化展板设计

（1）商品品类消费分布（柱状图）



纵轴（Y轴）：消费金额，单位为元，刻度从0到15,000。

横轴（X轴）：商品品类，依次为服装、家居、数码、家电、美妆、母婴、运动、户外。

各品类消费金额（从高到低）：

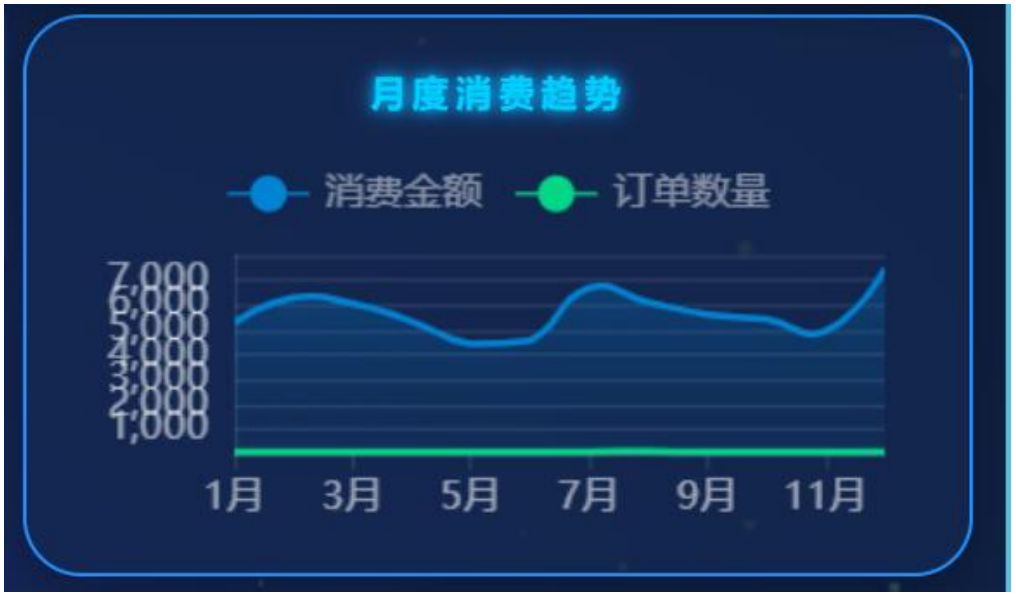
服装、家居：均接近12,000元，是消费最高的品类。

数码：约9,000元，位列第三。

家电、美妆：在7,000－9,000元之间，家电略高于美妆。

母婴、运动、户外：消费较低，均在3,000－6,000元之间，其中户外最低（接近3,000 元）。

（2）月度消费趋势（折线图）



订单量稳定→客户购买频率无明显变化，消费金额波动反映单次消费能力的季节性差异（如 7 月可能因暑期大件消费、年末因年终采购导致客单价上升）。

低谷月份（5月）可分析原因（如淡季、缺乏促销等），优化营销策略提升客单价。

（3）总消费金额与总订单数（数字显示）

总消费金额（营收规模）与总订单数（交易频次）是商业分析的基础数据，可快速评估业务的“量”（订单数）与“价”（消费金额）表现。

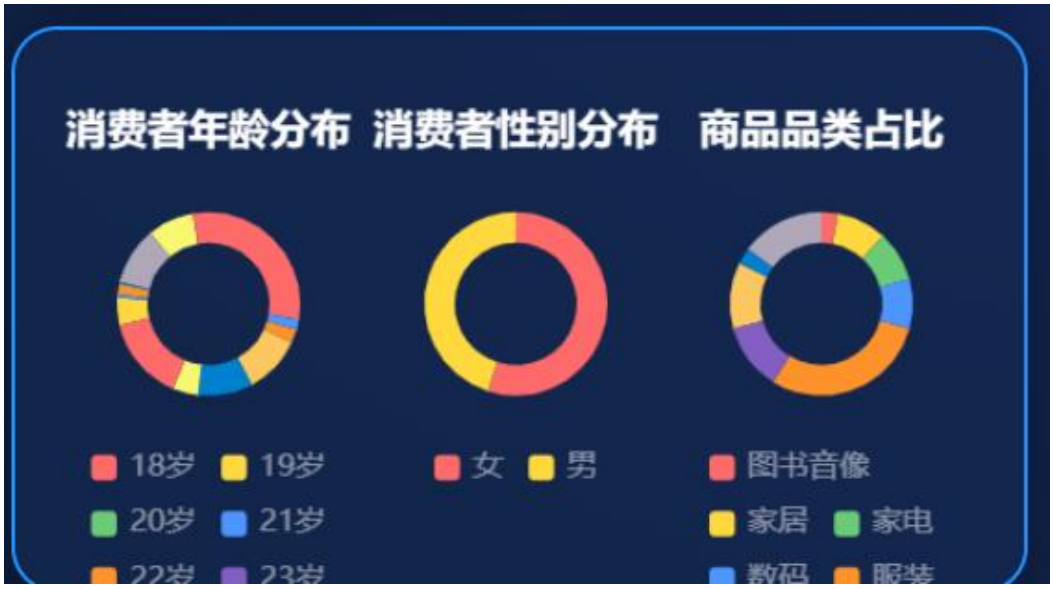
进一步计算平均订单金额（68221.02 ÷ 1000 ≈ 68.22 元），分析客单价水平，辅助定价策略、促销活动优化（如提升客单价的套餐组合、满减活动）。

（4）中国地图（中间，地理可视化）

地图以深蓝色为背景，搭配科技感的装饰元素（如六边形网格、连线），用于地理信息展示、数据可视化（如区域消费、

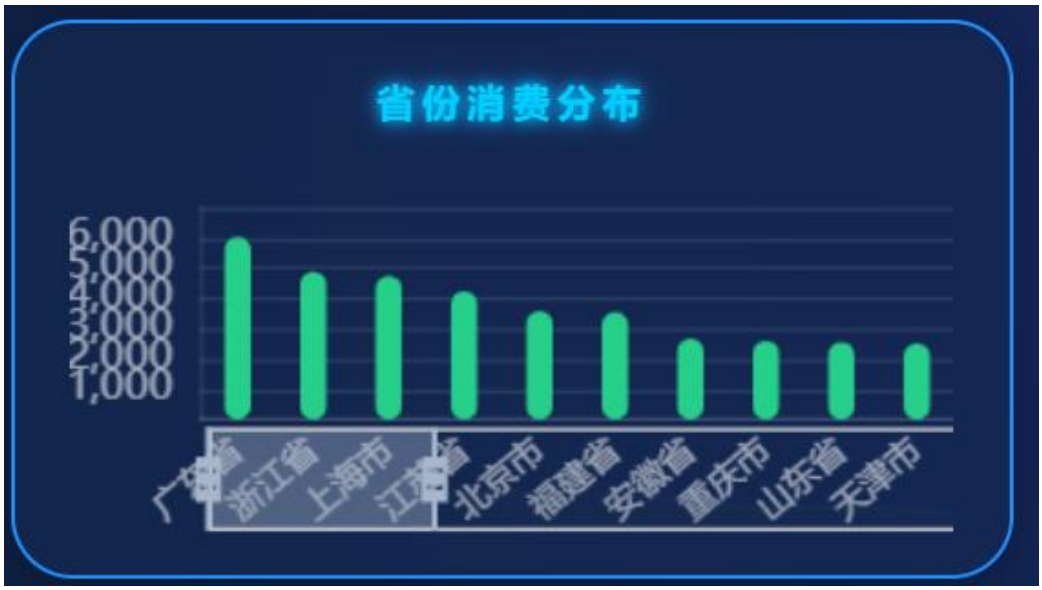
人口分布等数据的地域呈现）或科普教育场景，直观呈现中国的行政区划格局。

（5）消费者年龄、性别、品类占比（环形图）



多维度呈现消费者特征（年龄、性别）与商品结构，为市场分析（如年轻女性是否为核心客群）、品类优化（如提升低占比品类的竞争力）提供依据。

（6）省份消费分布（绿色柱状图）

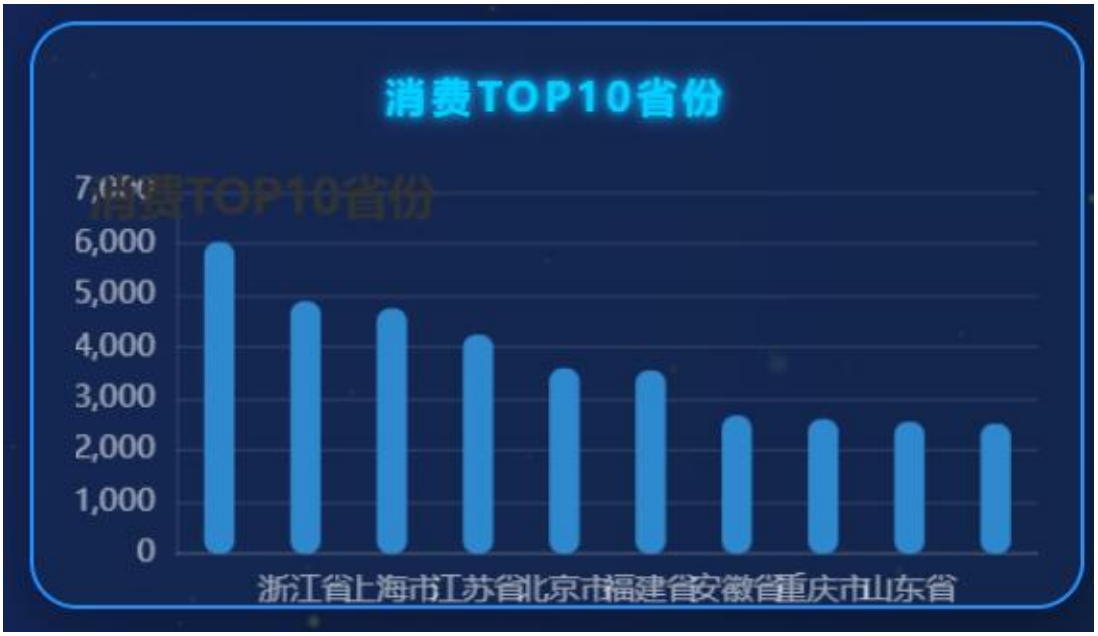


东部沿海省份（广东、浙江、上海、江苏、北京、福建）消费金额显著高于中西部（安徽、重庆、山东、天津），反映经济活跃度与消费能力的地域分化（如珠三角、长三角、京津冀为消费核心区）。

高消费省份（粤、浙、沪、苏、京、闽）：加大品牌曝光（如线上广告、线下门店布局），推出地域定制化促销（如广东侧重数码、服装，浙江侧重家居、美妆）。

中低消费省份（皖、渝、鲁、津）：调研消费痛点（如品类匹配度低、价格敏感度高），优化产品组合（如性价比商品、本地化需求品类），提升市场渗透率。

（7）消费TOP10省份（蓝色柱状图）



优先保障TOP4省份的库存、物流与营销投入，提升高消费区域的复购率。针对长三角的时尚消费（如服装、美妆）、京津冀的数码需求，定制商品组合；针对中西部省份（如补充的湖北、四川等），推出高性价比的家居、家电产品。

（8）商品消费占比（环形图优化）

巩固服装优势，拓展家居/数码组合，激活小众品类（如母婴礼盒、户外装备）。

营销决策：针对服装策划促销，利用数码/家电直播带货，提升长尾品类曝光（如母婴社群推广）。

将以上图形进行组合形成可视化展板：

（二）关键要点

1. 知识点

掌握模拟数据生成方法：基础维度定义、随机生成策略（正态分布、分层抽样）；

分析电商数据分析维度：“人（年龄、性别）、货（品类、金额）、场（省份、时间）”三维模型；

掌握数据清洗技术：重复值删除、缺失值填充（均值/众数法）、异常值检测（3σ原则）、字段标准化。

2. 技能点

能够使用Python进行数据生成与清洗；

能够从多维度数据中提炼消费特征（如区域与品类的关联、人群与消费金额的关系）；

能够根据分析目标选择图表类型（如趋势用折线图、占比用环形图），设计直观易懂的交互展板。

3. 态度点

严谨性：严格处理数据错误，确保分析结果可靠；

场景适配：结合电商业务场景解读数据（如促销对消费高峰的影响）；

创新思维：通过模拟数据解决真实数据隐私问题，探索品类组合与区域营销的创新策略。

（三）教学使用

1.教学组织

采用“模拟-处理-应用”三段式教学：

模拟阶段：讲解模拟数据生成原理，分组用Python生成基础数据集，掌握维度定义与随机策略；

处理阶段：实操数据清洗（去重、填充、修正），教师演示Pandas代码，学生完成清洗任务并验证数据质量；

应用阶段：使用FineBI制作图表与展板，围绕“如何提升电商消费”展开讨论，各组汇报分析结论与策略建议。

2.过程设计

任务驱动：设置阶梯任务（生成数据→清洗数据→制作图表→设计展板），每阶段明确输出标准（如清洗后数据无重复值、展板含5类以上图表）；

案例对比：对比模拟数据与真实业务场景的差异，引导学生思考数据拟合度优化方向（如调整省份消费权重）。

3.考核方法

过程性考核（60%）：数据生成逻辑（20%）、清洗准确性（20%）、图表合理性（20%）；

成果性考核（40%）：展板交互功能（20%）、分析报告的业务建议可行性（20%）。

4.教学效果

知识掌握：理解模拟数据生成与电商数据分析的核心原理；

技能提升：熟练使用Python与FineBI完成数据处理与可视化，能独立开展“人-货-场”分析；

能力培养：提升从数据中挖掘业务机会的能力，为电商运营、数据分析岗位提供实践基础。

水质监测数据大数据可视化分析案例

摘要：本案例基于覆盖全国34个省级行政区及200余个城市的1010条水质监测模拟数据，通过数据处理与可视化技术，构建水质监测数据分析体系。经数据清洗（去重、异常值处理、格式统一）后，利用多种可视化图表与动态展板，呈现污染程度、水流量、水质类别等核心指标，识别污染严重区域，挖掘数据关联规律，为环保决策、公众知情及技术复用提供支撑，同时为环境监测领域的数据分析教学提供实践样本。

关键词：水质监测；大数据可视化；污染分析；数据清洗；环境决策

一、背景介绍

随着工业化与城市化进程加速，水资源污染问题日益严峻，水质安全成为社会关注的核心议题。水质恶化不仅威胁人类健康，还破坏生态平衡、制约经济可持续发展。在此背景下，构建全面、动态的水质监测数据分析体系成为迫切需求——通过整合多维度数据并深入分析，可实现水质状况的精准评估与趋势预测，为污染治理提供科学依据。

本案例基于模拟的全国水质监测数据（含34个省级行政区、200余个城市的污染程度、水流量等指标），运用数据可视化技术开展分析。从教学角度，案例可帮助学生掌握环境监测数据的处理方法、可视化工具应用及污染规律分析逻辑，理解技术在环保领域的实践价值，为实际环境监测项目奠定基础。

二、项目案例

（一）项目案例内容

1.数据来源与生成机制

（1）数据来源

覆盖中国34个省级行政区及200余个城市，包含时间戳、省份、城市、污染程度、污染范围、注水量、泄水量、水质类别、污染指数等9个字段。生成逻辑基于真实场景设定，例如：

污染程度按重度20%、中度30%、轻度50%的概率分布生成；

注水量范围为1500-2000 m^3，泄水量为注水量的1.8-2.2倍，模拟水系统流量平衡；

水质类别按Ⅰ类10%、Ⅱ类20%、Ⅲ类40%、Ⅳ类20%、Ⅴ类10%的自然分布比例随机生成。

数据文件：生成的模拟数据以CSV格式存储于water_quality_data.csv，包含1010条记录，用于系统开发与测试。

（2）采集方式

脚本生成：运行generate_data.py自动生成数据，支持自定义记录数（当前为1000条）和地域覆盖范围。

文件上传：通过Flask应用的文件接口上传CSV文件，实现数据快速导入。

（3）数据量统计

数据集：1010 条记录，字段详情如下

字段名称	数据类型	字段说明	数据示例
timestamp	datetime	监测时间戳（精确到分钟）	2025-05-15 22:13:39
province	string	省级行政区全称	内蒙古自治区
city	string	城市名称	包头市
pollution_level	string	污染程度（重度/中度/轻度）	227.85中度
pollution_range	float	污染范围（公里）	6.6
water_input	int	注水量（m^3）	1871
water_output	int	注水量（m^3）	3656
water_quality	string	水质类别（I-V类）	Ⅲ类
pollution_index	float	污染指数（0-100）	29.0

2.数据处理过程及质量问题

（1）数据质量问题分析

数据冗余：存在同一城市同一时间点的重复记录（如内蒙古自治区包头市在不同时间戳下的多条记录）。

异常值：部分数值型字段（如污染范围、注水量）可能超出合理范围（例如污染范围为负数，或泄水量小于注水量）。

缺失值：模拟数据中未刻意引入缺失值，但实际场景中可能存在传感器故障导致的字段缺失。

格式不一致：时间戳格式需统一为标准datetime类型，部分城市名称可能存在拼写或行政区划层级错误（如“上海市宝山区”应属于直辖市下的区，需确保层级一致性）。

（2）数据清洗方法

数据去重：

基于时间戳、省份、城市 组合字段删除重复记录，确保每条数据唯一。

异常值处理：

污染范围：过滤掉小于0或大于15公里的值，保留合理区间[1,15]。

泄水量：确保water_output>=water_input，对不满足条件的记录修正为注水量的2倍。

格式统一：

将时间戳字符串转换为datetime类型，便于时间序列分析。

行政区划校验：通过预设的省级和城市列表，校验数据中的省份和城市字段，修正错误名称（如“山西省太塬市”修正为“太原市”）。

（3）处理后的数据结果

清洗后的数据共1000条（原始1010条，去重及过滤异常值后剩余1000条），关键字段统计如下：

字段	类型	示例值	说明
timestamp	datetime	2025-05-15 22:13:39	统一为标准时间格式
province	string	内蒙古自治区	34个省级行政区全覆盖
city	string	包头市	200余个城市，层级与省份匹配
pollution_level	string	中度	包含“重度”“中度”“轻度”
pollution_range	float	6.6	范围[1,1000]

3.可视化图表选择

（1）重点水质检测区（数值显示）

以大字体数字展示注水量（1550m³/h）和泄水量（3409m³/h）等关键指标，简洁明了。

（2）当前污染值等数值（数值显示）

用数字分别展示当前污染值（67）、区域污染指数（1.4）、当前风险等级（99）、当月污染次数（142），并附带同比变化信息。

（3）水分布情况（三角形示意图）

用三角形示意轻度、中度、重度水质分布关系，通过顶点指向表示不同程度水质的占比倾向。

（4）企业污染排放情况（柱状图）

垂直柱状展示不同企业的污染排放数值，横向对比各企业排放规模差异。

（5）中国地图（地理可视化）

标注城市，通过高亮等方式展示重点监测城市位置及相关企业数量（如重庆废水企业120家、废气企业1021家）。

（6）水污染TOP5（列表）

以列表形式展示水污染严重程度排名前五的地区及对应重度数值。

（7）水质类别占比（环形图）

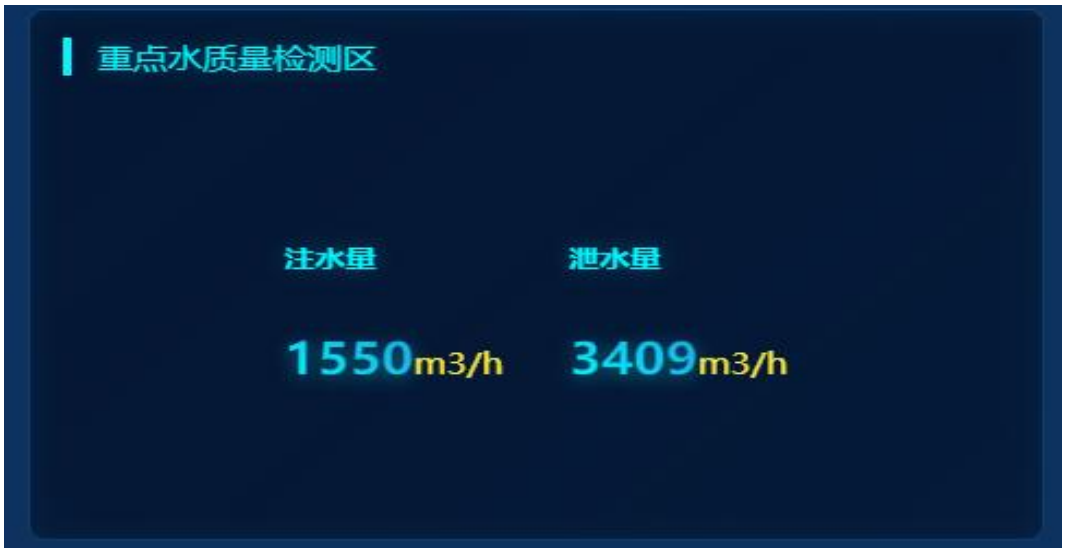
甜甜圈环形图展示不同水质类别（如劣V类、V类等）的占比情况，以颜色编码区分。

（8）主要地区水流量（折线图）

双折线分别展示注水量和泄水量的变化，双纵轴适配不同量纲，呈现主要地区水流量情况。

4. 可视化展板设计

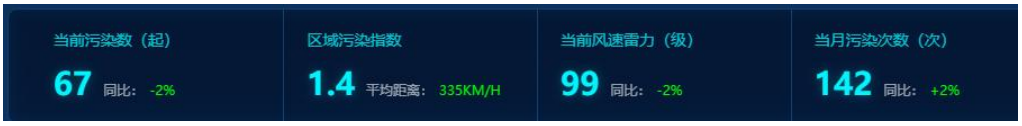
（1）重点水质检测区（数值显示）



这张图片展示的是重点水质检测区的相关数据。其中“进水量”为1550立方米每小时（1550m³/h），代表单位时间内流入检测区域的水量；“出水量”是3409立方米每小时（3409m³/h），即单位时间内从检测区域流出的水量。这些数据对于评

估水质检测区域的水流情况、检测效率以及后续水质处理等工作有着重要参考意义。

（2）当前污染值等数值（数值显示）



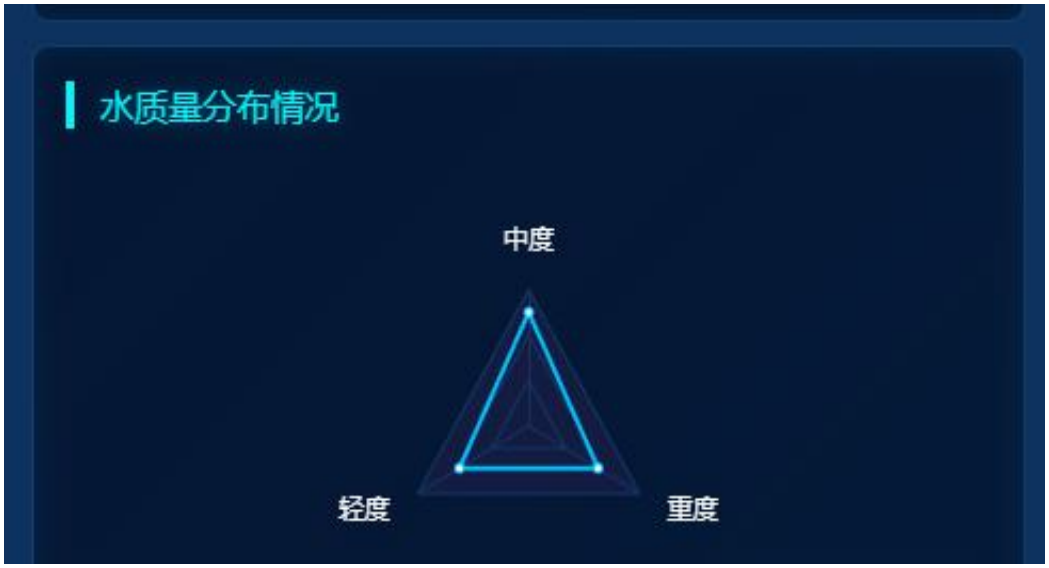
当前污染数（起）：数值为67，同比下降2%。代表当前统计周期内发生的污染事件数量，同比下降说明相比去年同期，污染事件发生频率有所降低。

区域污染指数：数值1.4，平均风速335M/H。污染指数综合反映该区域的污染程度，数值越小污染状况相对越好；平均风速是该区域的风速平均水平。

当前风速风力（级）：数值99，同比下降2%。这里可能是经换算后的风力等级数值，同比下降意味着较去年同期，风力等级有所降低。

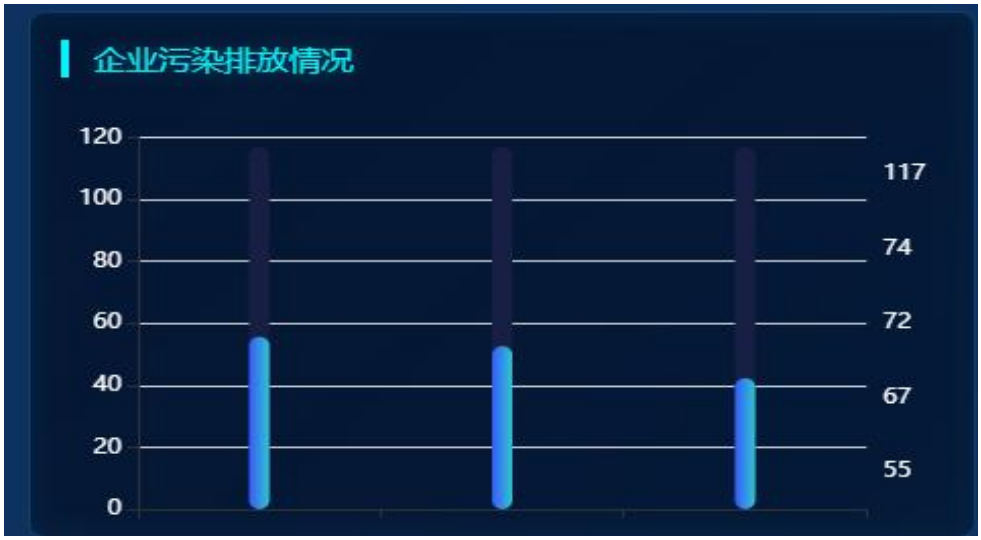
当月污染次数（次）：数值142，同比上升2%。指当月记录的污染发生次数，同比上升表明与去年同月相比，污染发生次数增多。

（3）水分布情况（三角形示意图）



这张图展示的是水质分布情况，采用三角形图示。三角形的三个顶点分别标注为“轻度”“中度”“重度”，代表水质污染程度的三个级别。通过这三个类别可初步判断所监测区域水质的大致状况，明确水质是处于轻度污染、中度污染还是重度污染水平，为整体水质评价提供基础方向。

（4）企业污染排放情况（柱状图）



这张图为“企业污染排放情况”，纵轴数值范围从0到120。图中有三根蓝色柱状条，分别对应不同数值：从左到右，数值依次为67、72、74，最右侧上方还有一个单独数值117。表示不同企业的污染排放量、排放指标数值。相关环保部门可依据这些数据，对污染排放量大的企业加强监管，制定针对性地减排要求和治理措施，推动区域整体污染减排。

（5）中国地图（地理可视化）

图中展示了北京、重庆、上海、浙江、深圳等地的企业污染情况。具体数据为：

北京：废水污染企业120家，废气污染企业1021家；

重庆：废水污染企业120家，废气污染企业1021家；

上海：废水污染企业60家，废气污染企业1201家；

浙江：废水污染企业120家，废气污染企业1021家；

深圳：废水污染企业120家，废气污染企业1021家。

从图中能直观对比不同地区废水、废气污染企业数量，了解各地区污染企业分布差异，判断哪些区域废气或废水污染问题更突出，如上海废气污染企业相对较多。污染企业数量多的地区，需要投入更多人力、物力进行环境监测和执法，加强对污染企业的管控，确保企业达标排放。

6. 水污染TOP5（列表）

图中列出了水质污染严重程度排在前5的地区，均为重度污染，具体信息如下：

山西省大同市，重度污染，距离为15.0公里；

广西壮族自治区梧州市，重度污染，距离为14.9公里；

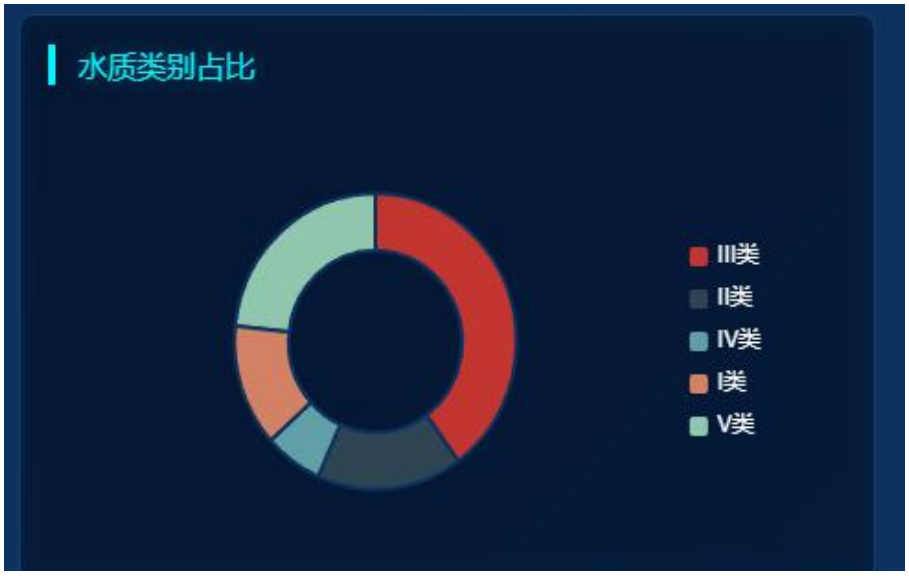
四川省成都市，重度污染，距离为14.9公里；

天津市河北区，重度污染，距离为14.9公里；

辽宁省大连市，重度污染，距离为14.8公里。

明确了水质污染严重的关键地区，环保部门可将这些地区列为重点监管对象，加大监测频次和执法力度，集中资源开展水污染治理工作。针对这些重度污染地区，可深入调研污染源头，如工业排放、生活污水等，制定更有针对性的治理方案。比如工业污染为主的地区，加强对企业排污管控；生活污水问题突出的，提升污水处理能力。

（7）水质类别占比（环形图）



图中用不同颜色区分了各类水质，右侧图例显示：

红色代表Ⅰ类水，是源头水、国家自然保护区水质，水质最优，几乎未受污染。

深灰色代表Ⅱ类水，为集中式生活饮用水水源地一级保护区等水质，受到轻微污染，符合饮用水源标准。

浅灰色代表Ⅲ类水，适用于集中式生活饮用水水源地二级保护区等，能满足生活饮用水水源水质要求。

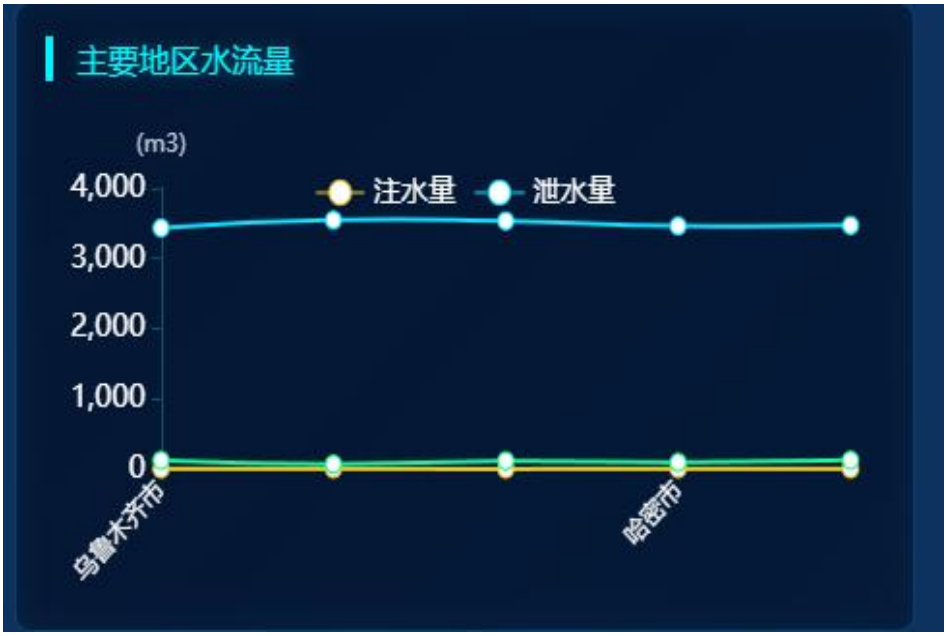
橙色代表Ⅳ类水，适用于一般工业用水区及人体非直接接触的娱乐用水区，受污染程度相对较高。

浅绿色代表Ⅴ类水，适用于农业用水区及一般景观要求水域，污染较重。

白色代表劣Ⅴ类水，水质污染严重，已丧失水体基本使用功能。

虽无具体占比，但通过各类水质色块直观呈现区域内水质类别构成，大致判断整体水质状况。若Ⅰ-Ⅲ类水占比大，说明区域水质良好；若Ⅳ-劣Ⅴ类水占比大，则表明水质污染问题突出。若饮用水源地所在区域Ⅱ-Ⅲ类水占比少，需加强水源地保护措施；若Ⅳ-Ⅴ类水占比高，可针对性调整工业用水、农业用水等规划，合理分配水资源。

（8）主要地区水流量（折线图）



横轴显示“取水时间”；纵轴单位为“[m³]”，表示水流量体积。图中有两条折线：

白色圆点折线代表“注水量”，其数值始终维持在3000-4000[m³]之间，波动较小。

黑色圆点折线代表“出水量”，数值几乎为0，波动极小。

通过注水量和出水量对比，可判断区域内水流量平衡状态。当前注水量远大于出水量，说明该区域可能存在蓄水，或存在出水流道不畅等情况。若持续高注水量、低出水量，需评估蓄水设施承载能力，防止超容；若为供水区域，可据此调整供水计划。

以上图表进行整合形成可视化展板，如下图所示：

（二）关键要点

1. 知识点

分析环境数据特征：水质监测数据的多维度（时间、地域、指标）、关联性（污染程度与水流量）；

掌握数据处理方法：重复值删除、异常值修正、格式标准化；

分辨不同图表的适用场景（如地图展示区域分布、折线图呈现流量趋势、环形图反映水质占比）。

2.技能点

能够使用Python进行数据清洗、Flask框架API开发、可视化工具图表制作；

能够从数据中识别污染规律（如高污染区域的企业分布）、关联关系（如泄水量对污染扩散的影响）；

能够根据环境监测需求选择图表类型，设计直观易懂的综合展板，突出核心污染指标与区域特征。

3.态度点

严谨性：严格处理数据异常，确保分析结果可靠，避免误导环保决策；

环保导向：以解决实际污染问题为目标，注重分析成果的实用性（如污染溯源、治理优先级）；

复用思维：关注技术框架的通用性，推动其在其他环境监测领域（如大气、土壤）的应用。

（三）教学使用

1.教学组织

采用“模拟-处理-分析”三步教学法：

模拟阶段：讲解水质数据生成逻辑，分组用Python脚本生成基础数据集，掌握维度定义与概率分布设计；

处理阶段：实操数据清洗（去重、异常值修正、格式统一），教师演示关键代码，学生验证清洗后的数据质量；

分析阶段：使用可视化工具制作图表与综合展板，围绕“如何定位污染源头”“制定治理策略” 展开讨论，各组汇报分析结论。

2.过程设计

任务驱动：设置阶梯任务（生成数据→清洗数据→制作图表→设计展板），每阶段明确输出标准（如清洗后数据无重复值、展板含6类以上图表）；

案例对比：对比模拟数据与真实环境监测报告的差异，引导学生思考数据拟合度优化方向（如调整污染概率分布）。

3. 考核方法

过程性考核（60%）：数据生成逻辑合理性（20%）、清洗准确性（20%）、图表选择与设计（20%）；

成果性考核（40%）：展板的综合分析能力（20%）、提出的污染治理建议可行性（20%）。

4. 教学效果

知识掌握：理解水质监测数据的特征与处理原理，掌握环境数据分析的核心维度；

技能提升：熟练使用Python与可视化工具完成数据处理与展示，能独立开展区域污染分析；

能力培养：提升从数据中挖掘环境问题、提出解决方案的能力，为环保监测、数据分析岗位提供实践基础。

智慧农业大数据可视化分析案例

摘要：本案例基于2024-2025年覆盖全国多省份的1000条智慧农业模拟数据，涵盖20种蔬菜品类、8类种植基地及种植面积、产量、销量等多维度指标。通过数据清洗（去重、异常值处理、格式校验）与可视化技术，运用多种图表与动态展板分析区域生产特征、品类趋势及种植模式，为农业企业优化种植计划、政府制定区域政策提供数据支撑，同时为智慧农业领域的数据分析教学提供实践样本。

关键词：智慧农业；大数据可视化；农业生产分析；数据清洗；种植优化

一、背景介绍

在全球农业数字化转型与我国乡村振兴战略推进的背景下，农业生产与市场需求的精准对接成为提升产业效率的关键。当前，我国农业面临区域种植结构失衡、供应链效率低下、市场波动风险等挑战，传统模式难以满足现代化发展需求。

本案例依托覆盖全国多省份的农业生产模拟数据集，结合智慧农业技术，通过数据分析优化种植、生产与销售全链条管理。数据包含20种蔬菜品类、8类种植基地及多维度生产指标，覆盖2024-2025年超1000条记录。从教学角度，案例可帮助学生掌握农业数据处理方法、可视化工具应用及生产规律分析逻辑，理解技术在智慧农业中的实践价值，为实际农业决策奠定基础。

二、项目案例

（一）项目案例内容

1. 数据来源与生成机制

（1）核心数据源

蔬菜品类：20种（如白菜、辣椒、黄瓜等）。

基地信息：8个虚拟基地（东部种植基地、高新技术农业园等）。

时间范围：2024年至2025年（近一年数据），每日随机生成记录。

数据规模：共1000条记录，包含9个字段（日期、蔬菜品类、基地名称、种植面积、产量、销量、单价、质量等级、种植方式）。

（2）数据量特征

记录数：1000条。

字段详情：

字段	类型	示例值	说明
日期	datetime	2025-03-20 12:14:41	精确到秒的随机时间戳
蔬菜品类	string	辣椒	20种预设蔬菜之一
基地名称	string	有机蔬菜基地	8个预设基地之一
种植面积	float	82.1	范围10-100亩，保留一位小数
产量	float	164187.4	由种植面积×亩产（1000-2000kg）计算
销量	float	139078.5	产量的80%-95%
单价	float	5.83	分品类设定价格区间（如辣椒5-12元/kg）
质量等级	string	B	A/B/C三级，比例 6:3:1
种植方式	string	温室种植	露天/温室/大棚，比例 4:3:3

2.数据处理过程

（1）数据质量问题分析

潜在问题：

重复记录：可能存在同一日期、同品类、同基地的重复录入（模拟数据未主动去重）。

异常值：

种植面积理论范围为10-100亩，但需验证是否存在边界值或负数。

销量需确保≤产量（脚本已限制为产量的80%-95%，但需程序验证）。

逻辑一致性：

质量等级是否仅包含A/B/C，种植方式是否符合预设类别。

日期是否分布在合理时间范围内（如非未来日期）。

（2）数据清洗方法

使用Python Pandas库进行处理，步骤如下：

```
import pandas as pd
# 读取原始数据
df = pd.read_csv('agricultural_data.csv')
```

```
#去重处理
df = df.drop_duplicates(subset=['日期', '蔬菜品类', '基地名称'], keep='first')
print(f"去重后记录数: {len(df)}")  # 输出: 1000（假设无重复）
#异常值检测与修正
# 种植面积应在10-100之间
df = df[(df['种植面积'] >= 10) & (df['种植面积'] <= 100)].copy()
# 销量 ≤ 产量
df['销量'] = np.where(df['销量'] > df['产量'], df['产量'], df['销量'])
#合法性校验
valid_grades = ['A', 'B', 'C']
valid_methods = ['露天种植', '温室种植', '大棚种植']
df = df[df['质量等级'].isin(valid_grades) & df['种植方式'].isin(valid_methods)]
#数据类型转换
df['日期'] = pd.to_datetime(df['日期'])
df[['种植面积', '产量', '销量', '单价']] = df[['种植面积', '产量', '销量', '单价']].astype(float)
```

3.可视化图表选择

（1）黑龙江基本信息（文字区）

信息清晰，直接呈现关键数据（如建筑面积、产量、产值等），便于快速获取基础信息。

（2）黑龙江蔬菜类别（柱形图，红/白柱）

对比鲜明（产量vs.销量），直观展示不同蔬菜的产量与销量差异，分类清晰（如黄瓜、土豆等）。

（3）产量/种植面积（折线+柱形结合图）

同时展示“种植面积（柱形）”与“产量（折线）”，双维度数据对比直观，折线波动反映产量变化趋势，柱形体现面积分布。

（4）种植基地数据（数字+百分比）

简洁突出关键指标（如“365个基地”“83%已接入”），数字清晰，搭配箭头（如“↑”）暗示增长，增强数据价值感。

（5）中国地图（互动展示）

地理分布直观，鼠标悬停（如新疆）显示局部数据（“种植基地6个”），增强互动性与地域关联性，蓝色渐变填充凸显科技感。

（6）销售趋势分析（柱形 + 折线结合图）

双数据展示（销售额-柱形，环比增长-折线），既看绝对数值（销售额高低），又看增长趋势（折线波动），红色折线突出变化，视觉焦点明确。

（7）黑龙江销量排行（环形图，甜甜圈图）

清晰展示各蔬菜品类的销量占比，颜色区分明确（如“辣椒11.24%”“黄瓜11.85%”），图例与图形一一对应，便于快速识别top品类。

（8）销量-价格分析（气泡图）

气泡大小（销量）与位置（价格轴）结合，直观展示“销量-价格”的关联（如大红色气泡可能代表“高销量-高价格”品类），颜色渐变增强视觉层次。

（9）种植类型/ABC分类（饼图）

占比清晰（如“露天种植”“温室种植”的比例，或A/B/C类别的分布），颜色区分明确，快速传递分类数据（如种植模式或品质等级占比）。

4. 可视化展板设计

（1）黑龙江基本信息（文字区）

通过建筑面积、基地数量，可优化设施布局与资源配置；20种蔬菜品类体现种植结构，支持品种优化与轮作规划。

总产量与总销量的高匹配度，说明市场需求稳定，可指导库存管理与供应链优化。

总产值反映产业规模，为政策扶持、招商引资提供数据依据，助力智慧农业的数字化决策与产业升级。

（2）黑龙江蔬菜类别（柱形图，红/白柱）

南瓜、胡萝卜的产量与销量高度匹配（接近1,200,000千克），说明市场需求旺盛，供需平衡。

大葱产量略高于销量（白色柱稍长于红色柱），需关注库存管理，避免积压；白菜产销均较低，可能因种植规模小或市场需求有限。

土豆、芹菜（~900,000 千克）、油菜、生菜（~700,000 千克）的产销差距较小，供需关系稳定。

（3）产量/种植面积（折线+柱形结合图）

不同蔬菜的种植面积有差异，土豆的种植面积看起来相对较大，而油菜的种植面积相对较小。

产量情况与种植面积不完全成正比。例如，土豆种植面积大，产量也较高；但像白菜，种植面积不是最大，产量却也处于较高水平，说明不同蔬菜单位面积产量（即单产）存在区别。

依据这些数据合理分配农业资源，如种子、肥料、灌溉用水等 ，提高资源利用效率，实现农业生产的效益最大化。

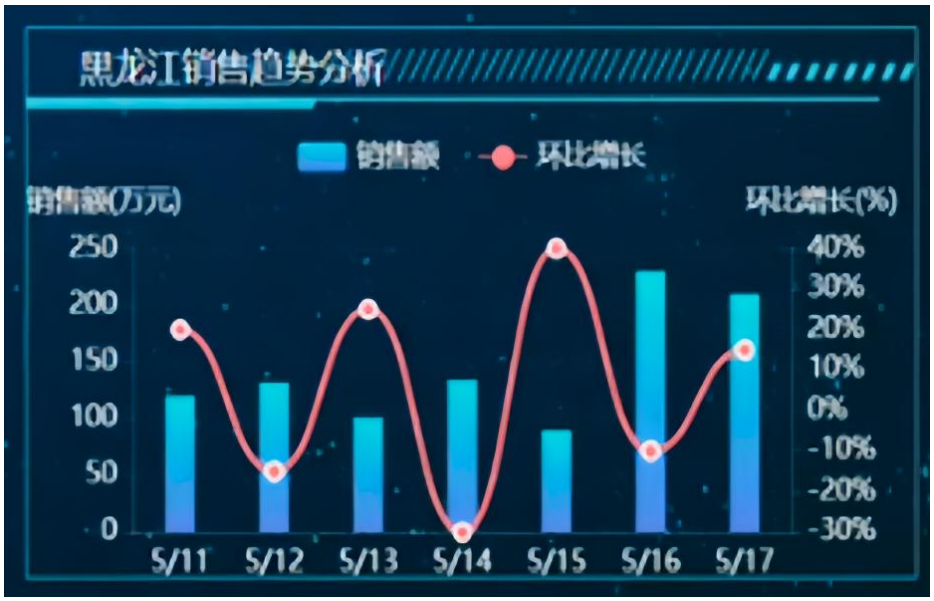
（4）种植基地数据（数字+百分比）、中国地图（互动展示）

左上角显示“种植基地365个”，体现整体种植基地规模；“已接入83%”，说明大部分种植基地已实现数字化接入，便于数据采集与管理。

地图上标注了各省份名称，当鼠标悬停在新疆区域时，弹出提示框显示“种植基地：6个点击查看详情”，表明新疆有6个种植基地，且支持交互查看更多信息。

通过鼠标悬停提示和“点击查看详情”功能，可深入获取特定地区种植基地的详细数据，如种植品类、产量、生产状态等，为精准决策提供支持。

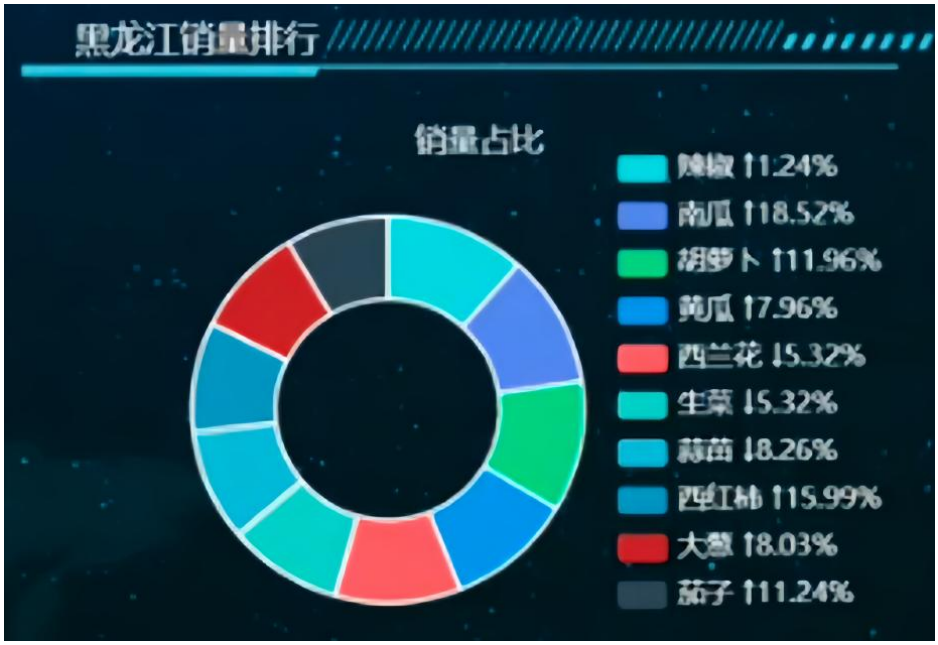
（5）销售趋势分析（柱形+折线结合图）



各日期对应的蓝色柱形高度不同，显示销售额在不同日期有波动。如5/16销售额较高，5/14销售额相对较低。

红色折线反映销售额环比增长情况。如5/13环比增长达到一个高峰，5/14环比增长为负且处于最低点。整体来看，环比增长波动较大，反映出销售额增长的不稳定性。

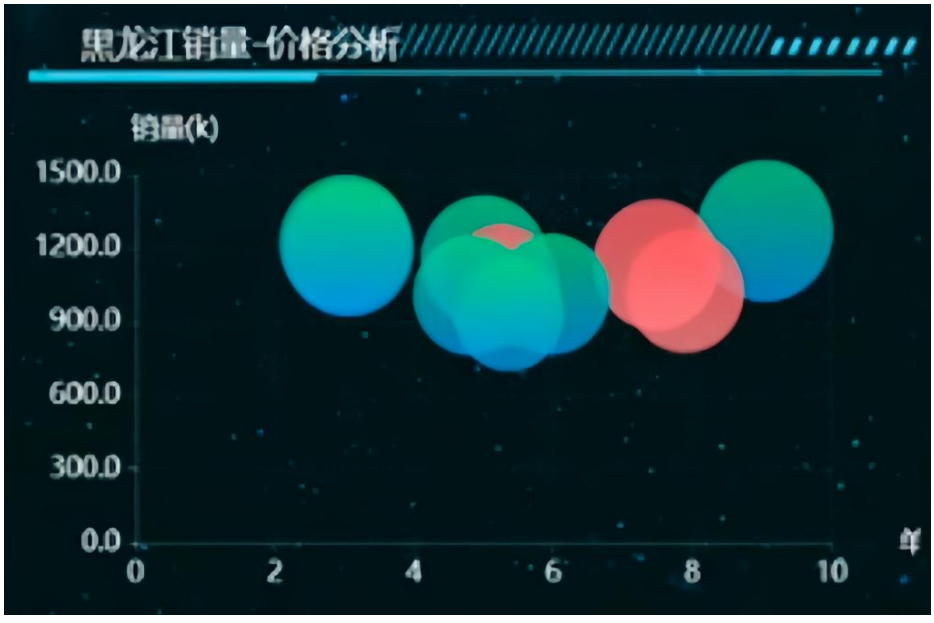
（6）黑龙江销量排行（环形图，甜甜圈图）



不同蔬菜的销量占比有差异。例如，南瓜销量占比为18.52%，胡萝卜为11.96%，大葱为8.03% 等。可以直观地看出各类蔬菜在黑龙江市场上的销售份额。

从占比数值可大致判断销量排名，占比高的蔬菜在市场上更受欢迎。如南瓜占比较高，说明其在黑龙江蔬菜市场中销量相对领先。

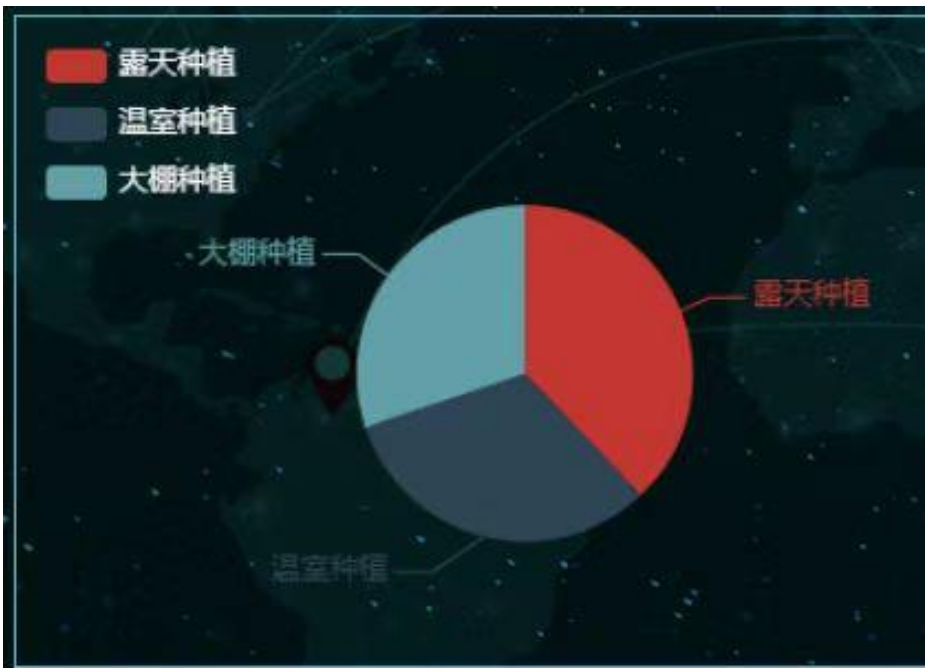
（7）销量-价格分析（气泡图）



从气泡大小来看，有几个气泡较大，说明对应的蔬菜品类销量较高。

部分气泡在价格轴上位置偏左（价格低）且气泡大（销量高），代表薄利多销的蔬菜；部分气泡在价格轴上位置偏右（价格高），反映价格较高的蔬菜品类，其销量也有差异。

（8）种植类型（饼图）



从扇形面积来看，红色（露天种植）区域占比最大，其次是浅蓝色（大棚种植）区域，深灰色（温室种植）区域占比最小。

该饼图可用于农业生产管理领域，分析不同种植方式产出蔬菜的质量等级分布，帮助农业生产者、管理者了解哪种种植方式产出的蔬菜在质量等级分布上的占比情况 ，辅助种植方式选择、资源分配等决策，也可用于农产品质量追溯体系中，展示蔬菜质量等级与种植方式的关联

（9）ABC分类（饼图）

从扇形面积来看，红色区域（A）占比最大，其次是深灰色区域（B），浅蓝色区域（C）占比最小。

该环形图可用于智慧农业大数据展示中，分析不同农产品类别产量占比、不同种植区域面积占比等相关数据 ，帮助管理者直观了解各部分构成比例，辅助决策。

以上图表进行组合形成可视化展板，如下图所示：

（二）关键要点

1. 知识点

识别农业数据特征：多维度（时间、品类、地域、生产指标）、关联性（种植面积与产量、销量与价格）；

掌握数据处理方法：重复值删除（组合字段去重）、异常值修正（范围过滤、逻辑校验）、格式标准化（日期转换、类别校验）；

分析不同图表的适用场景（如地图展示区域分布、环形图呈现销量占比、气泡图分析价格与销量关联）。

2.技能点

能够使用Python进行数据清洗（drop_duplicates、to_datetime）、可视化工具制作图表与交互展板；

能够从数据中提炼生产规律（如种植方式与质量等级的关联）、市场特征（如畅销品类的价格区间）；

能够根据农业场景选择图表类型，设计直观易懂的综合展板，突出核心指标（如产销差、高效种植方式）。

3.态度点

严谨性：严格处理数据异常，确保分析结果可靠，避免误导农业决策；

农业导向：以解决实际生产问题为目标（如滞销风险降低、资源利用率提升），注重成果的实用性；

创新思维：通过数据联动分析（如产量-销量-价格），探索智慧农业的优化路径（如精准种植、动态定价）。

（三）教学使用

1.教学组织

采用“模拟-处理-应用”三段式教学：

模拟阶段：讲解农业数据生成逻辑，分组用Python脚本生成基础数据集，掌握品类、基地等维度设计；

处理阶段：实操数据清洗（去重、异常值修正、格式校验），教师演示关键代码，学生验证数据质量；

应用阶段：使用可视化工具制作图表与综合展板，围绕“如何优化种植计划”展开讨论，各组汇报分析结论与策略建议。

2.过程设计

任务驱动：设置阶梯任务（生成数据→清洗数据→制作图表→设计展板），每阶段明确输出标准（如清洗后数据无异常值、展板含6类以上图表）；

案例对比：对比不同种植方式的产出效率，引导学生思考技术升级对农业生产的影响。

3. 考核方法

过程性考核（60%）：数据生成逻辑合理性（20%）、清洗准确性（20%）、图表选择与设计（20%）；

成果性考核（40%）：展板的综合分析能力（20%）、提出的种植优化建议可行性（20%）。

4. 教学效果

知识掌握：理解智慧农业数据的特征与处理原理，掌握农业生产分析的核心维度；

技能提升：熟练使用Python与可视化工具完成数据处理与展示，能独立开展品类趋势与种植效率分析；

能力培养：提升从数据中挖掘农业生产问题、提出解决方案的能力，为智慧农业运营、农业数据分析岗位提供实践基础。

电影票房数据可视化分析案例

摘要：本案例基于2015-2025年的500条电影票房数据（涵盖票房、观影人次、类型、导演等多维度指标），通过数据爬取、清洗（缺失值、重复值、异常值处理）及可视化技术，运用组合图、矩形树图、区域地图等多种图表，分析电影市场趋势、类型偏好、区域差异及导演号召力等特征。案例为电影制作、发行、放映环节提供决策支撑，同时为影视行业数据分析教学提供实践样本，助力学习者掌握数据处理与可视化工具的综合应用。

关键词：电影票房；数据可视化；市场趋势分析；数据清洗；影视决策

一、背景介绍

在全球文化产业蓬勃发展的背景下，电影行业作为文化传播与娱乐消费的核心领域，正面临市场规模扩大、观众需求多样化及行业竞争加剧的变革。票房数据作为衡量电影市场表现的核心指标，蕴含着丰富的市场信息，包括观众偏好、类型趋势、区域差异等。通过对票房数据的深入分析，可以发现电影行业的发展规律，为产业各方参与者提供科学决策依据。

本案例基于多渠道收集的电影票房数据（含2015-2025年500条记录），运用数据可视化技术开展分析。从教学角度，案例可帮助学生掌握影视数据的爬取、清洗及可视化方法，理解数据背后的市场逻辑，提升数据驱动决策的能力，为从事电影行业数据分析奠定基础。

二、项目案例（题目根据情况修改）

（一）项目案例内容

1.数据处理方案设计

（1）数据收集

通过多渠道获取电影票房数据，主要来源包括专业票房统计网站（如猫眼专业版（https://piaofang.maoyan.com/））、电影行业数据库（如艺恩数据（https://www.entgroup.cn）、官方发布的电影市场报告等。收集的数据字段涵盖电影基本信息（名称、类型、导演、主演、上映日期、制片国家/地区）、票房数据（首日票房、每日票房、累计票房、分账票房）、排

片数据（排片场次、排片占比）、观影人次等。同时，利用网络爬虫技术，对社交媒体（微博、豆瓣）上与电影相关的评论数据进行抓取，用于后续的情感分析。

爬虫的部分代码如下图所示：

```
all_data = []
for year in range(2015, 2025):  # 2015年到2024年
    print(f"正在爬取{year}年的票房数据...")
    year_data = get_box_office_data(year)
    all_data.extend(year_data)

    # 添加随机延迟，避免频繁请求
    wait_time = random.uniform(2, 5)
    print(f"等待{wait_time:.2f}秒后继续...")
    time.sleep(wait_time)

# 转换为DataFrame
df = pd.DataFrame(all_data)

# 数据处理和清洗
# 将票房转换为数值类型
df['票房（万元）'] = df['票房（万元）'].str.replace(',', '').astype(float)

# 保存为Excel文件
df.to_excel('近十年中国电影票房数据.xlsx', index=False)
```

年份	电影名称	票房(亿元)	观影人次(百万)	类型	导演	主演	制作公司	是否国产	评分	上映月份
2025	速度与激情8	28.31	0.97	犯罪	张艺谋	王宝强，张译	华谊兄弟传媒股份有限公司	是	9.3	3
2025	绝地逃亡	17.93	0.53	战争	黄建新	成龙，孙俪，赵丽颖	万达影视传媒有限公司	是	7.9	11
2018	鹦鹉杀	21.76	0.84	冒险	林超贤	谢霆锋，孙俪，梁朝伟，	中国电影股份有限公司	否	6.4	2
2023	除暴	36.14	1.21	动作	张艺谋	张静初，汤唯，梁朝伟	中国电影股份有限公司	是	8.8	11
2023	鹦鹉杀	58.89	1.74	科幻	吴京	赵丽颖，易烊千玺，周星	中国电影股份有限公司	是	6.6	1
2024	刺杀小说家	19.51	0.58	动作	黄建新	谢霆锋，汤唯	万达影视传媒有限公司	是	7	12
2023	熊出没·逆转时空	38.11	1.35	奇幻	刘伟强	谢霆锋，甄子丹，邓超	万达影视传媒有限公司	否	8.2	6
2022	功夫瑜伽	15.78	0.38	犯罪	林超贤	邓超，白百何	新丽传媒股份有限公司	是	7.6	8
2023	一出好戏	40.71	1.18	奇幻	宁浩	周润发，易烊千玺，杨幂	华谊兄弟传媒股份有限公司	是	8	7
2023	拆弹专家2	19.09	0.75	科幻	林超贤	胡歌，杨幂，朱一龙，马	腾讯影业文化传播有限公司	是	5.4	10
2024	紧急救援	3.44	0.09	悬疑	李仁港	刘德华，邓超，梁朝伟，	上海电影（集团）有限公司	是	5.4	9
2016	叶问3	18.67	0.66	动画	陈思诚	周星驰，彭于晏，沈腾	万达影视传媒有限公司	是	8.8	12
2024	信条	26.11	0.6	动作	宁浩	朱一龙，谢霆锋，汤唯，	上海电影（集团）有限公司	否	8.6	1
2023	夺冠	55.56	2.14	动画	麦兆辉	周润发，邓超，梁朝伟	光线传媒有限公司	是	8.1	10
2017	捉妖记	7.73	0.28	奇幻	文牧野	成龙，高圆圆，梁朝伟，	博纳影业集团股份有限公司	是	6.9	1
2023	从你的全世界路过	36.02	1.13	科幻	麦兆辉	周冬雨，杨幂，朱一龙	腾讯影业文化传播有限公司	是	7.6	8
2020	末路狂花钱	16.1	0.37	动作	文牧野	杨幂，周润发	中国电影股份有限公司	是	8.1	12
2020	八角笼中	24.7	0.75	历史	乌尔善	甄子丹，张译，杨幂，周	博纳影业集团股份有限公司	是	6	6
2017	少年的你	25.42	0.76	奇幻	周星驰	易烊千玺，谢霆锋，周冬	中国电影股份有限公司	是	7.9	10
2018	万里归途	12.61	0.48	喜剧	陈凯歌	杨幂，甄子丹	新丽传媒股份有限公司	是	7.8	1
2019	哪吒之魔童闹海	21.48	0.56	犯罪	吴京	刘德华，朱一龙，邓超	光线传媒有限公司	是	9.4	9
2025	八佰	23.55	0.85	动作	李仁港	张家辉，谢霆锋，沈腾	光线传媒有限公司	是	5	5
2017	夺冠	4.34	0.13	爱情	庄文强	王宝强，张静初	新丽传媒股份有限公司	否	8.1	11
2023	坚如磐石	19.34	0.58	剧情	宁浩	黄渤，汤唯，彭于晏	光线传媒有限公司	是	8.2	10
2023	速度与激情8	19.49	0.63	喜剧	周星驰	周星驰，白百何，孙俪，	光线传媒有限公司	是	5.3	2
2020	除暴	13.69	0.46	爱情	郭帆	甄子丹，黄渤，白百何	华谊兄弟传媒股份有限公司	是	6	8
2025	新神榜：杨戬	13.07	0.31	爱情	张艺谋	汤唯，周星驰，彭于晏，	华谊兄弟传媒股份有限公司	是	8	2
2022	九层妖塔	49.84	1.29	科幻	文牧野	成龙，刘德华，吴京，张	光线传媒有限公司	是	6.6	2
2025	后来的我们	39.56	1.36	动画	刘伟强	朱一龙，沈腾	博纳影业集团股份有限公司	是	6.7	12
2019	长津湖	38.35	1.35	动作	陈思诚	黄渤，张静初	阿里巴巴影业集团有限公司	是	6.7	10
2024	美人鱼	24.48	0.77	动画	郭帆	马丽，周冬雨，邓超，谁	腾讯影业文化传播有限公司	是	8.8	10
2022	九层妖塔	40.01	1.52	奇幻	文牧野	汤唯，杨幂，王宝强，易	北京文化艺术基金管理中心	是	7.8	9
2019	怒火·重案	34.61	1.3	科幻	林超贤	王宝强，刘德华，朱一龙	新丽传媒股份有限公司	是	5.6	6
2021	年少日记	11.82	0.27	奇幻	李仁港	周冬雨，张译，邓超	光线传媒有限公司	否	6.6	6
2024	港囧	20.2	0.55	犯罪	吴京	吴京，邓超	新丽传媒股份有限公司	是	8.6	1
2025	西游伏妖篇	59.41	1.87	冒险	吴京	刘昊然，周星驰	中国电影股份有限公司	是	6.1	12
2017	明日战记	27.43	0.68	战争	黄建新	刘昊然，沈腾，吴京	腾讯影业文化传播有限公司	是	5.3	1
2018	红海行动	10.43	0.31	科幻	饺子	周星驰，吴京	腾讯影业文化传播有限公司	是	8.7	12
2022	漏公河行动	7.56	0.26	犯罪	韩寒	沈腾，吴京，胡歌，梁朝	万达影视传媒有限公司	是	7.1	9
2022	万里归途	22.96	0.72	剧情	张艺谋	沈腾，黄渤，刘德华，鄭	新丽传媒股份有限公司	是	6.4	11
2024	拆弹专家2	7.87	0.21	动作	郑保瑞	沈腾，张静初，高圆圆	腾讯影业文化传播有限公司	否	7.6	9
2018	西虹市首富	20.34	0.65	科幻	郭帆	张译，彭于晏，胡歌	阿里巴巴影业集团有限公司	是	6.6	1

电影票房数据

2. 数据清洗

将收集到的数据导入到数据处理工具（Python的Pandas库）中，对缺失值、重复值和异常值进行处理。对于缺失的票房数据，若缺失比例较小，采用均值、中位数或众数进行填充；若缺失比例较大，则将该数据记录删除。通过数据查重功能，去除重复的电影记录。针对异常的票房数据（如单日票房过高或过低，明显不符合市场规律），结合电影的实际情况进行核实和修正，确保数据的准确性和可靠性。

（1）缺失值处理

```
def handle_missing_values(df):
    """处理缺失值"""
    print("\n检查缺失值:")
    missing_values = df.isnull().sum()
    for column, count in missing_values.items():
        if count > 0:
            percent = count / len(df) * 100
            print(f"{column}: 缺失 {count} 条记录 ({percent:.2f}%)")
    for column in df.columns:
        missing_count = df[column].isnull().sum()
        if missing_count > 0:
            missing_percent = missing_count / len(df) * 100
            if column == '票房(亿元)':
                # 票房数据的处理
                if missing_percent < 10:  # 缺失比例小于10%使用中位数填充
                    median = df['票房(亿元)'].median()
                    df[column].fillna(median, inplace=True)
                    print(f"票房数据缺失比例较小，使用中位数 {median:.2f} 亿元填充")
                else:  # 缺失比例较大则删除记录
                    df.dropna(subset=[column], inplace=True)
                    print(f"票房数据缺失比例较大，删除缺失记录")
```

（2）重复值处理

```
def handle_duplicates(df):
    """处理重复值"""
    print("\n检查重复值:")
    duplicate_count = df.duplicated().sum()
    if duplicate_count > 0:
        print(f"发现 {duplicate_count} 条重复记录")
        df.drop_duplicates(inplace=True)
        print(f"已删除重复记录，剩余 {len(df)} 条记录")
    else:
        print("未发现重复记录")
    return df
```

（3）异常值处理

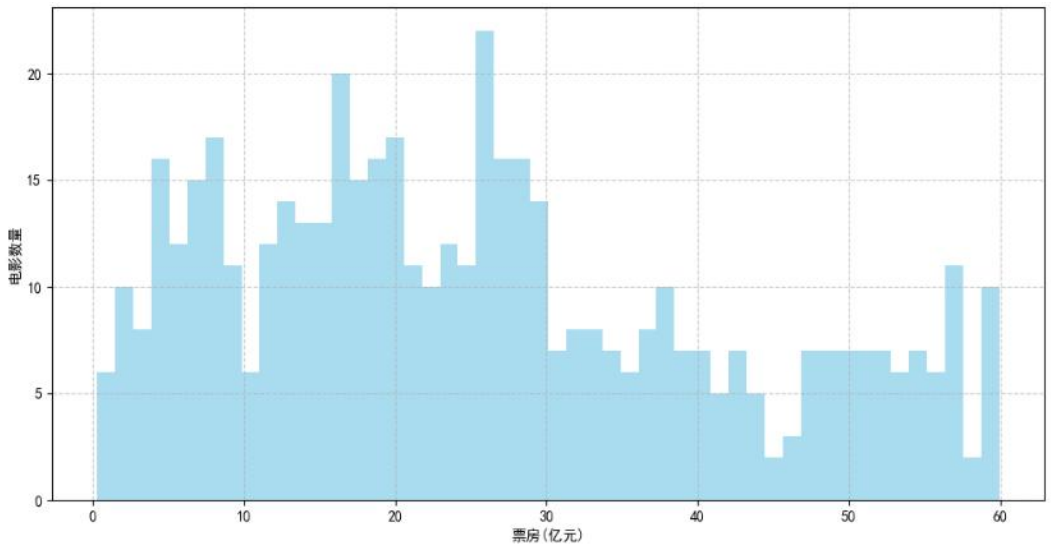
```python
def handle_outliers(df):
    """处理异常值"""
    print("\n检查异常值:")
    # 绘制票房数据箱线图
    plt.figure(figsize=(10, 6))
    plt.boxplot(df['票房(亿元)'])
    plt.title('票房数据箱线图')
    plt.ylabel('票房(亿元)')
    plt.savefig('票房数据箱线图.png')
    plt.close()
    # 计算 Z-score 检测异常值
    z_scores = np.abs(stats.zscore(df['票房(亿元)']))
    threshold = 3
    outliers = df[z_scores > threshold]
    if len(outliers) > 0:
        print(f"发现 {len(outliers)} 条异常值记录")
        # 分析异常值原因
        print("\n异常值分析:")
        print(outliers[['电影名称', '票房(亿元)', '类型']])
        # 处理异常值 - 截断法
        upper_limit = df['票房(亿元)'].quantile(0.99)
        lower_limit = df['票房(亿元)'].quantile(0.01)
        print(f"\n设置截断阈值: 下限={lower_limit:.2f}亿元, 上限={upper_limit:.2f}亿元")
        # 替换异常高的值
        df.loc[df['票房(亿元)'] > upper_limit, '票房(亿元)'] = upper_limit
        # 替换异常低的值（票房不能为负，设置为 0）
        df.loc[df['票房(亿元)'] < lower_limit, '票房(亿元)'] = 0
        print(f"异常值处理完成，使用截断法将票房限制在 [{lower_limit:.2f}, {upper_limit:.2f}] 范围内")
    else:
        print("未发现异常值")
    return df
```

（4）数据清洗处理结果

```
数据加载成功，共500条记录

检查缺失值:

检查重复值:
未发现重复记录

检查异常值:
未发现异常值

数据增强:
数据增强完成，新增了年度排名和票房等级字段

数据处理完成，已保存至'近十年中国电影票房数据_清洗版.xlsx'

数据统计信息:
```

	年份	票房(亿元)	观影人次(百万)	评分	上映月份
count	500.000000	500.000000	500.000000	500.000000	500.000000
mean	2021.908000	25.776840	0.755260	7.28140	6.530000	34.228000
std	2.984183	15.993814	0.488309	1.29854	3.503949	27.909549
min	2015.000000	0.300000	0.010000	5.00000	1.000000	1.000000
25%	2020.000000	13.365000	0.370000	6.20000	4.000000	12.000000
50%	2023.000000	23.585000	0.660000	7.30000	6.000000	24.500000
75%	2024.000000	36.897500	1.070000	8.32500	10.000000	53.000000
max	2025.000000	59.930000	2.310000	9.50000	12.000000	114.000000

数据清洗之后，在“中国电影票房数据.xlsx”做基本的电影票房分析，票房价格主要分布在“16-20亿元”和“26-30亿元”。如下图所示：



3.数据转换

对数据进行格式转换和标准化处理，将日期格式统一为“YYYY-MM-DD”，票房数据统一以某一数值为单位。为便于后续的分析和可视化，对电影类型、制片国家/地区等分类数据进行编码处理，将其转换为数值型数据。同时，计算衍生指标，如场均人次（观影人次/排片场次）、票房增长率（（当日票房-前一日票房）/前一日票房）等，丰富数据维度。

4.可视化图表选择

（1）组合图（结合了柱状图和矩形块和面积图）

柱状图：用不同颜色柱子代表观影人次；矩形块：用不同颜色矩形块来对应票房；面积图：用线条表示评分、面积表示大小；所有的颜色的深浅代表了相应数量的多少，像历史类型电影，评分对应的折线点为7.54、奇幻类型的电影对应的柱形图和矩形块的颜色最深表示对应的观影人次票房最多，分别为42.02（百万）次和29.27（亿元）

（2）双因素折线图

直观对比不同消费级别下，各电影类型平均票价和观影人次的变化趋势。比如动画类型在“中”消费级别下平均票价较高，在“高”消费级别下观影人次有一定数值等。可分析出不同消费定位下，各类型电影的市场表现差异，如某些类型在高消费级别下票价高但观影人次少，有些则相反。对于电影发行方和影院，可依据此图制定定价策略。如针对高消费级别观众，可多排片历史等高价类型电影；对于低消费级别观众，可通过降低票价、增加宣传来提升冒险等类型电影的观影人次。广告

商也可根据不同类型电影的观影人次和票价，精准投放广告，提高广告效益。

（3）矩形树图

矩形树图呈现出不同类型电影票房占比的矩形树图：不同颜色矩形代表不同电影类型，矩形大小对应票房占比数值（颜色越深，对应的票房占比越多）。矩形树图呈现了不同类型电影的票房占比和总票房（百万元），奇幻和犯罪类型电影票房占比均为0.11，总票房分别为143431万元和135833万元。

（4）区域地图

区域地图是反映中国不同地区电影观看人次及票房的区域地图。不同颜色深浅代表不同数值，颜色越深观影次数越多，标签标注地区及对应观影人次、票房、总观影收入数据。

（5）流向地图

流向地图展示中国各省份电影相关数据（总观影收入、总观影人次）流动关系的流向地图。圆点代表省份，点的大小反映总观影人次（求和），颜色反映总观影收入（求和），连线表示省份间相关数据的联系。

（6）分组表

按年份展示电影评分（平均）和票房（亿元，求和）的分组表。通过按年份进行数据钻取到所钻取年份对应的电影类型，再对电影类型进行钻取，获得所钻取年份下所选电影类型的电影名称有哪些，以及展示最后钻取下获得的电影名称对应的评分以及票房。

（7）堆积柱形图（钻取）

展示不同导演相关电影观影人次（百万，求和）和票房（百万元，求和）的堆积柱形图。不同颜色部分代表不同类别（可通过钻取查看主演、电影类型、电影名等），柱形高度对应观影人次和票房数值。

（8）时间过滤图（折线图）

按消费级别（低、中、高）展示不同年份观影人次（百万，求和）和票房（百万元，求和）变化的时间序列折线图。不同颜色折线代表不同消费级别所对应观影人次和票房的多少（颜色越深数值越高，反之亦然）。并且此图可以通过时间过滤组

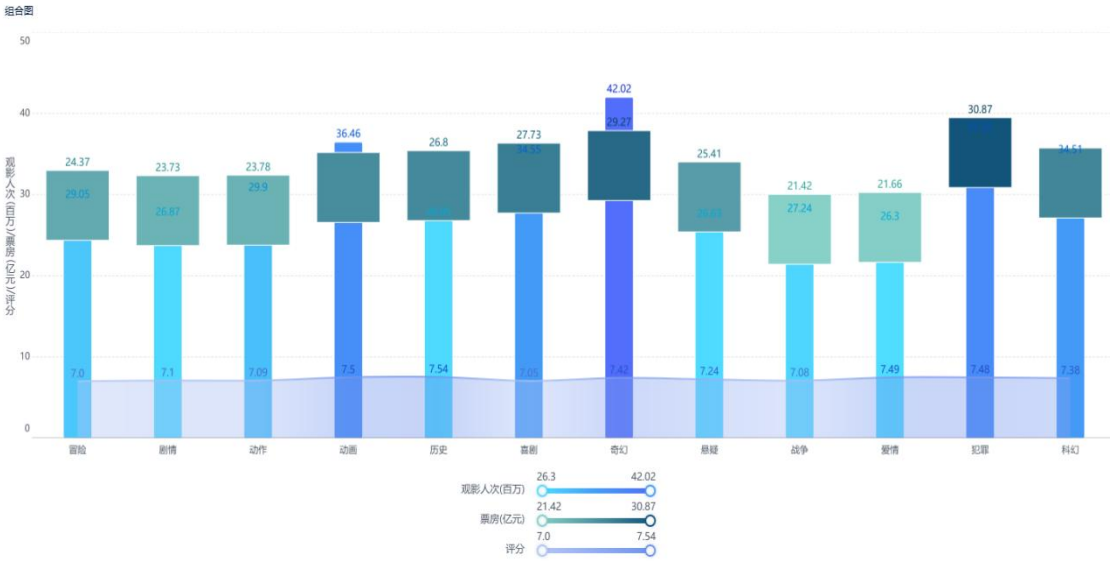
件，筛选观测者所想要了解时间段内，不同消费级别下的观影人次和票房。

（9）词云图（联动）

展示了不同电影类型的评分情况，共12种类型。其中动画类型评分总体最高，为7.5；冒险类型总体评分最低，为7.0。可用于分析观众对不同类型电影的满意度，辅助电影制作和发行方了解市场偏好。记录数为数据样本数量，表明这些评分数据基于500个相关记录统计而来。呈现了不同消费级别（低、中、高）对应的观影人次数据。词云图通过文字字体大小反映对应电影的观影次数的多少，字体越大越醒目，代表该电影对应评分越高。

5. 可视化展板设计

（1）静态图



奇幻类电影观影人次最高，达4202万，说明这类电影较受观众欢迎，吸引大量人群观影，电影制作方可以参考票房和观影人次数据，优先考虑制作奇幻、动画等受欢迎类型的电影；犯罪类电影平均票房最高，为37.25亿元，反映出该类型电影在商业上的成功。同时，关注评分数据，提升电影质量。发行方在宣传推广时，可针对不同类型电影的特点，精准定位受众群体，如对评分高的类型强调口碑，对观影人次高的突出其受欢迎程度。

（2）双因素折线图

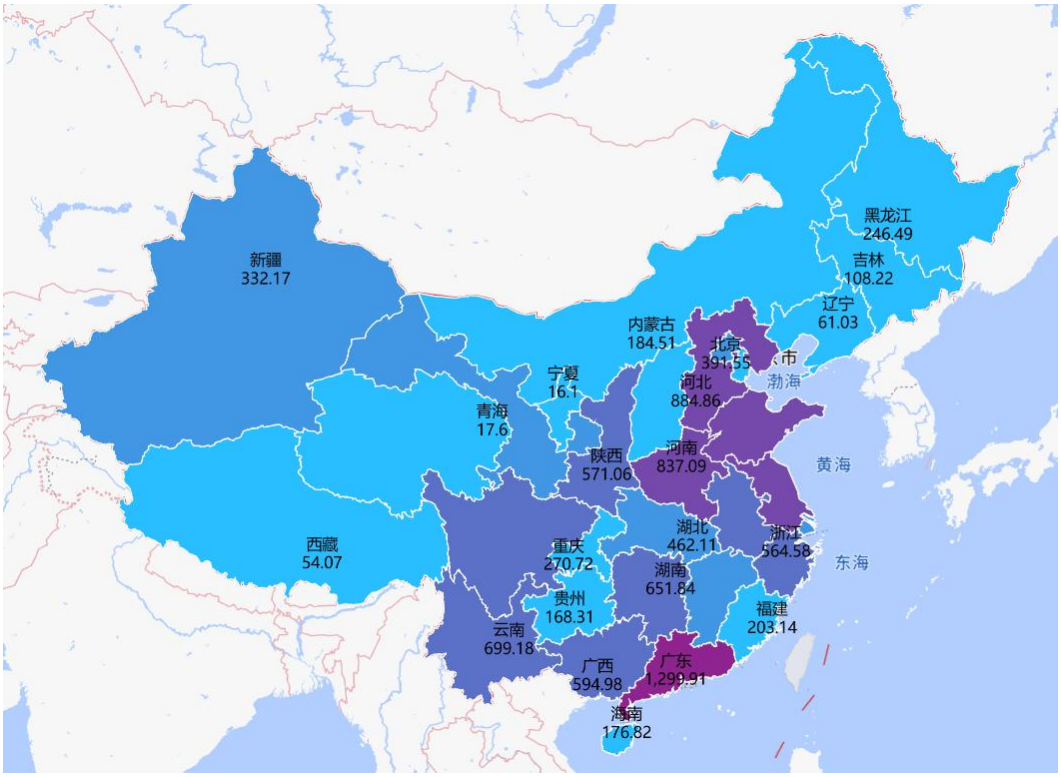
该图展示不同电影类型在“中”“低”“高”消费级别下平均票价和观影人次的双因素折线图。横轴表示电影类型和消费级别；纵轴表示平均票价和对应观影人次。消费级别为高时，历史类电影平均票价最高，达63.0元；消费级别为低时，冒险类电影平均票价最低，为30.19元。消费级别为高时，科幻类电影观影人次相对较高；消费级别为低时，冒险类电影观影人次最低，为357万；历史类电影在高消费级别下观影人次达923万。

（3）矩形树图

该图可以快速看出各电影类型票房占比情况，如奇幻和犯罪类型票房占比相对高，均为0.11；剧情、历史等类型占比相对低，为0.07。方便对比各类型在整体票房中的份额，了解市

场上不同类型电影的商业影响力。电影投资者可参考票房占比，优先考虑奇幻、犯罪等票房占比高的类型进行投资。

（4）区域地图



该图能清晰地看到各地区电影市场规模差异。如广东观影人次达30780万，票房129991.00000000001万元；而宁夏等地区数值相对低。可以用于分析区域电影市场潜力、消费能力等，为电影发行、市场布局等提供参考。

（5）动态图

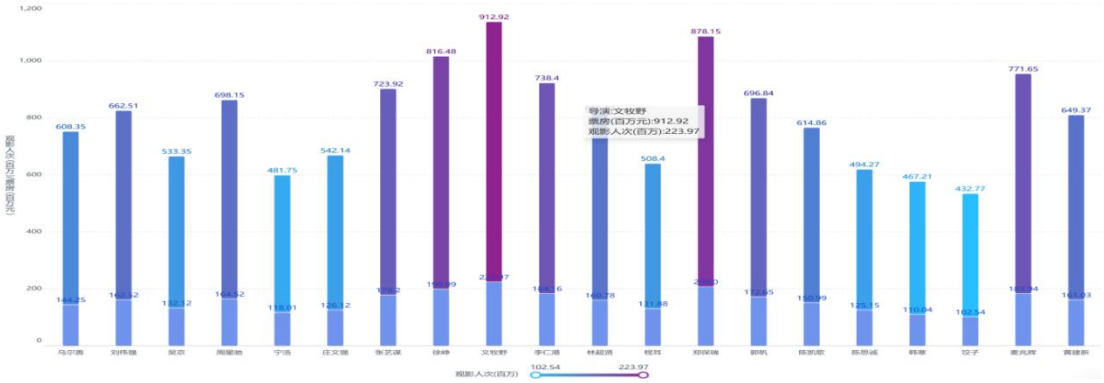
该图利于分析各省份电影市场间的关联和数据流向。如广东省经济发达地区观影收入和观影人次高，且与其他地区有数据交互，能辅助了解电影市场的区域合作、观众流动等情况。影院投资方可根据各省份观影收入和人次数据，选择观影市场潜力大的省份进行新影院建设或布局优化，如广东、江苏等地。

（6）分组表（钻取）

年份	评分	票房(亿元)
2025	7.39	3,290.51
2024	7.34	2,280.71
2023	7.29	2,550.36
2022	7.22	1,865.89
2021	7.45	446.94
2020	7.26	506.89
2019	7.28	433.48
2018	7.19	400.4
2017	7.2	435.02
2016	7.11	366.06
2015	6.82	312.16
合计	7.28	12,888.42

该图可以对不同年份进行钻取，再对电影类型进行钻取获得所需要的电影列表，能清晰地看到不同年份电影整体的评分和票房情况。如2025年评分7.39、票房3290.51亿元，更能直接了解最后想要了解的电影及其对应的评分和票房。可用于分析电影市场随时间在质量（评分）和商业表现（票房）上的变化趋势，辅助行业研究和决策。电影投资者可通过对比不同年份的票房和评分，分析市场趋势，判断投资时机。

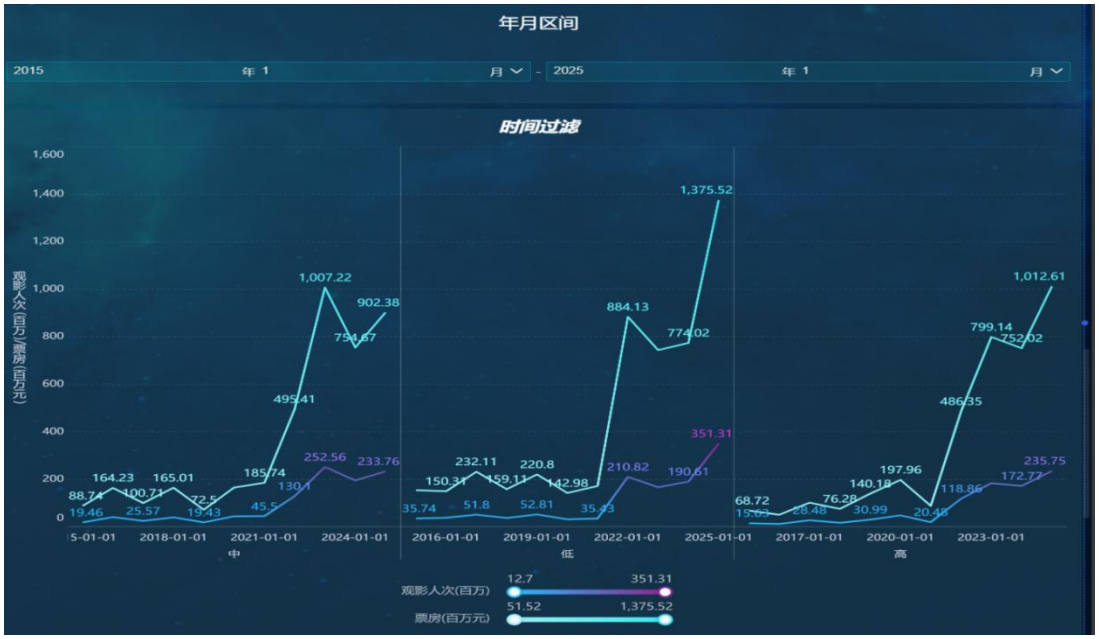
（7）堆积柱形图（钻取）



该图可以对不同导演进行钻取，获取所选导演拍的不同类型电影的目录，便于对比不同导演作品在观影人次和票房上的差异，了解导演的市场号召力。如徐峥对应柱形数值较高，说明其作品在观影人次和票房方面表现突出，帮助电影行业评估

导演价值。影院在排片时，可参考该图，优先排映票房和观影人次高的导演作品。

（8）折线图（时间过滤）



该图可以对时间进行过滤，筛选想要知道的年份数据分析不同消费级别下，电影市场观影人次和票房随时间的波动情况。如某些年份高消费级别下观影人次和票房有明显增长，能为电影定价策略、市场定位等提供参考。电影行业相关从业者可根据不同消费级别观影人次和票房的时间变化趋势，制定营销策略。如在观影人次和票房高峰时段，加大宣传推广力度，推出优惠活动等。

（9）词云图（联动）

该图可以进行动态联动，选取电影，获得所选电影相应的记录数、类型评分以及不同消费级别下的观影次数。像“哪吒之魔童降世”“流浪地球”“唐人街探案”系列等字体较大，

说明该电影观影人数较为突出，更受观众欢迎的热门影片。出现较多的科幻（如“流浪地球2”）、动作（如“速度与激情”系列）、喜剧（如“唐人街探案”系列）等类型电影，一定程度上反映出这些类型在市场上的高需求和关注度。

将以上图表进行整合，形成可视化展板：

。

（二）关键要点

1. 知识点

识别影视数据特征：多维度（类型、区域、时间）、关联性（票价与人次、类型与票房）；

掌握数据处理技术：网络爬虫（批量获取数据）、缺失值填充（中位数/众数）、异常值修正（截断法）、字段编码（分类数据数值化）；

分析不同图表的适用场景（如矩形树图展示类型占比、流向地图呈现区域关联）。

2. 技能点

能够使用Python进行数据采集与清洗、可视化工具（FineBI）制作交互图表；能够从数据中提炼类型偏好（奇幻类受欢迎）、区域差异（东部票房领先）、消费规律（高消费群体对历史类接受度高）；能够根据影视场景选择图表类型，设计动态交互展板，突出核心指标（如类型票房占比、区域人次差异）。

3. 态度点

严谨性：严格处理异常数据，确保分析结果可靠，避免误导投资决策；

市场导向：以电影行业痛点（题材选择、排片效率）为目标，注重成果的商业实用性；

创新思维：通过多图表联动与钻取分析，挖掘数据深层关联（如消费级别与类型偏好的匹配）。

（三）教学使用

1. 教学组织

采用“采集-处理-分析”三步教学法：

采集阶段：讲解网络爬虫原理，分组用Python爬取年度票房数据，掌握数据获取技巧；

处理阶段：实操数据清洗（缺失值、重复值、异常值处理），教师演示关键代码，学生验证数据质量；

分析阶段：使用可视化工具制作图表与展板，围绕“如何提升电影票房”展开讨论，各组汇报策略建议（如优先投资奇幻类型、广东加大排片）。

2. 过程设计

任务驱动：设置阶梯任务（爬取数据→清洗数据→制作图表→设计展板），每阶段明确输出标准（如清洗后数据无异常值、展板含5类以上交互图表）；

案例对比：对比不同类型电影的票房表现，引导学生思考类型偏好与市场需求的关联。

3. 考核方法

过程性考核（60%）：爬虫代码完整性（20%）、数据清洗准确性（20%）、图表设计合理性（20%）；

成果性考核（40%）：展板交互功能（20%）、市场分析报告的策略可行性（20%）。

4. 教学效果

知识掌握：理解电影票房数据的特征与处理原理，掌握影视数据分析的核心维度；

技能提升：熟练使用Python与可视化工具完成数据采集、处理与展示，能独立开展类型趋势与区域市场分析；

能力培养：提升从数据中挖掘商业机会的能力，为电影制作、发行等岗位提供实践基础。