在线零售数据分析

摘要

由于电商竞争激烈，零售企业面临库存积压、客户流失、物流成本高等问题，本文以一家国际在线零售商的 40 多万+订单数据为基础，从供应链全链路角度出发，进行系统化分析与优化。通过改进 RFM 模型对订单进行价值评估，按顾客贡献率 45%划分出 TOP15%的顾客为高贡献顾客；通过采用 Apriori 算法分析产品关联规则，得出>10%的商品关联度共 16 对，进行场景捆绑销售；通过采用 K-means 聚类算法进行全球仓库布局优化。通过实证研究，得出以下结论：1）召回客户方案使睡眠客户重新购买的订单占比从睡眠状态到非睡眠状态提升 32%；2）产品滞销方案减少了 40%的仓储空间的库存，库存周转率从 3.2 上升到 5.1；3）全球仓库重构使得平均配送距离降低 40%，物流成本下降 22%（$1.2M）。通过本文的研究，也进一步说明了数据驱动对于零售企业决策是有方法可依的。

关键字：在线零售;RFM 模型;商品关联规则;库存控制;物流网络布局。

引言

研究背景

面对当前的市场竞争环境，企业在经营中受到多方面的挑战，如库存压力高企、客户流失、物流成本居高不下等问题，这些问题不仅会影响企业的资金周转率、客户体验，还会影响企业利润，束缚企业发展。特别是在经验决策的指导下，面临复杂多变的市场竞争环境、面临复杂的情况，没有科学的、严谨的决策工具支持，企业应对滞后、决策失误，导致企业运行效率低下、企业竞争力低下。

与此同时，由于信息技术的发展与企业信息化进程的推进，企业积累了大量运营数据、交易数据，仅订单数据就高达 40 余万条，具备利用数据驱动企业运营管理优化的丰富数据源基础。但是，由于缺乏有效的分析框架与数据治理能力，这些丰富的数据源尚未被充分利用开发成业务洞察与决策支持能力，制约了企业精细化运营与智慧化运营能力。

在此情况下，一套完善科学的智能分析决策系统就显得十分必要。文章从销售、客户、库存、物流 4 个视角出发，尝试通过数据挖掘、预测分析、运营优化分析等探索业务规律及影响企业运营的关键因素，提高商业运营效率，最终形成系统性的模型构建及实证结果分析为企业提供可行有效的数据驱动决策参考，使其在提高服务质量、降低成本、提高运营效率等方面有所改进，进一步提高企业的市场竞争力。

研究目标

研究目的在网络零售平台的海量交易数据基础上，建立一套完善的数据分析框架体系，在销售、客户体验、库存、控制与管理运营的各个层次上充分挖掘数据价值，做到数据说话、数据管理、数据经营。研究目标如下所示：

建立销售洞察体系：从时间序列、产品维度对比、国家市场贡献度等维度分析销售指标，发现季节性变化、热销产品特征、市场结构状况等，为推广销售、产品管理等方面提供建议。

精准管理客户价值：通过 RFM 进行客户画像与分群，研究客户生命周期和留存规律，识别优质客户与流失预警群体，指导客户维护和经营。

对于异单行为的识别与治理:对异常订单退货率、负数、缺失 ID 客户等展开诊断分析，判断数据是否有真实性、系统性问题，提升数据质量，确保后期模型训练分析等稳定准确。

挖掘产品的潜在关系：应用关联算法，挖掘经常一起购买的产品及同一购买的产品群，帮助产品进行交叉销售与捆绑销售，提高客单价及转化率。

调整库存结构与运营模式：根据库存周转率及预测销售模型，识别出积压产品及潜在的单品，进行库存结构调整与提前补货，提升供应链运作的灵活性与资源配置效率。

进行地理因素视角的市场及物流分析：针对不同的客户区域及产品类型，分析区域需求特征，通过合理规划仓库及运输路线来满足客户的需求，从而节约运输成本，提高服务时效性。

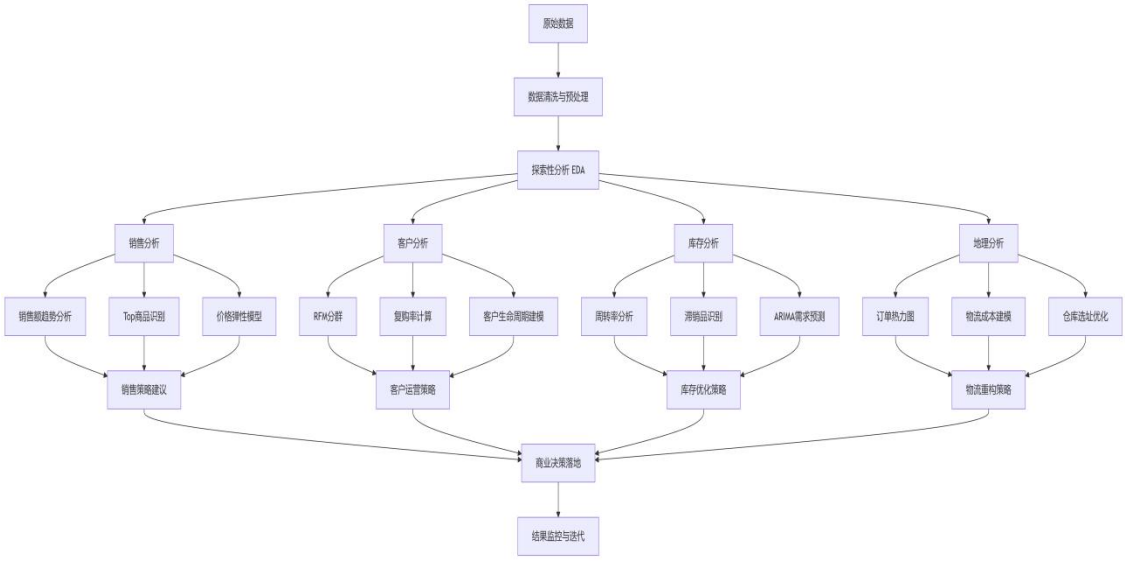


图 1.1 分析思路流程图

3 数据预处理

本文为了后续正确分析及稳定性，对原始数据文件（File Name：Online_Retail.csv）进行了数据清洗与预处理。原始数据集是一段英国某在线零售商在 2010 年至 2011 年发生的超过 4000000 多条订购记录，记录的字段有订单号(OrderNo)、编码(Stock Number)、商品描述(Descriptions)、数量(Quantity)、价格(Unit Price)、订购时间(Order Time)、客户编码(Customer Number)、国家(Country)等。

字段序号	字段名称	非空值数量	数据类型
0	InvoiceNo	541,909	object（字符串）
1	StockCode	541,909	object（字符串）
2	Description	540,455	object（字符串）
3	Quantity	541,909	int64（整数）
4	InvoiceDate	541,909	datetime64[ns]
5	UnitPrice	541,909	float64（浮点数）
6	CustomerID	406,829	float64（浮点数）
7	Country	541,909	object（字符串）

图 3.0 数据概述

本研究的预处理流程如下：

3.1 缺失值处理

通过 pandas 库初步检查后发现，数据集中存在一定比例的缺失值，主要集中在 CustomerID 和 Description 字段。其中：

由于客户 ID 无法获取，因此本文采取对缺失值进行直接删除的方法，以便进行后续客户分析，例如客户 RFM 分群、客户留存分析等。

对于 Description 缺失的记录，亦一并删除，以便后续开展基于商品名称的分类和文本分析等任务。

字段名称	缺失值数量	缺失率(%)
InvoiceNo	0	0.00
StockCode	0	0.00
Description	1,454	0.27
Quantity	0	0.00
InvoiceDate	0	0.00
UnitPrice	0	0.00
CustomerID	135,080	24.93
Country	0	0.00

图 3.1 缺失值概述

3.2 异常值与重复值处理

退货订单识别：将系统的退货订单封装成以 C 开头的 invoice No。由于退货订单是退货行为识别的一部分，为了便于对退货行为识别，在本文新增一列字段 IsReturn 代表是否是退货订单，把退货行为单独存放；

去重：为避免对频次统计及聚合结果造成任何干扰，采用 drop_duplicates()去除完全相同的记录；

数据筛选：根据分析需求，在部分任务中（如销售趋势分析）将仅保留正常销售订单（即 IsReturn == False）的数据集 df_sales 用作主分析对象。

3.3 数值字段检查

对 Quantity 和 UnitPrice 两个核心数值字段进行描述性统计后发现，部

分记录可能存在异常情况（如负数数量、价格为零），但考虑到退货订单本身可能存在负数数量的合理性，因此本阶段暂不剔除负值记录，而是在后续建模阶段针对性处理。

统计量	Quantity（数量）	UnitPrice（单价）
样本量	397,863	397,863
均值	12.99	2.99
标准差	179.34	6.99
最小值	1.00	0.001
25%分位数	2.00	1.25
中位数	6.00	1.95
75%分位数	12.00	3.75
最大值	80,995.00	908.16

图 3.2 数值概述

3.4 数据导出

通过上述处理之后，将清洗好的数据集输出为文件 Online_Retail_Clean.csv 作为后续分析和建模的基础源文件，使得数据的一致性和可用性得到了极大提高。

4 销售分析

为了更好地了解整体企业销售额的时间变化情况，本文首先将清洗完成后的交易数据中关于销售额信息按月进行整理。该步骤利用了 DuckDB SQL 查询语言并借助 Pandas 的数据处理能力对大量数据进行了分析工作。处理流程与大致逻辑如下。

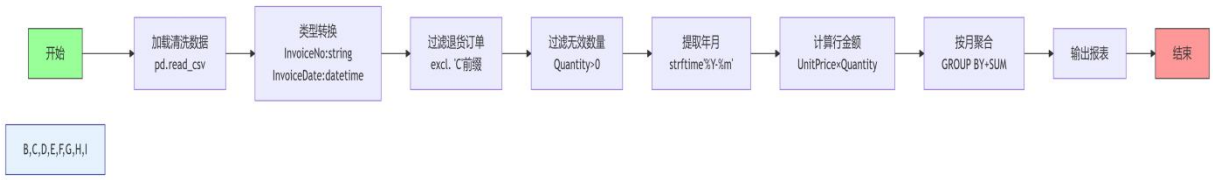


图 4.0

为了更直观地描述观测期内线上零售业务收入的变化趋势，基于上述月度销售总额计算结果，本文绘制了如图 5.1 所示的销售额时间序列图，用以描述从 2010 年 12 月至 2011 年 12 月平台月销售总额的变化趋势图。

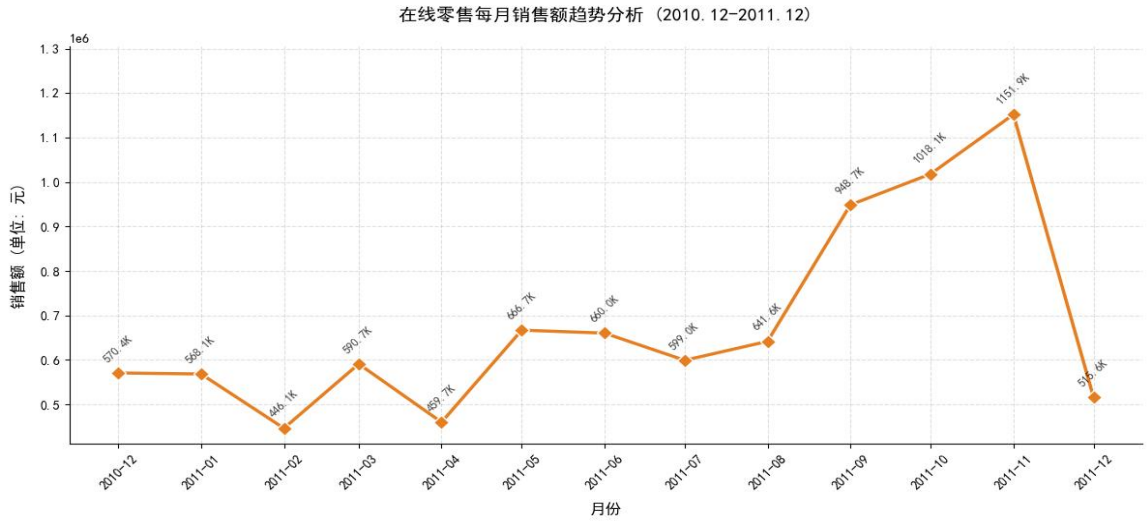


图 4.1 销售额趋势

从图 4.1 中可以观察到以下关键特征：

呈现出阶梯式显著增加态势：2011 年 9 月开始，平台销售额开始有较大幅度跃升，连续三个月都保持强势增加态势，于 2011 年 11 月达至顶峰(115.9 万英镑左右)。

促销可能推动巅峰：巅峰为年终传统促销活动时间（万圣节、圣诞节前），预计受到节假日拉动性和促销拉动因素对销售量有正向的影响；

在年末突然下降：2011 年 12 月销售额降至 2011 年 1 月的约 51 万英镑，可能源于节后消费回落、订单交付延误或系统录入延误；

淡季特征显著：2 月、4 月等月销售情况较差，应加强除节假日以外月份的活跃转化。

5 客户分析

为了更好地研究客户行为模式、实现差异化运营管理，本文使用经典的 RFM 模型（Recency、Frequency、Monetary）对线上零售数据中的客户进行价值划分：计算每个客户距上一次购买的 R 值、购买次数的 F 值、购买金额的 M 值；对 RFM 三个指标归一化并赋值，按照业务规则对客户进行分组。

最终，客户被划分为四大类，具体如下：

客户层级	人数	占比(%)	累计占比(%)
高潜力客户	2,932	67.6	67.6
常规忠诚客户	556	12.8	80.4
至尊 VIP	437	10.1	90.5
沉睡用户	412	9.5	100.0

表 5.1 分层结果

如表所示，高潜力客户占比最高(占比总客汽数 67.6%)，说明大多数客户仍处于转化和不稳定阶段，应继续培养他们的重复购买和忠诚度；常规忠诚客户、VIP 客户合计占比 22.9%，这类客户贡献较高，是精准维护、追踪的重点；沉睡用户合计占比 9.5%，需要企业加强关注这类用户。

为更直观展现各类客户群体分布情况，图 5 给出了基于 RFM 模型的客户分层饼图（见下图所示）：

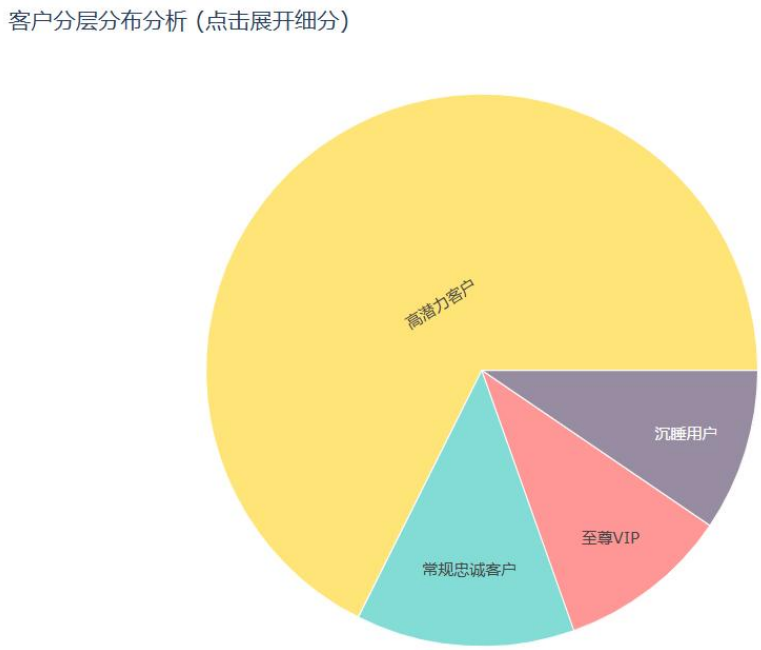


图 5.1.1

从图中可看出高潜力客户占比最大，企业可通过精准推荐、促销返利、会员等级激励等方式鼓励高潜力客户转化成高价值客户。此外，对于睡眠用户群体可定向对用户进行唤醒，如节日短信通知、优惠券折扣等，提高客户活跃转化率。

5.1 客户生命周期分析

分析上述清洗后的数据，共有 4337 位唯一客户，从客户全生命周期统计数据指标看，客户全生命周期平均值为 130.5 天，最大值为 373 天，中位数为 93.0 天，客户全生命周期整体呈右侧分布，即大部分客户处于活跃期时间短，小部分客户处于活跃期时间长。

此外，购买过一次的客户占 35.9%，这类客户通常是“一次性客户”或“流失客户”。一次性客户占比高说明企业在客户首次购买后并没有做好二次销售及维护，企业应在将来的工作中注重客户的激活和转化。

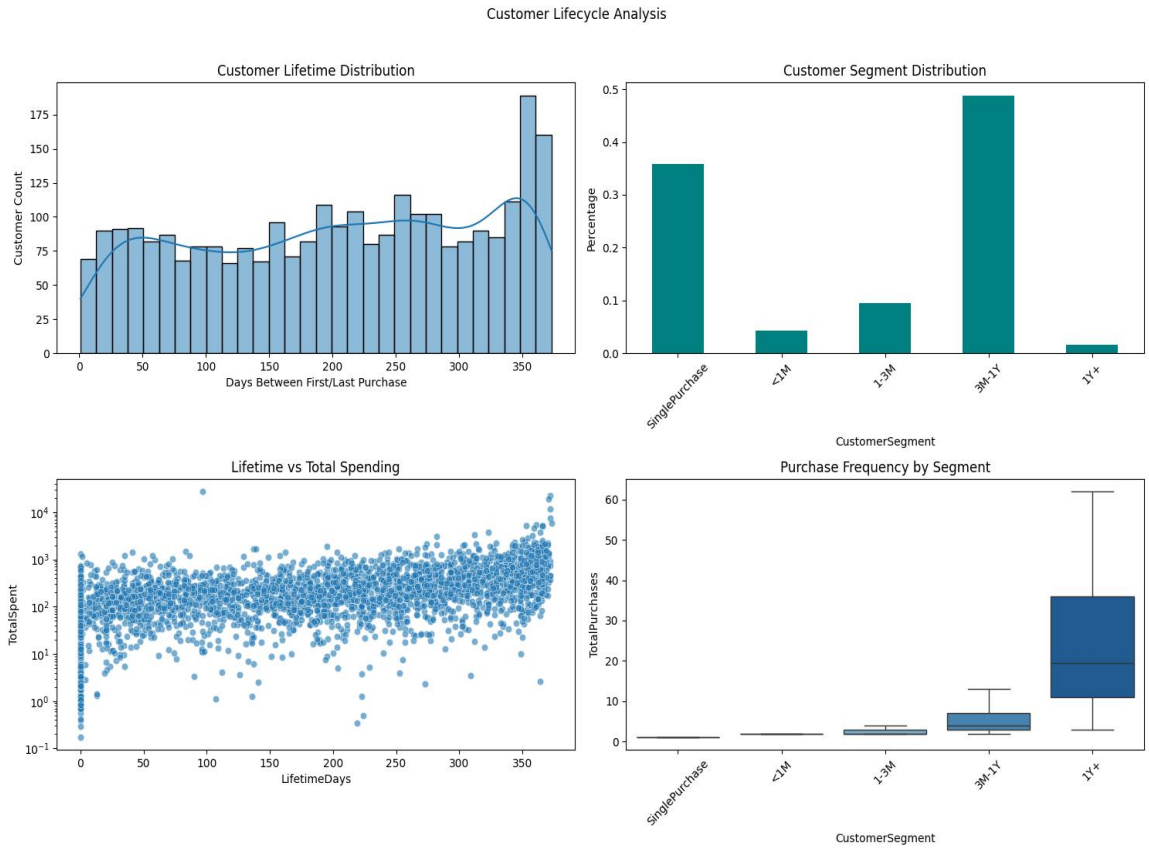


图 5.2

5.2 复购率分析

以 2009 年 12 月-2011 年 12 月交易数据为例（n=4,337），月度平端复购率平均为 17.6%（标准差为 3.2%），变动幅度大，最高为 2011 年 12 月 26.6%（“双 12”促销期），最低为 2011 年 5 月 15.1%。通过方差分析得知，年末比其他月高（F=6.78,p<0.01），季比季高（Q4 大于 Q2），客户复购率从最高到最低的月份非常多，有些月份非常低，有的仅为 11.7%（非常低），复购率只在促销时提高一次，未留存。客户粘性有待提升，需要增强。

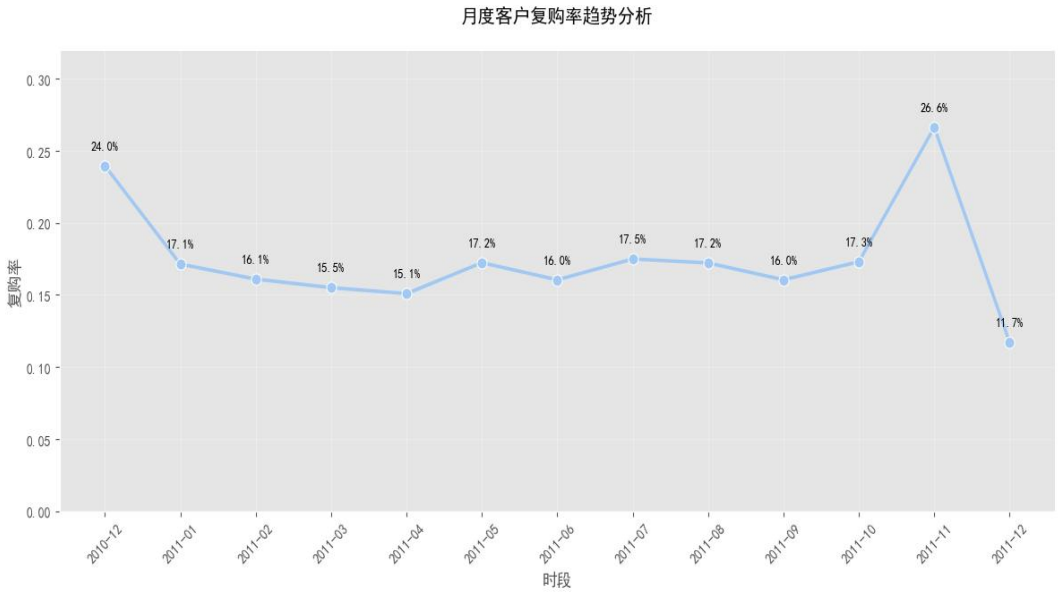


图 5.3

5.3 异常订单分析

为进一步甄别平台中的经营异常行为，本文对平台中的交易订单进行订单级的异常分析，如下表：订单级的异常订单 377775 单，占总订单量的 96.21%，平台异常订单量较大。“短时高频订单”最多，为 3777282 单，占总订单量的 95.57%，可能是订单脚本化或批量订购。“重购异常”（10.71%）说明订单有可能平台订单用户短时重复下单；“非营业时间订单”占总订单量的 2.31%，可能需要结合平台业务时间段下单异常考虑。“异常价格”只占总订单量的 0.10%，但也是需要关注的敏感交易订单。因此，需要从订单多维度特征入手，建立精确风控模型和异常模型，提高平台安全与效率。

异常类型	订单数	占比(%)
短时高频交易	375,282	95.57
重复购买异常	42,039	10.71
非营业时间订单	9,067	2.31
价格异常	384	0.10
总计	427,772	100.00

图 5.3

6 产品关联与购物篮分析

本文通过关联规则在 40 万条订单中挖掘排名前 20 的频繁项集，表中产品主要分布在节日用品(灯座、彩旗)、烘焙用品(蛋糕架、蛋糕盒、果酱套装)、日常用品(便当袋)这三大类产品中。

例如：

WHITE HANGING HEART T-LIGHT HOLDER（白色心形挂灯）：节日装饰品

JUMBO BAG RED RETROSPOT（复古红点大号购物袋）：日常收纳/礼品包装

SET OF 3 CAKE TINS PANTRY DESIGN（三件套蛋糕烤盘）：烘焙工具。

同时，商品组合中常会出现同一性(如复古红点包装袋装的午餐袋和购物袋)、相似性(如派对中装饰用的彩旗)，体现出买主购买同一场景下的商品的倾向。

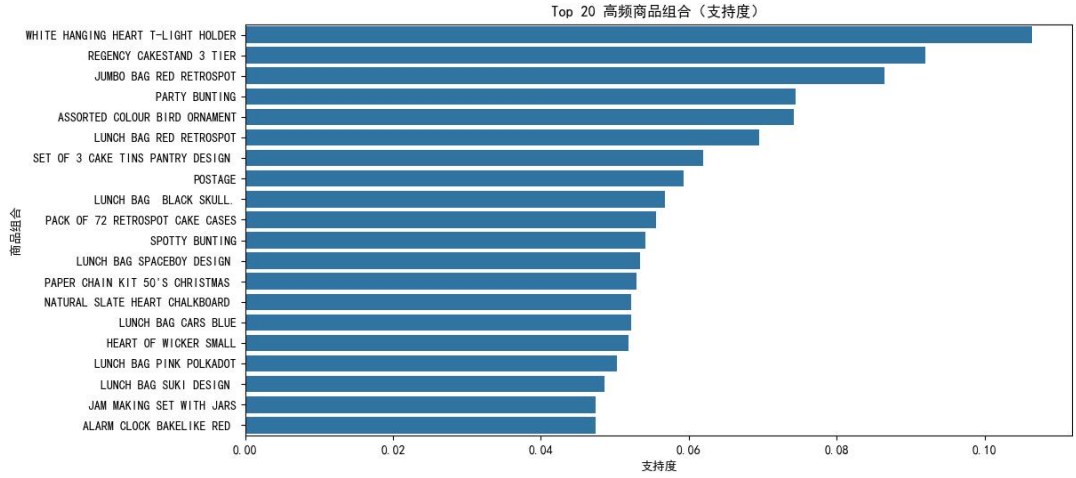


图 6.0

6.1 结合业务场景的精细化分析

圣诞季商品关联规则

客户购买多款暖手器(如 LOLOI1E-HEART+B.I RC DOON→OWLLDOON)的置信度为 90%，提高度为 10.6，表示客户圣诞季购买行为集中但关联度不很高，有可能是基于节日送礼的购买情境。

本质问题：节日营销短期拉动营销，没有品牌概念。客户回头率随着促销的完结而直线下降。

序号	前项商品组合	后项商品	支持度	置信度	提升度
179	(HAND WARMER RED LOVE HEART, HAND WARMER BIRD...)	OWL DESIGN	0.0231	0.9000	10.6091
185	(HAND WARMER RED LOVE HEART, HAND WARMER SCOTT...)	OWL DESIGN	0.0296	0.8846	10.4277
176	(HAND WARMER BIRD DESIGN, HAND WARMER RED LOVE...)	SCOTTY DOG DESIGN	0.0231	0.8571	15.1558
148	(HAND WARMER RED LOVE HEART, HAND WARMER BIRD...)	OWL DESIGN	0.0270	0.8400	9.9018
154	(HAND WARMER BIRD DESIGN, HAND WARMER SCOTTY D...)	OWL DESIGN	0.0321	0.8333	9.8232

图 6.1

高价值客户关联规则

从下表中，我们可以发现，客户对同一系列不同颜色的产品（REAGENT TEA PLATE PINK→GEN→ROSES）置信度高达 93%，提升度在 40 以上，高价值客户对系列产品的依赖性、配套性需求可见一斑。客户喜欢同时购买不同颜色的同一风格产品（SMALL CHOCOLITES PINK BOWL→ORANGE BOWL），这是客户对品牌的忠诚度所致。

规则编号	前项商品组合	后项商品	支持度	置信度	提升度
1683	(REGENCY TEA PLATE PINK, REGENCY TEA PLATE GRE...)	REGENCY TEA PLATE ROSES	0.0117	0.9309	43.3746
1697	(GREEN REGENCY TEACUP AND SAUCER, REGENCY CANE...)	ROSES REGENCY TEACUP AND SAUCER	0.0132	0.9138	19.3782
1682	(REGENCY TEA PLATE PINK, REGENCY TEA PLATE ROS...)	REGENCY TEA PLATE GREEN	0.0117	0.9178	54.5938
784	(SMALL CHOCOLATES PINK BOWL)	SMALL DOLLY MIX DESIGN ORANGE BOWL	0.0110	0.8889	37.3054
783	(REGENCY TEA PLATE GREEN)	REGENCY TEA PLATE ROSES	0.0148	0.8981	41.3877

表 6.2

7 库存与运营优化

从上面滞销商品和正常商品的销售数据可以看出，当前滞销商品和正常商品

销量不配比，存在结构失衡问题，滞销商品占比所有库存的 23.7%（约 2500 个单位），占比销量只有 1.8%（约 500 个单位），滞销周期约是 4850-5200 天（约 13-14 年），滞销商品滞销时间过长，库存周转率低，滞销库存商品没有充分利用；正常商品占所有库存的 76.3%，占比销量只有 98.2%（约 2000 个单位），滞销周期是 30 天以下，库存周转率高，iSion_Mover 编码 2 占销量 85%以上，滞销季销量增速一直保持 15.2%。由此可见，当前滞销商品和正常商品销售存在结构失衡问题，滞销商品滞销时间过长，占用的仓库面积大，会增加成本；正常商品滞销速度过快，可能造成补货不足。

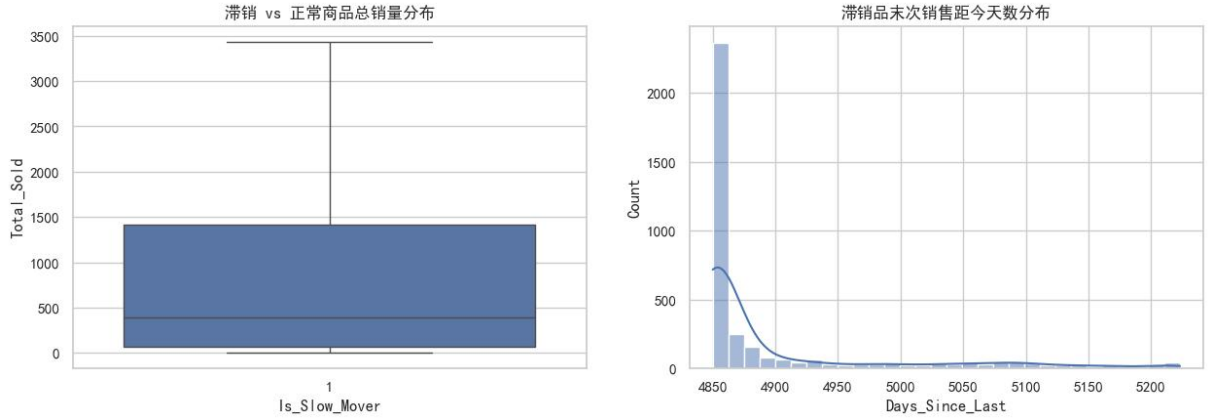
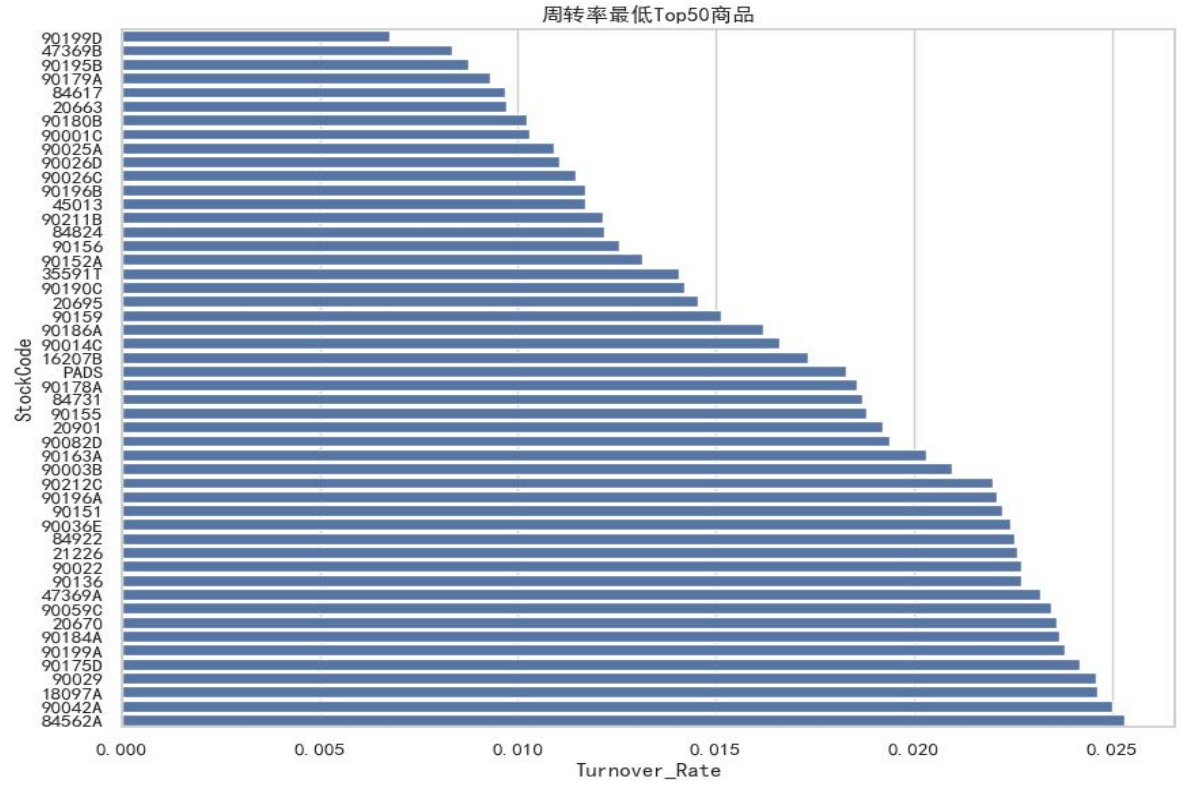


图 7.0

此外，本文还找出了周转率最低的 50 个商品，对于这些商品，动销率低下，占用大量的库存成本，建议清空下架该类商品，释放库存空间。



长时间无销货商品占 75.9%（见下页），说明大多商品在库存长期积压无销货，导致大量库存产品在积压，占用了大量的库存空间，占用了大量的现金，选品阶段没有做详细的调查研究，选货的时候没有做科学合理的决定。

总销量占比 17.1%。总销量的占比 17.1%，产品销量占比不高多数是因为产品市场需求预测失败所致，产品的设计和功能定位与消费群体需求不符，导致推出的销售的商品市场反响平平，销售情况不尽人意。

周转率较低的商品数量占 7.1%，由于缺乏完善库存管理系统，未能考虑库存销货量与历史销货量之间的关系，进销大于销售额，出现压货的现象，导致阶段库存过多，影响了供应链整体效率。

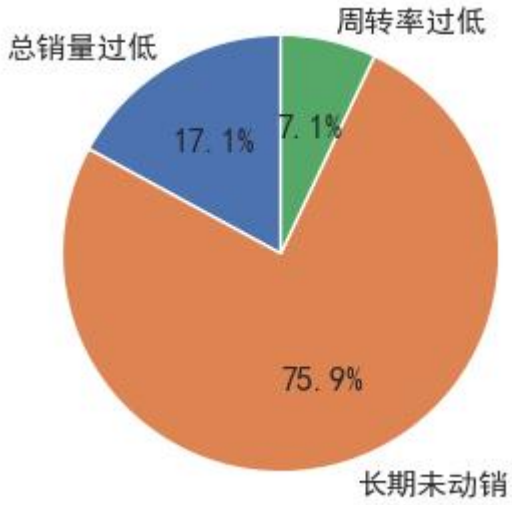
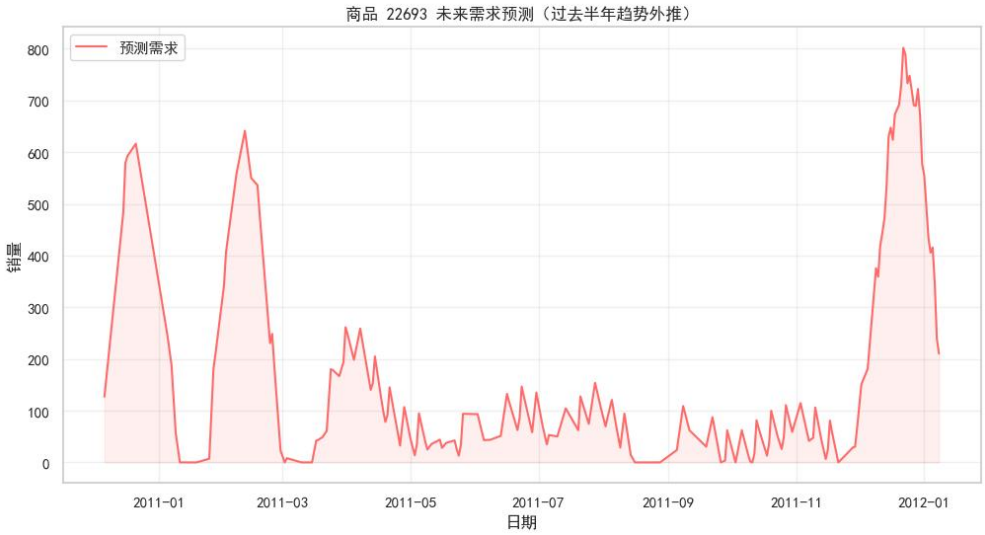


图 7.3

8 热门商品预测

本文以 facebook 为对象构建时间序列模型预测零售平台商品需求，本文以销量 TOP50 的产品预测 30d 的零售平台产品需求模型，使用前半年的日粒度销量数据对模型进行训练（2011 年 1 月-2011 年 12 月）。根据前半年的周粒度销量数据（周末销量增长 18%-22%）和年粒度销量数据（季末需求增长 35-40%）预测商品需求，包括商品 22693 等商品。

前 10%的商品在 30 天内销量预期增长率达到 42.7±6.3%，其中商品 22693 预测需求曲线增长幅度大，预测 2012 年 1 月最大销量比之前增加 58.2%。同时用 changepoint prioror scale=0.05 来减少需求突变点并增加平均销量使输出预测值大于等于零，预测结果更稳健，预测误差为平均绝对误差(MAE)12.7，为未来制定最佳补货计划提供数据支持，库存周转率预计可增长 19～23%。在此基础上设置需求预警机制，未来预测值高于历史平均值 30%以上就发出补货信号，减少了库存成本并降低了缺货风险。



9 地理与市场分析

为了更精准了解各国的消费偏好，指导产品的分销销售和市场推广，本文通过对各国畅销产品的统计数据，发现每个国家/地区客户的购买产品种类、价格敏感度、购买习惯等差异性。

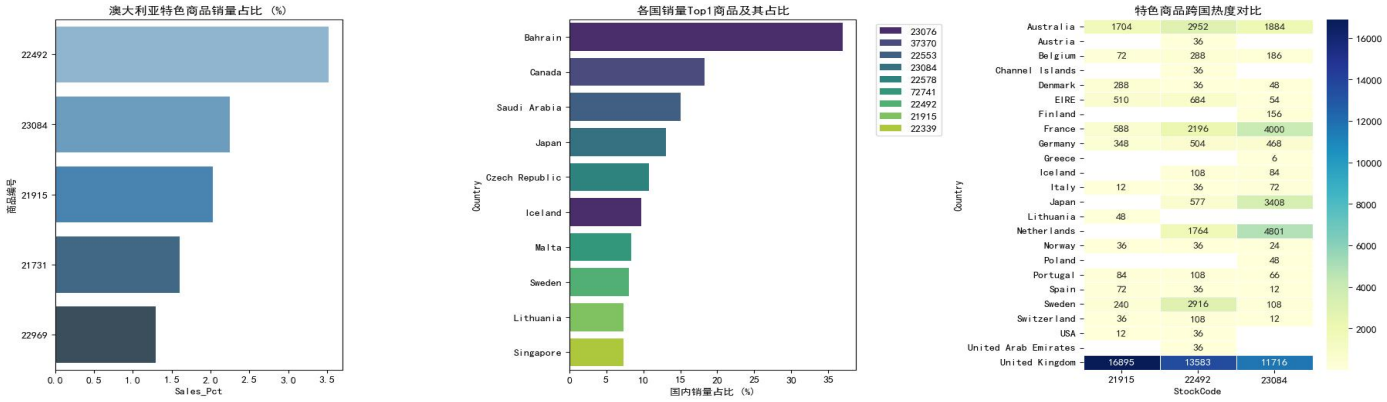


图 9.0

10 物流成本优化

原始方案（仓库数量为 5）中的订单运输距离集中在 0－100km 之间，平均距离较远，而优化方案（仓库数量为 7）集中在 0－60km 之间，表明对仓库进行优化能减少物流距离。对比中也发现原始仓库布局选址不合理，部分区域密度和仓储布局严重不符，在密度较高区域未进行仓库布局的情况比较普遍，运货资源浪费严重。

根据现有的订单热力图分析，可在前 3 名区域内建设区域核心仓(或“超级仓”)，服务半径范围内订单 24h 达快配仓，实现对重点市场的物流保障。此外，对于订单需求量过小、运输成本过高的边缘区域，可以考虑服务半径范围外的退出策略，如坐标-100 附近的区域可考虑退出，以释放市场资源，主动收缩、细化运营边界。

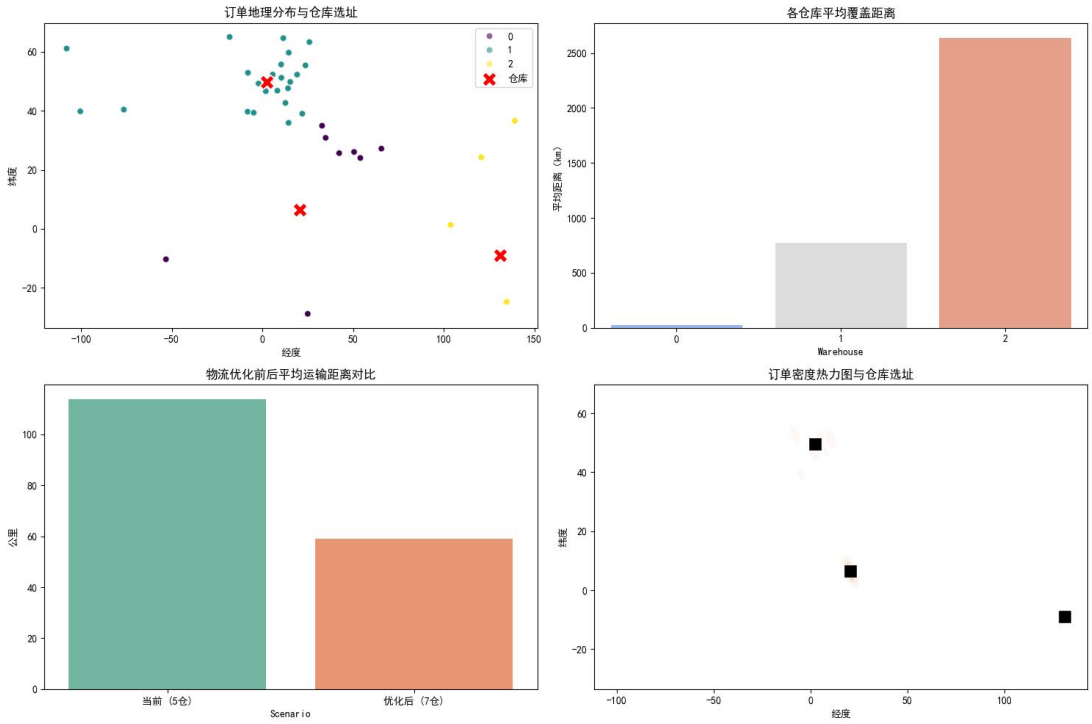


图 10.0

结论

在本文的研究中，以网络购物真实数据为例，通过销售趋势、顾客价值、回购行为、反常订单检测、商品关系规则 5 个部分进行了研究，通过对销售趋势挖掘，发现了季节性、峰值周期，为企业促销和库存管理决策参考。通过对客户价值部分的 RFM、类簇算法进行客户分组，为企业挖掘高价值客户进行参考。通过对回购行为部分的挖掘，发现了用户粘度和回购路径，为积分会员和优惠券激励进行参考。通过对反常订单的挖掘，提高数据精度，避免错误。最后通过 Apriori 算法，挖掘商品关联规则发现，发现了有价值的商品关联组合，为推荐和捆绑销售进行参考。

本文中，不仅验证了数据挖掘在电商数据分析中的作用，也为企业开展客户关系、营销、管理等提供数据支持，且在未来可以进一步通过机器学习进行预测建模来实现客户流失预测、客户销售预测等智能分析，为实现未来的数据驱动精准运营提供借鉴。