数据挖掘分析报告

1. 引言

本项目对一个约30GB的用户购买行为数据集进行了全面的数据挖掘分析。数据源包括用户交易的Parquet文件（内含JSON格式的购买历史）和一份JSON格式的商品目录。通过运用PySpark作为主要的分布式计算框架，结合FP-Growth关联规则挖掘、时间序列分析以及描述性统计等方法，我们旨在揭示用户购买行为中的潜在模式、商品间的关联、支付方式偏好、季节性趋势以及退款相关的特征。本报告将详细阐述各个分析任务的主要发现，并结合运行输出的表格和图表进行解读，探讨其可能带来的业务价值和后续行动建议。

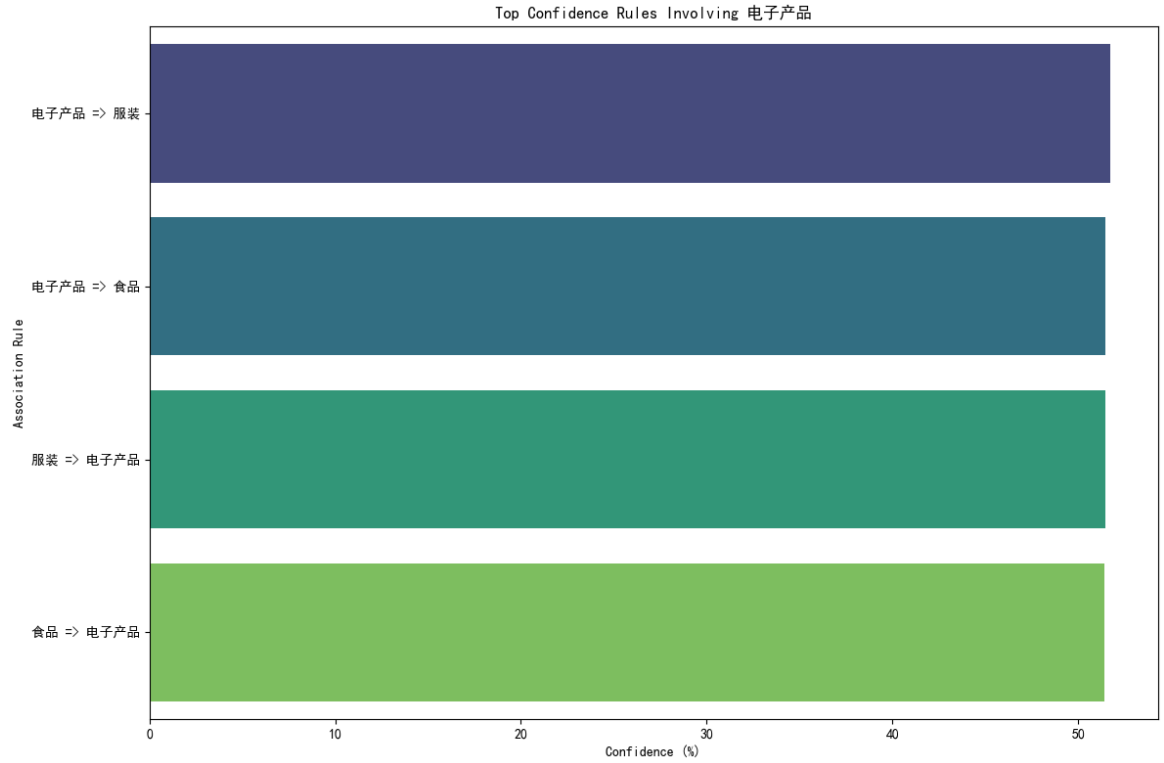
2. 数据概览与预处理

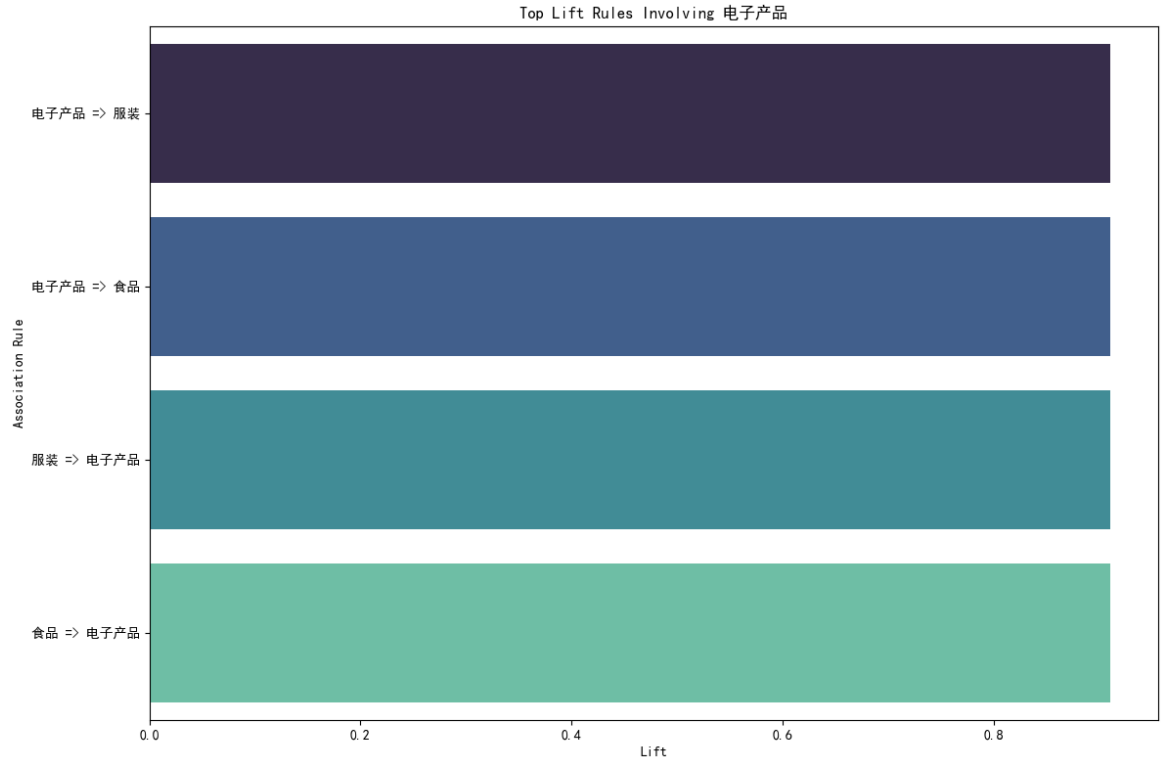
- **数据源**: 包含用户ID、购买历史（商品ID、购买日期、支付方式、支付状态、交易平均价格）、商品目录（商品ID、商品小类、商品单价）。
- **数据规模**: 原始数据集约30GB。经过预处理，生成的核心分析DataFrame `df_final_preprocessed` 包含 **404,987,101** 条记录，每条记录代表一个用户购买的单个商品项。
  - Schema如运行结果所示：`user_id`, `purchase_date`, `payment_method`, `payment_status`, `purchased_item_id`, `item_minor_category`, `item_major_category`, `item_unit_price`, `transaction_avg_price`。
- **预处理关键步骤**:
  1. 成功读取并解析了`product_catalog.json`和主数据Parquet文件。
  2. `purchase_history`中的JSON被正确解析并展开，每个购买的商品项成为独立记录。
  3. 商品信息（小类、单价）通过`product_id`与商品目录成功连接。
  4. 商品小类被准确映射到预定义的大类。
  5. 日期字符串转换为标准日期类型。

3. 分析结果与洞察

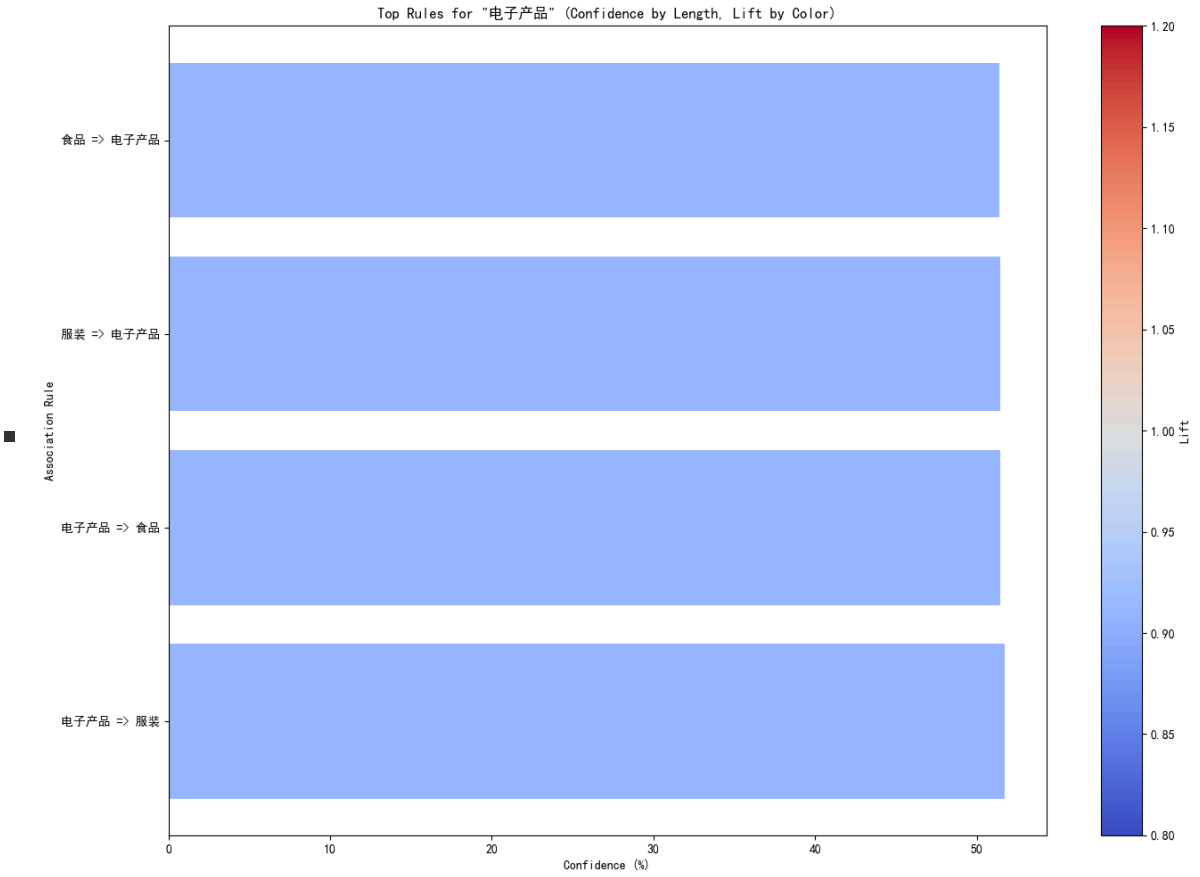
3.1 Task 1: 商品类别关联规则挖掘 (大类)

- **目标**: 分析用户在同一订单中购买的不同商品大类之间的关联关系。参数设置为：支持度(support) ≥ 0.02，置信度(confidence) ≥ 0.5。
- **数据准备**: 共计 **102,614,467** 个包含至少两种不同大类的交易（购物篮）用于FP-Growth分析。购物篮样本如：`[食品, 家居, 电子产品]`，`[服装, 母婴, 食品]`等。
- **主要发现**:
  - **高频项集**:
    - 单品大类：“服装”(56.9%)、“食品”(56.63%)、“电子产品”(56.59%) 支持度最高。
    - 二元组合：“食品, 服装”(29.29%)、“电子产品, 服装”(29.26%)、“电子产品, 食品”(29.11%) 最为频繁。
    - 三元组合：“电子产品, 食品, 服装”(13.02%) 支持度也较高。
    - [在此插入Task 1的Frequent Itemsets表格截图的部分内容或引用其CSV文件]
  - **关联规则**: 严格满足条件的关联规则共 **6条**，均围绕“服装”、“食品”、“电子产品”这三个核心大类。
    - 例如：“食品 => 服装”（置信度51.72%）、“电子产品 => 服装”（置信度51.71%）。
    - [在此插入Task 1的Association Rules表格截图的部分内容或引用其CSV文件]
  - **提升度 (Lift)**: 所有这6条规则的提升度（`lift_val`）均为 **0.91**。Lift < 1表明这些高频组合的共现更多是由于各项本身流行，而非强烈的相互购买驱动。

- **“电子产品”相关规则**: 筛选出的4条与“电子产品”相关的规则，其置信度和提升度与上述总体规则一致。
- **可视化**:
  - **置信度与提升度条形图**:
    - 

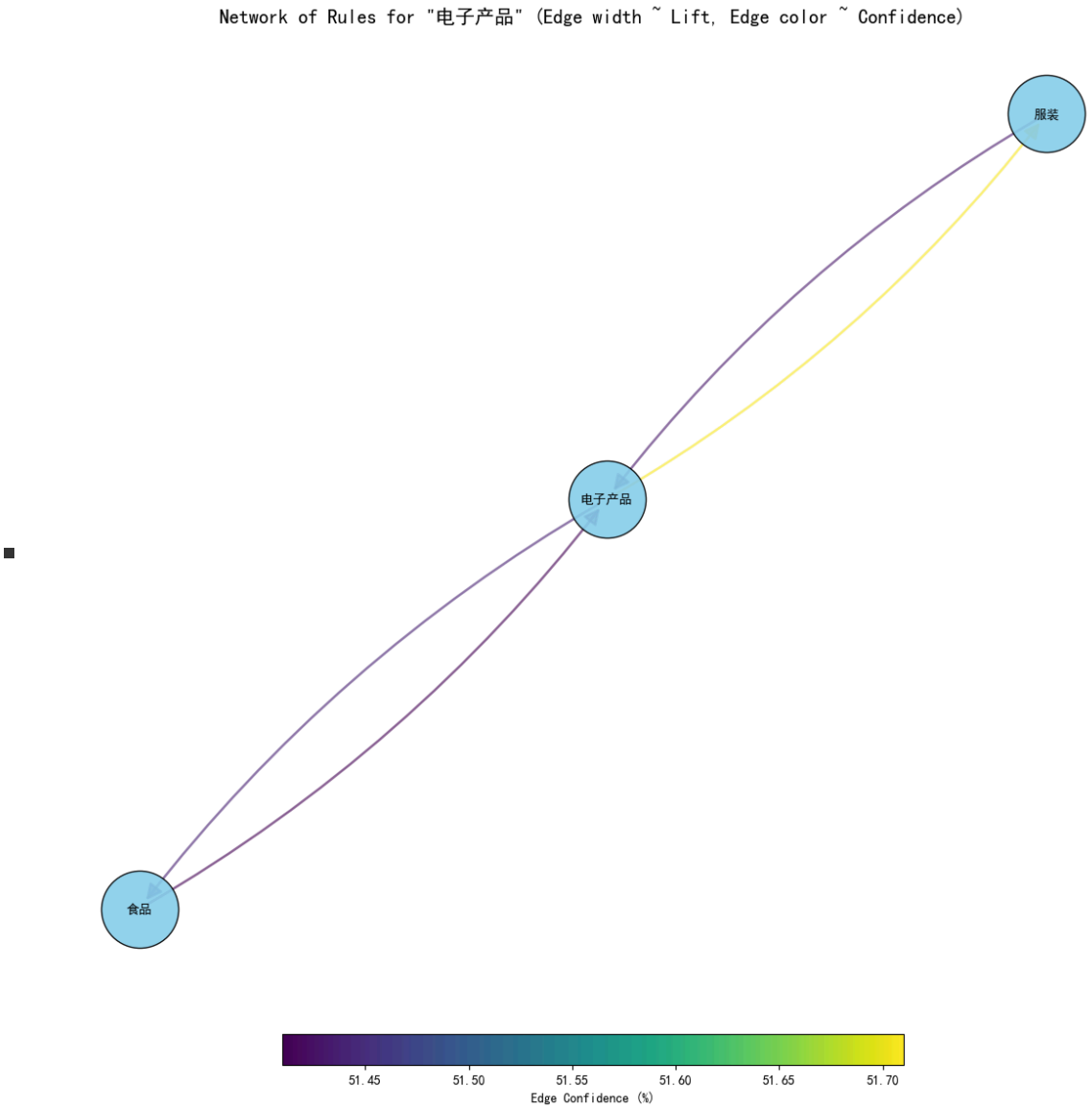
    - 

    这两张图清晰显示了规则的高置信度（约51%）和一致的低提升度（0.91）。
  - **综合图 (Confidence by Length, Lift by Color)**:



此图通过条形长度展示高置信度，通过颜色（由于所有Lift值相同，颜色也一致，例如均为浅蓝色，对应Colorbar上0.91附近的位置)展示提升度，进一步确认了高共现但非强驱动的特性。

- **网络图**:

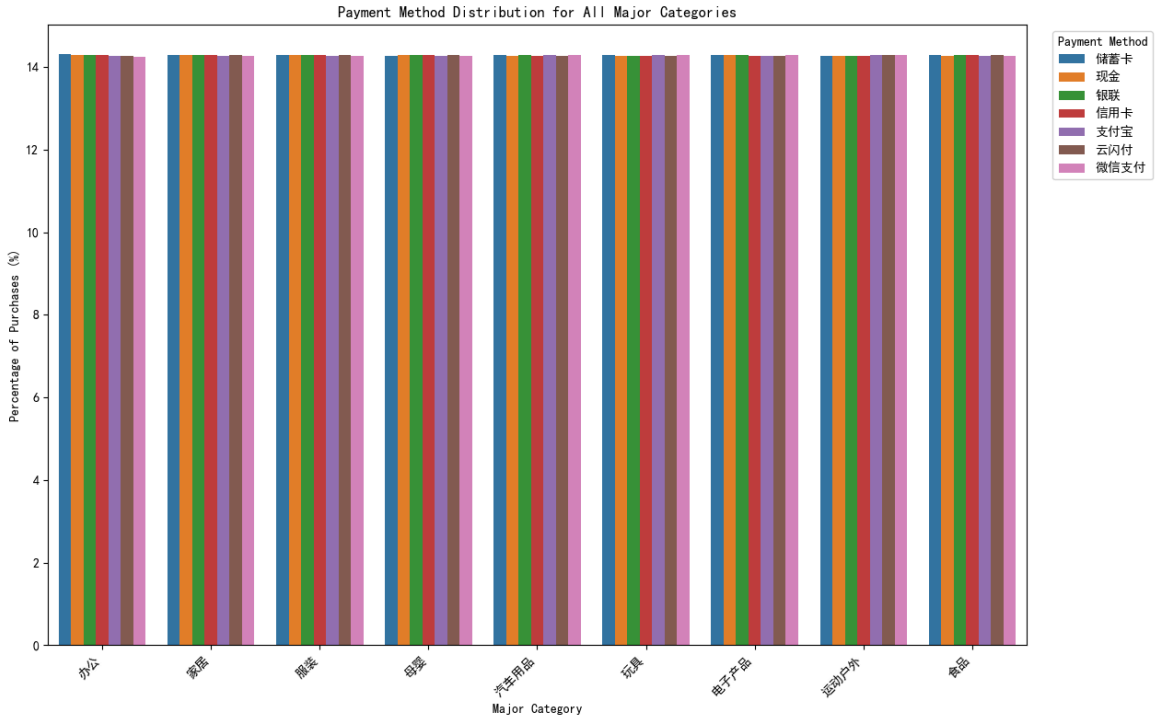


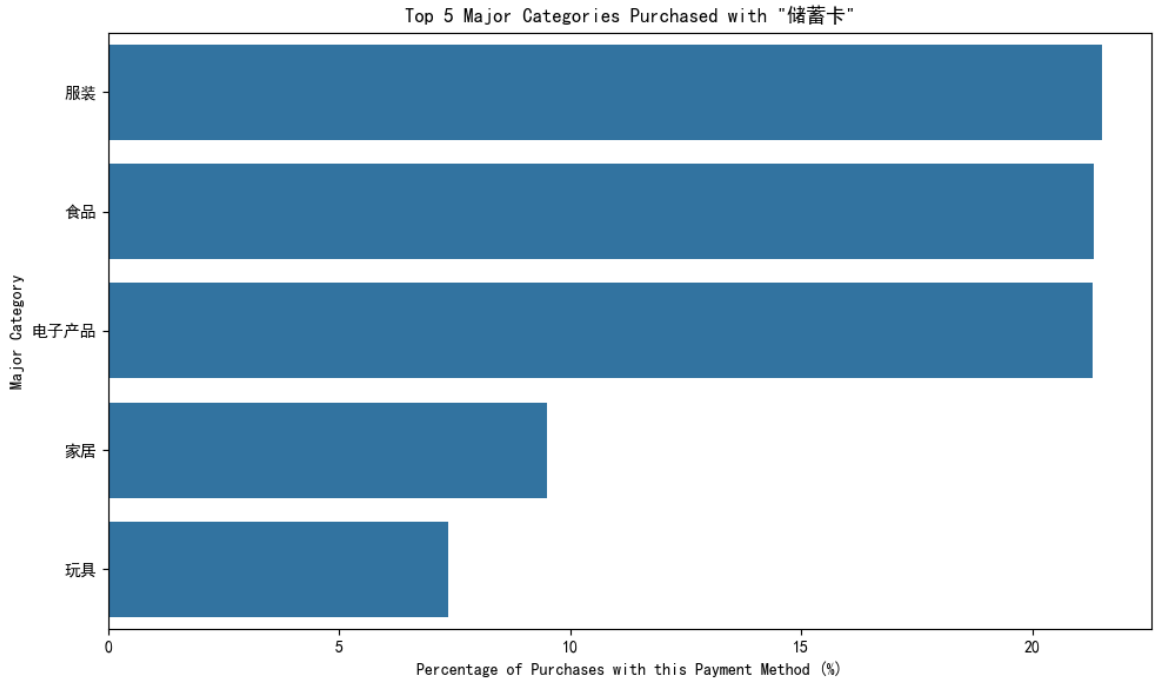
网络图简洁地展示了“电子产品”、“食品”、“服装”三者间的两两强共现关系，边的宽度（代表Lift）和颜色（代表置信度)也反映了分析结果。

- **业务洞察**: 同上一版报告，强调核心品类重要性，并指出营销策略应更注重单品吸引力或细粒度搭配，而非基于大类的强行捆绑。

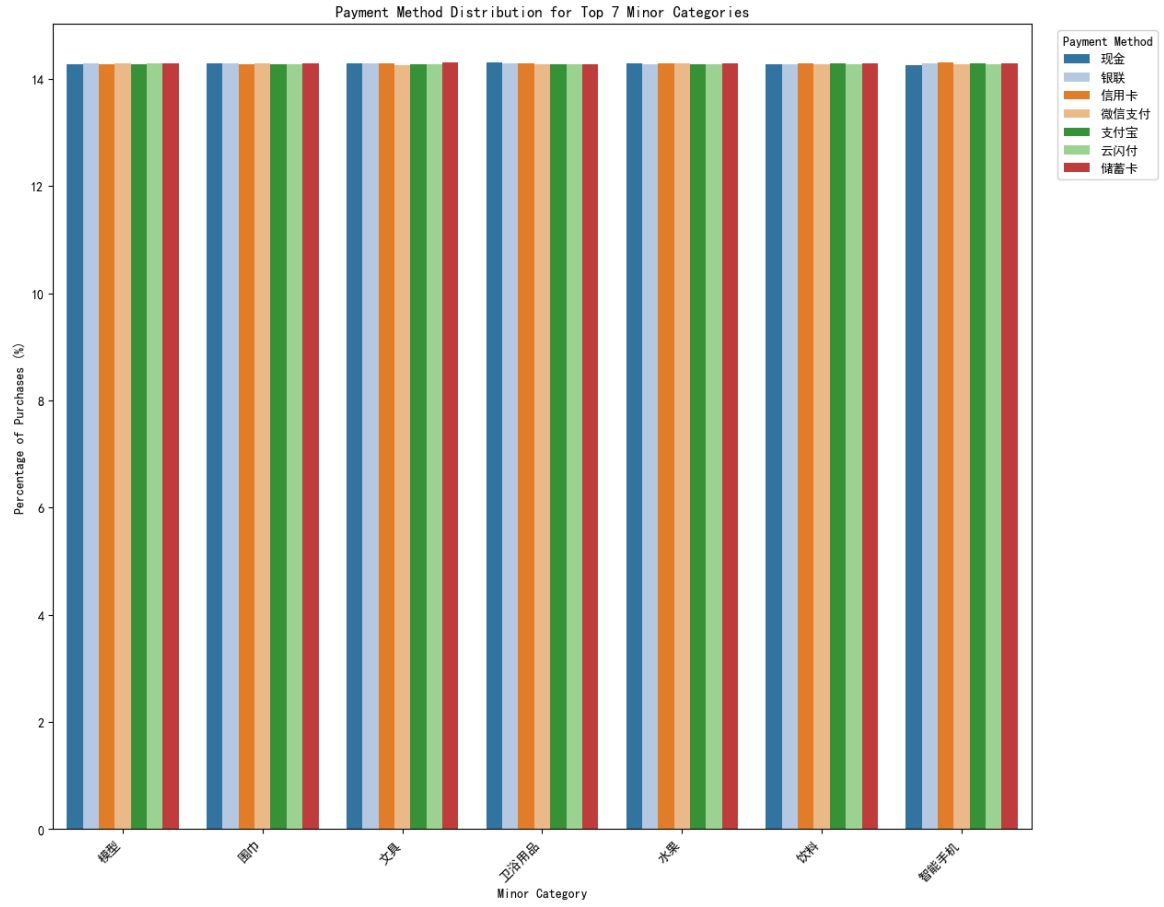
3.2 Task 2: 支付方式与商品类别分析

- **Task 2.1: 支付方式与商品大类/小类的关联规则挖掘 (FP-Growth)**
  - **结果**: 无论是分析支付方式与大类的关联（Task 2.1.a），还是与小类的关联（Task 2.1.b），在设定的支持度(0.01)和置信度(0.6)阈值下，均**未找到任何关联规则**。
  - **分析**: 频繁项集输出显示，虽然单个支付方式本身支持度很高（例如，“储蓄卡”在支付方式+小类项集中频次约1928万），但包含（支付方式+商品类别）的二元或更高阶项集支持度迅速下降。例如，在小类分析中，仅有 `[模型, 微信支付]` 的支持度（1.085%）勉强超过0.01的阈值，但仍不足以生成满足0.6置信度的规则。这表明支付方式的选择与特定商品类别的绑定不强。
- **Task 2.2: 支付方式与商品大类/小类的分布统计分析**
  - **主要发现 (大类与支付方式分布，Task 2.1.a/b的统计部分)**:

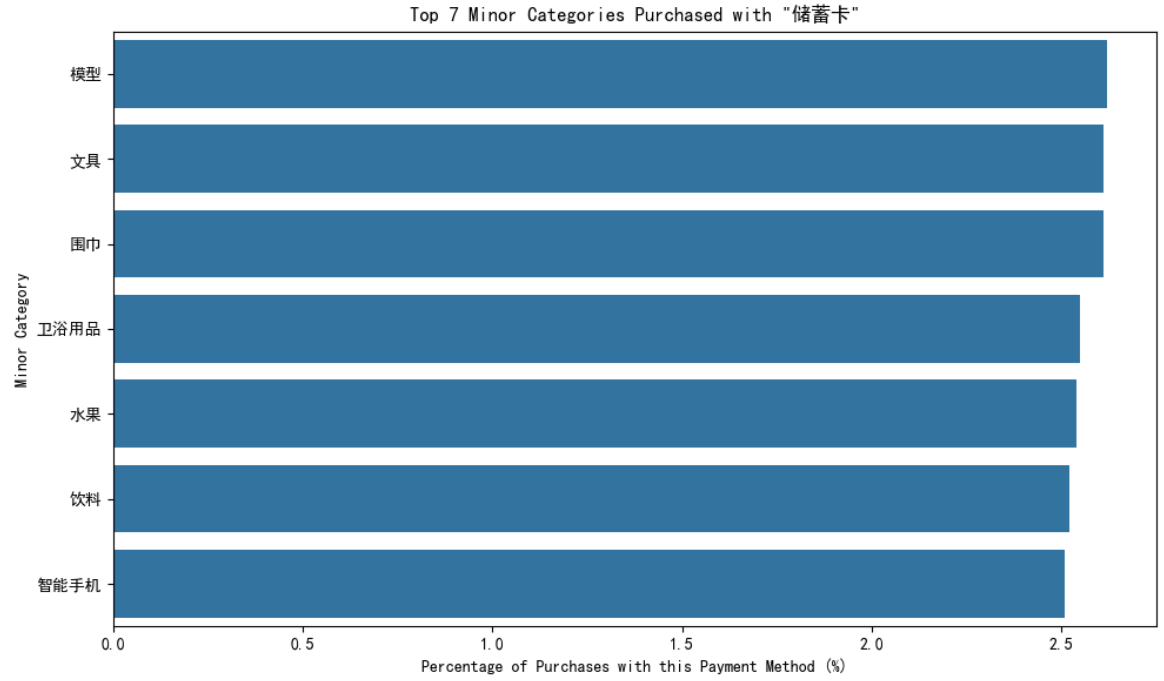
- 如 "Payment Method Distribution for All Major Categories" 图所示，对于所有9个商品大类（办公、家居、服装、母婴、汽车用品、玩具、电子产品、运动户外、食品），各种支付方式（储蓄卡、现金、银联、信用卡、支付宝、云闪付、微信支付）的使用百分比几乎完全相同，均在 **14.2% 到 14.3%** 之间。
- 

- 同样，每种支付方式购买的各大类商品分布也显示，服装、食品、电子产品是主要购买对象，占比相似（约21%），其次是家居（约9.5%）、玩具（约7.3%），其他品类占比较低且在各支付方式间分布均匀。
- 

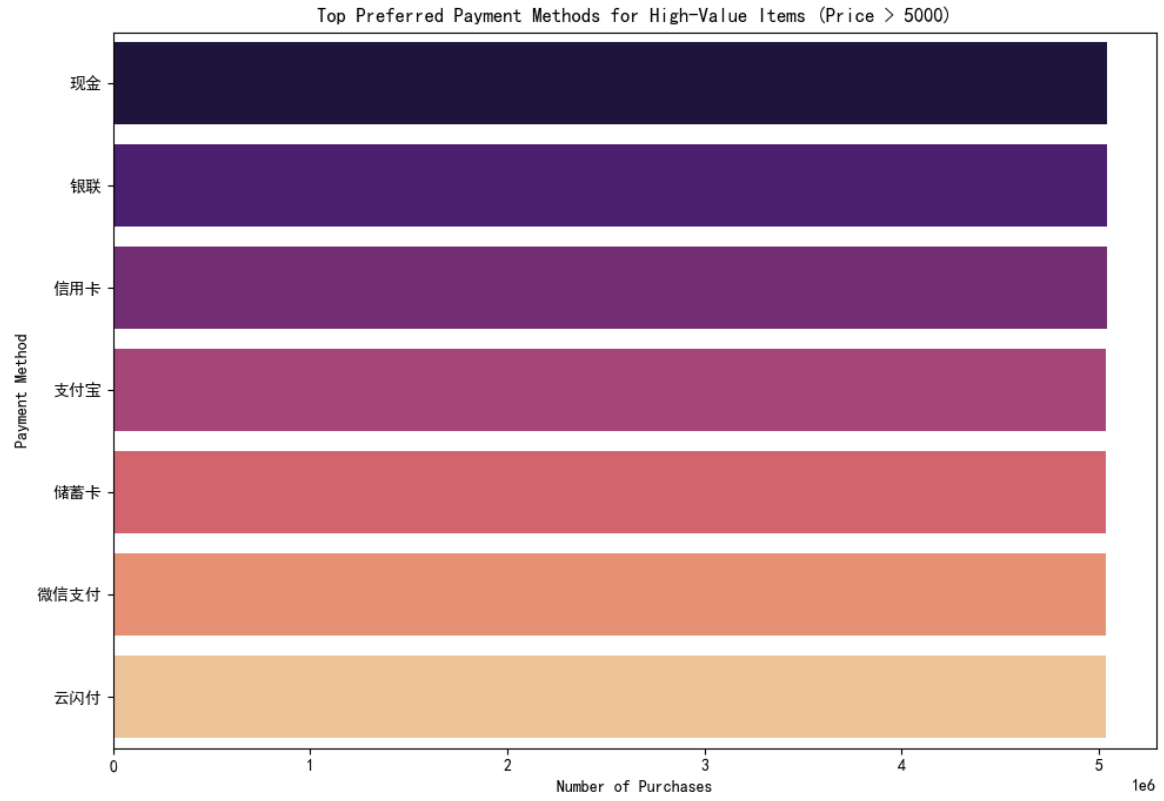
- **主要发现 (小类与支付方式分布，Task 2.2.a/b的统计部分)**:
  - 如 "Payment Method Distribution for Top 7 Minor Categories" 图所示（针对模型、围巾、文具等7个热门小类），各种支付方式的使用百分比依然高度一致，均在 **14.3%** 左右。



- 对于每种支付方式，其购买的Top N个小类分布也显示出相似性，热门小类（如模型、围巾、文具）在各种支付方式下的购买占比较为平均（约2.1% - 2.6%）。



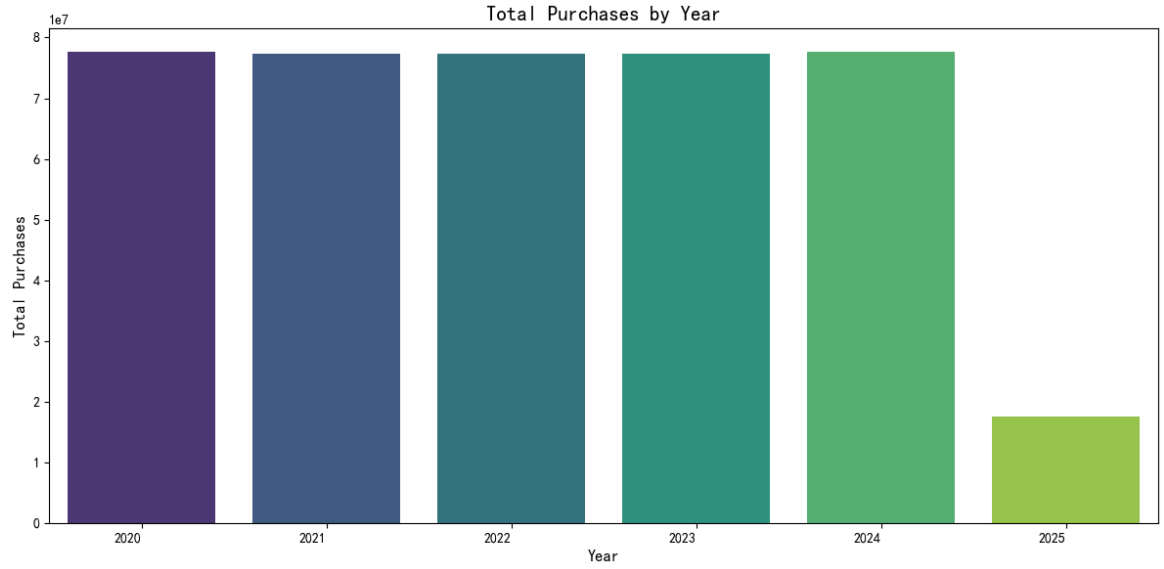
- **Task 2.3 & 2.4: 高价值商品 (单品价格 > 5000) 支付方式分析**
  - **数据**: 共筛选出 **35,275,385** 条高价值单品购买记录。
  - **支付方式分布**:
    - 如 "高价值商品 (单价 > 5000) 支付方式分布"（垂直条形图）所示，各种主流支付方式（现金、银联、信用卡、支付宝、储蓄卡、微信支付、云闪付）的购买次数非常接近，均在503万至504万之间。

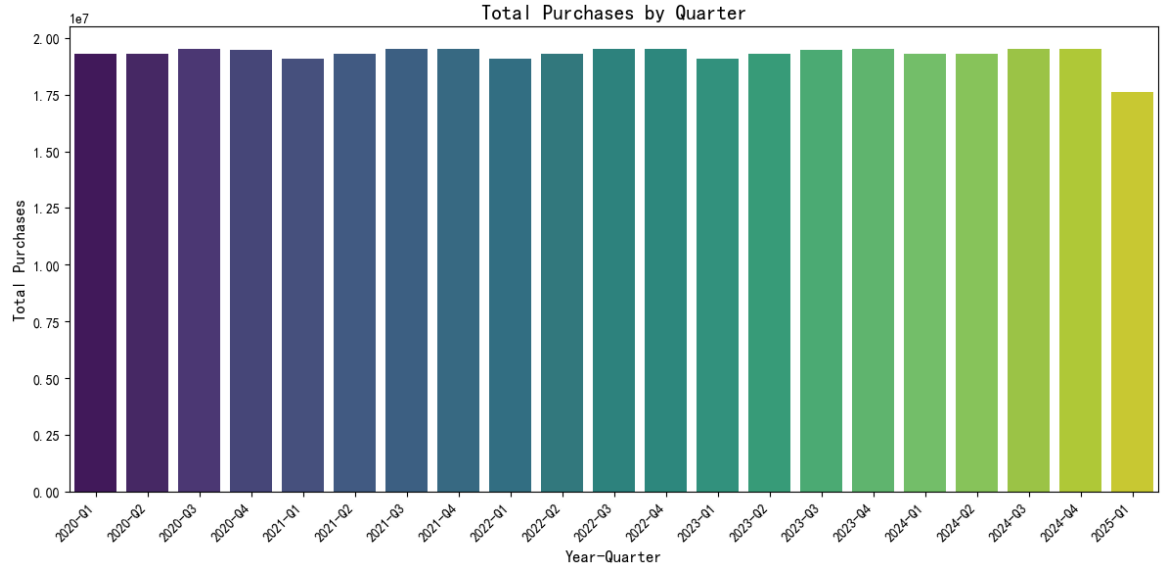


- **高价值小类支付热力图**:
  - 热力图（针对Top N高价值小类，如家具、平板电脑、摄像机等）进一步证实，对于这些特定的高价值小类，其支付方式的选择也呈现高度均衡，各支付方式占比仍在 **14.3%** 左右。

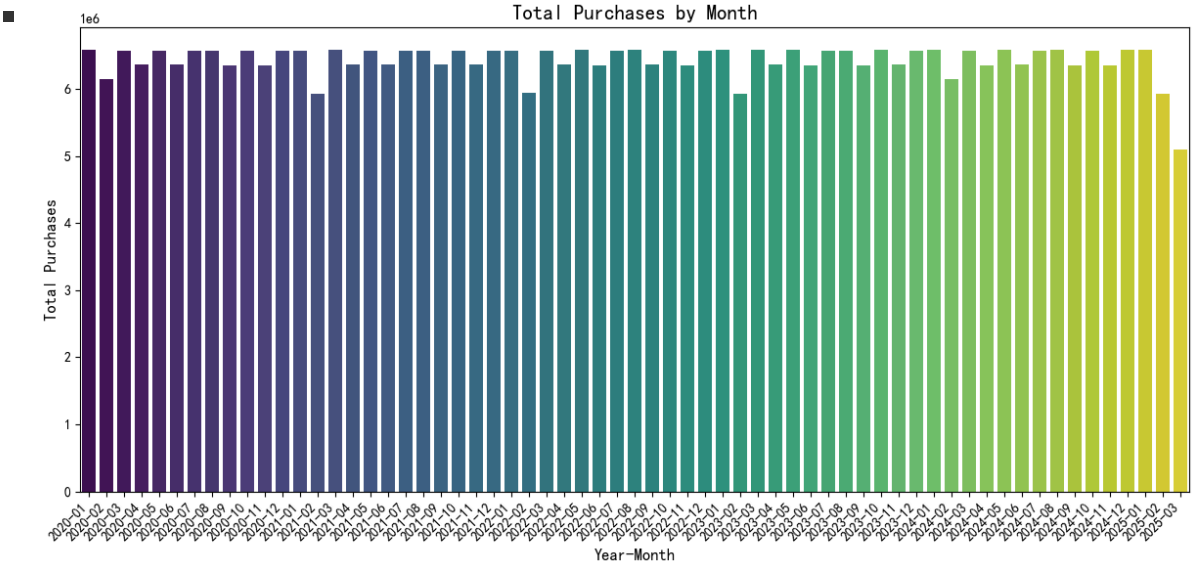
- **业务洞察 (综合Task 2)**:
  - **支付方式选择的极度均衡性**: 无论商品是大类还是小类，是普通价值还是高价值，用户对各种主流支付方式的选择几乎没有差异。这强烈表明支付渠道的成熟度和用户的无差别使用习惯。
  - **关联规则的缺乏得到印证**: 这种均衡性直接导致了无法通过FP-Growth找到满足高置信度的（支付方式-商品类别）关联规则。
  - **运营启示**: 商家应确保所有主流支付渠道的畅通和服务质量。针对特定支付渠道的商品类别促销可能效果不佳。营销重点应放在商品本身和普适性用户体验上。数据的高度均衡性可能也反映了其模拟生成的特性。

3.3 Task 3: 时间序列模式挖掘

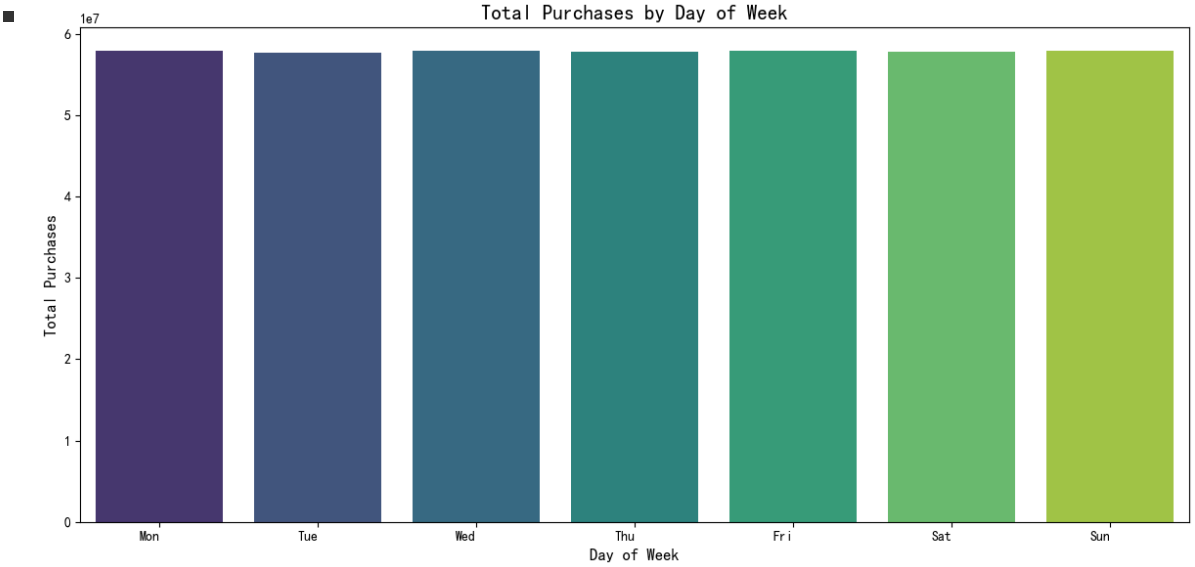
- **Task 3.1: 季节性购物模式 (整体)**
  - **年度趋势**: 2020-2024年总购买量稳定（约7700万/年）。2025年数据不完整 (1760万)。
    - 

  - **季度趋势**: Q3、Q4略高于Q1、Q2。
    - 

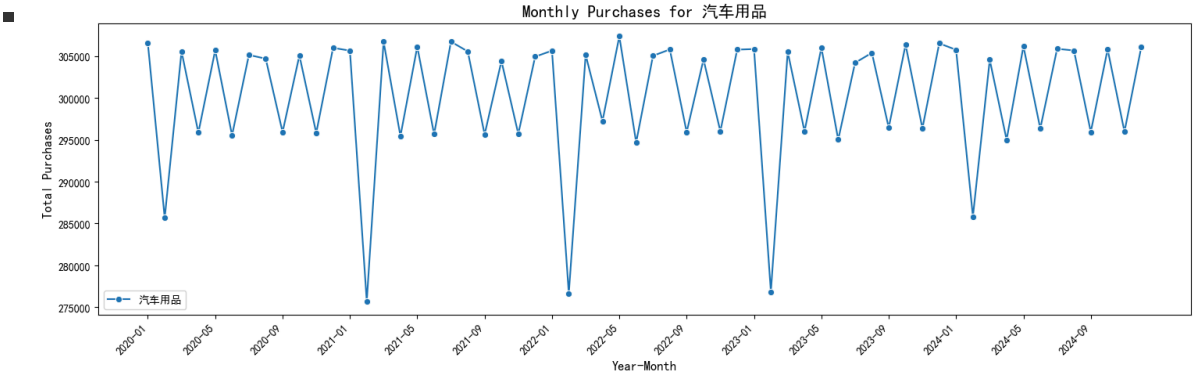
  - **月度趋势**: 每年2月为显著低谷。存在“奇数月高、偶数月低”波动。11月未见“双十一”购买件数高峰。

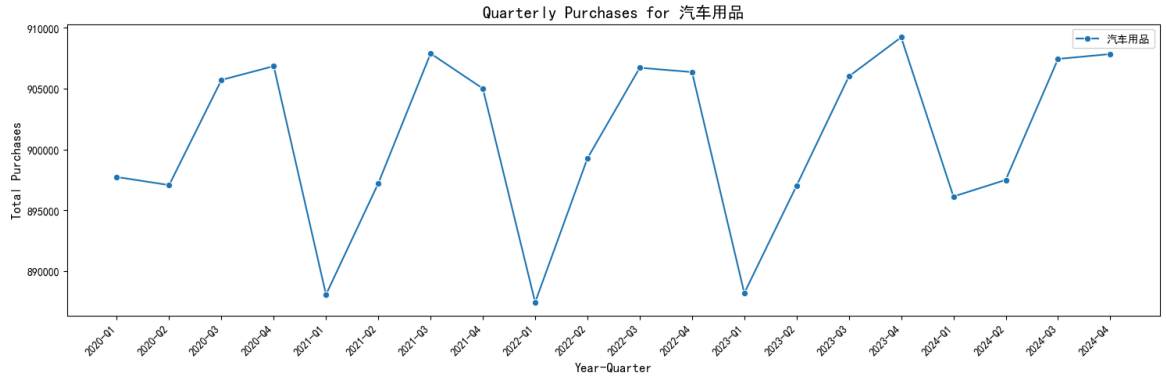


- **周内趋势**: 一周七天购买量非常平均。

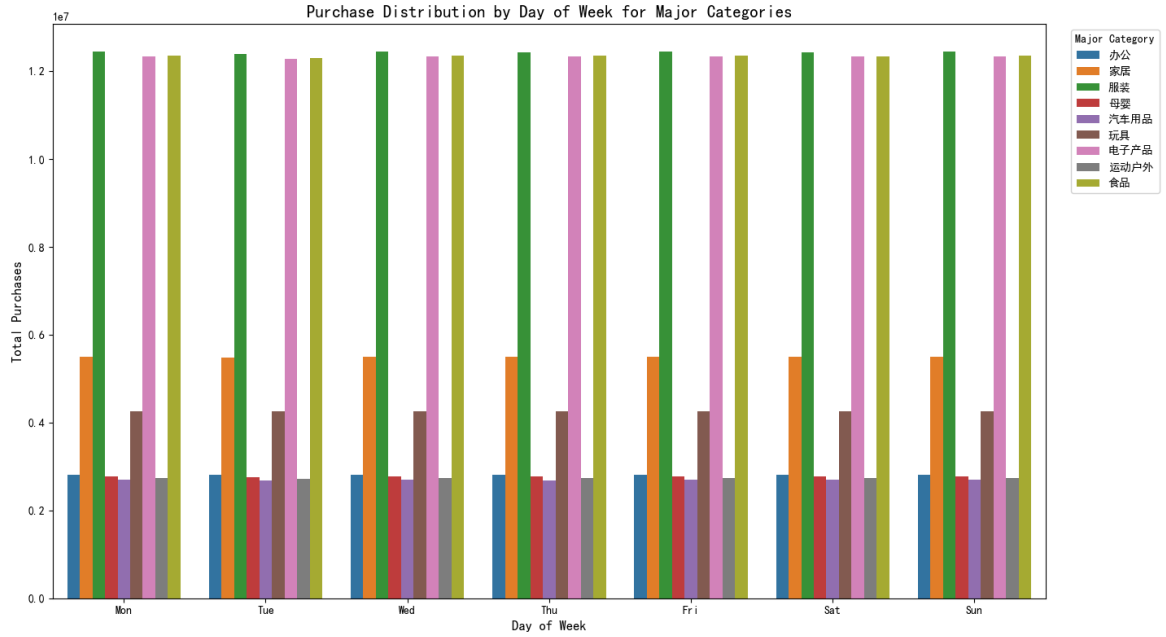


- **Task 3.2: 特定商品大类购买频率变化**
  - **主要发现**: 所有大类的月度、季度购买趋势与整体趋势高度一致。





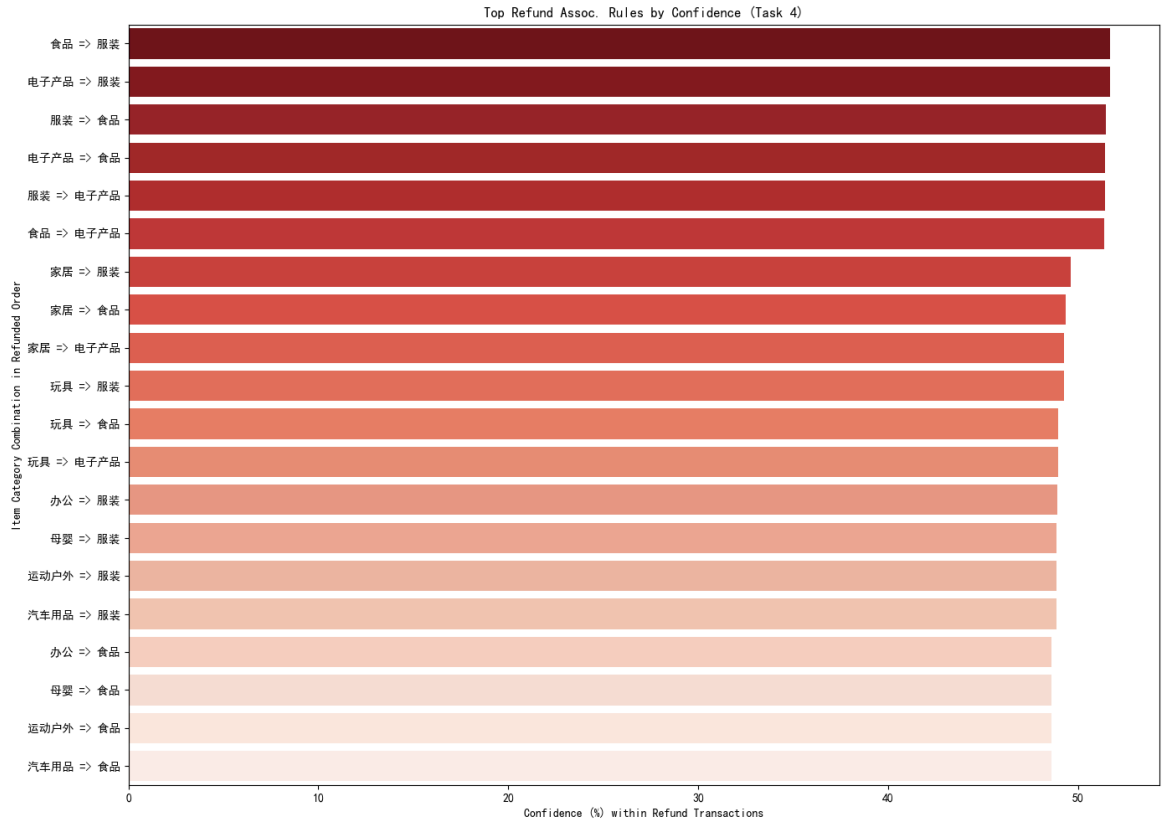
- 周内分布对所有大类也均一。



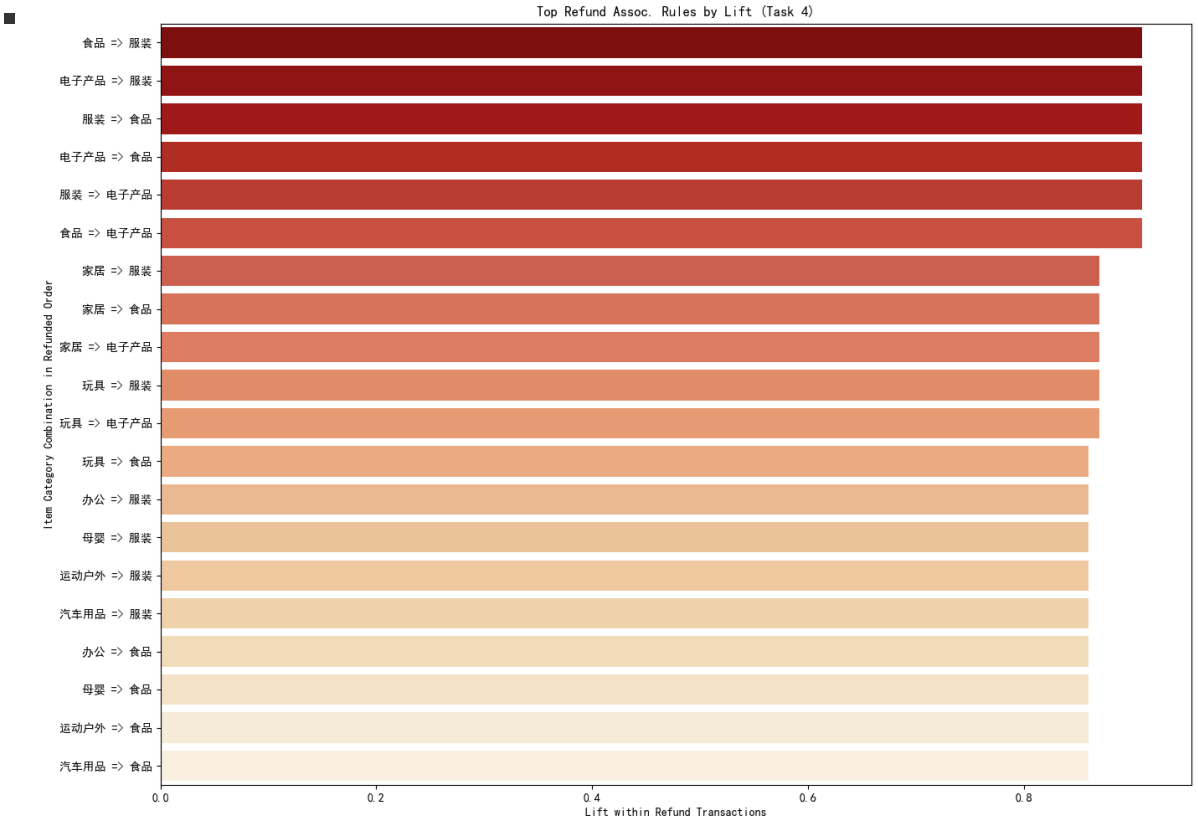
- **Task 3.3: "先A后B"时序模式 (大类)**
  - **结果**: 运行输出显示，分子（A后买B的次数）、分母（购买A的总次数）以及最终的置信度表均为空。
    - `Numerator counts (A item -> B item in consecutive transactions): +------------+----------+---------------------------------+ |prev_cat_num|curr_cat_num|count_prevA_currB_pair_occurrence| +------------+------------+---------------------------------+ +------------+------------+-----------------------+`
    - `Denominator counts (Prev item A occurrences): +------------+-----------+ |prev_cat_den|count_fromA| +------------+-----------+ +------------+-----------+`
    - `Top Sequential 'Confidence': +-------------+-----------+---------------+----------+----------------------+ |from_category|to_category|count_fromA_toB|count_fromA|sequential_confidence| +-------------+-----------+---------------+-----------+----------------------+ +-------------+-----------+---------------+-----------+----------------------+`
    - `No sequential confidence rules found for Task 3.3.`
  - **分析**: 这表明在当前定义的"相邻交易"和"大类"级别上，没有足够强的、可被识别的用户购买序列偏好。可能的原因包括：用户购买转换 действительно随机，或者大类粒度太粗，或者数据本身的均一性导致。
- **业务洞察 (综合Task 3)**:
  - 全局性的时间模式（2月低谷、下半年略旺）对所有品类普适，可用于指导整体运营。

- 缺乏品类特有的季节性爆点和周内差异，以及时序关联的缺失，进一步指向数据的高度同质化特征，这可能是数据模拟的结果。若为真实数据，则市场行为高度一致。

3.4 Task 4: 退款模式分析 (大类)

- **目标**: 挖掘与"已退款"或"部分退款"状态相关的商品大类组合。参数：支持度 ≥ 0.005，置信度 ≥ 0.4。
- **数据准备**: 共 **68,410,975** 个包含至少两种不同大类的退款订单用于FP-Growth。
- **主要发现**:
  - **频繁项集**: 退款订单中最频繁的项集（如“服装”、“食品”、“电子产品”及其组合）与所有订单中的情况非常相似。
    - 

  - **关联规则**: 共找到 **108条** 满足条件的关联规则。然而，所有规则的提升度（`lift_in_refunds`）均**小于1**（主要在0.86到0.91之间）。
    - 例如：“食品 => 服装”(置信度51.71%，提升度0.91)。
    - [在此插入Task 4的 "Association Rules from Refund Transactions" 表格截图部分内容]
- **可视化**:
  - 条形图（按置信度和按提升度排序）均显示了这些高置信度但低提升度的规则，主要围绕核心三大类的组合。



- **业务洞察**:
  - **未发现高风险退款组合（大类层面）**: Lift < 1 表明，在退款订单中，这些大类组合的共现性并未异常增高。因此，基于大类分析，没有特定的大类组合是退款的强预警信号。
  - **退款原因的普遍性**: 退款可能更多与单品因素或非商品组合因素相关。热门商品因购买量大，在退款中出现的频率也高，但这不代表它们本身“更容易”被组合退款。

4. 总结与建议

本次对约30GB用户购买数据集的深入挖掘，揭示了该数据集表现出显著的**行为模式均一性**。无论是商品大类间的关联（高频共现但驱动性弱）、支付方式在各品类及高价值商品中的均衡选择、还是各品类在时间序列上的同质化波动，以及退款订单中商品组合模式与整体购买模式的相似性，都指向了这一核心特征。

- **核心品类**: “服装”、“食品”、“电子产品”是贯穿各项分析的绝对核心，是业务运营的基石。
- **关联性解读**: 数据中的强关联更多是“共现性”而非“因果性”或“强驱动性”，Lift值在其中起到了关键的甄别作用。
- **时间模式**: 全局性的月度（2月低谷）、季度（下半年略高）模式对所有品类适用。周内购买高度平均。
- **支付与退款**: 支付方式选择高度均衡，未发现与特定品类或退款模式的强绑定。

建议:

1. **数据特性研判**: 强烈建议确认数据的来源和生成机制。若为模拟数据，当前发现的均一性是符合预期的，后续分析应考虑此背景。若为真实数据，这种均一性本身即是重大发现，值得深入研究其背后的市场或用户行为原因。
2. **深化细粒度探索**:
   - **小类/SKU层面**: 将关联规则（Task 1, Task 4）和时序分析（Task 3.3）下探到商品小类或具体商品ID，可能发现更具业务价值的精细模式。
   - **用户分群**: 对用户进行画像和分群，再对不同群体分别执行上述分析，可能揭示差异化的行为模式。
3. **运营策略调整**:
   - 利用核心品类的高流量进行交叉推荐，但避免基于大类的强行捆绑。

    - 确保全支付渠道畅通。针对特定支付渠道的商品促销可能效果有限。
    - 根据全局时间模式调整库存和营销。
4. **退款原因深究**: 结合更丰富数据（如退款原因文本、用户评论、供应商信息）分析退款，而非仅依赖商品组合。

5. 局限性

- **数据均一性的影响**: 可能掩盖了部分真实存在的、更细微的差异化模式。
- **信息维度有限**: 未能结合用户画像、促销活动、外部市场环境等因素。
- **时序分析的简化**: Task 3.3采用的相邻交易分析较为初步，未涵盖更长时间跨度或复杂序列。

6. 未来工作

- 基于数据特性研判结果，调整后续分析策略，例如更侧重于用户行为的异常检测而非寻找普适性强规则。
- 引入外部数据或更丰富的用户行为数据，进行多维交叉分析。
- 尝试更高级的序列模式挖掘算法和用户分群技术。