互评作业 2：频繁模式挖掘

周明轩 3120240977

GitHub：https://github.com/Zhoumingx/BIT-DataMining

1. 数据说明

本次作业使用作业 1 给定的 30G 数据，重点分析 purchase_history 字段。该字段由一个 JSON 对象构成，如下所示：

字段名	数据类型	示例
purchase_history	object	{"avg_price":9496,"categories":"零食","items":[{"id":7265}],"payment_method":"现金","payment_status":"已支付","purchase_date":"2023-07-30"}

具体来说，purchase_history 详细记录用户的购买行为，包括：avg_price：平均消费金额（如：9496）；categories：消费品类（如："零食"）；items：购买商品列表（每项包含商品 ID）；payment_method 与 payment_status：支付方式（如"现金"）及支付状态；purchase_date：最近一次购买日期（如："2023-07-30"）。

针对 items 购买商品列表，给定商品 ID 和种类的对照表，如下所示：


```
{
  "products": [
    {
      "category": "上衣",
      "id": 1,
```

```
    "price": 231.75
  },
  {
    "category": "儿童课外读物",
    "id": 2,
    "price": 96.95
  },
  {
    "category": "帽子",
    "id": 3,
    "price": 381.72
  },
  ……
}
```


根据商品 ID 可以唯一确定一个商品品类和对应的商品价格。对于商品品类，定义如下所示的商品大类，对商品进行进一步的区分：

- 电子产品：智能手机、笔记本电脑、平板电脑、智能手表、耳机、音响、相机、摄像机、游戏机
- 服装：上衣、裤子、裙子、内衣、鞋子、帽子、手套、围巾、外套
- 食品：零食、饮料、调味品、米面、水产、肉类、蛋奶、水果、蔬菜

- 家居：家具、床上用品、厨具、卫浴用品
- 办公：文具、办公用品
- 运动户外：健身器材、户外装备
- 玩具：玩具、模型、益智玩具
- 母婴：婴儿用品、儿童课外读物
- 汽车用品：车载电子、汽车装饰

2 商品类别关联规则挖掘

2.1 任务目标

分析用户在同一订单中购买的不同商品类别之间的关联关系；找出支持度（support）≥ 0.02、置信度（confidence）≥ 0.5 的频繁项集和关联规则；特别关注电子产品与其他类别之间的关联关系。

2.2 具体实现

数据加载与预处理：从多个 .parquet 分片文件中高效读取用户交易数据。每笔订单包含一个 purchase_history 字段，为嵌套 JSON 格式，记录用户购买商品的 id 列表。结合提供的 product_catalog.json 文件，将商品 id 映射为具体的商品子类别。

商品类别归类映射：为便于分析，将商品子类别进一步归入 10 大商品类别，包括“电子产品”、“服装”、“食品”、“家居”、“办公”、“运动户外”、“玩具”、“母婴”、“汽车用品”等。构建商品子类别 → 商品大类的映射表，用于统一订单中商品的分类粒度。

构建事务数据集：对每笔订单中的商品列表进行处理，将其转换为用户在该订单中购买的商品大类集合。构造满足关联规则挖掘算法

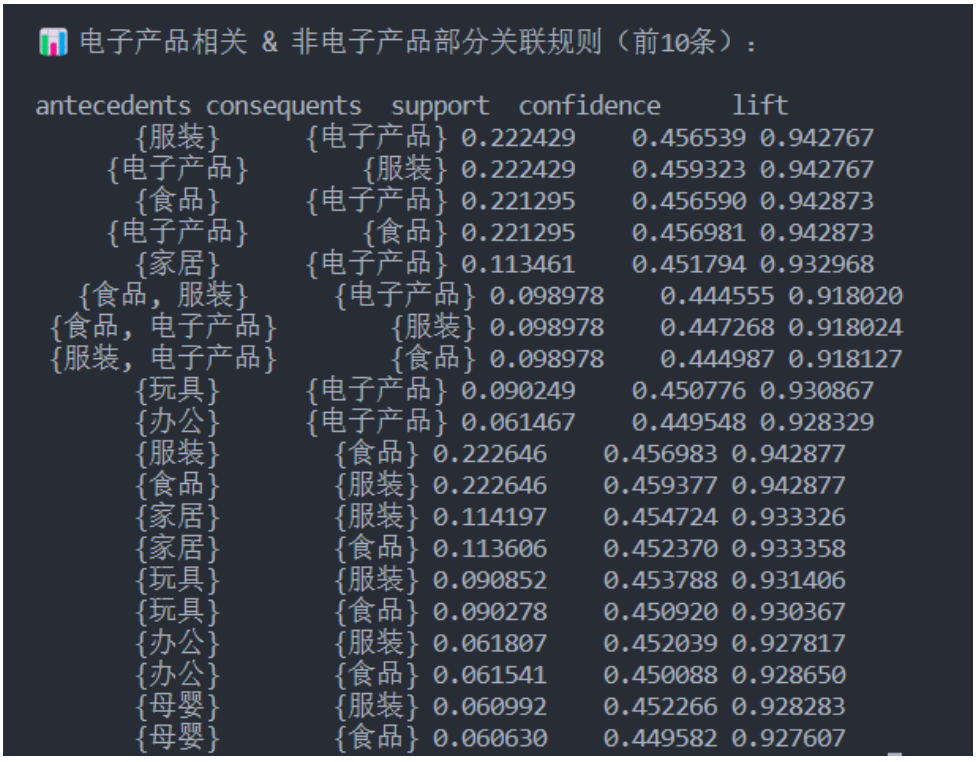
要求的事务型数据结构（Transaction List）。

频繁项集与关联规则挖掘：使用 mlxtend 提供的 Apriori 算法，对订单级事务数据进行频繁项集挖掘。设置支持度阈值为 0.02，生成所有满足条件的频繁项集。基于频繁项集生成关联规则，筛选置信度 ≥ 0.5 的规则。

聚焦电子产品关联关系：对所有生成的关联规则进行筛选，保留“电子产品”出现在前件（Antecedent）或后件（Consequent）中的规则。分析“电子产品”与其他商品类别之间的购买相关性、信赖度和提升度（Lift），识别高相关的联合购买模式。

2.3 结果展示

按照给定的支持度（support）≥0.02、置信度（confidence）≥0.5 的频繁项集和关联规则要求进行数据挖掘，发现无法得到满足上述条件的结果。降低置信度要求到 0.4，再次进行数据挖掘，结果如下：



3 支付方式与商品类别的关联分析

3.1 任务目标

挖掘不同支付方式与商品类别之间的关联规则；分析高价值商品的首选支付方式；找出支持度 ≥ 0.01、置信度 ≥ 0.6 的规则。

3.2 具体实现

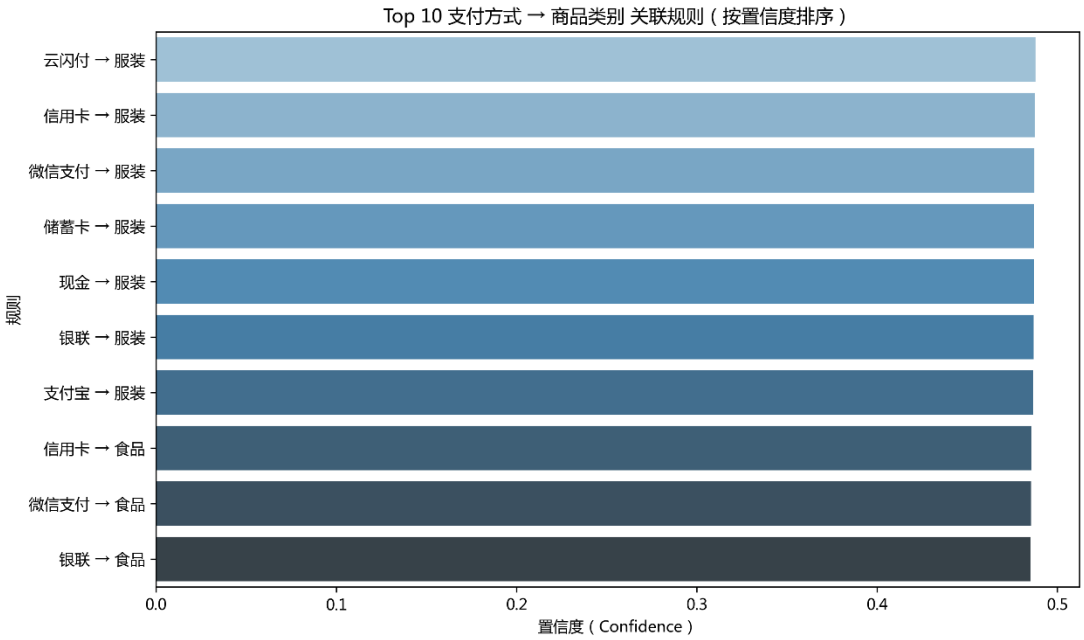
基本与上一任务相同，在数据加载与预处理、商品类别归类映射、构建事务数据集之后，使用 mlxtend.frequent_patterns.apriori() 函数挖掘支持度大于 0.01 的频繁项集，使用 mlxtend.frequent_patterns.association_rules() 提取置信度大于 0.6 的关联规则，筛选出有效规则：antecedent 为支付方式，consequent 为商品类别，代表“使用某种支付方式购买某个商品类别”的行为模式。遍历所有购买记录，判断所购商品中是否为高价值商品。统计所有含高价值商品的订单所采用的支付方式，计算频率与分布。使用 matplotlib 与 seaborn 可视化。

3.3 结果展示

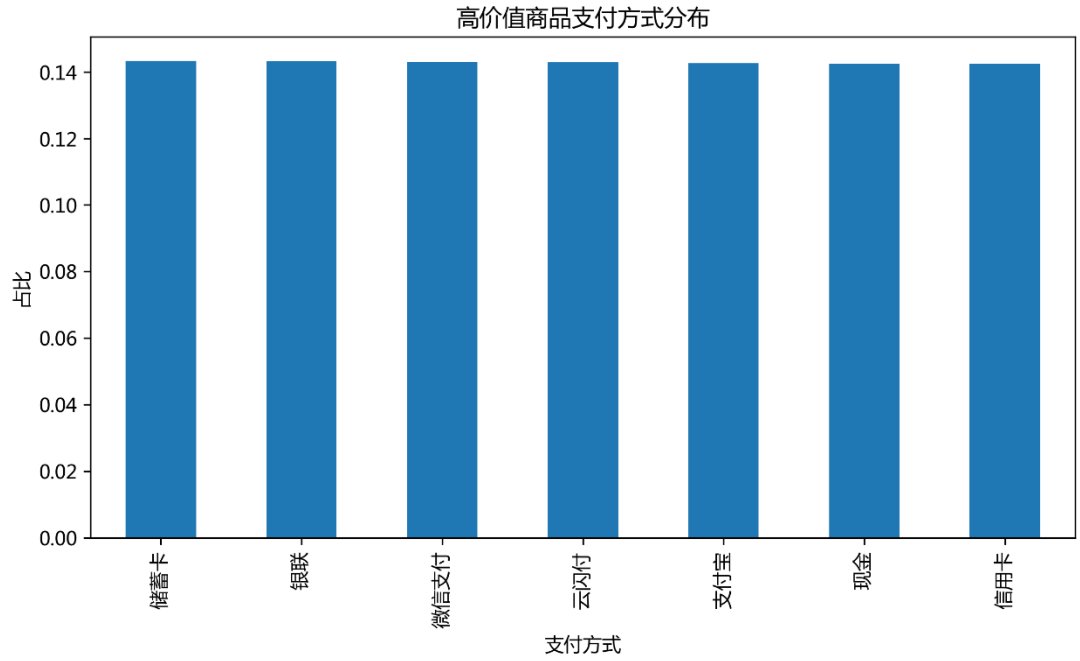
按照给定的支持度（support）≥0.01、置信度（confidence）≥0.6 的要求进行数据挖掘，发现无法得到满足上述条件的结果。降低置信度到 0.4，挖掘规则如下：

	antecedents	consequents	support	confidence	lift
5	(信用卡)	(食品)	0.069306	0.485655	1.001600
0	(云闪付)	(服装)	0.069690	0.487840	1.001432
19	(支付宝)	(电子产品)	0.069191	0.484664	1.001301
17	(微信支付)	(食品)	0.069336	0.485344	1.000958
3	(信用卡)	(服装)	0.069571	0.487510	1.000756
1	(云闪付)	(电子产品)	0.069195	0.484376	1.000706
44	(银联)	(食品)	0.069388	0.484990	1.000228
40	(银联)	(电子产品)	0.069263	0.484118	1.000174
15	(微信支付)	(服装)	0.069594	0.487149	1.000014
6	(储蓄卡)	(服装)	0.069663	0.487130	0.999975

按照置信度选取前 10 条规则，可视化如下：



高价值商品支付方式呈现均匀分布，可视化如下：



4 时间序列模式挖掘

4.1 任务目标

分析用户购物行为的季节性模式（按季度、月份或星期）；识别特定商品类别在不同时间段的购买频率变化；探索“先购买 A 类别，

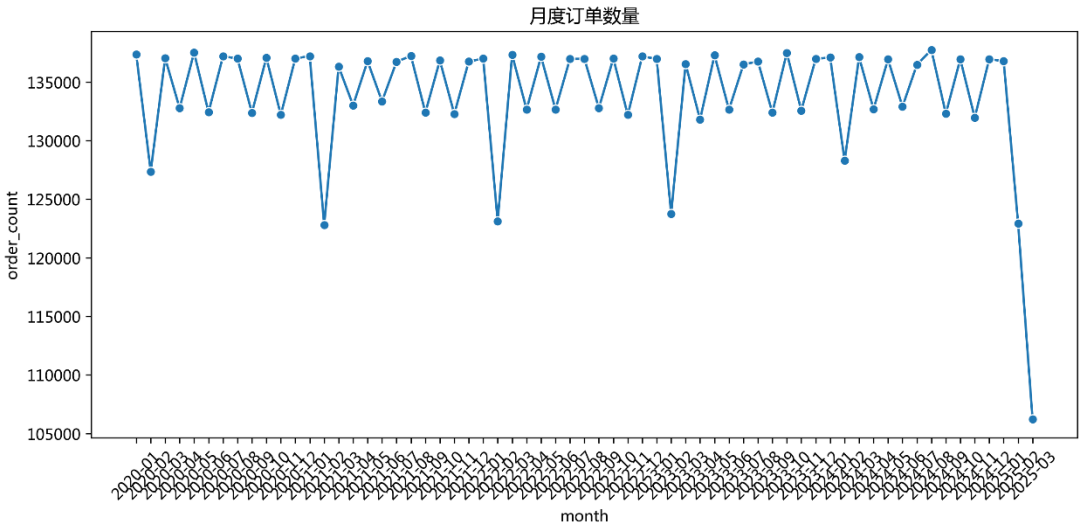
后购买 B 类别”的时序模式。

4.2 具体实现

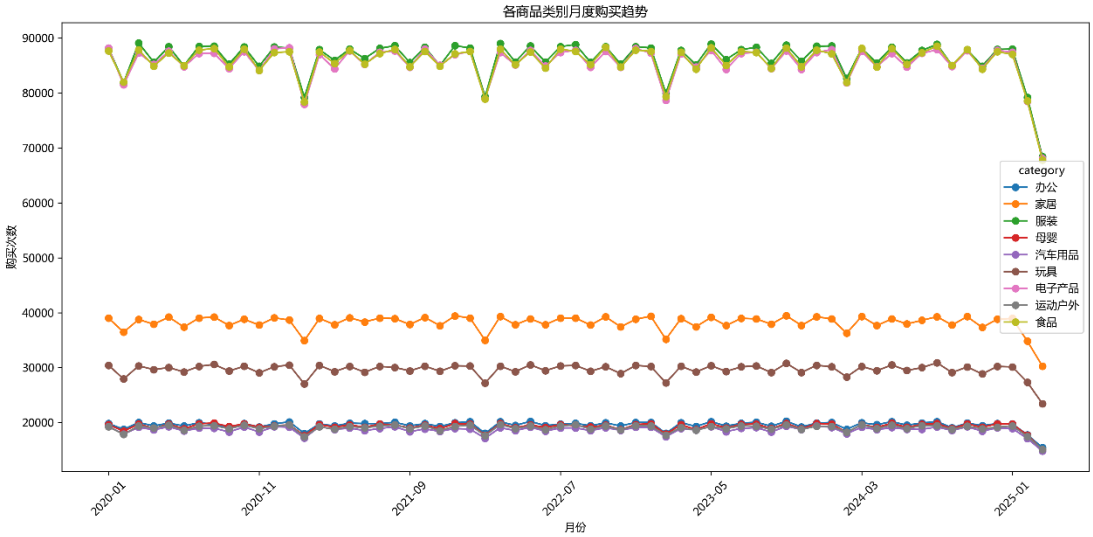
基于 purchase_history 字段中 purchase_date 和 items 的长度来统计每月的订单总量（即购买商品的次数），并进行时间序列可视化。统计每月购买行为中 item 的总数量（items 中每个元素视为一次购买）并映射到对应类别，绘制月度订单数量以及商品类别购买数量随时间变化的曲线图。对用户在不同时间的购买记录构建序列，找出常见的 A→B 类购买顺序。

4.3 结果展示

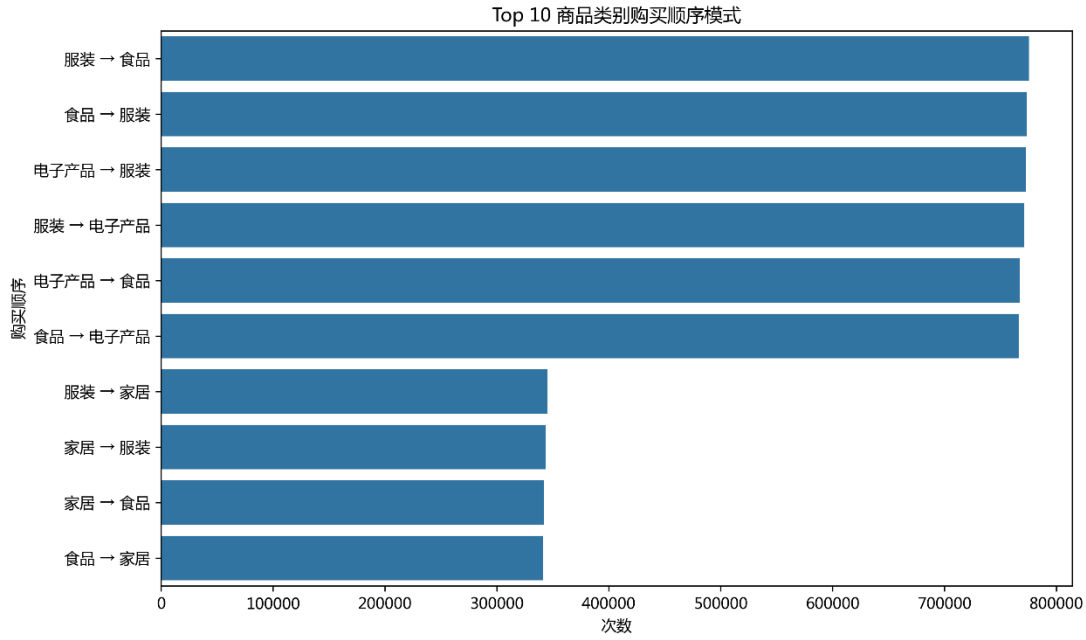
月度订单量变化折线图：



商品类别购买量随月度变化折线图：



用户商品购买顺序模式（前十）：



5 退款模式分析

5.1 任务目标

挖掘与“已退款”或“部分退款”状态相关的商品类别组合；分析导致退款的可能商品组合模式；找出支持度 ≥ 0.005、置信度 ≥ 0.4 的规则。

5.2 具体实现

具体实现基本同任务一和任务二。使用商品类别作为关联规则挖掘的 antecedents。payment_status ∈ {"已退款", "部分退款"} 作为 consequents。挖掘支持度 ≥ 0.005，置信度 ≥ 0.4 的组合。使用 mlxtend 提供的 Apriori 算法实现。

5.3 结果展示

按照给定的支持度（support）和置信度（confidence）的频繁项集和关联规则要求进行数据挖掘，结果如下：

	antecedents	consequents	support	confidence	lift
445	(玩具, 电子产品, 运动户外)	(状态:已退款)	0.005082	0.504482	1.008401
65	(办公, 母婴)	(状态:已退款)	0.008442	0.504162	1.007761
425	(玩具, 电子产品, 母婴)	(状态:已退款)	0.005131	0.503920	1.007277
298	(家居, 服装, 汽车用品)	(状态:已退款)	0.006316	0.503351	1.006140
449	(食品, 玩具, 运动户外)	(状态:已退款)	0.005100	0.503201	1.005841
73	(玩具, 办公)	(状态:已退款)	0.012521	0.503043	1.005524
49	(家居, 办公)	(状态:已退款)	0.015785	0.502871	1.005182
328	(食品, 家居, 母婴)	(状态:部分退款)	0.006472	0.502179	1.004920
368	(玩具, 服装, 母婴)	(状态:部分退款)	0.005155	0.502173	1.004907
252	(家居, 办公, 服装)	(状态:已退款)	0.006616	0.502701	1.004842