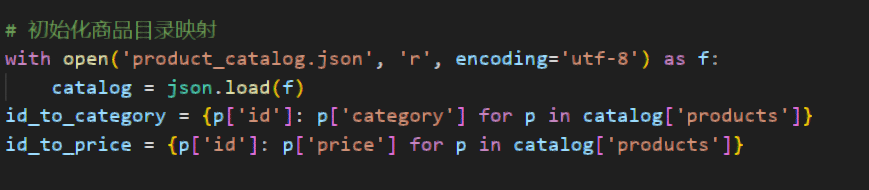
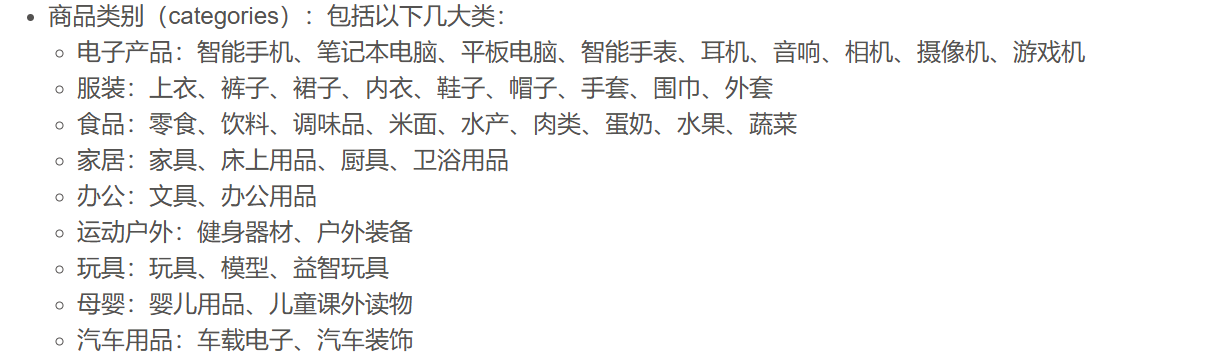
数据挖掘第二次互评作业

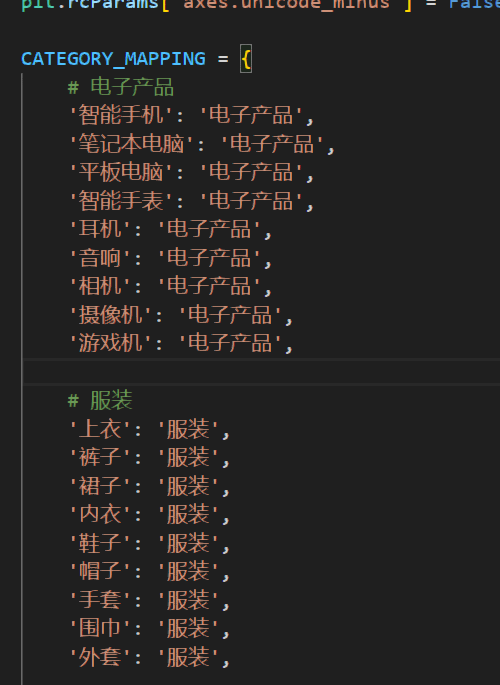
1. 数据预处理

对于该任务主要使用到的是之前作业的parquet文件中的’purchase\_history’字段，这个字段是一个json格式的信息存储，主要包含的内容是平均价格，商品类别，商品id，支付状态，支付方式和购买日期，而我们之后需要用到的内容中的商品价格是由另一个单独的json文件将商品类别，id，价格一一对应的，所以首先需要构建id对类别和价格的唯一映射。



另一方面，对于类别而言我们并不需要分析小的具体类别，所以还需要将id所对应到的具体的小类别和另一个大类别进行映射，映射关系由题目所指导。通过简单的对应进行类别归属的映射。





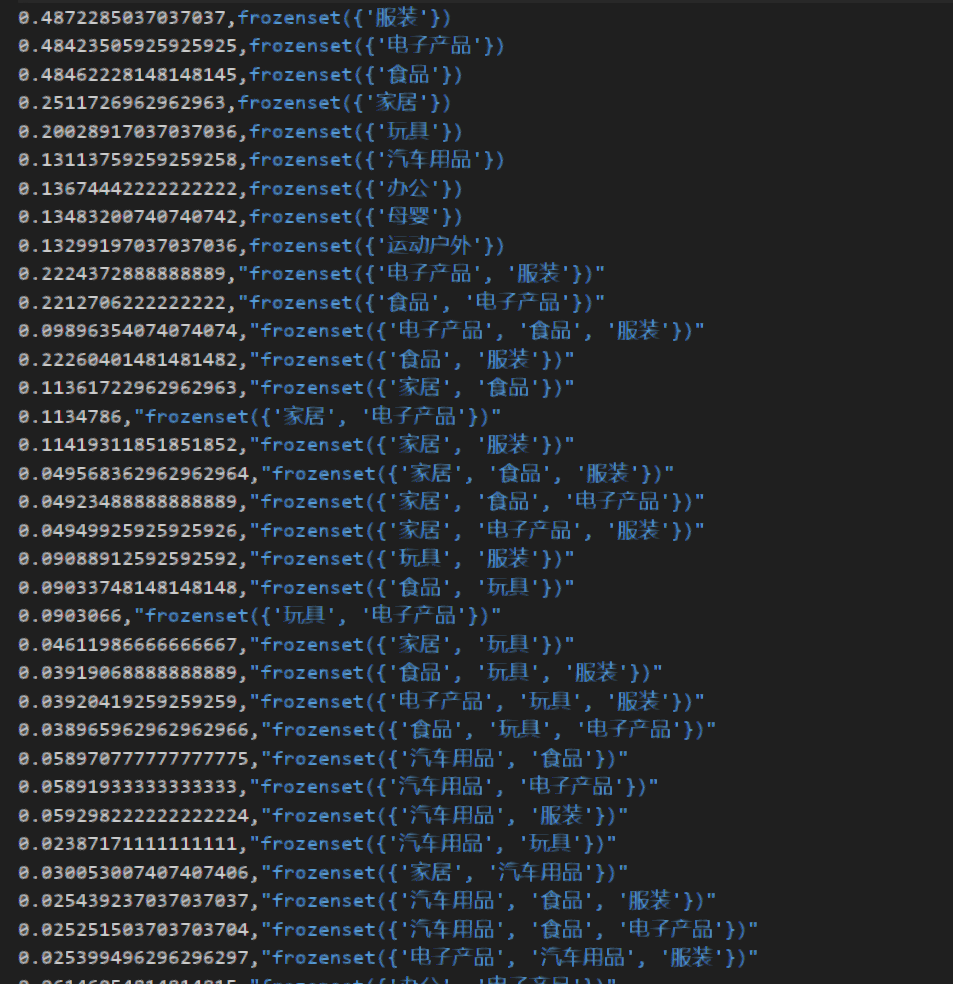
1. 商品类别关联规则挖掘

为了进行关联规则的挖掘，首先生成事务型的数据，即是每行代表一个订单的商品类别集合的数据，关键步骤是通过流式读取分批次读取文件夹内的大型parquet文件，防止内存溢出，然后通过映射对原始分类进行标准化重映射，最后转化为逗号分割的类别字符串写入csv事务文件。然后是通过fpgrowth算法计算频繁项集，需要注意的是，该算法需要所有事务信息，而直接一次性加载所有事务会导致内存的溢出，因此采用了稀疏矩阵编码的方式，先将事务数据转换为布尔矩阵然后分块加载并且补充稀疏矩阵拼接，成功解决了内存溢出的问题。



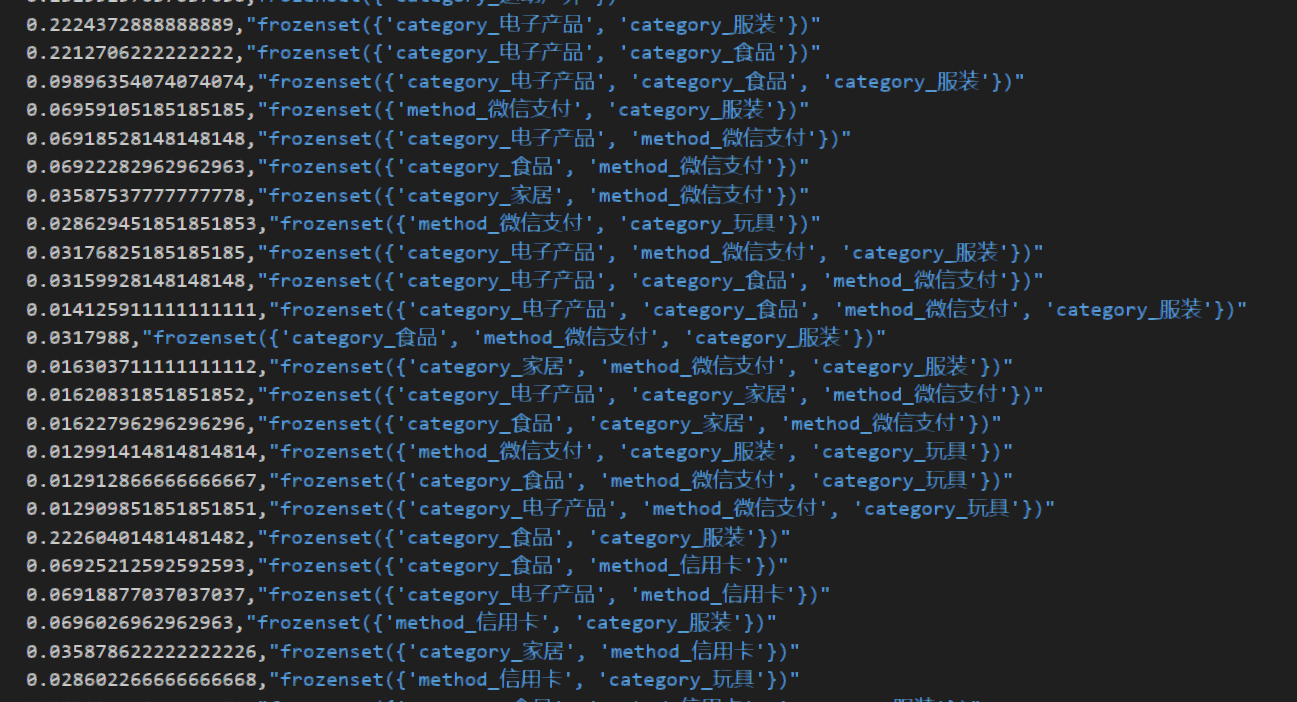
不过对分析出的结果有一定的疑惑，对于所要求的支持度0.02确实存在不少的频繁项集，但是经过计算并没有找到置信度大于0.5的关联规则，并且尝试验证了置信度大于0.4的关联规则是存在较多的从而验证了算法的正确性，并确定了确实没有大于0.5的关联规则，于是只展现了频繁项集的结果。

降序截取了一部分，其实满足支持度的频繁项集一共也只有58项左右，可能是数据的合理性还不太足够，说明多种类的商品组合的分布较为均匀，出现的频率并不高。

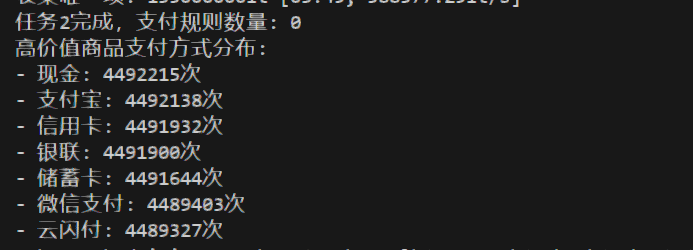


1. 支付方式与商品类别的关联分析

对于预处理部分和事务生成部分基本与上一任务一致就不多赘述，稍微有点差距的部分在于生成csv事务文件的内容格式，每个事务中包含一种支付方式和商品类别组合，然后为了防止内存溢出同样和上一任务一样采用了稀疏矩阵编码的方式，然后是按照要求的数值分析频繁项集和关联规则，遇到了相同的问题，并没有符合要求支持度和置信度的关联规则，同样也小样本降低置信度发现就存在了，但是由于数据量较大没有足够的时间全量样本重新进行关联规则的探索因此也只展示频繁项集的结果。随便从中取出一部分，在代码中没有显示的对这些频繁项集进行有一定要求的筛选，其实最后要从中选出只有一个method然后其他category的频繁项集，经过可视化之后能展示更明显的结果。



然后是分析高价值商品的首选支付方式，这部分直接读取高价值商品的相关事务然后可视化即可，这里采用的简单的降序打印：

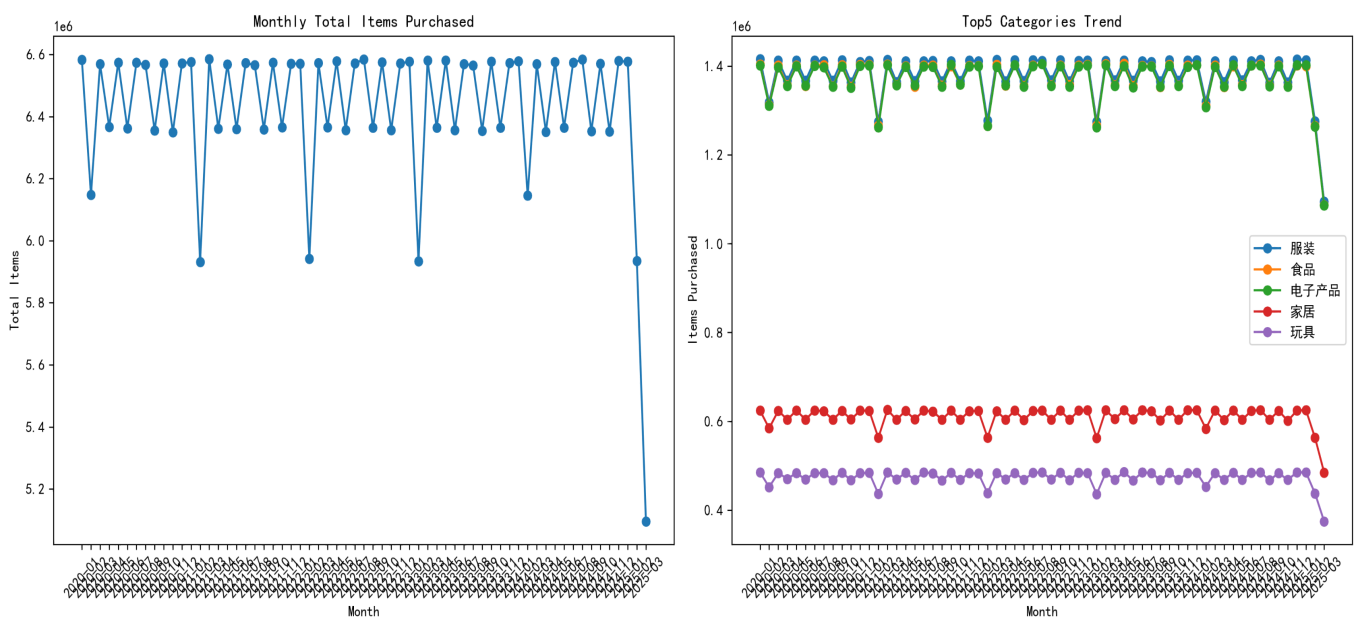


可以看到基本上在高价值商品支付方式分布中的这七种支付方式的分布较为平均，数量都是比较接近的。



1. 时间序列模式挖掘

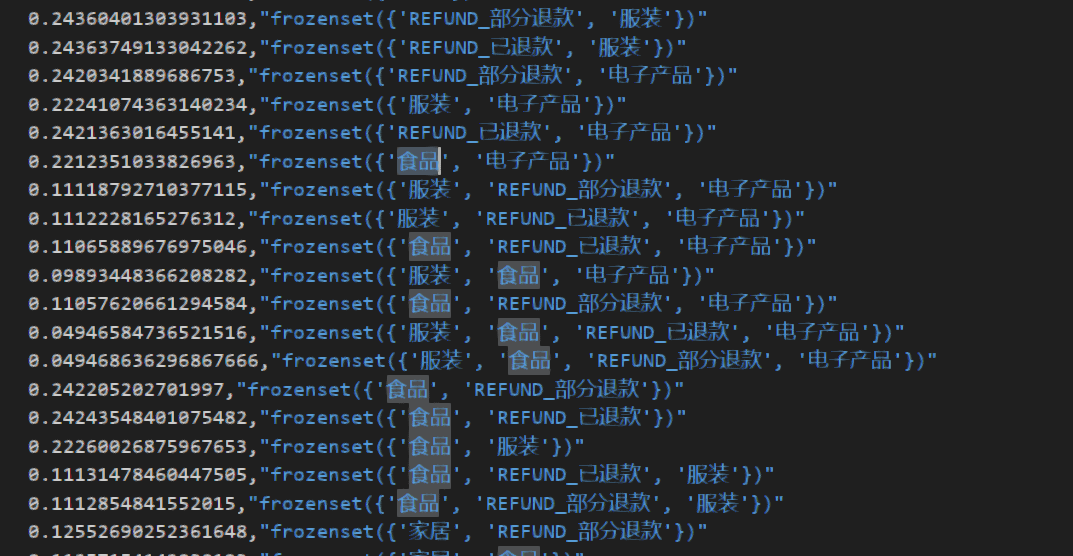
对于时间序列的分析直接采用结构化数据，对于季节性购物模式采用的方法是按照月度进行分析，分析的主要是根据当月总数和当月各品类数量来展示不同物品在不同时间段的销量：



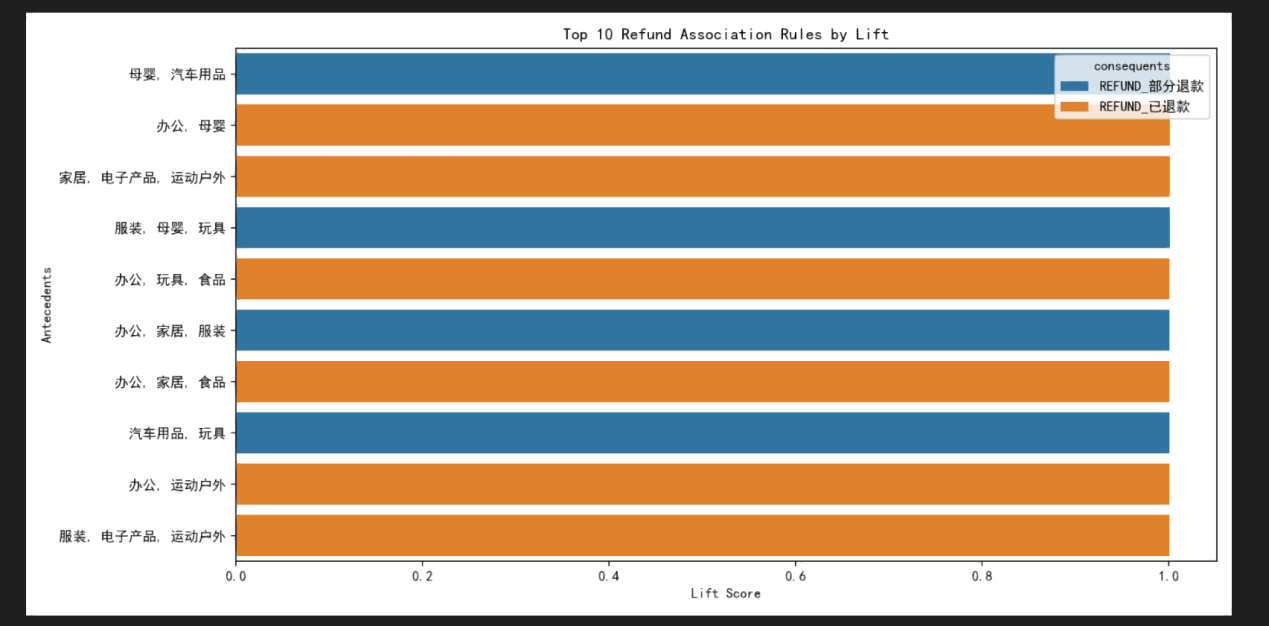
根据结果可以发现服装，食品和电子产品的购买量普遍比家居和玩具高得多，另外就月份而言基本的趋势是一个月购物量高下一个月购物量就会下降，在2025年3月出现了断崖式下降，暂不清楚原因，可能是由于数据本身不足。

1. 退款模式分析

最后一个任务是挖掘与退款相关的商品类别的关联规则，同样是同任务1,2一样的完成思路，先生成事务文件，然后分批加载事务文件通过稀疏矩阵防止内存溢出，也同样是在生成的事务格式方面进行一定的修改然后使用fpgrowth算法寻找频繁项集然后寻找关联规则。对于该任务的要求而言能找到很多关联规则，简单展示部分，其余可在csv文件里详细找到。







可以看到提升度排行最高的十条规则，主要是以上几种组合退款最常出现。

1. 总结

通过这次作业了解了什么是频繁项集和关联规则，主要学习到的内容是如何处理内存溢出问题，对于如此大的数据将其加载进内存需要通过分块再组合的方式尽可能节省内存的消耗，并且对于数据的一些格式化处理技巧也得到了加强，主要的问题在于对于很多应该保存的关键分析数据没有进行保存导致在处理大量数据失败后还需要耗费大量的时间重新处理，延长了作业的时间，导致时间略有不足。