C

2020 MCM 周末 2 问题 C：数据的财富

在其创建的在线市场中，亚马逊为客户提供了对购买进行评分和评价的机会。个人评级-称为“星级”-使购买者可以使用 1（低评级，低满意度）到 5（高评级，高满意度）的等级来表示他们对产品的满意度。此外，客户可以提交基于文本的消息（称为“评论”），以表达有关产品的更多意见和信息。其他客户可以在这些评论中提交有帮助或无帮助的评分（称为“帮助评分”），以协助他们自己的产品购买决策。公司使用这些数据来深入了解其参与的市场，参与的时间以及产品设计功能选择的潜在成功。

阳光公司计划在在线市场上推出和销售三种新产品：微波炉，婴儿奶嘴和吹风机。他们已聘请您的团队作为顾问，以在与其他竞争产品相关的客户提供的过去评级和评论中识别关键模式，关系，度量和参数，以：1）告知其在线销售策略；2）识别潜在的重要设计功能，以增强产品的合意性。阳光公司过去曾使用数据为销售策略提供信息，但他们以前从未使用过这种特殊的组合和数据类型。阳光公司特别感兴趣的是这些数据中的基于时间的模式，以及它们是否以有助于该公司制造成功产品的方式进行交互。

为了帮助您，阳光公司的数据中心为您提供了该项目的三个数据文件：hair_dryer.tsv，microwave.tsv 和 pacifier.tsv。这些数据代表在数据指示的时间段内在亚马逊市场上出售的微波炉，婴儿奶嘴和吹风机的客户提供的评分和评论。还提供了数据标签定义的词汇表。提供的数据文件包含您应用于此问题的唯一数据。

要求：

1. 分析提供的三个产品数据集，以使用数学证据来识别，描述和支持有意义的定量和/或定性模式，关系，量度和参数，这些数据将在有助于评估阳光公司的星级，评论和帮助等级之内和之间 在三个新的在线市场产品中都取得了成功。

2. 使用您的分析来解决阳光公司市场总监的以下特定问题和要求：

a.一旦三种产品在在线市场上出售后，就可以根据评级和评论确定最能为阳关公司跟踪的数据度量。

b.在每个数据集中识别并讨论基于时间的度量和模式，这些度量和模式可能表明产品在在线市场中的声誉在上升或下降。

c.确定最能表明潜在成功或失败产品的基于文本的度量和基于评级的度量的组合。

d.特定星级会引起更多评论吗？例如，在看到一系列低星级评级之后，客户是否更有可能撰写某种类型的评论？

e.诸如“热情”，“失望”之类的基于文本的评论的特定质量描述符是否与评分水平紧密相关？

3.写一两页的信给阳光公司市场总监，总结您团队的分析和结果。包括针对您的团队最有信心地推荐给市场总监的结果的具体理由。

您提交的内容应包括：
•一页摘要表
•目录
•一到两页的信函
•您的解决方案不超过 20 页，最多包含摘要页，目录和两页信函的 24 页。
注意：参考列表和任何附录不计入页数限制，应在完成解决方案后出现。您不应使用未经版权法限制使用的未经授权的图像和材料。确保您引用了想法的来源和报告中使用的材料。

词汇表
帮助等级：表示在决定是否购买该产品时特定产品评论的价值。
奶嘴：一种橡胶或塑料的舒缓装置，通常为乳头状，提供给婴儿吸吮或咬咬。
审查：对产品的书面评估。
星级：在系统中给出的分数，该分数使人们可以对具有多个星级的产品进行评分。

附件：问题数据集
Problem_C_Data.zip
所提供的三个数据集包含产品用户评分和通过 Amazon Simple Storage Service（Amazon S3）从 Amazon 客户评论数据集提取的评论。
hair_dryer.tsv
microwave.tsv
pacifier.tsv

数据集定义：每行代表划分为以下几列的数据。
●市场（字符串）：撰写评论的市场的 2 个字母的国家代码。
●customer_id（字符串）：随机标识符，可用于汇总单个作者撰写的评论。
●review_id（字符串）：评论的唯一 ID。
●product_id（字符串）：审核所属的唯一产品 ID。
●product_parent（字符串）：随机标识符，可用于汇总同一产品的评论。
●product_title（字符串）：产品的标题。
●product_category（字符串）：产品的主要消费者类别。
●star_rating（int）：评论的 1-5 星评级。
●helpful_votes（int）：有用的投票数。
●total_votes（int）：评论收到的总票数。
●vine（字符串）：基于客户在撰写准确而有见地的评论方面所获得的信任，邀请客户成为 Amazon Vine Voices。亚马逊为 Amazon Vine 成员提供了供应商已提交给该程序的产品的免费副本。Amazon 不会影响 Amazon Vine 成员的意见，也不会修改或编辑评论。
●verify_purchase（字符串）：“ Y”表示亚马逊已验证撰写评论的人在亚马逊上购买了该产品，并且没有以大幅折扣收到该产品。
●review_headline（字符串）：评论的标题。
●review_body（字符串）：评论文本。
●review_date（bigint）：撰写评论的日期