**练习四（11月29日）**

**第一题：**Demo\_1.csv文件中存储了某网红奶茶品牌2024年第四季度各个门店每天的销售量，数据包含以下字段，请分析下列问题：

|  |  |  |  |
| --- | --- | --- | --- |
| date | store | sales | orders |
| 销售日期 | 门店编号 | 销售额 | 订单数量 |

（1）（15分）已知该品牌正处于门店扩张阶段，请您统计10月1日之后新增门店最多的是哪天并计算门店的日均增长数量，预测到2025年1月末时门店的总数；

（2）（15分）已知每个门店在开始营业的第一个月会进行促销，因此第一个月的平均利润率为15%，此后的平均利润率为30%，请您预测该品牌什么时候每日总利润能够达到50万；

（3）（20分）月单量与单均价为衡量每个门店销售情况的重要标准，已知月单量较高且单均价较高的店铺为金牛型店铺，月单量较低且单均价较低的店铺为瘦狗型店铺，请您识别出该品牌的金牛店和瘦狗店（不考虑开业天数小于15天的店铺）。

**第二题：**Demo\_2.csv文件主要是关于亚马逊客户行为的调查，包含了客户的个人信息（如年龄、性别）、购物行为（如购买频率、购买类别、浏览频率等）、对产品和服务的评价（如个性化推荐频率、客户评论重要性、评分准确性、购物满意度等）以及一些其他相关信息（如购物车放弃因素、改进领域等）。

（1）（20分）在Demo\_3.csv文件中，根据客户行为将Demo\_2.csv中前600个用户划分为三个类型：A 忠诚型，B 常规性，C 厌恶型，请您以这些数据为基础选择两种不同的有监督模型（分类模型）进行训练，选取75%训练集，25%测试集，选取合适指标比较两种方法的优劣；

（2）（10分）使用效果较好的一个模型对csv中最后两个用户进行分类，判断其属于哪种类型；

（3）（10分）判断哪项指标对于用户类别的影响最大；

（4）（10分）请您根据客户行为数据重新对客户进行分类，要求所分类别不少于4类，并计算新类别与旧类型的相关系数。