2024 年（第 7 届）“泰迪杯”数据分析技能赛

B 题　特殊医学用途配方食品数据分析

一、背景

特殊医学用途配方食品简称特医食品，是指为满足进食受限、消化吸收障碍、代谢紊乱或者特定疾病状态人群对营养素或者膳食的特殊需要，专门加工配置而成的配方食品，包括 0 月龄至 12 月龄的特殊医学用途婴儿配方食品和适用于 1 岁以上的特殊医学用途配方食品。在医学营养管理与治疗方面起着重要作用。

特殊医学用途配方食品在生产和销售前需要经过严格的审批和注册过程，包括安全性、有效性的评估。所以在我国对于特殊医学用途配方食品的审核有着非常严格的规定。截至 2024 年 4 月，国内仅审批通过了 182 款特医食品（含已注销）。

二、目标

1. 提取 182 款特殊医学用途配方食品产品标签、说明书（以下简称特医食品说明书）中的相关数据，并对提取的数据及 data.xlsx 数据进行预处理。

2. 统计 182 款特医食品生产概况并可视化。

3. 构建特医食品推荐系统。

三、任务

data.xlsx 记录了 182 款特医食品的基本信息，特医食品说明书文件夹中包含 182 款特医食品说明书。请根据提供的数据，完成以下任务并撰写报告，在报告中详细描述各项任务的处理思路、过程及必要的结果。

任务 1 数据预处理

任务 1.1 读取 182 款特医食品说明书，按照表 1 的要求提取【营养成分表】中“每 100kJ”列的指定营养成分数据，将提取的数据保存到文件“result1.xlsx”中，同时在报告中列出每 100kJ（千焦）中蛋白质含量最高的三种特医食品，格式如表 1（注意营养成分的单位）。

表 1

注册证号	能量(kJ)	脂肪(g)	碳水化合物(g)	蛋白质(g)	钠(mg)	氯(mg)	钾(mg)	磷(mg)
国食注字 TY20175001	100	1.2	2.5	0.72	10.1	20.8	28.5	21.3
……	……	……	……	……	……	……	……	……

注 1　若该特医食品没有对应营养成分，填充为 0。

任务 1.2 提取 182 款特医食品说明书中【产品类别】、【组织状态】、【适用人群】的数据，在 data.xlsx 数据中新增“产品类别”、“组织状态”、“适用人群”三列。以表 2 的格式将提取的数据保存到文件“result2.xlsx”中，同时在报告中列前 5 款特医食品的结果（须说明特殊情况的处理）。

表 2

序号	企业名称	产品名称	注册证号	有效期至	产品类别	组织状态	适用人群
1	SHS INTERNATIONAL LTD	纽康特特殊医学用途婴儿氨	国食注字 TY20175001	2027 年 10	氨基酸配方	粉状	食物蛋白过敏婴儿

		基酸配方食品		月 13 日			
……	……	……	……	……	……	……	……

注 2 若该特医食品没有对应信息，留空即可。

任务 1.3 根据提取的【适用人群】信息，在 result2.xlsx 中新增“适用人群类别”列，对 182 款特医食品的适用人群进行归类，类别分为“特医婴配食品”和“1 岁以上特医食品”两种，将结果保存到文件“result2.xlsx”中。

注 3 “特医婴配食品”是针对 0-12 月龄人群的特殊医学用途配方食品，“婴儿”特指 0-12 月龄人群。

任务 1.4 特殊医学用途配方食品注册号的格式为：国食注字 TY+4 位年号+4 位顺序号，顺序号第 1 位数字为“5”表示该食品为进口产品，顺序号第 1 位数字为“0”表示该食品为国产产品；4 位年号为该食品的登记年份。基于任务 1.3 的 result2.xlsx 文件，新增“产品来源”和“登记年份”两列，提取 182 款特医食品的产品来源和登记年份数据，其中产品来源分为“国产产品”和“进口产品”两种。以表 3 的格式将结果保存到文件“result2.xlsx”中，同时在报告中列出前 5 款特医食品任务 1.3 和任务 1.4 的结果。

表 3

序号	企业名称	产品名称	注册证号	有效期至	……	适用人群类别	产品来源	登记年份
1	SHS INTERNATIONAL LTD	纽康特特殊医学用途婴儿氨基酸配方粉	国食注字 TY20175001	2027 年 10 月 13 日	……	特医婴配食品	进口产品	2017
……	……	……	……	……	……	……	……	……

任务 2 生产概况可视化

任务 2.1 统计不同登记年份不同产品来源的特医食品获批量，绘制双折线图，并在报告中对结果进行必要分析。

任务 2.2 根据特医食品产品来源与适用人群类别绘制内层为饼图的旭日图，其中内层表示适用人群类别，外层表示不同适用人群类别的产品来源分布，并在报告中对结果进行必要分析。

任务 2.3 统计不同产品类别的特医食品获批量，按获批量进行降序排列，绘制柱状图，x 轴为产品类别，y 轴为获批量，并在报告中对结果进行必要分析。

任务 2.4 在同一坐标系中，分别用不同颜色绘制 182 款特医食品脂肪和蛋白质含量的频数分布直方图，并在报告中对结果进行必要分析。

任务 2.5 根据 182 款特医食品的“适用人群”绘制词云图，并在报告中分析特医食品适用人群特征。

任务 3 特医食品推荐

在**任务 1 和任务 2** 的基础上，合理运用现有数据完成推荐任务。基于客户的需求描述（如年龄段、症状、特殊说明），从 182 款特医食品中自动筛选出符合条件的产品选项，为客户提供个性化的特医食品推荐服务。实现方式不限，可以使用推荐算法或大模型，但须在报告中详细描述实现过程、推荐逻辑以及推荐结果。

基于构建的智能推荐模型（或系统），根据下列的客户需求进行特医食品的推荐。

（1）客户 1：婴儿、蛋白质过敏。

（2）客户 2：10 岁儿童、需要补充蛋白质、乳糖不耐受。

四、数据说明

赛题数据文件夹具体内容如下所示。

表 4 数据文件说明

文件名	内容说明
特医食品说明书	包含 182 款特医食品说明书（PDF 文件）。
data.xlsx	每款特医食品的具体信息，包括该产品的企业名称、产品名称、注册证号、有效期至等。