2021 年“泰迪杯”数据分析技能赛

B 题

肥料登记数据分析

一、背景

肥料是农业生产中一种重要的生产资料，其生产销售必须遵循《肥料登记管理办法》，依法在农业行政管理部门进行登记。各省、自治区、直辖市人民政府农业行政主管部门主要负责本行政区域内销售的肥料登记工作，相关数据可从政府网站上自由下载。

二、目标

1. 对肥料登记数据进行预处理。
2. 根据养分的百分比对肥料产品进行细分。
3. 从省份、日期、生产商、肥料构成等维度对肥料登记数据进行对比分析。
4. 对非结构化数据进行结构化处理。

三、任务

请根据附件 1~附件 4 中提供的数据，自行选择分析工具完成以下任务，并撰写报告。

任务 1　数据的预处理

任务 1.1　附件 1 的产品通用名称存在不规范的情况。请按照复混肥料（掺混肥料归入这一类）、有机-无机复混肥料、有机肥料和床土调酸剂这 4 种类别对附件 1 进行规范化处理。请在报告中给出处理思路、过程及必要的结果，同时将完整的结果保存到文件“**result1_1.xlsx**”中。

任务 1.2　计算附件 1 中各肥料产品的氮、磷、钾养分百分比之和，称为总无机养分百分比。请在报告中给出处理思路、过程及必要的结果，同时将完整的结果保存到文件“**result1_2.xlsx**”中，结果保留 3 位小数（例如 1.0%，即 0.010）。

任务 2　肥料产品的数据分析

任务 2.1　从附件 2 中筛选出复混肥料的产品，将所有复混肥料按照总无机养分百分比的取值等距分为 10 组。根据每个产品所在的分组，为其打上分组标签（标签用 1~10 表示），将完整的结果保存到文件“**result2_1.xlsx**”中。分析复混肥料产品的分布特点，在报告中绘制产品登记数量的直方图，给出处理思路及过程，并按登记数量从大到小列出登记数量最大的前 3 个分组及相应的产品登记数量。

任务 2.2　从附件 2 中筛选出有机肥料的产品，将产品按照总无机养分百分比和有机质百分比分别等距分为 10 组，并为每个产品打上分组标签 (1,1), (1,2), …, (10,10)，将完整的结果保存到文件“**result2_2.xlsx**”中。请在报告中给出处理思路及过程，并根据分组情况绘制有机肥料产品的分布热力图，其中横轴代表总无机养分分组，纵轴代表有机质分组。在此基础上，分析有机肥料产品的分布特点，并按登记数量从大到小列出登记数量最大的前 3 个分组及相应的产品登记数量。

任务 2.3　从附件 2 中筛选出复混肥料的产品，按照氮、磷、钾养分的百分比，使用聚类算法将这些产品分为 4 类。根据聚类结果为每个产品打上聚类标签（标签用 1~4 表示），并将完整的结果保存到文件“**result2_3.xlsx**”中。请在报告中给出处理思路及过程，根据聚类标签绘制肥料产品的三维散点图和散点图矩阵，并通过绘制聚类结果的雷达图分析每个聚类的特征。

任务 3　肥料产品的多维度对比分析

任务 3.1　从文件“result2_1.xlsx”中提取发证日期中的年份，分析比较复混肥料中各组别不同年份产品登记数量的变化趋势。请在报告中给出处理思路及分析过程，使用合适的图表对结果进行可视化。

任务 3.2　从文件“result2_2.xlsx”中提取 2021 年 9 月 30 日仍有效的有机肥料产品，将完整的结果保存到文件“**result3_2.xlsx**”中。从有效产品中分别筛选出广西和湖北（根据正式登记证号区分）产品登记数量在前 5 的组别，分析两个省份上述组别的分布差异。请在报告中给出处理过程及分析结果。

任务 3.3　从附件 3 中提取产品登记数量大于 10 的肥料企业，给出这些企业所用到的原料集合（发酵菌剂除外）。以各企业用到的原料作为特征，计算企业

之间的杰卡德相似系数矩阵，并将结果（保留4位小数）保存到文件“**result3_3.xlsx**”中（不提供模板文件，格式见表1）。请在报告中给出处理思路、过程及相似系数矩阵。

注 集合 A 与 B 的杰卡德相似系数定义为 $J(A,B)=\frac{|A\cap B|}{|A\cup B|}$，其中 $|S|$ 表示集合 S 中元素的个数。

表1　结果文件 result3_3.xlsx 的格式

	企业名称 1	企业名称 2	…	企业名称 n
企业名称 1				
企业名称 2				
…				
企业名称 n				

任务 4　肥料产品的多维度对比分析

任务 4.1　设计算法或处理流程，从附件 4 技术指标中提取出氮、磷、钾养分和有机质的百分比，以及肥料含氯的程度。请在报告中给出处理思路及过程，并将结果保存到文件“**result4_1.xlsx**”中。

注　如果技术指标中只给出总养分百分比（“≥”按照“=”处理）而无明细数据，则氮、磷、钾养分的百分比按照总百分比的 1/3 来计算，结果保留 3 位小数（例如 1.0%，即 0.010）。复混肥料属于无机肥料，它的有机质百分比设定为 0。含氯情况分为“无氯”、“低氯”、“中氯”和“高氯”4 种。如果肥料产品的技术指标中没有给出含氯情况，则视为“无氯”；如果注明“含氯”，则视为“低氯”。

任务 4.2　设计算法或处理流程，从附件 4 原料与百分比中提取各种原料的名称及其百分比。请在报告中给出处理思路及过程，并将结果保存到文件“**result4_2.xlsx**”中（参见表 2）。

表 2　结果文件 result4_2.xlsx 的示例

序号	原料名称	百分比
001	酒精废液	45%
001	蔗渣	3.52%
…	…	…
002	滤泥	0.153%

四、数据说明

附件 1~附件 4 的数据收集自农业部门官方网站，部分数据细节与实际有差别，仅供比赛使用。

附件 1 为安徽肥料登记数据，附件 2 为广西、湖北肥料登记数据。这两个附件中表的主要字段有企业名称、产品通用名称、正式登记证编号、发证日期、有效时间、产品形态、营养成分百分比、含氯情况等。其中产品通用名称实际上是肥料产品的类型，需要在省级农业行政主管部门登记的肥料有复混肥料（包括掺混肥料）、有机-无机复混肥料、有机肥料和床土调酸剂这 4 类。肥料的营养成分百分比指标，通常标记出属于无机成分的氮、磷、钾的含量，以及有机质的含量。我国规定，氮肥成分以总氮的质量来计算含量，磷肥成分按磷元素的量折算成五氧化二磷（P2O5）的质量来计算含量，钾肥成分按钾元素的量折算成氧化钾（K2O）的质量来计算含量。注意，肥料正式登记证有效期为 5 年，可以续期，会出现有效期距发证日期大于 5 年的情况。

附件 3 给出了某省登记肥料的产品配方，相比附件 1 和附件 2 增加了关于肥料原料的信息。

附件 4 给出了某省肥料登记数据中营养成分及原料构成的原始数据。字段技术指标以字符串的形式给出了肥料的营养成分的百分比。例如某复混肥料的技术指标字段取值为“N+P2O5+K2O≥20%（7-10-3） 有机质≥20% 含氯”，表示肥料中氮磷钾三大元素的总养分含量不小于 20%；“（7-10-3）”指的是氮磷钾的配比，氮含量为 7%，磷肥成分（折算为 P2O5）含量为 10%，钾肥成分（折算为 K2O）含量为 3%；“有机质≥20%”表示肥料中有机质的含量不小于 20%；“含氯”表示肥料中含有氯元素。有机肥料由于不含无机养分或含量较少，有些产品

只在技术指标中标明“总养分≥…%”，没有给出氮、磷、钾 3 大元素的具体含量。字段原料与百分比以字符串的形式给出了肥料的原料构成及质量百分比，例如某有机肥料的原料与百分比字段取值为“糖蜜酒精废液（占 25%),发酵菌种(占 1%),木糠（占 25%),滤泥（占 49%)”，表明了该有机肥料由蜜糖酒精废液、发酵菌种、木糠和滤泥四种原料构成，质量百分比分别是 25%、1%、25%及 49%。

注 本赛题中，不同的正式登记证号代表不同的产品。

五、关于竞赛成果提交的说明

1. 登录方式

请使用**队长**的账号登录数睿思网站（www.tipdm.org），进入第四届技能赛页面（https://www.tipdm.org:10010/#/competition/1422385224767152128/introduce）。为保证成功提交，**请使用谷歌浏览器无痕模式**。

2. 作品提交

报告以 PDF 格式提交，文件名为“**report.pdf**”，要求逻辑清晰、条理分明，内容包括每个任务的完成思路、操作步骤、必要的中间过程、任务的结果及分析。

3. 附件提交

3.1 如使用编程实现，将任务 1、2、3、4 的源程序分别保存到“program1”，“program2”，“program3”，“program4”文件夹，然后存放到“**program**”文件夹中。

3.2 将任务 1、2、3、4 所得到的结果文件存放到“**result**”文件夹中。

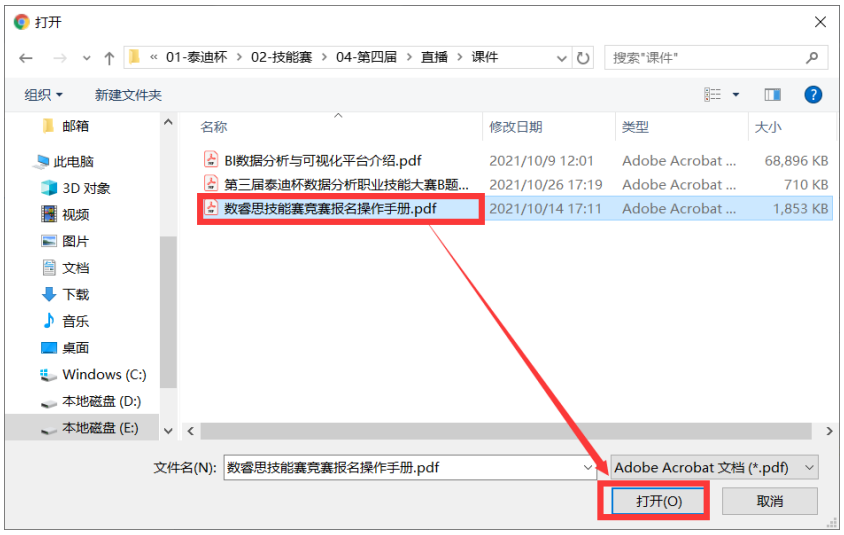
3.3 将程序文件夹“**program**”、结果文件夹“**result**”以及报告的 **word** 版本打包成“**appendix.zip**”，作为附件提交。

4. 提交界面

4.1 找到赛题提交入口。

4.2　点击“点击上传”按钮。

4.3　选择需要上传的对应文件，点击“打开”。



4.4　进度条加载完成后会有“上传成功”提示。

4.5　页面如下图即为上传提交成功，多次提交会以最后一次为准。