肥料登记数据分析

摘要

中国的种植粮食的土地是有限的，而肥料的管理与应用对我国的粮食安全来说，至关重要。尽管粮食作物的生长需要磷、钾和氮等元素，但是一旦这些元素过量就会对土地造成伤害。因此，肥料的生产销售必须遵循我国的《肥料登记管理办法》，依法在农业行政管理部门进行登记，而各地方人民政府也需要做好带头作用，从而更好地为我国的粮食安全做出贡献。

一、问题重述

题目要求我们根据给出的数据，完成以下的大目标：

1. 对肥料登记文件的数据进行预清洗与预处理。
2. 根据养分的百分比对文件当中的肥料产品进行分类。
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。

二、解决任务的步骤

2.1 任务一

2.1.1

题目要求将肥料分为复混肥料（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 4 种类别。

2.1.2

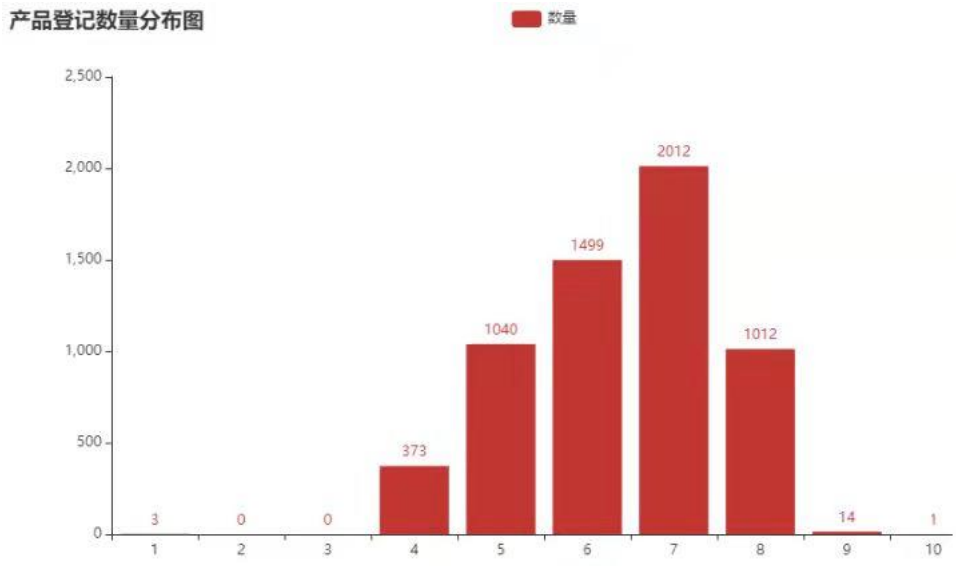
计算附件 1 中各肥料产品的氮、磷、钾养分百分比之和，结果保留 3 位小数（例如 1.0%，即 0.010）。

2.2 任务二

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。绘制产品登记数量的直方图，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。

图：



图：复混肥料产品登记数量直方图

分析复混肥料产品的分布特点：

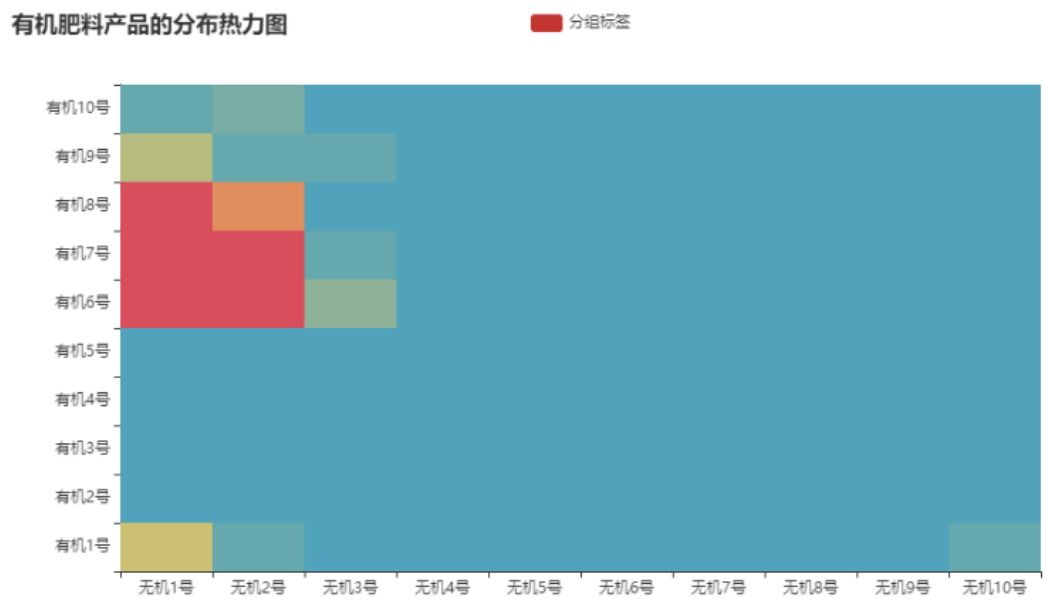
直方图(Histogram)又称质量分布图，是一种统计报告图，由一系列高度不等的纵向条纹或线段表示数据分布的情况。 一般用横轴表示数据类型，纵轴表示分布情况。通过直方图，用户可以很直观的看出数据分布的形状、中心位置以及数据的离散程度等。各组产品登记数量存在较大的差异，第 1,2,3,9,10 组产品登记数量非常少，其中 2 组和 3 组的产品登记数量为 0，而产品登记数量最大的是第 7 组，该组有 2012 个产品登记数量。

表：复混肥料数量最大的前 3 个分组及相应的产品登记数量

排名	一	二	三
分组标签	7	6	5
产品登记数量	2012	1499	1040

任务 2.2

以横轴为总无机养分分组，纵轴为有机质分组绘制有机肥料产品热力图



图：有机肥料产品分布热力图

任务 2.2.3

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。

我们使用的聚类算法是 K-means，k-means 算法又称 k 均值算法,K-means 算法中的 k 表示的是聚类为 k 个簇，means 代表取每一个聚类中数据值的均值作为该簇的中心。

K-means 聚类的算法思想大致为：先从样本集中随机选取 k 个样本作为簇中心，并计算所有样本与这 k 个“簇中心”的距离，对于每一个样本，将其划分到与其距离最近的“簇中心”所在的簇中，对于新的簇计算各个簇的新的“簇中心”[1]

K-means 聚类的算法流程[2]：

2.3 任务三肥料产品的多维度对比分析

2.3.1 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。