

# 肥料登记数据分析

## 摘要

肥料是农业生产中一种重要的生产资料,可以为多种植物提供他们所需要的营养元素,并且可以改善土壤的质量、提高土壤肥力的一类物质,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,肥料质量的优劣直接关系到农业能否增收,农产品质量是否安全,并且直接影响着农业生态环境的安全,对增加农民收入起着重要作用,加强对肥料管理非常必要,因此,利用数据分析技术对肥料登记进行研究具有重大意义。

**关键词:** K-means 聚类, python

## **Abstract**

Fertilizer is an important means of production in agricultural production. It can provide nutrients for a variety of plants, improve soil quality and improve soil fertility. Its production and sales must comply with the measures for the administration of fertilizer registration and be registered in the agricultural administrative department according to law. The agricultural administrative departments of the people's governments of all provinces, autonomous regions and municipalities directly under the central government are mainly responsible for the registration of fertilizers sold in their respective administrative regions. The quality of fertilizers is directly related to whether agriculture can increase income and whether the quality of agricultural products is safe, and directly affects the safety of agricultural ecological environment. It plays an important role in increasing farmers' income. Therefore, it is very necessary to strengthen fertilizer management, Using data analysis technology to study fertilizer registration is of great significance.

# 目录

|  |   |
|--|---|
| 问题重述.....                                  | 4 |
| 问题背景.....                                  | 4 |
| 问题分析.....                                  | 4 |
| 任务一：数据预处理.....                             | 5 |
| 任务 1.1 规范化处理 .....                         | 5 |
| 任务 1.2 计算总无机养分百分比 .....                    | 5 |
| 任务二：肥料产品的数据分析.....                         | 5 |
| 任务 2.1 对复混肥料产品分组标签并绘制直方图 .....             | 5 |
| 2.1.1 筛选产品分组标签 .....                       | 5 |
| 2.1.2 产品登记数量直方图 .....                      | 5 |
| 任务 2.2 对有机肥料产品分组标签并绘制热力图 .....             | 6 |
| 2.2.1 筛选产品分组标签 .....                       | 6 |
| 2.2.2 有机肥料产品分布热力图 .....                    | 6 |
| 任务 2.3 对复混肥料产品分类标签并绘制三维散点图和散点图矩阵、雷达图 ..... | 6 |
| 2.3.1K-means 聚类算法原理 .....                  | 6 |
| 任务三：肥料产品的多维度对比分析.....                      | 7 |
| 任务 3.1 可视化分析产品登记数量的变化趋势 .....              | 7 |

# 问题重述

## 问题背景

肥料是一种可以为多种植物提供他们所需要的营养元素,是农业生产中一种重要的生产资料,并且可以改善土壤的质量、提高土壤肥力的一类物质.合理施用肥料能够使农作物生长态势良好,品质优良,农民能够获得大丰收.但是,若是不合理使用肥料,如使用过少或使用过多,都不能达到理想的效果,一定会在不同程度上影响农产品的质量与品质.农业增产的核心问题是粮食产量的增加,而肥料又是粮食产量增加的重要因素。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,肥料质量的优劣直接关系到农业能否增收,农产品质量是否安全,并且直接影响着农业生态环境的安全,对增加农民收入起着重要作用,加强对肥料管理非常必要,在登记管理等多个环节仍存在许多需要解决的问题,需要进一步加强肥料监管,解决肥料问题,因此,利用数据分析技术对肥料登记进行研究具有重大意义。

## 问题分析

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

# 任务一：数据预处理

## 任务 1.1 规范化处理

## 任务 1.2 计算总无机养分百分比

# 任务二：肥料产品的数据分析

## 任务 2.1 对复混肥料产品分组标签并绘制直方图

### 2.1.1 筛选产品分组标签

### 2.1.2 产品登记数量直方图

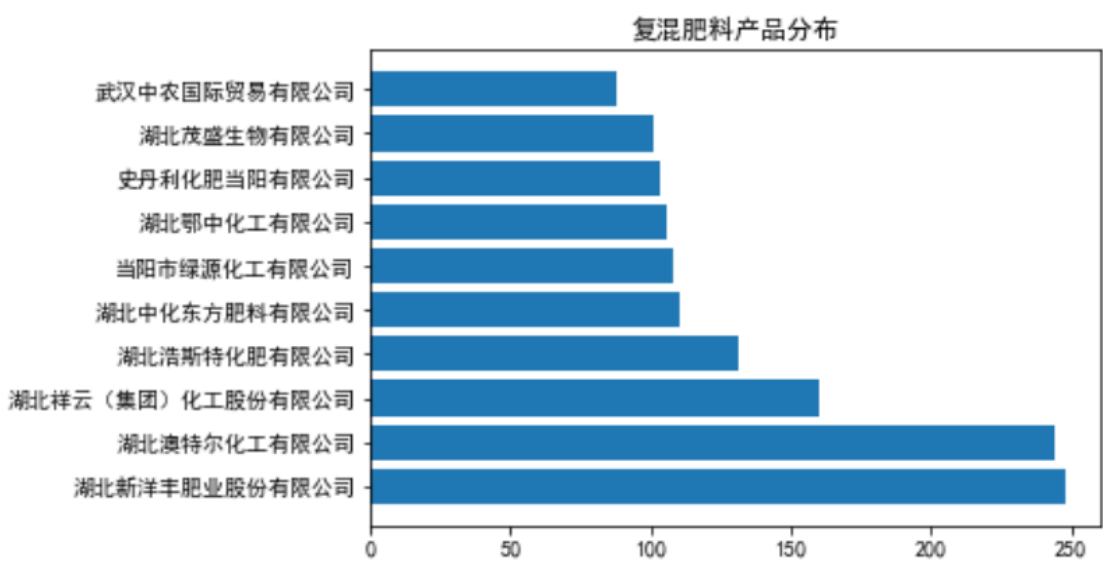


图 12 复混肥料产品登记数量前十名的企业

|        |      |      |      |
|--------|------|------|------|
| 排名     | 一    | 二    | 三    |
| 分组标签   | 7    | 6    | 5    |
| 产品登记数量 | 2012 | 1501 | 1038 |

## 任务 2.2 对有机肥料产品分组标签并绘制热力图

### 2.2.1 筛选产品分组标签

### 2.2.2 有机肥料产品分布热力图

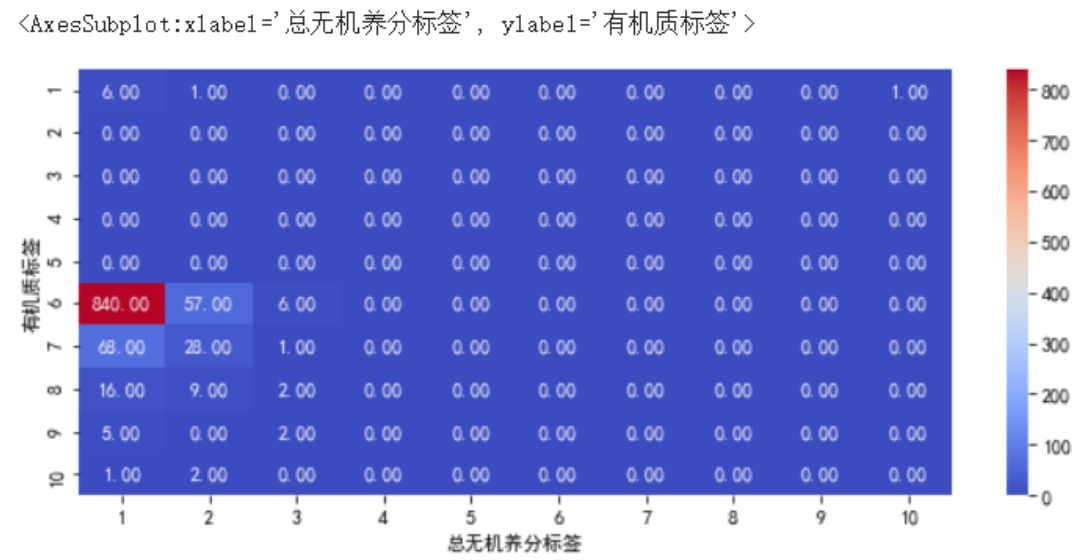


图 20 有机肥料产品的分布热力图

用 seaborn 库里的 heatmap 绘制热力图

分析：由上图分析可知，有机质标签为 6 和无机养分标签为 1 的有机肥料产品比较多，也看得出有机肥料产品中虽然有机质含量较高，但无机养分也相对比较少。

## 任务 2.3 对复混肥料产品分类标签并绘制三维散点图和散点图矩阵、雷达图

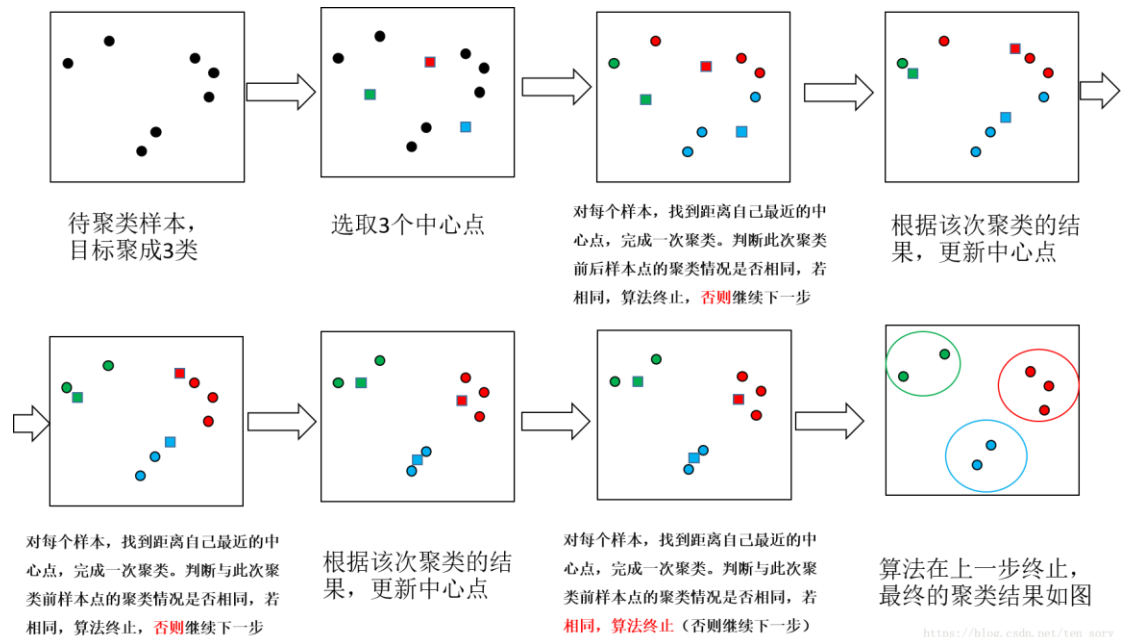
### 2.3.1K-means 聚类算法原理

K-means 算法是一种聚类算法，其中 K 表示类别数，Means 表示均值。K-Means 是一种通过均值对数据点进行聚类的算法。k-平均聚类的目的是：把 n 个点划分到 k 个聚类中，使得每个点都属于离他最近的均值（此即聚类中心）对应的聚类，以之作为聚类的标准。K-Means 算法通过预先设定的 K 值及每个类别的初始质心对相似的数据点进行划分。并通过划分后的均值迭代优化获得最优的聚类结果，算法的输入为一个样本集（或者称为点集），通过该算法可以将样本进行聚类，具有相似特征的样本聚为一类。

针对每个点，计算这个点距离所有中心点最近的那个中心点，然后将这个点归为这个中心点代表的簇。一次迭代结束之后，针对每个簇类，重新计算中心点，

然后针对每个点，重新寻找距离自己最近的中心点。如此循环，直到前后两次迭代的簇类没有变化。

K-means 算法的过程：如下图所示，目标是将样本点聚类成 3 个类别



已知观测集，其中每个观测都是一个  $d$ -维实向量， $k$ -平均聚类要把这  $n$  个观测划分到  $k$  个集合中( $k \leq n$ ),使得组内平方和最小。换句话说，它的目标是找到使得下式满足的聚类 $S_i$

簇：所有数据的点集合，簇中的对象是相似的。

质心：簇中所有点的中心（计算所有点的均值而来）。

SSE: Sum of Squared Error（误差平方和），它被用来评估模型的好坏，SSE 值越小，表示越接近它们的质心。聚类效果越好。由于对误差取了平方，因此更加注重那些远离中心的点（一般为边界点或离群点）

k-means 聚类基本的步骤为：

step1: 选定要聚类的类别数目  $k$ （如上例的  $k=3$  类），选择  $k$  个中心点。

step2: 针对每个样本点，找到距离其最近的中心点（寻找组织），距离同一中心点最近的点为一个类，这样完成了一次聚类。

step3: 判断聚类前后的样本点的类别情况是否相同，如果相同，则算法终止，否则进入 step4。

step4: 针对每个类别中的样本点，计算这些样本点的中心点，当做该类的新的中心点，继续 step2。

## 任务三：肥料产品的多维度对比分析

### 任务 3.1 可视化分析产品登记数量的变化趋势