肥料登记数据分析

摘要

肥料是农业生产中的一种重要生产资料，肥料是作物的粮食，是增产的物质基础，我国近年来的土壤肥力监测结果表明，肥料对农产品产量的贡献率，全国平均为57.8%。中国以占世界7%的耕地养活占世界22%的人口，应该说一半归功于肥料的作用。但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。本文主要是对附件所给的数据进行肥料进行数据分析。

关键词：数据预处理、数据多维度分析、数据可视化

目录

一、任务分析

1.1 背景介绍

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。

1.2 目标分析

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。

1.3 软件与技术介绍

大数据（big data）或称巨量资料，是指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。大数据有大量（Volume）、高速（Velocity）、多样（Variety）、低价值密度（Value）、真实性（Veracity）五大特点。它并没有统计学的抽样方法，只是观察和追踪发生的事情。大数据的用法倾向于预测分析、用户行为分析或某些其他高级数据分析方法的使用。

Python 为数据分析提供诸多工具,anaconda 是其中一个著名的科学计算发行版，包括近 200 多个工具包,常见的有 numpy，scipy，pandas，ipython，matplotlib，sklearn 等。Python 数据分析流程如图 1 所示。

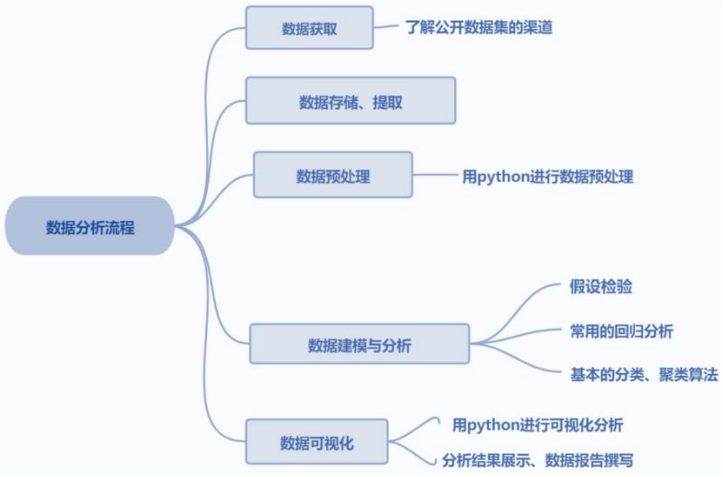

图 1 数据分析流程图

二、任务一 数据预处理

2.1 规范数据

未经过处理的附件 1 如图 2 所示，看起来很杂乱。我们要使用一些处理让他看起来更加的清楚明了。

2.2 计算总无机养分百分比

三、任务二 肥料产品的数据分析

3.1 复混肥料分析

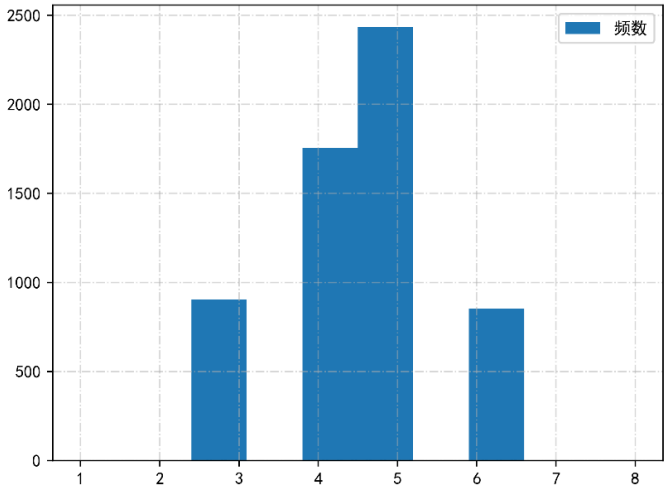

图 10 登记数量直方图

由表 4 可以看出登记数量最大的三个组依次是 5 组、4 组和 3 组。所以登记数量最大的前 3 个小组的相关数据如表 4 所示。

表 4 登记数量前 3 组数据表

排名	一	二	三
分组标签	5	4	3
产品登记数量	2435	1756	904

3.2 有机肥料分析

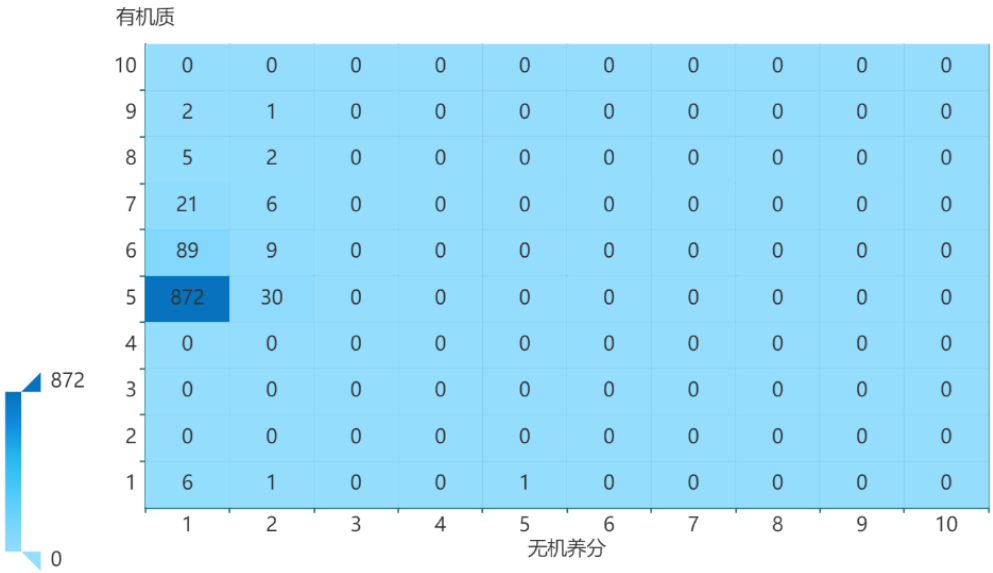

图 11 有机肥料分布热力图

3.3 复混肥料的中氮、磷、钾的养分百分比分析

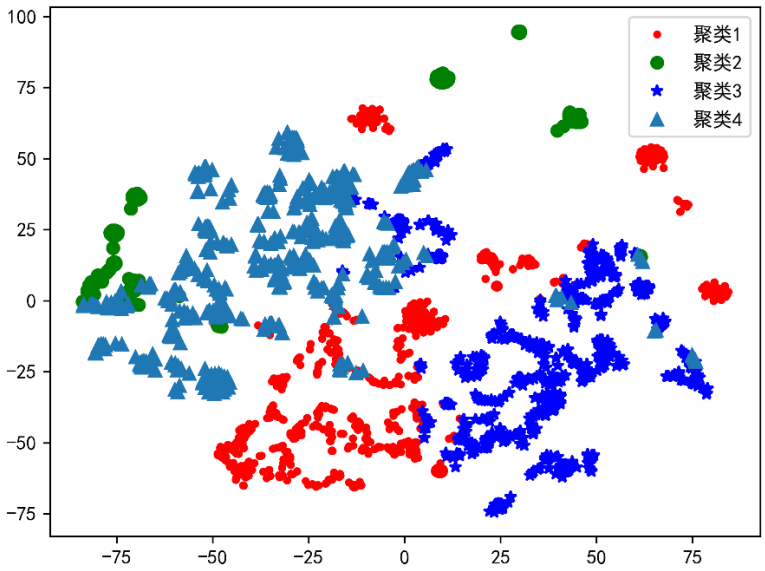

图 13 散点图

3.4 另一种聚类算法——K-Means。

k 均值聚类算法（k-means clustering algorithm）是一种迭代求解的聚类分析算法。其步骤是，预将数据分为 K 组，则随机选取 K 个对象作为初始的聚类中心，然后计算每个对象与各个种子聚类中心之间的距离，把每个对象分配给距离它最近的聚类中心。聚类中心以及分配给它们的对象就代表一个聚类。每分配一个样本，聚类的聚类中心会根据聚类中现有的对象被重新计算。这个过程将不断重复直到满足某个终止条件。终止条件可以是没有（或最小数目）对象被重新分配给不同的聚类，没有（或最小数目）聚类中心再发生变化，误差平方和局部最小。

用数据表达式表示，假设簇划分为(C1,C2,...Ck)(C1,C2,...Ck)，则我们的目标是最小化平方误差 E。如公式（2）所示。

其中 μiμi 是簇 CiCi 的均值向量，有时也称为质心，表达式如公式（3）所示。

四、任务三 肥料产品的多维度对比分析

4.1 不同年份产登记数量分析

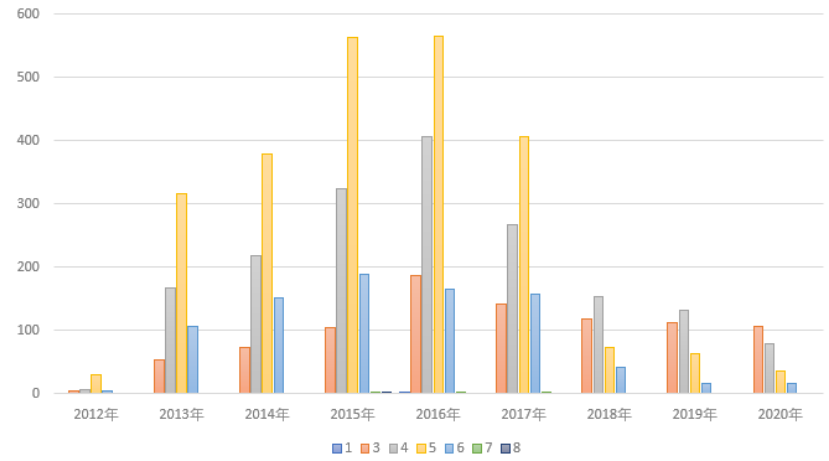

图 19 各年度各组别登记数量对比图

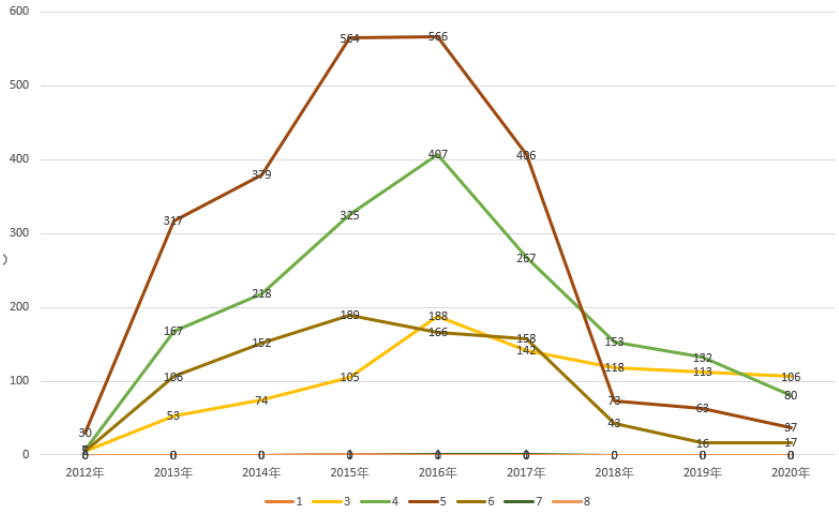

图 20 各年度各组别登记数量变化趋势图

4.2 广西与湖北的登记数量分布

4.3 杰卡德相似系数矩阵

Jaccard 相似指数用来度量两个集合之间的相似性，它被定义为两个集合交集的元素个数除以并集的元素个数。如公式（4）。

Jaccard 距离用来度量两个集合之间的差异性，它是 Jaccard 的相似系数的补集，被定义为 1 减去 Jaccard 相似系数。如公式（5）

五、任务四 肥料产品原料与养分比的多维度对比分析

5.1 养分百分比分析

参考文献

[1]刘勘, 周晓峥, 周洞汝. 数据可视化的研究与发展[J]. 计算机工程, 2002, 28(8):3.
[2]黄志澄. 给数据以形象 给信息以智能 数据可视化技术及其应用展望[J]. 电子展望与决策, 1999(6):3-9.
[3]李莉. 大数据下的智能数据分析技术研究[J]. 电脑迷, 2018, 000(016):193.
[4]彭燕妮, 樊晓平, 赵颖,等. 时间事件序列数据可视化综述[J]. 计算机辅助设计与图形学学报, 2019(10):1698-1710.
[5]汪飞, 李强, 左伍衡. Visual Search User Interfaces[J]. 计算机辅助设计与图形学学报, 2014, 026(005):708-716.
[6]朱建平, 章贵军, 刘晓葳. 大数据时代下数据分析理念的辨析[J]. 统计研究, 2014, 031(002):10-19.
[7]曾悠. 大数据时代背景下的数据可视化概念研究[D]. 浙江大学, 2014.
[8]王永周, 邓燕. 基于大数据预测的消费者购买决策行为分析[J]. 商业经济研究, 2016(23):3.
[9]何强, 董志勇. 利用互联网大数据预测季度 GDP 增速的方法研究[J]. 统计研究, 2021, 37(12):14.
[10]何强, 董志勇. 利用互联网大数据预测季度 GDP 增速的方法研究[J]. 统计研究, 2021, 37(12):14