参赛报告

作品名称 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，其生产销售必须遵守《肥料登记管理办法》，依法在农业部门进行登记。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。因此，利用数据分析技术对肥料产品的研究工作有重大的意义。

ABSTRACT

Fertilizer is an important means of production in agricultural production. Its production and sales must comply with the measures for the administration of fertilizer registration and be registered in the agricultural department according to law. The competent 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 within their respective administrative regions. Therefore, the use of data analysis technology is of great significance to the research of fertilizer products.

For task one, to preprocess a given data set, first normalize the classification name with Python, and then classify it. The contents of nitrogen, phosphorus and potassium are added and a new term is generated, which is the percentage of total inorganic nutrients. Complete all task requirements.

For task two, firstly, the pandas Library in Python is used to process the data in Annex 2. For task 2.1, first group and label the data, and draw the histogram for visualization according to the product distribution characteristics. For task 2.2, the operation of grouping and labeling is also carried out, and the data visualization is realized by drawing the thermal diagram. The distribution characteristics of correlation were analyzed. For task 2.3, use clustering algorithm to classify products, use numpy Library in Python, use disteclud() function to calculate Euclidean distance, and use randcenter() function to build a set containing K random centroids for a given data set. Each product is labeled with cluster classification, and the data is visualized by drawing relevant charts. All requirements were finally completed.

For task three, according to the year in the certificate issuing date, the change trend of product registration quantity of each group of compound fertilizer in different years is analyzed. Extract the data of organic fertilizer products valid after September 30, 2021, and screen out the top five groups of products registered in Hubei and Guangxi. If the registered quantity of extracted products is greater than 10, the raw materials shall be given. Taking the raw materials used by each enterprise as the eigenvalue, the jacquard coefficient matrix is calculated. Through the above three dimensions to compare and analyze.

For task 4, first clean the data. The columns' technical index 'and' raw material and proportion 'need to be split by using the split() function. The relevant data such as the percentage of nitrogen, phosphorus, potassium nutrients and organic matter and the nitrogen content of fertilizer were extracted, and a new data set was generated.

Key words: clustering algorithm Python numpy kujkaard similarity coefficient matrix

目　　录

1 分析目标

（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

2 任务一 数据预处理

2.1 数据的规范化处理

在本次数据分析过程中，主要存在产品名称不规范的情况。主要为产品类别的分类，需要将“掺混肥料”归入“复混肥料”这一分类。再按照四种肥料类别进行分类。通过 python 运用 pandas 库对数据进行操作，从而得出 result1_1。

2.2 统计分析

对个肥料产品中的氮、磷、钾养分百分比进行统计求和，并输出为总无机养分百分比。通过 python 先对数据进行了解，之后通过 Excel 对 N,P,K 养分百分比进行计算，公式如下：

总氮百分比+P205 百分比*0.436+K20 百分比*0.827

图 2、3 数据表数值型数据相关信息展示

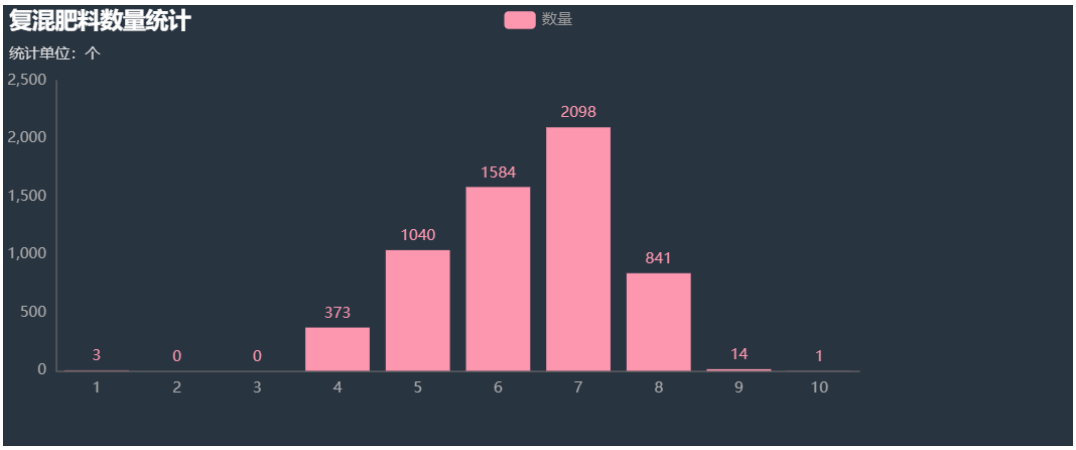
	序号	总氮百分比	P2O5百分比	K2O百分比	有机质百分比
count	2925.000000	2925.000000	2925.000000	2925.000000	2925.000000
mean	1463.000000	0.191366	0.108651	0.111827	0.037617
std	844.519094	0.069265	0.056421	0.060933	0.117844
min	1.000000	0.016700	0.000000	0.000000	0.000000
25%	732.000000	0.150000	0.060000	0.060000	0.000000
50%	1463.000000	0.190000	0.100000	0.100000	0.000000
75%	2194.000000	0.250000	0.150000	0.160000	0.000000
max	2925.000000	0.400000	0.500000	0.400000	0.700000

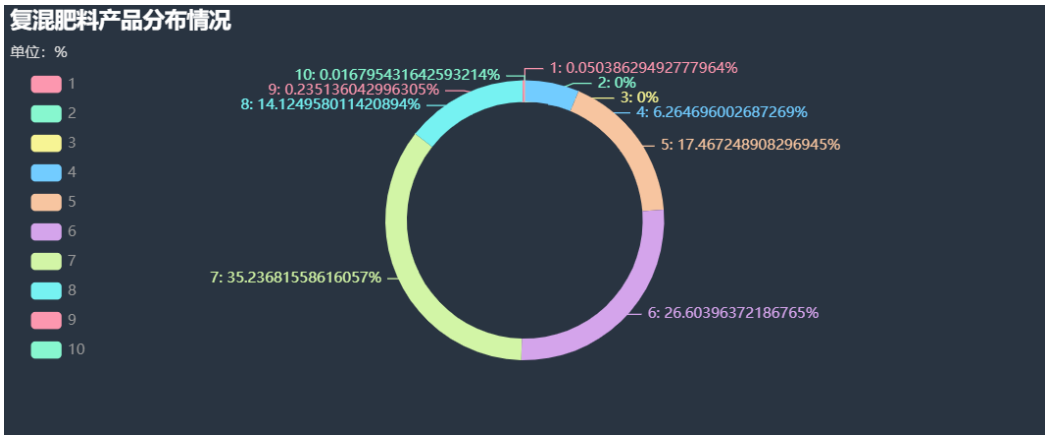
	序号	总氮百分比	P2O5百分比	K2O百分比	有机质百分比	总无机养分百分比
count	7619.000000	7619.000000	7619.000000	7619.000000	7619.000000	7619.000000
mean	2230.627379	0.162622	0.089254	0.102964	0.078918	0.354840
std	1530.790543	0.075677	0.058090	0.065122	0.166392	0.144362
min	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	953.000000	0.130000	0.050000	0.050000	0.000000	0.300000
50%	1905.000000	0.170000	0.080000	0.090000	0.000000	0.400000
75%	3479.500000	0.210000	0.120000	0.150000	0.000000	0.450000
max	5384.000000	0.390000	0.520000	0.460000	0.900000	0.720000

3 任务二 肥料产品分析

3.1 复混肥料产品分布分析

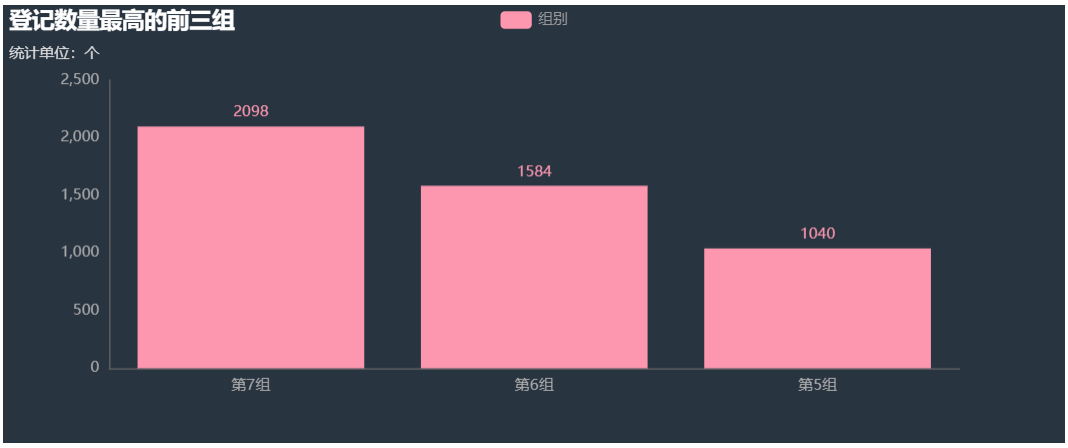
利用 Python 中 pandas 库的 cut 方法对数据打上分组标签，通过绘制直方图以及环形图直观展示产品的分布特点。





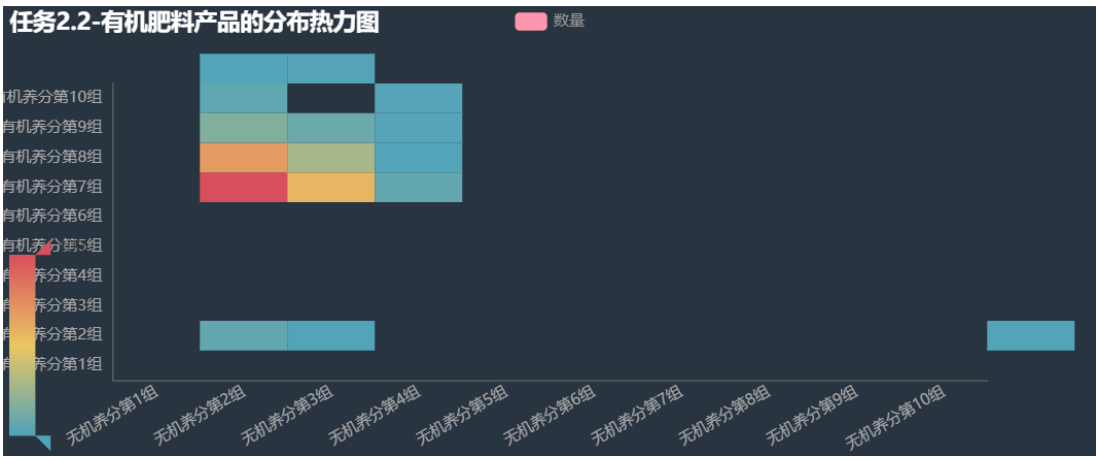
登记数量最大的前三个分组为第 7、6、5 组，相应的产品的登记数量分别

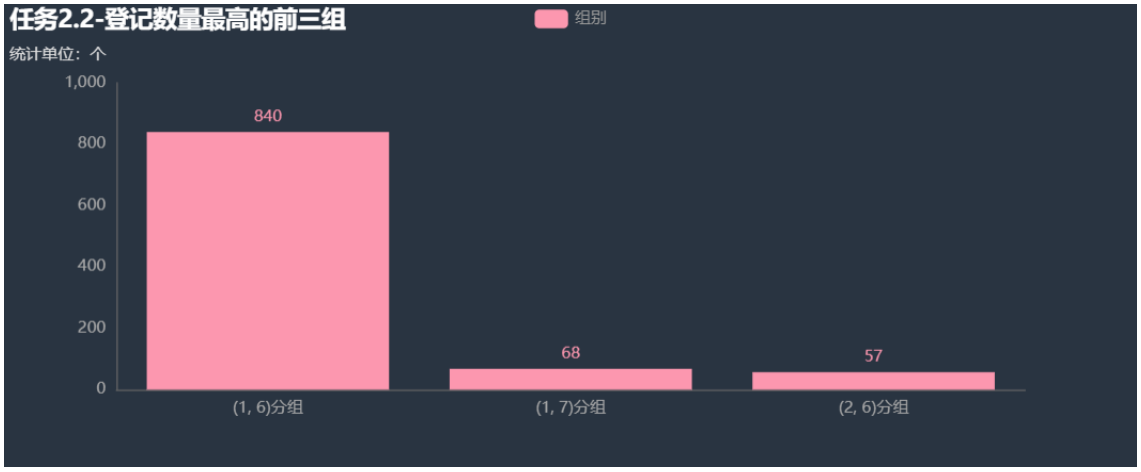
为 2098、1584、1040。



排名	一	二	三
分组标签	7	6	5
产品登记数量	2098	1584	1040

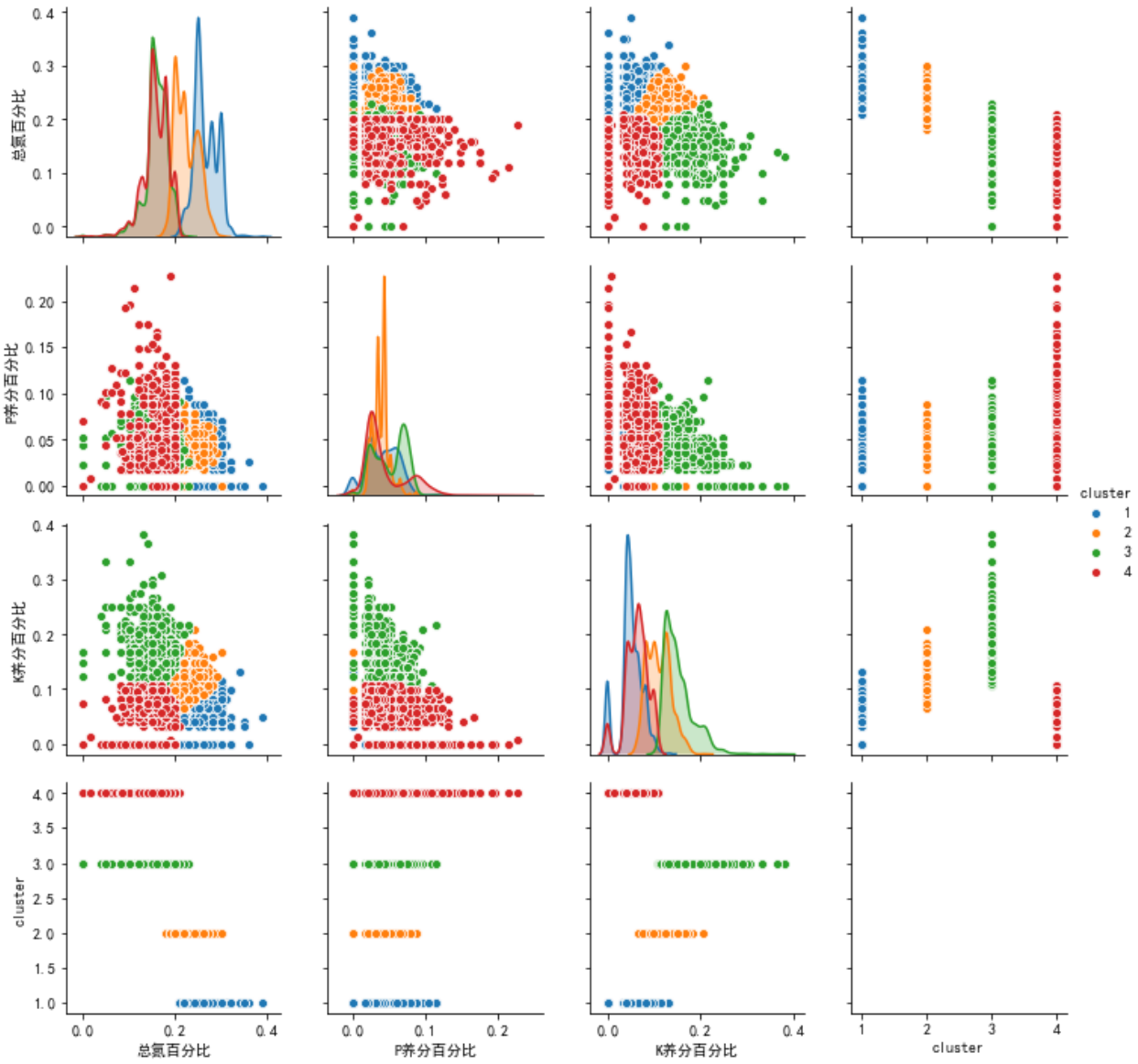
3.2 有机肥料产品分布特点



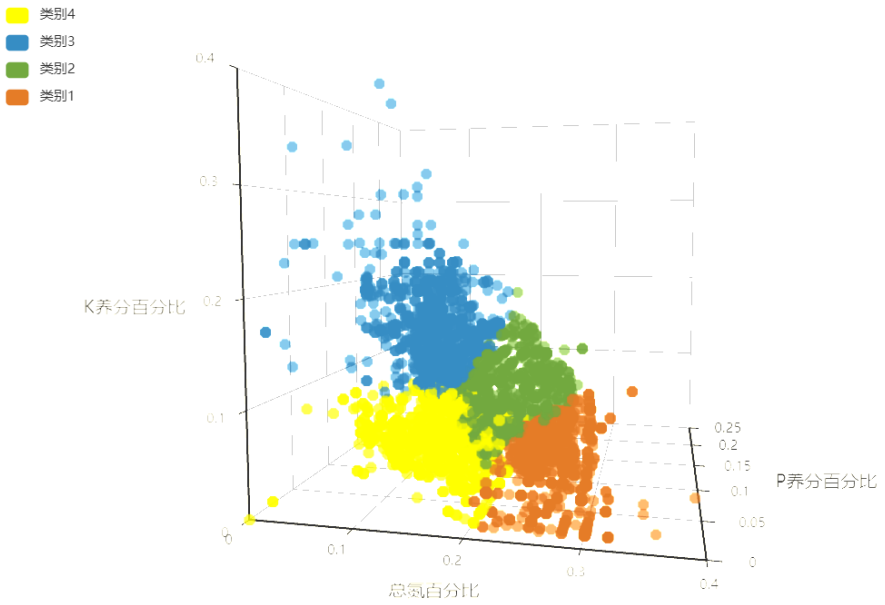


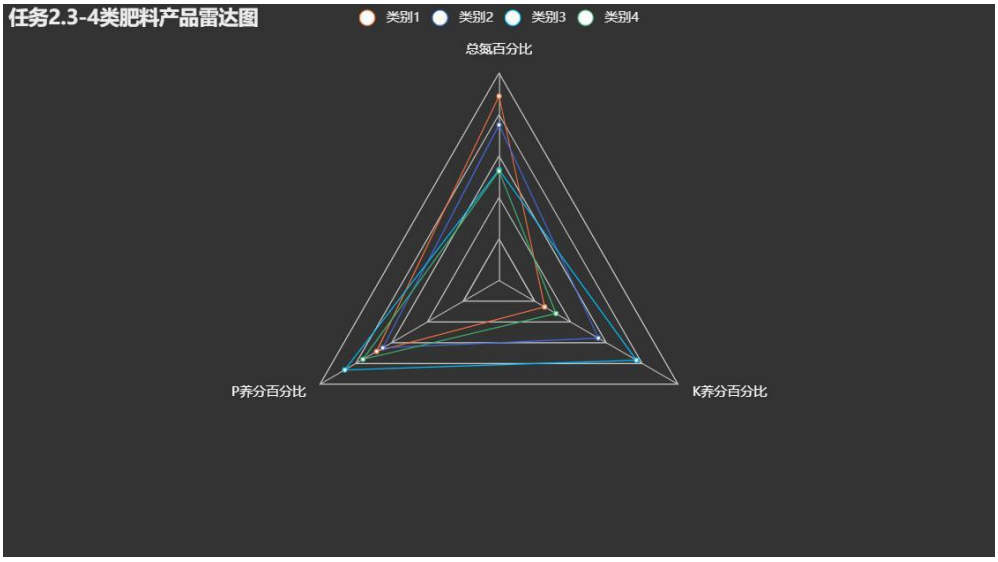
3.3 聚类算法

图 5 四类肥料散点图矩阵



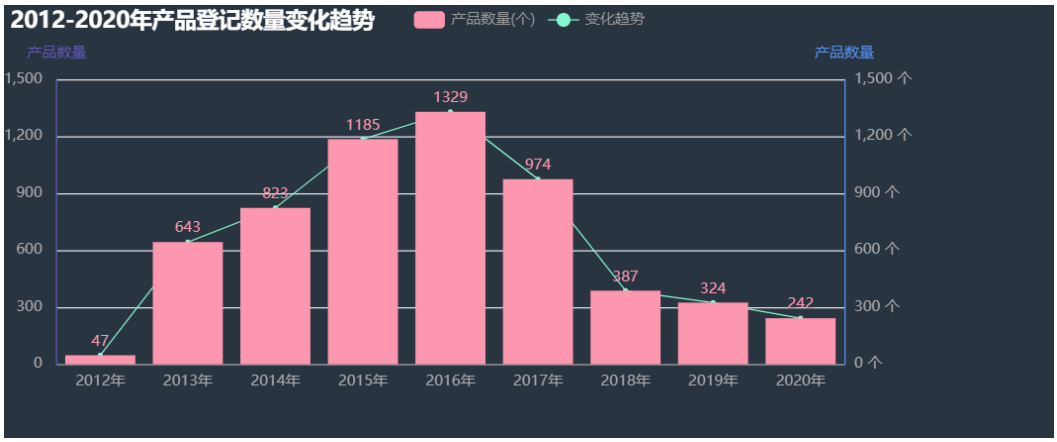
任务2.3-肥料产品3D散点图

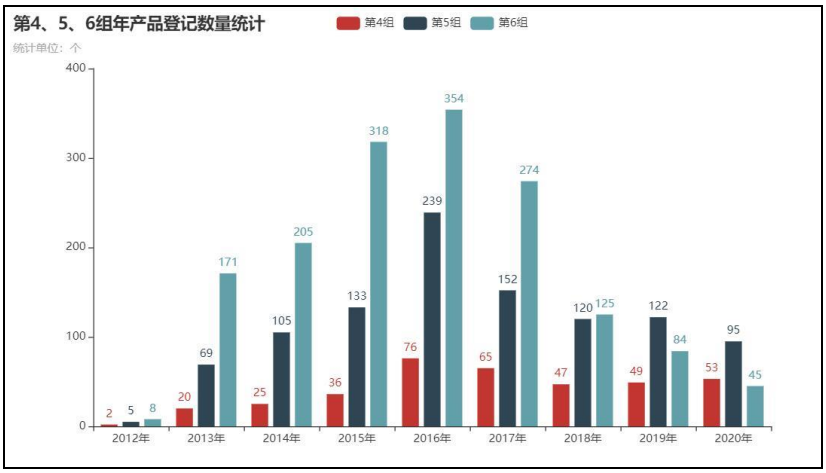


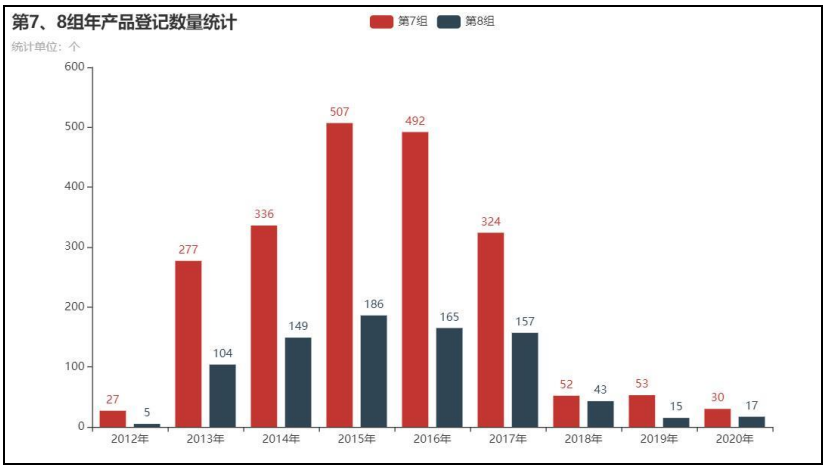


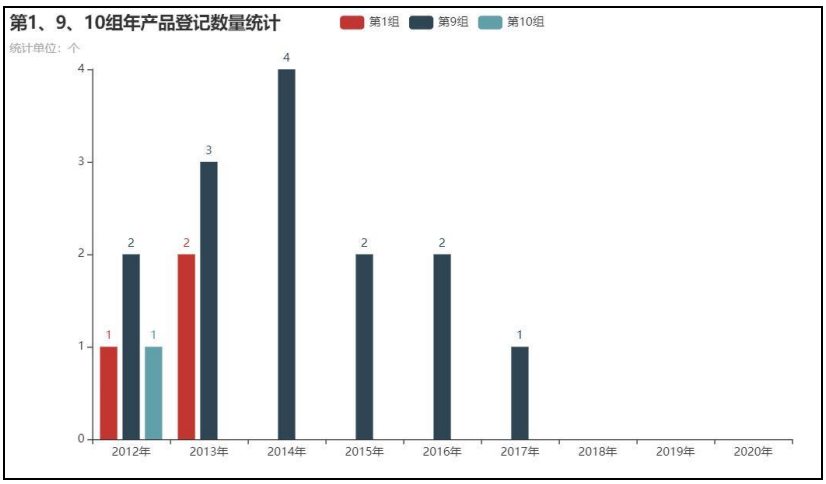
4 任务三 肥料产品的多维度对比分析

4.1 复混肥料数量变化趋势分析

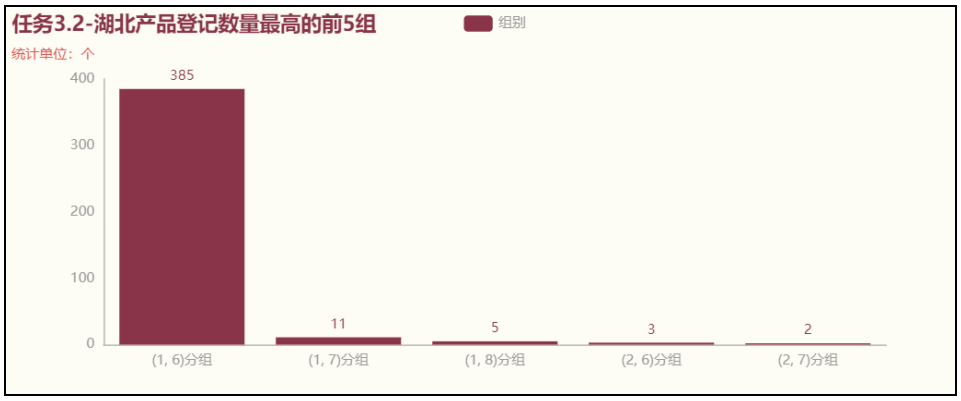


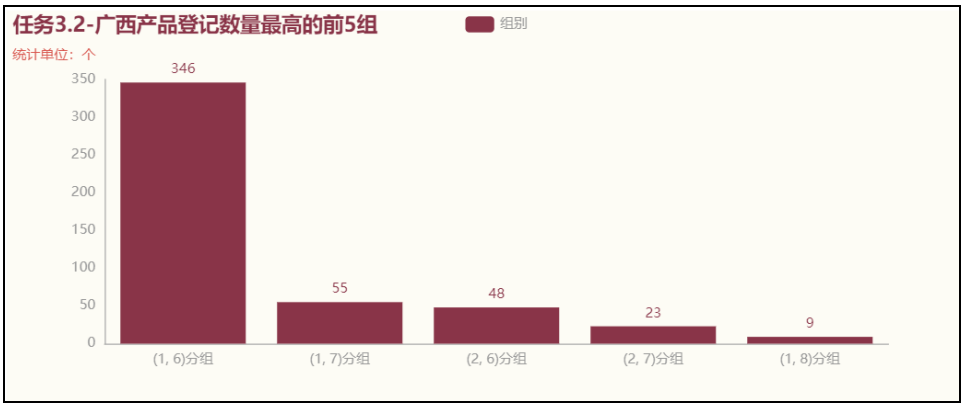






4.2 省份分布差异分析





4.3 原料分析

杰卡德相似系数说明：又称为 Jaccard 相似系数（Jaccard similarity coefficient）用于比较有限样本集之间的相似性与差异性。Jaccard 系数值越大，样本相似度越高。

给定两个集合 A,B，Jaccard 系数定义为 A 与 B 交集的大小与 A 与 B 并集的大小的比值，定义如下：

对附件 3 进行相关数据提取操作后，以各企业所用到的原料作为特征进行相关系数的计算。可以看出相似系数矩阵是一个对称矩阵，并且看以看出 ID10 和 ID13、ID23 相似度比较高。

图 6 相似系数矩阵

1.0000	0.2500	0.2500	0.1765	0.2222	0.1667
0.2500	1.0000	0.5333	0.3846	0.5385	0.2667
0.2500	0.5333	1.0000	0.3529	0.4706	0.2632
0.1765	0.3846	0.3529	1.0000	0.4286	0.2667
0.2222	0.5385	0.4706	0.4286	1.0000	0.3125
0.1667	0.2667	0.2632	0.2667	0.3125	1.0000

5 任务四 肥料产品的多维度对比分析

6 参考文献

[1] Mark Lutz，《Python 学习手册（第 4 版）》.机械工业出版社，2011

[2] Peter Harrington，《Machine Learning in Action》.中国工信出版社，2013

[3] 姜启源，数学模型（第二版）.北京：高等教育出版社出版社，1993

[4] 范金城，梅长林，数据分析.科学出版社，2002

[5] 优化初始聚类中心的 K-means 聚类算法[J]. 郭永坤,章新友,刘莉萍,丁亮,牛晓录. 计算机工程与应用. 2020(15)

[6] 基于学科竞赛的统计学应用型人才培养研究[J]. 程从华,王静. 山西青年. 2020(17)

[7] 信息的计算机分析算法[J]. 马玉荣,朱勇. 实用心电学杂志. 1999(04)

[8] 基于二元与三元模型相结合的句法规则层次化分析算法[J]. 张海玲,邵玉斌,贾继康,龙华,杜庆治. 计算机工程与科学. 2021(07)