

**长春市企业状况分析报告**

**目录**

[1.摘要 3](#_Toc161509682)

[2.引言 3](#_Toc161509683)

[3.数据获取 3](#_Toc161509684)

[4.数据清洗 3](#_Toc161509685)

[4.1数据标准化 3](#_Toc161509686)

[5.数据探索 4](#_Toc161509687)

[5.1简单描述 4](#_Toc161509688)

[5.2简单统计 4](#_Toc161509689)

[5.3文本特征处理 16](#_Toc161509690)

[6.建模和分析 18](#_Toc161509691)

[6.1可视化结果分析 21](#_Toc161509692)

[7.结论与思考 23](#_Toc161509693)

# 摘要

使用的方法

TF-IDF，LDA主题模型，独热算法，KMeans聚类算法，PCA降维算法等

得出的结论

1.企业规模大都以小型为主

2.企业制度以有限责任公司为主，制度形式较为多元

3.公司分布较为集中，都集中在几个区域内

4.通过聚类分析发现，偏科技的大都创建时间较晚，所属区县较为集中，以小型为主。

# 引言

# 数据获取

通过互联网收集长春市内2655家科技型中小企业的数据，主要包括企业名字、成立时间、企业规模、企业简历、发展历程、企业类别、联系方式、主要产品和应用案例等信息，收集到的数据保存在csv文件中。

# 数据清洗

对收集到数据中大多数属性的值都是字符串，并且统一属性的取值有很多都是相同的，所以本研究对缺失值采用众数插入的方式进行清洗。

## 4.1 数据标准化

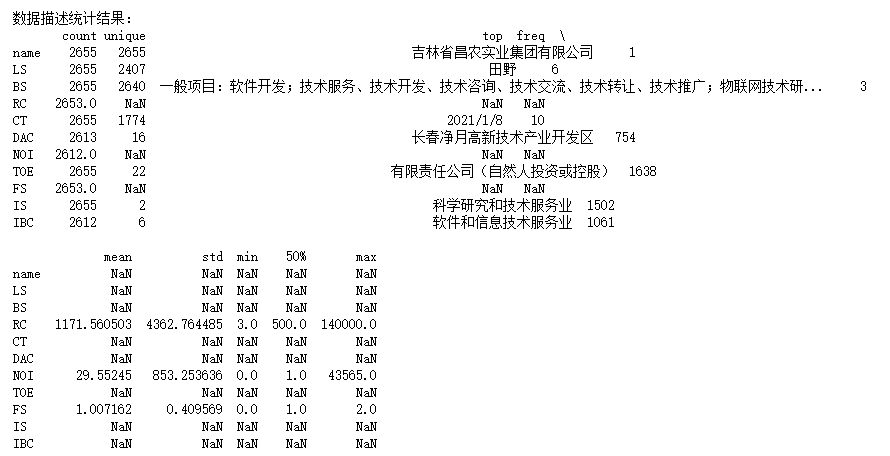
为了防止数据中不同属性的差异太大，本研究采用python中的Z-sore方法对数据进行标准化，主要处理内容如下所示：

1. 采用python中的z-score算法将数据进行标准化，对收集到的数据只保留经营范围cluster，创建时间CT，注册资本RC，所属区县DAC，企业类型TOE，企业规模FS和国标行业大类IBC。
2. 将企业创建时间CT从字符串类型转换成数值类型；
3. 采用独热编码处理分类数据，并生成映射关系表。

# 数据探索

## 5.1 简单描述

使用python的describe方法对数据进行简单的统计，并输统计结果，包括平均值、方差、最小值、最大值、中位数等信息，如下图所示。

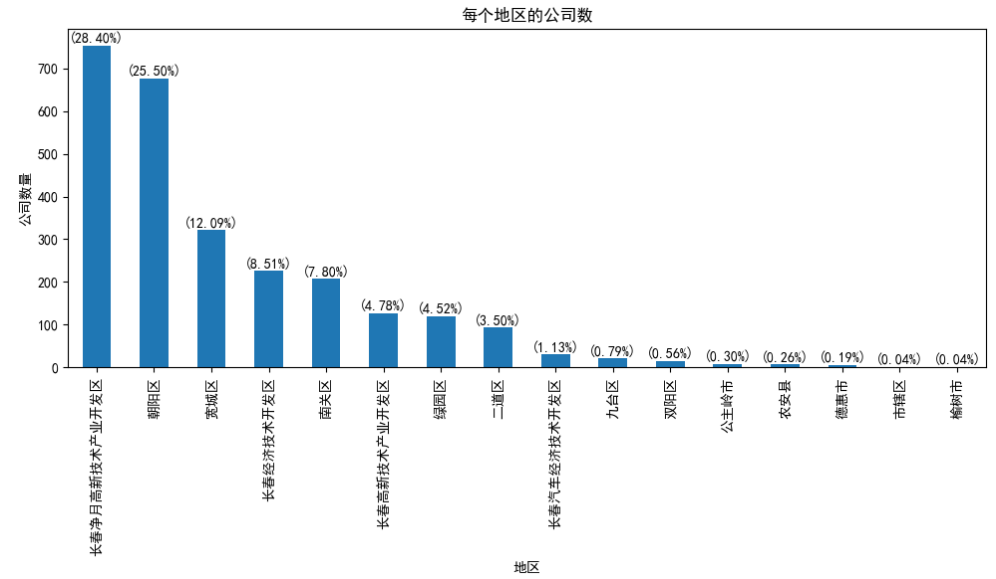


## 5.2 简单统计

为方便对数据进行分析，首先使用python对收集到的数据进行简单的统计分析，并对分析结果进行可视化，将可视化结果保存在image文件夹中。

1. **统计各个地区公司数量**

首先对长春市内每个地区的公司数量进行统计，结果如下图所示：

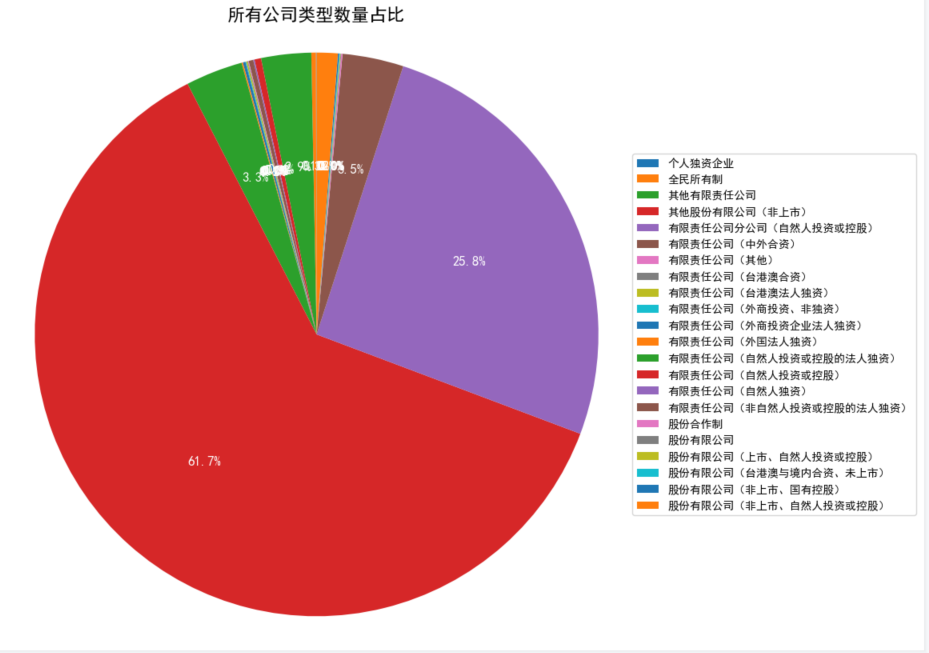


从上图可以看出，长春市内的公司主要集中在长春净月高新技术产业开发区和朝阳区，两个地区的公司占全部公司的一半以上，产生这样现象的原因可能和长春市内的政策和地理位置有关。

1. **统计各个地区公司类型**

接着对每个地区的公司类型进行分析，并对结果进行可视化。

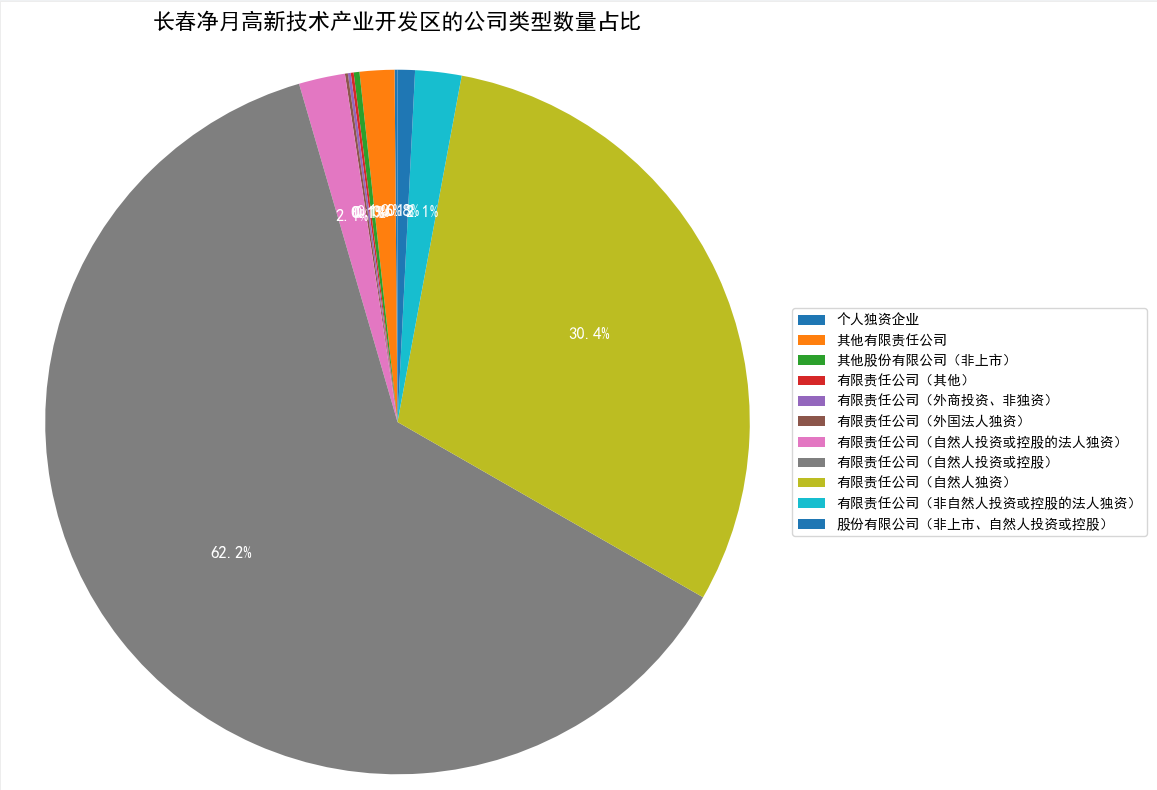
1. **所有公司类型数量及占比**



由上图可以看出，公司类型以有限责任公司为主，其中自然人投资或控股是最主要的方式，占比为61.7%，其次是自然人独资的方式，占比为25.8%。

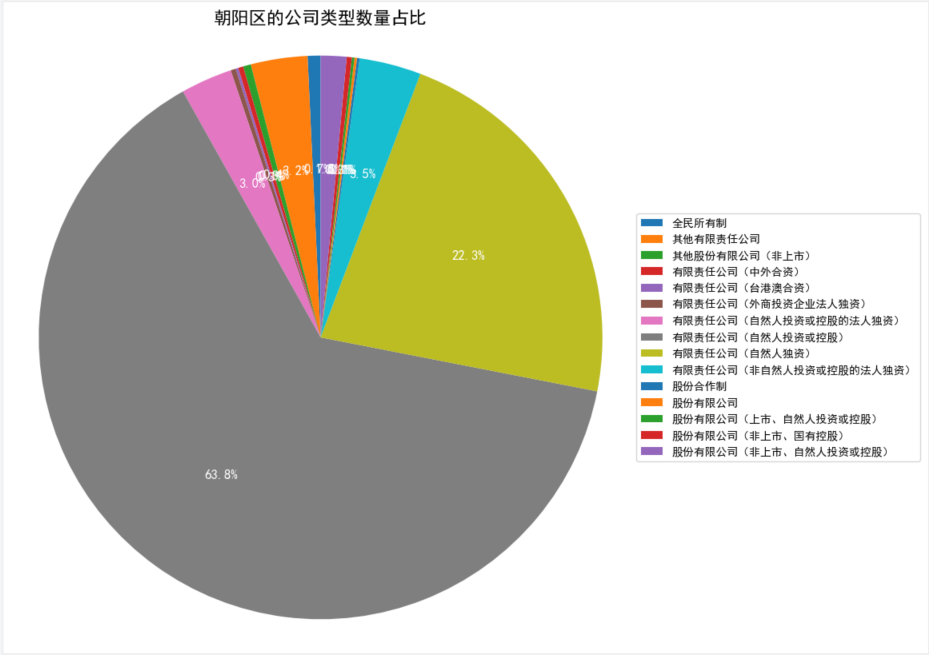
1. **长春净月高新技术产业开发区**

长春净月高新技术产业开发区的公司类型分布如下图所示：



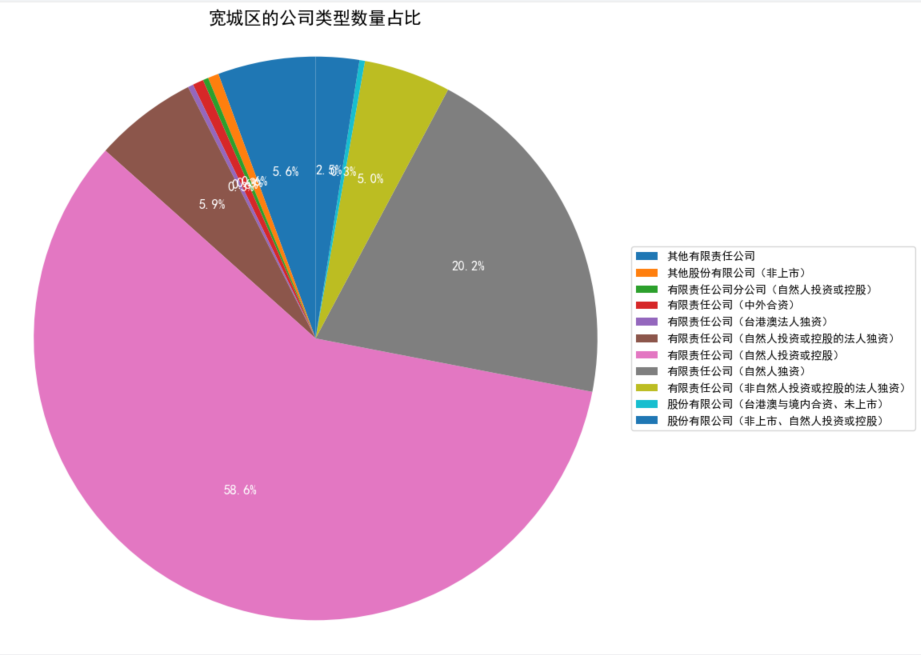
从上图可以看出，长春净月高新技术产业开发区的公司种类多样，其中“有限责任公司（自然人投资或控股）”占比高达62.2%，其次是“有限责任公司（自然人独资）”，占比达30.4%。

1. **朝阳区**



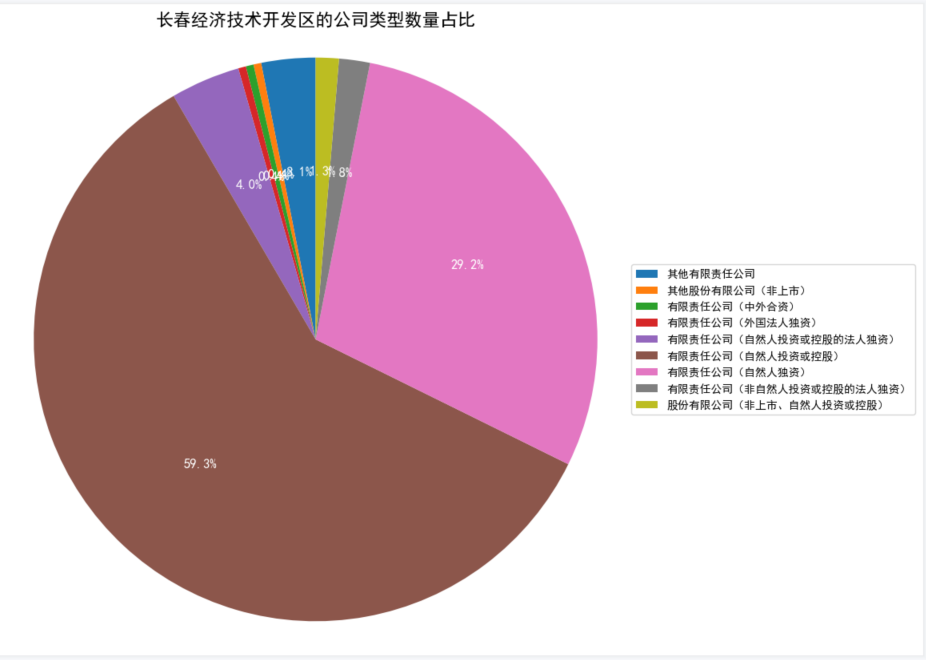
从上图可以看出，朝阳区的公司种类多样，其中“有限责任公司（自然人投资或控股）”占比高达63.8%，其次是“有限责任公司（自然人独资）”，占比达22.3%。

1. **宽城区**



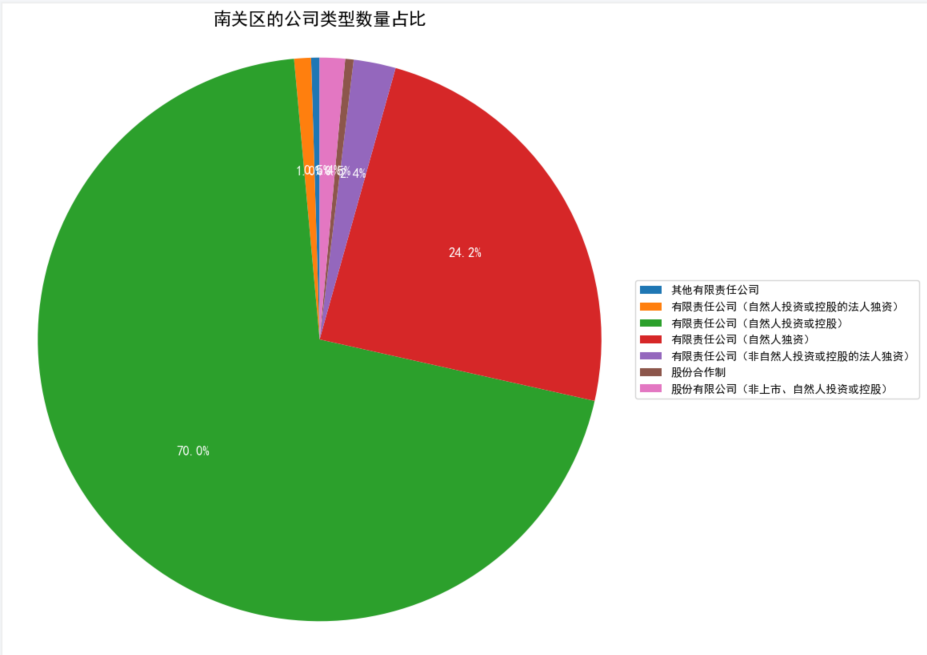
从上图可以看出，宽城区的公司种类多样，其中“有限责任公司（自然人投资或控股）”占比高达58.6%，其次是“有限责任公司（自然人独资）”，占比达20.2%。

1. **长春经济技术开发区**



从上图可以看出，长春经济技术开发区的公司种类多样，其中“有限责任公司（自然人投资或控股）”占比高达59.3%，其次是“有限责任公司（自然人独资）”，占比达29.2%。

1. **南关区**



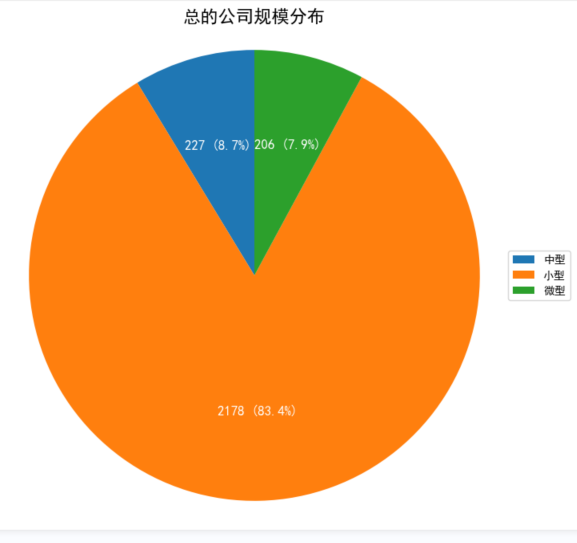
从上图可以看出，南关区的公司种类多样，其中“有限责任公司（自然人投资或控股）”占比高达70.0%，其次是“有限责任公司（自然人独资）”，占比达24.2%。

从可视化结果可以看出每个地区公司种类多样，公司数量越多，种类越是丰富，而且几乎所有地区的公司都以“有限责任公司（自然人投资或控股）”为主，大体上都以有限责任公司制度为最主要的公司制度。产生这种原因可能和公司制度的利弊有关系。

1. **统计各个地区公司规模**

接着对按照公司的规模对各个地区的公司规模进行统计，公司规模分为0-微型 1-小型 2-中型 3-大型4种类型。

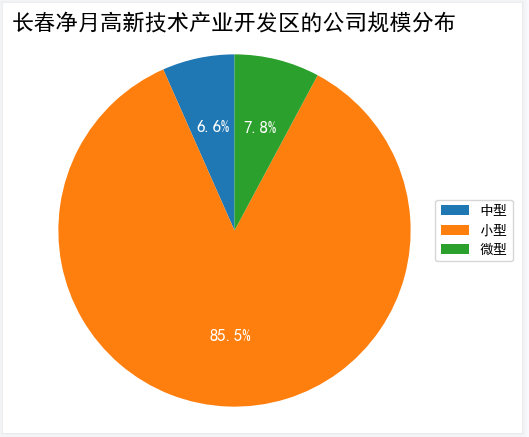
1. **总的公司规模分布**



从上图可以得知，各种公司大都以小型公司为主，占比83.4%，有2176家小型企业，其他规模企业都相对较小，微型有206家，占比7.9%，中型有227家，占比8.7%。

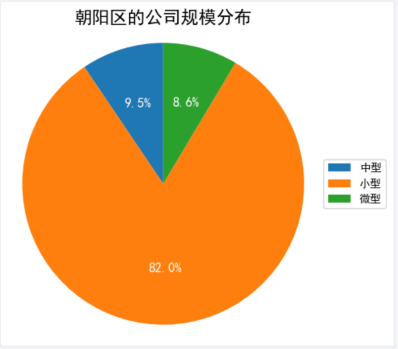
1. **长春净月高新技术产业开发区**

长春净月高新技术产业开发区公司规模如下图所示：



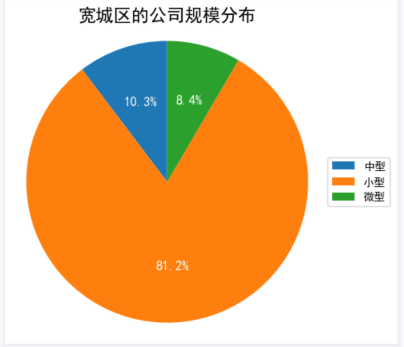
从上图可以看出，长春净月高新技术产业开发区的公司规模以小型为主，占比为85.5%，微型和中型都相对较少，分别占比7.8%和6.6%。

1. **朝阳区**

****

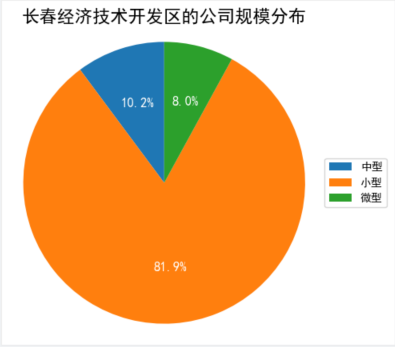
从上图可以看出，朝阳区的公司规模以小型为主，占比为82.0%，微型和中型都相对较少，分别占比8.6%和9.5%。

1. **宽城区**

****

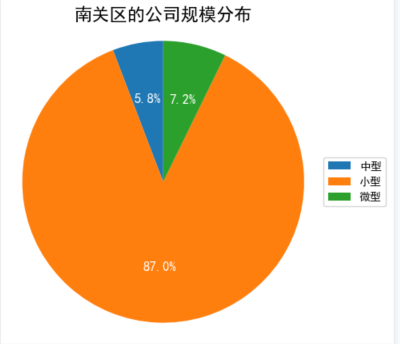
从上图可以看出，宽城区的公司规模以小型为主，占比为81.2%，微型和中型都相对较少，分别占比8.4%和10.3%。

1. **长春经济技术开发区**

****

从上图可以看出，长春经济技术开发区的公司规模以小型为主，占比为81.9%，微型和中型都相对较少，分别占比8.0%和10.2%。

1. **南关区**



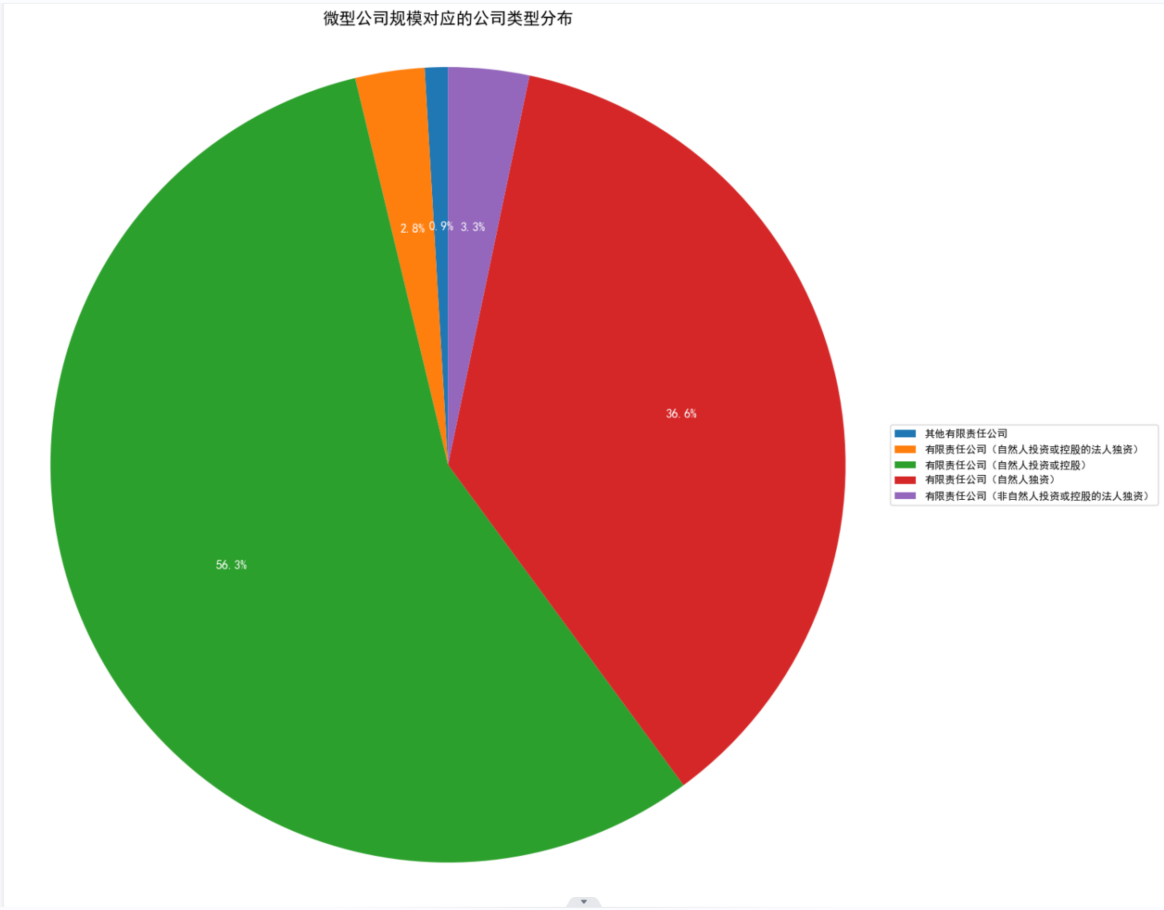
从上图可以看出，南关区的公司规模以小型为主，占比为87.0%，微型和中型都相对较少，分别占比7.2%和5.8%。

每个地区大都有中小型企业，各地区大都是以小型企业为主，小型企业占比最多，其次是中型企业，最后是微型企业。产生这种情况的原因可能和当地的环境、经济、政策等有关系。

1. **分析公司规模和公司类型的关系**

接下对公司规模和公司类型的关系进行分析。

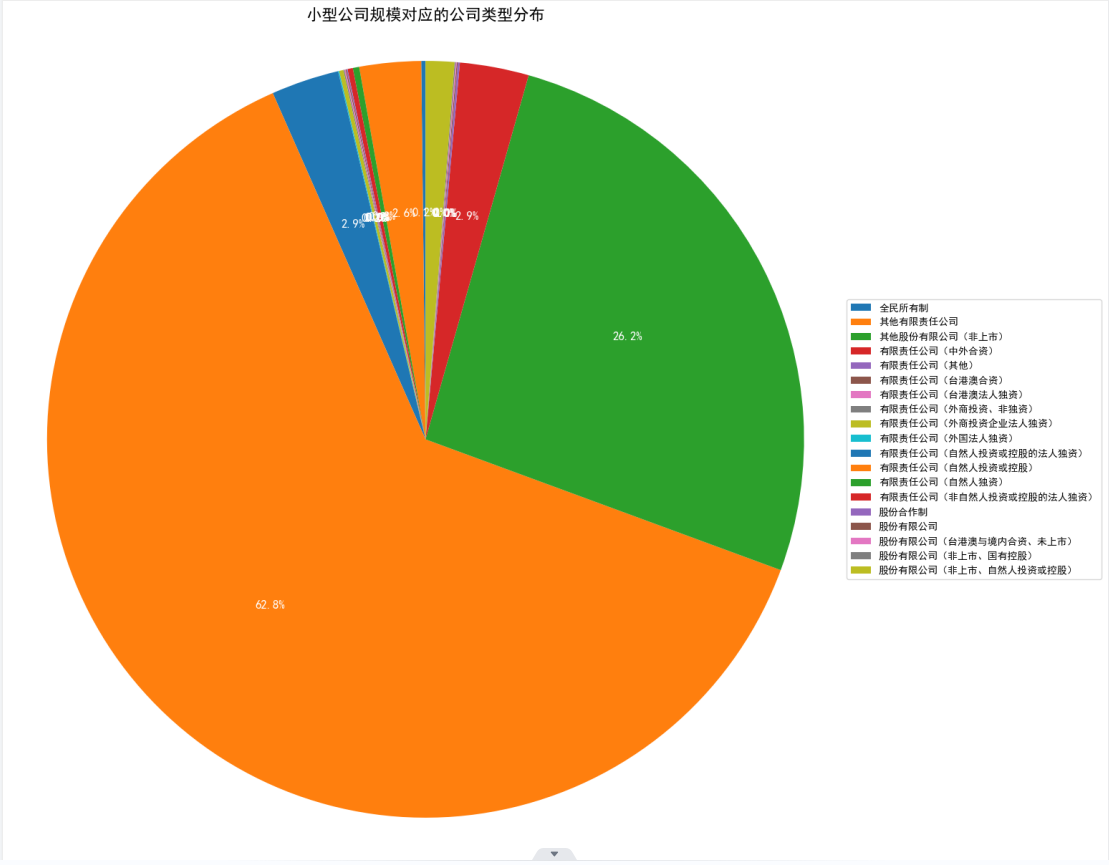
1. **微型公司**



如上图所示，微型公司的公司类型只有有限责任公司，其中自然人投资或控股占比高达56.3%，其次是自然人独资，占比达36.6%。

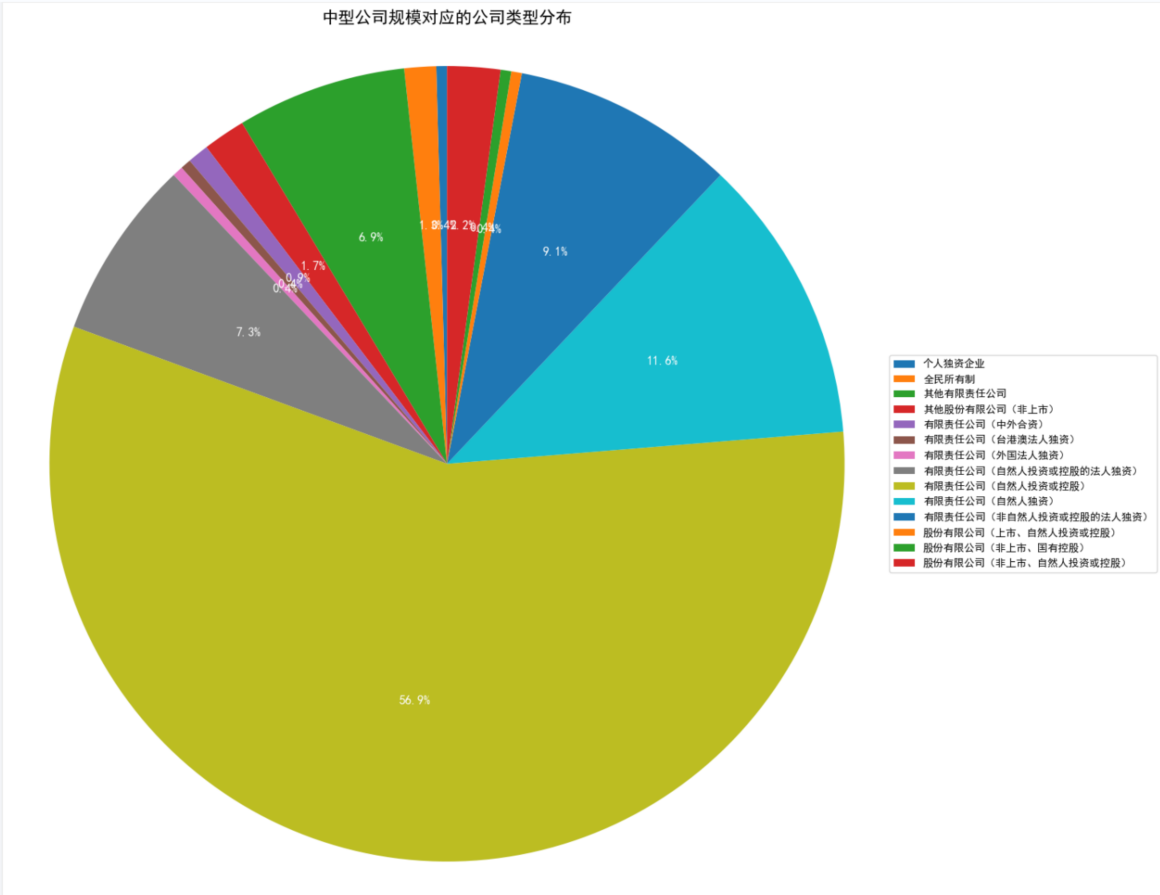
1. **小型公司**

小型公司规模和对应公司类型的关系如下图所示：



如上图所示，小型公司仍然以有限责任公司制为主，但是出现一定比例的股份有限公司和全民所有制，其中有限责任公司（自然人投资或控股）占比高达62.8%，有限责任公司（自然人独资）占比26.2%，

1. **中型公司**



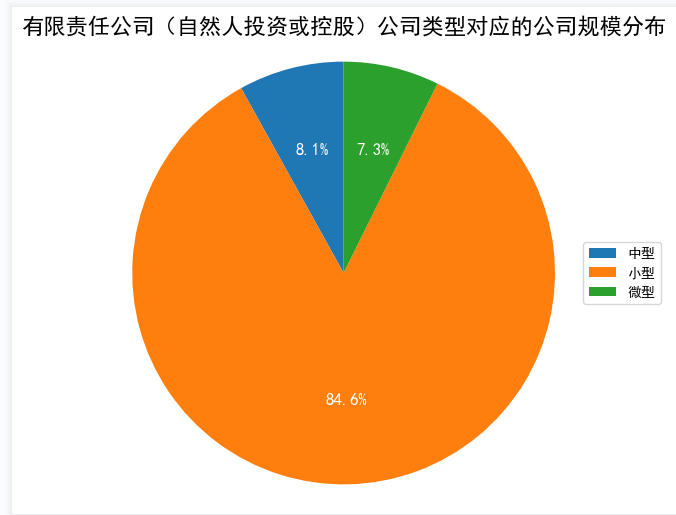
从上图可知，中型公司公司类型出现个人独资企业，有限责任公司比例缩小，其中有限责任公司（自然人投资或控股）占比达56.9%，有限责任公司（自然人独资）占比11.6%，出现中外合资的有限责任公司，占比9.1%。

无论是那种公司类型，大都以“有限责任公司”为主，尤其是微型公司，公司类型只有有限责任公司制，小型公司和中型公司开始出现股份有限公司和全民所有制，有趣的是在中型公司当中出现了一定比例的个人独资企业。而且，随着公司规模变大有限责任公司制比例逐渐缩小，开始出现多元化的制度形式。

1. **分析公司类型和公司规模的关系**

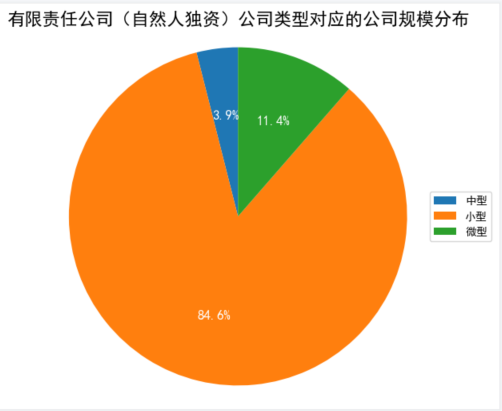
本研究对长春市内占比较大的公司类型和公司规模之间的关系进行了分析。

1. **有限责任公司（自然人投资或控股）**



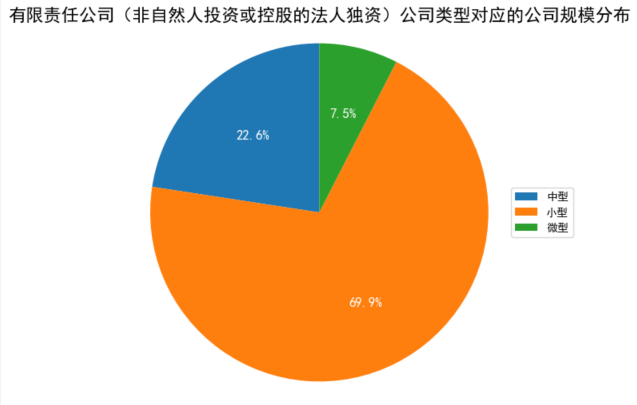
如上图所示，自然人投资或控股的有限责任公司以小型为主，占比84.6%，中型和微型公司占比相对较少。

1. **有限责任公司（自然人独资）**



如上图所示，自然人独资的有限责任公司以小型为主，占比84.6%，中型和微型公司占比相对较少，但，相较于中型而言，微型居多。

1. **有限责任公司（非自然人投资或控股的法人独资）**

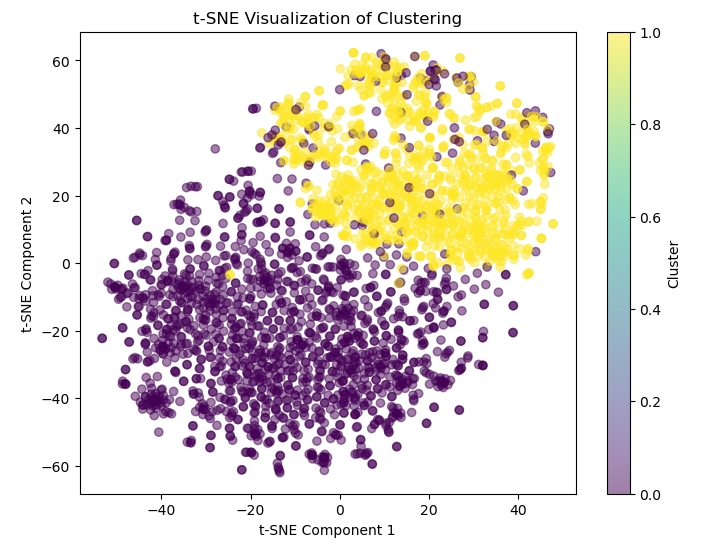


如上图所示，非自然人投资或控股的法人独资的有限责任公司以小型为主，占比69.9%，中型和微型公司占比相对较少。但，相较于微型而言，中型居多。

每种公司类型大都有中小型企业，各种类型大都是以小型企业为主，小型企业占比最多。

1. **对企业经营范围BS进行聚类**

采用TF-IDF将企业经营范围BS进行向量化，并使用KMeans聚类算法对BS进行聚类，并将聚类结果进行可视化，如下图所示：



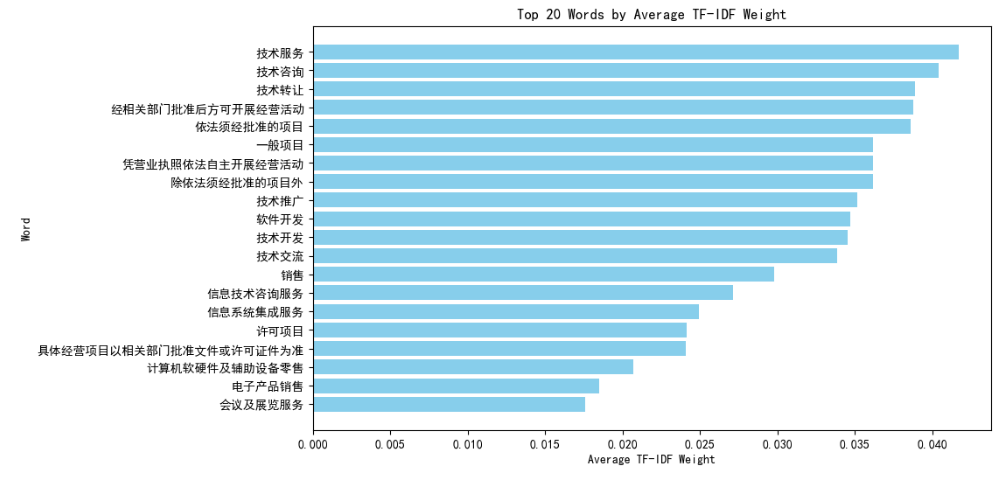
如上图所示，KMeans算法将数据分成两类，对应的BS（经营范围）如下表所示（只选取了主要的分析关键词）：

|  |  |
| --- | --- |
| 1类 | 许可经营项目凭有效许可证或批准文件经营；一般经营项目可自主选择经营等 |
| 2类 | 一般项目：软件开发；技术服务、技术开发、技术咨询、技术交流、技术转让、技术推广等 |

第1类有1594个样例，与计算机技术关系不大；第2类有1061个样例，大多与计算机技术有关。

## 5.3 文本特征处理

本研究主要对企业的经营范围进行探索，首先采用TF-IDF方法对文本数据进行向量化，完成关键词提取，之后采用LDA模型完成主题分析，并生成可视化图像。使用TF-IDF提取关键词的结果如下图所示：



从上图可以发现，企业经营范围主要以技术服务、技术咨询、技术转让为主，技术和法相关的关键字在企业经营范围中占了很大的分量。

使用LDA模型进行主题分析的结果如下表所示：

|  |  |
| --- | --- |
| Topic 1: | ['水性涂料研发', '道路救援', '通讯设备销售', '智能光电系统', '室内空气污染治理', '大数据科技领域内的技术开发', '电器材料批发零售', '电子图书销售', '信息产业项目管理', '塑料加工专用设备制造'] |
| Topic 2: | ['城乡规划编制资质乙级', '办公设备及耗材技术服务', '水资源管理与保护技术咨询', '覆膜砂', '自动化系统工程与技术咨询', '数据维护', '云平台服务', '城市路灯和亮化工程设计', '非专控通讯设备', '组织文化交流'] |
| Topic 3: | ['农林工程', '版权代理服务', '住宅水电安装维护服务', '模具标准件', '住宅小区宽带通信工程', '机电设备的设计', '各类生产设备', '净水器', '大数据资源服务', '生物科技领域的技术研发'] |
| Topic 4: | ['通信系统与计算机网络工程系统集成的设计', '智能交通产品技术开发', '信息网络', '品牌推广', '光学加工设备租赁', '能源与环保技术开发', '导航终端制造', '房地产咨询', '林业空间信息技术应用与开发', '汽车贴膜服务'] |
| Topic 5: | ['公共卫生场所监测', '混凝土添加剂', '凭环保许可证', '门锁', '办公现代化设备及其配件与相关耗材', '固体废物检测', '营商环境提供场地', '润滑油销售', '脱硝催化剂', '生物分离纯化试剂盒'] |

从上图可知，企业的经营范围大致可以分为5个主题：

1. **主题1: 科技领域相关的公司活动**

该主题的关键词包括“水性涂料研发”、“智能光电系统”、“大数据科技领域内的技术开发”等，表明这些公司可能在从事科技研发方面的工作，如新材料、智能系统等。

这些公司可能专注于新技术的研究与开发，如水性涂料、智能光电系统等，这些领域在现代科技中具有重要的应用前景。

1. **主题2: 城市规划、云服务等相关活动**

该主题的关键词包括“城乡规划编制资质乙级”、“云平台服务”、“城市路灯和亮化工程设计”等，表明这些公司可能在从事城市规划、云服务等方面的工作。

这些公司可能参与城市基础设施建设、城市规划设计、云服务提供等领域的活动，致力于提升城市的智能化水平和服务质量。

1. **主题3: 农林工程、生物科技等相关活动**

该主题的关键词包括“农林工程”、“生物科技领域的技术研发”、“净水器”等，表明这些公司可能在从事农林工程、生物科技等领域的工作。

这些公司可能致力于农林资源的开发利用、生物科技产品的研发生产，以及环境保护与治理等方面的工作。

1. **主题4: 智能交通、能源环保等相关活动**

该主题的关键词包括“智能交通产品技术开发”、“能源与环保技术开发”、“汽车贴膜服务”等，表明这些公司可能在从事智能交通、能源环保等领域的工作。

这些公司可能参与智能交通产品的研发与制造、环保技术的开发与应用，以及相关服务行业的提供等活动。

1. **主题5: 公共卫生监测、办公设备服务等相关活动**

该主题的关键词包括“公共卫生场所监测”、“办公现代化设备及其配件与相关耗材”、“润滑油销售”等，表明这些公司可能在从事公共卫生监测、办公设备服务等领域的工作。

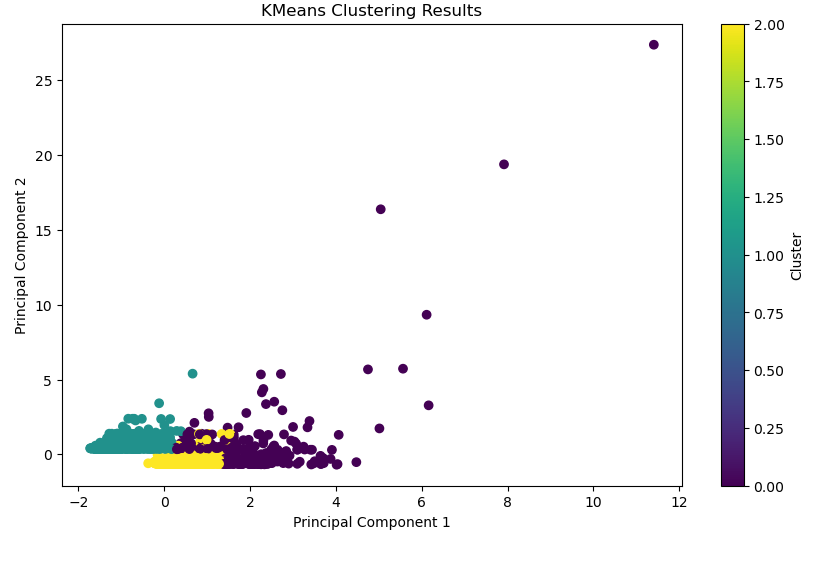
这些公司可能参与公共卫生设施的监测与维护、办公设备的销售与维修等活动，为社会提供相关的服务与支持。

# 建模和分析

本部分从数据清洗后的数据中选取合适特征，把选取的数据成两个数据集，采用KMeans聚类算法对数据进行分析。

1. **选取经营范围Zcluster、创建时间ZCT和注册资本ZRC对数据进行聚类**

聚类结果并如下图所示：



**散点图分析：**

散点图是一种用于展示两个变量之间关系的常见图表类型。在本例中，散点图用于展示数据样本在两个主成分上的分布情况，以及聚类结果的可视化。

每个散点代表一个数据样本，其位置由其对应的两个主成分的值决定。

聚类结果通过颜色来表示，相同颜色的散点属于同一类别，不同颜色的散点属于不同的类别。

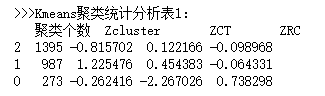
**横轴和纵轴的含义：**

横轴（X轴）：表示第一个主成分（Principal Component 1，PC1），是经过主成分分析（PCA）后得到的数据在第一个主成分上的投影值。主成分分析是一种降维技术，用于将原始数据映射到新的特征空间，其中每个特征称为主成分，以捕捉原始数据中的大部分方差。

纵轴（Y轴）：表示第二个主成分（Principal Component 2，PC2），同样是通过主成分分析得到的数据在第二个主成分上的投影值。

主成分通常是数据中最重要的方差方向，因此，通过对主成分的分析，可以更好地理解数据的结构和特征。

**聚类后主要数据如下：**



并将对应关系存入csv文件中

根据聚类结果可以发现，通过KMeans算法可以将原数据分为3类。

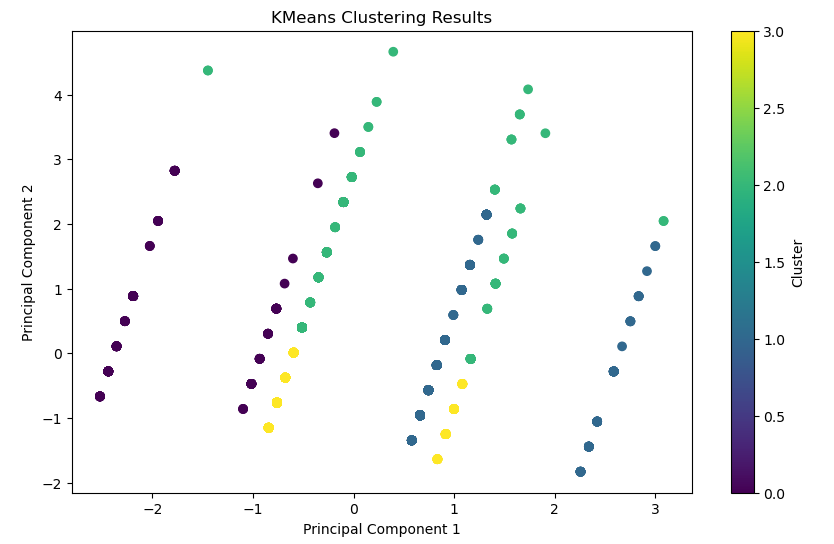
第0类：有273个，聚类中心为-0.262416、-2.267026、0.738298

第1类：有987个，聚类中心为1.225476、0.454383、-0.064331

第2类：有1395个，聚类中心为-0.815702、0.122166、-0.098968

1. **选取经营范围Zcluster、所属区县ZDAC和企业规模ZFS进行聚类**

聚类结果如下图所示：



**散点图分析：**

散点图是一种用于展示两个变量之间关系的常见图表类型。在本例中，散点图用于展示数据样本在两个主成分上的分布情况，以及聚类结果的可视化。

每个散点代表一个数据样本，其位置由其对应的两个主成分的值决定。

聚类结果通过颜色来表示，相同颜色的散点属于同一类别，不同颜色的散点属于不同的类别。

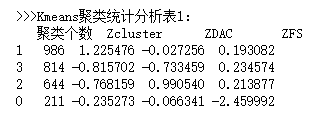
**横轴和纵轴的含义：**

横轴（X轴）：表示第一个主成分（Principal Component 1，PC1），是经过主成分分析（PCA）后得到的数据在第一个主成分上的投影值。主成分分析是一种降维技术，用于将原始数据映射到新的特征空间，其中每个特征称为主成分，以捕捉原始数据中的大部分方差。

纵轴（Y轴）：表示第二个主成分（Principal Component 2，PC2），同样是通过主成分分析得到的数据在第二个主成分上的投影值。

主成分通常是数据中最重要的方差方向，因此，通过对主成分的分析，可以更好地理解数据的结构和特征。

**聚类后主要数据如下：**



并将对应关系存入csv文件中

根据聚类结果可以发现，通过KMeans算法可以将原数据分为4类。

第0类：有211个，聚类中心为-0.2635273、-0.066341、-2.4559992

第1类：有986个，聚类中心为1.225476、-0.027256、-0.193082

第2类：有644个，聚类中心为-0.768159、0.990540、0.213877

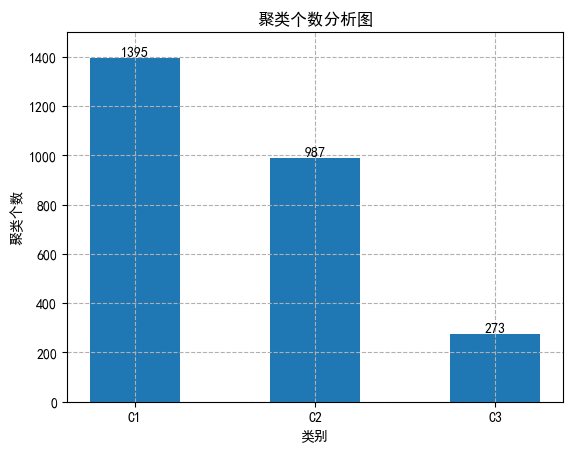
第3类：有814个，聚类中心为-0.815702、-0.733459、0.234574

## 6.1 可视化结果分析

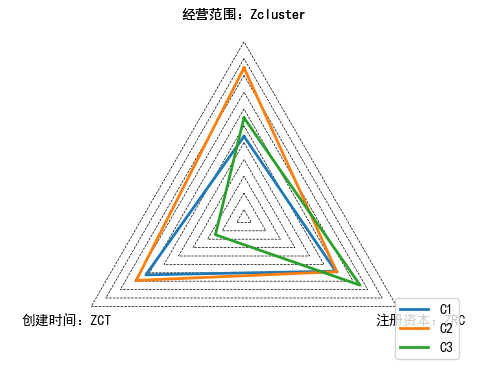
最后对于两种不同的聚类情况进行可视化分析

* 一个：'Zcluster','ZCT','ZRC'（经营范围，创建时间，注册资本）

柱形图：



雷达图：



根据数据和图像，可以简单的发现经营范围越和计算机技术想关的创建时间较晚，注册资本也较少，其中C3表现出注册时间很早，注册资本较高，经营范围偏重1类，但是数量较少。而且根据聚类数据可以发现注册时间较早的往往资本更多。

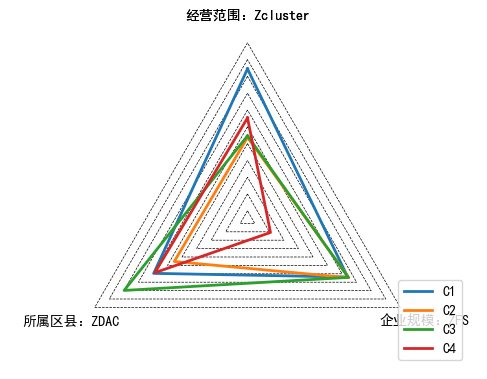
* 一个：'Zcluster','ZDAC','ZFS'（经营范围，所属区县，企业规模）

柱形图：

图表, 条形图

描述已自动生成

雷达图：



根据图像和相关映射关系，可以简单发现经营范围越和计算机技术相关的所属区县往往企业数较多，大多数企业以小型企业为主。

# 结论与思考

在所选数据集中，企业以小型企业为主，以“有限责任公司（自然人投资或控股）”TF-IDF进行处理可以发现经营范围以“技术服务”，“技术咨询”，“技术转让”等为主，进行LDA主题模型分析可以发现经营范围也和农林，生物，公共卫生，环境保护有关。

通过独热算法将字符串转换为数值型，并采用Z-sore进行标准化，最后通过对不同特征进行聚类分析可以看出，经营范围和计算机相关的很多创建时间相对较晚，所属区县较为集中，企业规模以小型为主。