# **企业数据分析**

[**企业数据分析** 1](#_Toc161398337)

[**摘要** 1](#_Toc161398338)

[**引言** 1](#_Toc161398339)

[**数据获取** 1](#_Toc161398340)

[**数据探索** 1](#_Toc161398341)

[**简单描述** 1](#_Toc161398342)

[**简单统计** 2](#_Toc161398343)

[**文本特征处理** 5](#_Toc161398344)

[**数据清洗** 6](#_Toc161398345)

[**数据标准化** 7](#_Toc161398346)

[**建模和分析** 8](#_Toc161398347)

[**可视化结果分析** 9](#_Toc161398348)

[**结论与思考** 11](#_Toc161398349)

## **摘要**

* 使用到的方法：

TF-IDF，LDA主题模型，独热算法，KMeans聚类算法等

* 得出的结论：

……

## **引言**

## **数据获取**

通过互联网收集企业的数据，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。

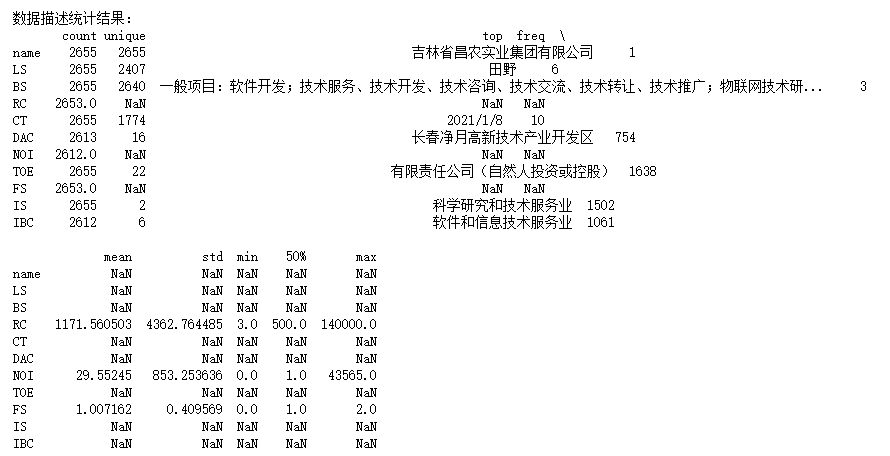
数据对象就是长春市入户科技型中小企业。

将数据放在data01中，经统计长春共有2655家中小微型企业。

## **数据探索**

### **简单描述**

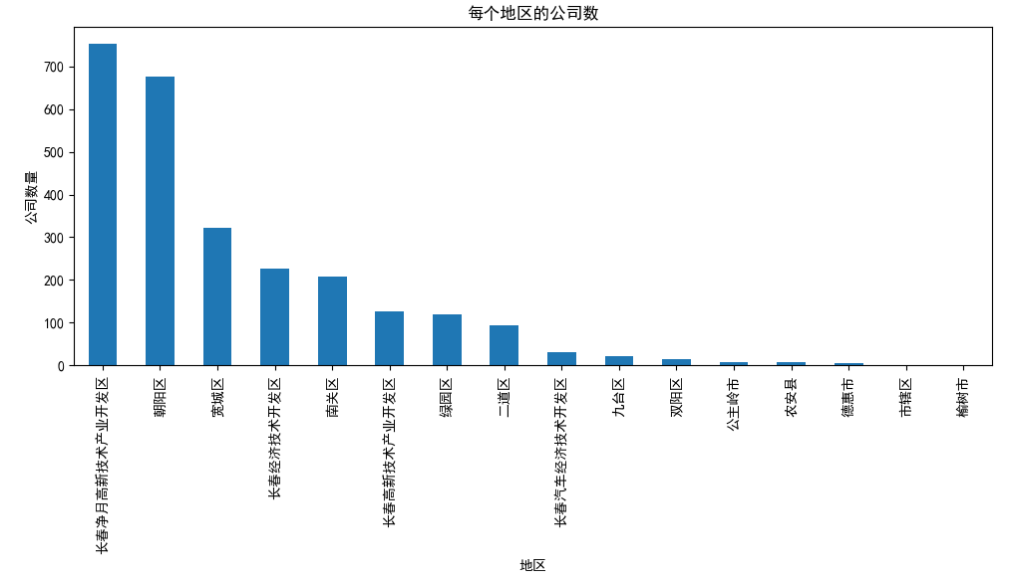
采用describe进行简单的数据描述，并输出有关数据相关的信息，包括平均值、方差、最小值、最大值、中位数等信息。



### **简单统计**

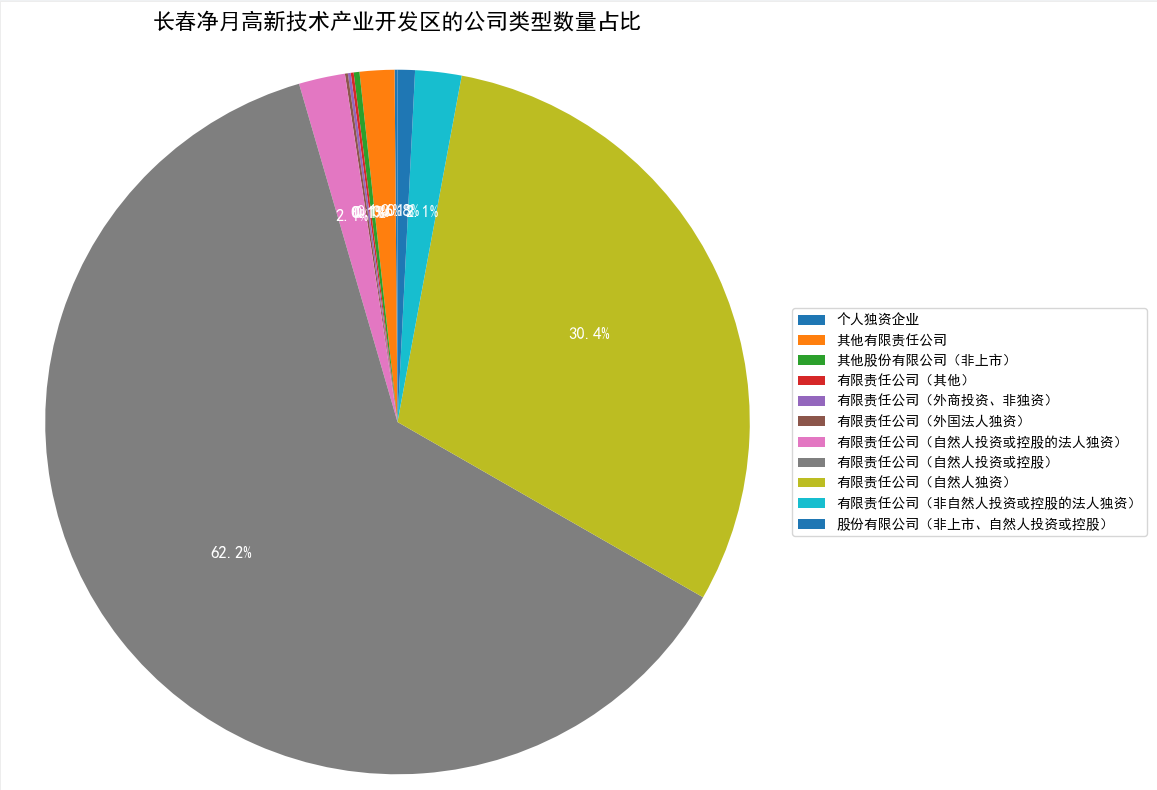
为方便后边的分析，在这里进行数据相关统计工作。在这里进行数据可视化工作，并存入image中。

* 统计每个地区有多少个公司



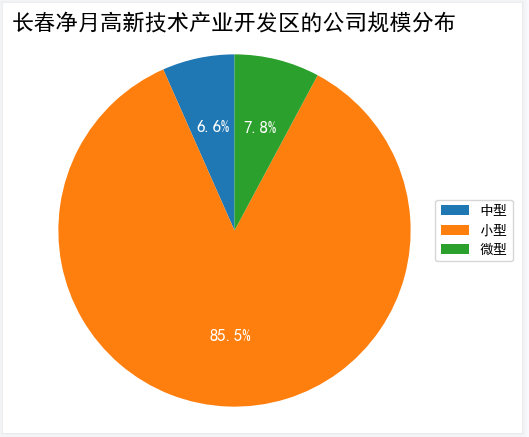
从图表中可以看到 长春的公司主要集中到“长春净月高新技术产业开发区” ，其次在“朝阳区”，两者求和将近一半之多，产生这样现象的原因可能和当地的政策和环境有关。

* 每个地区的公司类型有哪些



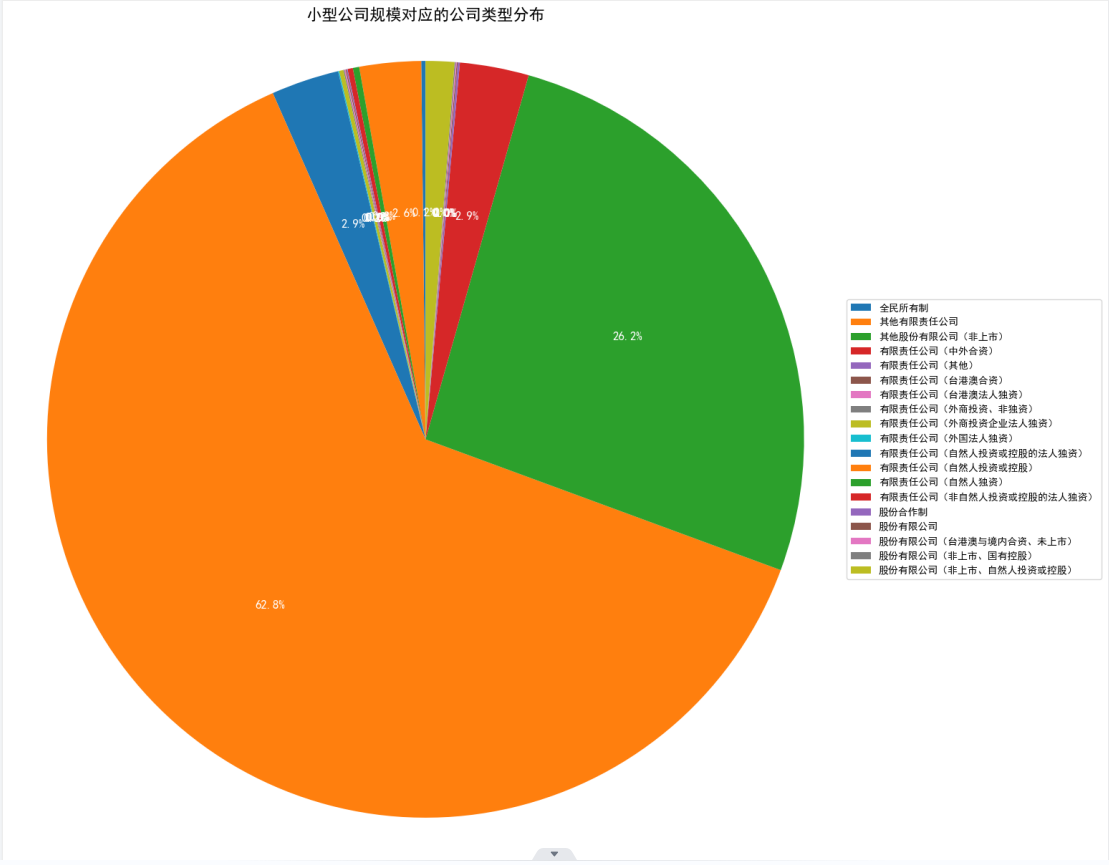
每个地区公司种类多样，公司数量越多，种类越是丰富，而且几乎所有地区的公司都以“有限责任公司（自然人投资或控股）”为主，产生这种原因可能和公司制度特点有关系。

* 每个地区的公司规模是什么情况（公司规模中： 0-微型 1-小型 2-中型 3-大型）



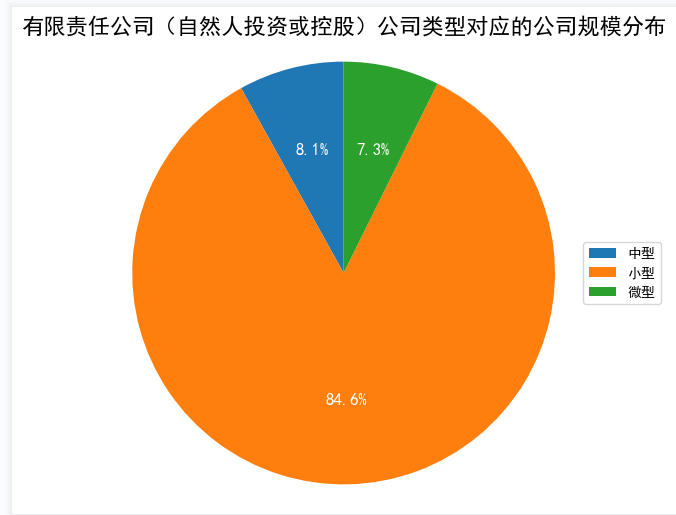
每个地区大都有中小型企业，各地区大都是以小型企业为主，小型企业占比最多。产生这种情况的原因可能和当地的环境、经济、政策等有关系。

* 公司规模和公司类型的关系



无论是那种公司类型，大都以“有限责任公司（自然人投资或控股）”为主

* 公司规模和公司类型的关系

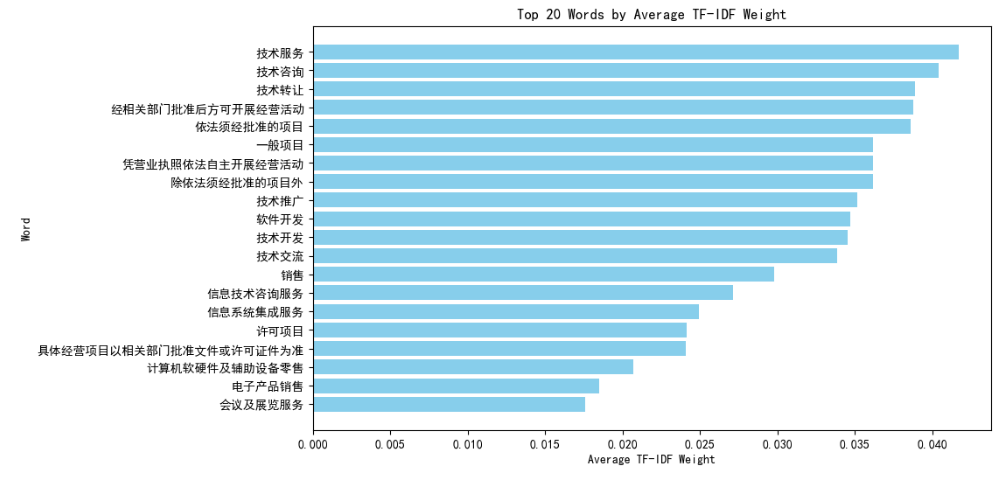


每种公司类型大都有中小型企业，各种类型大都是以小型企业为主，小型企业占比最多。

### **文本特征处理**

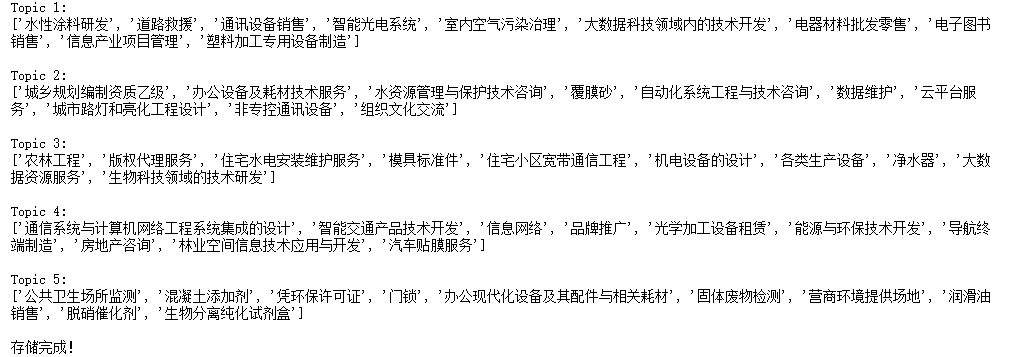
本部分主要对于经营范围进行处理，对经营范围进行简单探索，采用TF-IDF方法进行向量分析，完成关键词提取，之后采用LDA主题模型完成主题分析，并生成可视化图像。

* TF-IDF处理结果



经过分析文本，并提取关键词，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”，“法”等在其中占了很大的分量。

* LDA主题模型主题分析



**Topic 1:**

这个主题可能与科技领域相关，涉及到水性涂料研发、智能光电系统、大数据科技等。

该主题的概率分布中，第4个词的概率最高，可能代表着该主题在文档中的出现频率较高。

**Topic 2:**

这个主题可能与城市规划、办公设备技术服务等相关。

该主题的概率分布中，第1个词的概率最高，可能代表着该主题在文档中的出现频率较高。

**Topic 3:**

这个主题可能与农林工程、生物科技研发等相关。

该主题的概率分布中，第4个词的概率最高，可能代表着该主题在文档中的出现频率较高。

**Topic 4:**

这个主题可能与通信技术、智能交通产品开发等相关。

该主题的概率分布中，第3个词的概率最高，可能代表着该主题在文档中的出现频率较高。

**Topic 5:**

这个主题可能与公共卫生、环境保护等相关。

该主题的概率分布中，第1个词的概率最高，可能代表着该主题在文档中的出现频率较高。

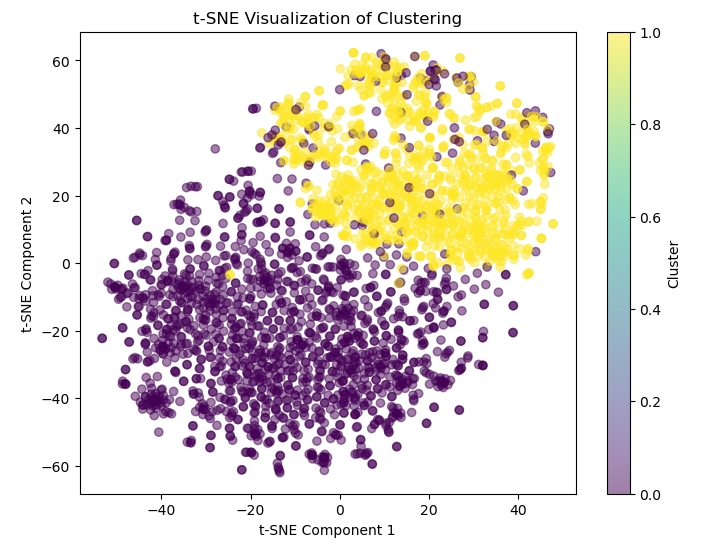
## **数据清洗**

对缺失值主要采用众数插入的方式，主要原因在于大多数据是字符串，并且很多都是一样的。

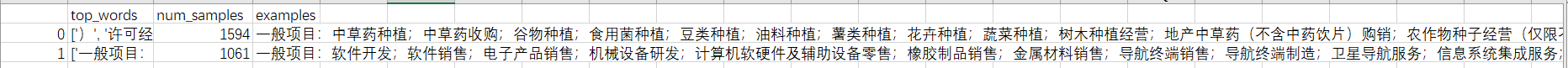
## **数据标准化**

主要采用Z-sore进行标准化，防止数据差异太大。主要完成以下工作内容：

1. 主要采用独热编码处理分类数据，并生成映射关系表。
2. 采用TF-IDF将BS（经营范围）转化，并使用KMeans算法聚类，并将其可视化。



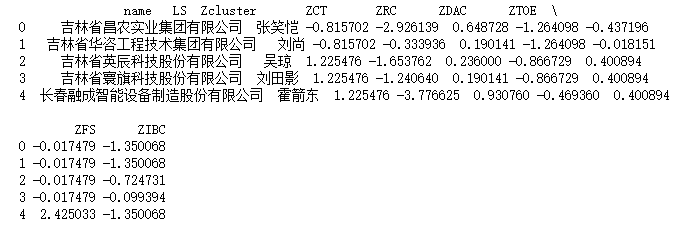
主要将数据分成两类：



第一类大多与计算机技术关系不大，有1594个样例

第二类大多与计算机技术有关，有1061个样例

1. 将CT（创建时间）从字符串类型转换成数值类型。
2. 然后采用z-score算法将数据进行标准化,只保留了'cluster'(经营范围),'CT'（创建时间）,'RC'（注册资本）,'DAC'（所属区县）,'TOE'（企业类型）,'FS'（企业规模）,'IBC'（国标行业大类）。



## **建模和分析**

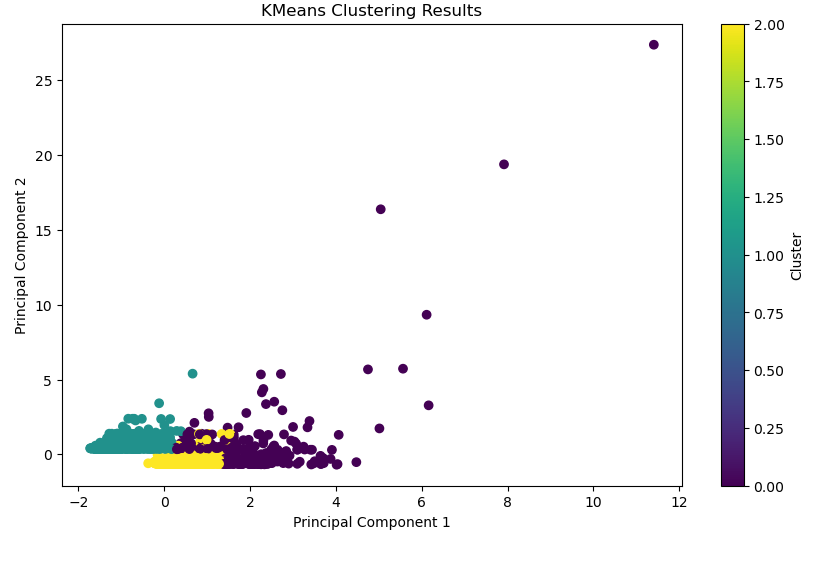
本部分主要采用KMeans算法，选取合适特征进行处理。

采用KMeans聚类分析的方法

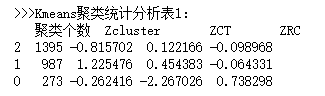
生成两个数据集，对两个数据集进行分析

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

聚类图像如下：



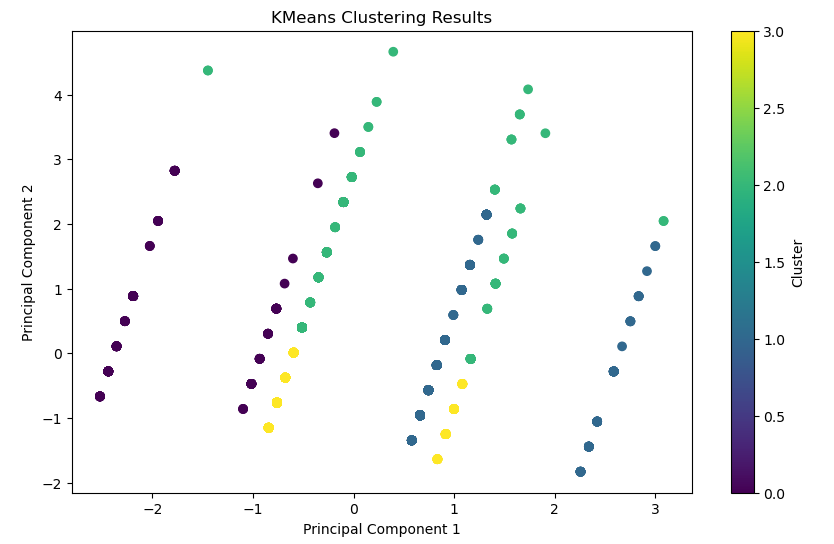
主要数据如下：



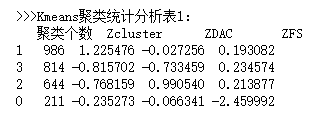
并将对应关系存入csv文件中

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

聚类图像如下：



主要数据如下：



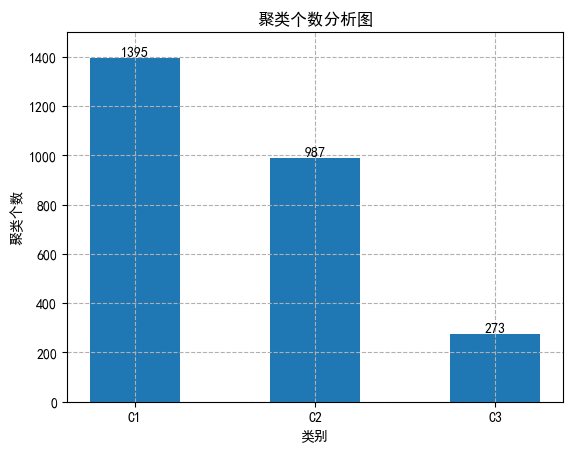
并将对应关系存入csv文件中

## **可视化结果分析**

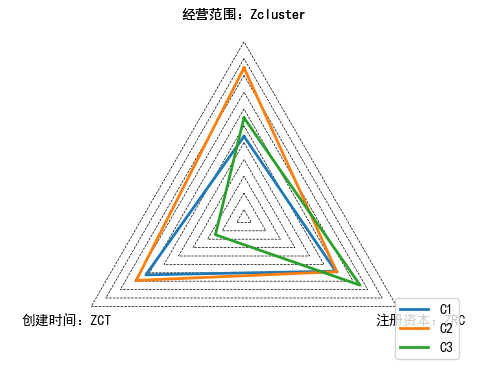
最后对于两种不同的聚类情况进行可视化分析

* 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。

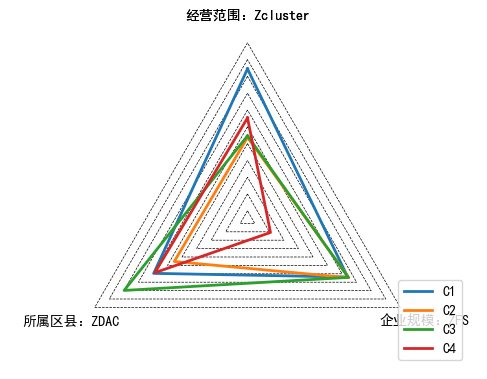
* 一个：'Zcluster','ZDAC','ZFS'（经营范围，所属区县，企业规模）

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。