

机器学习

课程报告

**学 院： 经济与金融学院**

**专业班级： 金融学汇丰班**

**课程名称： 机器学习**

**学生姓名： 吴泓栎 王馨怡**

**孙智敏 杨子涵**

日期：2021年7月2日

目录

[一、概述 1](#_Toc76141013)

[二、数据来源 1](#_Toc76141014)

[1. 整体初始指标情况 1](#_Toc76141015)

[2.各部分获取 1](#_Toc76141016)

[三、数据预处理 5](#_Toc76141017)

[1.数据整合： 5](#_Toc76141018)

[2.数据处理与转换 6](#_Toc76141019)

[3. 样本数据可视化展示 7](#_Toc76141020)

[四、特征工程构建 7](#_Toc76141021)

[1.特征构建： 7](#_Toc76141022)

[2.特征提取 7](#_Toc76141023)

[3.特征选择 9](#_Toc76141024)

[五、模型构建 10](#_Toc76141025)

[1.模型选择过程 10](#_Toc76141026)

[2.模型计算 11](#_Toc76141027)

[3.模型聚类结果及解释 13](#_Toc76141028)

[六、 成员分工 19](#_Toc76141029)

**淘宝网电商卖家的信用评估**

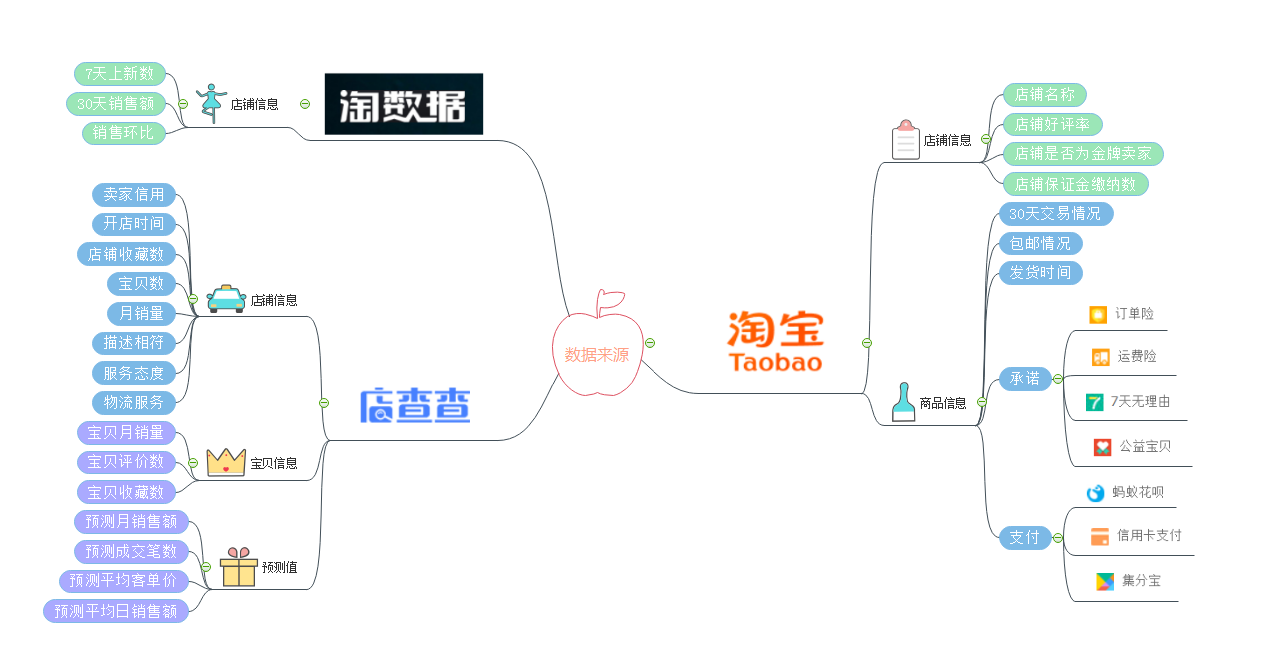
# 一、概述

电商平台卖家的信用评估体系一直在持续不断地完善与发展之中，本研究结合因子分析以及kprototype机器学习算法，以淘宝网服装业电商卖家为样本，建立对淘宝网电商卖家的信用评估方案。

工具：Python、SPSS、Excel、dataV

# 二、数据来源

## 1. 整体初始指标情况



## 2.各部分获取

（1）**淘宝网**(<https://www.taobao.com/>)

在进入淘宝网搜索页后，输入“服装”以及筛选店铺为淘宝店铺，排序依据淘宝平台信用评分从高到低返回结果。结果显示有100页，每页24家店铺。利用Python爬虫技术爬取每页的前三家店铺的部分基础信息，共获取300家店铺信息。获取部分如下：

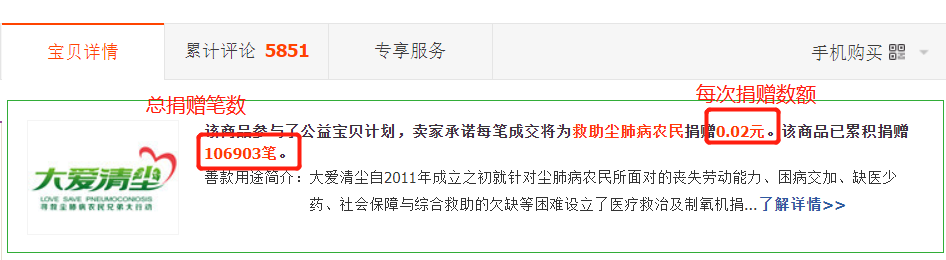


对每家店铺，获取进入店铺链接后的部分信息，包括商家基础信息以及商品部分数据，对应每家店铺一张表，爬取部分如下：





同时对商品承诺中含有公益宝贝计划的商品，获取对应信息：



**（2）店查查**(<https://www.dianchacha.com/>)

根据店铺id进行网址拼接，爬取店铺的相关信息：

i.基础信息



ii.店铺7天销量



iii.综合搜索情况



iv.宝贝列表



v.月销售预测



**（3）淘数据**(<https://www.taosj.com/>)

获取店铺部分销售数据：拼接网址，爬虫实现+手动填充



# 三、数据预处理

## 1.数据整合：

（1）淘宝网店铺基础信息表：

提取好评率的小数，去除网址

（2）店铺宝贝列表：



计算月销量均值、评价数中位数、收藏数中位数、价格均值作为店铺的宝贝信息

（3）店铺热度和曝光度表：



宝贝曝光度和热度取均值作为店铺曝光度和热度

（4）店铺月销预测



转换数据类型

**数据整合方式**：由于淘宝店铺信息获取时，获取了店铺的id，作为店查查的网址拼接，保存数据时会用id作为第一列或者用id命名进行保存，最终利用id进行整个数据表的拼接。

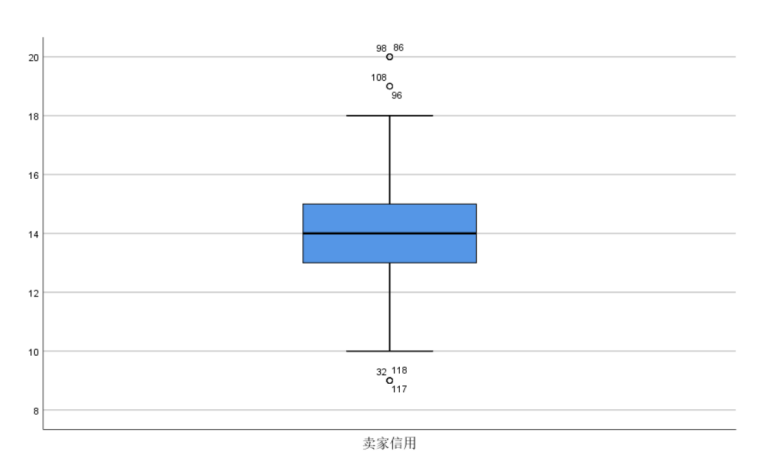
## 2.数据处理与转换

（1）**缺失值**处理

对于**缺失值过多的店铺**选择从数据表中删除；少部分空缺属于爬虫无法爬取的值，手动填充该缺失值；部分无法查询的值，选择通过前后值或中位值填充。

（2）**异常值**检验和处理

采取箱线图逐个对数值型指标进行异常值检验，以卖家信用为例：



存在异常高和异常低的离群点，经过仔细的数据对比，源数据中的真实值就是如此，对于过于离群的数据选择整行删除，部分选择保留。

（3）**指标转换**

将指标开店时间转化成距今的天数，同时删除方差小于设定阈值的数值型指标。

整合所有数据表后，得到初始数据表如下：



最终保留228条商家数据

## 样本数据可视化展示

利用dataV在阿里云平台展示样本数据整体情况，网址如下：

<https://datav.aliyuncs.com/share/c8d8eec6add658169735f0a681294961>

可视化结果：



# 四、特征工程构建

## **1.特征构建**：

**引入哑变量**

对于部分分类变量（是否金牌卖家、承诺、男装或女装、支付手段）拆分为多个哑变量，取值0、1。

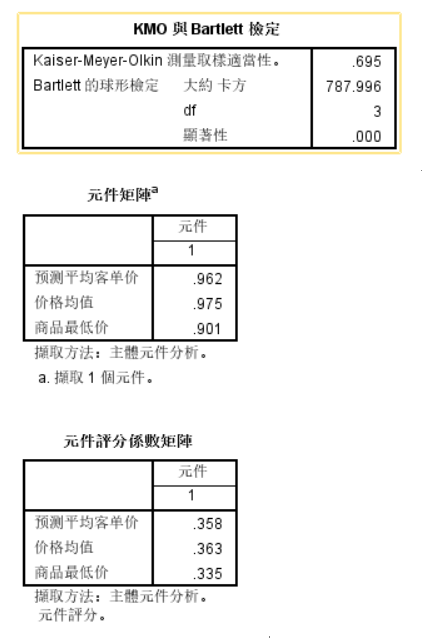
## 2.特征提取

综合后得到43个特征变量，首先利用因子之间的相关性，对各个数值型因子进行聚类划分，将两两之间相关度大于0.5的划分到同一类，根据KMO和Bartlett球体检验对此类别下指标数据是否适合做因子分析进行判断，对于KMO的值大于0.6的且Bartlett球体检验显著性P值小于0.05的指标大类选择保留主因子载荷量最大的那一项。

分类结果：

|  |  |  |  |
| --- | --- | --- | --- |
| 类别 | 指标 | 类别 | 指标 |
| 第一类 | 交易成功比率 | 第六类 | 预测平均客单价 |
| 第二类 | 销售环比 | 价格均值 |
| 第三类 | 平均月销量 | 商品最低价 |
| 第四类 | 距今天数 | 第七类 | 描述相符 |
| 第五类 | 好评率 | 服务态度 |
| 第八类 | 发货时间 | 物流服务 |
| 第九类 | 曝光度 | 第十一类 | 店铺收藏数 |
| 热度 | 月销量 |
| 第十类 | 保证金 | 预测月销售额 |
| 7天上新数 | 预测成交笔数 |
| 第十二类 | 卖家信用 | 预测平均日销售额 |
| 宝贝数 | 收藏数中位数 |
| 第十三类 | 宝贝公益计划 | 评价数中位数 |
| 第十四类 | 30天销售额 |  |  |

KMO与球型检验结果，以第六类为例：



最终提取的主要数值型因子如下：

|  |  |  |  |
| --- | --- | --- | --- |
| 类别 | 指标 | 类别 | 指标 |
| 第一类 | 交易成功比率 | 第六类 | 价格均值 |
| 第二类 | 销售环比 | 第七类 | 描述相符 |
| 第三类 | 平均月销量 | 第八类 | 发货时间 |
| 第四类 | 距今天数 | 第九类 | 曝光度 |
| 第五类 | 好评率 | 热度 |
| 第十一类 | 店铺收藏数 | 第十类 | 保证金 |
| 第十二类 | 宝贝数 | 7天上新数 |
| 卖家信用 | 第十三类 | 宝贝公益计划 |
|  |  | 第十四类 | 30天销售额 |

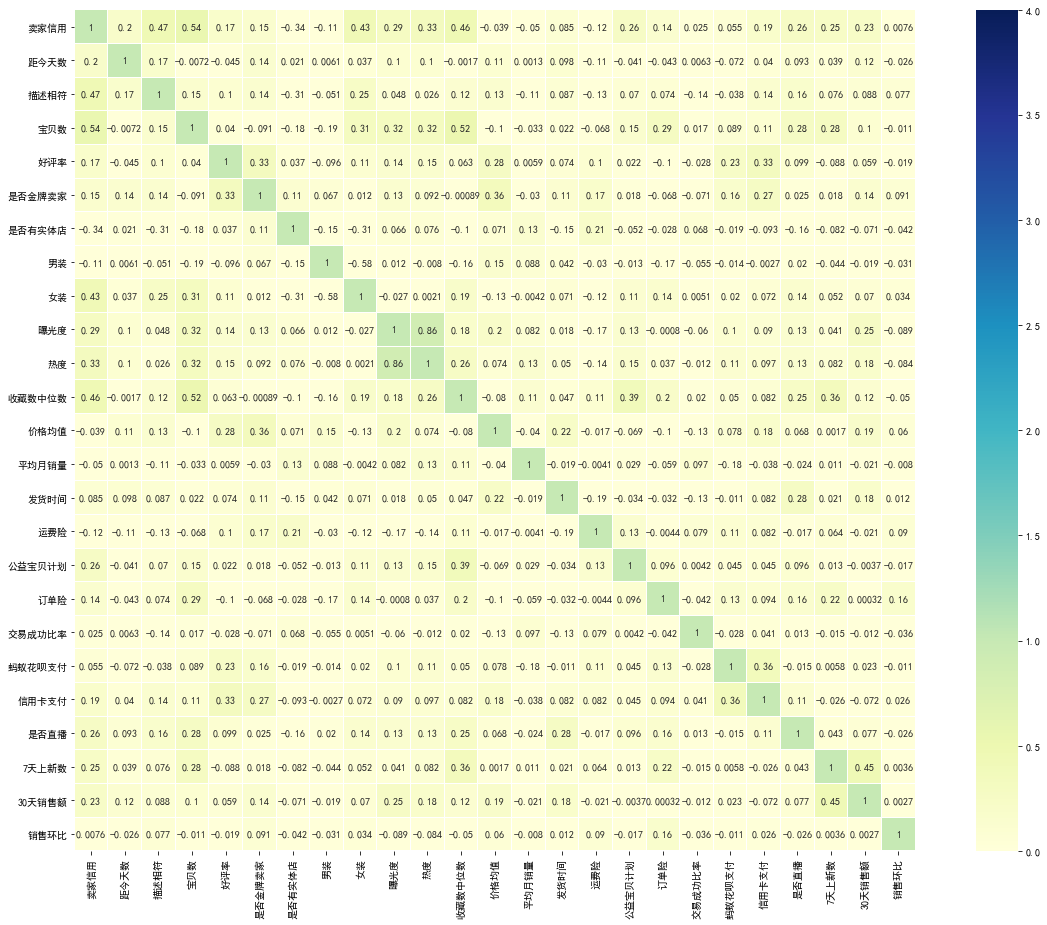
## 3.特征选择

对于频率小于设定阈值的分类型指标进行删除，降低维度，最终得到的特征变量如下：

|  |  |
| --- | --- |
| 分类型 | |
| 是否金牌卖家 | 订单险 |
| 是否有实体店 | 蚂蚁花呗支付 |
| 男装 | 信用卡支付 |
| 女装 | 是否直播 |
| 运费险 |  |

|  |  |
| --- | --- |
| 数值型 | |
| 交易成功比率 | 价格均值 |
| 销售环比 | 描述相符 |
| 平均月销量 | 发货时间 |
| 距今天数 | 曝光度 |
| 好评率 | 热度 |
| 店铺收藏数 | 保证金 |
| 宝贝数 | 7天上新数 |
| 卖家信用 | 宝贝公益计划 |
|  | 30天销售额 |

同时检验各类指标的相关性,特征工程后剩下指标的相关性热图如下：



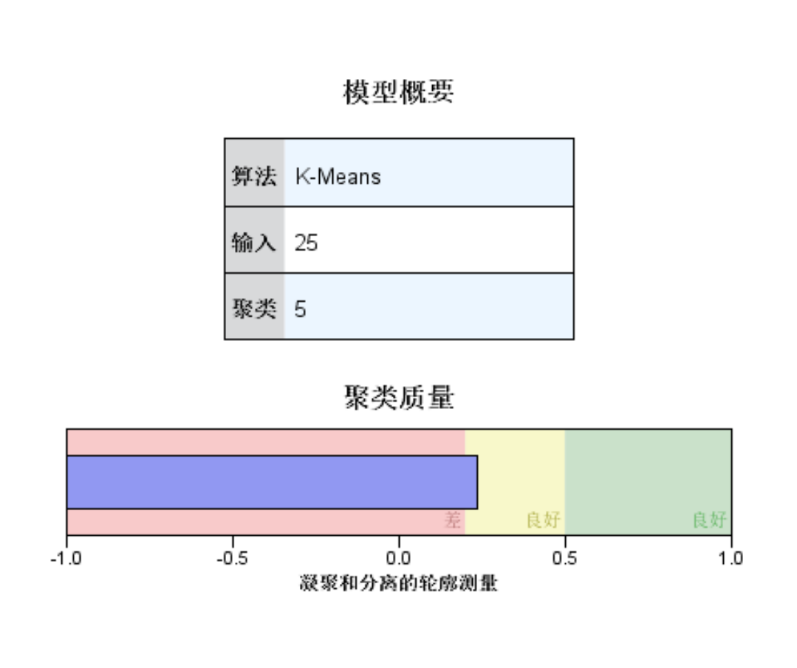
可看出剩余特征相关性都不高，特征工程构建完成。

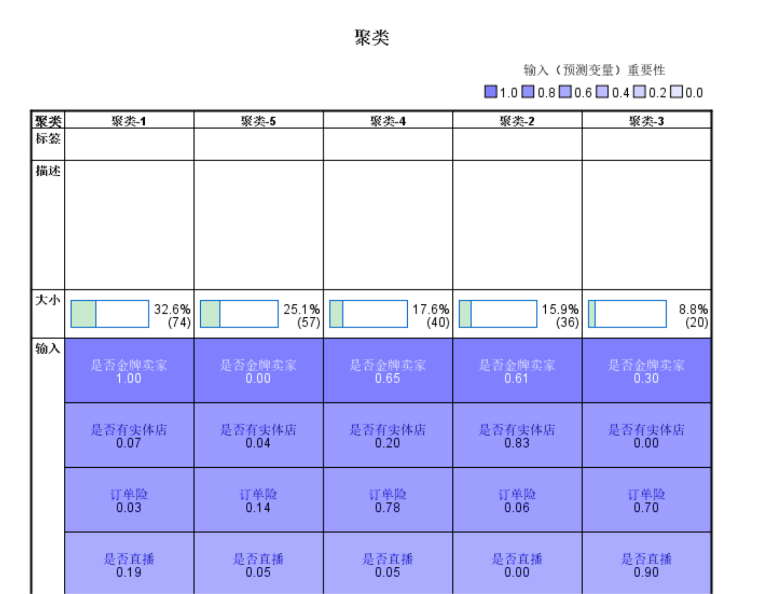
# 五、模型构建

## 1.模型选择过程

**（1）模型选用**

最初预备采用k-means聚类算法进行聚类分析，同时在进入模型前特征工程构建时，选用PCA、KPCA、LLE等多种方法对原高维数据进行降维，然而在进行了初步模型建立(利用SPSS)后，模型的聚类效果并不理想，且轮廓系数均较低，以PCA+k-means结果为例：





考虑到k-means受到分类变量的影响较大，决定更改模型，经过检验，选用K-prototype聚类算法对模型进行改进。

**（2）K-prototype聚类算法介绍**

K-Prototype算法是结合K-Means与K-modes算法，针对混合属性的，解决2个核心问题如下：

* 度量具有混合属性的方法是，数值属性采用K-means方法得到P1，分类属性采用K-modes方法P2，那么D=P1+a\*P2，a是权重。如果觉得分类属性重要，则增加a，否则减少a，a=0时即只有数值属性；
* 2.更新一个簇的中心的方法，方法是结合K-Means与K-modes的更新。

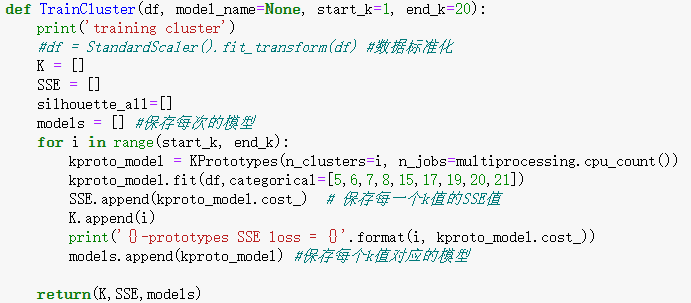
**（3）数值属性距离计算方式**

K-Prototype算法默认采用欧氏距离对数值型属性进行距离计算，但在高维欧式空间中，各数据点的距离变得十分相近，距离计算结果较差。在参考多篇文献后，发现选择使用余弦距离进行数值型属性的距离计算更合理，因此，在模型的数值型距离计算中，采用余弦距离进行计算。

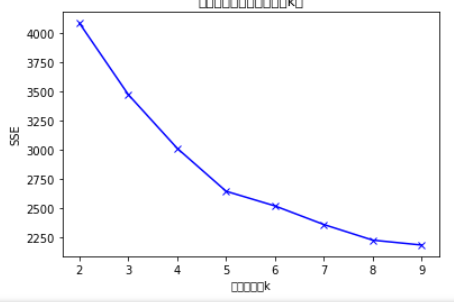
## 2.模型计算

（1）k值的确定

采用**肘部法则**对k值进行确定，



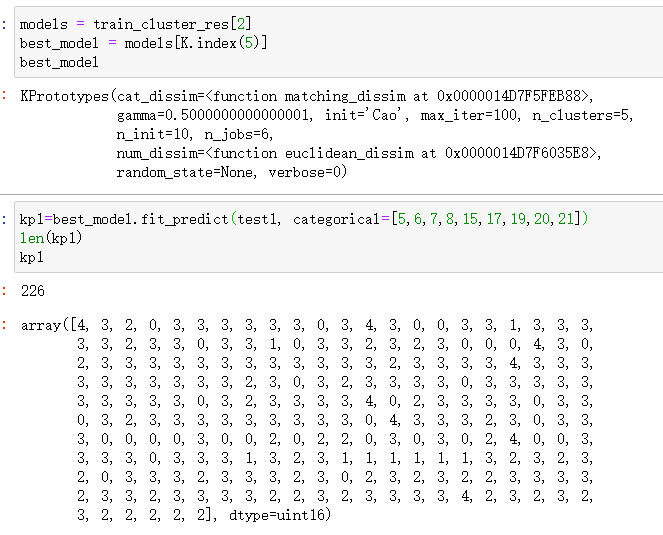
在python中画出不同k值下，整体SSE的变化情况：



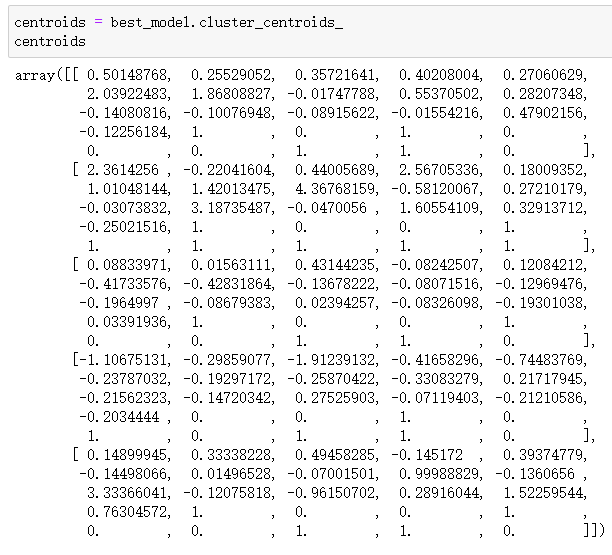
可以看到，在k=5时，效果较好，确定k=5

（2）聚类分析

将得到的数据放入模型中，使k=5，输出聚类结果如下：

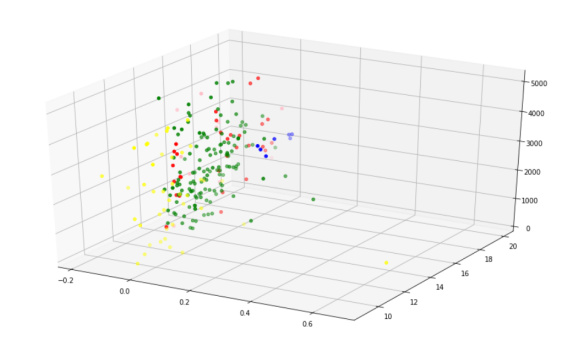


可看出，数据较好地相互分离成5个类别。各类中心点：

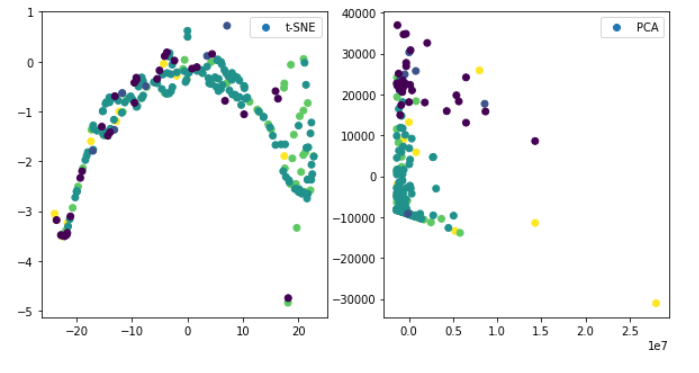


同时对聚类结果进行可视化：

KPCA三维图：



1. SNE以及PCA二维图：



## 3.模型聚类结果及解释

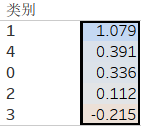
为了使得建模聚类的结果具有应用价值，我们参考了商业银行贷款管理的相关论文，为所有类别生成一个评分，以对每个类进行具有实际意义的区分。经过一定桌面研究和资料整理，我们得知目前银行放贷时对中小企业信用着重参考获利能力、营运能力、获现能力、发展潜力、偿债能力、履约状况、企业规模、领导者素质、市场竞争能力、创新投入、职工素质与能力、行业状况、管理水平、产业链、合同担保情况等等，具体如下图所示：

结合之前所获取的40个特征，无法体现的方面有获利能力、领导者素质、创新投入，无需考虑的方面有行业状况、产业链，因为所有的样本都源于同一个样本行业。由此筛除以上指标，重新根据重要性对各个保留的能力赋予权重。同时在我们所获取的特征中，选取相应能代表那部分能力的单个指标，与能力权重加权汇总，最终生成适合于本文样本的指标评价体系具体如下所示。

|  |  |  |
| --- | --- | --- |
| **能力指标** | **权重** | **代表特征** |
| 营运能力 | 0.05 | 7天上新数量 |
| 获现能力 | 0.01 | 30天销售额 |
| 发展潜力 | 0.1 | 销售环比 |
| 偿债能力 | 0.1 | 运费险 |
| 履约状况 | 0.15 | 卖家信用 |
| 管理水平 | 0.05 | 服务态度 |
| 合同担保情况 | 0.05 | 保证金 |
| 职工素质与能力 | 0.15 | 是否金牌卖家 |
| 市场竞争能力 | 0.15 | 平均月销量 |
| 企业规模 | 0.1 | 宝贝数 |

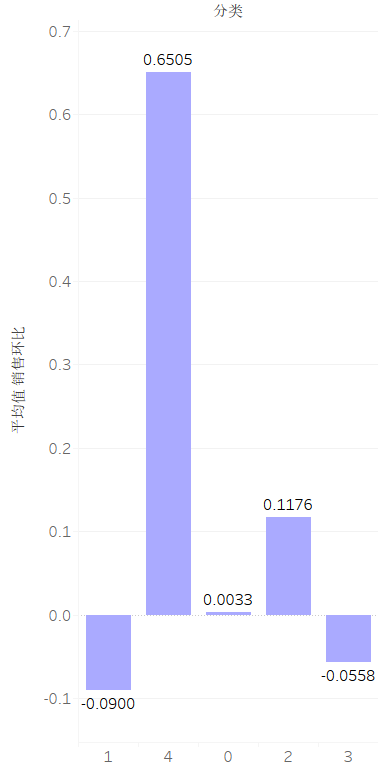
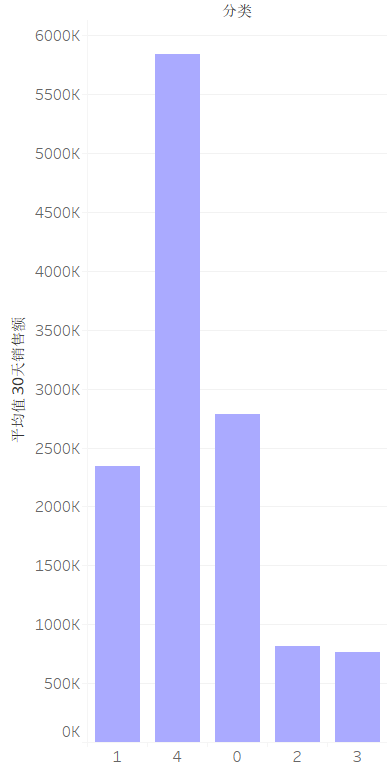
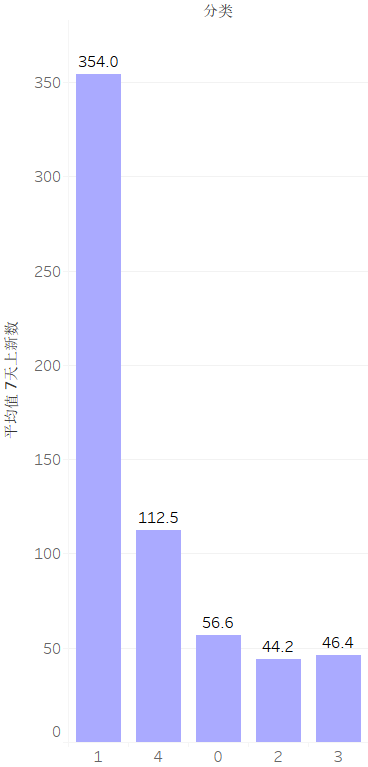
生成评分的时候，首先得到每个类的以上10个指标的平均值，将数值型的标准化，分类型的不做改变，将对应值乘以相应的能力权重，生成每个类的评分，该个评分代表了该类店家的信用好坏程度，具体如下所图所示：

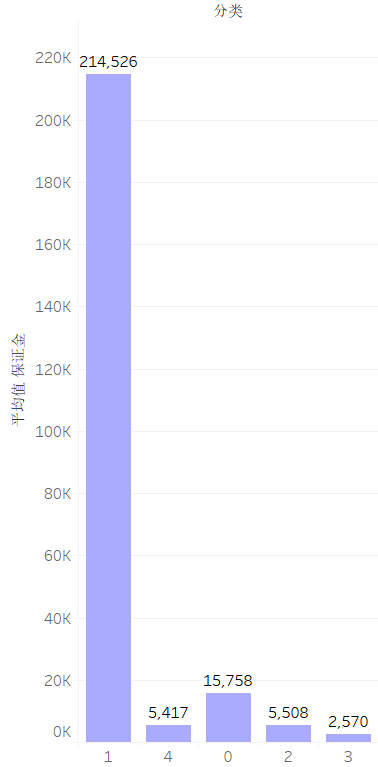
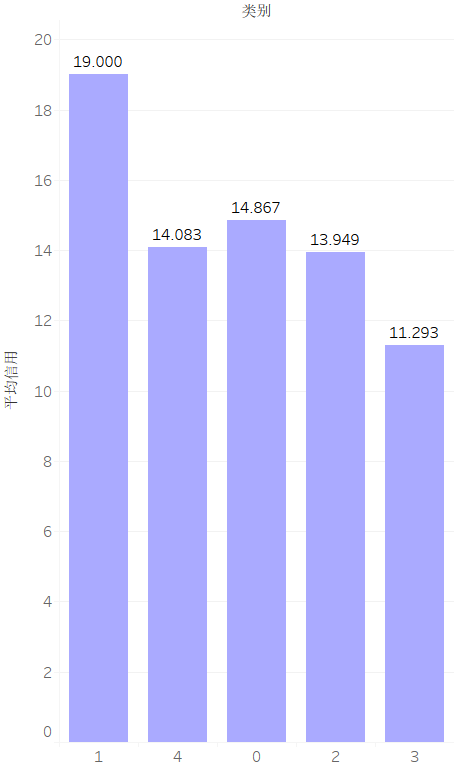
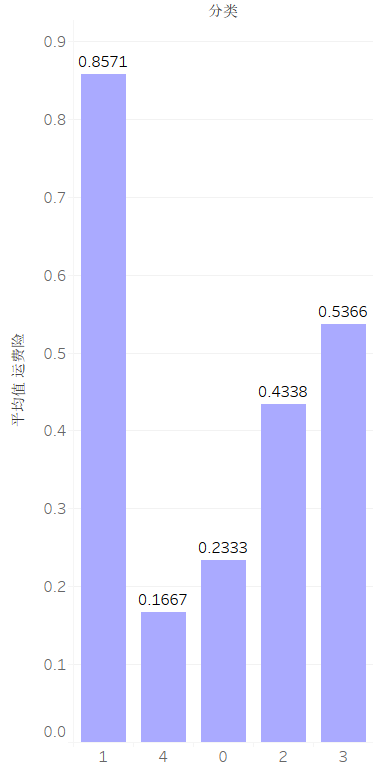


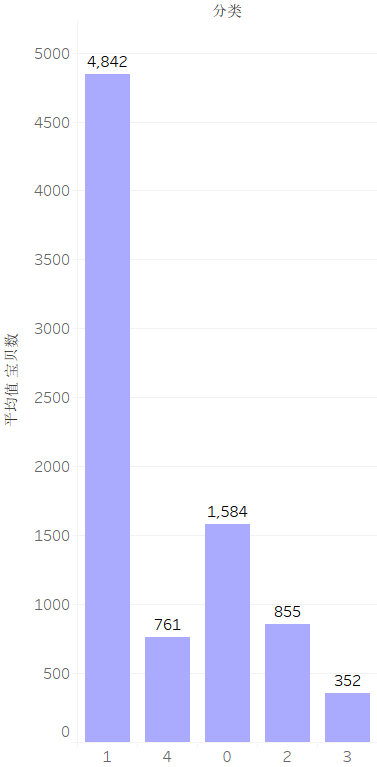
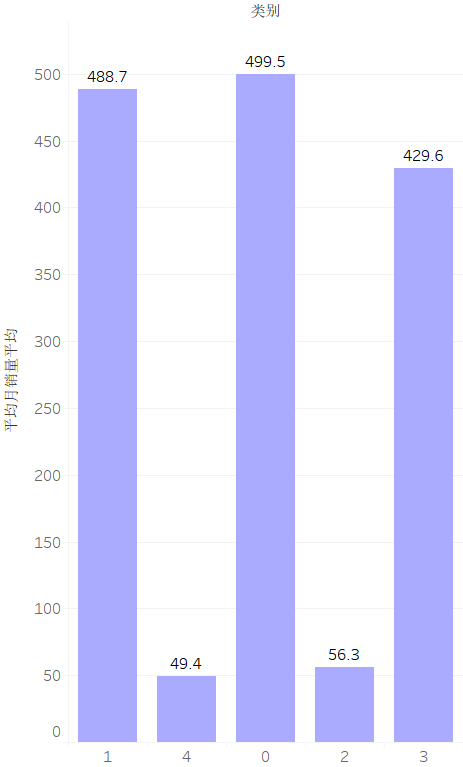
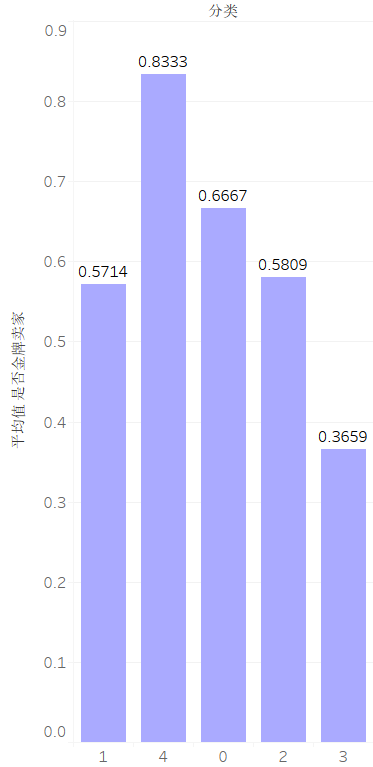
可以看到，第一类店家中的信用水平最佳，第二类次之，以此类推，第三类店家的信用状况最令人担忧，下图比较了类别1和类别3的指标平均值，可以看到，几乎类别1的所有指标都远远优于类别3，这也侧面证明了我们模型聚类效果的可靠性。



对于每个类别的其他信用评价指标的均值如下所示：

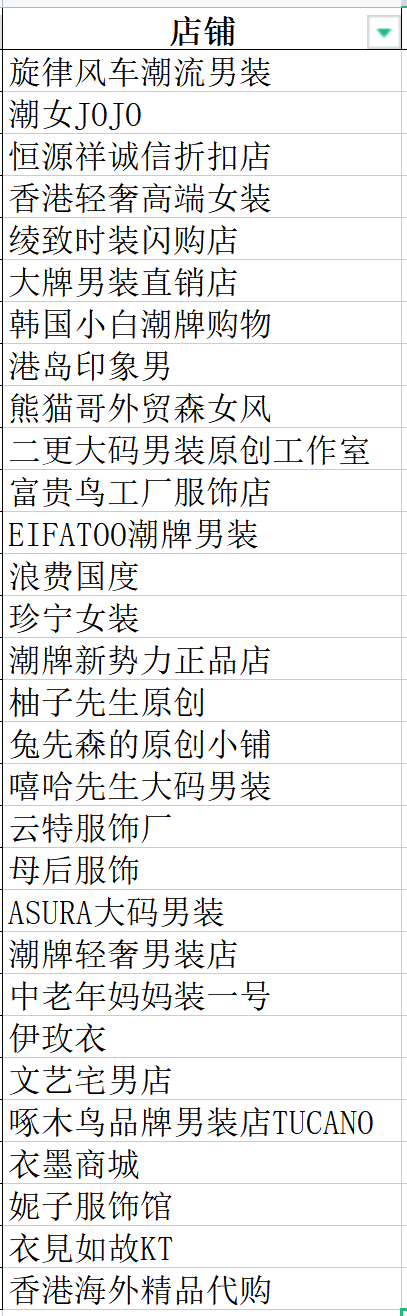
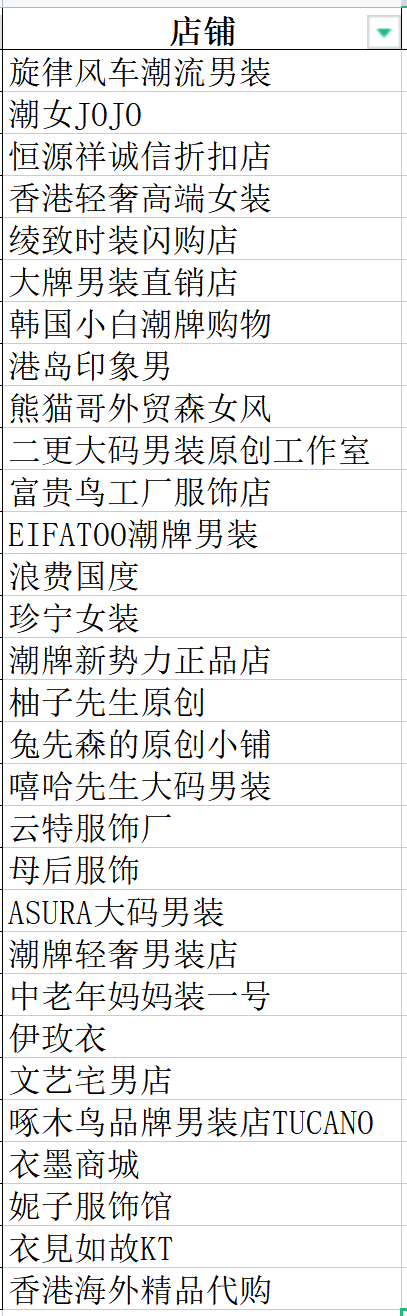




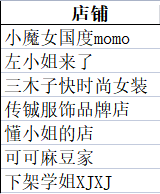


具体列别的店铺列表如下所示：

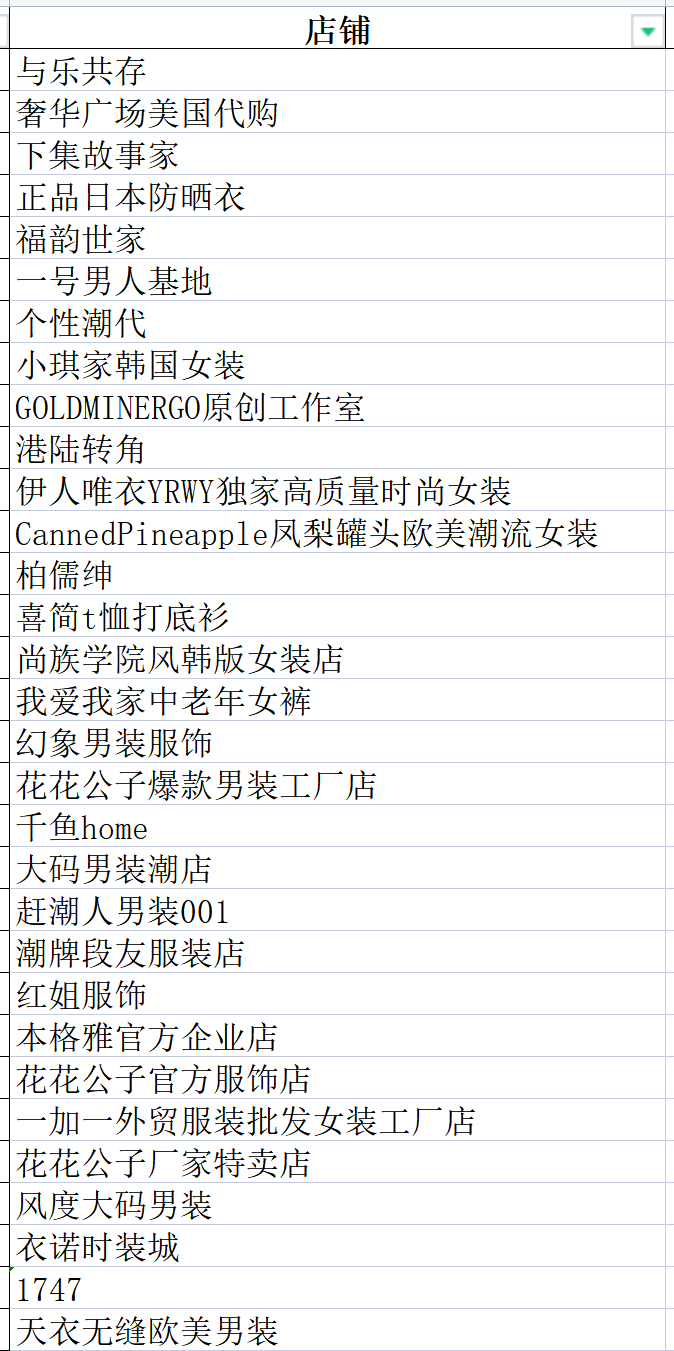
类别0：

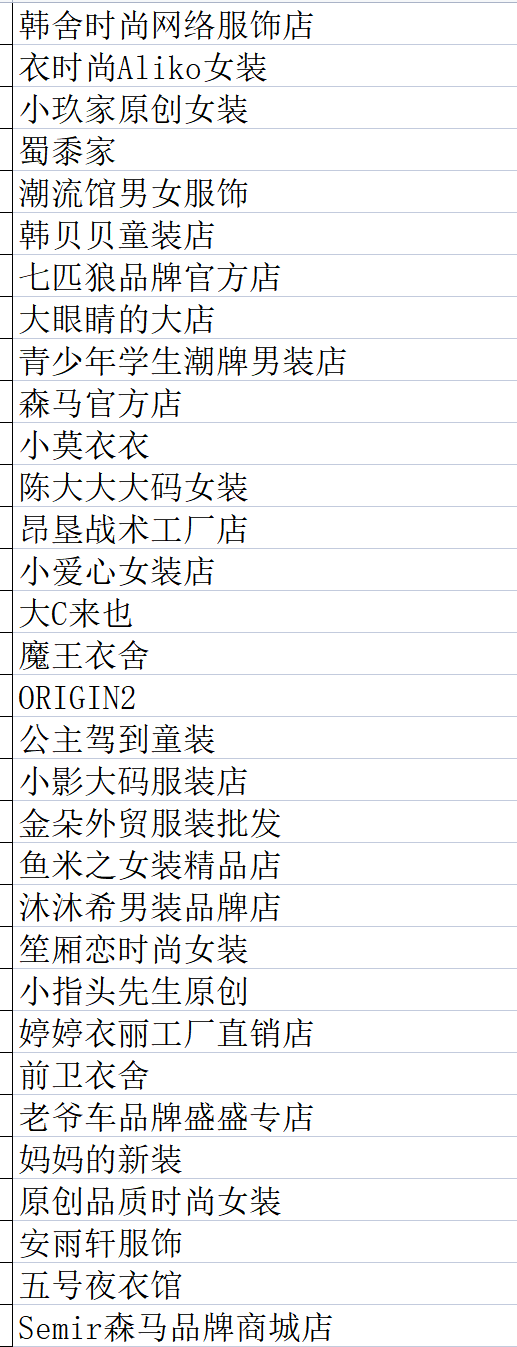
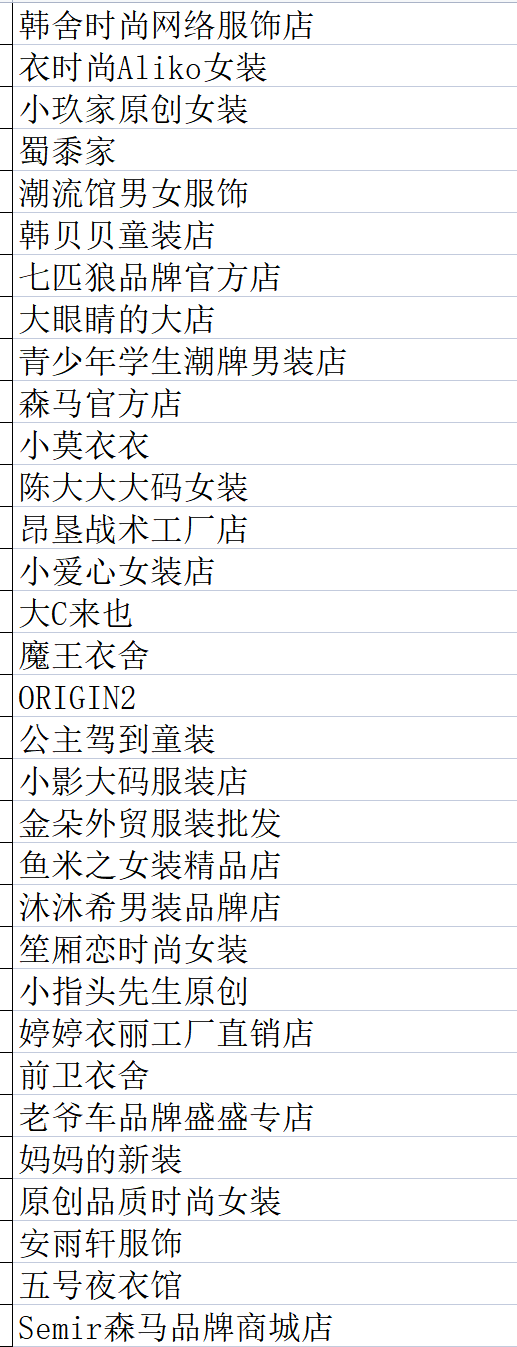
 

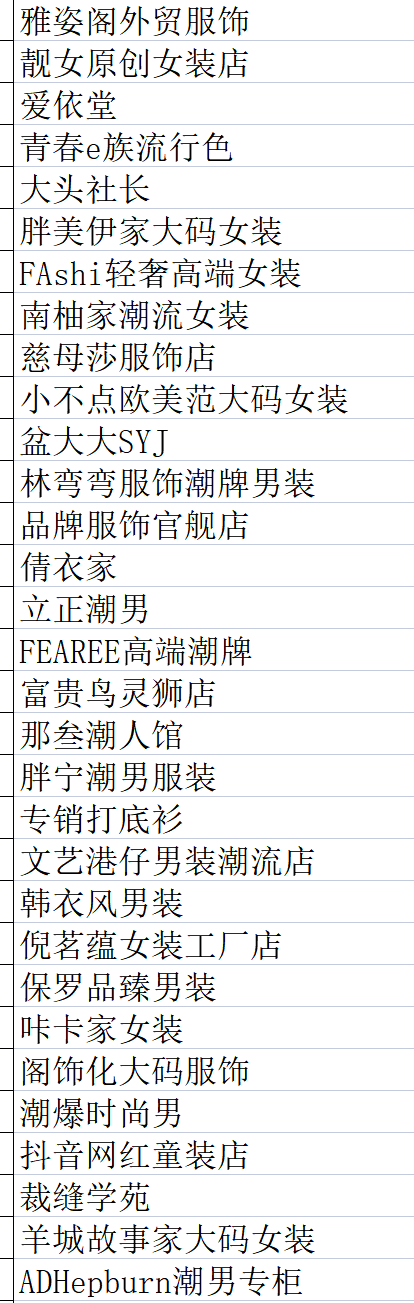
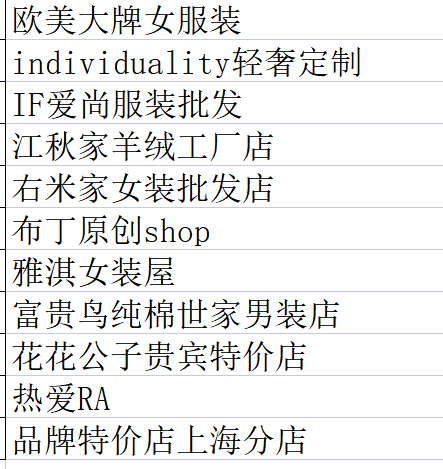
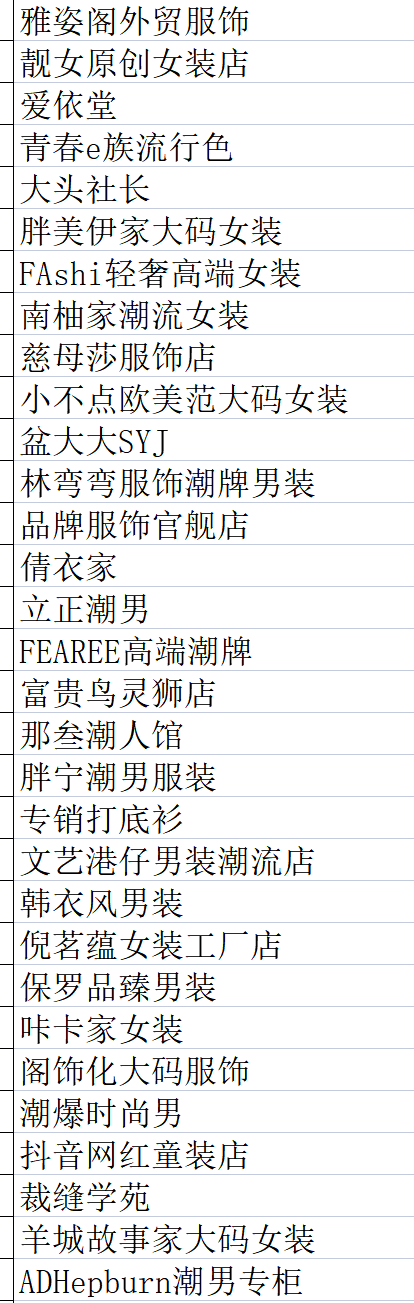
类别1：



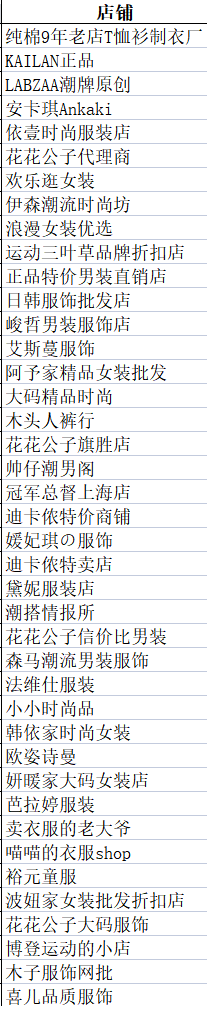
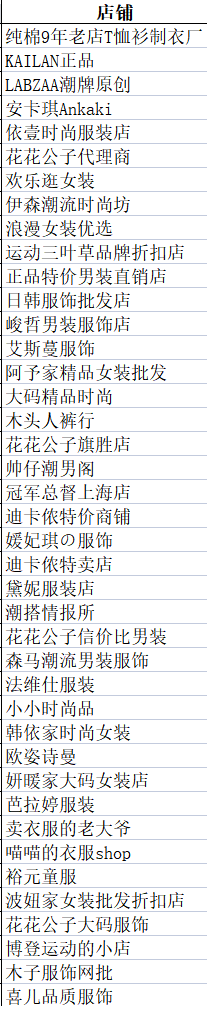
类别2：

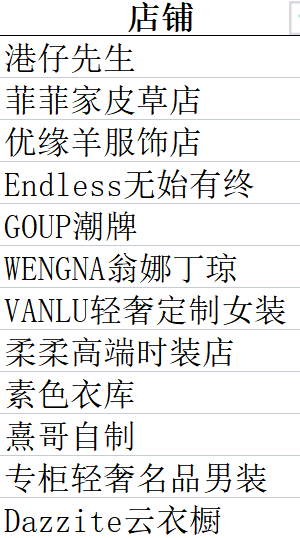
 

类别3：

类别4：



# 成员分工

|  |  |  |
| --- | --- | --- |
| 内容 | 具体内容 | 成员 |
| 数据获取 | 淘宝网店铺列表以及可获取的基础信息 | 吴泓栎 |
| 店查查店家基础信息+月销售整体值 | 孙智敏 |
| 店查查各店铺宝贝信息 | 王馨怡、杨子涵 |
| 数据预处理 | 整合数据 | 吴泓栎、孙智敏、杨子涵、王馨怡 |
| 特征工程构建 | 因子分析 | 孙智敏、杨子涵 |
| 特征提取 | 吴泓栎、王馨怡 |
| 模型构建 | 数据建模 | 吴泓栎、王馨怡 |
| 模型解释 | 王馨怡 |
| 模型结果可视化 | 吴泓栎 |
| 样本数据可视化 | 整体数据大屏显示 | 孙智敏、杨子涵 |

注：排名不分先后