数据挖掘分析报告

—— 基于批发商客户的聚类实验

学年学期：2022-2023-2

课程名称：数据挖掘

学号：42023017

姓名：常远

专业：金融数学实验班

日期：2023 年 6 月 25 日

第一章 概述

1. 背景介绍

对批发商的客户类型进行聚类，以便批发商更好的组织物流和服务。

2. 数据描述

本项目数据集来源于 UC Irvine，感谢来自葡萄牙的 Margarida Cardoso 予以捐赠。其中包含了葡萄牙调查当地批发商一年内不同商品的销售情况，此数据集包含了 441 条数据，每条数据包含 6 个数值属性和 2 个用数字表示的标称属性共 8 个字段属性，具体属性信息见**表 1。**

表 1 数据集属性信息表

序号	属性名称	含义
1	Chanel	客户的渠道-Horeca(酒店/餐厅/咖啡厅的简称)或零售渠道
2	Region	客户区域- Lisnon 或 Oporto 或其他区域
3	Fresh	新鲜产品的年度支出
4	Milk	奶制品的年度支出
5	Grocery	杂货产品的年度支出
6	Frozen	冷冻产品的年度支出
7	Detergents_Paper	洗涤剂及纸制品的年度支出
8	Delicassen	熟食产品的年度支出

报告内容：1、项目内容及意义 2、数据集描述 3、模型及实验过程 4、结果展示

第二章 探索性数据分析

1. 数据质量分析

- 特殊字符

  通过对数据的预览，数据集的列名和字段值不存在特殊符号。

- 缺失值

  数据集总体上缺失缺失属性很少且缺失比例很低，共有两个属性存

在缺失值，分别为第 232 行的“Grocery”和第 241 行的“Frozen”，缺失值计数条形图见**图 1**。

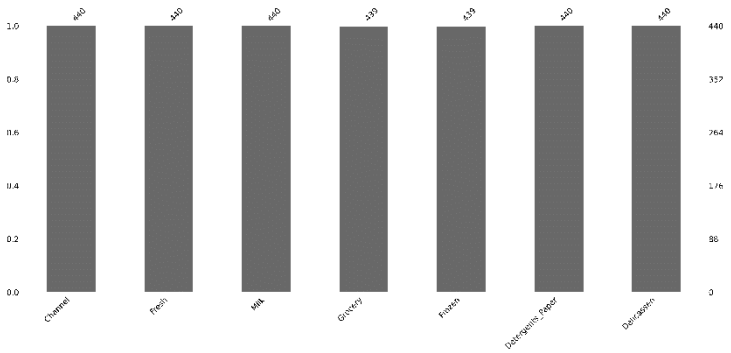


图 1 数据缺失值计数条形图

	Channel	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
232	1	25962	1780	NaN	638.0	284	834

图 2 查看 Grocery 缺失值的所在行

	Channel	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
241	1	24929	1801	2475.0	NaN	412	1047

图 3 查看 Frozen 缺失值的所在行

- 异常值

数据集数值属性的描述统计见**表 2** 以及箱形图见**图 4**。其中每个字段都存在一些数值较大或较小，差异比较大，但是可以视为合理的情况，本文进行两次实验，第一次将极远点视为异常值删除后做聚类，第二次视为处在合理范围之内做聚类实验。

表 2 数据集描述性统计表

	count	mean	std	min	25%	50%	75%	max
Channel	440.0	1.322727	0.468052	1.0	1.00	1.0	2.00	2.0
Fresh	440.0	12000.297727	12647.328865	3.0	3127.75	8504.0	16933.75	112151.0
Milk	440.0	5796.265909	7380.377175	55.0	1533.00	3627.0	7190.25	73498.0
Grocery	439.0	7960.646925	9511.970103	3.0	2151.00	4757.0	10665.50	92780.0
Frozen	439.0	3073.881549	4860.039567	25.0	740.50	1517.0	3559.50	60869.0
Detergents_Paper	440.0	2881.493182	4767.854448	3.0	256.75	816.5	3922.00	40827.0
Delicassen	440.0	1524.870455	2820.105937	3.0	408.25	965.5	1820.25	47943.0

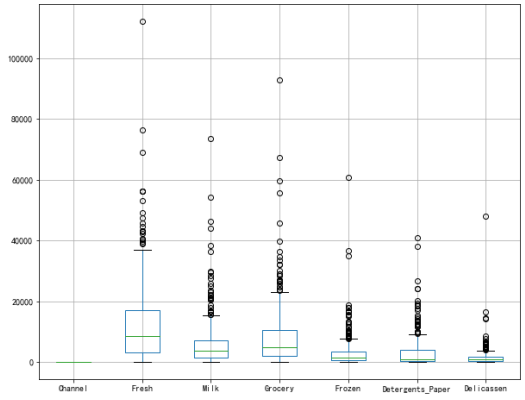


图 4 6 种产品的箱线图

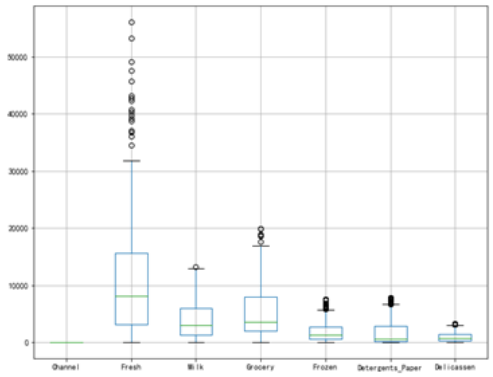


图 5 删除部分异常值后 6 种产品的箱线图

2. 数据特征分析

首先分析“Region”数据，分为三个地区，根据地区分类后对任意两个产品的关系画图。具体结果如下，关系图见**图 6。**

可以发现，不同地区在新鲜产品、奶制品、杂货产品、冷冻产品、洗涤剂及纸制品、熟食产品的年度支出的分布几乎相同，分布轨迹重合度非常高，则认为“Region”属性与聚类结果关系不大，我们可以删除该属性而不影响聚类实验结果。

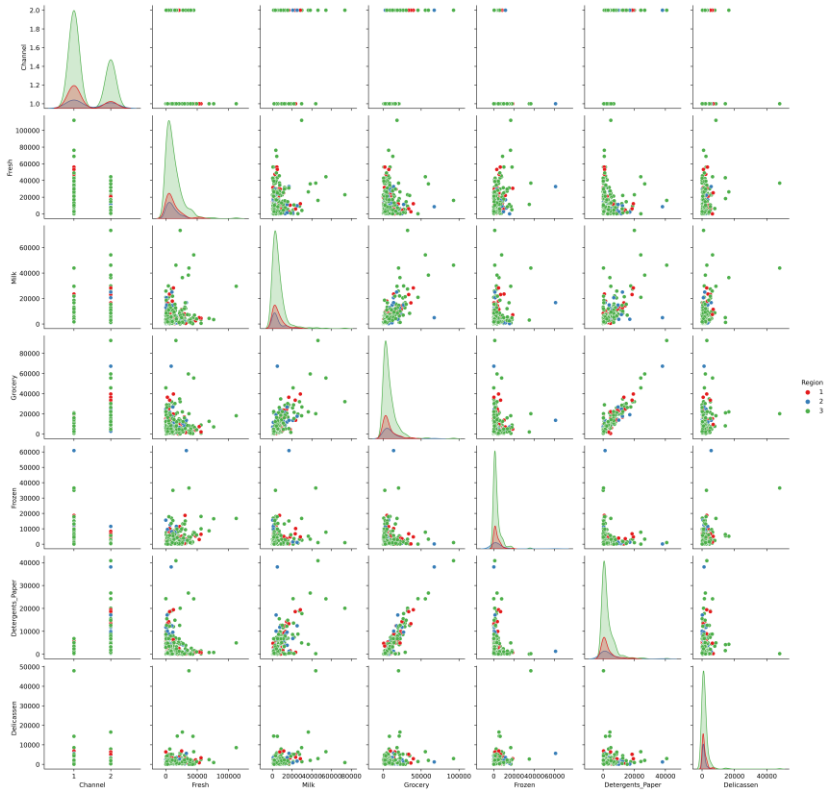
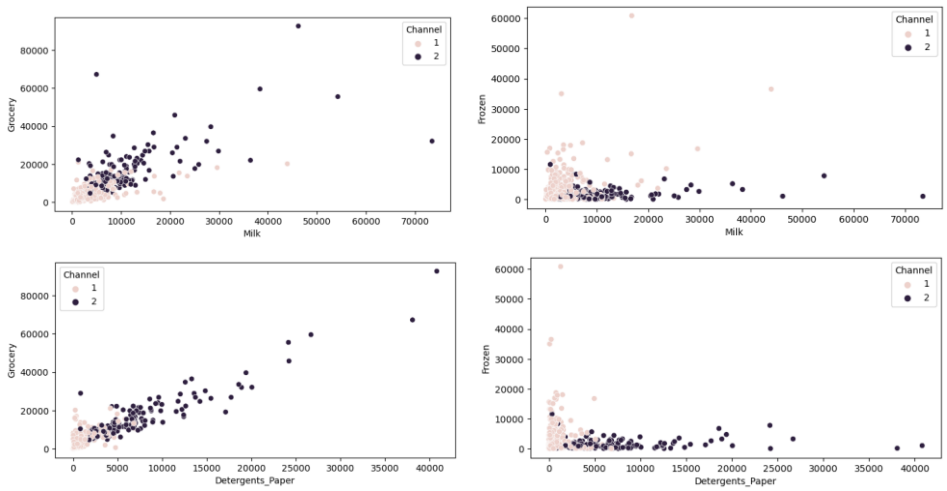


图 6 根据地区分类后其他产品的两两关系图

由数据类型可知，批发商的客户渠道分为两种，现在我们通过分析所有客户购买其他产品的关系在两个渠道占的权重来认知“Channel”。具体结果如下，相关系数图见**图 7**。

- 对于奶制品和冷冻产品的出售，一般为零售的客户买牛奶会更多。
- 对于杂货和奶制品的出售，零售的客户买的都较多。
- 对于洗涤剂及纸制品和杂货的出售，“Horeca”渠道的客户购买都很少，零售渠道的客户购买都较多。
- 对于新鲜产品和冷冻产品的出售，零售渠道的客户购买很少，“Horeca”渠道的客户购买非常多。



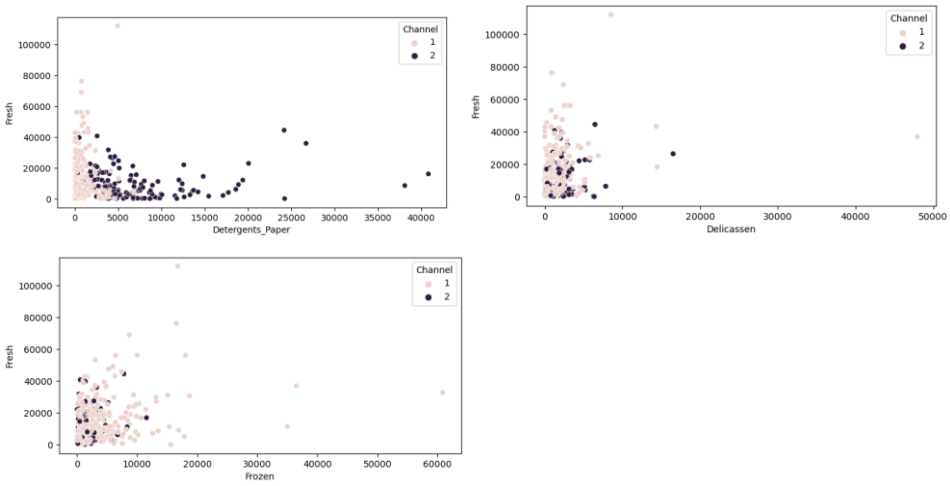


图 7 渠道分类后的两类产品关系图

该图更加明显的显示渠道 1 和渠道 2 的客户在不同产品中的占比情况，以及在不同年支出水平的占比。

- 对于新鲜产品和冷冻产品，可以直观看出“Channel1”无论在哪个年度支出水平都高于“Channel2”。
- 对于杂货和洗涤剂及纸制品，“Channel2”的年度支出更加高。
- 对于奶制品，“Channel1”的客户集中在年度支出较少一带，“Channel2”的客户集中在年度支出较大一带。

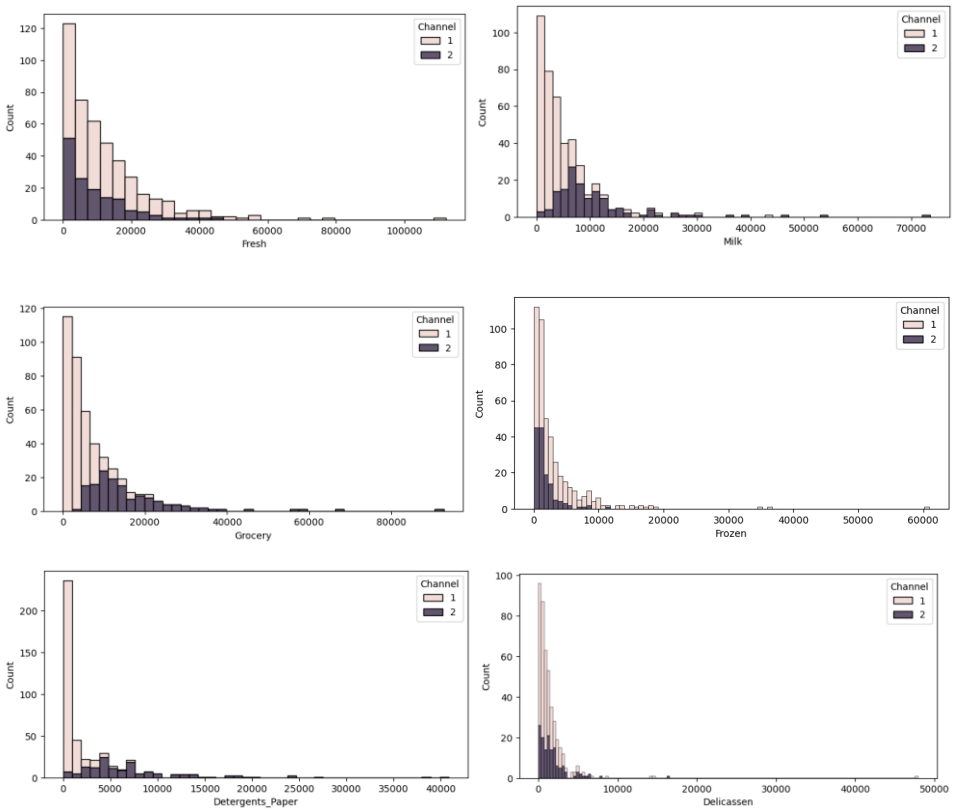


图 8 各类产品渠道分类后占比直方图

总体上看 6 种产品的数据分布较为平衡，可以比较直观的看出“Fresh”、“Milk”、“Grocery”的销量较高，且销售产量的广度较大，可以对比的是“Delicassen”的销售产量范围较小，批发商容易确定销售此产品的数量，该产品目标特征明确。产品数据本身的分布特征和结构见**图 9**。

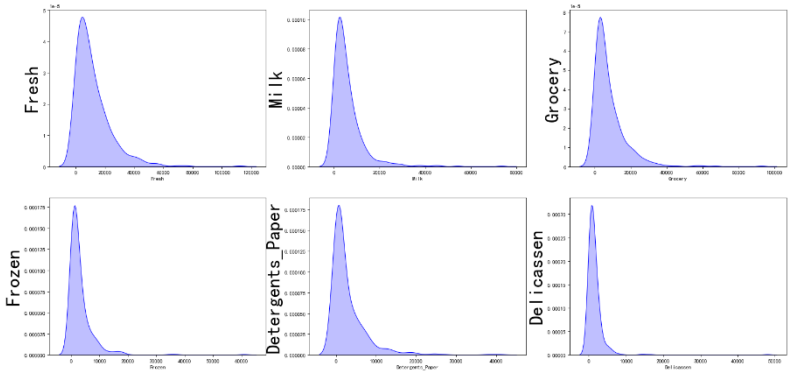


图 9 核密度图

3. 相关性分析

本实验对目标属性进行了二元编码，计算了变量间和变量与目标间的相关系数，具体结果如下，相关系数热力图见**图 10**。

- 奶制品和冷冻产品的销售有为微弱的正相关关系，相关系数为 0.12。
- 奶制品和杂货的销售有较强的正相关关系，相关系数为 0.73。
- 杂货和洗涤剂和纸制品的销售有很强的正相关关系，相关系数为 0.92。
- 洗涤剂及纸制品和冷冻产品的销售有微弱的负相关关系，相关系数为-0.13。
- 洗涤剂及纸制品和新鲜产品的销售有微弱的负相关关系，相关系数为-0.1。
- 熟食产品和新鲜产品的销售有较弱的正相关关系，相关系数为 0.24。
- 新鲜产品和冷冻产品的销售有适中的正相关关系，相关系数为 0.35。

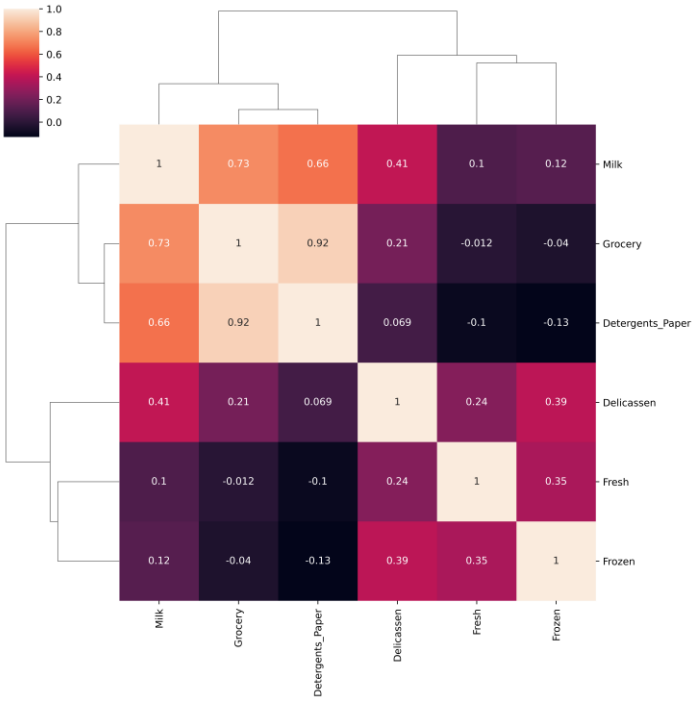


图 10 相关系数热力图

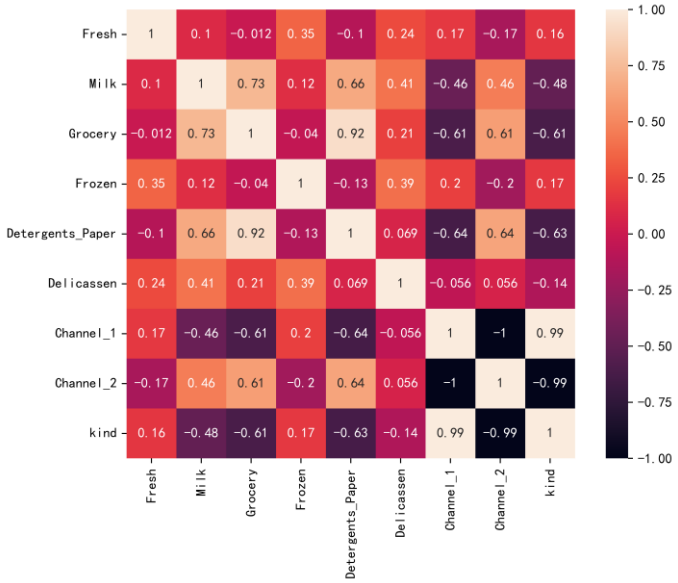


图 11 规范化及数值化的相关系数热力图

第三章 数据预处理

1. 数据清洗

- 特殊字符

本项目中没有特殊字符，不需要进行特别处理。

- 缺失值

在该数据集中共两个属性存在缺失，分别为第 232 行的“Grocery”和第 241 行的“Frozen”，由于缺失值较少，我们分别采用 Grocery 中位数、Frozen 中位数填充缺失值。

- 异常值

在数据质量分析中未发现不合理的异常值，故在此不做处理。

2. 特征变换

2.1 标称属性

客户的渠道具有标称特征，通过哑变量变换，转换为二元数值特征类型，客户渠道为“Horeca”时，设为(Channel1,Channel2)=(1,0);客户渠道为零售时，设为(Channel1,Channel2)=(0,1)。

2.2 数值属性

在全部转换为数值属性后，进行标准化，这里由于数据集的标准差较大，极端值较多，对于极端值和异常值敏感的最小-最大规范化不适用，且该数据集特征大致符合正态分布，则我们采用零-均值规范化，按照均值中心化，再利用标准差重新缩放数据，使数据服从均值为 0，方差为 1 的正态分布。

3. 特征提取

在该文中，有关特征提取方法采用维度约束中的降维方法，通过线性或非线性变换方法，将数据投影到低维空间中，从而减少数据特征个数。本文采用 t-SNE 方法降维，把高纬度的数据点之间的距离转化为高斯分布概率，保留一定的局部特征。

表 3 数据处理后的属性信息表

属性	属性名称
Fresh	新鲜产品的年度支出
Milk	奶制品的年度支出
Grocery	杂货产品的年度支出
Frozen	冷冻产品的年度支出
Detergents_Paper	洗涤剂及纸制品的年度支出
Delicassen	熟食产品的年度支出
Fresh	新鲜产品的年度支出
Channel 1	客户的渠道-Horeca(酒店/餐厅/咖啡厅的简称)
Channel 2	客户的渠道-零售渠道

第四章 模型训练

本问对数据集进行聚类实验，分别尝试用 K-means 算法和 DBSCAN 算法进行实验并进行比较。

1. K-means

K-means 算法是基于划分的聚类算法，核心思想是将 n 个数据对象划分到 k 个簇中，使得每个对象到其所属簇中心的距离平方和最小。K-means 算法存在一定的局限性，如提前设定聚类数量，对初始簇中心敏感，容易陷入局部最优解等。

该批发商客户数据集属于凸数据集，且该数据集的异常点非常少，同时我们使用 TSNE 方法降维，该方法将数据变为高斯分布的形式，K-means 算法在处理此分布时效果非常不错。

2. DBSCAN

DBSCAN 是基于密度的聚类算法。其核心思想是基于样本点之间的密度分布来划分聚类簇，同时能够识别出噪声点，在我们对将降维后的数据进行 DBSCAN 聚类时，我们可以发现噪声点只有一个。但是当数据集密度差异较大时，选择合适的参数（邻域半径和最小密度阈值）比较困难。

对于 K-means 算法，容易受到异常值的影响。而在该实验中，我们观测箱线图无法判断部分极大值是否是异常值，而 DBSCAN 不要求指定集群的数量，避免了异常值，所以我们采用该算法排除异常值的困扰，同时验证 K-means 算法的正确性。

第五章 模型讨论

1. 箱线图异常值处理前后比较

2.1 异常值处理后

在异常值处理后，我们可以看到当 K<2 时,曲线极速下降；当 K>2 时，曲线趋于平缓，通过手肘法我们判定拐点 2 为 K 的最佳值。

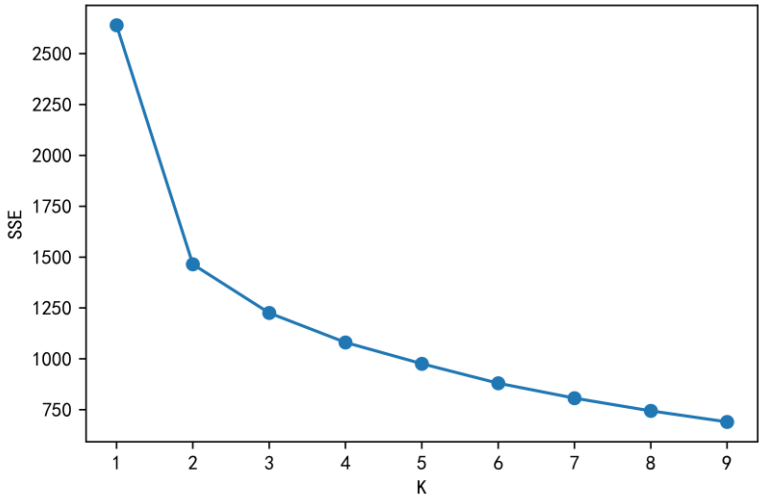


图 12 手肘原则确定 K 值图

之后我们根据 K-means 算法得出各类别的聚类中心。

表 4 各类别的聚类中心及类别数目表

序号	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen	Channel_1	Channel_2	类别数目
0	0.0612	-0.3820	-0.4411	0.1235	-0.4857	-0.1289	0.5843	-0.5843	242
1	-0.1681	1.0505	1.2130	-0.3397	1.3358	0.3545	-1.6069	1.6069	88

由表 4 可知，在异常值处理后，剩下的 330 个数据分为两个簇，分别为 242 个和 88 个。第一个簇表示批发商可将新鲜产品、冷冻产品一块进货，一般为 Horeca 渠道购买比较多；第二个簇表示批发商可将奶制品、杂货、洗涤剂及纸制品、熟食品一块进货，一般零售渠道购买比较多。

根据 K-means 算法得出聚类图，可以看出聚类 2 的簇的聚集效果比较好，但是聚类 1 的簇比较分散。

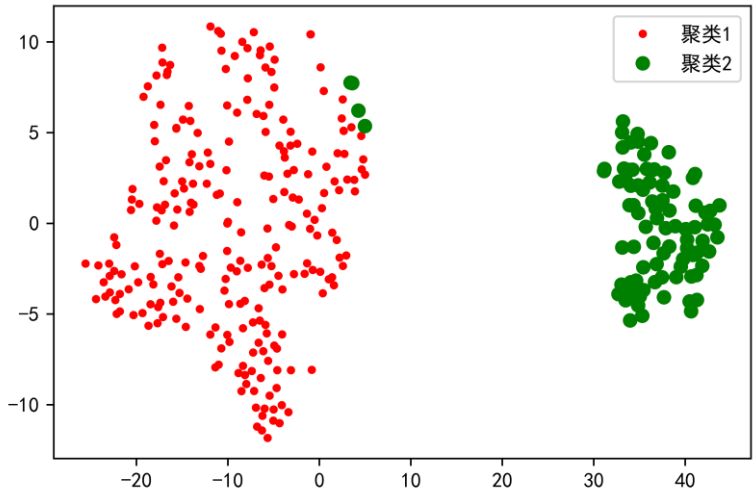


图 13 降维后 K-means 聚类图

2.2 没有视为异常值

当没有做异常值处理时，我们可以看到当 K<2 时,曲线极速下降；当 K>2 时，曲线趋于平缓，通过手肘法我们判定拐点 2 为 K 的最佳值。

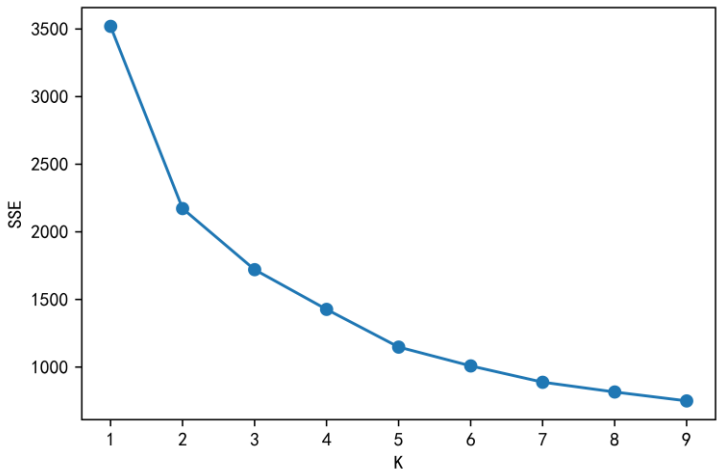


图 14 手肘原则确定 K 值图

之后我们根据 K-means 算法得出各类别的聚类中心。根据数据我们可以将其分为两类作为批发商判断客户的依据，第一类为 Horeca 渠道类型，购买新鲜产品、冷冻产品较多，且 Horeca 渠道类型客户更多，为 297 家客户，几乎为零售渠道客户的 2 倍；第二类为零售渠道类型，购买奶制品、杂货产品、 洗涤剂及纸制品、熟食产品较多，但是购买的客户较少。

表 5 各类别的聚类中心及类别数目表

序号	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen	Channel_1	Channel_2	类别数目
0	0.1105	-0.3365	-0.4259	0.1165	-0.4387	-0.0943	0.6903	-0.6903	297
1	-0.2296	0.6989	0.8846	-0.2420	0.9111	0.1958	-1.4337	1.4337	143

由表 5 可知，没有进行异常值处理，共分为两个簇，分别为 297 个和 143 个。第一个簇表示批发商可将新鲜产品、冷冻产品一块进货，一般为 Horeca 渠道购买比较多；第二个簇表示批发商可将奶制品、杂货、洗涤剂及纸制品、熟食品一块进货，一般零售渠道购买比较多。与表 4 相比，簇的分类是相同的，但是奶制品、杂货和洗涤剂及纸制品的聚类中心均低于进行异常值处理后的数据。

根据 K-means 算法得出聚类图，可以明显看出在没有删除较大数值时，聚类产生的簇更加集中，聚类效果更加明显，所以我们选择在之后的实验中不将极大值视为异常值，保留数据的完整性。

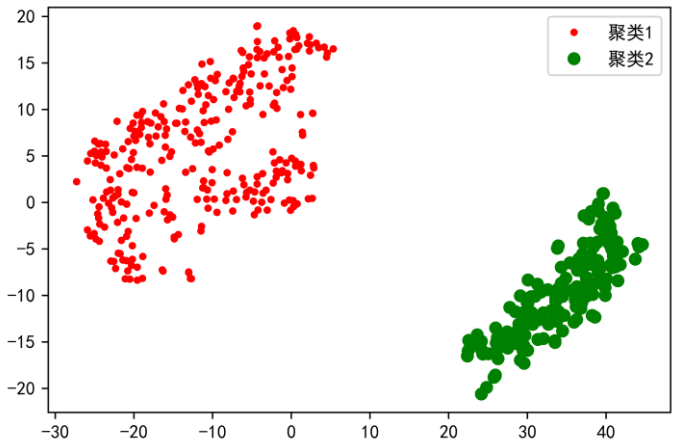


图 15 降维后 K-means 聚类图

2. K-means 与 DBSCAN 模型比较

由上一问可知，在箱线图中距离较远的点不应该视为异常值，没有删除较大较小值的聚类实验结果较好，接下来的实验皆选用所给数据集整体。

2.1 K-means

从图 16 可以看出，基于 Median，客户群 0 对于奶制品、杂货产品、洗涤剂及纸制品的反应很强，对于熟食产品有一个中等的反应；客户群 1 则是对新鲜产品和冷冻产品的反应更强。

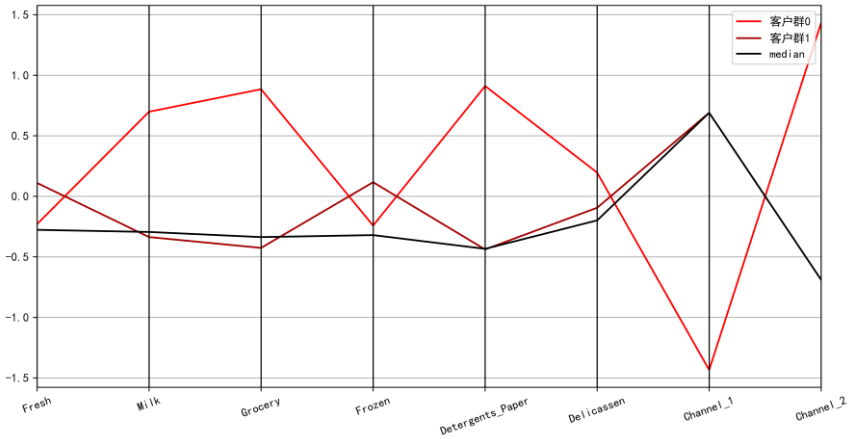


图 16 K-means 聚类平行坐标系图

从图 17 可以看出，

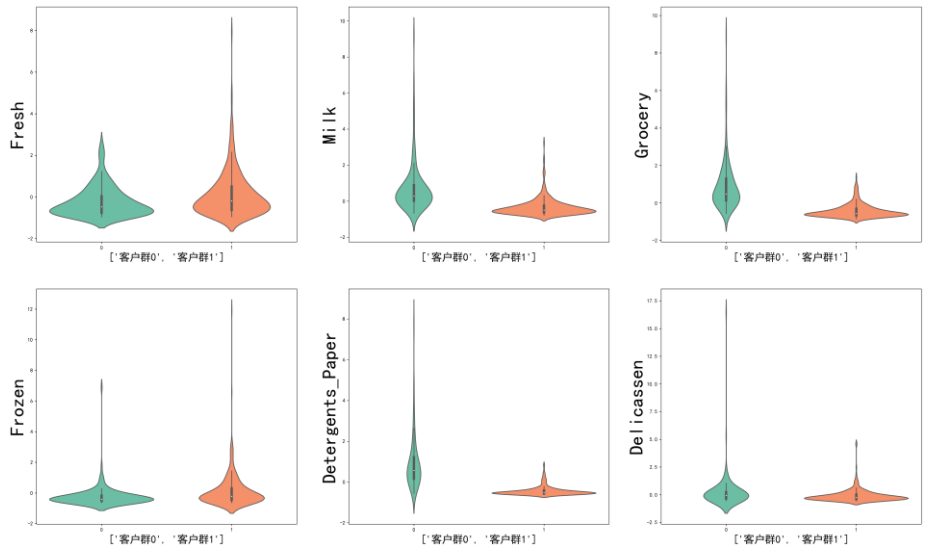


图 17 K-means 聚类小提琴图

从图 18 可以看出，数据的密度分布特征具有正态中的偏态现象，同时我们可以看出在新鲜产品上，Horeca 渠道是主要的购买方，但是零售渠道对于批发商也有很大的市场，所以，批发商如果更加迎合零售渠道的客人，可以适当增加新鲜产品进货；对于洗涤剂及纸制品，我们可以看出 Horeca 渠道的数据密度非常小，所以对于客户主要来源与 Horeca 渠道的批发商，尽量不要进货洗涤剂及纸制品。

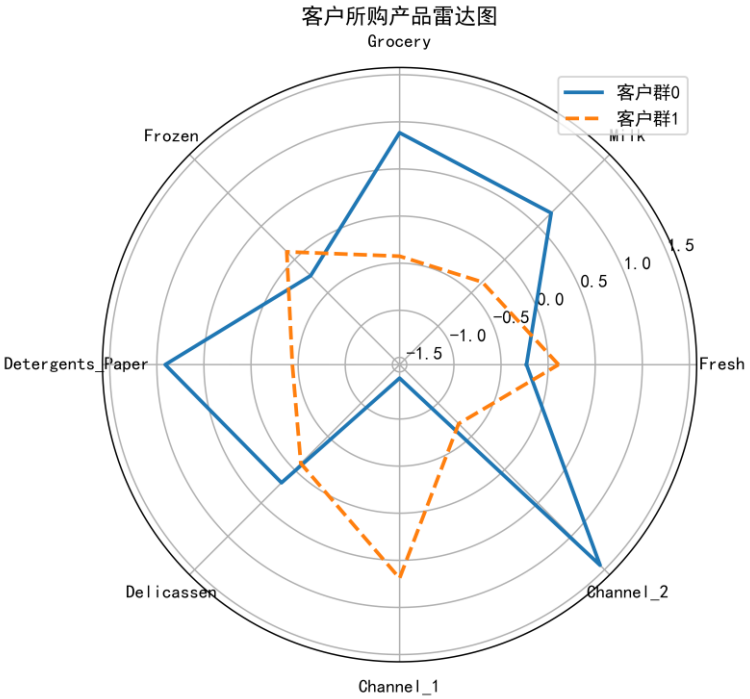


图 18 K-means 聚类客户所购产品雷达图

从图 19 可以看出，对于客户群 0，新鲜产品和冷冻产品的波动更加明显。对于客户群 1，奶制品、杂货和洗涤剂及纸制品的波动非常剧烈。

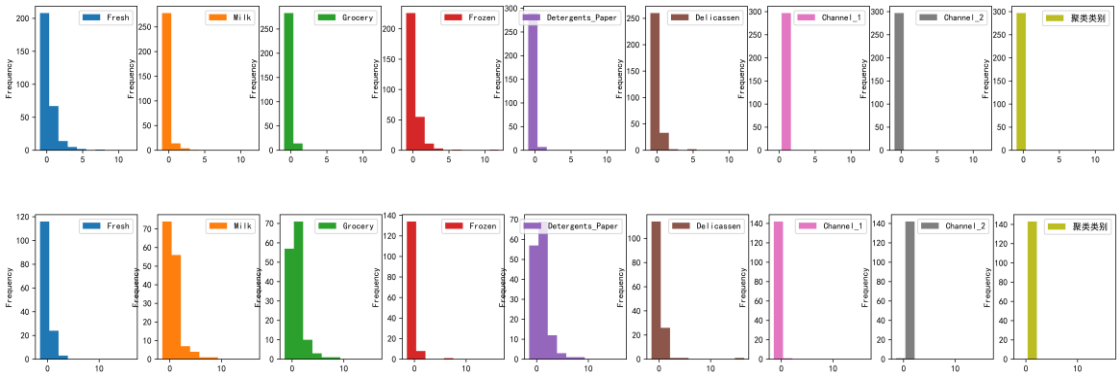


图 19 K-means 聚类结果图

2.2 DBSCAN

这里我们采用离群点百分比的大小来决定 epsilon 值，同样是选择离群点百分比拐点的位置。在这里我们选择 epsilon=2 作为最优点。

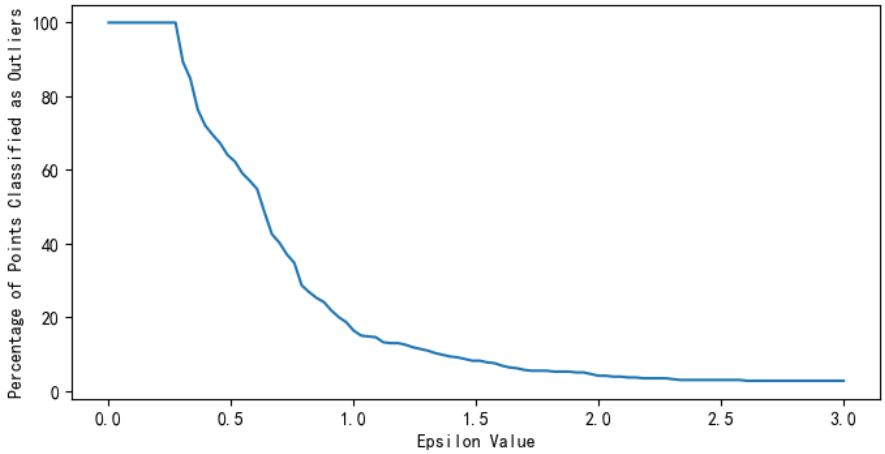
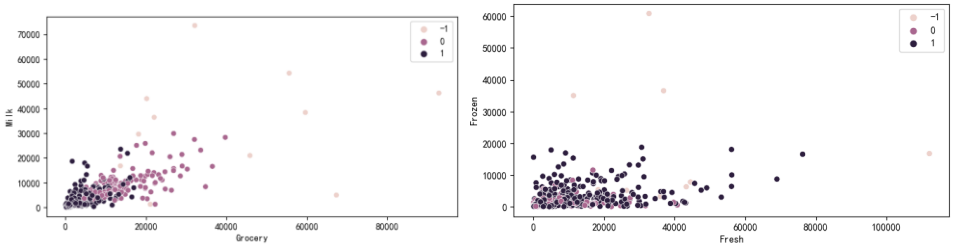


图 20 未降维的离群点百分比与 epsilon 值选择的线形图

图 21 为 DBSCAN 聚类的结果图，分析了多种产品的聚类关系，比如说通过杂货和牛奶可以分为两类，但是这两类分离度较差，且离散点较多，通过这些图聚类的结果不太好，这也与 DBSCAN 聚类维数不能太多有关，之后我们对数据进行降维。



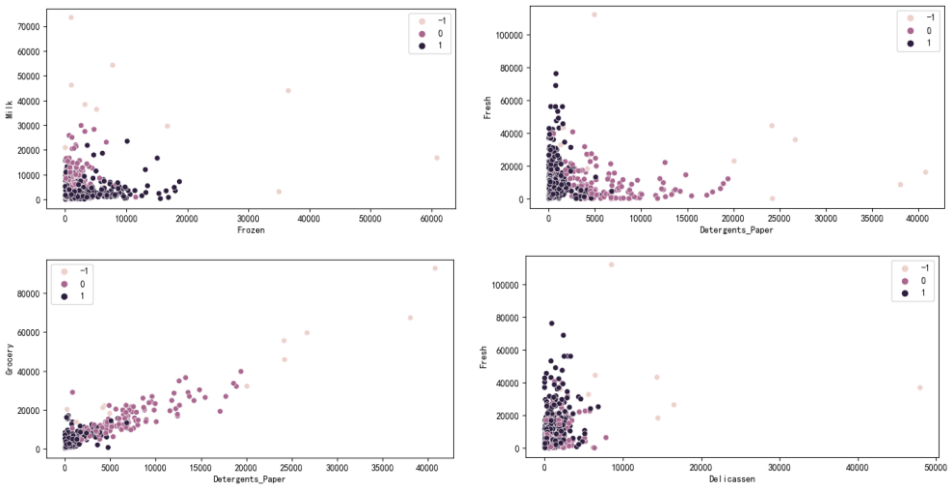


图 21 DBSCAN 聚类分布图

降维后的数据离群点百分比更低，同样是当 epsilon=0.2 时处于拐点，但是离群点百分比已经接近于 0。

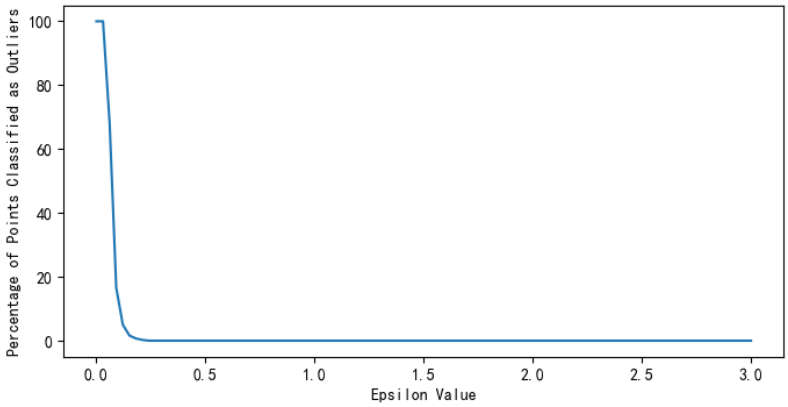


图 22 降维后的离群点百分比与 epsilon 值选择的线形图

图 23 为 DBSCAN 聚类图，可以看出明显的分为两类，而且离群点只有一个，该聚类方法的效果非常好。

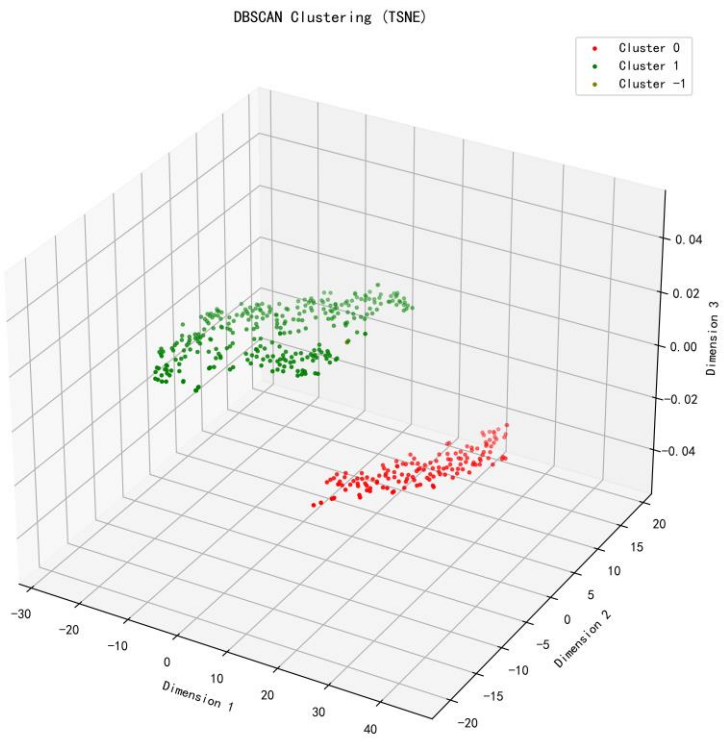


图 23 DBSCAN 聚类分布三维散点图

表 6 为 DBSCAN 聚类得出的统计平均值，由统计平均值可分为两个簇，第一个簇表示批发商可将新鲜产品、冷冻产品一块进货，一般为 Horeca 渠道购买比较多；第二个簇表示批发商可将奶制品、杂货、洗涤剂及纸制品、熟食品一块进货，一般零售渠道购买比较多。我们通过比较该表和表 5 可知，K-means 聚类中心的值和 DBSCAN 聚类平均值非常近似，结合聚类图直观判断，可以得出 K-means 和 DBSCAN 聚类算法效果都比较好。

表 6 集群和离群值在不同产品上支出金额的统计平均值表

序号	Fresh	Milk	Grocery	Frozen	Detergents_ Paper	Delicassen	Channel_1	Channel_2
1	0.0697	-0.3547	-0.4416	0.0416	-0.4445	-0.1422	0.6903	-0.6903
0	-0.2991	0.4668	0.6781	-0.3027	0.7230	0.0306	-1.4487	1.4487
-1	1.5409	3.1207	2.8769	2.2095	2.4755	2.8752	-0.4614	0.4614

第六章 总结

本项目基于葡萄牙批发商客户数据进行了数据挖掘，实验过程中尝试了多种数据预处理方法，比较了数据集在 K-means 聚类实验和 DBSCAN 聚类实验下效果表现，最后对两个模型进行了模型解释和讨论。

总体上项目中模型均表现出较佳的性能，但是没有深挖出数据更深层次的含义，对于其隐含特征没有深入特征处理。

通过聚类算法，将客户群体按照不同购买行为划分为 2 个簇，从而更好地了解各个客户群体的特点，以及他们在不同商品类别上的消费偏好。有助于针对不同客户群体提供更加个性化的服务和优惠策略。通过比较 K-means 聚类和 DBSCAN 聚类的结果，可以进一步了解并挖掘客户购买行为的规律。比如在第一个簇中，新鲜产品和冷冻产品的购买量普遍较高，其中 Horeca 渠道占大多数，可能代表这类客户偏爱这两类产品，或许是生活节奏较快、注重工作效率的群体。分析这些规律有助于批发商更准确地把握市场脉络，调整销售策略。

同时，通过聚类实验，我们可以识别出哪些商品类别之间存在较高的关联度。例如，某个客户群体可能在购买杂货时，同时还会购买洗涤剂及纸制品。了解这些关联关系可以帮助批发商针对性地进行捆绑销售、组合推荐等活动，提高销售额。

对于离散值，DBSCAN 聚类算法具有较好的异常值检测功能。我们可以找出那些在购买行为上与其他客户明显不同的客户，这些客户可能是潜在的大客户、恶意退货者或者存在其他特殊情况的客户。针对这些客户，批发商可以采取不同策略对待。比如我们在降维后聚类发现有一个点为离散值，该值需要特别讨论。

总之，通过使用 K-means 聚类和 DBSCAN 聚类算法分析批发商客户数据，我们可以更深入地了解客户购买行为，挖掘潜在的市场机会，并有效指导批发商优化销售策略。