哈爾濱工業大學(深圳)

HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

数学建模课程论文

题　　目: 自动定价与补货策略的探索与分析

学　　院: 理学院

专　　业: 数据科学与大数据

成　　员: 莫力炬　210810419

吴阳诚　210810507

姜欣妍　210810308

2024 年 6 月 16 日

小 组 分 工 与 题 目 来 源

莫力炬 **210810419**：负责问题一的分析内容及论文部分，完成各种数据处理工作。

姜欣妍 **210810308**：负责问题二中的相关性分析、拟合与预测，并完成相关的论文内容。

吴阳诚 **210810507**：负责整体框架设计，以及问题二、问题三中的优化模型的建立与求解，并完成相应的论文内容。

题目来源：2023 年全国大学生数学建模竞赛 C 题

摘　要

本文目的是分析历史数据的规律，从而建立蔬菜类商品的自动定价与补货决策。

【背景】生鲜商超需要结合历史数据，得到市场需求及进货价格的规律，从而在不确切知道具体单品和进货价格的情况下，做出当日各蔬菜品类的补货决策，并确定“成本加成定价”，最终达到总利润最大化的目的。

【方法】本文主要采用了可视化分析、k-means 聚类、GNU 模型、0-1 规划等方法，从高维度、与时间高度相关的历史流水数据中分析规律，并结合供给侧和需求侧进行逻辑分析，最终构建模型最大化总利润。

【分析过程】问题一：首先对历史数据进行全面的预处理，再进行可视化探索性分析以及统计检验，并使用通过 kMeans 方法，探究数据中各品类及单品的分布关系，并得到“互补性”“替代性”等相关关系。

问题二：第一小问中，针对题设的目标时间段，采用 sigmoid 模型等探索标准化后的成本加成定价与销量之间的关系。另外，使用包含反比例的 MLR 多元线性回归模型，并采用放缩系数，得到各品类的销售总量与成本加成定价的约束关系，用于后续的优化模型。

第二小问中，首先通过 GNU 模型预测出各品类未来一周的销量以及进货价，再结合第一小问得到的销售总量与成本加成定价的约束关系，最终构建非线性规划模型，使用 scipy.optimize 的 minimize 模块得出总利润最大化的各品类定价与补货策略。

问题三：筛选出符合供给需求的单品集合后，类似地完成相关信息的预测，再设置逻辑变量，结合题设约束构建 0-1 规划模型或多目标规划模型，使用问题二中类似方法求解，得出最大化总利润的单品补货量和定价策略。

【评价与改进】由于时间关系，本文并未处理原题的问题四，同时问题三的优化模型也没有得到完全的代码结果，有待未来的进一步探究。

关 键 词：自动定价与补货策略，可视化分析，kMeans，GNU，非线性规划，0-1 规划

目 录

1 问题重述与分析

1.1 问题背景

在现代零售业中，蔬菜类商品由于其易腐性和需求波动，如何有效管理其定价和库存，直接关系到企业的盈利能力和运营效率。本赛题聚焦于蔬菜类商品的自动定价与补货决策，蔬菜的进货交易时间通常在凌晨 3:00-4:00，为此商家须在不确切知道具体单品和进货价格的情况下，做出当日各蔬菜品类的补货决策。传统的经验定价和人工补货方法已经难以适应日益复杂的市场环境，因此，利用历史数据进行深入分析、建立科学的预测模型、制定优化的定价和补货策略，显得尤为重要。

1.2 问题重述

本题所用的数据包括来自 2020 年 7 月到 2023 年 6 月的历史销售流水数据，一共包含来自 6 个商品大类的约 250 种单品及其各种销售信息，我们需要利用这些数据来求解以下的问题。

问题一要求对各蔬菜品类及其单品的销量和价格进行详细的统计分析，旨在发掘不同品类或不同单品之间潜在的关联关系。具体来说，我们需要从历史数据中提取和清洗信息，在品类层面，可以通过分析品类蔬菜的销量分布情况得出品类间的相关性及其分布规律。在单品层面，需要进一步细化分析不同单品的销量分布情况。通过聚类分析方法，可以将销量特征相似的单品归类，简化后续的建模和预测工作。问题一的目的在于通过对历史数据的深入分析，全面了解各品类及单品的销量和价格变化规律，为后续的预测和决策奠定基础。

问题二要求对未来的销量和价格进行预测，基于问题一的分析结果，选择适当的模型对销量和成本加成定价之间的关系进行建模。针对销量这一变量，我们预计采用多元线性回归等模型进行尝试。通过比较不同模型的预测性能指标，选出最优模型。接下来基于预测的结果，我们将构建以利润最大化为目标的优化问题并求解。问题二的目的在于通过构建和验证合适的销量和价格预测模型。

问题三在问题一和问题二的基础上，要求制定一个结合定价和补货的综合优化策略，以实现利润最大化和成本最小化的目标。这个问题相较于问题二更加综合，因为决策层次从品类进一步细化到每个单品。需要满足的限制条件也更为复杂:: 可售单品总数限制在 27-33 个; 每个单品最低订购量不少于 2.5 千克，这些属于硬性约束。同时，还需要考虑近期商品的市场需求，还需要判断每个单品在目标时期是否是有供应的。

1.3 问题分析

1.3.1 问题一：销量和价格变化规律分析

在分析蔬菜类商品的销量和价格变化规律时，首先需要对给定的历史数据进行全面的预处理。原始数据包含噪声、缺失值和异常值，因此数据清洗是数据分析的第一步。完成数据清洗后，接下来是对各蔬菜品类的销量进行分布分析。销量数据通常具有明显的季节性特征和趋势性变化，首先按日对数据进行分组和探索性数据分析（EDA），通过绘制直方图、箱线图等图表，直观展示各品类的销量分布情况。

在品类分析的基础上，进一步细化到单品层面的销量分析是必要的。不同单品的销量可能受到多种因素影响，例如节假日促销活动、天气变化等。通过对单品销量数据进行聚类分析，可以将具有相似销量特征的单品归为一类，从而简化后续的建模和预测工作。此外，从其他的粒度（月份、季度、年度）进行规律探索也很有价值。

1.3.2 问题二：基于品类的总销量与定价关系探究以及补货和定价策略制定

本题要求以品类为单位做补货计划，并给出各蔬菜品类未来一周 (2023 年 7 月 1-7 日) 的日补货总量和定价策略，使得商超收益最大。

在对未来的销量和价格进行预测时，模型的选择至关重要。基于问题的特点和数据特征，我们计划选择多元线性回归或者岭回归、lasso 回归等带有正则惩罚项的鲁棒性更优的回归模型来拟合销量与成本加权定价之间的关系。在价格预测方面，需要采用时间序列的分析方法，使用 ARIMA 或 SARIMA 模型进行线性预测，若数据拟合效果不好，则改用 SVM、GRU 等复杂度更加高的模型进行尝试。

然后，建立以补货总量和定价为决策变量，以周销售收益最大化为目标的非线性规划模型。最后，求解该规划模型，得到每类蔬菜在未来一周的优化补货总量和最优定价策略。

1.3.3 问题三：基于单品的定价和补货策略优化

问题三需要基于 2023 年 6 月 24 日至 6 月 30 日的可售品种数据，制定 7 月 1 日的单品层面补货策略。首先筛选出近一周内有销售的单品。然后根据题设要求，需要控制可售单品总数在 27-33 个之间，每个单品的订购量满足最小陈列量 2.5 千克。

为了符合限制条件，首先进行单品的筛选，得到符合要求的单品候选集合，在此基础上，构建基本的最优化模型，0-1 规划模型 [1]。通过逻辑变量表示是否进货该单品，同时，结合单品的损耗率以及时间序列得到的预测信息，构建利润模型，以单品订购量和定价为决策变量，以在约束条件下最大化商超收益为目标，求解即可得到商超收益最大化的补货计划。

2 模型假设与符号说明

2.1 模型假设

为确保模型的简化和可操作性，我们需要设定一些模型假设。以下是模型假设的详细说明：

1. 供需平衡假设：需求侧的市场需求和供给侧的供给量全部体现于销量（假设供需平衡始终成立，即进货全部卖出）中，即一种单品在某段时期销量较高，直接反映出该时期中该单品的市场需求量较高，同时也反应当前该单品在该时期的供给量较高。

2. 时间序列假设：销量和价格数据均为时间序列数据，具有一定的季节性和趋势性。我们假设这些时间序列可以通过 ARIMA（自回归积分滑动平均）或 LSTM（长短期记忆）等模型进行有效预测。

3. 库存管理假设：库存管理中只考虑当天的补货量作为库存，即不存在“隔夜菜”。

4. 价格波动假设：价格波动遵循一定的规律，且可以通过时间序列分析方法预测。价格变化主要受市场供需关系影响，而外部干扰（如政策变动、突发事件）在模型中暂不考虑。

5. 数据完备性假设：假设所使用的历史数据是完备且准确的，经清洗后没有缺失值或异常值，数据预处理后的结果能够真实反映市场情况。

6. 无竞争假设：模型假设市场上只有单一供应商，忽略了竞争对手的影响。这意味着我们只需考虑自身的定价和补货策略，而无需考虑市场竞争。

7. 短期供给假设：假定 2023 年 6 月 24 日-30 日的可售品种即为 2023 年 7 月 1 日可售品种。

2.2 符号说明

表 2-1 符号说明

符号	含义	单位
K_i	第 i 个品类 $i \in \{1, 2, \cdots, 6\}$	-
J_i	i 品类包含的单品总数	个
$g_{i,j}$	i 品类下第 j 个单品，$j \in \{1, 2, \cdots, J_i\}$	-
$w_{i,j}$	目标时期中 i 品类下第 j 个单品在 i 品类中的权值	-
Pro	目标时期的总利润	元
Pro_i	i 品类目标时期的总利润	元
P_i	i 品类目标时期等效单位定价	元/千克
B_i	i 品类目标时期的等效进货单价	元/千克
M_i	i 品类目标时期的进货量	千克
N_i	i 品类目标时期的销售量	千克
R_i	i 品类目标时期的等效损耗率	-
D_i	i 品类目标时期的等效折后价与等效原价之比	-
Ω_A	目标时间段内符合供应量需求的单品的指标对 (i, j) 集合	-
$p_{i,j}$	i 品类中第 j 个单品目标时期的单位定价	元/千克
$b_{i,j}$	i 品类中第 j 个单品目标时期的进货单价	元/千克
$m_{i,j}$	i 品类中第 j 个单品目标时期的进货量	千克
$n_{i,j}$	i 品类下第 j 个单品目标时期的总销量	千克
$r_{i,j}$	i 品类中第 j 个单品目标时期的损耗率	-
$d_{i,j}$	i 品类中第 j 个单品目标时期的折扣率	-

3 问题一的模型建立与求解

3.1 数据预处理与分组

在处理和分析蔬菜类商品的销量和价格数据之前，必须对数据进行预处理，以确保其完整性和一致性。数据预处理的步骤包括数据清洗、缺失值处理和数据标准化。

首先，数据清洗是指去除数据中的噪音和错误值。对于异常数据点（如极端高或低的销量和价格），我们采用三倍标准差法或 IQR（四分位距）方法进行异常值检测和处理。假设我们有销量数据 $\{S_t\}$ 和售价数据 $\{P_t\}$，我们定义异常值检测公式如下：

$$\text{上限} = Q_3 + 1.5 \times \text{IQR}$$

$$\text{下限} = Q_1 - 1.5 \times \text{IQR}$$

其中，Q_1 和 Q_3 分别为第 25 百分位数和第 75 百分位数，IQR 为四分位距。超出上限和下限的数据点将被视为异常值并进行处理。

其次，缺失值处理是保证数据完整性的关键步骤。经过对数据集的检查，缺失值的数量很少，因此我们直接对缺失值做删除处理，几乎不影响数据集的完整性。这也印证了我们的模型假设，这组销售数据的完整性和可信度很高。

在数据预处理之后，我们可以根据蔬菜的品类对数据进行分组。分组后的数据将用于后续的探索性数据分析和建模。对于问题一，我们感兴趣的变量只有销售时间和销量，因此我们做的处理有：

1. 变量解析：我们从附件 2 中提取出需要的列（单品编号、销售日期、销量 (千克) 等），并从日期字段中提取需要的时间特征，如年、月、日、季度、星期等。这样可以根据不同的时间维度进行分组。
2. 聚合操作：同时读取附件 1，将其与附件 2 的表格合并，以获取单品编号对应的单品名称、所属大类等信息。
3. 数据分组：根据提取的时间特征对数据进行分组。可以按天、周、月、季度或年等不同时间单位进行分组。使用 pandas 库的 ‘groupby‘函数，可以按指定的时间特征对数据进行分组。

3.2 数据探索性分析

数据探索性分析（Exploratory Data Analysis，EDA）是理解数据内在特征和规律的重要步骤。通过统计描述和可视化手段，我们可以初步了解销量和价格的分布、趋势及其变化规律。

首先，对每个蔬菜品类及单品的销量和价格进行统计描述。我们通过计算各类蔬菜的主要统计量来理解数据的分布和特性。常用的统计指标包括均值、标准差、中位数、偏度、峰度等。以蔬菜品类 i 的销量 S_i 为例，其统计描述如下：

1. 均值（mean）：均值是数据的平均值，表示总体数据的中心趋势。

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

2. 标准差（std）：标准差表示数据的离散程度，衡量数据点偏离均值的平均程度。

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2}$$

3. 中位数（median）：中位数是将数据排序后位于中间的值，表示数据的中间趋势。

$$\text{median}(X) = \begin{cases} x_{\frac{n+1}{2}} & \text{if } n \text{ is odd} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} & \text{if } n \text{ is even} \end{cases}$$

4. 峰度（kurt）：峰度衡量数据分布的陡峭程度，相对于正态分布的峰度。

$$\text{kurt}(X) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{X}}{\sigma}\right)^4 - 3$$

5. 偏度（skew）：偏度衡量数据分布的对称性，偏度为正表示右偏，负表示左偏。

$$\text{skew}(X) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{X}}{\sigma}\right)^3$$

6. 变异系数（coefficient of variation）：简称 cv，变异系数是标准差与均值的比值，用于比较不同数据集的离散程度。

$$\text{CV} = \frac{\sigma}{\bar{X}}$$

将计算结果列表如下：

表 3-1 各类蔬菜的统计指标

分类名称	mean	std	median	kurt	skew	cv
水生根茎类	30.876	24.243	26.354	12.822	2.525	0.785
花叶类	169.649	79.112	156.287	23.648	2.830	0.466
花菜类	38.164	22.313	33.824	3.843	1.461	0.585
茄类	20.833	12.584	18.408	5.291	1.650	0.604
辣椒类	84.084	53.078	72.537	18.257	3.121	0.631
食用菌	69.651	47.586	57.321	17.139	2.978	0.683

通过上述统计量的计算，我们可以对各类蔬菜的销量和价格数据进行深入分析，识别其特点和规律。

水生根茎类的均值为 30.876，该类蔬菜的日均销量较低，但标准差为 24.243，显示出销量波动较大。这意味着在某些时段，销量可能极高或极低。中位数为 26.354，说明大部分时段的销量低于均值。峰度为 12.822，偏度为 2.525，表明数据分布有较高的峰值且右偏。变异系数为 0.785，说明该类蔬菜的销量具有较大的离散性，波动较为剧烈。

花叶类的均值为 169.649，日均销量较高，但标准差为 79.112，显示出销量波动较大。中位数为 156.287，说明多数时段的销量集中在较高水平。峰度为 23.648，偏度为 2.830，显示出极高的峰值和显著的右偏。变异系数为 0.466，表明该类蔬菜的销量波动相对较小，数据较为集中。

花菜类的均值为 38.164，日均销量较低。标准差为 22.313，显示出销量波动较大。中位数为 33.824，说明大部分时段的销量低于均值。峰度为 3.843，偏度为 1.461，显示出相对较低的峰值和右偏。变异系数为 0.585，表明该类蔬菜的销量具有中等的离散性。

茄类的均值为 20.833，日均销量较低。标准差为 12.584，显示出销量波动较小。中位数为 18.408，说明大部分时段的销量低于均值。峰度为 5.291，偏度为 1.650，显示出较高的峰值和右偏。变异系数为 0.604，表明该类蔬菜的销量具有中等的离散性。

辣椒类的均值为 84.084，日平均销量较高。标准差为 53.078，显示出销量波动较大。中位数为 72.537，说明多数时段的销量集中在较高水平。峰度为 18.257，偏度为 3.121，显示出极高的峰值和显著的右偏。变异系数为 0.631，表明该类蔬菜的销量具有较大的离散性。

食用菌的均值为 69.651，日平均销量较高。标准差为 47.586，显示出销量波动较大。中位数为 57.321，说明多数时段的销量集中在较高水平。峰度为 17.139，偏度为 2.978，显示出极高的峰值和显著的右偏。变异系数为 0.683，表明该类蔬菜的销量具有较大的离散性。

3.3 相关性分析

相关性分析旨在研究不同蔬菜品类的销量和价格之间的线性关系。通过计算各变量之间的相关系数，可以了解哪些因素对销量和价格有显著影响，并为后续的建模提供依据。

在进行相关性分析之前，需要对数据做标准化处理。数据标准化是为了消除不同特征量纲的影响，使得数据在同一尺度下进行分析。常用的方法是 Z-score 标准化和 Min-Max 归一化，本题中我们采用 Z-score 标准化。假设 $\{X_i\}$ 为原始数据，Z 为标准化后的数据，Z-score 标准化公式如下：

$$Z = \frac{X_i - \mu}{\sigma}$$

其中，μ 为均值，σ 为标准差。

3.3.1 皮尔逊相关系数

皮尔逊相关系数是最常用的相关性指标，衡量两个变量之间的线性关系。其值介于-1 和 1 之间，其中 1 表示完全正相关，-1 表示完全负相关，0 表示无线性关系。皮尔逊相关系数的计算公式为：

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3-1}$$

其中，x_i 和 y_i 分别是变量 x 和 y 的观测值，$\bar{x}$ 和 $\bar{y}$ 是变量 x 和 y 的均值。

3.3.2 相关性分析结果

表3-2显示了各蔬菜品类销量和价格之间的皮尔逊相关系数。

表 3-2 各蔬菜品类销量和价格的皮尔逊相关系数

分类名称	水生根茎类	花叶类	花菜类	茄类	辣椒类	食用菌
水生根茎类	1.000	0.484	0.472	-0.467	0.459	0.653
花叶类	0.484	1.000	0.747	-0.034	0.624	0.546
花菜类	0.472	0.747	1.000	0.057	0.425	0.428
茄类	-0.467	-0.034	0.057	1.000	-0.191	-0.409
辣椒类	0.459	0.624	0.425	-0.191	1.000	0.575
食用菌	0.653	0.546	0.428	-0.409	0.575	1.000

画出热力图如图3-1：

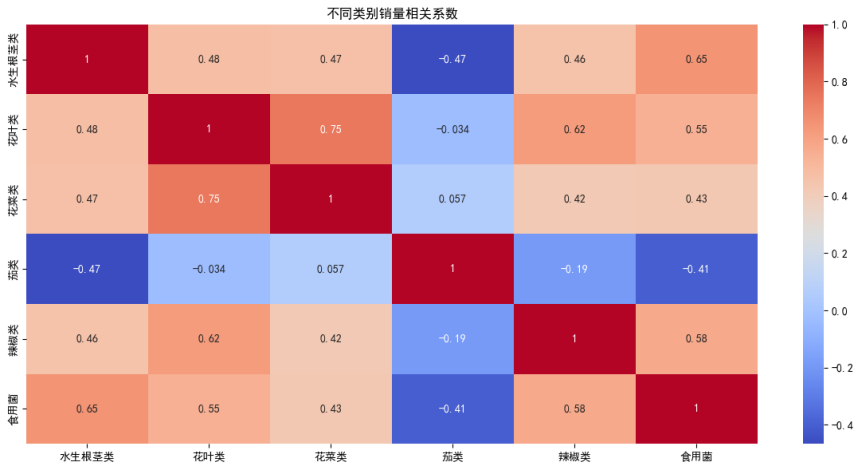

图 3-1 类别销量相关系数热力图

根据表3-2中的数据，我们可以观察到不同蔬菜品类之间的相关性情况。以下是对各品类间相关性的详细分析：

水生根茎类与其他品类的相关性总体较高，其中与食用菌的相关系数最高，达到0.653，表明这两类蔬菜的销量之间存在显著的正相关关系。这可能是因为这两类蔬菜在市场上具有相似的消费群体或受相似的季节性因素影响。水生根茎类与茄类之间存在显著的负相关关系，相关系数为-0.467，这表明当水生根茎类的销量增加时，茄类的销量可能会减少。这种负相关关系可能反映了消费者在这两类蔬菜之间的替代性选择。

花叶类与花菜类的相关系数为 0.747，显示出非常强的正相关关系，这可能是由于它们在生长周期、销售季节和消费习惯上具有高度相似性。此外，花叶类与辣椒类、食用菌的相关系数分别为 0.624 和 0.546，表明这些品类之间也存在较强的正相关关系，可能是因为它们在市场需求和供应链方面具有一定的共性。

花菜类与其他品类的相关性相对较弱，除了与花叶类的强正相关关系（0.747）外，其他相关系数均较低。特别是与茄类的相关系数为 0.057，几乎没有线性关系，这说明花菜类和茄类在销量上几乎不受对方影响。

茄类与其他品类的相关性普遍较低，尤其是与水生根茎类和食用菌分别呈现显著的负相关关系，相关系数分别为-0.467 和-0.409。这种负相关性可能反映了市场上茄类与这些蔬菜的竞争关系，当茄类销量增加时，消费者可能减少对水生根茎类和食用菌的购买。

辣椒类与花叶类、食用菌的相关系数分别为 0.624 和 0.575，显示出较强的正相关关系，可能反映了市场上这些蔬菜在某些消费场景下的共同需求。此外，辣椒类与茄类的相关系数为-0.191，表明两者之间存在一定的负相关关系，反映了市场上这两类蔬菜的替代性。

食用菌与其他品类的相关性总体较高，尤其是与水生根茎类和辣椒类的相关系数分别为 0.653 和 0.575，显示出较强的正相关关系。这表明这些品类的销量在一定程度上可能受到相同因素的影响，如季节变化、促销活动等。

3.4 数据可视化

通过绘制直方图、箱线图等可视化图表，我们可以展示和发掘销量的分布及其变化趋势。

3.4.1 总销量分析

首先，我们画出了三年来各个品类的总销量条形图和饼图，如图3-2、3-3所示：

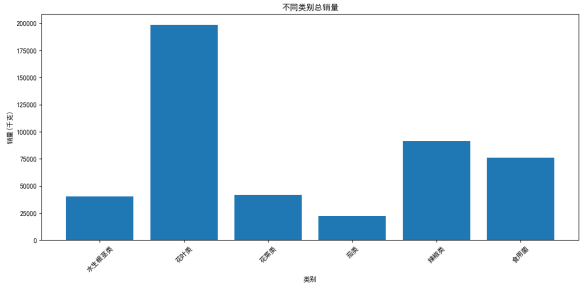

图 3-2 总销量条形图

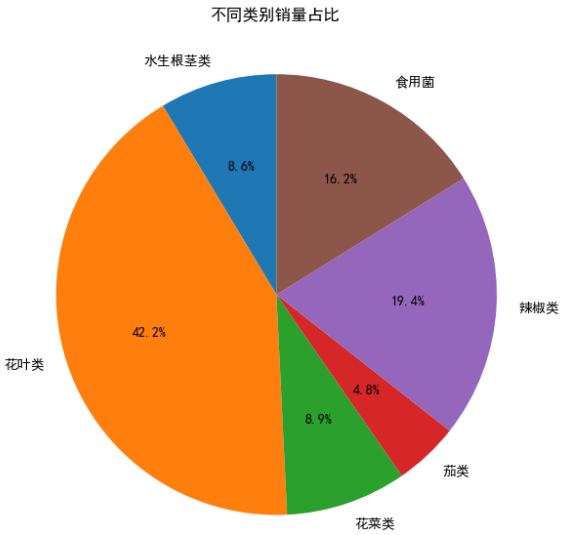

图 3-3 总销量饼图

从图中可以看出，花叶类的蔬菜销量远高于其他种类，总销量占比高达 42.2%；而茄类的蔬菜销量最低，仅有 4.8%。这反映出不同种类的蔬菜在供给量、受欢迎程度等方面都有着显著的区别。

3.4.2 月均大类销量分析

我们绘制了不同类别蔬菜每个月的月均销量走势图，如图3-4所示：

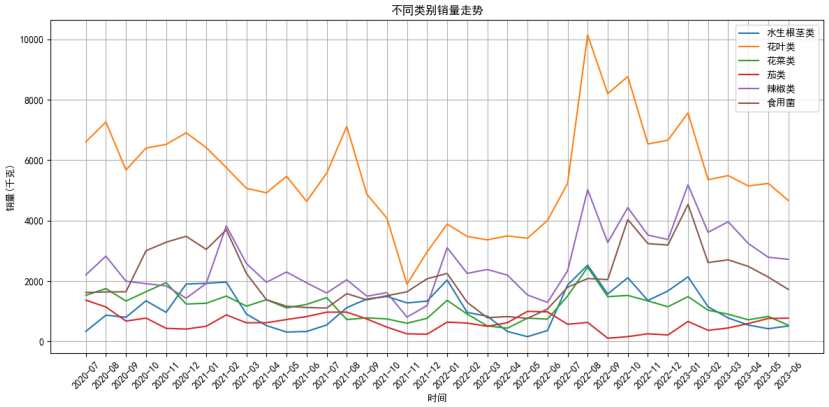

图 3-4 不同类别销量走势图

从折线图中可以看出，花叶类蔬菜的销量始终最高，其余品类的蔬菜中，辣椒类和食用菌类蔬菜的销量在 2022 年 7 月开始有明显增长。而且观察发现在 2021 年 9 月到 2022 年 6 月这段时间里，所有蔬菜的销量都明显下降，几乎达到最低点，结合时事推测，这可能是由于疫情等特殊情况的影响。但是从图中的波动情况我们没有发现明显的季节性特征。

3.4.3 月均单品销量分析

为了深入分析每个单品的月均销量，我们绘制了不同类别蔬菜单品的销量散点图，如图3-5所示：

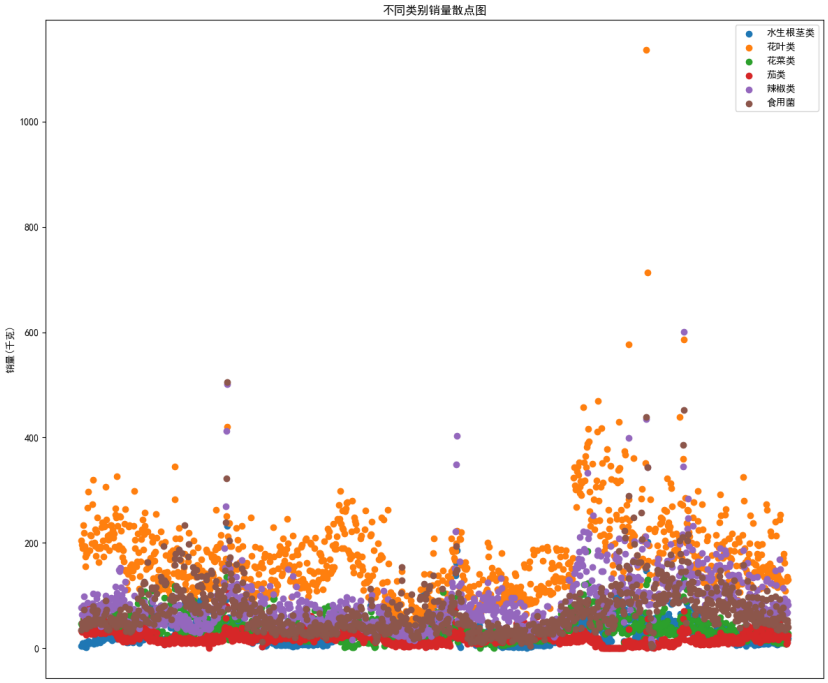

图 3-5 不同类别销量散点图

从散点图中可以看出，即使属于同一个品类，其中不同的蔬菜单品销量存在显著的离散性。部分单品销量稳定，而另一些单品销量波动较大。

3.4.4 日均销量分析

为了分析不同类别蔬菜在每日的销量情况，我们绘制了各类蔬菜的日均销量箱线图，如图3-6所示；以及所有单品的小提琴图，由于单品数量过大，我们只展示总销量最大的几个单品，如图3-7所示。

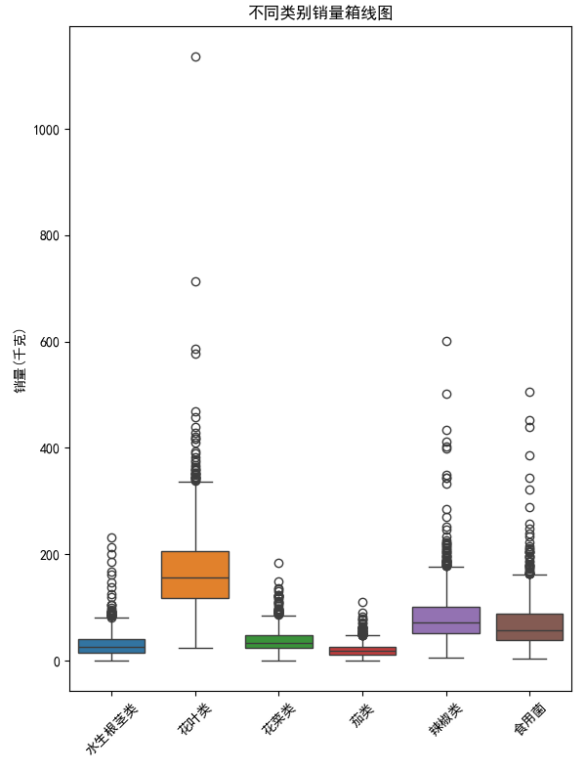

图 3-6 不同类别销量箱线图

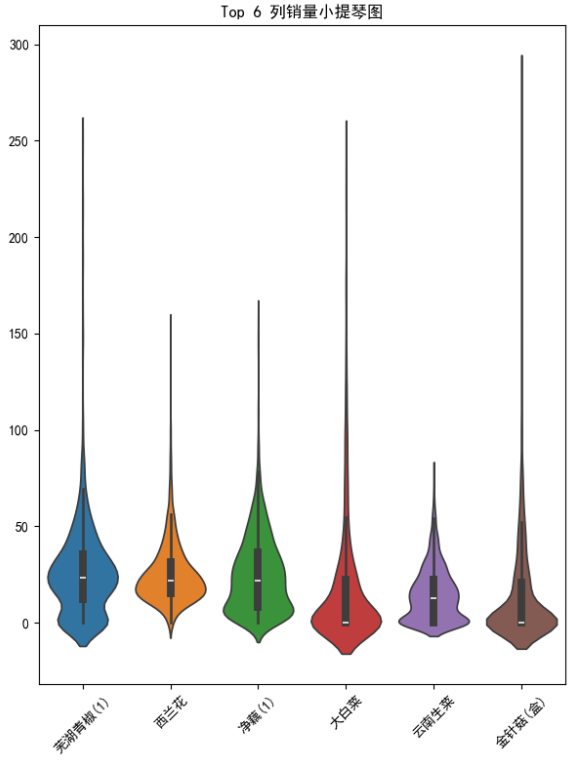

图 3-7 Top 6 销量小提琴图

箱线图展示了不同类别蔬菜日均销量的中位数、上下四分位数范围及潜在的异常值。我们可以看到，花叶类蔬菜的日均销量中位数较高，且数据分布相对集中，而茄类蔬菜的销量分布较为分散，存在较多的异常值，说明其销量十分不稳定。这个结论和先前分析月均销量时一致。

小提琴图结合了箱线图和密度图的特点，展示了销量数据的概率密度分布。通过小提琴图，我们可以观察到各个蔬菜单品销量的分布形态和对称性。芜湖青椒（1）和云南生菜的销量分布呈现出明显的双峰特征，说明这两类蔬菜在某些时间段内存在销量高峰。这几种蔬菜虽属于不同的品类，但是他们的销量并没有显著的差距，可以推测：品类间销量差距主要是由于类内的单品数量、供给稳定程度等影响较大。

3.4.5 不同季节销量走势

为了更好地了解各蔬菜品类的季节性变化，我们绘制了不同类别蔬菜的月均销量雷达图，如图3-8所示：

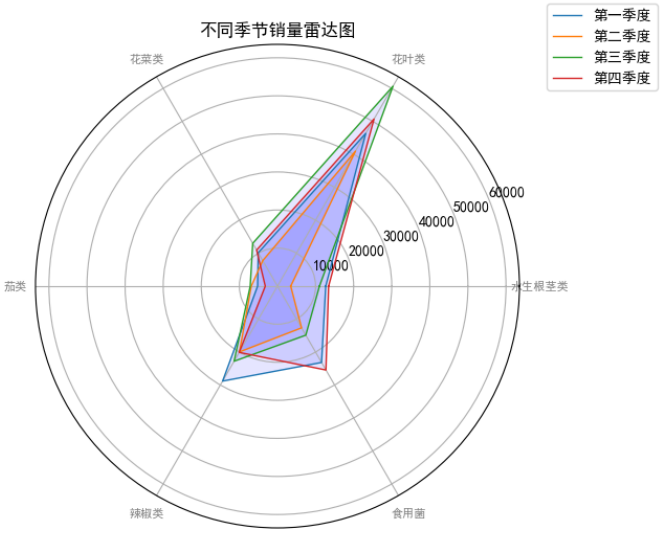

图 3-8 不同季节销量雷达图

其中，四个季度分别对应一年中的四个季节：第一季度：1 月、2 月、3 月；第二季度：4 月、5 月、6 月；第三季度：7 月、8 月、9 月；第四季度：10 月、11 月、12 月。

从雷达图中可以看出，各类蔬菜在不同季节的销量存在明显的波动。例如，辣椒类蔬菜在第一季度的销量较高，而在其他季度销量较低。水生根茎类蔬菜的销量在第二季度明显低于其他三个季度，这与消费者的饮食习惯和季节性需求变化有关。

3.5 分布拟合

为了检验各类蔬菜销量数据是否符合正态分布，我们采用了 Shapiro-Wilk 检验和 Kolmogorov-Smirnov 检验。这两种方法广泛应用于统计分析中，用于评估样本数据与正态分布的拟合程度。

3.5.1 Shapiro-Wilk 检验

Shapiro-Wilk 检验是一种有效的小样本正态性检验方法。其检验统计量 W 值在 0 到 1 之间。对于大样本，Shapiro-Wilk 检验的 p 值可以更敏感地反映偏离正态分布的程度。当 p 值小于显著性水平（设定为 0.05）时，我们拒绝原假设，认为数据不符合正态分布。其检验统计量 W 的计算公式如下：

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3-2}$$

其中：

- $x_{(i)}$ 表示第 i 个顺序统计量，即排序后的第 i 个样本值。
- $\bar{x}$ 表示样本均值，即 $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$。
- a_i 是常数，由样本大小 n 和样本的期望值及协方差矩阵决定。

W 的值越接近 1，表明数据越符合正态分布。

3.5.2 Kolmogorov-Smirnov 检验

Kolmogorov-Smirnov 检验是一种无参数检验，用于确定两个样本是否来自相同的分布，或者样本是否与特定的分布（如正态分布）一致。其检验统计量 D 值表示样本累积分布函数与期望累积分布函数之间的最大差异。当 p 值小于显著性水平时，拒绝原假设，认为数据不符合正态分布。其检验统计量 D 的计算公式如下：

$$D = \sup_x |F_n(x) - F(x)| \tag{3-3}$$

其中：

- $\sup_x$ 表示在所有 x 值上的最大值。
- $F_n(x)$ 表示样本数据的经验累积分布函数（ECDF）。
- $F(x)$ 表示理论分布的累积分布函数（CDF）。

D 值越大，表明样本分布与理论分布之间的差异越显著。

3.5.3 结果分析

表 3-3 各类蔬菜的正态性检验结果

	SW-W	SW-p	KS-D	KS-p
水生根茎类	0.948	0.000	1.000	0.000
花叶类	0.983	0.000	1.000	0.000
花菜类	0.954	0.000	1.000	0.000
茄类	0.979	0.000	1.000	0.000
辣椒类	0.943	0.000	1.000	0.000
食用菌	0.944	0.000	1.000	0.000

根据表3-3, 所有蔬菜品类的 Kolmogorov-Smirnov D 值均为 1.000, 且 p 值为 0.000。这说明各类蔬菜的销量数据与正态分布之间存在显著差异；各类蔬菜的 Shapiro-Wilk W 值都小于 1，并且 p 值均为 0.000。这表明无论是哪类蔬菜，其销量数据均显著偏离正态分布。

通过以上两种正态性检验方法，我们可以得出结论：各类蔬菜的销量数据均不符合正态分布。这一结果对于后续的数据分析和模型选择具有重要意义。在进行回归分

析或其他统计建模时，我们需要考虑数据的非正态性，选择适合的统计方法或对数据进行必要的变换，如对数变换或非参数方法，以提高模型的适用性和预测准确性。

3.6 聚类分析

3.6.1 KMeans 聚类算法介绍

KMeans 聚类是一种广泛使用的无监督学习算法，主要用于数据分组，其目标是将数据集分成 K 个簇，每个簇由数据的相似性决定。

KMeans 算法的具体步骤可以表示如下：

1. 初始化质心：

$$\mu_k^{(0)}, \quad k = 1, 2, \ldots, K$$

2. 簇分配：对于每个数据点 x_i，找到最近的质心：

$$c_i^{(t)} = \arg\min_k \|x_i - \mu_k^{(t-1)}\|^2$$

3. 质心更新：重新计算每个簇的质心：

$$\mu_k^{(t)} = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i$$

其中，C_k 是簇 k 的点集合，$|C_k|$ 是簇 k 中的点的数量。

3.6.2 使用手肘法确定 K 值

在进行 KMeans 聚类时，确定合适的簇数 K 是一个关键步骤。我们使用手肘法（Elbow Method）来确定最佳的 K 值。手肘法的基本思想是：随着簇数 K 的增加，簇内误差平方和（Sum of Squared Errors, SSE）会逐渐减小，当 K 达到某个临界点时，SSE 的下降幅度会显著减缓，该临界点对应的 K 值即为最佳簇数。

具体步骤如下：

1. 计算不同 K 值（如 1 到 10）的 KMeans 聚类结果，并记录每个 K 值对应的 SSE。2. 绘制 K 值与 SSE 的关系图。3. 找到 SSE 曲线明显拐点的位置，该位置的 K 值即为最佳簇数。

在我们的分析中，手肘图如图3-10所示：

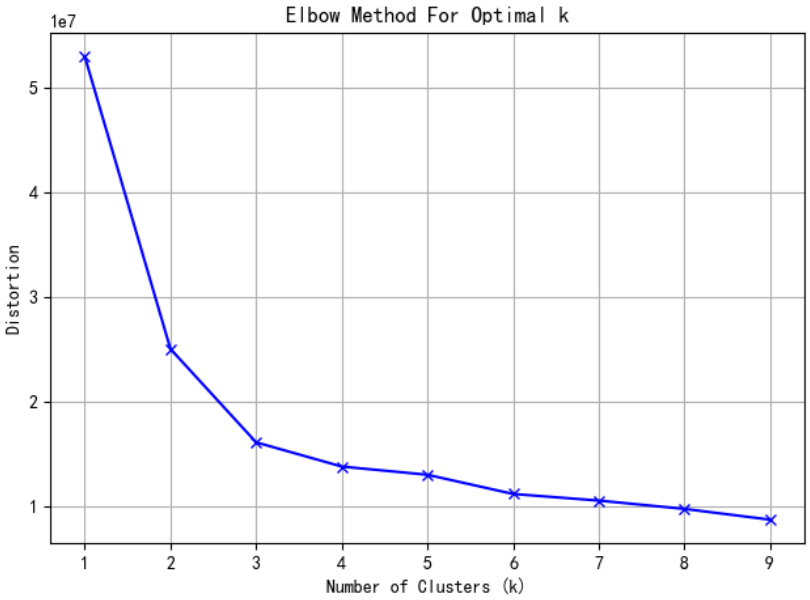

图 3-9 手肘法确定最佳簇数

从手肘图中可以看出，当 K 值为 4 时，SSE 的下降幅度明显减缓，因此我们选择 $K = 4$ 作为最佳簇数。

3.6.3 KMeans 聚类结果分析

通过对数据进行 KMeans 聚类分析，我们得到了四个簇，每个簇的均值如表3-4所示。

表 3-4 KMeans 聚类结果

label	0	1	2	...	10	11
0	17.858	17.056	14.630	...	12.881	14.900
1	188.904	209.255	231.930	...	122.371	122.384
2	999.486	846.685	635.883	...	844.132	886.798
3	310.034	203.706	163.089	...	233.880	244.420

为了更直观地展示聚类之间的区别，我们将其绘制在一张条形图上以便对比，如图

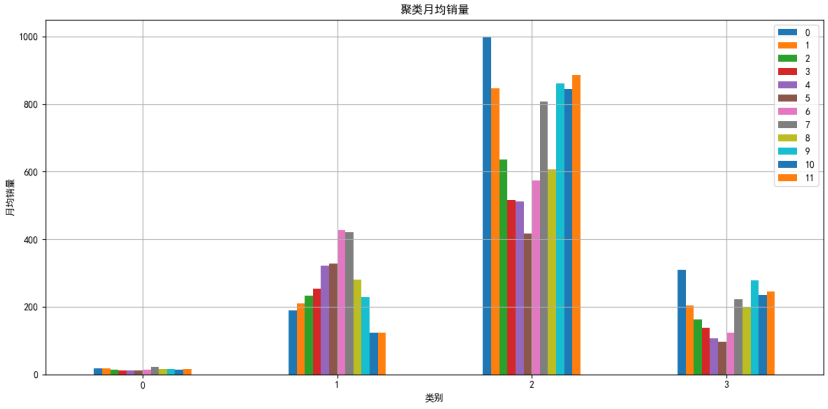

图 3-10 聚类月均销量条形图

聚类结果可以分为以下几类进行详细分析：

1. 簇 0：这一簇的数据点的销量普遍较低，月份之间的变化幅度也不大，销量在 11 到 22 之间波动。这表明这一类蔬菜的需求较为稳定，可能是一些受季节影响较小、销售量不大的品种。

2. 簇 1：这一簇的数据点的销量相对较高，且在夏季（6-8 月）销量达到峰值。这可能是一些受季节影响较大的蔬菜，如夏季上市的时令蔬菜。

3. 簇 2：这一簇的数据点的销量最高，且在 12 月份达到最大值。显然，这类蔬菜在冬季需求量巨大，可能是一些冬季主打的蔬菜品种，如大白菜、萝卜等。

4. 簇 3：这一簇的数据点的销量较低，但在 7 月份和 9 月份有两个显著的销售高峰。这类蔬菜可能是某些特定月份需求量激增的品种，或是有些季节性促销活动导致的销量上升。

通过 KMeans 聚类分析，我们可以识别出不同类别蔬菜的销售模式，为后续的问题解决提供重要思路。根据不同簇的销售特点，可以制定相应的采购和销售策略，提高经营效率和市场响应速度。例如，对于簇 2 中的高销量蔬菜，应提前准备充足库存，并加强冬季的市场宣传和促销力度。对于簇 0 中的稳定销量蔬菜，则可以采取精细化库存管理，避免积压和缺货。

4 问题二的模型建立与求解

4.1 基于品类的总销量与定价关系探究

4.1.1 建立品类的等效单位定价模型

在同一个品类中，不同单品的销售量具有较大差异。因此，每种单品的定价对品类的等效定价的影响不同，销售量大的单品的定价对品类的等效定价影响比较大，应赋予较大的权重。本文采用销量权重赋值法，根据单品销量在所属品类总销量中的占比赋予单品定价在品类等效定价中的权重。

针对第 i 个品类 K_i，其下包含 J_i 个单品、总销量为 N_i，第 j 个单品的单位定价为 $p_{i,j}$、总销量为 $n_{i,j}$。

构建品类的等效单位定价模型:

$$P_i = \sum_{j=1}^{J_i} w_{i,j} * p_{i,j} = \frac{n_{i,j}}{N_i} * p_{i,j}$$

根据预处理得到的按月分类的单品销量数据、按月分类的品类销量数据和原附件 1 中的类别信息，计算得到各个单品的权重和所属类别标签，存储在文件 $goods\_ij\_weight.xlsx$ 中。

根据权重数据集 $goods\_ij\_weight.xlsx$、按月分组的单品单价数据集，计算按月分组的各品类的等效单位定价，将结果保存在文件“按月分组品类单价$.xlsx$”中。

表 4-1 不同蔬菜种类的月平均单价 (元/斤)

Year-Month	花叶类	花菜类	水生根茎类	茄类	辣椒类	食用菌
2020-07	6.171	11.185	14.274	5.622	4.893	5.681
2020-08	6.534	10.582	12.947	6.371	4.002	6.348
2020-09	5.909	10.386	12.122	7.432	8.003	7.966
2020-10	5.697	9.535	5.928	5.591	5.272	6.519
2020-11	3.561	5.233	3.436	5.370	6.131	5.349
2020-12	3.228	6.987	5.945	7.066	7.957	6.682
2021-01	6.748	10.262	4.065	10.056	10.358	6.935
2021-02	7.705	7.483	8.852	11.657	12.301	7.122

4.1.2 分析品类销量与等效单位定价的相关性

按照时间戳合并按月分组的品类等效单价、按月分组的品类销量的数据集，针对各个品类依次绘制销量与等等效单位定价的散点图，观察两个变量之间的关系。得到各品类的散点图如图 X:

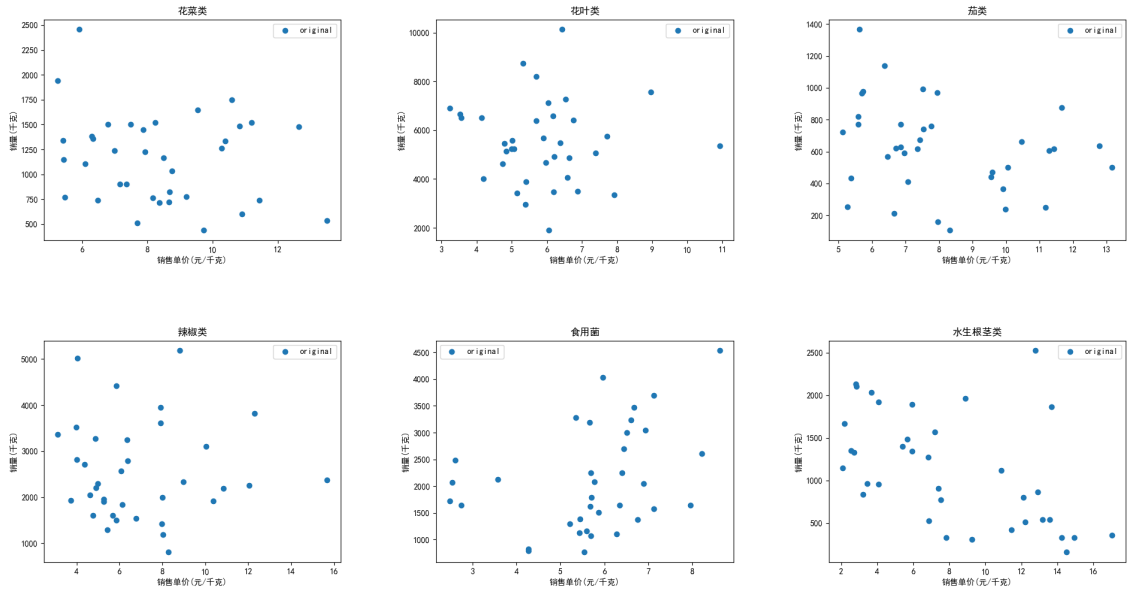

图 4-1 各品类散点图

可见，对于不同品类，销量与销售等效单价之间均存在一定的关联，但两变量间的分布规律和函数关系并不一致。

4.1.3 对销量与等效单价进行函数拟合

由于不同品类中销量与销售单价间的关系存在差异，对各个品类依次进行销量与销售单价的函数拟合。依次采用线性拟合、多项式拟合 [2] 和 sigmod 函数拟合的方法进行尝试。下面展示用每种函数拟合的效果，以水生根茎类和花叶类为例。

线性拟合：

$$y = kx + b$$

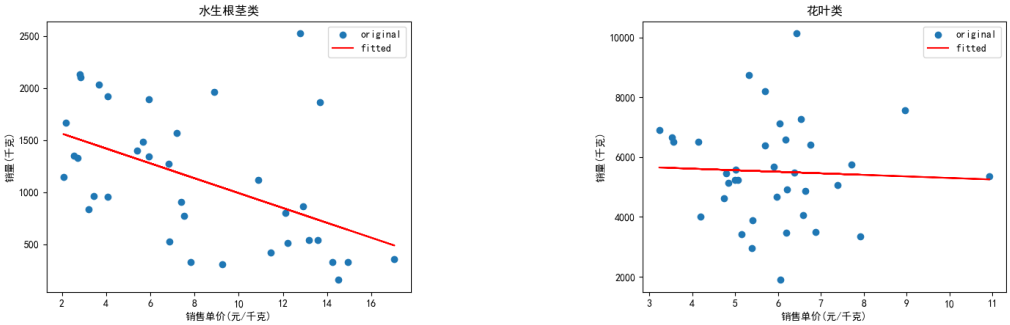

图 4-2 线性拟合结果图

多项式拟合：

$$y = p_0 + p_1 x + p_2 x^2 + \cdots + p_n x^n$$

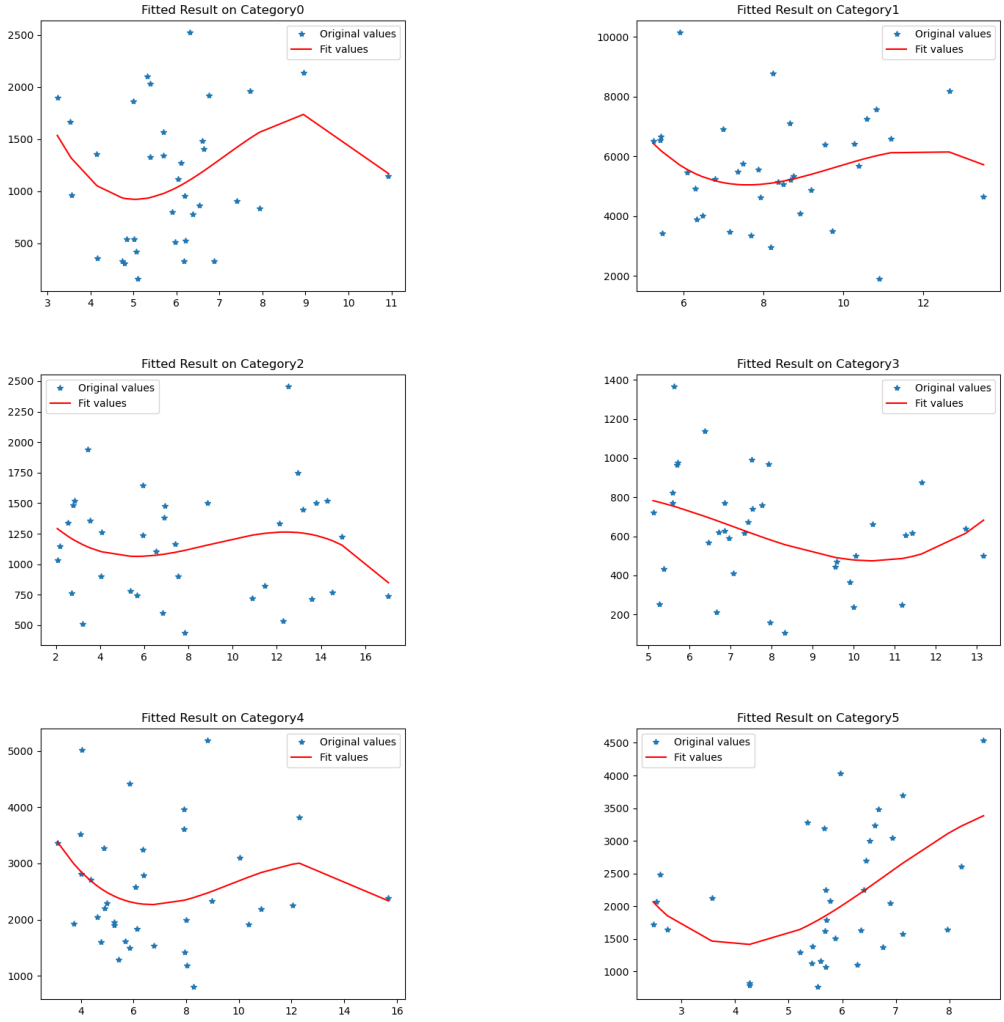

图 4-3 各品类销量与成本加成定价的多项式约束关系

可见，各品类销量与单价的散点样本比较分散，用线性拟合难以达到比较好的效果；对于多项式函数，随着次数增加模型在训练集上的拟合准确度提高，但会导致过拟合和出现指数爆炸问题，不利于模型的拓展应用。

sigmod 函数拟合：

$$y = f(x) = \frac{e^{ax+b}}{1+e^{ax+b}}(x_{max} - x_{min}) + x_{min}$$

$$N_i = f_i(P_i) = \frac{e^{a_i P_i + b_i}}{1+e^{a_i P_i + b_i}}(P_{max} - P_{min}) + P_{min}$$

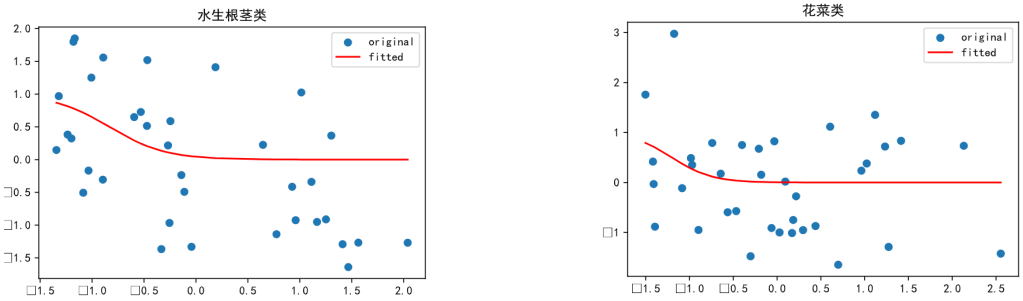

图 4-4 sigmoid 函数拟合结果图

可见，sigmod 函数对销量与销售单价的函数拟合效果更优，且不存在指数爆炸的问题，更适合销量与销售单价的预测。调用 $scipy.optimize$ 的 $curve\_fit$ 函数进行参数估计，得到各品类的结果如下：

表 4-2 各品类销量与销售单价的拟合结果

分类名称	a	b
花叶类	-765.059	-1217.528
花菜类	-4.845	-7.478
水生根茎类	-2.983	-6.582
茄类	-2.979	-3.226
辣椒类	-201.962	-187.389
食用菌	3.288	-3.135

然而，实际上销量与成本加成定价之间的关系是比较模糊的（从散点图可见）。我们应该找到合适的上下限曲线来有效地符合大多数情形。于是，调整拟合函数为 $N_i = f_i(P_i) = a_i/P_i + b_iP_i + c_i(i = 1, 2, \cdots, 6)$，并针对该拟合函数设置上下限放缩系数 l_i 与 u_i，即销量与加成成本定价大部分情形下会有如下控制关系：

$$N_i \in [l_i \cdot f_i(P_i),\ u_i \cdot f_i(P_i)], \quad i = 1, 2, \cdots, 6$$

由程序可得合适的参数如下表 4-3：

表 4-3 各品类销量与成本加成定价的拟合结果

分类名称	a_i	b_i	c_i	l_i	u_i
花叶类	516.3904	12.8409	15.0696	0.5	1.8
花菜类	369.9957	4.1393	-42.9657	0.4	1.6
水生根茎类	1.8895	-2.2526	55.3172	0.3	1.8
茄类	179.3935	1.4922	-15.1570	0.3	1.8
辣椒类	289.08373	5.4042	-0.5120	0.4	1.8
食用菌	417.7931	27.3609	-166.8525	0.5	2.0

得到各品类销量与成本加成定价的约束关系，如下图 4-5 所示：

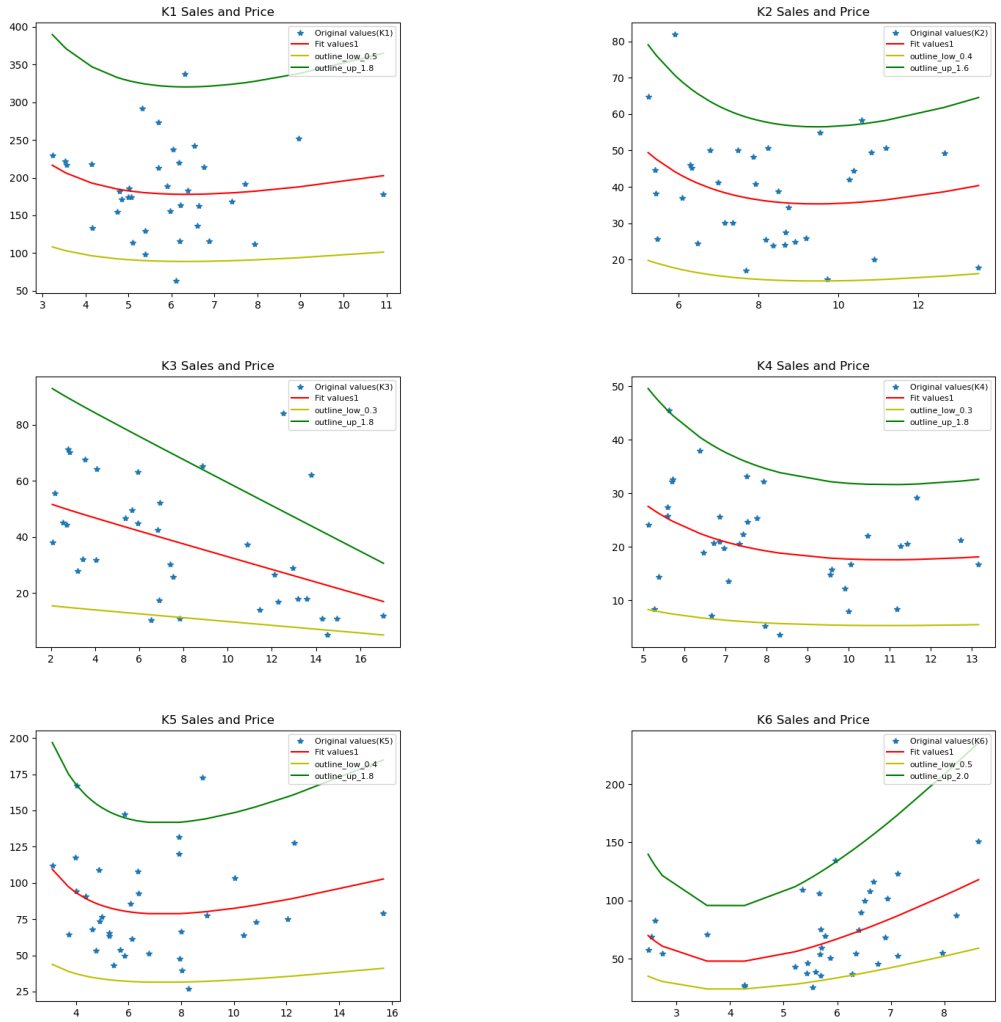

图 4-5 各品类销量与成本加成定价的约束关系

4.2 基于季节性分析和 GRU 模型的目标时段数据预测

4.2.1 针对销量数据的季节性分析

原数据集的时间戳包括 2022 年 7 月-2023 年 6 月三年中的日期，数据量很大、且具有节律性，通过按月分组的形式对数据集进行压缩，以月份为单位（共 36 个样本）对数据进行季节性分析。调用 *statsmodels.tsa.seasonal* 库引入 *seasonal_decompose* 函数，对原数据做季节性分解，得到结果如下：

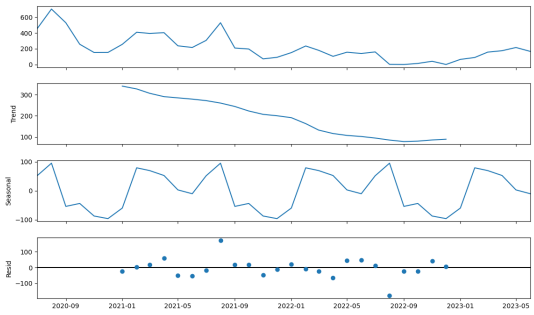

图 4-6 按月分组的品类数据季节性分解结果

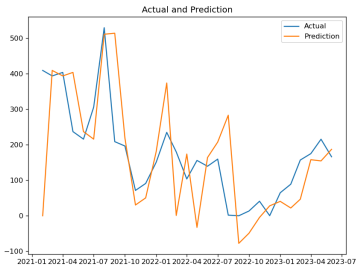

图 4-7

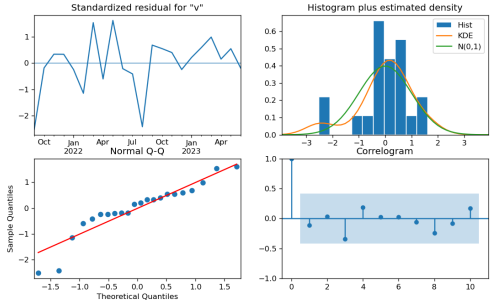

图 4-8

由以上分解结果，数据具有一定的季节性和周期性，周期为 12，即一年。因此，预测目标时段的数据时，可以只采用同期（即每年）6 月中旬到 7 月中旬（6.15-7.15）的数据作为历史数据进行预测。

4.2.2 GRU 模型介绍与目标时段数据预测

4.2.2.1 GRU 模型介绍

门控循环单元（Gated Recurrent Unit, GRU）是 Recurrent Neural Network (RNN) 的一种变体，旨在解决传统 RNN 在处理长序列数据时存在的梯度消失和梯度爆炸问题。GRU 通过引入门控机制，能够更有效地捕捉时间序列中的长期依赖关系。

4.2.2.2 GRU 模型原理与公式

GRU 模型主要由更新门（update gate）和重置门（reset gate）组成，这两个门控制了信息在神经网络中的流动。他们的作用如下：

1. 更新门：决定了前一步的隐状态信息有多少需要传递到当前状态。
2. 重置门：决定了前一步的隐状态信息有多少需要重置。

GRU 的一些关键计算公式如下：

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t])$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t])$$

$$\tilde{h}_t = \tanh(W_h \cdot [r_t \odot h_{t-1}, x_t])$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$$

其中：x_t 是当前时间步的输入。h_t 是当前时间步的隐状态。h_{t-1} 是前一时间步的隐状态。z_t 是更新门。r_t 是重置门。$\tilde{h}_t$ 是候选隐状态。σ 是 sigmoid 激活函数。$\odot$ 表示元素级的乘法。

更新门的作用是控制当前隐状态中有多少信息来自于之前的隐状态，而重置门则是控制当前隐状态中有多少信息需要被忽略或重置。

4.2.2.3 GRU 模型的结构与流程图

为了更清晰地理解 GRU 的工作机制，我们可以参考图4-9所示的 GRU 结构图。

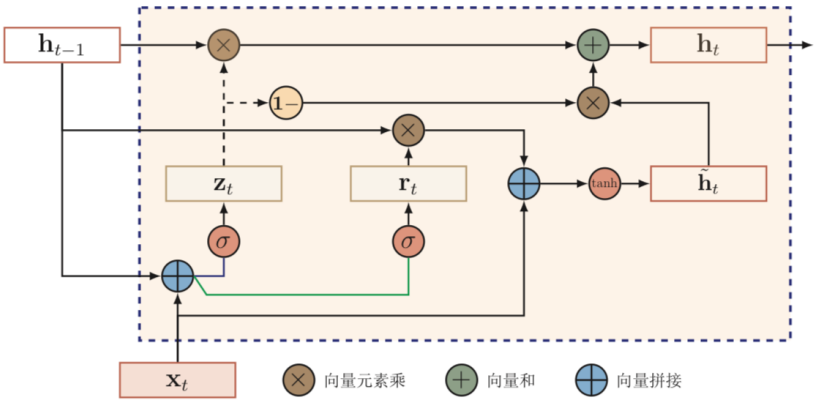

图 4-9 GRU 网络结构图

如图所示，GRU 的结构较为简洁，主要包括以下几个步骤：

1. 输入处理：当前时间步的输入 x_t 和前一时间步的隐状态 h_{t-1} 一起输入到重置门和更新门。

2. 计算重置门和更新门：通过 sigmoid 激活函数计算重置门 r_t 和更新门 z_t。

3. 生成候选隐状态：前一时间步的隐状态 h_{t-1} 经过重置门调制后与当前输入 x_t 一起生成候选隐状态 $\tilde{h}_t$。

4. 计算当前隐状态：利用更新门将前一时间步的隐状态和候选隐状态结合，生成当前时间步的隐状态 h_t。

GRU 通过其简化的门控机制，与 LSTM 相比，具有计算效率更高、训练时间更短的优点，但在捕捉长时间依赖性方面，性能差异不大。GRU 的这种平衡性使其在许多时间序列预测任务中得到了广泛应用。

综上所述，GRU 通过其独特的门控机制，有效地缓解了传统 RNN 的梯度消失问题，能够更好地捕捉序列数据中的长期依赖关系。在实际应用中，GRU 广泛用于自然语言处理、时间序列预测等领域，显示出了其强大的建模能力和实用性。我们选用其进行本文的所有预测任务。

4.2.3 目标时段数据预测

分别对目标时段（2023 年 7 月 1-7 日）分别七天的单品销售量 $n_{i,j}$、单品等效进货单价 $b_{i,j}$、品类销售量 N_i、品类等效进货单价 B_i 进行预测。

我们首先对每个品类的销量时间序列进行训练和预测，部分结果如图4-10和图4-11所展示。

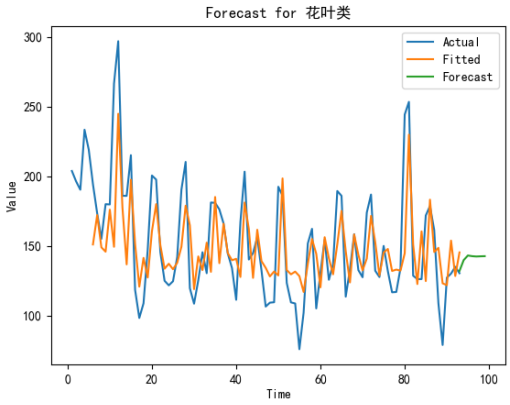

图 4-10

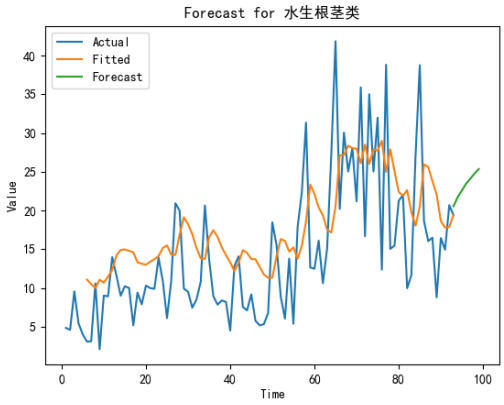

图 4-11

可以发现, GRU 模型很好地学习到了时间序列数据的趋势和波动信息, 预测结果也基本符合预期。因需要预测的目标时段以一日为单位, 需要采取日为单位的历史数据。采用预处理得到的按日分类的单品销量数据、按日分类的单品进价数据、按日分类的品类销量数据, 分别利用 GRU 模型预测出未来 7 天的单品销量 $n_{i,j}$、单品进价 $b_{i,j}$ 和品类销量数据 N_i, 将结果分别存储在文件 $sales\_forecast\_gru.xlsx$、$cost\_forecast\_gru.xlsx$、$Categorysales\_forecast\_gru.xlsx$ 中。

根据权重数据集 $goods\_ij\_weight.xlsx$、预测得到的目标时段单品进价数据集 $cost\_forecast\_gru.xlsx$，与 Chapter 4.1.1 中品类的等效单位定价模型类似，用单品进价 $b_{i,j}$ 代替单品定价 $p_{i,j}$，计算各品类的等效单位进价 B_i，将结果保存在文件 $Categorycost\_forecast.xlsx$ 中。

表 4-4 不同蔬菜每日销量变化

蔬菜	**Day 1**	**Day 2**	**Day 3**	**Day 4**	**Day 5**	**Day 6**	**Day 7**
苋菜	7.685	7.763	7.869	7.886	7.920	7.927	7.934
云南生菜	2.054	2.007	2.058	2.083	2.091	2.122	2.125
竹叶菜	13.650	13.917	13.994	14.023	14.033	14.037	14.040
小白菜	-0.151	-0.179	-0.208	-0.229	-0.246	-0.261	-0.270
南瓜尖	0.157	0.157	0.157	0.157	0.157	0.157	0.157
上海青	4.378	4.430	4.372	4.223	4.126	3.986	3.942
菜心	1.622	2.096	2.609	3.087	3.588	4.077	4.537
木耳菜	5.171	5.073	4.989	4.959	4.944	4.934	4.929
大白菜	-0.156	-0.272	-0.431	-0.636	-0.901	-1.238	-1.673

可以发现，其中还有一些预测值为负数的情况，我们在后续处理中将预测值小于 0 的单品也视为不可供应的单品进行处理。

4.3 基于品类的补货和定价策略制定

4.3.1 步骤一：基于品类的利润模型构建

以下构建以品类单位，2023 年 7 月 1-7 日的周利润模型。

各品类的周利润 Pro_i 由周等效定价 P_i 、周进货量 M_I 、周等效进货单价 B_i 、周销售量 N_i 、等效损耗率 R_i 、等效折扣率 D_i 共同决定。

各品类的利润模型如下:

$$Pro_i = N_i[(1-R_i)P_i + R_i(Pi \cdot D_i)] - B_i M_i \quad (M_i \geq N_i)$$

总的利润模型即为:

$$Pro = \sum_{i=1}^{6} \hat{N}_i \cdot [(1-R_i) \cdot P_i + R_i \cdot (P_i \cdot D_i) - \hat{B}_i \cdot M_i]$$

品类的周等效定价 P_i 、周进货量 M_I 、周等效进货单价 B_i 、周销售量 N_i 的范围都要预测得到。其中，周进货量 M_I 和周销售量 N_i 的范围实际上是统一的（历史数据的供需平衡假设），不过实际取值可以不同。

结合上一小节中 GNU 模型的预测结果，周销售量 N_i 与周等效进货单价 B_i 的预期范围应如下区间所示。区间为预测值上下 5% 波动区间的非负部分，若波动区间上限为负则区间两端均为 0.

$$\begin{cases} \hat{N}_i \in [n_i^-, n_i^+] \\ \hat{B}_i \in [b_i^-, b_i^+] \end{cases}$$

而 $\hat{N}_i$, P_i 间的约束关系 f_i 在上一小节得到，如下所示：

$$\hat{N}_i \in [l_i \cdot f_i(P_i),\, u_i \cdot f_i(P_i)], \quad i = 1, 2, \cdots, 6$$

4.3.2 步骤二：品类等效折扣率和等效损耗率计算

生成等效折扣率 D_i 所选取数据范围为 2023 年 6 月 24-30 日。先得到该时期有供应的单品的 $d_{i,j}$ ，再基于各单品累积销量得到其在品类的权重 $w_{i,j}$ 。类似的，也可得到 R_i 。于是 D_i ， R_i 为：

$$D_i = \sum_{i=1}^{6} w_{i,j} d_{i,j}$$
$$R_i = \sum_{i=1}^{6} w_{i,j} r_{i,j}$$

其中 $d_{i,j}$ 为各单品在目标时期的打折均价与未打折均价比值（若无打折情形，则为 1；若仅有打折情形，则默认取为 0.8）。而 $r_{i,j}$ 直接由附件四得到。

得到各品类等效折扣率如下表 4-5 所示：

表 4-5 各品类等效折扣率和等效损耗率

品类	D_i	R_i
花叶类	0.7986	0.1551
花菜类	0.9374	0.1365
水生根茎类	0.7037	0.1283
茄类	1.0000	0.0945
辣椒类	0.7911	0.0924
食用菌	0.6770	0.0668

4.3.3 步骤三：非线性规划的建立和求解

于是以商超总收益作为目标函数，并取最大化，得到完整的非线性规划模型如下：

$$\max_{M_i, P_i} Pro = \sum_{i=1}^{6} \hat{N}_i \cdot [(1 - R_i) \cdot P_i + R_i \cdot (P_i \cdot D_i) - \hat{B}_i \cdot M_i]$$

$$s.t. \begin{cases} M_i \in [\hat{N}_i, n_i^+] \\ \hat{N}_i \in [n_i^-, n_i^+] \\ \hat{B}_i \in [b_i^-, b_i^+] \\ \hat{N}_i \in [l_i \cdot f_i(P_i),\, u_i \cdot f_i(P_i)] \\ n_i^-, n_i^+, b_i^-, b_i^+, P_i \geq 0 \\ i = 1, 2, \cdots, 6 \end{cases}$$

该非线性规划模型使用 python 的 scipy 库中 minimize 模块求解。

最终得到的基于品类的下周日补货总量和定价策略如下表 4-6 所示（七天都采取相同的策略）：

表 4-6 基于品类的日补货和定价策略

品类	M_i	P_i
花叶类	25.9535	6.4197
花菜类	118.2640	3.6330
水生根茎类	30.5768	17.0649
茄类	31.6600	10.0000
辣椒类	66.9781	8.0000
食用菌	43.1209	7.1764

对应得到单日最高利润平均约为 511 元。

5 问题三的模型建立与求解

5.1 基于单品的定价和补货策略优化

相较于问题二，问题三有更细致的要求：以单品为单位，满足单品总数约束，最小陈列量约束，尽量满足市场对各品类商品的需求。在满足条件的前提下，制定 2023 年 7 月 1 日的单品补货量和定价策略，以达到商超利润最大化。

5.1.1 步骤一：有效单品的筛选

为了便于操作，首先进行单品种类的筛选：选择目标时期有供应量的单品，同时剔除 2023 年 6 月 24-30 日供应量 (即日销售量) 上限没有达到 2.5kg 的单品，最终得到符合供应要求的单品的指标 (i,j) 集合，记为 Ω_A 。总共有 82 个有效单品。

5.1.2 步骤二：单品种信息的预测

类似于问题二，此处仍需通过时间序列模型等基于历年相关时期（6 月 15 日至 7 月 15 日）的数据，预测 2023 年 7 月 1 日的有效单品的信息，包括：单品的预计进价范围，单品的预计销量范围（即供给量和需求量范围），单品的定价范围。同时采用问题二中约束关系的模型参数，得到单品的销量与加成成本定价的约束关系。

类似问题二，得到的预测区间记号如下：

$$\begin{cases} \hat{n}_{i,j} \in [n_{i,j}^{-}, n_{i,j}^{+}] \\ \hat{b}_{i,j} \in [b_{i,j}^{-}, b_{i,j}^{+}] \\ p_{i,j} \in [l_i(\hat{n}_{i,j}), u_i(\hat{n}_{i,j})] \end{cases}$$

其中 $(i,j) \in \Omega_A$ 。

5.1.3 步骤三：基于单品的利润最优化模型的建立与求解

为了从 Ω_A 中选择合适的单品，可以采用逻辑变量 $x_{i,j}$ 构建 0-1 规划模型。

$$x_{i,j} = \begin{cases} 1, & \text{选取 } i \text{ 品类中第 } j \text{ 个单品} \\ 0, & \text{舍弃 } i \text{ 品类中第 } j \text{ 个单品} \end{cases}$$

其中 $(i,j) \in \Omega_A$.

类似问题二，总利润 Pro 与所选单品在 2023 年 7 月 1 日的定价 $p_{i,j}$ 、进货量 $m_{i,j}$ 、进货单价 $b_{i,j}$ 、销售量 $n_{i,j}$ 、损耗率 $r_{i,j}$ 、折扣率 $d_{i,j}$ 共同决定。

则模型的目标函数总利润为:

$$Pro = \sum_{(i,j)\in\Omega_A} \left\{[(1-r_{i,j})\cdot p_{i,j} + r_{i,j}(p_{i,j}\cdot d_{i,j})]\cdot \hat{n}_{i,j} - \hat{b}_{i,j}\cdot m_{i,j}\right\}\cdot x_{i,j} \quad (m_{i,j}\geq \hat{n}_{i,j})$$

由于采用筛选过的有效单品，还需要补充约束条件如下:

单品总数的条件:

$$\sum_{(i,j)\in\Omega_A} x_{i,j} \in [27,33]$$

满足各品类都有供应的条件:

$$\sum_{j} x_{i,j} \geq 1 \quad (i=1,2,\cdots,6)$$

于是完整的 0-1 规划模型如下:

$$\max_{x_{i,j},p_{i,j},m_{i,j}} Pro = \sum_{(i,j)\in\Omega_A} \left([(1-r_{i,j})\cdot p_{i,j} + r_{i,j}(p_{i,j}\cdot d_{i,j})]\cdot \hat{n}_{i,j} - \hat{b}_{i,j}\cdot m_{i,j}\right)\cdot x_{i,j}$$

$$s.t.\begin{cases} m_{i,j}\in[\hat{n}_{i,j}, n^+_{i,j}] \\ \hat{n}_{i,j}\in[n^-_{i,j}, n^+_{i,j}] \\ \hat{b}_{i,j}\in[b^-_{i,j}, b^+_{i,j}] \\ p_{i,j}\in[l_i(\hat{n}_{i,j}), u_i(\hat{n}_{i,j})], \quad i=1,2,\cdots,6 \\ \sum_{(i,j)\in\Omega_A} x_{i,j}\in[27,33] \\ \sum_j x_{i,j}\geq 1 \quad (i=1,2,\cdots,6) \\ n^-_{i,j}, n^+_{i,j}, b^-_{i,j}, b^+_{i,j}\geq 0 \\ x_{i,j}\in\{0,1\} \\ (i,j)\in\Omega_A \end{cases}$$

对于该模型，仍然使用 python 中 scipy 库的 minimize 模块进行最优化求解。得到 2023 年 7 月 1 日的单品日补货和定价策略。

6 总结

本文完成了关于商超的实时补货和定价策略的基础性探索与分析。

在问题一中，对历史数据进行全面的预处理后，采用丰富可视化的探索性分析和统计性检验等方法探索数据分布规律，并使用 kMeans 方法，探究数据中各品类及单品的分布关系，得到“互补性”“替代性”等相关关系。

在问题二中，以品类为基础进行探究：一方面，通过 MLR、sigmoid 等模型进行销量与加成成本定价的关系分析；另一方面，基于 GRU 模型完成了目标时期的品类相关信息的预测。在此基础上，构建品类为单位的非线性规划模型，找到目标时期的利润最大化的补货与定价策略。

而在问题三中，更进一步细化到单品层面的探究。完成了基于单品的利润模型构建，在满足市场需求、供应需求等约束条件下，完成 0-1 规划模型的最优化。

综合本文的探究与分析，能充分理解商超实时补货和定价策略的重要性与困难程度。同时，本文的探究与分析也给出了一些基本的模型和方法，为该议题给出了一些粗浅的解答。而正如问题四所要探究的，真实情形的商超实时补货和定价策略显然更为复杂，销量的影响因素还有很多，等等。这些都有待未来进一步的探究。

参考文献

[1] 姜启源　谢金星. 数学建模算法与应用（第 2 版）[M]. [出版地不详]: 高等教育出版社, 2012.
[2] 司守奎　. 数学建模算法与应用（第 2 版）[M]. [出版地不详]: 国防工业出版社, 2015.