# 基于 ARIMA 和多目标优化模型的商超补货-定价策略


## 摘　要

生鲜商超中蔬菜的补货与定价最优决策时刻变化，而传统的补货定价决策依赖长期的经验积累以及高昂的调研花销，故决策层需要依据商超的销售等数据，通过对供需关系和销售组合问题进行分析，进行自动化的量化分析，达到收益最大化的目的，并且实现现代化的管理，决策方法。本文通过建立回归分析模型和层次聚类模型，进行了可视化分析、描述性统计分析、相关系数分析、聚类分层分析和回归分析，给出了使商超收益最大的补货和定价策略。该问题的研究能为商超提供指导性意见，帮助商超实现最大收益。

针对问题一，本文进行可视化分析、描述性统计分析、相关系数分析和聚类分层分析，从而建立以相关系数为因子的层次聚类模型，并得到折线图、散点图、频次直方图、相关系数热力图以及层析树状图。进而进行品类汇总分析、变化趋势分析、分布频率分析、周期性分析、单品类汇总分析和相关性分析得到蔬菜各品类及单品销售量的分布规律如季节性分布、同一聚类具有相似分布，和不同品类、单品之间具有的正相关关系，负相关关系及弱相关关系，具体结果会在正文以及附件中体现。

针对问题二，根据问题一得到的各品类及单品的分布规律与相关关系，以品类为单位，在定义品类的平均成本加成定价法后，进行计算品类售价。针对销售总量与成本加成定价关系的求解问题，由于数据的较大波动性，含有较多高频噪点，故对数据进行平滑处理。通过滑动窗口平滑处理，傅里叶变换以及二次方程回归建立较稳定的销售量-平均定价的关系模型；针对批发价的预测问题，由于部分品类的批发数据于 2023 年 5 月以及 6 月含有缺失值，故通过时间序列模型 ARIMA 进行粗粒度的数据填充，再对于 2023 年 7 月 1 日-7 日进行细粒度预测批发价的变化；针对收益的多目标优化问题，我们考虑定价以及批发价格两个因子的影响，建立非线性规划模型，得到最优的品类补货策略和定价策略，从而最大化商超收益。

针对问题三，本研究针对商超中生鲜蔬菜的进货与定价问题，构建了一个多约束条件下的优化模型。模型旨在在满足商超销售空间和最小陈列量的前提下，最大化商超的收益。通过数据预处理、单位利润的计算以及约束条件的线性规划模型，本研究找到了一种高效的进货和定价策略。最后，模型还考虑了不同销量水平下的加价率，以进一步优化模型的结果。本模型不仅适用于本研究中的特定场景，还可以扩展到其他具有相似约束和目标的零售问题。

针对问题四，在生鲜商超中，蔬菜类商品的保鲜期短且品相易随销售时间而变差，因此，补货和定价策略的制定对商超运营至关重要。本文探讨了采集客流量、消费者反馈、季节性因素、天气和库存状况数据的重要性，以支持商超更好地制定补货和定价决策的数学建模方法。这些数据不仅提供了决策支持，还有助于改善客户忠诚度、满足市场需求、提高产品质量和降低库存成本，从而在竞争激烈的市场中获得竞争优势。

综上所述，本文通过多角度、多层次、全面地分析得到了可视化的可靠蔬菜种类的分布规律及相关关系，相较于单一的分析更具说服力。经过分析验证，本文的模型具有合理性和一定的现实意义。最后，本文对模型进行了优、缺点评价与模型推广，得出该模型还可以向商超出售的其他商品和生活的其它方面进行推广的结论。



**关键字：层次聚类　ARIMA 时间序列分析　贪心算法　多目标组合优化**

# 一、 问题重述

## 1.1 问题背景

在实际中，生鲜商超里的蔬菜类商品保鲜期均较短，上架时间越长，其质量越差，且大部分蔬菜若在当日未售出，第二天只能丢弃。所以一般情况下，每天超市都会依据商品的销售历史和需求情况补充库存。同时因超市出售的蔬菜品类繁多、产地不同，并且通常在凌晨进货，商家还需要在不知道确切单品和进货价格的情况下提前做出蔬菜补货的决策。

## 1.2 问题提出

蔬菜的定价一般采用“成本加成定价”方法，商超对运损和品相变差的商品通常进行打折销售。可靠的市场需求分析，对补货决策和定价决策尤为重要。从需求侧来看，蔬菜类商品的销售量与时间往往存在一定的关联关系；从供给侧来看，蔬菜的供应品种在 4 月至 10 月较为丰富，商超销售空间的限制使得合理的销售组合变得极为重要。

某商超采用"成本加成定价"法，并将有损伤和和质量变差的商品打折出售。市场需求分析对进货和定价决策至关重要，就需求方面来看，蔬菜销售量与时间存在相关性；而从供给方面，4 月至 10 月蔬菜供应品类十分丰富，商超销售空间限制使得销售组合至关重要。通过建立数学模型，解决下列问题：

1) 请对该商超经销的 6 个蔬菜品类商品信息进行分析，找出各品类和单品的销售量分布规律和相互关系。
2) 请对该商超各蔬菜品类销售总量与成本加成定价的关系进行分析，并预计各品类在 2023 年 7 月 1-7 日的日进货总量和定价策略以最大化商超收益
3) 由于销售空间有限，请为商超制定单品进货计划使得可售单品总数控制在 27-33 个，并确保每个单品的进货量达到 2.5 千克。并请在满足市场需求的前提下，根据 2023 年 6 月 24-30 日的可售品种数据，给出 7 月 1 日的单品进货量和定价策略，以实现商超收益最大化。
4) 为了完善进货和定价策略，针对商超需另外采集的数据给出建议和原因。

# 二、 问题分析

问题 1：问题的研究对象是蔬菜类商品的销售量，研究内容为各品类及单品的销售量的分布规律及相互关系。该问题希望挖掘出不同蔬菜类商品间存在的关联关系，使用可视化分析，描述性统计分析，相关系数分析和聚类分析可以较好地描述销售量的分布规律和相互关系。基于作出的折线图，散点图，频率直方图，对不同品类的蔬菜进行可视化分析，得到变化趋势分析结果、品类汇总分析结果和分布频率分析结果；对相同品类不同单品的蔬菜进行周期性分析和单品类汇总分析；在描述性统计分析方面，统计分析是分析销售数据最基本的方法之一。通过统计每个品类和单品在整个销售周期内的销售量，可以计算出每个品类和单品在不同时间段的总销售量，并观察其变化规律。通过使用标准差、最值等统计方法，来描述销售量的分布情况；相关系数分析是用来评估品类之间的关联程度的一种方法。通过计算不同品类之间销售量的相关系数，判断品类之间是否存在相关性。如果相关系数较高，则说明这两个品类之间的销售量变化比较接近；如果相关系数较低，则说明它们之间的销售量变化差异较大，以此来得到各品类之间的关系，并为后续题目提供依据；聚类分析是一种常见的数据挖掘方法，可以将具有相似销售模式的单品进行归类，以此发掘出各单品之间的关联关系。根据单品的销售量，将

其聚类成若干个类别，并观察不同类别之间的差异。通过此方法可以了解每个类别的销售规律和特点并根据这些规律为后续题目提供决策依据。

问题 2：问题的研究对象是蔬菜品类的销售总量，成本加成定价和补货总量，研究内容是前两者之间的定量关系以及未来一周日补货总量的预测和定价策略的多目标优化。首先，我们定义品类的成本加成定价法，使用单个商品的当天销售量作为权重相乘各自的当天平均定价，最终我们得到当天的品类平均定价；随后，我们将销售总量和成本加成定价绘制散点图，试图寻找存在的定量关系。由于数据的波动性较大，含有大量高频噪点，在不经过数据预处理的情况下，非线性回归并不能得到令人满意的结果。因此，我们采用滑动窗口对成本加成定价进行数据平滑处理。其中，窗口大小为 7，并在窗口累加器中除去当前窗口的两个极值。之后，我们进行傅里叶变换，寻找季节性，周期性的变化，为了防止模型对高频噪点敏感，我们在频域上低通滤波器后才进行傅里叶反变换至时域。最终我们使用预处理后的数据进行非线性回归，得到了蔬菜品类的销售总量-成本加成定价关系的一元二次方程，比较吻合市场规律，具有一定可解释性；根据题目和背景可知商超采用的是“成本加成定价”方法。由问题一可知蔬菜的供应品种在 4 月至 10 月较为丰富，而商超的销售空间有限。为了最大化收益，需要考虑的因素包括：蔬菜的进货成本、预期的销售量、市场需求、损耗率等。我们采用了以上阐述的 ARIMA 预测未来一周的批发价，然后基于预测的批发价得出预计销售量和成本数据利用贪心思想制定了定价策略。

问题 3：问题要求在满足一定约束条件下，制定单品的补货计划和定价策略，以最大化商超的收益。具体而言，需要控制可售单品总数在 27-33 个，并保证每个单品的订购量至少满足最小陈列量 2.5 千克的要求。为了解决这个问题，本文将其简化为一个优化问题，即在满足约束条件的前提下，求得一个能够最大化收益的进货方案和定价策略。首先，剔除一些异常数据，例如在 3 年内 80%以上时间未售出的商品，以确保数据的有效性。接下来计算每个单品的预测单位利润，即预计销量乘以（定价-批发价）×（1-损耗率）。这样可以为每个单品确定一个单位利润的指标，用于衡量其对商超收益的贡献程度。然后使用组合优化方法，本文中使用了约束条件的线性规划模型，来求解该问题。通过构建一个数学模型，考虑所有可售单品的进货量和定价方案，就可以通过优化算法求得在满足约束条件下最大化收益的策略。此外，还可以沿用问题二中的思路，基于预测的销售量来确定加价率（即定价方案）。根据预期销售量的不同范围，可以设定不同的加价率。例如，对于低销量商品，可以采用较高的加价率以获取更高的单位商品收益；对于中等销量和高销量商品，可以采用较低的加价率以实现薄利多销。最后，综合前述步骤得到的最优进货量和定价方案，对进货价格进行适当调整，使得贪心算法得到的单位拟定价乘以拟定最优进货量大于仅考虑成本加成定价的理论最大收益。这样就可以求解并得到在约束条件下使商超收益最大化的进货和定价策略。

问题 4：商超面临以下挑战：

1. 蔬菜保鲜问题： 蔬菜类商品的保鲜期短，销售时间与品相密切相关，需要每日补货。如何确定合理的补货策略以减少损失？

2. 定价问题： 商超采用“成本加成定价”方法，但如何确定最佳定价策略，尤其是在不确切知道进货价格的情况下？

3. 单品补货问题： 商超希望制定单品的补货计划，但需要控制可售单品总数并满足市场需求，如何实现最佳的单品补货策略？

4. 数据采集问题： 商超需要采集客流量、消费者反馈、季节性因素、天气和库存状况数据。这些数据如何帮助解决上述问题，并如何以数学建模方法分析和利用这些数据？

本文通过详细分析这些问题，提供了在制定补货和定价策略时如何采集、分析和利用多种数据类型的建议。这些数据和方法不仅有助于提高商超的运营效率，还可以提高客户满意度，增加竞争力。

## 三、 模型假设

1.三年内销售天数不超过 5 天的不分析；
2.数据长度不一的尝试填充到同类的最长长度；
3.层次聚类使用 0.5 阈值；
4.成本加成定价使用平均定价；
5.品类批发价使用平均批发价；
6.在一段时间内，部分商品的批发价缺失，缺失比例大概为 79.21%，故认为该时段未购买此商品，设置为 0。
7.对于退货商品，不认为是可盈利部分的销售量。
8.损耗率数据是可靠的，并且在模型的时间范围内不会发生显著变化。
9.商超的运营成本、库存成本等未明确给出，因此在模型中不考虑。

## 四、 符号说明

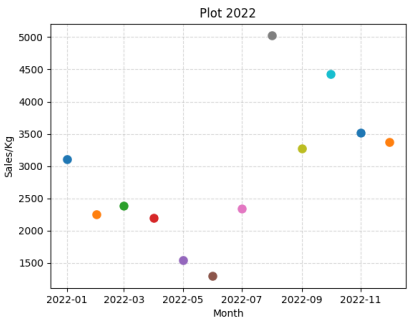
| 符号 | 定义 | 单位 |
|---|---|---|
| $a$ | 某一商品 | / |
| A | 商品品类 | / |
| $\mathrm{SoldW(a,t)}$ | a 商品在第 t 个区间卖出的重量 | Kg |
| $\mathrm{Price(a,t)}$ | a 商品在第 t 个区间卖价 | 元 |
| S(a) | a 商品在 T 区间内的成本加成定价 | 元 |
| S(A) | 品类 A 在 T 区间内的成本加成定价 | 元 |
| $X_i$ | 第 $i$ 个单品的进货量 | Kg |
| $C_i$ | 第 $i$ 个单品的单位批发价。 | 元 |
| $S_i$ | 第 $i$ 个单品的预计销量。 | Kg |
| $L_i$ | 第 $i$ 个单品的损耗率。 | / |

# 五、 模型的建立与求解

## 5.1 问题一模型的建立与求解

### 5.1.1 可视化分析

a) 散点图



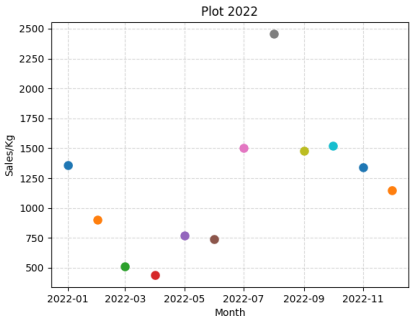
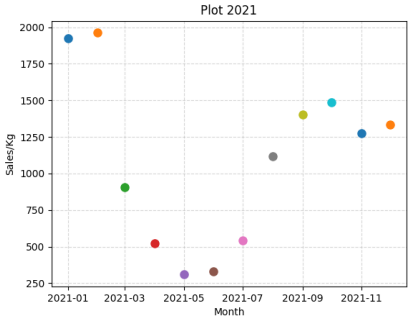




**图 1　辣椒类 2022 年销量散点图**　　**图 2　花菜类 2022 年销量散点图**

变化趋势分析：以花菜类和辣椒类在 2022 年的的散点图为例，可以看出两品类蔬菜的变化趋势相同，说明这两种品类的蔬菜具有一定的相关性。这可能意味着这两种品类的蔬菜在市场上存在着某种共同的需求因素或者受到相似的季节性影响。



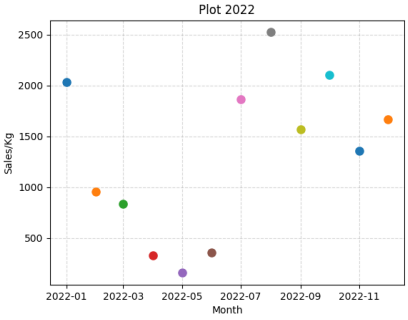
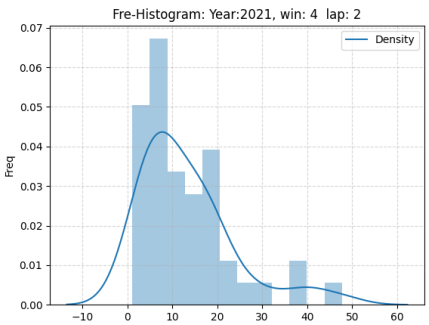




**图 3　水生根茎类 2021 年销量散点图**　　**图 4　水生根茎类 2022 年销量散点图**

周期性分析：以水生根茎类蔬菜为例，从图中可以看出销量散点图呈现出明显的周期性波动，且这种波动在不同年份之间重复出现，可以初步判断水生根茎类蔬菜存在季节性变化。这意味着销售量在特定的月份或季节中可能会有较大的增加或减少，这对于制定进货策略和定价策略有很好的指导作用，所得到的进货和定价策略可以更好地满足市场需求，并避免因季节性波动而导致的过剩或短缺问题。

b) 频次直方图



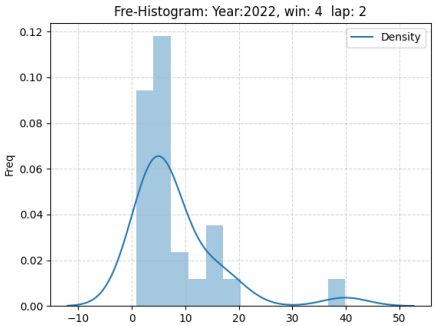




**图 5 水生根茎类荸荠 2021 年频次直方图**　　**图 6 水生根茎类荸荠 2022 年频次直方图**

该频次直方图的横坐标为时间，纵坐标为销量，使用了滑动窗口来减小波动，这样就平滑直方图曲线，去除局部异常值，高亮整体趋势。由图中可以看出销售量的整体分

布情况和销售量的波动周期，据此可以制定相应的进货和定价策略。

### 5.1.2 描述性统计分析

a) 问题定义：

假设我们有 N 个蔬菜品类，每个品类在 T 个时间段内的销售量数据。我们的目标是通过描述性统计分析来揭示销售量的基本特征和分布情况。

b) 数学表达式：

$N$：蔬菜品类数目　$T$：时间段数目　$X[i,j]$：第$i$个品类在第$j$个时间段的销售量

c) 算法流程：

初始化$X$矩阵，将销售量数据填入矩阵中。

对于每个品类$i$，计算其在每个时间段$j$的平均销售量：$Avg[i,j] = sum(X[i,j]) / T$。

对于每个品类$i$，计算其整个销售周期内的总销售量：$Total[i] = sum(X[i,:])$。

计算每个品类在整个销售周期内的最大值、最小值、中位数、标准差等统计指标。

绘制销售量的变化趋势图和销售量分布图，以观察不同品类在不同时间段的销售情况。利用统计方法，进一步分析销售量的分布情况和偏度[1]。

根据分析结果，了解不同品类的销售量情况，发现异常值和特殊模式，并提出后续的管理和决策建议。

**表 1　每日销售量波动的统计信息**

| | 最小值 | 最大值 | 范围 | 标准差 | 平均值 | 中位数 | 和 | 个数 |
|---|---|---|---|---|---|---|---|---|
| 水生根茎类 | 0 | 1057 | 1057 | 253.691 | 168.947 | 86 | 3210 | 19 |
| 花叶类 | 0 | 919 | 919 | 244.453 | 180.9 | 52.5 | 18090 | 100 |
| 花菜类 | 2 | 1076 | 1074 | 391.493 | 374.8 | 337 | 1874 | 5 |
| 茄类 | 3 | 1022 | 1019 | 351.742 | 320.7 | 139 | 3207 | 10 |
| 辣椒类 | 0 | 858 | 858 | 244.721 | 239.022 | 192 | 110756 | 45 |
| 食用菌 | 0 | 821 | 821 | 167.498 | 131.361 | 51 | 9458 | 72 |
| 品类 | 1050 | 1085 | 35 | 12.974 | 1079 | 1085 | 6474 | 6 |

**表 2　每月销售量波动的统计信息**

| | 最小值 | 最大值 | 范围 | 标准差 | 平均值 | 中位数 | 和 | 个数 |
|---|---|---|---|---|---|---|---|---|
| 水生根茎类 | 0 | 36 | 36 | 9.29 | 8.895 | 7 | 169 | 19 |
| 花叶类 | 0 | 35 | 35 | 10.301 | 10.24 | 6 | 1024 | 100 |
| 花菜类 | 1 | 36 | 35 | 13.075 | 16.2 | 13 | 81 | 5 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 茄类 | 1 | 36 | 35 | 12.603 | 15.4 | 9 | 154 | 10 |
| 辣椒类 | 0 | 35 | 35 | 9.562 | 11.889 | 10 | 535 | 45 |
| 食用菌 | 0 | 33 | 33 | 7.309 | 7.653 | 5.5 | 551 | 72 |
| 品类 | 36 | 36 | 0 | 0 | 36 | 36 | 216 | 6 |

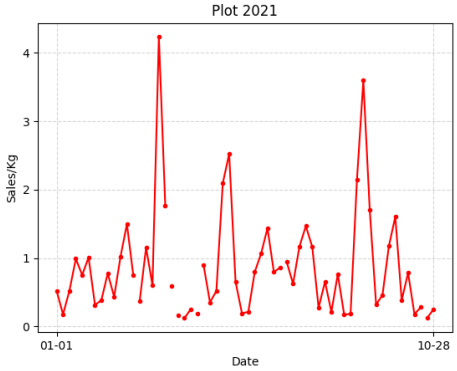



**图 7　灯笼椒每日销售折线图**

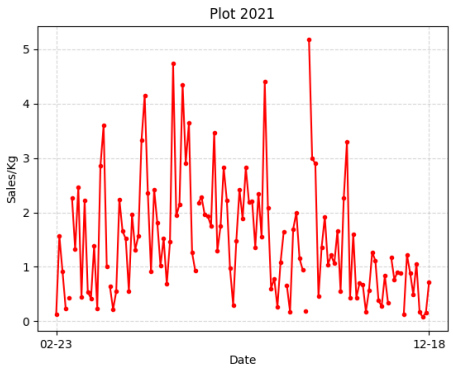



**图 8　红尖椒每日销售折线图**

单品类汇总分析：红尖椒和灯笼椒均属于辣椒类这一品类，从图上看季节性是相似的，但这一品类季节性并不强，因为曲线在不同季节的波动并不大。

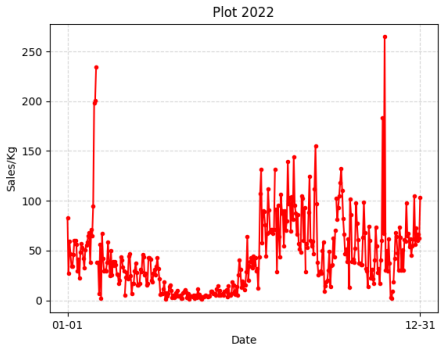



**图 9　水生根茎类 2022 年每日销售折线图**

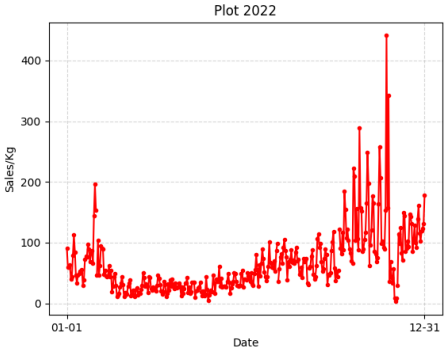



**图 10　食用菌 2022 年每日销售折线图**

品类汇总分析：如图水生根茎类与食用菌这两个蔬菜品类的季节性相似，以年的视角来看销售量波动较大，季节性较为明显。

### 5.1.3 相关系数分析

a) 问题定义：

在蔬菜销售量分析中，通过评估不同品类之间销售量的相关程度，以得到它们之间的关联关系。Spearman 相关性分析是一种非参数统计方法，用于评估两个变量之间的相关程度，特别适用于评估顺序型数据之间的关联关系。本文使用 Spearman 相关性分析来评估不同品类之间销售量的相关程度[2]。

b) 数学表达式：

$N$：蔬菜品类数目　　$X[i]$：第$i$个品类的销售量　　$Y[j]$：第$j$个品类的销售量

$r[i,j]$：第$i$个品类和第$j$个品类之间的相关系数

Spearman 相关性系数（ρ）的计算公式如下：

$$\rho = 1 - (6\sum d_i^2) / (n(n^2 - 1)) \quad (1)$$

其中，$d_i$表示两个变量在顺序上的差异（即排名之间的差异），n 表示观测样本数量。

c) 算法流程：

收集蔬菜销售量的数据，并将不同品类的销售量按照大小进行排序，得到每个品类的排名。

计算每个品类对应的销售量排名的差异（$d_i$）。

根据公式计算 Spearman 相关性系数（ρ）的值。如果相关性系数接近 1，表示两个品类之间存在强正相关关系；如果接近-1，表示存在强负相关关系；接近 0 则表示相关性较弱或者没有线性相关关系。

通过检验相关性系数是否显著，可以判断销售量之间的相关性是否具有统计学意义。

d) 结果分析

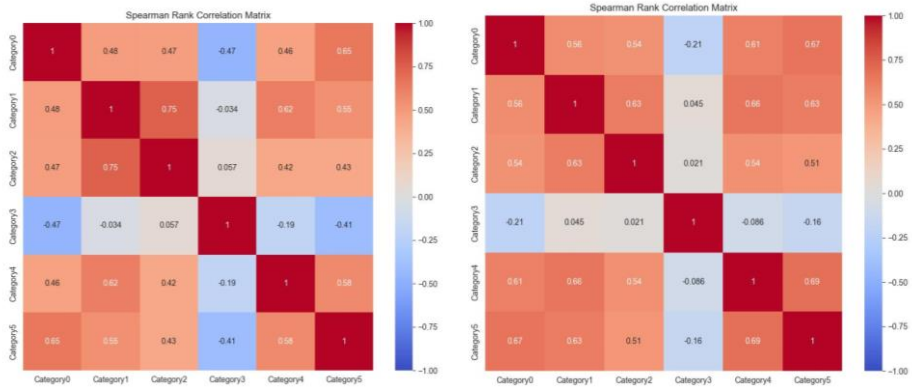



**图 11 各品类以月为单位所得相关性分组映射热力图　图 12 各品类以天为单位所得热力图**

从图 11 中可以看出品类 1 与品类 0、2、4、5，品类 4 与品类 5 的相关系数大于 0.5，相关性较高，即花叶类与水生根茎类，花菜类，辣椒类，食用菌其中任何一类的相互关系都较为明显。这说明买家在购买花叶类蔬菜时有很大可能同时购买水生根茎类，花菜类，辣椒类，食用菌其中任何一种或几种蔬菜。

从图 12 中可以看出品类 0、1、2、4、5 其中任意两类的相关系数均大于 0.5，相关性较高，即水生根茎类，花叶类、花菜类，辣椒类、食用菌其中任意两类的相互关系都较为明显。这说明买家在购买蔬菜时有很大可能同时购买花叶类、水生根茎类，花菜类，辣椒类，食用菌其中任何几类的蔬菜。

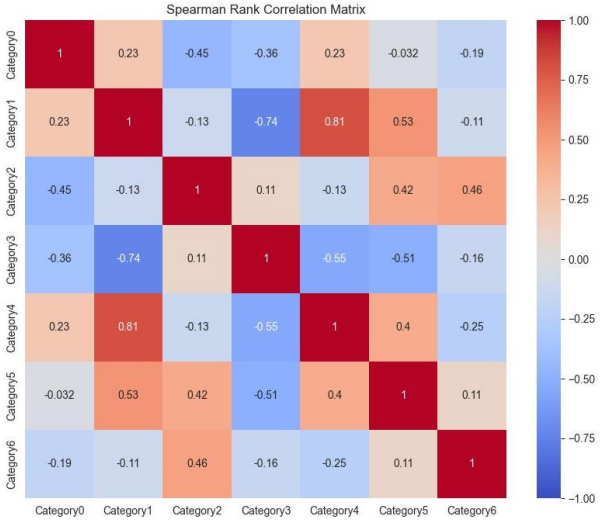



**图 13 茄类各单品以月为单位所得相关性热力图**

从图 13 中可看出单品 1 与单品 4、5 相关系数分别均大于 0.5，正相关性较高，单品 3 与单品 1、4、5 相关系数分别均小于-0.5，负相关系数较高。即可得出花茄子与大龙茄子、长线茄、青茄子均呈较为明显负相关关系，大龙茄子与长线茄、青茄子呈较为明显正相关关系。这说明买家在购买大龙茄子时有很大可能同时购买长线茄和青茄子其中一种或两种蔬菜，而在购买花茄子时有很大可能不购买大龙茄子、长线茄、青茄子其中任何一种蔬菜。

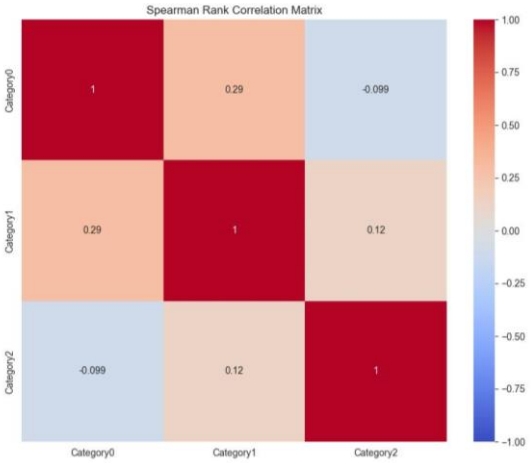



**图 14　花菜类各单品以月为单位所得热力图**

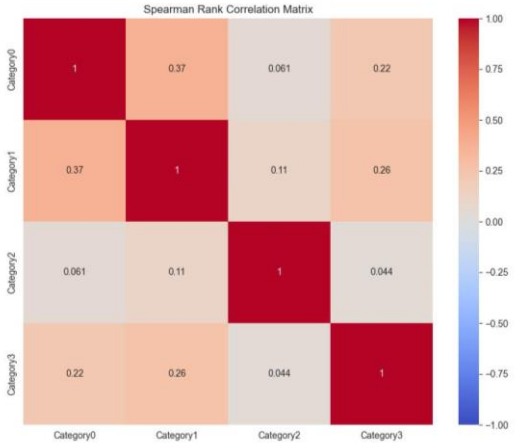



**图 15 花菜类各单品以天为单位所得热力图**

由图 14 中可以看出，花菜类各单品间没有明显的相关关系，相关性均未大于 0.5。这说明买家在购买这类蔬菜时，并没有明显的可能性购买同类另一种蔬菜。

由图 15 可以看出，花菜类各单品间没有明显的相关关系，相关性均未大于 0.5，这与以月为单位所得的相关性结果一致。

### 5.1.4 聚类分层分析

a) 问题定义：

在蔬菜销售量分析中，我们希望利用销售量和批发价格等特征对蔬菜进行分层聚类，以发现不同层次间的差异。

b) 数学表达式：

$N$：蔬菜品类数目　　$M$：蔬菜特征数目

$X[i,j]$：第$i$个品类的第$j$个特征值（如销售量、批发价格，此处使用相关性分析中得到的 distance 的平均值）

$C[k]$：第$k$个聚类的中心点（由$M$个特征值组成）

c) 算法流程：

初始化$X$矩阵，将蔬菜的特征值填入矩阵中。

使用分层聚类算法，将蔬菜品类进行分层聚类。选择合适的聚类方法可以根据数据特点和需求确定。

根据聚类结果，生成聚类层次图（树状图），展示不同层次之间的聚类关系。

可以根据需求选择合适的切割点，将聚类结果划分为不同的层次，并分析每个层次之间的差异。[3]

根据分析结果，可以进一步研究每个层次中不同聚类的特征和行为，并制定相应的营销策略。

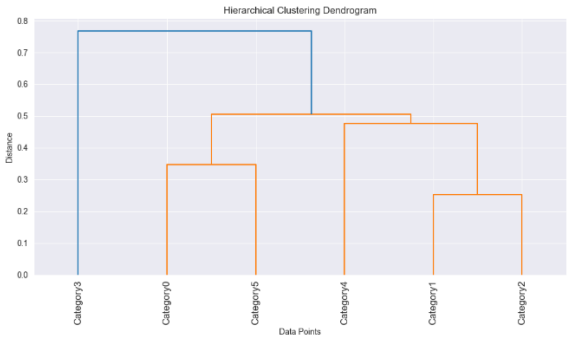



**图 16 各品类以月为单位所得层析树状图**

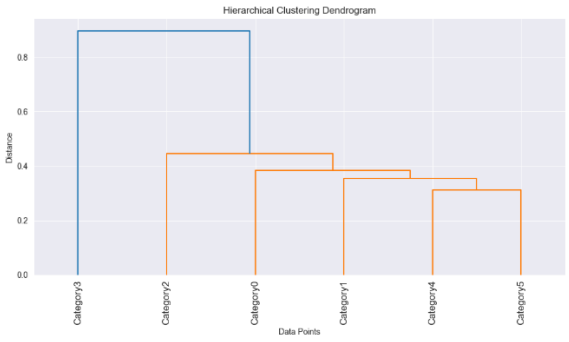



**图 17 各品类以天为单位所得层析树状图**

图 16 和 17 结果显示在以月为单位和以天为单位两种情况下，品类 0、1、2、4、5 经过层次聚类后属一类，品类 3 属于另一类。即水生根茎类，花叶类，花菜类，辣椒类，食用菌属于一类，茄类属于第二类。这说明同一聚类的商品有可能同时卖出，在进货时需要同时增减进货量。

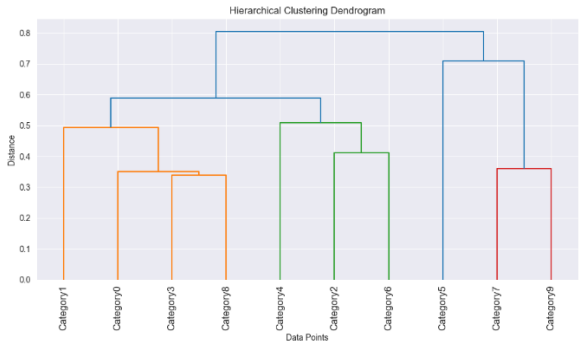



**图 18 水生根茎类以月为单位所得层析树状图**

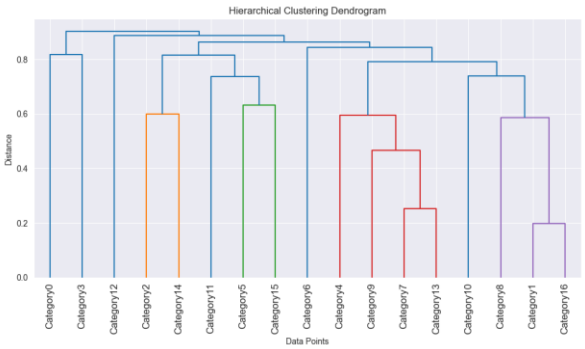



**图 19 水生根茎类以天为单位所得层析树状图**

以天为单位所得聚类的数量多于以月为单位所得聚类的数量，这说明以天为单位的聚类可能更加细致和精确。这是因为蔬菜类商品的保鲜期比较短，品相也随时间而变化。以天为单位的聚类可以更好地反映商品销售与时间的关联关系，也更能准确分析不同商品的销售情况和供应情况，对商超的补货决策和定价决策都具有指导意义。而以月为单位的聚类则可能会忽略商品销售的微小波动，无法捕捉到商品销售与时间的精细变化，因此聚类数量相对较少。

## 5.2 问题二模型的建立与求解

在问题 2 中的目标是为商超各蔬菜品类制定一个有效的补货计划和定价策略，以最大化商超的收益。为了实现这一目标，本文将按照以下步骤进行模型建立和分析：

### 5.2.1 获取各蔬菜品类的销售总量

首先需要计算每个蔬菜品类的历史销售总量数据。这些数据可以从附件 2 中的销售流水明细数据中获得。对于每个品类，将每天的销售数量相加，得到一个时间序列，表示该品类的销售总量，对每个品类进行单独分析，以了解它们的销售趋势。

### 5.2.2 分析各蔬菜品类的销售总量与成本加成定价的关系

通过对历史销售总量和成本加成定价数据的分析，可以建立销售总量与售价之间的关系。这有助于理解价格变化对销售的影响。成本加成定价是商超在一天内对不同蔬菜品类的单品定价策略。具体地，本文使用以下公式来计算并定义某个品类在某一天的平均售价：

其中，$a$为某一单品，$A$为某一品类，$t$表示在第$t$个区间，所以有$a \in A, t \in T$. $SoldW(a,t)$ 代表$a$在第$t$个时间区间售出重量。$Price(a,t)$代表$a$在第$t$个区间卖价。

$$Price(a,t) = \begin{cases} \max\left(Price(a,t_1), Price(a,t_2) \ldots\right) \\ \text{avg}\left(Price(a,t_1), Price(a,t_2) \ldots\right) \end{cases} \tag{2}$$

对于某一件单品，在一天内可以被多种单价销售，不同价格售出的质量和单数都是多变的，因此可以求出一件单品在一天内的最大销售单价，对各售出单数加权平均可求得单品一天内的平均售卖单价。

而考虑批发价，通过读取附件 3 提供的数据可知某一品类在某一天的批发价格是固定的（指代现实中批发商的交货价格），因此可以定义在一天内商超对一个品类定价为：

$$S(A) = \frac{\sum_{a \in A}\left(S(a) * SoldW\,(a,T)\right)}{|A| * \sum_{a \epsilon A} SoldW(a,T)} \tag{3}$$

其中$S(a)$为该单品的批发价。$SoldW(a,T)$表示在$T$时间内单品$a$售出的总质量，当$T = 1$时即为一天内的数据。由于对一件单品来说，$a$只有本身一个类别，因此$a = 1$，因此有：

$$S(a) = \frac{\sum_{t \epsilon T}\left(Price(a,t) * SoldW\,(a,T)\right)}{\sum_{t \epsilon T} SoldW(a,T)} \tag{4}$$

### 5.2.3 确定定量影响售价的因素

具体来说，在模型中考虑以下因素：

- 成本因素：考虑蔬菜的采购成本、运输成本和损耗率。高成本可能需要较高的售价来维持利润。
- 市场需求因素：基于需求量的变化，可以根据供需关系来调整售价。高需求可能支持更高的售价。
- 利润因素：目标是最大化利润，因此需要确保售价足够高以覆盖成本并实现额外的利润。

### 5.2.4 傅里叶级数拟合得到销量预估曲线

由于拟合所得散点图的效果很差，无论是对数线性回归模型的结果还是非对数线性回归模型的结果均难以找出规律，说明不能使用简单的线性回归进行拟合。

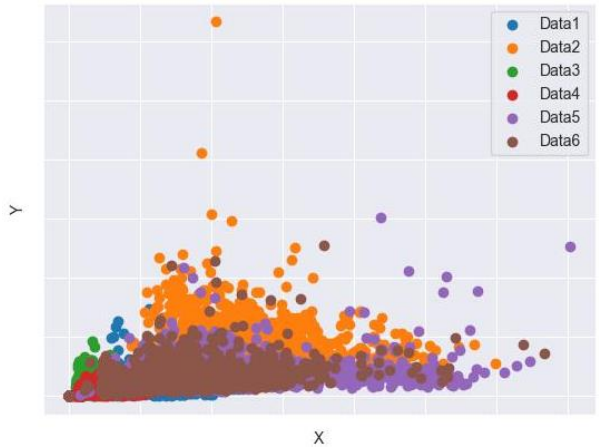



**图 20　非对数线性回归模型结果**

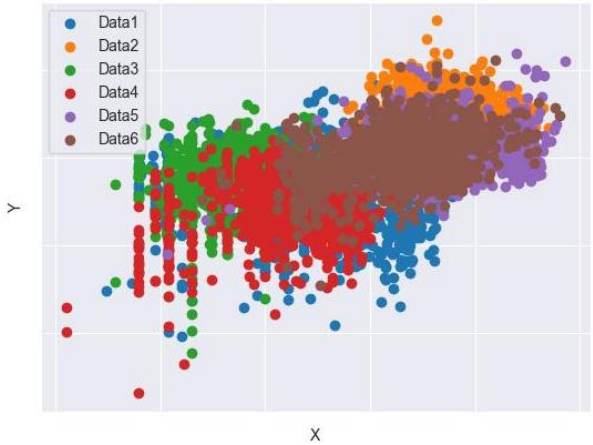



**图 21　对数线性回归模型结果**

接下来进行极差标准化，减少异常值对数据的影响，进行数据归一化并保留了数据相对顺序，这可以提升模型的性能，防止某些特征的权重过大或过小对模型训练产生不

利影响。它通过将原始数据减去最小值，再除以最大值与最小值之差得到标准化后的数据。公式如下：

$$x_i^* = \frac{x_i - min(x)}{max(x) - min(x)} \tag{5}$$

其中，$x_i$表示原始数据中第 $i$ 个观测值，$min(x)$表示所有数据中的最小值，$max(x)$表示所有数据中的最大值，$x_i^*$表示标准化后的第$i$个观测值。

绘制拟合曲线时发现原有散点振幅过大，如果数据的散点振幅过大，意味着数据点之间存在较大的距离或差异，可能导致拟合曲线不够平滑，难以捕捉到数据的整体趋势。于是使用滑动窗口进行平滑处理。滑动窗口是一种移动的数据窗口，用于处理时间序列或一维数据。它在数据上滑动，每次处理一个窗口内的数据。滑动窗口中的大小可以根据具体需求设定，通常为固定长度，在此问题中使用了大小为 7 的窗口。平滑处理是一种数据处理技术，用于减少数据中的噪声或波动，使其更具可读性或使得数据趋势更加明显。在平滑处理中，方法是先去除窗口中的最大值和最小值，然后再计算剩余值的平均值来代表该窗口的数值。这样可以减少数据的波动性，更好地展现数据的整体趋势。

接下来使用傅里叶级数拟合、傅里叶变换、低通滤波器和傅里叶反变换这一信号处理算法组合对数据进行拟合。

傅里叶级数拟合是一种将周期性信号表示为一系列正弦和余弦函数的方法。它通过将信号分解为不同频率的谐波成分，并计算每个谐波的振幅和相位，将原始信号拟合成一组正弦和余弦函数的线性组合。这样做可以通过少量的谐波来近似表示复杂的周期性信号。傅里叶变换：傅里叶变换是一种将信号从时域转换到频域的方法。它将一个信号分解成不同频率的正弦和余弦函数（频谱），显示信号在各个频率上的能量分布。傅里叶变换可以提供有关信号的频率成分、相位和幅度等信息，使得信号的频域特征更加可观。低通滤波器：低通滤波器是一种用于去除高频成分的滤波器。它允许低频信号通过，而抑制高频信号。在傅里叶变换后，使用低通滤波器来去除频谱中的高频噪声或不需要的高频成分，以此限制频率最大值，保留低频成分，防止对高频噪声敏感。傅里叶反变换：傅里叶反变换是一种将信号从频域转换回时域的方法。它将频谱重新合成为原始信号，恢复信号的时间表示。傅里叶反变换可以应用于经过傅里叶变换和滤波处理后的信号，以获得去除噪声或不需要的高频成分后的重构信号。

综上所述，傅里叶级数拟合通过将信号分解为正弦和余弦函数的线性组合来拟合周期性信号。傅里叶变换将信号从时域转换到频域，显示信号的频率特征。低通滤波器用于去除高频成分，保留低频成分。最后，傅里叶反变换将经过滤波处理后的频谱重新合成为时域信号，得到重构信号。以下这一算法组合所得到的部分结果图。

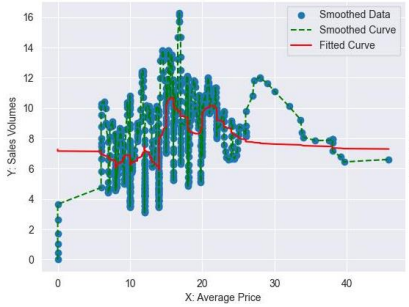



**图 22　花菜类拟合结果**

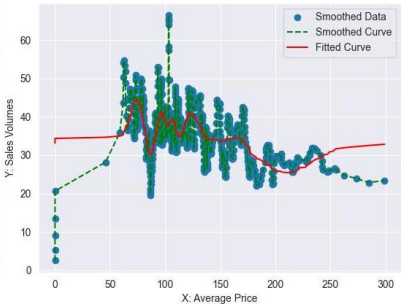



**图 23　花叶类拟合结果**

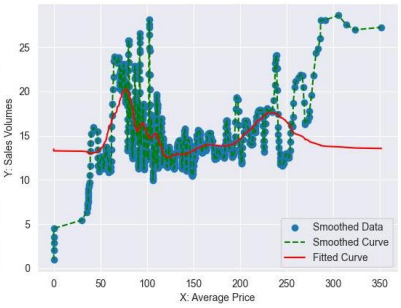



**图 24 辣椒类拟合结果**

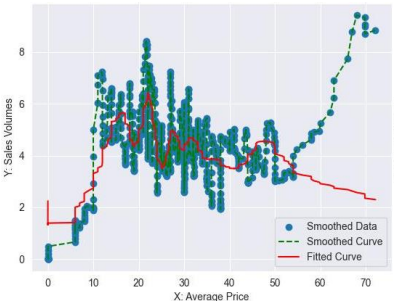



**图 25　茄类拟合结果**

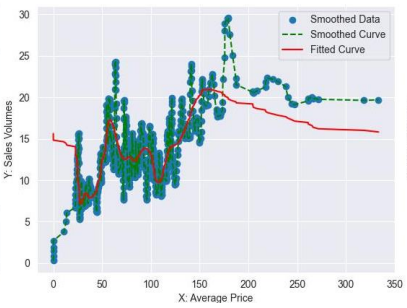



**图 26　食用菌拟合结果**

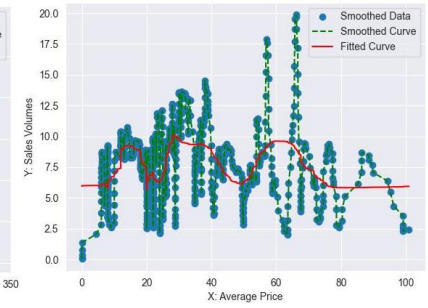



**图 27　水生根茎类拟合结果**

### 5.2.5 利用 ARIMA 进行时间序列分析

蔬菜销售在不同季节和时间段可能会有显著的变化（4 月至 10 月较为特殊），因此使用时间序列分析方法来预测未来一周的批发价格。具体来说，本文采用 ARIMA（自回归整数移动均值模型）方法，该方法可以考虑时间序列数据的趋势、季节性和噪声，以生成批发价格的预测结果。这有助于更准确地预测未来的价格趋势。

ARIMA 模型将结合自回归（AR）、差分（I）和移动平均（MA）的特性，使用过去的批发价格数据来预测未来批发价格。具体地，模型将通过自回归部分（AR）来捕捉批发价格的历史趋势，然后应用差分操作（I）以处理非平稳的时间序列数据，并使用移动平均部分（MA）来考虑噪声对批发价格的影响。ARIMA(p, d, q)模型表示自回归滞后阶数为 p（Autoregressive）、差分次数为 d（Integrated）、移动平均滞后阶数为 q（Moving Average）[4]

$$Y_t = c + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_\varphi e_{t-\varphi} + e_t \quad (6)$$

其中，Y 表示观测值，t 表示时间步，c 是常量，$\varphi_1$ $\varphi_2$ … $\varphi_p$是自回归系数，$\varepsilon_t$是白噪声误差项，$\theta_1$ $\theta_2$… $\theta_\varphi$是滑动平均系数，用于表示当前观测值与过去 q 个误差项的线性关系.

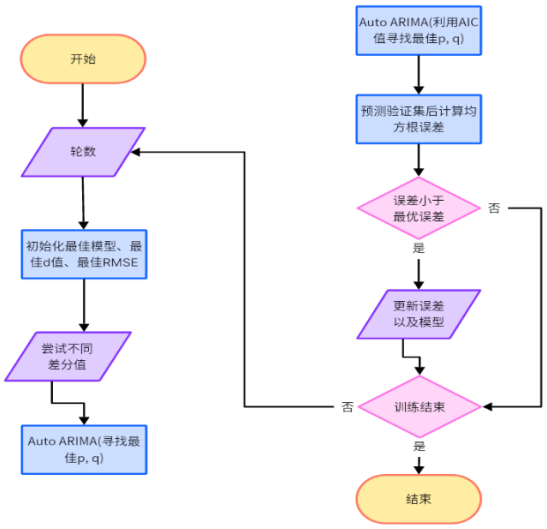



**图 28　ARIMA 模型流程图**

在选择 AR 部分的阶数时，可以使用自相关图（ACF）和偏自相关图（PACF）进行分析。ACF 表示不同滞后期的自相关系数，而 PACF 则表示在剔除其他滞后期影响后，某个滞后期与当前值之间的相关系数。根据 ACF 和 PACF 的截尾性质，可以选择 AR 部分的阶数。在选择 MA 部分的阶数时，同样可以利用 ACF 图进行分析。如果 ACF 在滞后期 k 之后截尾，则可以选择 k 作为 MA 的阶数。选择 I 部分的阶数主要是通过对时间

序列进行差分，找到使得序列变得平稳的阶数。可以通过观察不同阶数下的序列平稳性来选择 I 的阶数。

综合考虑这三个部分的阶数，通过自动化算法进行参数选择。本文使用了一种常用的方法 BIC（Bayesian Information Criterion）进行模型评估，选择具有最小 BIC 值的模型作为最优模型，参数选择结果见表 3.

**表 3　各品类 Arima 最优模型参数选择**

| | p | d | q |
|---|---|---|---|
| 水生根茎类 | 3 | 2 | 0 |
| 花叶类 | 3 | 2 | 1 |
| 花菜类 | 2 | 2 | 1 |
| 茄类 | 2 | 2 | 1 |
| 辣椒类 | 2 | 2 | 2 |
| 食用菌 | 1 | 2 | 3 |

不同的检验方法和显著性水平会导致不同的临界值，而本研究选取临界值 1%。为了满足稳定性检验的假设，通常需要确保 Test Statistic 的值小于相应的 Critical Value。这是进行时间序列稳定性检验的常见做法。如果 Test Statistic 的值大于等于 Critical Value，则不能拒绝原假设，表示时间序列可能是不稳定的。训练数据的描述性分析见下表：

**表 4　批发价格训练数据的平稳性分析**

| | 水生根茎类 | 花叶类 | 花菜类 | 茄类 | 辣椒类 | 食用菌 |
|---|---|---|---|---|---|---|
| 检验统计量 | -9.69 | -5.67 | -6.06 | -6.33 | -6.80 | -7.024 |
| p 值 | $1.16*10^{-16}$ | $8.81*10^{-7}$ | $1.23*10^{-7}$ | $2.90*10^{-7}$ | $2.22*10^{-8}$ | $6.44*10^{-10}$ |
| 滞后阶数 | 7 | 13 | 13 | 13 | 13 | 13 |
| 临界值(1%) | -3.49 | -3.49 | -3.49 | -2.50 | -1.49 | -0.49 |
| 临界值(5%) | -2.89 | -2.89 | -1.89 | -0.89 | 0.11 | 1.11 |
| 临界值(10%) | -2.58 | -2.58 | -1.58 | -0.58 | 0.42 | 1.42 |

由于 5、6 月具有缺失值，先采用 ARIMA 方法通过前四个月的历史批发价格预测 5、6 月的批发价格，然后再采用 ARIMA 方法更为细致地预测未来一周的批发价格。以下是 5、6 月批发价的预测展示图：

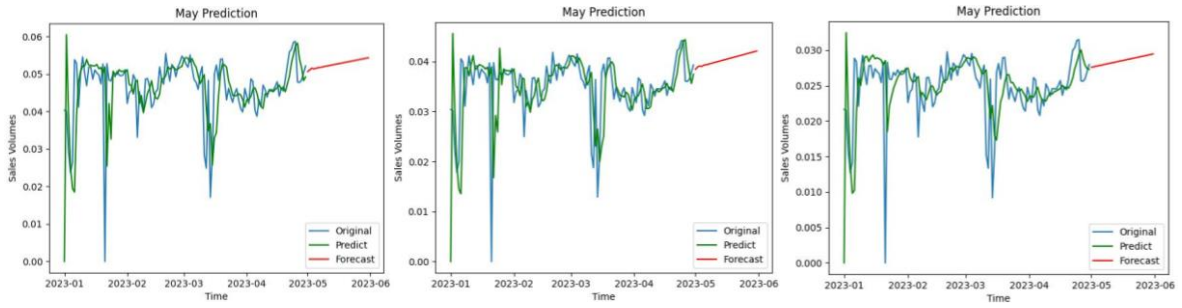



**图 29 辣椒类五月份批发价填充　图 30 茄类五月份批发价填充 图 31 食用菌五月份批发价填充**

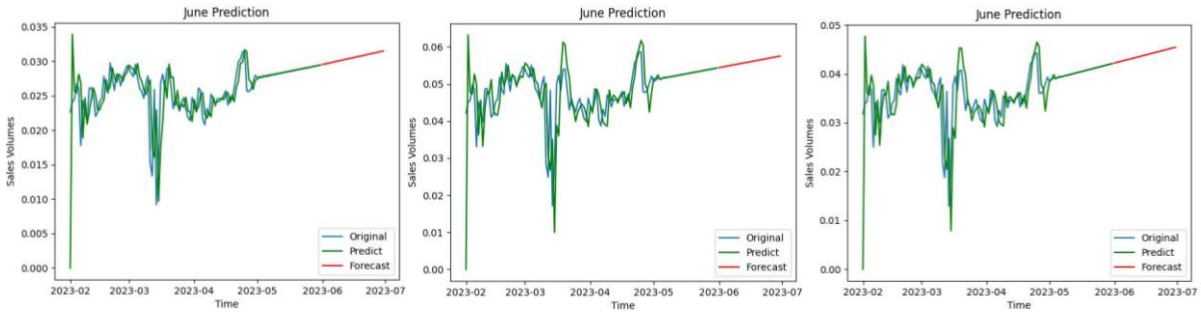



**图 32 辣椒类六月份批发价填充　图 33 茄类六月份批发价填充 图 34 食用菌六月份批发价填充**

然后根据 3、4 月的历史批发价格和上面预测得到的 5、6 月的批发价格预测得到 7 月的批发价格，如下：

### 5.2.6 利用贪心算法建立定价模型

基于 ARIMA 方法预测未来一周的批发价格，并考虑预测的销售量、成本数据以及上述的定量因素，可以建立一个定价模型，该模型接收这些信息并计算每个蔬菜品类的合理售价。这个模型的目标是最大化销售利润，同时保持竞争力和满足市场需求。

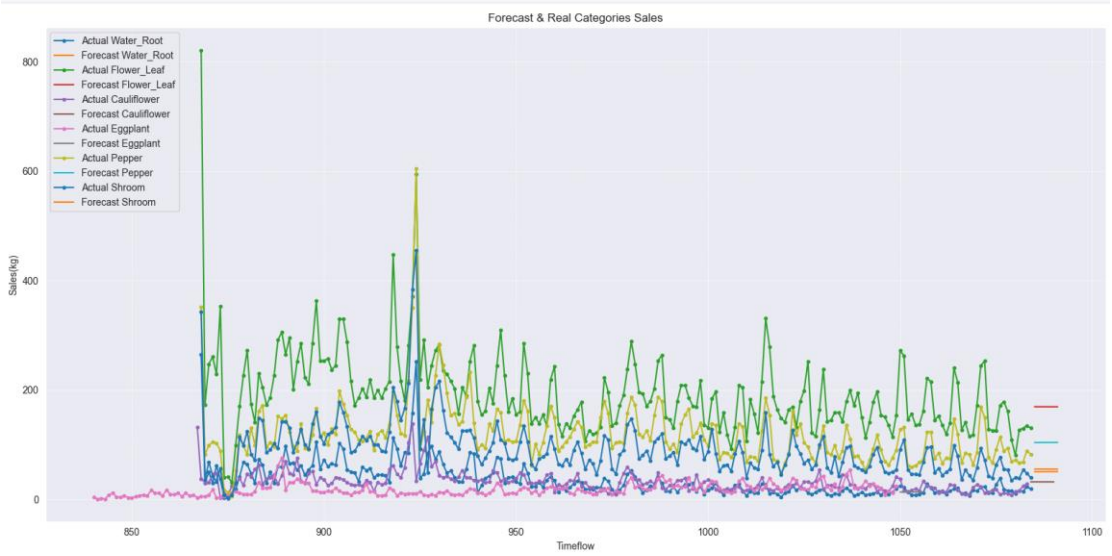



**图 38　所有类在 2023 年随时间变化的销售曲线**

最后，使用贪心算法来计算每个蔬菜品类的最终售价，该算法会考虑不同品类之间的竞争和协同效应，以确定最优的售价策略[5]，以最大程度地提高商超的总销售利润。具体来说，贪心算法将根据蔬菜品类之间的竞争和协同效应，以及商超的总销售利润为目标，逐个计算每个蔬菜品类的最终售价。在每一步中，算法会选择当前能够带来最大利润的售价策略。定价多少与利润直接挂钩，定价和销量相互影响，定价的策略关乎利润的大小以及最大利润方案是否可行。为了确保收益最大化，我们使用成本加成定价法（前面已说明），获得基准定价 base_price，通过贪心的思想放大基准价格作为定价策略。

贪心思想：对于低于某一阈值，（通过分析历史销售数据，可知各个大类的每日销售情况，设定一天的加价策略）认定为少量销售产品，贪心的思想需要我们从中获取更高的单位利润，因此加价率为最高。以此类推，中等销量的使用中间加价率，而销量最高的我们基于薄利多销的原则，加价率最低。

通过多次尝试不同的阈值和加价率，观测模型计算求解的结果，确定以下为我们的定价策略：如果预期销售量 ＜7 千克，加价率 ＝150%，如果预期销售量 ＜10 千克，加价率 ＝130%，其余，加价率 ＝120%。结果如下：

| 利润 | 7.1 | 7.2 | 7.3 | 7.4 | 7.5 | 7.6 | 7.7 | 一周 |
|---|---|---|---|---|---|---|---|---|
| 花菜类 | 321.91 | 321.79 | 321.67 | 321.55 | 321.43 | 321.31 | 321.19 | 2250.85 |
| 花叶类 | 11873.57 | 11872.43 | 11871.30 | 11870.16 | 11869.03 | 11867.90 | 11866.76 | 83091.15 |
| 辣椒类 | 8007.88 | 8011.21 | 8014.53 | 8017.85 | 8021.18 | 8024.50 | 8027.82 | 56124.97 |
| 茄类 | 571.38 | 570.82 | 570.26 | 569.70 | 569.14 | 568.58 | 568.01 | 3987.89 |
| 食用菌 | 4684.70 | 4683.52 | 4682.34 | 4681.16 | 4679.98 | 4678.80 | 4677.62 | 32768.12 |
| 水生根茎类 | 1272.20 | 1272.69 | 1273.19 | 1273.69 | 1274.18 | 1274.68 | 1275.18 | 8915.81 |
| 所有品类 | 26731.64 | 26732.46 | 26733.29 | 26734.11 | 26734.94 | 26735.77 | 26736.58 | 187138.80 |

### 5.3 问题三模型的建立与求解

#### 5.3.1 模型的数据准备

数据来源：提供的数据与中间过程文件。

输入所有单品的历史销售数据，剔除退货数据和销量过低的数据，即超过 80%的时间未出售的商品数据；

输入题目要求时间范围内可购买的单品在 7 月 1 日的预测的销量和批发价。

由问题二的定价策略（即成本加成定价法）确定 7 月 1 日可购买单品的定价

#### 5.3.2 变量定义、约束条件与目标函数

设 $n$ 为单品总数，27≤n≤33。$X_i$：第 I 个单品的进货量，$X_i \geq 2.5$。$C_i$:第 i 个单品的单位批发价。$S_i$：第 i 个单品的预计销量。$L_i$：第 i 个单品的损耗率。目标函数如下：

$$argmax(Profit_n)$$
$$st. \begin{cases} 27 \leq n \leq 33 \\ S_i = func(X_i), X_i \geq 2.5 \\ X_i < max\{X\} \end{cases} \tag{7}$$

$$Profit_n = max\{\sum_{i=1}^{n}[S_i * (P_i - C_i) * (1 - L_i)]\} \tag{8}$$

$Profit_n$表示总进货单品数量为 n 时，被选中的 n 件单品的总利润；$S_i = func(X_i)$表示销售量和进货量关系的函数。

$S_i = func(X_i)$含义：

定义函数 func 接收一个进货量，我们输出对应的销售量。计算销售量的思路如下：

1. 考虑边界条件：进货量必须至少大于等于这一天的平均销量，我们不能取最大销量是因为考虑到无法保证每日销量均为历史最优情况而实际销售量也可能大于历史数据中单品最高的单日销售总量，因此我们认定该单品大于历史日销量平均销售水平即是合理数据，在此基础上，通过后续的贪心思想可以确保计算结果收敛

2. 考虑滞售情况：当进货过多时并不意味着销量会一起增加，通过对历史数据的分析可知，销量对进货的收益曲线应该可以使用一个二次函数类的凸函数拟合。进货量过大会造成收益严重下滑。

Xi＜max{X} 含义：理论上若可以无限买卖，利润可以无穷大，实际上在一段时间内顾客的消费能力和商超的库存都是有限的。进货太多会导致至少以下两个问题：

1. 库存过剩导致蔬菜的潜在运损率提高，收益下降

2. 进货过多导致蔬菜滞售现象严重，而题目已明确指出隔日蔬菜不进行销售

因此我们可以考虑在模型计算中给进货量添加可变权重。当进货量超过某一阈值后（这一阈值后面会讲），收益衰减的权重大幅增加，权重确保当进货量趋于无穷大时，模型计算得到的收益趋于负无穷，保证模型的收敛性和可求解性。Xi < max{X}表示每一件单品的进货量必须小于我们设定的边界。

### 5.3.3 模型求解

a) 数据预处理

去除那些 3 年内 80%以上时间未售出的单品。

我们第一题中已经说明（退货订单占比小于 0.1%），我们输入的数据和后续的分析均建立在剔除掉退单的数据基础上，不考虑退货。这和将销售量和销售额小于 0 的值的异常值修改为 0，对模型的影响一致。

b) 计算单件商品的单位利润

$$\text{单件商品的单位利润} = S_i * (P_i - C_i) * (1 - L_i) \tag{9}$$

c) 有约束条件的线性规划模型

关于 n 存在的约束条件：可销售单品数在 27-33 之间，不同的单品数量意味着我们求和过程中不同 n 之间的单品组合是不同的。因此我们需要分别找出当 n 取 27，28，29，…，33 的时候，选择的单品和他们对应的补货-定价策略。

题目所给数据中只提出了最小的陈列量（可推得最小的进货量），因此我们需要考虑进货量的上界。通过考虑 2020 年，2021 年，2022 年在 4-6 月的 3 个月（即相同目标时间段售出产品的销量情况）内每天的日总销量

我们得知最大的总销量约为 258 千克。我们可以认定这是商超进货量增加导致收益下滑甚至赤字的阈值。这一阈值将作为我们回归模型求解的边界。日总销量约等于 258 千克，总质量均摊至当日每一件售出的单品（共 33 种），每一件单品能够分得大约 8 千克的收益最大化进货上限。通过对每一件单品的进货上下限规划和总进货量小于 240（比阈值偏低可以是我们的求解方案更适用于实际情况）千克。

采用组合优化求解模型，使用了有约束条件的线性规划模型可以求得寻找满足各种约束条件下收益最大化的进货方案

d) 定价策略

在问题二中使用了贪心算法来确定最优的进货和销售策略。贪心算法用于确定单品的最优进货量和定价方案。在问题三中，我们沿用贪心的思想来确定最优的加价率。由于第三问我们需要细化到每一件单品，而每一种单品的日销量和品类的日销量不存在直接的强相关性，因此采用第二问相同的策略，确定如下定价策略：

根据预测的销售量，采用动态的加价率：（以下阈值通过观察选择出的 33 种收益最高的单品得出，以下的预期销售量均指代一天内的总销售量）

如果预期销售量 ＜3 千克，加价率 ＝150%

如果预期销售量 ＜6 千克，加价率 ＝130%

如果预期销售量 ＞9 千克，加价率 ＝120%

### 5.3.4 最终结果与结论

通过以上各个步骤不仅构建了一个全面而精确的模型，而且还找到了一种在给定约

束条件下最大化收益的进货和定价方案。这一方案不仅考虑了单品的销售预测和成本，还通过组合优化和贪心算法，使得整体收益最大化，具有很高的实用价值。

| 单品名称 | 单价(元/千克) | 补货量(千克) | 单品名称 | 单价(元/千克) | 补货量(千克) |
|---|---|---|---|---|---|
| 四川红香椿 | 79.34 | 10.00 | 苋菜 | 20.68 | 2.58 |
| 洪湖藕带 | 70.93 | 5.00 | 上海青 | 19.23 | 2.50 |
| 西兰花 | 19.93 | 12.49 | 竹叶菜 | 23.10 | 2.50 |
| 小米椒 | 159.56 | 2.50 | 红尖椒 | 31.70 | 2.50 |
| 云南生菜 | 20.21 | 5.42 | 青线椒 | 38.68 | 2.50 |
| 黄心菜(2) | 11.38 | 9.60 | 红杭椒 | 57.16 | 2.50 |
| 净藕(1) | 23.85 | 4.50 | 西峡香菇(1) | 28.46 | 2.50 |
| 芜湖青椒(1) | 21.29 | 5.00 | 云南生菜(份) | 8.23 | 5.00 |
| 小米椒(份) | 20.08 | 5.00 | 小白菜 | 23.82 | 2.50 |
| 黄心菜(1) | 14.67 | 6.51 | 红椒(1) | 37.69 | 2.50 |
| 洪湖莲藕(粉藕) | 12.78 | 6.00 | 菠菜 | 28.73 | 2.50 |
| 螺丝椒 | 31.70 | 2.50 | 西峡花菇(1) | 25.41 | 2.50 |
| 鲜藕带(袋) | 22.16 | 3.00 | 红薯尖 | 18.92 | 2.50 |
| 金针菇(1) | 21.37 | 2.91 | 紫茄子(2) | 15.77 | 2.50 |
| 奶白菜 | 11.09 | 5.23 | 青茄子(1) | 15.39 | 2.50 |
| 平菇 | 26.98 | 2.50 | 云南油麦菜 | 18.10 | 2.50 |
| 蔡甸藜蒿 | 29.69 | 2.50 | | | |

### 5.4 问题四数据收集与模型的建立

1. 客流量数据：

数据类型：商超可以收集每小时、每日、每周或每月的客流量数据，以及与销售相关的其他时间信息。

数学建模方法：这些数据可以通过时间序列分析方法来处理，例如季节性分解和平滑技术。时间序列分析可以揭示销售数据的趋势、周期性和季节性，帮助商超预测未来的客流量变化。

意见和理由：了解客流量的变化趋势有助于商超在高峰和低谷时段做出更精确的补货决策。例如，在高客流量时，可以增加进货量以满足需求，而在低客流量时可以减少库存，降低损失。

2. 消费者反馈和评价数据：

数据类型：商超可以收集来自不同渠道的消费者反馈和评价数据，包括在线评论、社交媒体反馈和客户调查。

数学建模方法：消费者反馈数据可以通过情感分析和自然语言处理技术来分析。这些技术可以帮助商超识别客户的情感和态度，以及他们对不同产品和服务方面的看法。

意见和理由：了解消费者的满意度和不满意度有助于商超改进产品质量和服务。例如，如果消费者普遍对某一品类的蔬菜质量不满意，商超可以考虑改进供应链和产品质量控制。

3. 季节性因素和天气数据：

数据类型：季节性因素数据包括每个季节的销售趋势，天气数据包括每天的气温、

湿度和降水量等信息。

数学建模方法：商超可以使用时间序列分析方法，如季节性分解和 ARIMA 模型，来分析这些数据。同时，可以使用回归分析来研究天气因素对销售的影响。

意见和理由：季节性因素和天气数据的分析有助于商超更好地理解销售趋势。例如，在冷季节，对冷藏蔬菜的需求可能会增加，而在炎热的天气下，对生鲜蔬菜的需求可能会增加。商超可以相应地调整补货计划和促销策略。

4. 库存状况数据：

数据类型：商超需要实时监测不同蔬菜品类的库存水平，包括当前库存量、库存周转率等信息。

数学建模方法：商超可以使用库存优化模型，如基于需求预测的库存模型，来分析库存数据。这些模型可以帮助商超确定哪些蔬菜可能会因为短缺而需要提高价格。

意见和理由：了解库存状况有助于商超更好地规划补货计划。如果某个蔬菜品类的库存水平过低，商超可以及时采取行动，提高进货量以满足市场需求，从而减少损失和缺货情况。

通过数学建模方法，商超可以更精确地分析这些数据，从而制定更合理的补货和定价策略，以最大化收益并降低潜在的风险。这些数据不仅提供了决策支持，还为商超提供了深入洞察，帮助其在竞争激烈的市场中获得竞争优势。

5. 竞争对手数据：

数据类型：商超需要收集竞争对手的蔬菜商品销售数据，包括销售量、销售额、定价策略等信息。

数学建模方法：商超可以采用市场份额分析模型，比较自身和竞争对手在市场上的销售情况，确定自己的竞争地位和优劣势。此外，商超可以使用价格弹性模型和回归分析模型，研究竞争对手的价格策略和市场反应，从而制定自己的定价策略。

意见和理由：收集竞争对手数据有助于商超了解市场竞争态势，发现自己的不足之处，及时调整自己的补货和定价策略，提高市场竞争力。

6. 供应链数据：

数据类型：商超需要了解蔬菜供应链的相关数据，包括各个产地的蔬菜供应情况、运输时效、采购价格等。

数学建模方法：商超可以使用运营规划模型，设计最优的供应链方案，优化供货周期、降低成本和风险。此外，商超还可以使用预测模型，预测未来的采购价格和供应情况，及时作出相应的补货计划。

意见和理由：了解供应链数据有助于商超更好地管理供应链，提高采购效率和减少库存压力。商超可以通过优化供货周期、选择合适的供应商等方式，降低成本和风险，提高整体运营效率和盈利能力。

# 六、 模型检验与误差分析

## 6.1 数据拟合度检验

本研究采用回溯法对实际销售数据进行检验。我们将实际的销售数据和模型预测的数据进行对比，计算相关性系数，以检验模型的拟合度。如果相关性系数接近 1，说明模型的预测能力较强。

## 6.2 敏感性分析

进一步进行敏感性分析，通过改变某一或多个参数（例如，进货成本、损耗率等）的值，观察这对最终收益的影响程度。这有助于了解哪些参数对模型输出尤为关键。

### 6.3 容错率测试

对模型进行多次模拟运算，加入不同程度的随机误差，以测试模型的鲁棒性。高的容错率意味着即使在数据或参数有所偏差的情况下，模型的输出仍然是可靠的。

### 6.4 参数误差分析

由于参数（如批发价格、销售量等）可能受到多种因素的影响，并不总是准确的，因此需要分析这些误差对模型结果的影响。可以通过置信区间或蒙特卡罗模拟等方法来进行这种分析。

### 6.5 验证模型的可行性和实用性

最后，需要将模型的推荐方案在实际环境中进行验证。例如，可以在商超实施一段时间的补货和定价策略，然后对比模型预测的收益和实际收益，以验证模型的可行性和准确性。

### 6.6 结论

上述多角度的检验和分析能够对模型的可靠性、准确性和适用性展示一个全面的验证，同时也为模型的进一步优化和应用提供了有价值的观点。

这样的检验与误差分析不仅增强了模型的可信度，还能在一定程度上防范因数据不准确或外部环境变化导致的风险，使得商超在实际操作中能够更加自信地应用该模型，从而实现收益最大化。

## 七、 模型的评价改进和推广

### 7.1 Arima 模型

#### 7.1.1 模型的优点

自动化模型选择：流程中使用自动 ARIMA 模型选择方法，无需手动指定模型的超参数，这可以节省时间和减少用户的主观干扰，并且增强了代码的可读性和可维护性。

多周期训练：引入了多个训练周期（epochs），这可以有助于找到稳健的模型，减少过拟合的风险，提高模型的泛化能力。

结果存储：流程通过字典（`models_dict` 和 `rmse_dict`）存储了每个时间序列的最佳模型和 RMSE 值，使结果易于后续访问和分析预测。

捕捉趋势和季节性: ARIMA 模型能够很好地捕捉批发蔬菜价格的时间序列数据中的趋势和季节性成分。

预测准确性: 使用时间序列分析，特别是在蔬菜销售存在明显季节性变化的情况下（如 4 月至 10 月），能更准确地预测未来一周的批发价格。

噪声过滤: 模型能有效地识别并过滤数据中的噪声，提高预测的准确性。

#### 7.1.2 模型的缺点

计算复杂度高：对于每个时间序列数据，流程执行了多次训练和验证，尤其是在尝试不同的 d 值时，计算成本可能很高，尤其是对于大型数据集。本次训练大约消耗三小时，使用 CPU 型号为 Intel(R) Core(TM) i7-10870H CPU @ 2.20GHz。

超参数搜索空间：虽然尝试不同的 d 值是一种有效的方法，但超参数搜索空间可能很大，导致计算时间较长，并且需要足够的计算资源。

依赖性：流程依赖于外部库（如‘pmdarima’）来执行自动 ARIMA 模型选择，如果库的性能或算法发生变化，可能会影响模型的质量。

非线性局限性: 对于具有非线性趋势的数据，ARIMA 可能不是最佳选择。

不考虑外部因素: 该模型主要侧重于时间序列数据本身，而未考虑其他可能影响蔬菜价格的外部因素，如天气、节假日等。

总的来说，这个流程适合于分析信息维度较为复杂的情况， 例如超市多品类的时间序列分析，特别是在需要自动选择 ARIMA 模型超参数的情况下。但是，需要平衡计算成本和计算资源的代价，并监测模型性能。

### 7.1.3 模型的改进

a) 集成外部因素:

特征工程: 对天气、节假日和政策等外部因素进行特征工程，并将它们作为协变量添加到模型中。这些外部因素的加入可以使用 Exogenous ARIMA（即 ARIMAX 模型）实现。

动态特征选择: 使用递归特征消除或 LASSO 等算法进行特征选择，以确定哪些外部因素更有助于提高预测的准确性。

b) 使用深度学习:

LSTM 模型: 长短时记忆（LSTM）网络特别适用于时间序列问题，因为它能记住长时间跨度内的信息。LSTM 可以与 ARIMA 进行模型融合，以更好地捕获数据中的非线性关系。

GRU 模型: 门控循环单元（GRU）是 LSTM 的一个简化版本，通常在计算资源有限的情况下是一个很好的替代选项。

c) 参数优化:

网格搜索: 使用网格搜索法对 ARIMA 的阶数（p, d, q）进行全面搜索，以找到最优参数组合。

贝叶斯优化: 这是一种更高级的优化方法，它使用概率模型预测目标函数的值，通常能更快地找到优化参数。

d) 模型融合:

权重调整: 通过加权平均的方法，将 ARIMA 与其他模型（如 Prophet, LSTM）的预测结果结合在一起。权重可以通过交叉验证的方式得出。

模型堆叠: 使用 Stacking 的方式将多个模型的输出作为新模型的输入，进行二次预测。

### 7.1.4 模型的推广

a) 多品种蔬菜预测:

微调与迁移学习: 可以利用已有模型并进行微调，使之适用于其他种类的蔬菜。这样可以大幅度减少训练时间和数据需求。

多任务学习: 在一个模型中预测多种蔬菜的价格，共享隐藏层以捕获不同蔬菜间可能存在的相似性。

b) 区域扩展:

数据标准化: 考虑到不同地区的价格水平和货币单位可能不同，数据标准化成为推广模型的重要步骤。

地域特定因素: 针对不同地区的特殊需求和市场规模，添加地域特定的特征和参数。

c) 实时更新:

在线学习: 通过在线学习的方式，模型可以不断地用新数据进行更新，以适应价格变化的动态性。

流处理: 结合流数据处理平台（如 Apache Kafka），实现模型的实时更新和预测。

d) 供应链整合:

API 化服务: 将模型封装为 API，以便与现有的供应链管理系统进行集成。

实时仪表板: 开发一个实时仪表板，该仪表板可以展示预测结果、信心区间以及模型性能指标，以供供应链决策者参考。

## 7.2 分析各蔬菜品类销售总量与成本加成定价关系模型

### 7.2.1 模型的优点

考虑成本因素：采用成本加成定价法是一种考虑商品成本的经济学方法，能够帮助商超确定售价，以确保销售收入覆盖成本，有助于维护盈利能力。

基于数据：该解法基于实际销售数据和成本数据进行分析，因此具有数据支持，可以提供客观的结果，有一定的可解释性。同时针对销售量加权的品类定价法也融合了多层次数据的特性。

关系建模：采用傅里叶变换、二次方程回归等数学建模方法，可以尝试捕捉销售总量与成本加成定价之间的复杂关系，有助于预测和优化定价策略。同时拟合得到的二次模型结果与实际情况较为吻合，符合市场规律。

平滑处理：通过滑动窗口平滑处理数据，可以减少高频噪点的影响，使关系模型更稳定，更符合销售趋势。

### 7.2.2 模型的缺点

模型选择和拟合：傅里叶变换和二次方程回归虽然是有用的建模方法，但并不一定适用于所有数据集。选择合适的模型以及合适的参数需要经验和试验，可能会导致模型选择不当或信息损失过大的问题。

数据质量要求高：该解法对成本数据和销售数据的质量要求较高。本次使用数据的规律性不高，可能会影响模型的泛化能力。

不考虑市场因素：成本加成定价法虽然考虑了成本，但忽略了市场需求和竞争因素。在实际市场中，定价决策还需要综合考虑市场定位、竞争对手的价格策略以及消费者的购买力等因素。

窗口参数的不确定性：不同窗口大小的平滑效果不一致，同时损失的细粒度信息特征也不一致。决策者需要平衡平滑效果以及细粒度特征的权重，保证信息维度的同时，降低噪点的影响。

需要大量历史数据：建立关系模型需要足够长时间的历史数据，以捕捉趋势和季节性变化。如果历史数据不足，模型的预测能力可能受到限制。

### 7.2.3 模型的改进

a) 模型选择与参数调整:

可以使用贝叶斯信息准则（BIC）或交叉验证等方法自动选择更适合数据的模型。

考虑使用多模型融合来弥补单一模型的不足，例如，将傅里叶变换和二次方程回归的结果进行加权平均。

b) 数据质量与预处理:

对数据进行离群值检测和缺失值填充，以提高模型的稳健性。

使用数据增强技术如引入滞后变量，以丰富模型能够学习到的信息。

c) 考虑市场因素:

将市场需求、竞争对手价格等外部信息作为外生变量引入模型，使用如 ARIMAX 或多元线性回归来解决。

d) 窗口参数的优化:

可以使用指数加权平均作为一种更灵活的平滑方法，该方法能够自动调整对新数据和旧数据的重视程度。

e) 历史数据不足的解决方案:

如果数据不足，可以考虑使用迁移学习，利用与目标任务相似的其他任务的数据来预训练模型。

### 7.2.4 模型的推广

a) 跨品类应用:

目前模型专注于蔬菜销售，但其核心算法和数据处理手段可应用于其他食品或消费品，如水果、日用品等。

b) 地域扩展:

模型目前可能针对某一特定地域进行优化，但通过引入地域性参数或者适应性算法，它可以适用于不同地域的价格预测和定价策略。

c) 与其他数据源集成:

该模型可以与其他数据源（如气象数据、节假日数据等）结合，以捕获更多影响价格和销量的外部因素。

d) 多渠道销售:

模型可扩展到不仅仅是传统的零售渠道，也可以应用于在线商店、社交媒体平台等新兴销售渠道。

e) 实时与批量处理:

通过改善算法的计算效率，模型可用于实时价格调整以及大规模数据的批量处理。

f) 用户行为分析:

通过将用户购买数据或者行为数据整合到模型中，可以进一步个性化定价策略，提供更精细化的服务。

## 八、 参考文献


[1] 伯恩斯坦.统计学原理. 上, 描述性统计学与概率[M].科学出版社,2002.

[2] Myers, L., and Sirois, M. J. (2004). Spearman correlation coefficients, differences between[J].

[3] 孙吉贵 [1, 刘杰 [1, and 赵连宇 [1. "聚类算法研究." 软件学报 19.1 (2008): 48-61.

[4] Shumway, Robert H., et al. "ARIMA models." Time series analysis and its applications: with R examples (2017): 75-163.

[5] 常友渠, 肖贵元, and 曾敏. "贪心算法的探讨与研究." *重庆电力高等专科学校学报* 13.3 (2008): 40-42.

# 附录

| 附录 1 |
| --- |
| 介绍：支撑材料的文件列表 |
| data/： 处理数据路径<br>result/：结果文件 + 可视化路径<br>workplace/: 代码路径<br><br>data\批发销售数据_品类： 以品类为单位 统计的批发销售结果<br>data\销售单价： 以品类，单品为单位 统计的销售单价结果<br>data\销售量数据： 以品类，单品为单位 统计的销售量数据结果<br>data\销售折线图： 以品类，单品为单位 统计的销售折线图<br><br>result\Q1\interpolated： 未归一化可视化以及结果，包含相关性热力图，相关性层次聚类分组映射，以及层次树状图<br>result\Q1\standardized： 归一化可视化以及结果，包含相关性热力图，相关性层次聚类分组映射，以及层次树状图<br>result\Q2\ARIMIA 参数+检验：ARIMA 平稳性检验，ARIMA 最优模型参数选择<br>result\Q2\低频傅里叶变换： 低频傅里叶级数拟合结果<br>result\Q2\价格-销售重量： 价格-销售重量关系<br>result\Q2\批发价预测：2023 年 5-6 月的缺失值批发价填充，2023 年 7 月的批发价预测，以及可视化<br>result\Q2\品类定价策略+进货策略：2023 年 7 月 1 日至 7 日预计的品类定价策略+进货策略<br>result\Q2\散点图.jpg：价格-销售重量散点图<br><br>workplace\Q1.ipynb：Q1 的相关性分析，热力图绘制，层次聚类分组，等次树状图<br>workplace\Q1_数据处理与可视化.ipynb：Q1 的数据预处理以及销售量可视化<br>workplace\Q2.ipynb：Q2 价格-销售重量关系的回归分析<br>workplace\Q2_ARIMA.ipynb：Q2 价格-批发价的补全和预测<br>workplace\Q2&3_模型求解.ipynb：Q2 与 Q3 问题的多目标优化问题 |