

基于价格弹性的蔬菜类商品自动定价与补货决策

摘要

首先，基于数据高维、庞杂的特征对其进行整合、分析与清洗。对多个附件自然连接 (NJ)、对不同变量进行 One-hot 编码与时间序列处理、将进货销货信息按日/周/月/年聚合；对整合完成的数据挖掘更多信息，实现供应时间、同货异源、折扣退货的分析，并完成蔬菜单品的常年性、季节性、时令性分类。为保证后续求解的准确性，我们设定规则对利润异常值与滞销类无关单品进行清洗。

对于问题一，为更好地分析，本文将问题聚焦为探究销售量的时间分布规律、销售分布规律以及相关关系，并从多个时间颗粒维度分析总体、类别、个体的综合信息。其一，我们使用 ACF 自相关函数与时间序列分解，发现了各销量在各时间颗粒维度上的波动规律和趋势，并结合生活与经济学理论完成了解释和检验。其二，分析销售规律时，发现单品销量呈现长尾分布、各品类销量比以及各单品在品类中的占比差异大、少部分单品有多量少笔或少量多笔的销售特征。其三，采用偏相关性分析发现类别间的替代与互补关系，采用 FP-Growth 关联度分析发现部分单品的强相关性。

对于问题二，为探究销量与成本加成定价的关系，本文选取经济学指标构建双对数需求模型对价格弹性进行估计，结果发现六种品类的销量与定价均呈现出或大或小的负相关性。为提供定价与补货的方案，我们先采用 LSTM 模型对未来七天各品类的成本与销量进行预测，平均 R 方均达到 0.8 以上，并设定价格弹性修正函数融入定价的影响。为求解连续型优化问题，本文使用 GBest-PSO 算法并对其改进引入先验知识，得到决策方案 (具体见附录 H) 后与平均情况对比，发现收益有明显提升。

对于问题三，本文对确定的可售单品进行滞销过滤、并以单品销量与销售次数为依据采用 VIKOR 评价方法确定其需求值，初步确定了 40 个预选目标。为解决混合多目标非线性规划问题，本文选用 NSGA-II 非支配排序遗传算法，对其改进引入先验知识、采用 BG-RI 混合编码，建立多重 CV 约束，确定双优化目标为：最大化收益与最大化市场需求。最终，在 Pareto 前沿解中确定了最优值，决策方案见附录 D。同样与一般收益情况进行对比，发现双效益提升。

对于问题四，本文参照在预测与分析中出现的问题、实际经验认知、经济学领域相关文献，从定价方法的角度提出三点兼顾成本、竞争与需求的数据采集建议。其一，收集市场平均定价及主要竞争对手的定价数据以平衡经济效益与竞争优势；其二，收集消费者的预期价格及价格敏感程度以直观了解目标消费群体的满意度；其三，了解当地居民收入及人均消费水平数据，了解居民的整体消费能力与消费意愿。

关键字：ACF 自相关函数 FP-Growth 关联度分析 LSTM 模型 VIKOR 评价 NSGA-II 算法

目录

一、问题重述	3
1.1 背景资料	3
1.2 题目信息	3
1.3 待求解问题	3
二、问题分析	3
2.1 问题一分析	4
2.2 问题二分析	4
2.3 问题三分析	4
2.4 问题四分析	4
三、模型的假设	5
四、符号说明	5
五、数据预处理	5
5.1 数据整合	6
5.2 数据分析	6
5.2.1 单品按时间供应分类	6
5.2.2 同货异源类蔬菜的分析	8
5.2.3 打折单品与退货单品分析	8
5.3 异常数据处理	9
5.4 无关数据处理	9
六、问题一的模型建立与求解	9
6.1 问题聚焦	9
6.1.1 分布规律之时间:	10
6.1.2 分布规律之销售:	10
6.1.3 相互关系:	10
6.2 时间分布规律的探究	10
6.2.1 基于自相关函数 ACF 的时间序列分解的模型建	10
6.2.2 时间分布规律的求解结果	11

一、问题重述

1.1 背景资料

在生鲜商超中，蔬菜类商品保鲜要求严格，隔日产品无法继续销售，因此商家需要每日在凌晨 3:00 至 4:00 之间进行补货。在单品种类与进价未知的情况下，如何根据历史销售需求和供应情况作出补货和定价决策使利润达到最大化尤为重要。

1.2 题目信息

- 一般而言，商家对蔬菜定价采用“成本加成定价”方法；
- 蔬菜商品运送过程中会产生损耗，商家会对运损严重的品质较差的商品进行打折；
- 蔬菜类商品的需求量、供给量与时间存在关联关系，在 4 月至 10 月供给量较为丰富；
- 商超销售空间有限制，只能进货一定数量的蔬菜种类；
- 附件中给出某商超经销的蔬菜品类与单品信息、2020 年 7 月 1 日至 2023 年 6 月 30 日各商品的销售明细与批发价、商品近期损耗率。

1.3 待求解问题

问题一 分析数据，找出各蔬菜品类及单品销售量的分布规律与相互关系；

问题二 分析各蔬菜品类的销售总量与成本加成定价的关系，并以商超利润最大为目标。

给出各蔬菜品类未来一周的日补货总量与定价策略；

问题三 根据过去一周的可售品种，在满足单品数在 27 到 33 的范围、且各单品订购量大于 2.5 千克的条件下，给出 7 月 1 日单品补货量和定价策略；

问题四 分析可能影响蔬菜补货与定价决策效果的其他因素，对上述问题进行改进说明。

二、问题分析

本题主要分为基于数据分析的生鲜商超销售量影响因素及呈现规律探究问题、生鲜商超销售量的预测问题、以及基于对销售量的认知为生鲜商超的定价与补货提供决策方案的最优化问题。

分析题目中所给的信息以及附件数据，发现附件数据较为零散，因此首先需要对数据进行整合；其次对信息进行初步分析，对变量进行分类，便于更直观的把握数据的特征；最后大量的数据意味着数据质量可能难以保证，需要对异常值与无关数据进行处理保证后文模型的效果。

2.1 问题一分析

问题一主要要求我们分析数据，探究各个品类及单品销售量存在的规律与关系。但是由于题目未明确给出具体的探究方向，因此需要先考虑从哪几个角度入手。考虑到销售量的分布规律可分为随时间的分布规律、随每个订单的分布规律、随每个品类的分布规律三个角度；相互关系也可分为品类间的相互关系以及单品间的相互关系，从这几个角度出发即可以较为完整并且清晰地刻画销售量地分布规律与相互关系。同时为了了解分析结果的准确性与其背后的影响因素，可以结合生活实际经验以及经济学理论对其进行解释验证。

2.2 问题二分析

问题二可以分为三步进行逐一求解。首先分析销售总量与成本加成定价的关系；其次对蔬菜各品类未来一周的销售水平、批发价格进行预测；最后综合销售总量与定价的关系分析结果以及未来一周的预测结果，构建收益最大化决策模型进行求解。其中需要注意的是，销量总量与成本加成定价可能受多种因素的影响，需要建立合适的模型刻画定价与销量的关系。

2.3 问题三分析

问题三是混合多目标规划问题。尽管题中要求在满足需求量的前提下进行最大收益优化，但是由于单品数量的限制，无法完全满足顾客的购买需求，因此本题应为一个多目标优化类问题，应对顾客需求满足程度与收益情况进行综合优化。但其中需要注意的是顾客的需求度并不能完全简单地与销售量对等，需寻找评价指标综合销量与销售次数等其他因素，得到需求度。此外需要选择合适的优化算法方能求解。

2.4 问题四分析

问题四是一个较为开放的问题，可以根据前三问数据分析结果、处理信息时遇到的难点、实际生活经验并结合相关专业领域文献寻找对商超决策有利的其他数据。

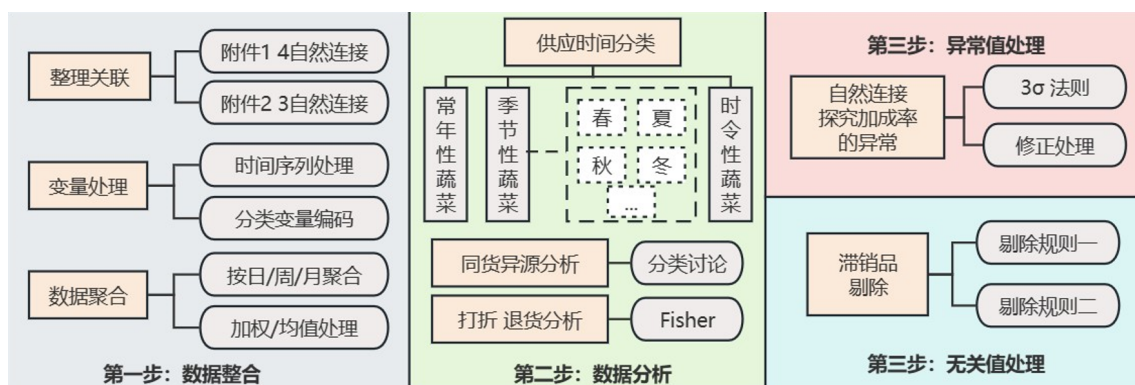
三、模型的假设

- 假设短期内商超的进货与售价不会对顾客忠诚度造成影响；
- 假设短期内，历史售价等客观因素不会对顾客当日需求量造成影响。
- 假设商超在这三年中的补货与定价策略已经一定程度上考虑了顾客消费需求；
- 该商超附近无其他竞争对手，不存在由恶意竞争引起的销量波动

四、符号说明

符号	意义	符号	意义
Q_i	品类 i 的总销量	Q_{ij}	品类 i 中单品 j 的销量
P_i	品类 i 的平均定价	P_{ij}	品类 i 中单品 j 的定价
L_i	品类 i 的平均损耗	L_{ij}	品类 i 中单品 j 的损耗
B_i	品类 i 的总补货量	B_{ij}	品类 i 中单品 j 的补货量
C_i	品类 i 的补货成本	C_{ij}	品类 i 中单品 j 的成本
w_i	品类 i 的平均加成率	w_{ij}	品类 i 中单品 j 的加成率
M	总销售额	T	总成本
w_i	品类 i 定价的加成率	w_{ij}	品类 i 中单品 j 的加成率
k	损耗折扣率	D_j	需求度
E_p	需求价格弹性	E_c	交叉弹性弹性
E_v	价格预期弹性	EV	交叉弹性弹性

五、数据预处理



5.1 数据整合

概览题设所给出的附件信息，可以发现其数据量庞大、信息冗杂琐碎，是典型的实际生活数据。对其进行整理关联、变量处理、数据聚合等初步工作能有效帮助后续数据的处理、分析与预测。

- **整理关联：**附件一、四共同描述了单品的编号、分类、近期损耗率信息，而附件二、三共同描述了单品的销售、进货信息，通过 Python 编程以实现表格间自然连接（NJ）。

- **变量处理**：将附件中的列“销售日期”与“销售时间”处理为时间序列、“销售类型”与“是否打折销售” one-hot 编码处理为分类变量。
- **数据聚合**：由于蔬菜单品的销售信息琐碎，故将其以单品和品类两种形式按日、周、月、年进行数据聚合，以品类的日聚合为例介绍聚合规则如下：

(a) 对于日销售量：

$$Q_i = \sum_j Q_{ij} \quad (1)$$

其中 O_k 表示每个订单的单品编号, A_i 表示品类 i 中的单品集合, Q_{ij} 表示品类 i 中的单品 j 的销量, Q_i 表示品类 i 的总销量

(b) 对于日销售单价、进价、加成率：

$$X_i = \frac{\sum_j Q_{ij} \times x_i}{Q_i} \quad (2)$$

其含义可以理解为以销量为权重进行加权求均值，其中加成率的计算公式如下：

$$\text{加成率} = \frac{\text{定价} - \text{成本}}{\text{成本}} \quad (3)$$

5.2 数据分析

5.2.1 单品按时间供应分类

对蔬菜的进货信息分析，发现蔬菜的供应季节、周期、时间存在明显差异，查阅相关资料后，本文将从供应时间的角度将蔬菜单品分为以下三类：

- 常年性蔬菜：这些产品可以在不同的地区或国家生产, 几乎整年都可以获得。
- 季节性蔬菜：受气候和地理条件影响，该类产品只在特定的季节内生长与销售。
- 时令性蔬菜：该类产品仅在某些特定的节日、假期、节气等短期时间销售。

结合题设信息，本文设定如下分类规则：

规则 1 三年中某一年可供天数大于 300 天，划分为常年性蔬菜；

规则 2 每年最大可供天数小于 15 天，划分为时令性蔬菜；

经过规则一处理了 30 个单品、经过规则二处理了 70 个单品，剩余单品划分为季节性蔬菜，对季节性蔬菜进行季节分析进一步统计其所出现的季节，发现如下信息：

- 若某单品每年可供天数虽不超过 300 天，但是在各季节均可供，应属于常年性蔬菜；
- 若某单品在某些季节可供，但可供天数极少，为异常数据，不能视为该季节蔬菜；
- 若某单品在某季节集中供应且其余季节无供应，则属于典型该季节蔬菜。

基于以上信息，增加如下规则：

规则 3 四季都可供，且各季节三年可供天数均大于 20 天，划分为常年性蔬菜；

规则 4 各个季节三年共计可供天数均小于 20 天，划分为时令性蔬菜；

经过规则 3、规则 4 修正，分别处理了 46 个单品、12 个单品，余下的 90 种蔬菜均划为季节性蔬菜，具体分类结果见支持材料，占比情况和部分结果如下：



图 1 各类型蔬菜占比

表 1 各类型蔬菜举例

蔬菜类型	举例
常年性 30%	大白菜、胡萝卜、土豆
季节性 37%	槐花、野藕、薄荷叶
时令性 33%	莲藕、蒲公英、荸荠

进一步，我们根据各季节性蔬菜在进货信息中常出现的月份，推测各个季节性蔬菜其所属的季节，具体结果见支撑材料。查阅相关资料，我们发现分类的结果十分贴合生活常识，验证了分类的精确性。这为之后蔬菜的分析提供可靠的参考。如“四川红香椿”，统计分析其为春季蔬菜，查阅资料后发现其 3 月末到 4 月初为采摘和食用的黄金时间。

5.2.2 同货异源类蔬菜的分析

题目给出的数据中包含不同供应地的同种蔬菜，为避免其对后续建模造成影响，进行其统计分析，发现主要可以分为两种类型：

- 销量均衡类：如黄心菜 (1) 与黄心菜 (2)，销量分别为 2911kg 与 1882kg, 共计 6 种
 - 销量失衡类：如紫茄子 (1) 与紫茄子 (2)，销量分别为 279kg 与 13602kg, 共计 13 种
- 但是由于同货异源类蔬菜的进价和售价并不相同，故不能将其合并，应当视作两种

不同的蔬菜，故本文不对其做过多处理。

5.2.3 打折单品与退货单品分析

应当注意到，少量（5.39%）蔬菜由于品相等原因打折销售，极少量（5‰）订单由于未知因素被退货。前者因为打折而导致销售单价及销量发生改变、后者因为销售单价为负与正常项不同，但经过数据按时间聚合后，两类数据均不会对结果造成影响。

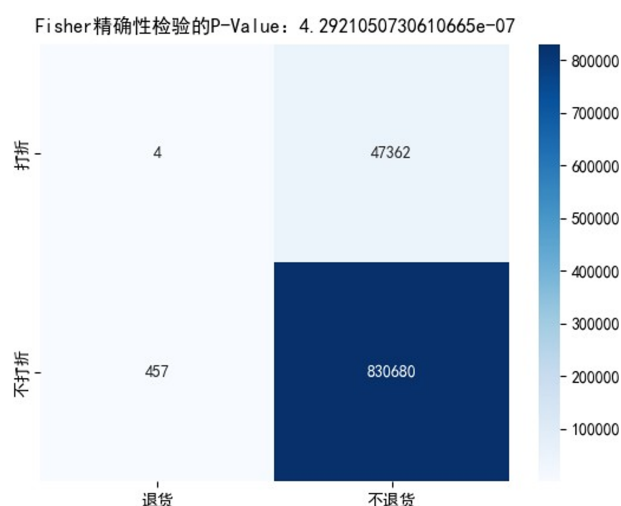


图2 打折与退货的关系

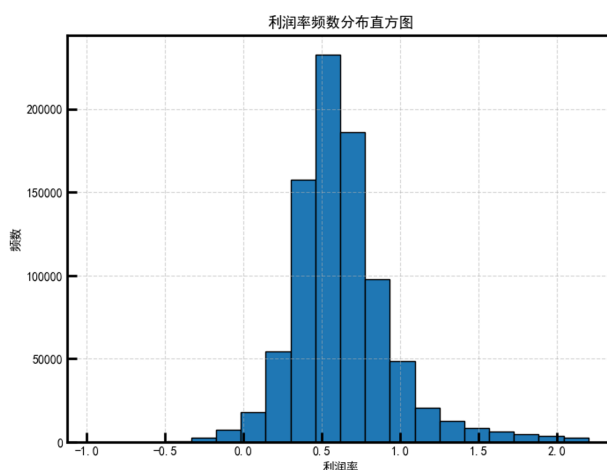


图3 商品加成率分布图

如图2, 对两者进行 Fisher 精确性检验可以发现 p 值小于 0.05, 说明两者存在相关关系, 具体来说即, 当打折销售时退货的概率会降低。

5.3 异常数据处理

将附件自然连接之后能发现不易被发现的异常数据, 如加成率 (利润率)。大部分单品的加成率为 0.5 左右, 然而存在部分单品加成率高达 200, 显然不符合蔬菜一般定

价规律。本文对附件二中的 87 余条销售记录进行如下处理：

- 3 σ 法则：剔除利润率偏差极大的数值
- 修正处理：将利润率大于 2 的销售记录剔除

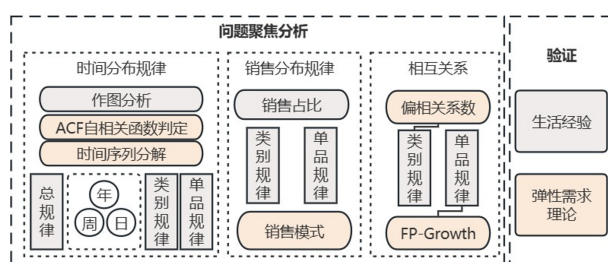
经过处理之后，删去 16430 条销售记录，对商品加成率画出分布图如图 3 所示。

5.4 无关数据处理

观察蔬菜销售数据发现，部分蔬菜在供应上未呈现周期规律性，并且供应天数极少销售量同样极少。本文认为该类蔬菜可能为某段时间突然出现的特殊品种蔬菜，存在供应链不稳定，或者市场需求量极少的情况，并不存在稳定销售以及供应规律，在数据分析中属于无关数据，对蔬菜总销量以及未来预测无有利贡献，因此本文不予进行考虑。本文对无关数据进行如下处理：

- 剔除完全没有销量数据的蔬菜单品，共计 5 个；
- 剔除销售天数小于 10，销售量小于 3‰，且无年周期规律的蔬菜单品，共计 48 个。

六、问题一的模型建立与求解



6.1 问题聚焦

问题一要求探究各个品类及单品的蔬菜销售量的分布规律以及相互关系。由于能够考虑的方面很多，我们将问题聚焦为探究以下三个部分：

6.1.1 分布规律之时间：

题中所给的蔬菜销售数据，存在长期变化趋势和周期变化规律，并且受到购买蔬菜的顾客生活习惯、工作时间以及气候等因素的影响。这些因素对于大部分单品与品类产生的影响是同向性的，能体现外部经济、环境因素的作用。

- 首先，对销售总量进行时间上的分布规律分析，观察外部因素对总体销售量的影响；
- 其次，对各品类不同时间维度上的销量分布分析，发掘人们对各个品类的不同需求；
- 最后，基于单品的时间分类对各单品探究其时间分布规律。

6.1.2 分布规律之销售：

商品销售模式根据其单笔平均销量以及出售次数的情况可以分为三种：多量少笔、少量多笔、以及零散销售。对每种单品以及品类的总销售量在每笔销售量上的分布情况分析，可以得到蔬菜销售分布规律。

6.1.3 相互关系：

探究品类、单品销售量的相互关系时需要注意去除外界因素的干扰，剔除经济、社会、生活习惯对每个品类、单品的影响。本文认为经济生活状况对各种品类销量的影响主要体现在总销量的变化，因此本文采用偏相关性分析剔除其他因素的影响。

6.2 时间分布规律的探究

为确定蔬菜总销售量在三年中的变化是否含有周期性的规律变化，本文通过以下三个步骤进行分析：

步骤一 作图分析：绘制时间序列图，观察曲线趋势，初步分析可能存在的趋势、周期。

步骤二 周期性判定：本文采用自相关函数（ACF）图进行时间序列的周期判定。

步骤三 时间序列分解：判定周期性后，仍不能确定一个周期内的变化规律，因此本文采用时间序列分解，分离时间序列的总体趋势，周期规律。

6.2.1 基于自相关函数 ACF 的时间序列分解的模型建

立自相关函数 ACF：

一个时间序列在不同时间点的值之间可能存在相关性。自相关性通过计算时间序列在不同滞后阶段之间的相关系数来衡量这种相关性。如果在某个滞后阶段存在显著的相关性，那么可以认为时间序列在该阶段具有自相关性。可以按照如下步骤操作：

1. ACF 的计算：自相关系数计算公式为：

$$\rho_k = \frac{Cov(X_t, X_{t-k})}{Var(X_t)}$$

其中， X_t 是时间序列在时间点 t 的值， X_{t-k} 是在 t 时刻滞后 k 期的值， Cov 表示协方差， Var 表示方差。

2. ACF 图的绘制：将计算得到的自相关系数绘制成 ACF 图。横轴表示滞后阶段（lag），纵轴表示自相关系数的值。

3. ACF 图的解释：相关系数在滞后阶段 0 处为 1，因为时间序列与自身在同一时间点的相关性始终是 1。其他滞后阶段的相关系数将在 ACF 图中显示。

- 若 ACF 图在之后的滞后阶段上有正相关性，可能表示存在季节性模式。

- 若 ACF 图在之后的滞后阶段上有负相关性，可能表示存在反向的季节性模式。
- 若 ACF 图在之后的滞后阶段上都接近零，可能表示时间序列随机，无自相关性。
- 若 ACF 图在某个滞后阶段上出现显著的峰值，相关性系数高于显著性阈值，则表示存在周期性或趋势。

乘法时间序列分解:

乘法分解假定时间序列是由趋势、季节性、周期性和噪声等组成部分相乘而成的，通常表示为：

$$Y_t = T_t \cdot S_t \cdot C_t \cdot E_t$$

其中 Y_t 为观测值， T_t 表示趋势成分， S_t 表示季节性成分， C_t 表示周期性成分， E_t 表示随机噪音。

6.2.2 时间分布规律的求解结果

(a) 总销量的分布规律

总销量在三年间的变化曲线如图 4 所示，总体销量在 2021.2-2022.5 年间处于低迷状态，并于 2023 年明显提高。