蔬菜类商品的自动定价与补货决策

摘要

对于问题一，题目要求分析蔬菜各品类及单品销售量的**分布规律**及**相互关系**，本文采用**统计学分析法**，对分布规律进行描述性表述，结论为蔬菜各品类中花叶类对于商超中的需求量比较大，辣椒类、食菌类需求量一般，水生根茎类、花菜类以及茄类需求量较小。单品中销量差距大，其中芜湖青椒（1）、西兰花、净藕（1）大白菜、云南生菜位于销量前五。蔬菜销量在全年的销售上，存在**季节性规律**。利用**时间序列分析法**，对蔬菜销售在时间序列上的规律进行分析，同样得出存在季节性规律，和销售时间趋势具有一定关系。并利用 **Tuday HSD 方法**和 **Pearson 相关性分析**，对蔬菜各品类及单品销量之间的差异性和相互关系进行分析，得出各品类和单品之间存在数据**差异性**和**相关性**，各品类和单品两两之间存在差异性，以及两两之间的**相关性程度**。

对于问题二，题目要求分析蔬菜各品类的销量与**成本加成定价**的关系，给出蔬菜品类未来一周的**日补货量和定价策略**，且使得商超收益最大。该题为**优化类问题**，通过所给的数据，本文建立了利润相关的方程，并通过**多元线性回归**建立销售价格、销量及进价之间的线性方程，联立两个方程建立利润最大化模型，再结合**报童模型**，求解出各品类未来一周的日补货量和定价作为本题的补货与定价决策。

对于问题三，题目要求分析以**利润最大化为目标**，在蔬菜商品的销售空间有限、单日可售单品总数范围在 27-33 个、单品订购量大于等于 2.5kg 且所选的单品仅为 2023 年 6 月 24-30 日的可售单品为**约束条件**的前提下，制定出 7 月 1 日的**单品补货量和定价策略**。本文采用**基于熵值法的秩和比综合分析**，得出 **RSR** 排名前 33 的单品，与问题二类似，建立相似的利润方程和变量相关性数方程，联合求解决策。围绕目标和约束条件，得到最终选购 30 个单品，以及相应的 7 月 1 日单品补货量和定价**策略**。

对于问题四，题目要求提供一定的数据采集上的意见，帮助商超更好的指定补货和定价决策，根据处理过程所产生的问题以及对于模型的假设，**收集四大类数据：商超本身的数据、需求端的数据、供货端的数据以及地域性的数据**，其能直接或间接影响分析补货和定价决策模型里的某值的大小，从而提升**利润**。

关键词：时间序列分析；报童模型；基于熵权法的秩和比综合分析法；

一、问题重述

1.1 问题背景

生鲜商超经营蔬菜类商品时面临着保鲜期短、品相变差等挑战。为了满足消费者需求，商超需要每天进行补货和定价决策。然而，在凌晨进货时，商家通常无法确切知道单品和进货价格，这就要求他们依赖可靠的市场需求分析来做出决策。

在需求侧，蔬菜类商品的销售量与时间存在一定的关联关系。这意味着商家可以通过历史销售数据和趋势分析来预测当天的需求量。了解消费者的购买习惯、季节性变化以及特殊节假日等因素都有助于更准确地预测需求，并做出相应的补货决策。

在供给侧，商超面临着蔬菜供应品种的多样性。尤其是在 4 月至 10 月期间，供应更为丰富。然而，商超的销售空间有限，因此需要进行合理的销售组合安排。这可能涉及到根据销售情况和消费者偏好来确定进货量和组合，以确保供应充足并最大程度地提高销售效益。

由于蔬菜的保鲜期限制和品相变差,商超通常会对运损和品质较差的商品进行打折销售。因此，商超对定价决策也变得重要。一般情况下，商超采用"成本加成定价"的方法，即根据商品的进货成本以及运营成本来确定售价，并在此基础上进行适当的调整。

1.2 问题提出

问题 1 蔬菜类商品不同品类或不同单品之间可能存在一定的关联关系，请分析蔬菜各品类及单品销售量的分布规律及相互关系。

问题 2 考虑商超以品类为单位做补货计划，请分析各蔬菜品类的销售总量与成本加成定价的关系,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。

问题 3 因蔬菜类商品的销售空间有限，商超希望进一步制定单品的补货计划,要求可售单品总数控制在 27-33 个,且各单品订购量满足最小陈列量 2.5 千克的要求。根据 2023 年 6 月 24-30 日的可售品种，给出 7 月 1 日的单品补货量和定价策略，在尽量满足市场对各品类蔬菜商品需求的前提下，使得商超收益最大。

问题 4 为了更好地制定蔬菜商品的补货和定价决策，商超还需要采集哪些相关数据，这些数据对解决上述问题有何帮助，请给出你们的意见和理由。

二、问题分析

2.1 问题一的分析

本问题要求分析蔬菜各品类及单品销售量的分布规律及相互关系。即分别对两个参数在销售量上的分布规律和相互关系进行分析，对于分布规律的分析，预采用统计学分析，将蔬菜各品类和单品中三年的所有销售量，做描述性统计，分别求出总销量，月度销量，以及销量占比等数据，再对数据进行直观性分析，进行描述得出规律。基于背景中说明，蔬菜类商品的销售量与时间往往存在一定的关联关系，因此预采用时间序列分析法，从时间关系上，对蔬菜各品类及单品销量进行分析，求解其在时间上所存在的规律。对于相互关系的分析，本题预在数据及模型可靠性良好，方法、模型可行性高的基础上进行分析，先后采用 ANOVA 检验和 Kendall 检验，检验销量之间是否差异性和相关性，再采用 Tukey HSD 方法和皮尔逊相关分析，对于销量之间的差异性和相关性进行近一步的分析，求知各品类及单品销量之间哪些含有显著性差异以及相关系数，即相关程度。

2.2 问题二的分析

本问题要求分析蔬菜品类的销售总量与成本加成定价之间的关系，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。本问题属于单目标优化类问题，通过进行多元线性回归方程以及联立方程，建立利润最大化模型，即对各品类的销售单价、销量、批发价格进行多元线性回归分析所得到的函数方程，以及利润方程进行联立得到最终利润最大化模型，以求得最佳利润收益下的定价。运用报童模型对日补货总量进行决策， 得到最佳补货量。

2.3 问题三的分析

本问题要求在蔬菜商品的销售空间有限、单日可售单品总数范围在 27-33 个、单品订购量大于等于 2.5kg、且所选取的单品仅为 2023 年 6 月 24-30 日的可售品种内的。根据以上要求并围绕商超收益最大化制定出 7 月 1 日的单品补货量和定价策略。该问题类型与问题二类似，都为优化类问题，且都是围绕收益最大化为目标。故建立相似模型，将六大蔬菜品类替换成 27-33 类单品，根据单品的各项数据与利润建立模型，具体操作同问题二。前提需要，利用秩和比综合评价法，在 2023 年 6 月 24-30 日的可售品种内，从销售价格、总销量、损耗率、收益、批发价格表现，分析各单品的权重，并取前 27-33 项作为分析对象，以利益最大化为目标，求解该 27-33 项的 7 月 1 日的单品补货量和定价策略，并绘制表格。

2.4 问题四的分析

本问题要求提供一定的数据采集上的意见，帮助商超更好的指定补货和定价决策，根据对前三问的分析处理上所包含的问题、对模型分析的优化、以及得出的分析结论。拟定收集商超本身的数据、需求端的数据、供货端的数据以及地域性的数据。

三、模型假设

1. 假设本题所提供的数据真实可靠，数据完整，没有主观因素进行破坏，即具有数据可靠性。
2. 假设模型使用时，所采用的样本是相互独立的，即具有相互独立性。
3. 假设数据之间存在的因果关系，关联性，可以用线性函数进行描述。
4. 假设在建立利润最大化模型时，不考虑库存成本的影响。

四、符号说明

符号	说明
y_t	时间序列模型观测值
adj_p	显著性水平
π	利润
Y	售价
x_1	销售量
x_2	进价
η	损耗率
x	日需求量
β	相关性系数
VIF	膨胀因子
Q	补货量
C_1、	批发价格
C_2	库存持有价格
q	前一天的库存剩余量
P	销售价格
$f(x)$	销量的概率密度函数
$F(x)$	销量的分布函数
w_i	指标权重
R_{ij}	行列元素的秩
RSR_i	秩和比
r	皮尔逊相关系数

五、模型的建立与求解

5.1 问题一模型的建立与求解

蔬菜产业中常对某段时间销量进行统计分析，通过数据可视化，对销售趋势和规律进行分析。对于蔬菜类产品，常常也与时间相关联。市场经济上，销售信息之间也普遍存在相关性。因此本题根据上述情况，对本题做统计学分析，进行规律分析，建立 pearson 相关性分析的。本题解题过程如下图图 1 所示。

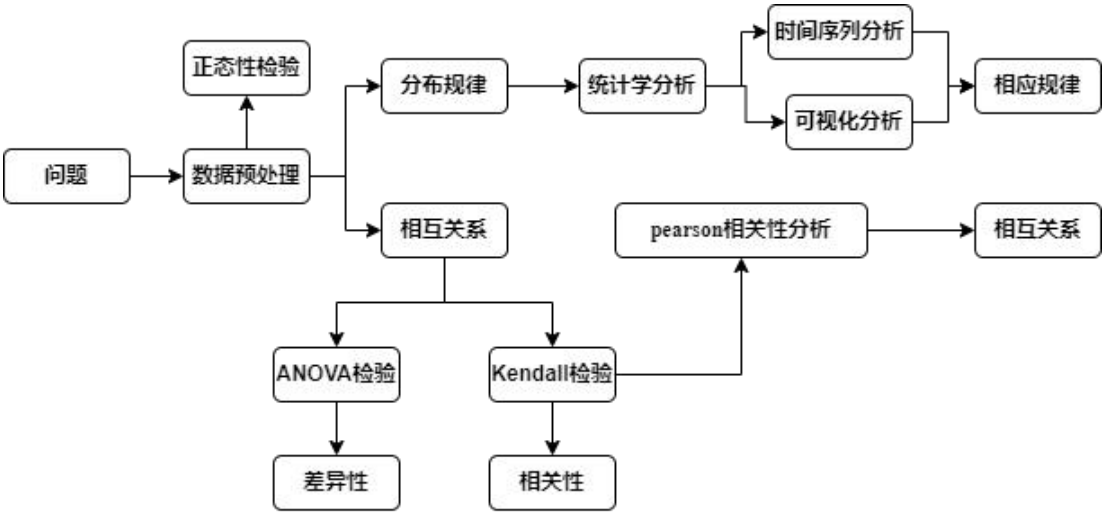

图 1 问题一解决流程

5.1.1 问题一模型的准备

本题为分析蔬菜各品类及单品销量的分布规律及相互关系，采用统计学分析，进行数据量统计，再进行可视化分析，对相应规律进行分析与解释。建立 pearson 相关性分析，对销量数据之间的关系进行线性相关性进行分析。进行分析与模型建立之前，进行数据预处理操作，将有关数据以及附件进行整合。对数据进行正态性检验，采用 SPSSPRO 软件中的 Shapiro-Wilk 检验，得出 K-S 检验结果为**0.106**，同时得出对应的 P-P 图与 Q-Q 图如下。

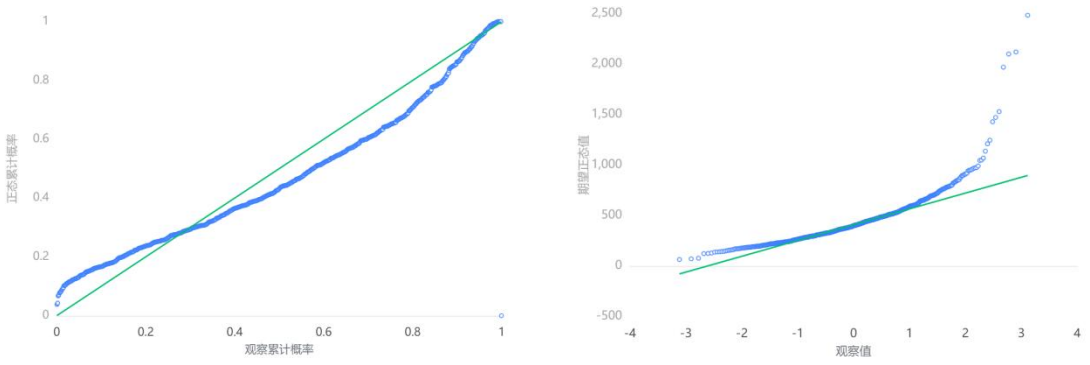

图 2 正态性分析 P-P 图及 Q-Q 图

根据 K-S 检验结果，需要该结果小于 0.05 才能证明该数据满足正态性分布，通过检验所得值为 **0.106**，略大于 **0.05**，故不能满足。但是通过观察分析 P-P 图与 Q-Q 图，该数据 P-P 图中观察的累积概率与正态累计概率的拟合度较高，Q-Q 图中观测值与预测值不同分位数的概率分布的散点重合率也较高。故可以认为该数据基本满足**正态分布**。同时通过得出的正态性检验直方图也能得出该结论。

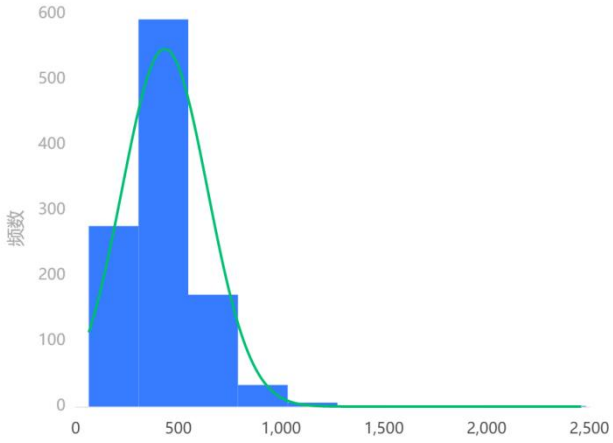

图 3 正态性检验直方图

如图所示该真太土基本上呈现中间高，两端低。说明该数据虽然不满足绝对正态分布，但基本可接受为正态分布。

5.1.2 问题一模型的建立

（1）分布规律

通过对蔬菜各品类及单品销量进行统计学分析。分别对 6 个品类在三年的总销量、6 个品类三年里 12 个月的月度销量、251 个单品在三年的总销量、251 个单品在三年的总销量占前五的五个单品销量分析。

为探究蔬菜各品类及单品销量在时间关系上的规律，采用时间序列分析模型，对该商超三年里每天销量进行时间序列分析。

ARIMA 时间序列模型可以用于预测未来的数据点，进而进行时间序列分析、需求预测，是一种常用有效的时间序列建模方法。在建立 ARIMA 时间序列模型[4]需要对数据进行 **ADF 检验**。

ARIMA 时间序列模型公式如下：

$$\text{ARIMA（p, d, q）：} y_t = c + \varnothing_p y_{t-1} + ... + \varnothing_p y_{t-p} + \theta_{1\in_{t-1}} + ... + \theta_{q\in_{t-q}} + \in_t$$

通过对最终所得到的时间序列图进行分析，直观的回答相应的时间规律。

（2）相互关系

本题为探究蔬菜各平类及单品销售量的相互关系，主要对数据之间的差异性

和相关性进行分析，先后采用 AVOVA 检验和 Kendall 检验，检验销量之间是否存在差异性和相关性，再利用 Tukey HSD 方法和 pearson 相关性分析，对销量之间的差异性和相关性进行进一步的分析。

进行后续检验和模型建立操作前，需进行正态性检验，在准备阶段，已经对数据进行了正态性检验，该数据满足正态性分布。

***AVOVA* 检验：**

ANONA 检验需要计算总平方和、组内平方和、组间平方和、均方差、F 统计量和自由度等关键指标，比较两个组之间的均值是否存在差异性显著。

***Kendall* 检验：**

Kendall 检验对统计量的显著关系进行检验，判断 P 值是否呈现出显著性(P<0.05)，若呈显著性，则说明数据呈现相关性。

***Tukey HSD* 方法：**

Tukey HSD 方法提供了一种有效的多重比较方法，它可以可靠地确定哪些组之间存在显著差异，并给出对应的显著差异区间。同时，通过调整后的显著性水平，可以控制进行多个比较时产生错误结果的概率。方法公式如下：

比较值（q 值）

$$q = q(\alpha, df\_w, k)$$

集中参数

$$CR = q * sqrt(MSW / n)$$

计算显著性差异区间

$$CI = (X\_i - X\_j) \pm CR$$

调整后的显著性水平（adj_p）的计算公式

$$adj\_p = p * k(k-1)/2$$

其中，α 代表显著性水平，df_w 代表组内自由度，k 代表组数，q 代表比较值 MSW 代表组内均方差，n 代表每组的样本量，X_i 和 X_j 分别代表第 i 组和第 j 组的均值，CR 代表参数，p 代表原始的显著性水平，k 代表组数。

通过检验数据之间的差异性，来表示进行统计量的操作具有意义。在市场经济上也有相对于的影响和区别。

***pearson* 相关性分析：**

本题所给的数据，满足正态性分布，且在时间序列上具有连续性。采用 Pearson 相关性分析是最优的。Pearson 相关分析用于衡量两个连续变量之间的线性相关关系。它衡量了这两个变量之间的线性关系的强度和方向。在本题中对蔬菜各品类及单品销售量中的相互关系，进行相关性检验，分析相关性程度。如果

两个变量完全正相关，数据之间皮尔逊相关系数为 1，如果完全负相关则为-1，如果没有线性关系则为 0。皮尔逊相关系数通常用符号 r 表示，设有 n 各样本。X 为第一个变量，Y 为第二个变量，m_X 和 m_Y 分别为 X 和 Y 的均值，则皮尔逊相关系数的公式为：

$$r=\frac{\sum_{i=1}^{n}(X_i-m_X)(Y_i-m_Y)}{\sqrt{\sum_{i=1}^{n}(X_i-m_X)^2}\sqrt{\sum_{i=1}^{n}(Y_i-m_Y)^2}}$$

最终得到相关系数，绘制相关热力图，根据热力图对各数据之间的相关性进行分析。

5.1.3 问题一模型的求解

（1）分布规律

统计学分析：

对蔬菜各品类及单品的销量，按照相应描述进行统计学分析，得到以下结果：

表一 蔬菜各品类 3 年销量表

品类	销售总量（kg）	平均单次销量
花叶类	198520.978	0.598
花菜类	41766.51	0.482
茄类	22431.782	0.500
辣椒类	91588.629	0.881
食用菌	76086.725	0.513
水生根茎类	40581.353	0.691

从表中分析可以知道，在所包含的 6 个品类的 3 年的销售总量中，花叶类的销售量远高于其余五类，其次是辣椒类和食菌类的销售量相对于剩余类也保持较高的销量，剩余三类的销售总量较低。但是含有个别品类的销售总量低，平均单次销量高的情况。

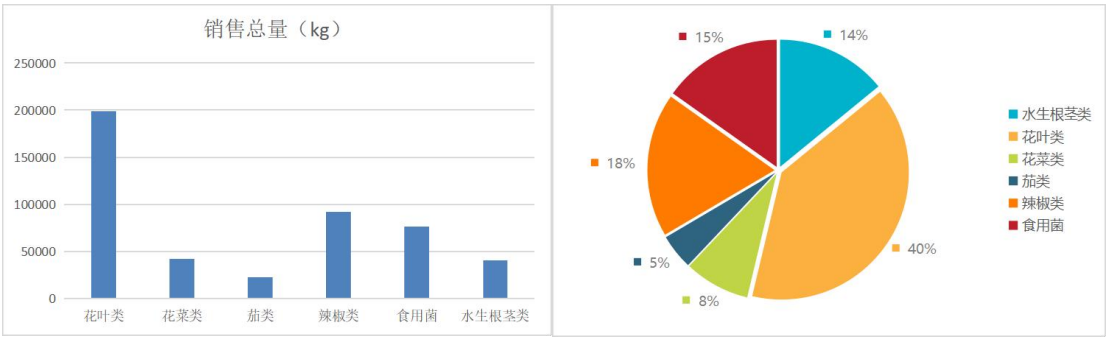

图 4 各品类销售总量直方图（左）、各品类销售总量占比图（右）

通过上图可以直观的观察出，花叶类的销售总量远大于其余品类，而茄类的销售总量远小于其余品类。

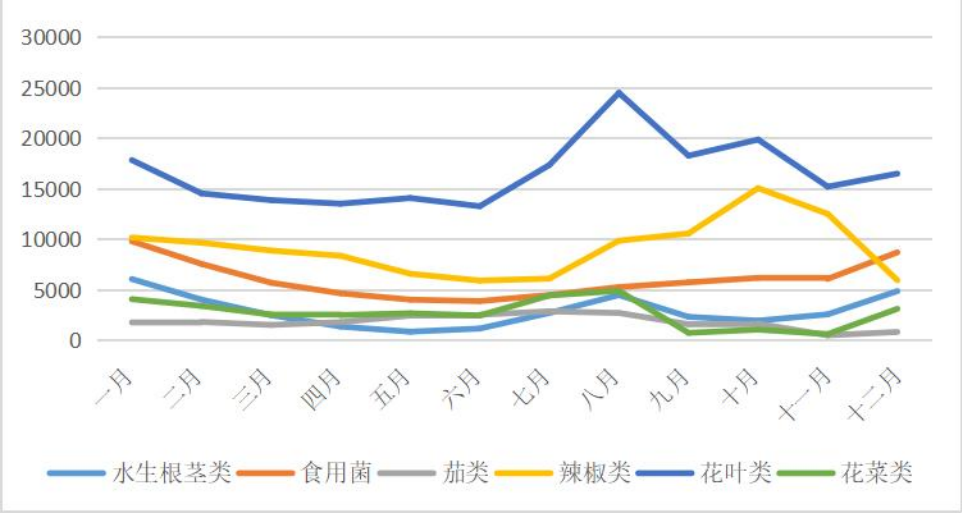

图 5 各品类各月份销量折线图

通过汇总各品类在三年内 12 个月的月度销量得到相关数据，并绘制相关趋势折线图。如上图，由图可观察出，花叶类在每个月的销量都高于其余品类，6 各品类在七月份到八月的销量都一定上升趋势，在十二月份到二月份的销量都有一定的下降趋势。

表二 各单品三年销量表

单品名称	销量（kg）
芜湖青椒（1）	28164.331
西兰花	27537.228
净藕（1）	27149.44
大白菜	19187.218
云南生菜	15910.461
…	…
红珊瑚（粗叶）	0.682
芥兰	0.671
紫白菜（2）	0.615
红橡叶	0.419
水果辣椒（橙色）	0.415

由上表可以得出，在 251 种单品中，总销量排名的首位销量差距巨大，其中排名前五位的单品分别为：芜湖青椒（1）、西兰花、净藕（1）、大白菜、云南生菜。排名后五位的单品分别为：红珊瑚（粗叶）、芥兰、紫白菜（2）、红橡叶、水果辣椒（橙色）。芜湖青椒（1）与水果辣椒（橙色）的销售差量为 **28163.916kg**，数值接近 3 万。

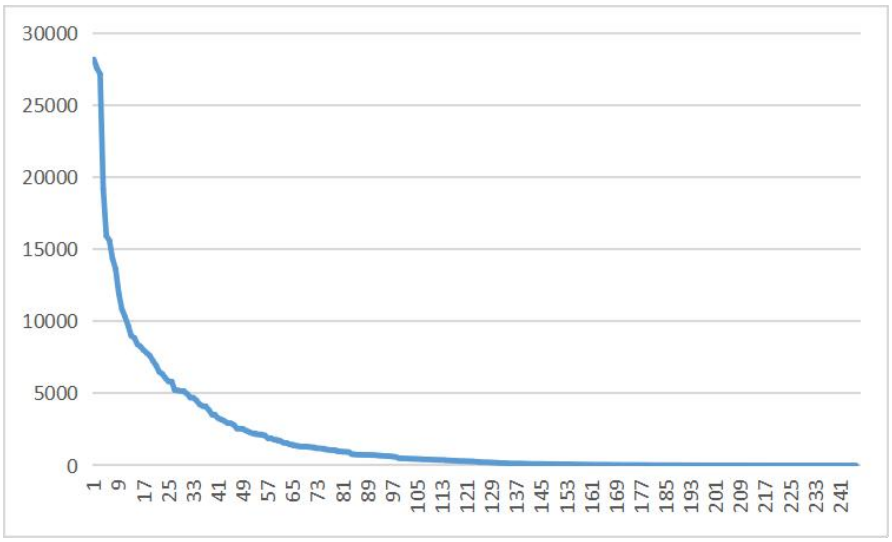

图 6 各单品销量汇总图

通过上图可以观察出，在 251 类单品中仅有少量，近 30 类单品的三年销售总量超过了 **5000kg**，且销量售量所呈现的趋势呈指数性下降。即销售量超过 5000kg 的品类中彼此差距也大，绝大多数的品类销售量极低，甚至接近于 0。

图 7 各单品词云图

上图为通过各单品的销售总量作为数据量绘制的词云图，通过上图也可以得知，各单品之间销量差距大。

时间序列分析：

通过 SPSSPRO 软件对蔬菜各品类及单品销量，对于时间关系的分析，得到以下结果。

表三 ADF 检验表

变量	差分阶数	P
销量（千克）	0	0.030
	1	0.000
	2	0.000

ARIMA 模型要求序列满足平稳性，通过 ADF 检验，检验数据是否满足时间序列数据，若呈现显著性**(P<0.05)**，则说明拒绝原假设，该序列为一个平稳的时间序列，反之则说明该序列为一个不平稳的时间序列。由上表可知，差分为 0、1、2 阶时，显著性 P 值分别为 **0.030**、**0.000**、**0.000**，这三个值都小于 0.05，故该数据满足时间序列数据，可建立 ARIMA 模型。

表四 ARIMA 模型检验表

项	符号	值
样本数量	N	1085
Q 统计量	Q6(P 值)	0.001(0.970)
	Q12(P 值)	42.736(0.000***)
	Q18(P 值)	133.06(0.000***)
	Q24(P 值)	218.49(0.000***)
	Q30(P 值)	275.828(0.000***)
拟合优度	R^2	0.444

通过 SPSSPRO 软件，软件基于 AIC 信息准则自动寻找最优参数，模型结果为 **ARIMA（1，1，1）**。基于变量：销量（千克），从 Q 统计量结果分析可以得到：Q6 在水平上不呈现显著性，不能拒绝模型的残差为白噪声序列的假设，同时模型的拟合度 R^2为 **0.444**。

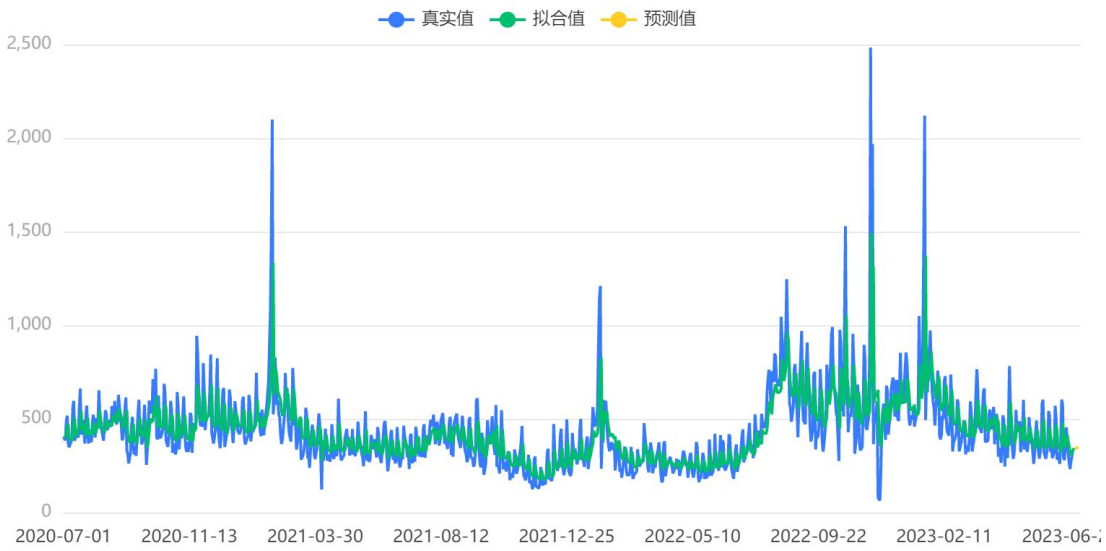

图 8 时间序列图（所有蔬菜）

通过 ARIMA 模型分析后，生成时间序列图，通过序列图可以观察出，将整条时间线分为三个区段：2020.07.01-2021.06.30、2021.07.01-2022.06.30、2022.07.01-2023.06.30。观察发现三个区段的相似性极高，仅在第三区段中的 2022.11.00-2022.12.00 区段出现较大差异。

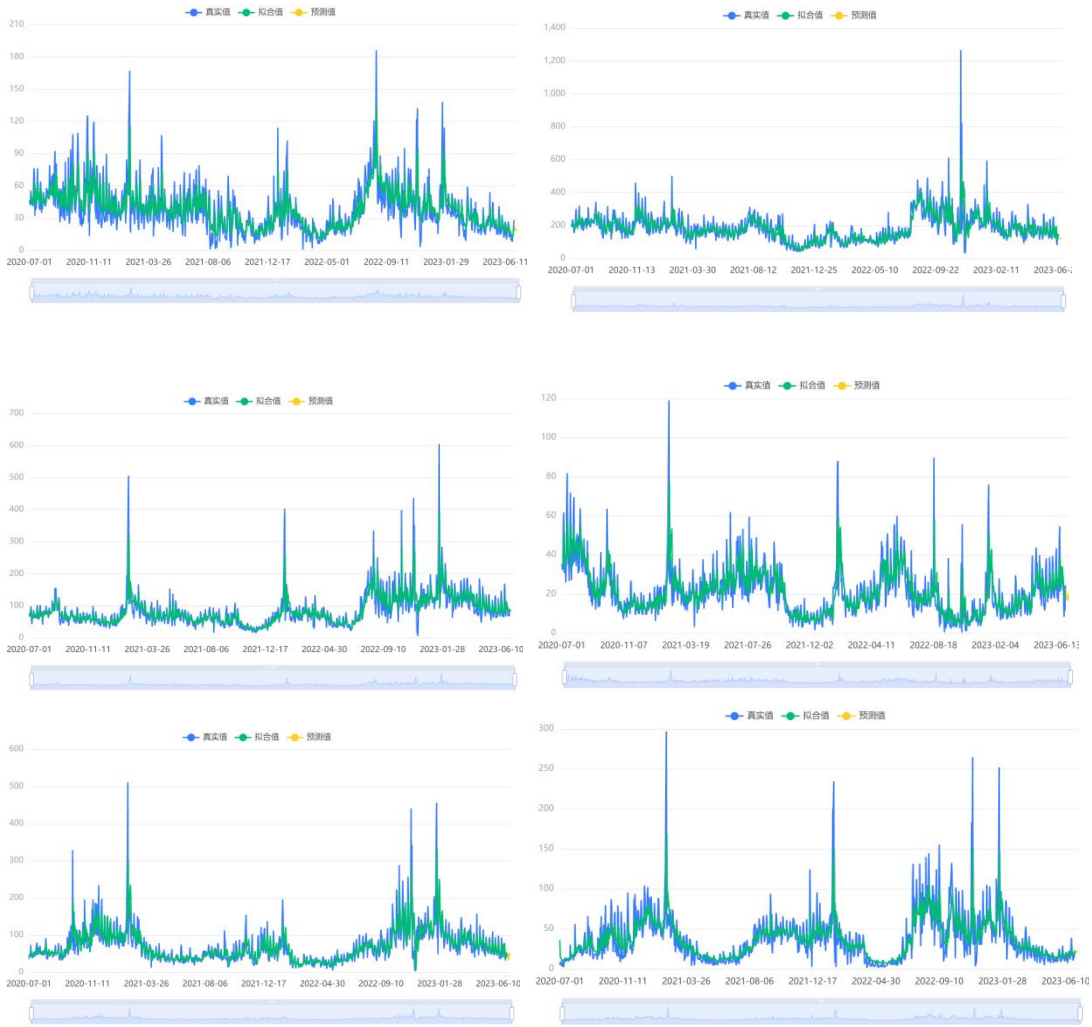

图 9 各品类销量时间序列图

（花叶：左上 花菜：右上 辣椒：左中 茄类：右中 食用菌：左下 水生根系：左下）

通过 ADF 检验，各品类的销量数据皆满足时间序列数据，并对各项数据进行 ARIMA 时间序列分析，各品类对应拟合度分别为：花菜类 **0.466**、辣椒类 **0.516**、茄类 **0.512**、食用菌 **0.531**、花叶类 **0.464**、水生根系 **0.495**。各品类所得拟合都都一般。通过时间序列分析生成上图，对比各品类时间序列图可知，花叶类销量随时间规律与茄类相符合、花菜类销量随时间规律与辣椒类相符合、食用菌类销量随时间规律与水生根系相符合。将各品类图形也分三个区段，同样体现一定的**季节性规律**。

（2）相互关系

差异性分析：

通过 python，对蔬菜各品类销量和单品销量排名前五的五类单品之间，进行 AVOVA 检验，得到 P 值分别为 **0.00** 和 **0.00**，说明数据之间存在差异性。在进行 AVOVA 检验的基础上，进行采用 Tukey HSD 方法，进行深度差异性检验，得到下表：

表五 Tukey HSD 分析结果表（蔬菜各品类）

Group1	*Group2*	*Meandiff*	*p-adj*	*lower*	*upper*	*reject*
水生根茎类	花叶类	0.0278	0.0	0.0227	0.0329	True
水生根茎类	花菜类	0.0225	0.0	0.0164	0.0286	True
水生根茎类	茄类	0.0261	0.0	0.0189	0.0332	True
水生根茎类	辣椒类	0.0226	0.0	0.0173	0.028	True
水生根茎类	食用菌	0.0229	0.0	0.0173	0.0284	True
花叶类	花菜类	-0.0053	0.0066	-0.0097	-0.001	True
花叶类	茄类	-0.0017	0.9582	-0.0074	0.004	False
花叶类	辣椒类	-0.0052	0.0011	-0.0084	-0.002	True
花叶类	食用菌类	-0.0049	0.0011	-0.0085	-0.0014	True
花菜类	茄类	0.0036	0.6306	-0.003	0.0102	False
花菜类	辣椒类	0.0002	1.0	-0.0045	0.0048	False
花菜类	食用菌类	0.0004	0.9999	-0.0045	0.0053	False
茄类	辣椒类	-0.0035	0.5584	-0.0094	0.0025	False
茄类	食用菌类	-0.0032	0.6641	-0.0094	0.0029	False
辣椒类	食用菌类	0.0002	1.0	-0.0037	0.0041	False

在结果中，每一行表示两组之间的比较结果。"groupl"和"group2"列显示了进行比较的两个组名称。"meandiff"列显示了两个组之间的平均差异值。"p-adj"列显示了经过多重校正后的 p 值，用于判断差异是否显著。"lower"和"upper"列分别显示了置信区间的下界和上界。最后一列"reject"表示是否拒绝零假设（即两组之间没有差异），如果拒绝，则表示两组之间差异显著。由表中可知，水生根茎类的销量数据跟与其余五个品类之间都存在显著性差异，花叶类与花菜类、辣椒类及食菌类之间的销量数据存在显著性差异，其余对于两两组别之间不存在差异性。

表六 Tukey HSD 分析结果表（单品销量前五）

Group1	*Group2*	*Meandiff*	*p-adj*	*lower*	*upper*	*reject*
云南生菜	净藕（1）	6.042	0.0000	3.8232	8.2594	Ture
云南生菜	芜湖青椒（1）	13.183	0.0000	10.8576	15.5084	Ture
云南生菜	西兰花	5.949	0.0000	3.7417	8.1577	Ture
云南生菜	西峡云菇（1）	-5.123	0.0000	-2.7727	-2.7727	Ture
净藕（1）	西峡云菇（1）	7.140	0.0000	9.3213	9.3213	Ture
净藕（1）	芜湖青椒（1）	-0.093	0.9999	1.9624	1.9624	False
净藕（1）	西兰花	-11.166	0.0000	-8.9581	-8.9581	Ture
芜湖青椒（1）	芜湖青椒（1）	-7.233	0.0000	-5.0608	-5.0608	Ture

芜湖青椒（1）	西兰花	-18.306	0.0000	-15.9891	-15.9891	Ture
西兰花	西峡云菇（1）	-11.073	0.0000	-8.8735	-8.8735	Ture

由表中可知，除了净藕（1）与芜湖青椒（1）之间不存在显著性差异外，其余组别之间都存在显著性差异。

相关性分析：

通过 SPSSPRO 软件，对蔬菜品类及单品销量排名前五的五类单品进行 Kendall 一致性检验，Kendall 系数一致性检验的结果显示，蔬菜品类销量的显著性 P 值为 **0.00**，水平上呈现显著性，拒绝原假设，因此数据呈现一致性，同时模型的 Kendall 协调系数 值为 **0.719**，因此相关性的程度为高度的一致性。单品销量排名前五的五类单品的显著性 P 值为 **0.00**，水平上同样呈现显著性，拒绝原假设，因此数据呈现一致性，同时模型的 Kendall 协调系数 值为 **0.194**，因此相关性的程度为极低的一致性。

通过 SPSSPRO 软件，对数据进行 Pearson 相关性分析，并绘制相关热力图。

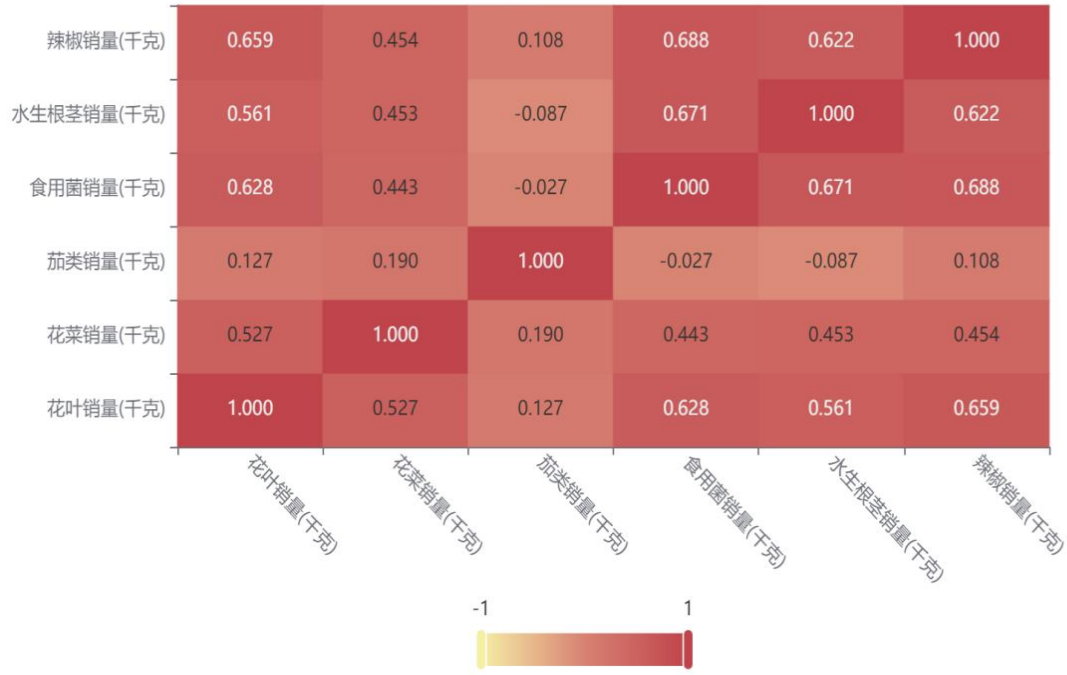

图 10 各品类销量相关性热力图

由上图可知，各品类销量之间都存在相关性，且相关性都较高。其中花叶类、辣椒类以及食用菌类对于其余类的相关更高，茄类对于其余类的相关性较低，且部分品类存在负相关。基本符合 Kendall 检验所得出的结论：Kendall 协调系数值为 **0.719**，因此相关性的程度为**高度的一致性**。

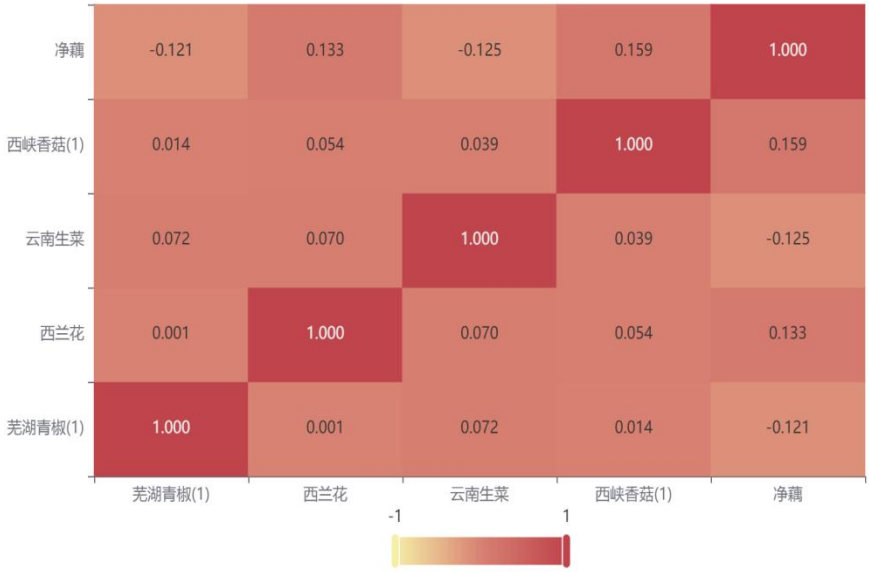

图 11 单品销量排名前五的五类单品相关性热力图

由上图可知，单品之间的存在相关性，但是五个单品所呈现的相关性都较低。基本符合 Kendall 检验所得出的结论：Kendall 协调系数 值为 **0.194**，因此相关性的程度为**极低的一致性**。

5.1.4 问题一模型的结果分析

（1）分布规律

基于模型求解所得数据进行分析，解释得出以下结论：

a、销量数据基本符合正态分布

b、花叶类蔬菜，是大众化的，民众比较倾向于购买花叶类蔬菜，对于茄类的购买比较少。部分品类单次销量高，总销量不高，说明蔬菜本身对于大众的容纳量也由一定关系。如：辣椒。

c、所有品类在一定区段都表现出相同的销售趋势，故蔬菜销售中存在季节性规律。

d、对于单品的销售量，由大到小进行排列之后，销售量呈现指数性下降，即蔬菜单品中，存在大众化蔬菜和小众化蔬菜。大众化蔬菜有：芜湖青椒（1）、西兰花、净藕（1）、大白菜、云南生菜等。小众化蔬菜有：红珊瑚（粗叶）、芥兰、紫白菜（2）、红橡叶、水果辣椒（橙色）等。

e、在相同品类中各单品销量也存在较大差异，如芜湖青椒与水果辣椒（橙色）差值为 28163.916kg。

f、通过分析时间序列图，三个区段在分布上呈现的规律相似性极高，仅在第三区段中的 2022.11.00-2022.12.00 区段出现较大差异。若不排除外界环境影响，则说明蔬菜销售在时间上存在季节性规律。

g、对比各品类时间序列图可知，花叶类销量随时间规律与茄类相符合、花

菜类销量随时间规律与辣椒类相符合、食用菌类销量随时间规律与水生根系相符合。将各品类图形也分三个区段，同样体现一定的季节性规律。

上述结论中，除季节性规律外，其余规律结论为统计学分析所得，基于数据量的支持，故结论具有可靠性。而季节性规律在统计学分析上和时间序列分析所得的规律相似，故结论也具有可靠性。

（2）相互关系

水生根茎类的销量数据跟与其余五个品类之间都存在显著性差异，花叶类与花菜类、辣椒类及食菌类之间的销量数据存在显著性差异，其余对于两两组别之间不存在差异性。净藕（1）与芜湖青椒（1）之间不存在显著性差异外，其余组别之间都存在显著性差异。

各品类销量之间都存在相关性，且相关性都较高。其中花叶类、辣椒类以及食用菌类对于其余类的相关更高，茄类对于其余类的相关性较低，且部分品类存在负相关。单品之间的存在相关性，但是五个单品所呈现的相关性都较低。Pearson 相关性分析所得的数据之间的相关性程度与 Kendall 一致性检验所的的一致性程度相符合，故该结论具有可靠性。

5.2 问题二模型的建立与求解

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，需要通过对已有数据进行分析处理，建立一个模型来对补货和定价策略进行预测，该模型要以利润最大化为目标，设置相关约束条件，通过有关方法，来最终确立模型。

5.2.1 问题二模型的准备

本题为围绕利润最大化为目标，设立相关约束条件，建立目标优化模型。先对所需分析数据进行整合，以确保后续，实现相关方程的建立。针对不同品类的销量、销售价格、进价以及损耗率进行分析，拟定约束条件，以求解最终利润最大化时的进货量以及定价决策。

5.2.2 问题二模型的建立

本问题为优化类问题，以求得在利润最大化为优化目标，建立单目标优化类模型。先对利润进行一个定义，在数据中选取对利润有相关的变量，围绕利润进行方程建立，得到利润方程：

$$\pi = x_1 Y - \frac{x_1}{1-\eta} x_2$$

其中 π 为利润，x_1 为销售量，x_2 进价，η 为损耗率。

再对销售价、销售量以及进价，进行多元线性回归分析，得到销售价、销售

量以及进价之间的相关性方程：

$$Y = \beta_1 x_1 + \beta_2 x_2 + \beta_0$$

其中 Y 为销售价，x_1 为销售量，x_2 进价，

联立利润方程和销售价、销售量以及进价之间的相关性方程得到利润最大化模型：

$$\pi = \beta_1 x_1^2 + x_1(\beta_2 x_2 + \beta_0 - \frac{x_2}{1-\eta})$$

其中β为相关性系数。

对该模型求偏导：

$$\frac{d\pi}{dx_1} = 2\beta_1 x_1 + \beta_2 x_2 + \beta_0 - \frac{x_2}{1-\eta}$$

将相关数据带入模型方程，进行求解，得到在利润在最大化时的销售量和销售价，即进货量和定价。

建立报童模型，对日补货总量进行决策，得到最佳补货量。

多元线性回归分析：

用于探究一个或多个自变量与一个连续因变量之间的关系。在多元线性回归中，假设存在一个因变量和多个自变量。通过建立一个数学模型，可以通过自变量的值来预测因变量的值。多元线性回归模型的基本形式如下：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$$

其中，Y 表示因变量，X1、X2、...、Xp 表示自变量，β0、β1、β2、...、βp 表示模型的参数，ε表示误差项。模型的目标是通过对参数进行估计，最小化观测值与模型预测值之间的误差。

对相关性方程所包含的数据量进行描述性统计。

VIF 检验：

多元线性回归中用于检验共线性问题的一种方法。VIF 检验通过计算每个自变量的方差膨胀因子（VIF 值），来评估该自变量与其他自变量的相关性程度。相关公式如下：

$$\mathrm{VIF_i} = \frac{1}{1 - \mathrm{R}^2_{1-k/i}}$$

$\mathrm{R}^2_{1-k/i}$ 是将第 i 个自变量作为因变量，剩下的 $k-1$ 个自变量回归得到的拟合度。

$$\mathrm{VIF} = \max\{\mathrm{VIF_i}\}$$

若 VIF ＜ 10 则不存在严重的多重共线问题，反之则存在。

报童模型：

报童模型一种库存管理模型，常用于决策如何合理订购货物数量的问题。它以一个类比为背景，假设有一个报童需要在每天早上决定购买多少份报纸以满足当天的需求，目标是最大化利润或最小化成本。

在本题中，采用报童模型基于利润最大化模型所求得的最佳售价和进价求得在利润最大化时的日补货量。公式及原理如下：

利润与最优销售价格、补货量、批发价格之间的关系

补货量大于需求量：

$$\pi = P(x-q) - C_2(Q+q-x) - C_1Q \quad , \quad x \le Q+q$$

需求量大于补货量

$$\pi = PQ - (x-Q-\mathrm{q})(P-C_1) - C_1Q \quad , \quad x > Q+q$$

上述式子中，π 表示利润，P 表示最优销售价格，Q 表示补货量，q 表示前几天的库存量，C_1 表示批发价格，C_2 表示库存成本。

商超当天的销售利润期望值：

$$E(\pi) = \int_0^Q [(P+C_2)x - (C_1+C_2)Q]f(x)dx + \int_Q^{+\infty} [(-P+C_1)x + 2(P-C_1)Q]f(x)dx$$

其中 $\int_0^{+\infty} xf(x)dx = \mu$，$\int_0^{+\infty} xf(x)dx = 1$，$f(x)$ 销量的概率密度函数，$F(x)$ 销量的分布函数：

$$E(\pi) = (2P - C_1 + C_2)\int_0^Q xf(x)dx - (2P - C_1 + C_2)Q\int_0^Q f(x)dx + (C_1 - P)\mu + 2(P - C_1)Q$$

上式两边对 Q 求一阶偏导得：

$$\frac{\partial E(\pi)}{\partial Q} = 2(P - C_1) - (2P - C_1 + C_2)F(Q)$$

再对上式两边对 Q 求二阶偏导得：

$$\frac{\partial^2 E(\pi)}{\partial Q^2} = -(2P - C_1 + C_2)F'(Q)$$

已知 $P > C_1 > C_2$，故 $-(2P - C_1 + C_2) < 0$，因此 $\frac{\partial^2 E(Q)}{\partial Q^2} < 0$ 可知 $E(\pi)$ 存在最大值，且当 $\frac{\partial^2 E(Q)}{\partial Q^2} = 0$ 时，存在最优补货量 Q*使得 $E(\pi)$ 最大。当 $\frac{\partial^2 E(Q)}{\partial Q^2} = 0$ 时带入可得：$F(Q^*) = \frac{2(P - C_1)}{2P - C_1 + C_2}$，当 $F(Q^*) = \frac{2(P - C_1)}{2P - C_1 + C_2} = T$ 查阅标准正太分布表可得：$\frac{Q^* - \mu}{\sigma} = t$ 最后解出最优补货量 Q*

5.2.3 问题二模型的求解

通过模型建立过程，对各操作，方法模型进行求解得到以下数据：

表七 描述性统计表

变量名称	数据量	平均值	标准差	最小值	最大值
销售价格	17943	7.418797	5.723521	1	119.9
总销量	17943	10.3245	13.62292	0.006	340
批发价格	17943	4.438432	3.941633	0.01	65.41

上表对于所需要分析的销售价格、销量、批发价格做了一个描述性统计。

表八 方差分析表（一）

方差来源	平方和	自由度	均方误差
回归	536940.065	2	268470.033
残差	50816.4615	17940	2.83257868
总计	587759.527	17942	32.7586962

表九 方差分析表（二）

数量级	F 值	P 值	R^2 值	均方根误差	判定系数
17943	94779.37	0.000	0.9135	0.9135	1.683

通过上述方差分析表可以得出所得拟合度指标 **Adj R-squared=0.9135** 大于 0.9000，拟合度优秀，故该回归模型有良好的解释性。

表十 VIF 检验结果表

变量名称	VIF	1/VIF
总销量	1.05	0.951784
批发价格	1.05	0.951784

由上表可得，上述表中所含的数据变量中，总销量和批发价格的 VIF 值都小于 10，故总销量与批发价格之间不存在多重共线性。

通过多元线性回归得到各品类关于销售价格与销量和批发价格之间的相关

系数方程：

花叶类：$Y=1.471+1.38x_2-0.15x_1$

花菜类:$Y=1.873+1.264x_2-0.006x_1$

茄类：$Y=2.170+1.265x_2-0.032x_1$

水生根系类：$Y=2.397+1.319x_2-0.043x_1$

食用菌：$Y=1.151+1.50x_2-0.023x_1$

辣椒类：$Y=2.992+1.376x_2-0.056x_1$

利润方程为：$\pi=x_1Y-\frac{x_1}{1-\eta}x_2$

分别对各品类的相关系数方程与利润方程进行联立，其中进价和损耗率都取总数据的平均值，得到利润最大化时，各品类的定价，然后再利用报童模型求解出各品类日补货量。

表十一 各品类未来七天补货量与定价

品类		7月1号	7月2号	7月3号	7月4号	7月5号	7月6号	7月7号
花叶类	补货量	193.31	182.97	178.98	128.75	178.84	165.28	172.42
	定价	2.15	3.42	5.66	5.33	5.33	9.27	7.97
花菜类	补货量	33.21	35.58	35.87	34.44	32.44	37.16	35.01
	定价	8.578	6.31	6.37	7.89	8.80	5.11	6.45
茄类	补货量	21.10	19.65	18.34	18.73	17.67	17.28	17.68
	定价	5.61	6.03	7.78	8.36	9.54	10.70	8.86
水生根茎	补货量	45.86	38.65	37.08	40.22	34.89	31.13	28.62
	定价	3.76	5.69	7.64	7.54	7.96	10.26	13.10
食用菌	补货量	69.64	69.67	61.39	61.43	61.20	59.94	58.94
	定价	5.03	4.35	7.81	9.98	10.25	14.00	18.04
辣椒类	补货量	94.03	86.01	78.00	72.12	74.79	75.86	80.13
	定价	5.77	7.82	11.44	17.47	14.47	13.02	10.57

如上表所示，表中含有各品类未来七天的补货量与定价数值，该表即为本题最终的日补货量与定价策略，在该日补货量和定价的情况下，未来一周每日的收益最大。

5.2.4 问题二模型的结果分析

通过建立利润最大化模型结合报童模型，求解出未来一周的日补货量和定价，表十一 各品类未来七天补货量与定价中所给出的日补货量以及定价即为日补货量和定价策略。在该日补货量和定价的情况下，未来一周每日的收益最大。

表十二 结果可靠性分析表

品类	花叶类	花菜类	茄类	水生根系	食用菌	辣椒类
总销量（真实值）	198520.978	41766.451	22431.782	40581.353	76086.725	91588.629
总销量（计算值）	185399.22	37635.787	20145.207	39603.207	68289.858	86625.162

上表为模型结果可靠性分析表，由表可知在销量上所呈现的趋势符合，问题一所得出的时间序列分析可以证明 7 月份的销售量开始从一个小低谷进行爬升趋势，故用该时间段所描述的三年总销量，要表现较低，符合上表。

5.3 问题三模型的建立与求解

本题类型与问题二一致，存在额外的约定条件，分析数据项由 6 个蔬菜品类转换成 27-33 个单品。根据单品的各项数据与利润建立模型。求解 27-33 个单品在 7 月 1 日的单品补货量和定价策略。

5.3.1 问题三模型的准备

将所需要的数据进行整合，进行分析筛选出 2023 年 6 月 24-30 日的可售单品，对这些可售单品进行评价类分析，选取排名前 33 项的单品进行分析，得到 33 项在利润最大化的情况下的单品补货量与单价决策。再通过利润最大化目标以及约束条件进行删选决定最终订购单品数量。

5.3.2 问题三模型的建立

本问题为优化类问题，以求得在利润最大化为优化目标，建立单目标优化类模型。先对利润进行一个定义，在数据中选取对利润有相关的变量，围绕利润进行方程建立，得到利润方程：

$$\pi = x_1 Y - \frac{x_1}{1-\eta} x_2$$

其中 π 为利润，x_1 为销售量，x_2 进价，η 为损耗率。

再对销售价、销售量以及进价，进行多元线性回归分析，得到销售价、销售量以及进价之间的相关性方程：

$$Y = \beta_1 x_1 + \beta_2 x_2 + \beta_0$$

其中 Y 为销售价，x_1 为销售量，x_2 进价，

联立利润方程和销售价、销售量以及进价之间的相关性方程得到利润最大化模型：

$$\pi = \beta_1 x_1^2 + x_1(\beta_2 x_2 + \beta_0 - \frac{x_2}{1-\eta})$$

其中β为相关性系数。

通过基于熵值法的秩和比综合评价法，从 2023 年 6 月 24-30 日可选单品中，从销售价格、总销量、损耗率、收益、批发价格表现，分析各单品的权重，取前 33 项作为分析对象。

基于熵值法的秩和比综合评价法：

秩和比综合评价法用于评估多个因素或方案的优劣，并确定最佳选择。该方法将多个因素或方案进行排名，然后计算各个排名之间的秩和比来比较它们的综合表现。本问采用基于熵值法的秩和比综合评价法，即引入了权重的考虑。在进行秩和比综合评价法的之前，需对数据进行归一化处理。秩和比综合评价法相关公式：

求得秩矩阵：$R = (R_{ij})_{m\times n}$，R_{ij} 表示第 i 行第 j 列元素的秩

对于效益型指标：

$$R_{ij} = 1 + (n-1)\frac{X_{ij} - \min(X_{1j}, X_{2j}, ..., X_{nj})}{\max(X_{1j}, X_{2j}, ..., X_{nj}) - \min(X_{1j}, X_{2j}, ..., X_{nj})}$$

对于成本型指标：

$$R_{ij} = 1 + (n-1)\frac{\max(X_1, X_{2j}, ..., X_{nj}) - X_{ij}}{\max(X_{1j}, X_{2j}, ..., X_{nj}) - \min(X_{1j}, X_{2j}, ..., X_{nj})}$$

计算 RSR_i 值：

$$RSR_i = \frac{1}{n}\sum_{j=1}^{m} w_j R_{ij}$$

$$\text{权重 } w_j = \frac{1}{m}$$

计算得到的 RSR，即可作为各对象综合评价得分。本文根据 RSR 数值的排名选取前 33 项单品。

5.3.3 问题三模型的求解

通过 SPSSPRO 软件中基于熵值法的秩和比综合性评价法，从可选单品中挑选下表内中的 33 种：

表十三 单品 RSR 排名表

单品名称	RSR 排名
西兰花	1
小米辣（份）	2
四川红香椿	3
西峡花菇	4
大白菜	5
红椒（1）	6
紫茄子（2）	7
净藕（1）	8
红杭椒	9
泡泡椒（精品）	10
娃娃菜	11
小米椒	12
黄花菜	13
金针菇（袋）（1）	14
保康高山大白菜	15
奶白菜（份）	16
金针菇（盒）	17
黄白菜（2）	18
虫草花	19
洪湖藕芋	20
紫苏	21
蒲公英	22
丝瓜尖	23
黑皮鸡枞菌	24
芜湖青椒（1）	25
云南生菜	26
云南油麦菜	27
枝江青梗	28
双孢菇	29
螺丝椒	30
蒜小米椒组合装	31
菠菜	32
竹叶菜	33

本题在后续模型的数据分析中，对上表所述的 33 中品类的各项数据进行模型求解，求得相应的单日补货量和定价决策。

通过多元线性回归得到 33 个单品关于销售价格与销量和批发价格之间的相关系数方程：

$$Y = \beta_1 x_1 + \beta_2 x_2 + \beta_0$$

分别对 33 个单品的相关系数方程与利润方程进行联立，其中进价和损耗率

都取总数据的平均值，得到利润最大化时，求得单日补货量和定价决策结果。

表十四 33 个单品的单日补货量和定价决策结果表

单品名称	日补货量（kg）	定价（kg/元）
西兰花	12.92	8.97
小米辣（份）	25.45	4.84
四川红香椿	5.45	14.70
西峡花菇	14.08	5.71
大白菜	10.18	4.42
红椒（1）	30.57	12.65
紫茄子（2）	16.78	7.41
净藕（1）	14.36	7.55
红杭椒	11.11	5.63
泡泡椒（精品）	12.73	9.31
娃娃菜	10.52	4.70
小米椒	10.62	6.92
黄花菜	0.81	7.89
金针菇（袋）（1）	30.90	2.74
保康高山大白菜	21.73	3.03
奶白菜（份）	15.37	2.90
金针菇（盒）	16.97	3.11
黄白菜（2）	13.41	5.88
虫草花	12.72	12.24
洪湖藕芋	16.03	8.15
紫苏	13.05	5.53
蒲公英	0.72	6.34
丝瓜尖	1.89	14.33
黑皮鸡枞菌	6.27	15.33
芜湖青椒（1）	21.58	6.67
云南生菜	25.62	3.79
云南油麦菜	23.27	3.52
枝江青梗	21.10	8.93
双孢菇	9.11	9.18
螺丝椒	15.34	4.75
蒜小米椒组合装	13.08	3.05
菠菜	25.35	4.43
竹叶菜	16.40	7.13

上表为经过模型分析求解，在利润最大化时的 33 个单品的单日补货量和定价决策结果表，商超对于上述表格中的各单品补货量和定价进行销售，可以得到最大化利润。其中黄花菜、蒲公英、丝瓜尖的单日日补货量少于 2.5kg，但问题要求单品订购量满足最小陈列量 2.5 千克，若黄花菜、蒲公英、丝瓜尖的单日补货量达到 2.5kg，则不满足利润最大化目标，可能造成负面影响，即亏损。故本题最终应选择 30 个单品数量。

5.3.3 问题三模型的结论分析

通过基于熵值法的秩和比综合分析得到预先选择的 33 个单品，对 33 个单品的数据量进行模型分析求解，得到表十四 33 个单品的单日补货量和定价决策结果表，再根据利润最大化目标原则以及相关的约束条件，最终应排除黄花菜、蒲公英、丝瓜尖，这 3 个单品，最终选取 30 个单品。得到最终商超在 2023 年 7 月 1 日的 30 个单品的单日补货量和定价决策。

表十五 30 个单品的单日补货量和定价决策结果表

单品名称	日补货量（kg）	定价（kg/元）
西兰花	65.92	8.97
小米辣（份）	241.45	4.84
四川红香椿	5.45	34.70
西峡花菇	139.08	18.71
大白菜	10.18	4.42
红椒（1）	184.57	12.65
紫茄子（2）	151.78	7.41
净藕（1）	98.36	7.55
红杭椒	11.11	23.63
泡泡椒（精品）	126.73	9.31
娃娃菜	181.52	4.70
小米椒	10.62	27.92
金针菇（袋）（1）	130.90	2.74
保康高山大白菜	120.73	3.03
奶白菜（份）	103.37	2.90
金针菇（盒）	693.97	3.11
黄白菜（2）	443.41	5.88
虫草花	11.72	24.24
洪湖藕芋	156.03	8.15
紫苏	7.05	20.53
黑皮鸡枞菌	23.27	87.33
芜湖青椒（1）	879.58	6.67
云南生菜	709.62	3.79
云南油麦菜	238.27	3.52
枝江青梗	21.10	8.93
双孢菇	9.11	20.18
螺丝椒	57.34	4.75
蒜小米椒组合装	276.08	3.05
菠菜	61.35	4.43
竹叶菜	28.40	7.13

模型在第二问的结论分析种已经得到了可靠性的解释，故在本题中也具有可靠性。

5.4 问题四的分析与结论

在解决前三问的时候，出现一定的问题，以及在对模型的构建中，做一系列合理假设的优化方向。

5.4.1 问题四的分析

对于问题一，本文在分布规律上进行了有关时间序列的分析，即从时间关系方向对数据进行规律分析，故在解决过程中，提出是否在空间关系上也存在规律。基于问题所给出的数据，显然不能够完成。若能收集顾客的反馈情况，即可以直观的分析顾客对于蔬菜的喜好规律。

对于问题二，本文在对数据之间的相互关系，建立方程与模型时，出现缺乏数据项，只能用其余数据项，经过一系列处理之后代替所需要的数据项。如：库存量若能收集蔬菜是否应季和保险周期等商品标签数据，则可以从蔬菜自身做一定的约束条件，从而可以降低蔬菜自身的损耗率，进而提升利润。若能收集商超周边一定范围内的住户对于蔬菜的单日需求量，则可以把控单日进货量的最大值的把控，从而减少总体的蔬菜损耗率，进而提升利润。若能收集地域性数据，如地方饮食习惯和经济发展，则可以对品类日补货量占比进行分析，从而促进库存量的销售，进而提升利润，而经济发展数据，则可以根据具有经济水平支持的前提下，保持销量，然后在符合市场对于蔬菜售价调控的基础上，适当提升蔬菜售价，从而获得等大利润。

对于问题三、本文所采用的模型与问题二基本相同，故出现相同的问题。

5.4.2 问题四的结论

为了更好的制定蔬菜商品的补货和定价决策，商超主要还需收集四类数据：商超本身的数据、需求端的数据、供货端的数据以及地域性的数据。

商超本身的数据：库存量，库存量可以更好的拟合模型的建立，从而使模型准确性更佳，进而得到更准确的日供货量和定价策略。

需求端的数据：顾客反馈情况、单日需求量，该类数据可以之间或间接影响单日补货量决策。

供货端的数据：蔬菜是否应季、保鲜周期等商品标签数据，该类数据都可以直接或间接的降低对于蔬菜自身的损耗率，从而提升利润。

地域性的数据：地方饮食习惯、经济发展，通过分析可以用于调整单日补货量和定价策略。

六、模型评级与优化

6.1 模型的优点

1、在采用分析方法时，前后两类分析能够相互印证，确保了结果的可靠性;2、在使用方法和建立模型之前，进行一定的数据检验，确保了方法和模型的可靠性。

6.2 模型的缺点

1、在选取样本时，对过于庞大的数据进行了简化，可能会导致结果的偏差、使得所建立的模型不能运用到其它同类问题中。2、本模型在进行分析处理时，仅对表中含有的数据进行处理，未考虑其它因素对于模型的影响。3、问题所给出的数据不包含库存量，在数据分析时，将其他数据代替库存量建立模型。

6.3 模型的优化及推广

综合考虑多方面因素，对数据进行分析，用更加严谨的数据关系，去建立模型，使得模型更加符合实际情况。

本模型在处理未给出库存量的优化类中的补货类问题，可行性高。

七、参考文献

[1] 丁若薇. A 餐饮企业生鲜水产品订货策略研究[D]. 北京交通大学,2021.DOI:10.26944/d.cnki.gbfju.2021.003825.

[2]谢如鹤,罗湖桥,陈冠名.基于初始新鲜度的生鲜农产品订货策略[J].包装工程,2020,41(13):179-184

[3] 马鹏, 张晨. 绿色供应链背景下互补品定价策略[J]. 控制与决策,2018,33(10):1861-1870.

[4]张文华.ARIMA 模型用于时间序列预测的算法改进与仿真[J].信息与电脑(理论版),2021,33(05):53-56.

[5]王学萍.运用秩和比法综合评价医疗质量[J].中国医院统计,1994(01):37-38.

八、附录

相关操作代码：

```
数据处理（和表分表）+皮尔逊相关分析
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import spearmanr

df1 = pd.read_excel(r'E:\数学建模\day05\附件 1.xlsx')
df2 = pd.read_excel(r'E:\数学建模\day05\附件 2.xlsx')
df3 = pd.read_excel(r'E:\数学建模\day05\附件 3.xlsx')

merged_df = df2.merge(df1[['单品编码', '单品名称','分类名称']], on='单品编码', how='left')
sales_by_category = merged_df.groupby('分类名称')['销量(千克)'].sum()

df = sales_by_category.to_frame()

data1 = pd.read_excel(r'D:\桌面\按菜种类划分  日期排序  12 合并.xlsx', parse_dates=['销售日期'],
date_parser=lambda x: pd.to_datetime(x, format='%Y-%m-%d %H:%M:%S'))
groups = [df["销量(千克)"].dropna().values for _, df in data1.groupby("分类名称")]

for group in groups:
    if np.isnan(group).any():
        print("该组含有空值")

for group in groups:
    print(len(group))

groups_1000 = [arr[:1000] for arr in groups]
num_groups = len(groups_1000)
correlation_matrix = np.zeros((num_groups, num_groups))

for i in range(num_groups):
    for j in range(num_groups):
        coef, _ = spearmanr(groups_1000[i], groups_1000[j])
        correlation_matrix[i, j] = coef

plt.figure(figsize=(8, 6))
plt.imshow(correlation_matrix, cmap='coolwarm', vmin=-1, vmax=1)

for i in range(num_groups):
    for j in range(num_groups):
```

```
            plt.text(j, i, f"{correlation_matrix[i, j]:.2f}",
                     ha="center", va="center", color="black")

plt.colorbar(label="Spearman Rank Correlation Coefficient")
plt.xlabel("Group")
plt.ylabel("Group")
plt.title("Correlation Matrix of Group Pairs")
plt.show()

product_groups = [df["销量(千克)"].dropna().values for _, df in data1.groupby("单品名称")]
num_groups = len(product_groups)
print("分组数目：", num_groups)

group_counts = data1.groupby("单品名称")["销量(千克)"].count()
top_five_groups = group_counts.sort_values(ascending=False).head(5)

print(top_five_groups)

top_five_group_names = top_five_groups.index.tolist()

filtered_data = data1[data1["单品名称"].isin(top_five_group_names)]

product_groups = [df["销量(千克)"].dropna().values for _, df in filtered_data.groupby("单品名称")]

groups_1000 = [arr[:100] for arr in product_groups]

num_groups = len(groups_1000)
correlation_matrix = np.zeros((num_groups, num_groups))

for i in range(num_groups):
    for j in range(num_groups):
        coef, _ = spearmanr(groups_1000[i], groups_1000[j])
        correlation_matrix[i, j] = coef

plt.figure(figsize=(8, 6))
plt.imshow(correlation_matrix, cmap='coolwarm', vmin=-1, vmax=1)

for i in range(num_groups):
    for j in range(num_groups):
        plt.text(j, i, f"{correlation_matrix[i, j]:.2f}",
                 ha="center", va="center", color="black")

plt.colorbar(label="Spearman Rank Correlation Coefficient")
plt.xlabel("Group")
```

```
plt.ylabel("Group")
plt.title("Correlation Matrix of Group Pairs")
plt.show()
```

```
import pandas as pd
from scipy.stats import f_oneway
from statsmodels.stats.multicomp import pairwise_tukeyhsd

# 创建一个空的 DataFrame
df = pd.DataFrame()

# 将每个产品组添加为 DataFrame 的列
for i, group in enumerate(product_groups):
    column_name = f'Group {i+1}'
    df[column_name] = pd.Series(group)

# 输出到 Excel 文件
df.to_excel(r'd:\桌面\product_groups.xlsx', index=False)

# 读取数据
data1 = pd.read_excel(r'D:\桌面\单品.xlsx')

# 进行 ANOVA 检验
groups = []
for group_name, group_data in data1.groupby('单品名称'):
    groups.append(group_data['销量(千克)'])
f_statistic, p_value = f_oneway(*groups)
print("ANOVA 检验结果：")
print("F 值:", f_statistic)
print("p 值:", p_value)

# 执行 Tukey's HSD 测试
tukey_results = pairwise_tukeyhsd(data1['销量(千克)'], data1['单品名称'])

# 将结果转换为 DataFrame 并打印出来
tukey_results_df = pd.DataFrame(data=tukey_results._results_table.data[1:],
columns=tukey_results._results_table.data[0])
print(tukey_results_df)
```

ANOVA 检验+差异性分析

```
import pandas as pd
from scipy.stats import f_oneway
from statsmodels.stats.multicomp import pairwise_tukeyhsd
```

数据处理+描述统计

```
# 创建一个空的 DataFrame
df = pd.DataFrame()

# 将每个产品组添加为 DataFrame 的列
for i, group in enumerate(product_groups):
    column_name = f'Group {i+1}'
    df[column_name] = pd.Series(group)

# 输出到 Excel 文件
df.to_excel(r'd:\桌面\product_groups.xlsx', index=False)

# 读取数据
data1 = pd.read_excel(r'D:\桌面\单品.xlsx')

# 进行 ANOVA 检验
groups = []
for group_name, group_data in data1.groupby('单品名称'):
    groups.append(group_data['销量(千克)'])
f_statistic, p_value = f_oneway(*groups)
print("ANOVA 检验结果：")
print("F 值:", f_statistic)
print("p 值:", p_value)

# 执行 Tukey's HSD 测试
tukey_results = pairwise_tukeyhsd(data1['销量(千克)'], data1['单品名称'])

# 将结果转换为 DataFrame 并打印出来
tukey_results_df = pd.DataFrame(data=tukey_results._results_table.data[1:],
columns=tukey_results._results_table.data[0])
print(tukey_results_df)
```

```
import pandas as pd
import matplotlib.pyplot as plt
from matplotlib.font_manager import FontProperties

# 读取数据
data = pd.read_excel(r'D:\桌面\按菜种类划分 日期排序 12 合并.xlsx', parse_dates=['销售日期
'], date_parser=lambda x: pd.to_datetime(x, format='%Y-%m-%d %H:%M:%S'))

# 检查列名是否包含"销售日期"
if '销售日期' in data.columns:
```

```
    print("数据表中存在'销售日期'这一列")
else:
    print("数据表中不存在'销售日期'这一列")

# 设置销售日期列为索引
data.set_index('销售日期', inplace=True)

# 设置中文字体
font = FontProperties(fname='/path/to/your/font.ttf', size=12)
plt.rcParams['font.family'] = font.get_name()

# 绘制销量排名前 10 的单品各年份季度对比图
years = [2020, 2021, 2022, 2023]

for year in years:
    # 筛选出指定年份的数据
    data_year = data[data.index.year == year]

    # 添加季度列
    data_year['季度'] = data_year.index.quarter

    # 按季度和单品对销量进行汇总
    quarterly_product_sales = data_year.groupby(['季度', '单品名称'])['销量(千克)'].sum().unstack()

    # 对每个季度的销量进行排名
    quarterly_product_sales_ranked = quarterly_product_sales.rank(axis=1, ascending=False)

    # 根据排序结果重新排列列顺序
    quarterly_product_sales_sorted_columns = quarterly_product_sales[quarterly_product_sales_ranked.columns]

    # 获取前 10 个单品名称
    top_10_products = quarterly_product_sales_ranked.columns[:10]

    # 绘制各季度排名前 10 单品销量对比图
    plt.rcParams['font.size'] = 10  # 设置字体大小为 10
    quarterly_product_sales_sorted_columns[top_10_products].plot(kind='bar', stacked=True)
    plt.xlabel('季度')
    plt.ylabel('销量')
    plt.title(f'{year}年各季度排名前 10 单品销量对比图')
    plt.legend(title='单品', bbox_to_anchor=(1, 1))
    plt.show()
```