问题重述

一、 问题背景

随着数字经济时代的到来，零售行业正经历着深刻的转型与升级，精细化运营已成为商超企业在激烈市场竞争中脱颖而出的核心能力。生鲜商品作为商超吸引客流、提升消费黏性的关键品类，其经营管理水平直接关系到企业的整体盈利能力与市场地位。

在生鲜商品中，蔬菜类商品因其固有的保鲜期短、易损耗、品相随时间快速下降等特性，给商超的日常运营带来了巨大挑战。商超通常需要在每日凌晨，即不完全掌握当日确切进货价格与市场需求的情况下，迅速制定出品种繁多、来源各异的蔬菜商品的补货与定价计划。

由此，商超经营的核心矛盾便凸显出来：一方面，若补货量过大或定价不合理，将导致商品滞销、损耗增加，最终侵蚀利润；另一方面，若补货量不足，则会错失销售机会，影响顾客满意度与忠诚度。如何在信息不充分的条件下，实现供给与需求的动态平衡，成为蔬菜供应链管理中亟待解决的现实难题。

因此，运用数学建模、数据分析等科学方法，深入挖掘历史销售数据、批发价格及损耗率中蕴含的规律与关联，构建精准的需求预测与优化决策模型，对于提升商超在蔬菜类商品经营上的科学决策水平、降低运营风险、实现收益最大化具有至关重要的理论价值与现实指导意义。

二、 问题要求

为应对上述挑战，并实现商超运营利润最大化的最终目标，我们需要基于所提供的附件数据与资料，建立数学模型，系统性地解决以下具体问题：

问题一：深入分析各蔬菜品类及单品销售量的统计分布规律，并量化不同品类之间、不同单品之间在销售量上可能存在的相互关联关系。

问题二：在以“品类”为单位进行补货的框架下，研究各蔬菜品类的销售总量与“成本加成定价”策略之间的关系。基于此关系，建立优化模型，求解并给出未来一周（2023年7月1日至7日）各品类的每日补货总量与最优定价策略，以实现商超总体收益的最大化。

问题三：考虑到销售空间的物理限制，进一步细化补货颗粒度至“单品”级别。要求在可售单品总数介于27至33个，且每个单品订购量不低于2.5千克最小陈列量的约束下，依据2023年6月24日至30日的可售品种信息，为7月1日制定具体的单品补货量与定价策略，其目标是在尽可能满足市场需求的前提下，最大化当日商超的预期收益。

问题四：从优化决策模型的实际应用与持续改进角度出发，探讨为更精准地制定蔬菜商品的补货与定价策略，商超还需要采集哪些额外的数据。并请详细阐述这些数据将如何帮助解决上述问题，并说明其重要性与理由。

问题分析

问题一的分析

针对问题一第一小问，首先，对原始数据进行预处理。合并附件1与附件2，将单品与其所属品类进行匹配，并对数据进行清洗，剔除无效或异常条目。接着，以“日”为单位，分别对6个蔬菜品类及所有单品的销售量进行聚合统计，形成后续分析所需的时间序列数据集。

其次，对于品类和代表单品， 通过描述性统计分析和数据可视化，对于统计结果进行定性分析，把握数据的集中趋势、离散程度和分布形状特征。最后通过自相关性分析，从统计学角度对前文观察到的周期性规律进行定量验证 。

第二小问对不同品类及单品间的相互关系进行深入挖掘。 首先，销售高峰之间可能存在的非线性单调关系，我们选择计算各品类及单品日销售量之间的Spearman等级相关系数，稳健地度量商品间的关联性。随后，利用热力图对该矩阵进行可视化，直观地识别出具有强正相关或负相关的商品对。 接着，以各单品的日销售量时间序列作为输入变量，通过PCA将高维的销售数据投影到少数几个关键的主成分上。 最后，通过分析不同单品在主要主成分上的载荷，识别出在同一主成分上均具有高载荷的商品组合。

#### 问题一模型的建立与求解

#### 数据整合与格式规范化

首先，我们对原始数据进行了整合与规范化。读取附件1（商品信息）与附件2（销售流水明细），以“单品编码”作为唯一键，将两个数据表进行合并，为每一条销售记录补充其对应的“分类名称”等维度信息，形成一个完整的数据集。

随后，对关键字段进行格式转换。将“销售日期”列由文本格式统一转换为标准的日期时间（datetime）格式，这是后续进行任何时间序列分析的基础。同时，对数据集中存在的缺失值和明显错误数据进行处理，对于数值型字段的缺失，我们采用0值填充，这在零售场景下符合“当日无销售记录”的业务逻辑；对于销量或价格为负数等明显不合逻辑的条目，则予以直接剔除，以保证数据的真实性。

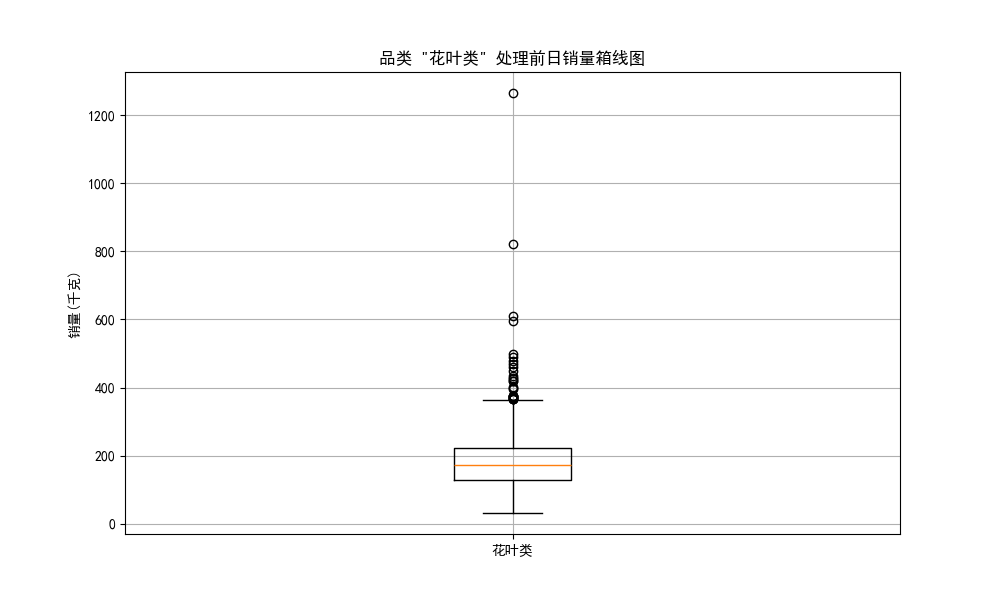
#### 数据聚合与特征提取

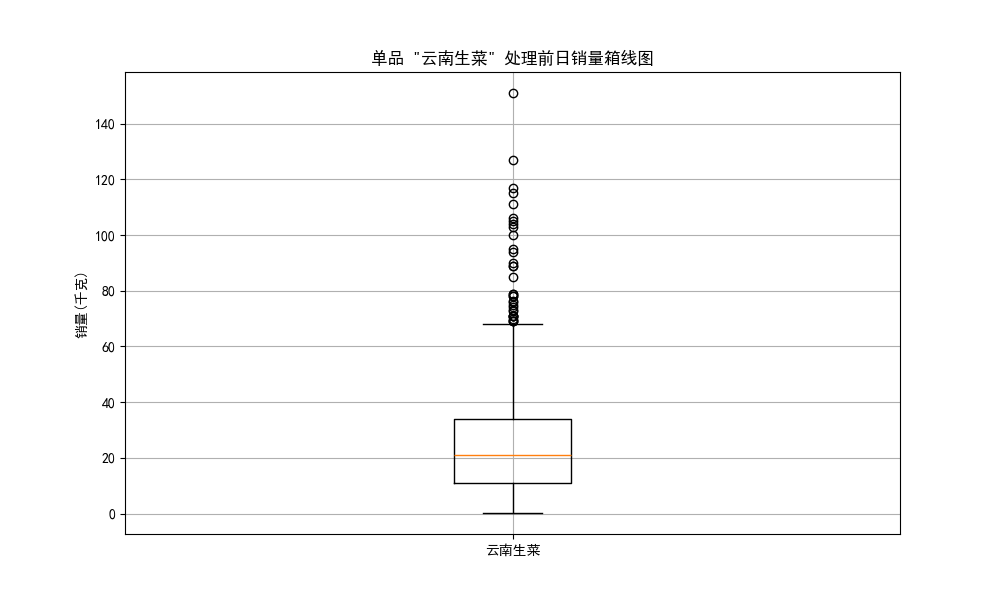
在数据清洗的基础上，为满足后续不同粒度模型的分析需求，我们对数据进行了聚合处理。以“销售日期”和“分类名称”为分组依据，计算出每个品类在每一天的销售总量，形成品类日销售数据集。同理，以“销售日期”和“单品名称”为分组，计算出每个单品在每一天的销售总量，形成单品日销售数据集。这两个聚合后的数据集将作为后续分析的核心。

#### 异常值识别与处理

考虑到蔬菜销售数据易受促销、节假日等因素影响而产生极端值，为避免这些异常值对模型训练造成不成比例的影响，我们对其进行了识别与处理。

我们采用箱型图（Box Plot）结合IQR（Interquartile Range）方法来识别异常值。该方法对数据分布没有特定要求，对于通常呈现偏态分布的销售数据而言，比传统的Z分数法更为稳健。我们选取了销量最高的品类“花叶类”及该品类下销量最高的单品“云南生菜”作为代表进行分析。如图X所示的箱型图清晰地展示了其销售数据的分布情况，其中远离箱体上下边缘的孤立点被识别为潜在异常值。



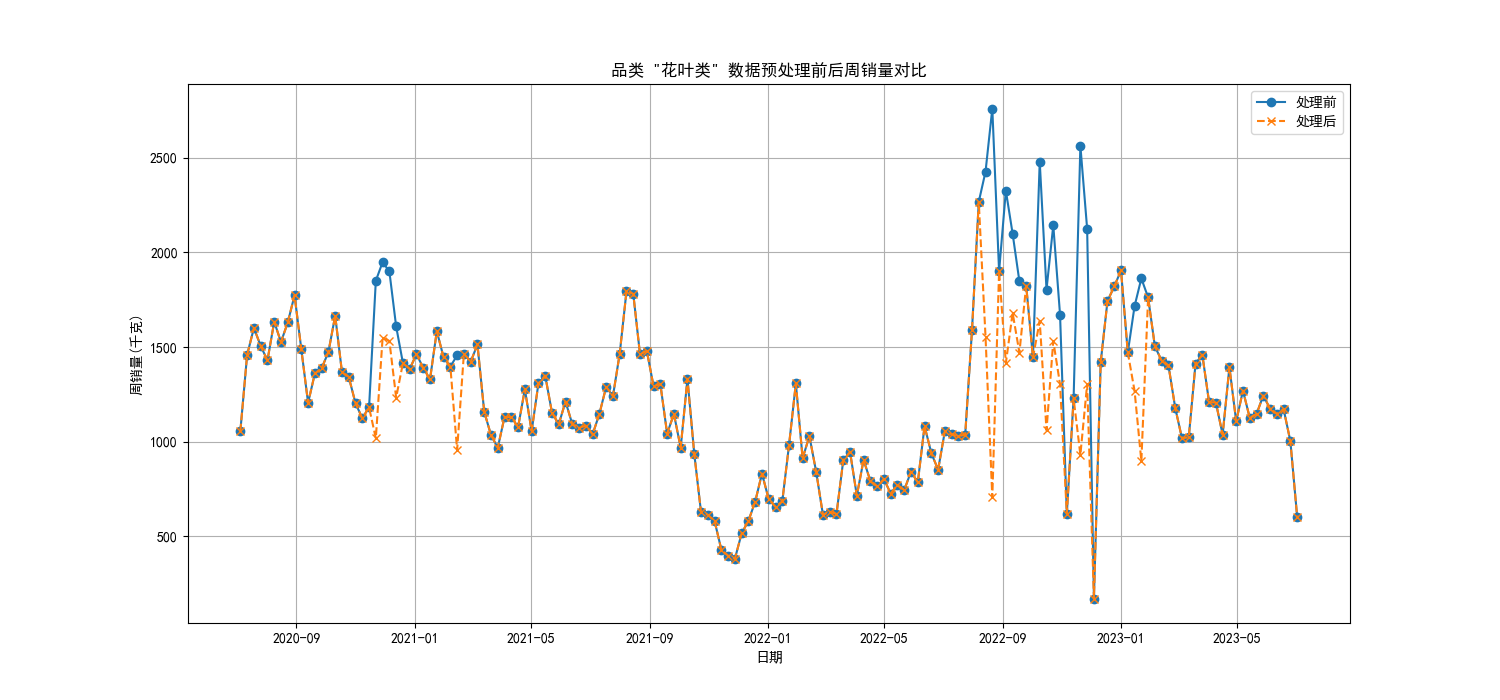
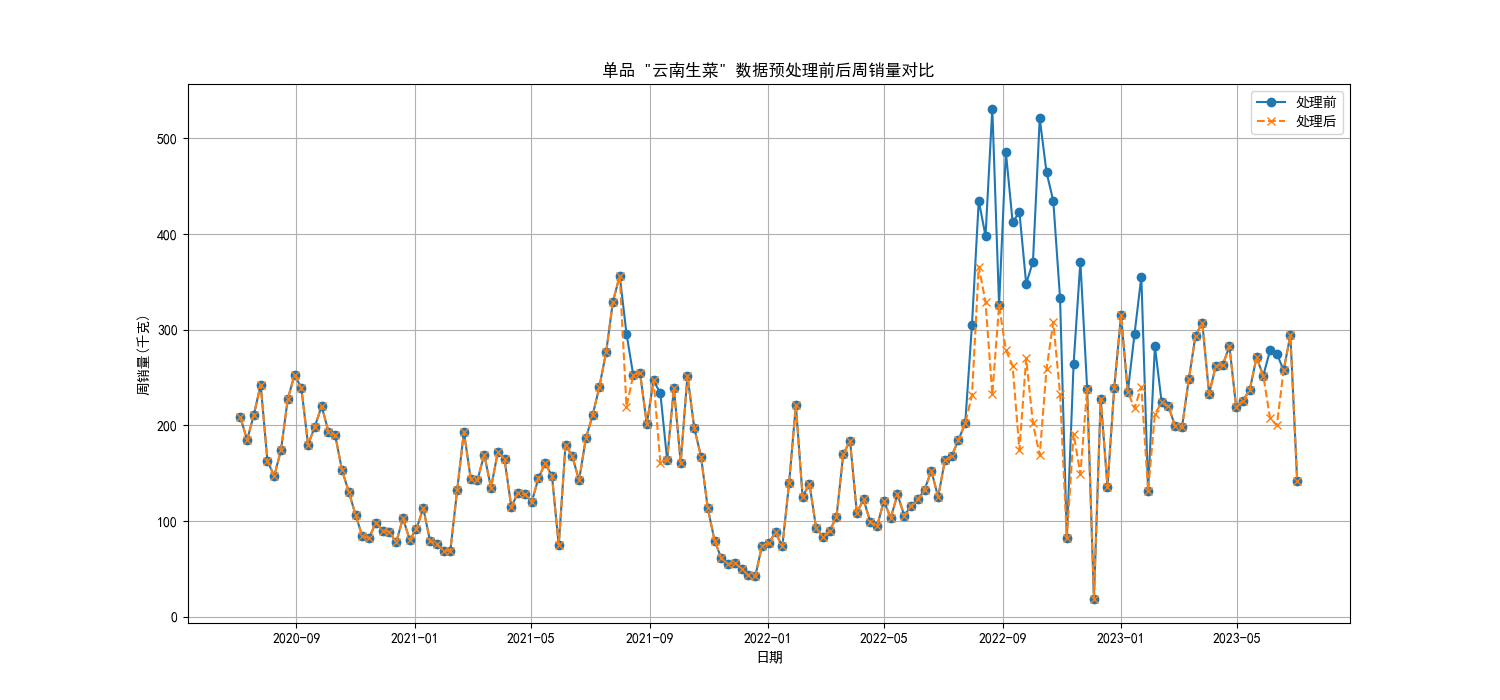
花叶类与云南生菜销量箱型图

基于IQR的计算结果，我们对数据集中所有超出 [Q1 - 1.5\*IQR, Q3 + 1.5\*IQR] 范围的数据点进行了剔除处理，以构建一个更稳定、更能反映常规销售模式的数据集用于后续分析。

#### 数据预处理效果验证

为直观展示异常值处理对数据平稳性的改善效果，我们绘制了“花叶类”和“云南生菜”在数据预处理前后的周销量时间序列对比图。如图Y所示，经过异常值剔除后，两条销量曲线的极端峰值得到了有效平滑，整体波动性降低，更能体现其内在的周期性与趋势性规律。

数据预处理前后周销量时间序列对比图



为深入探究蔬菜类商品的销售规律及内在关联，本研究对附件数据进行了系统的探索性数据分析（EDA）。我们首先在品类和单品两个层级上，对销售数据的时间序列特性和季节性模式进行了深入剖-析。

4.1 蔬菜品类销售规律分析

4.1.1 品类销售总体趋势与波动性

我们首先对所有六个蔬菜品类的日销售量进行了宏观审视。如图1所示的对比时间序列图显示，各品类的销售体量存在显著差异。“花叶类”蔬菜作为核心品类，其日销售量远高于其他品类，构成了该商超蔬菜销售的基本盘。其余品类销量依次为“辣椒类”、“食用菌”、“水生根茎类”、“花菜类”和“茄类”。

从时间维度上看，所有品类的销售均表现出明显的周期性波动和随机扰动，但未观察到持续性的增长或下降趋势，表明在过去三年内，蔬菜销售市场整体处于一个相对成熟和稳定的状态。其中，“花叶类”和“辣椒类”的销量波动尤为剧烈，时常出现销售尖峰，这可能与季节性消费、促销活动或市场供给变化密切相关。

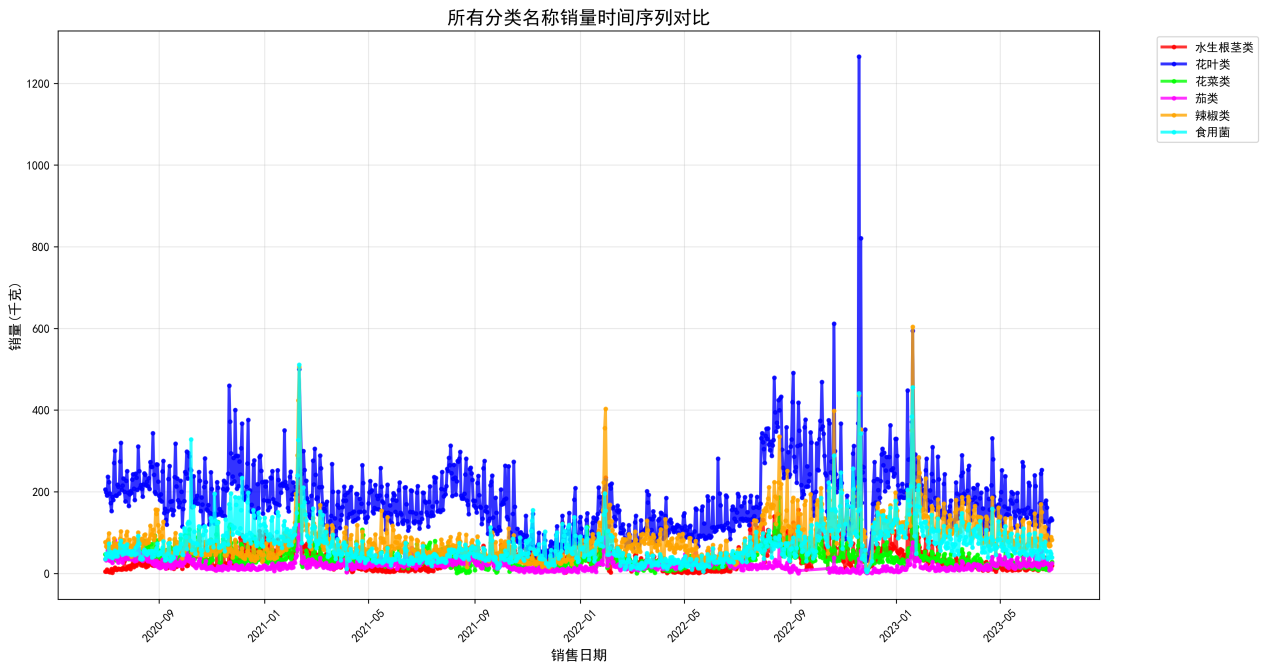


图1：所有分类名称销量时间序列对比

为更清晰地观察各品类的独立销售模式，我们绘制了个体时间序列图（如图2所示）。图中可见，“花叶类”的日均销量最高，达到了182.97千克，但其波动也最为显著。“辣椒类”和“食用菌”的销量同样表现出较强的波动性。相比之下，“茄类”和“花菜类”的日销量则相对较低且更为平稳。

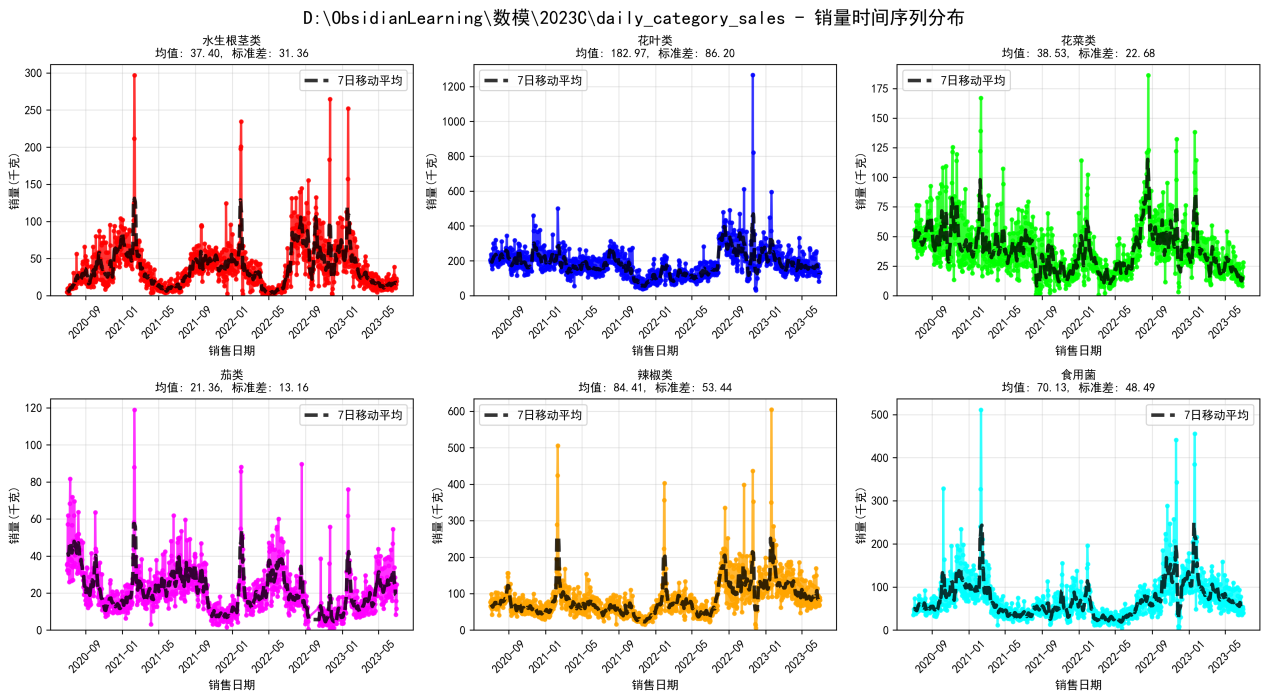


图2：各分类名称销量时间序列分布

4.1.2 品类销售的季节性模式

为探究销售波动背后的驱动因素，我们对各品类的月度和周度销售模式进行了分析。如图3所示，季节性规律清晰地浮现出来：

月度模式：存在显著的季节性消费偏好。例如，“水生根茎类”和“花菜类”的销售高峰集中在秋冬季节（10月至次年1月），而“辣椒类”则在夏季（6-8月）更受欢迎。“花叶类”作为基础品类，全年销量维持在较高水平，但在春季（3-5月）和秋季（9-11月）达到顶峰。

周度模式：所有品类均表现出高度一致的周内销售节律。销量从周一到周五呈现平稳或小幅增长，并在周五达到峰值，随后在周末（周六、周日）逐步回落。这一“周末效应”清晰地反映了消费者的购物习惯，即倾向于在周末前集中采购生鲜商品。

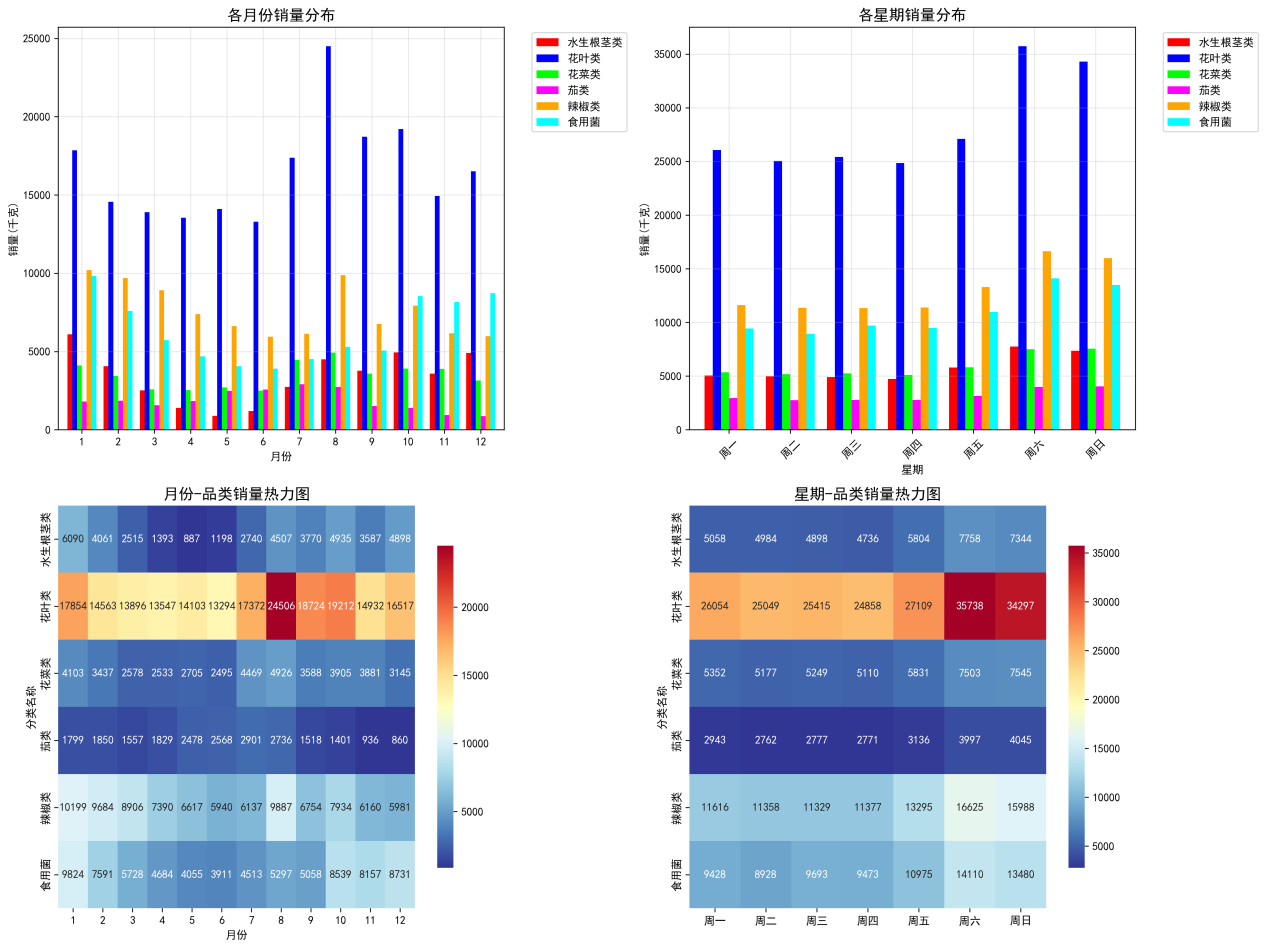


图3：品类销售季节性模式分析

4.2 代表性单品销售规律分析

在品类分析的基础上，我们进一步选取了各品类中累计销量排名前列的代表性单品进行深入分析，以探究微观层面的销售特性。

4.2.1 单品销售趋势与波动性

如图4所示的单品个体时间序列图显示，大部分核心单品的销售同样呈现出无明显长期趋势但具有强周期性的特点。例如，“云南生菜”、“白萝卜”和“黄瓜”等单品，其销量在三年间围绕一个稳定的均值上下波动。部分单品如“大白菜”的销售则表现出更强的季节性，在特定年份的秋冬季出现集中的销售高峰。

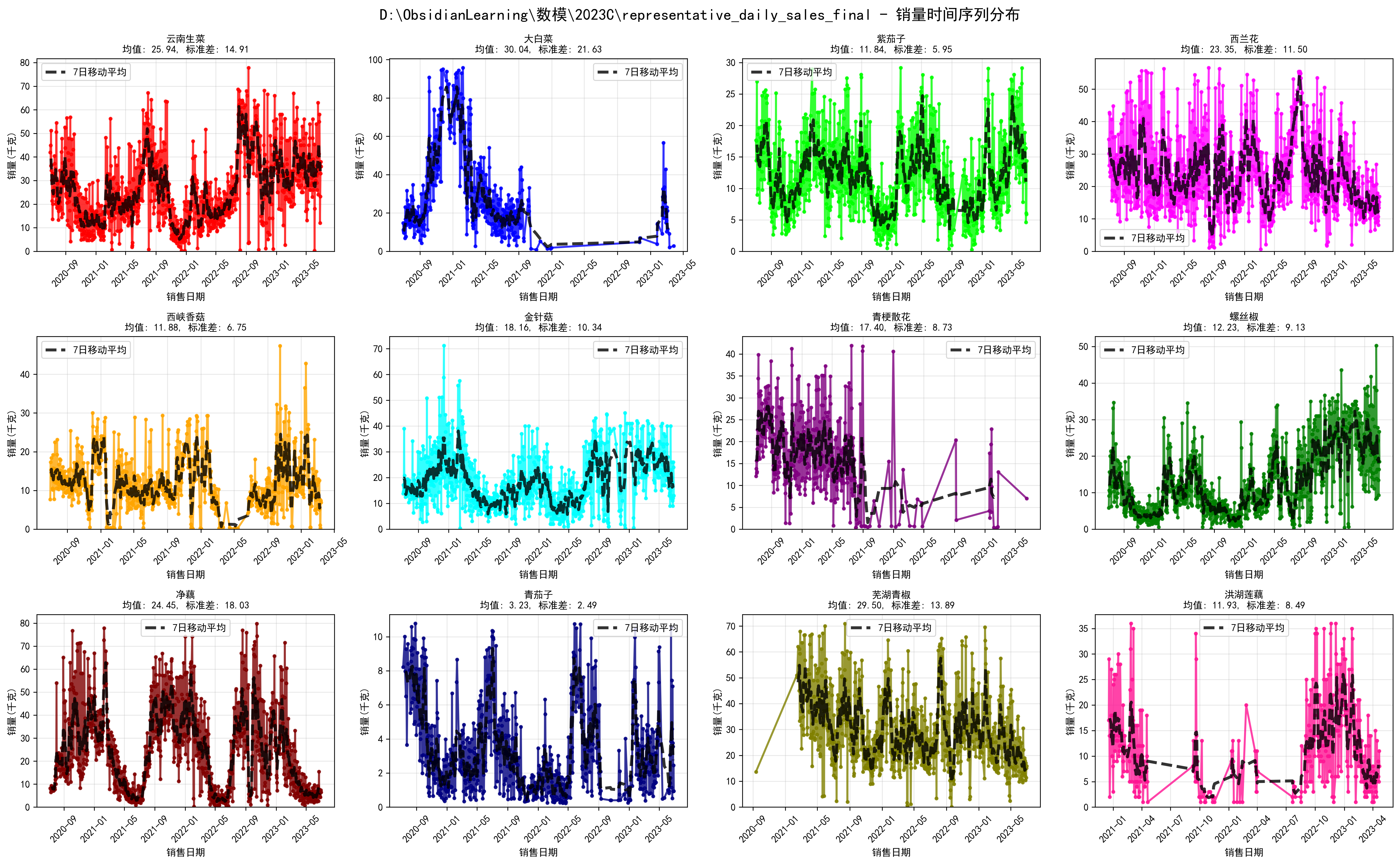


图4：代表性单品个体时间序列分布

4.2.2 单品销售的季节性模式

单品层面的季节性模式（如图5所示）为品类分析提供了更精细的注脚：

月度模式：单品的季节性特征更为鲜明。例如，“云南生菜”和“上海青”等绿叶蔬菜在春秋两季销量最高，而“白萝卜”的销售高峰则严格地出现在冬季的11月至1月，这与消费者的季节性饮食习惯高度吻合。

周度模式：与品类层面观察到的规律完全一致，所有代表性单品的销量均在周五达到一周内的最高点，再次验证了“周末采购高峰”这一核心消费行为模式。

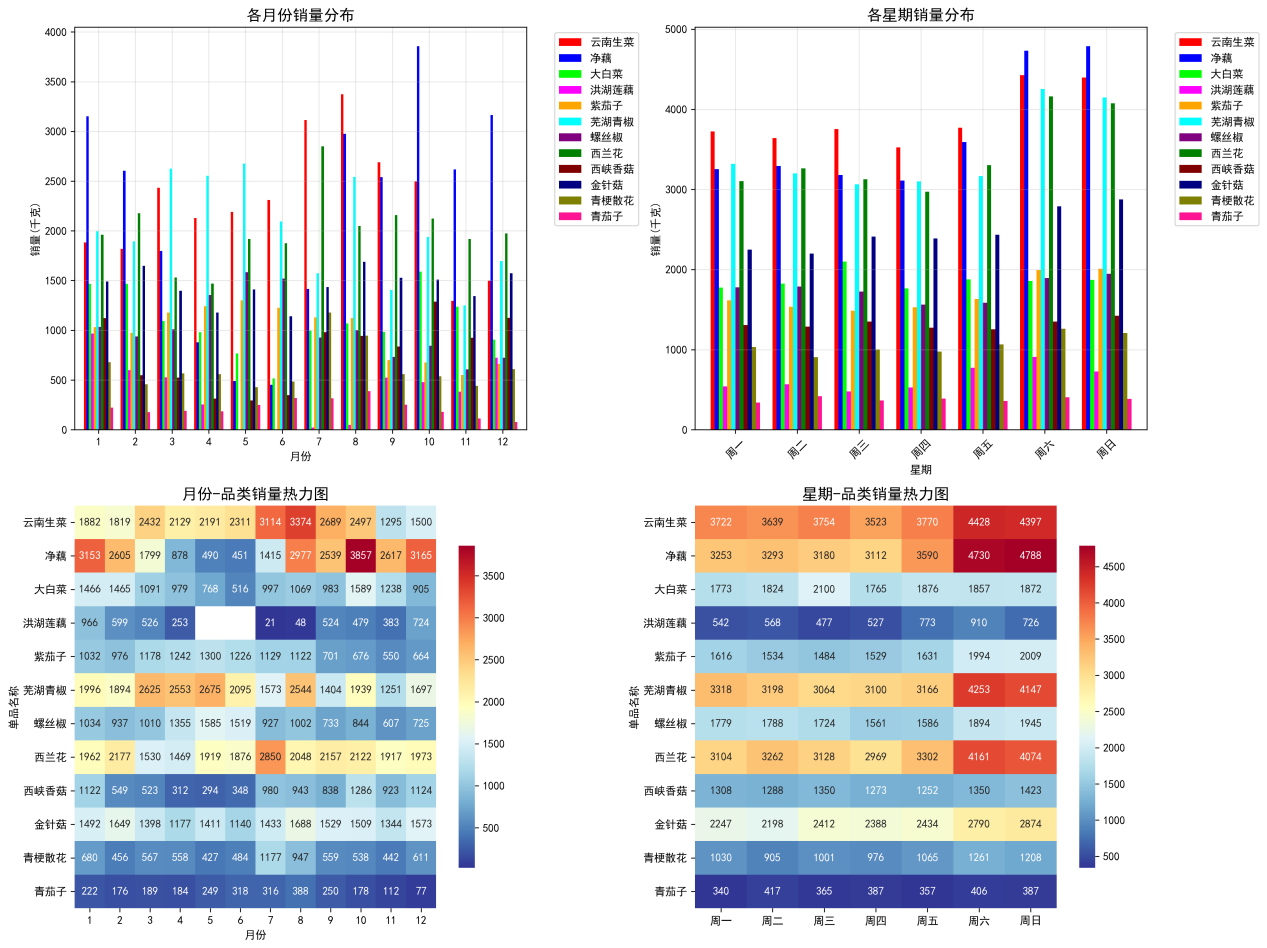


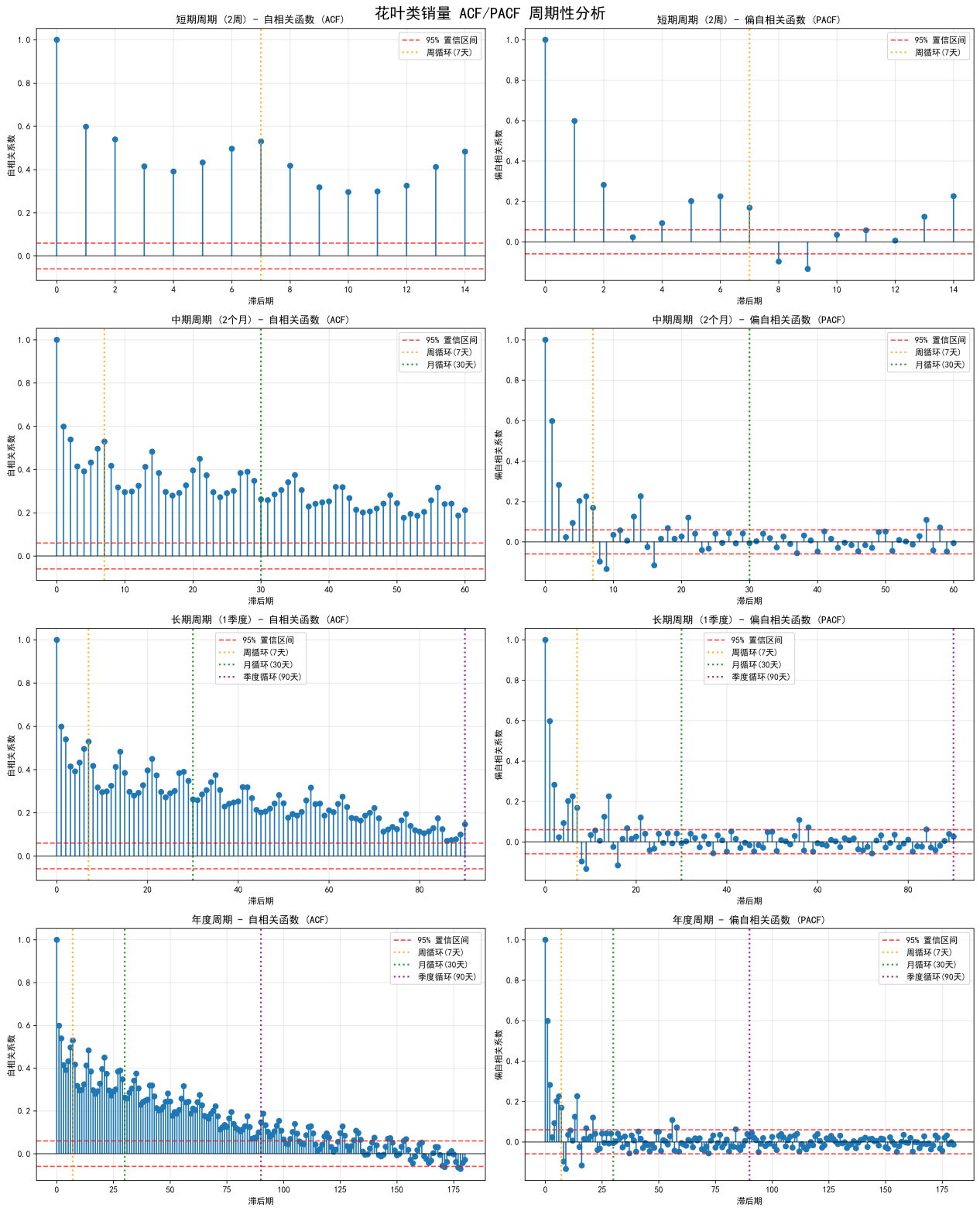
图5：单品销售季节性模式分析

#### 4.3 销售数据的时间自相关性分析 (定量验证)

为从统计学角度对前文观察到的周期性规律进行定量验证，我们选取最具代表性的“花叶类”品类与核心单品“云南生菜”的日销售量时间序列进行自相关性分析。

##### 4.3.1 品类层面的自相关性分析（以“花叶类”为例）

“花叶类”作为销量最高的品类，其时间序列（如图6所示）表现出极强的波动性。我们对其进行了自相关（ACF）与偏自相关（PACF）分析，结果如图7所示。

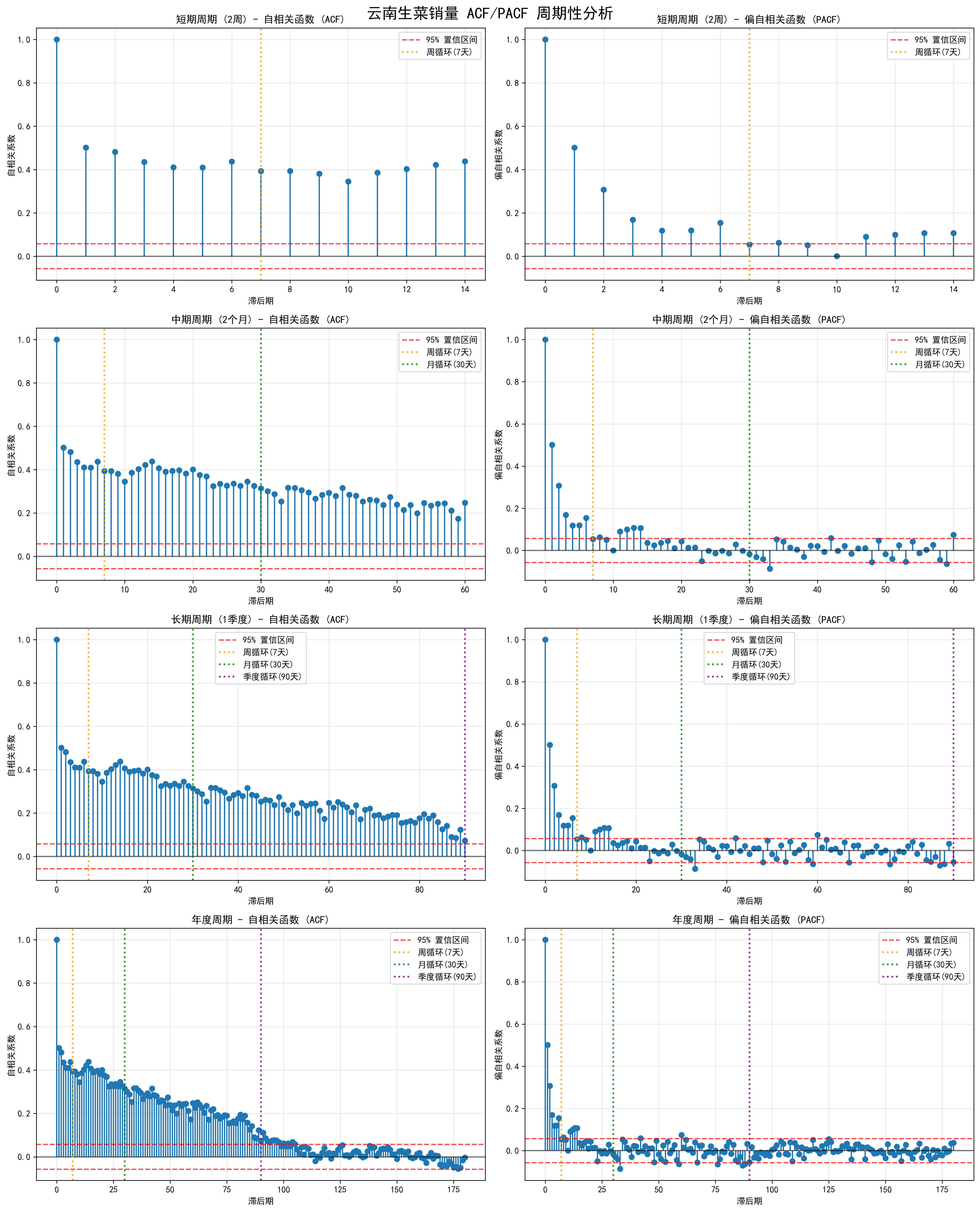


!图7：花叶类ACF/PACF综合分析图

1. 自相关函数（ACF）分析： ACF图呈现出显著的拖尾和周期性波动，表明序列具有强记忆性。数值分析结果显示，滞后7天（ACF=0.5294）、14天（ACF=0.4830）和30天（ACF=0.2620）的相关系数均远超95%置信区间（±0.0595）。这一发现为我们在前文通过可视化观察到的7日为周期的销售规律和月度消费模式提供了强有力的统计学证据。

2. 偏自相关函数（PACF）分析： PACF图同样揭示了显著的周期性，其在滞后7天处的相关系数（PACF=0.1690）显著为正，这表明即使在剔除了前6天的影响后，上周同日的销量对当日销量仍有直接的正面影响。此外，PACF在滞后1阶（PACF=0.5985）和2阶（PACF=0.2825）后迅速下降，表现出短期的截尾特征。

##### 4.3.2 单品层面的自相关性分析（以“云南生菜”为例）

为验证该规律在单品层面同样适用，我们对核心单品“云南生菜”进行了相同的分析。ACF图清晰地展示了其日销量序列的内部关联结构，呈现出与“花叶类”相似的缓慢衰减的拖尾形态。数值分析结果显示，滞后7天（ACF=0.3935）和14天（ACF=0.4377）的相关系数均显著为正，再次印证了周度周期的存在。其PACF图也呈现出清晰的截尾特征，在滞后2阶后迅速落入置信区间内，表明其同样具有典型的自回归（AR）过程特征。

综上所述，通过对品类和代表性单品的双重分析，并结合自相关性检验，我们清晰地刻画出了该商超蔬菜销售的核心规律：整体市场稳定，无长期趋势，但存在由消费者购物习惯驱动的、有统计学显著性的强周度周期（周五高峰），以及由气候和饮食文化驱动的显著季节性周期。同时，数据表现出的强自回归特性，为后续建立ARIMA等时间序列预测模型奠定了坚实的基础。