基于预测模型的自动定价与补货决策问题

摘要

针对问题 1：先对数据进行预处理，过滤掉数据的异常值，对数据各组分类求和。对于分布规律，先对各品类进行**描述性统计分析**，画出了饼状图分析销售总量，用折线图等对蔬菜销量进一步探究分布规律，发现其销量存在**季节性分布**和**销售时间段效应**。再对各单品计算统计性指标，对其进行**均值分析**和**稳定性分析**探索其分布规律；对于相互关系，各品类和单品都采用 Spearman **相关系数**画出相关系数热力图，直观表示相互关系。

针对问题 2：根据成本加成定价理论，将题目中所求关系转换成求销售总量与进货成本和销售价格的关系。做出两两之间的散点图初步判断是否线性，对于线性关系，采用**线性回归**；对于非线性关系，采用**多项式回归**，用回归模型给出题中所求关系。求解收益最大化问题，建立 ARIMA **模型**对总销售量和进货成本做一个预测，进而利用成本加成定价理论中的相关公式解得最大收益。

针对问题 3：根据问题三给出的一系列限制约束条件制定单品的补货计划，先根据 2023 年 6 月 24-30 日的可售品种进行数据预处理，再使用随机森林预测销售量和进货成本，建立非线性规划模型，根据题目要求和数据特点给定一系列约束条件，将函数和约束导入 python 进行求解，得出利润最大的定价策略。

针对问题 4：结合前三个问题所不可或缺的数据以及实际生活经验，提出补充气候季节、库存、市场行业、地域风俗等数据，以便商超更好地解决定价与补货决策问题。

关键词：Spearman 相关系数；回归模型；ARIMA 模型；随机森林预测模型

一、问题重述

1.1 问题背景

在生鲜商超中，蔬菜类商品的保鲜期通常都比较短，品相会随销售时间的增加而变差，同时有很大一部分蔬菜只能当天售卖。因此，商超会根据商品的历史销售和需求情况每天补货。

由于生鲜商超销售的蔬菜品种多样、产地各异，在不确切了解具体单品和进货价格的情况下，商家需要做出每日蔬菜品类的补货决策。蔬菜的定价通常采用“成本加成定价”的方法，并对运损和品质变差的商品进行打折销售。对于需求侧，蔬菜销售量通常与时间相关；对于供给侧，蔬菜的供应品种在 4 月至 10 月较为丰富。因此，商超在制定补货决策和定价决策时，既需要可靠的市场需求分析，同时由于商超销售空间有限，也需考虑合理的销售组合。

1.2 问题提出

针对问题 1：蔬菜类商品不同品类之间或者不同单品之间也许存在一定联系，分析蔬菜各品类和单品之间销售量的分布规律以及它们之间的相互关系。

针对问题 2：考虑商超以品类为单位制定补货计划，分析各蔬菜品类的销售总量与成本加成定价之间的关系。在此基础上，为了使商超收益最大化，给出蔬菜品类未来一周（2023 年 7 月 1 日至 7 日）的每日补货总量和定价策略。

针对问题 3：由于蔬菜类商品销售空间有限，商超希望优化单品的补货计划，要求可售单品总数在 27-33 个之间，且每个单品的订购量必须满足最小陈列量 2.5 千克的要求。根据 2023 年 6 月 24 日至 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，使得商超收益最大化。

针对问题 4：以便能够更好地制定出蔬菜商品的补货和定价决策，商超还需要收集哪些有关的数据，这些相关数据对解决以上问题是否有帮助，请给出意见和理由。

二、问题分析

2.1 问题 1 的分析

问题 1 要求分析蔬菜类商品销量的分布规律以及相互关系，本文对各品类和各单品分别进行分析。首先对数据进行异常值的处理，并按照需求整理出需要的表格和数据。对于各品类的蔬菜销售量进行描述性统计分析，从季节性、周期性等方面探究分布规律；对于单品销售量计算相关统计指标来探究分布规律。接着对各单品以及品类做相关性分析，画相关系数热力图直观表示相关关系。

2.2 问题 2 的分析

问题 2 第一个小问要求分析各蔬菜品类的总销量和成本加成定价的关系，根据成本加成定价理论可以把题目问题转为与进货成本和销售价格之间的关系，画出相关关系散点图，大致观测这两者的关系，并给出回归模型。根据问题二第二小问要求给出各蔬菜品类未来一周的日补货总量和定价策略采用时间序列模型来预测未来一周的销售量和进货成本，进行平稳性检验通过后再利用 ARIMA 模型求解，求解出结果后，优化加成率矩阵，使得商超收益最大。

2.3 问题 3 的分析

问题三相较于问题二多增加了几个约束的条件，然后需要给出 7 月 1 日的补货量与定价策略。首先进行数据预处理，不同于问题二，本问使用随机森林模型来预测销售量和进货成本，然后建立非线性规划模型，给定一系列约束条件，通过 python 进行求解。

2.4 问题 4 的分析

根据问题四需求，结合上面三个问题，通过多个维度方面来满足商超补货与定价决策的不足，分别采取了气候与季节、库存、市场行业、活动与节假日、地域风情、消费

者行为等六大方面的数据来解决以上问题。

三、模型假设

1. 完整性假设：所有给定数据都是准确无误、完整的，没有包含错误与遗漏。
2. 市场稳定假设：假设市场上不会出现恶意竞争。
3. 活动与节假日假设：销售可能因为特别节假日或者搞活动出现销量增加或减少的情况。
4. 销售量与定价假设：假设商超根据销售量来设定售卖价格。

四、符号说明

符号	意义
IQR	四分位距
r_s	等级相关系数
R^2	拟合优度
r	加成率
β_i	回归系数

五、模型的建立与求解

5.1 问题 1

5.1.1 数据预处理

Step1)附件 2 销售流水明细中，退货数据和相应买入数据存在信息的重叠，又由于退货商品这部分的数据对于数据量庞大的总体来说影响较微，所以为简化运算，本文把退货数据和相应买入数据删去得到一个新的表格。

Step2)在步骤 1 得到的新表中，**依据销量 2.5IQR** 对表格数据进行过滤，过滤掉了偏离总体的**严重异常值**，得到第一问预处理数据（见附录）。

Step3)分组分类求和，得到每天、每月、每季度的销售总量。

5.1.2 分布规律

1.蔬菜各品类销售量

首先将蔬菜各品类商品的销量数据整合成一张新的表格，将每个品类销售量加和计算百分比得到各品类蔬菜销售量饼图如下：

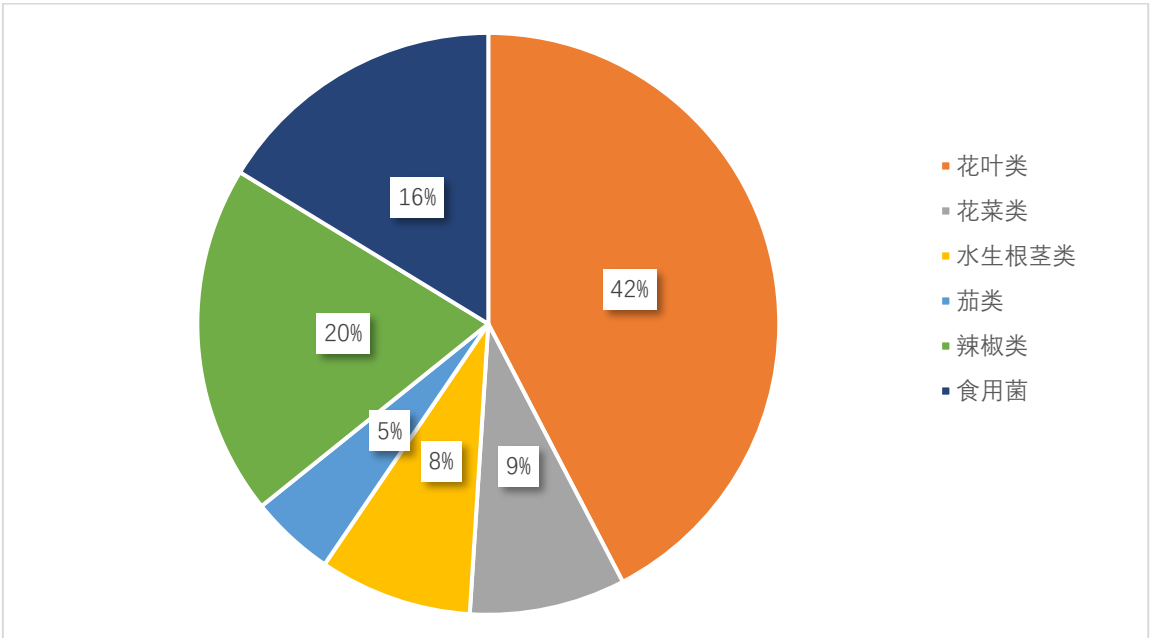

图 1：各品类蔬菜销售量饼图

由上图可知，在 6 个品类中，**花叶类蔬菜的销售总量最高**，其次是辣椒类。众所周知，花叶类蔬菜是日常生活中最常见的蔬菜品类，例如菠菜、白菜等都是家庭膳食中广泛可见的蔬菜品种，比较受人们喜爱；并且，花叶类蔬菜通常有多个丰收季节，这种季节性供应使得花叶类蔬菜在市场上具有更好的可获得性和较为稳定的供应量。因此对比其他品类，花叶类的蔬菜具有更高的销售量。而辣椒类作为一种广泛使用的配菜销量也较高，尤其是在喜食辣椒的地区。

• 季节性分析

由于蔬菜的生长和产量**受季节变化的影响较大**，且人们的饮食习惯和消费行为也会受季节变化的影响，因此本文对 6 个品类的销售量**按季度划分**，分别为 Q1、Q2、Q3 和 Q4，得到折线图如下：

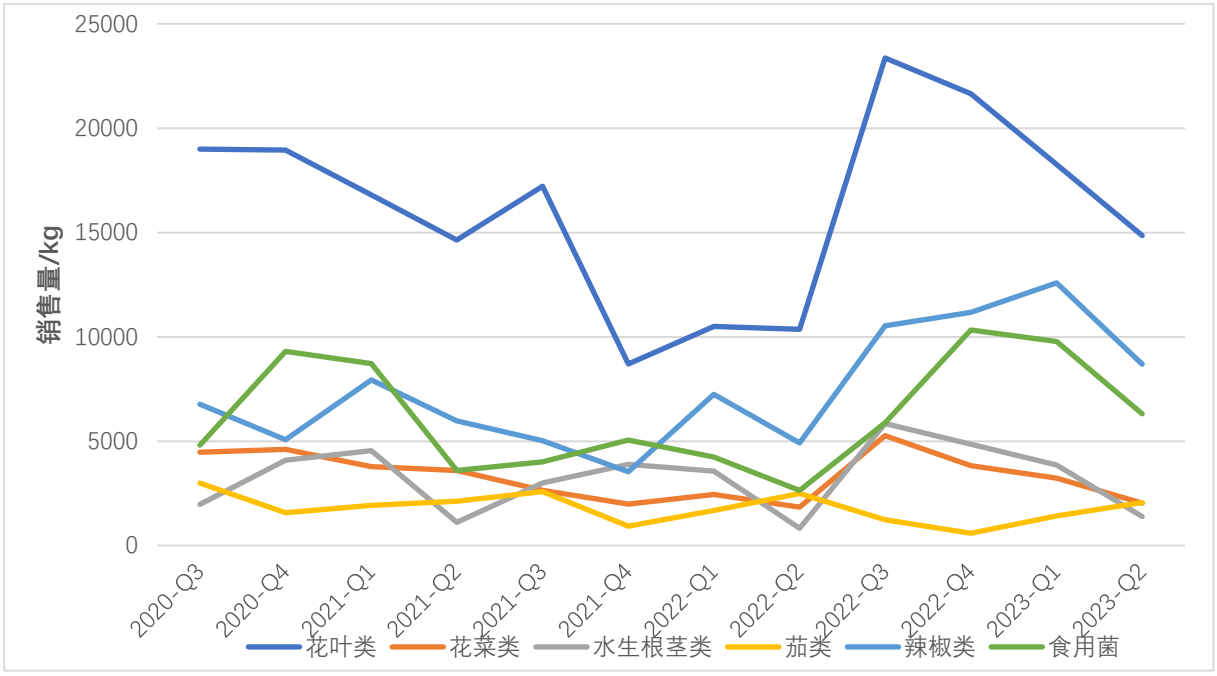

图 2：各品类蔬菜销售量随时间变化折线图

从上面折线图可以更加直观地观察出各品类销售量的变化，可以看出各品类销售量具有明显的季节性。

1）就**花叶类蔬菜**而言，**第三季度**往往是销售的**高峰期**，这是由于第三季度处于夏季和秋季的过渡时期，此时正是许多花叶类蔬菜生长和收获的季节。完整的生长周期和适宜的气温条件使得产量较大，因此在第三季度花叶类蔬菜供应充足，销售量相对较高。而到了第四季度和第一季度，花叶类的销售量则较为低迷，这是由于第四和第一季度处在冬季和春季，这个时期气温较低、天气寒冷，这样的气候条件不适于蔬菜类农作物的生长和生产，于是销量偏低。与花叶类蔬菜类似的还有花菜类、茄类。

2）就**食用菌**而言，第四季度为它的一个峰值时间段，食用菌大多数生长在气候冷凉、湿度较高的环境中，适合于在秋季和冬季生长和收获。在这个季节，食用菌的产量会相对较高，因此供应充足，销售量也相对较高。

3）就**辣椒类和水生根茎类蔬菜**而言，销售量在第一季度达到最高点，这是由于辣椒类和水生根茎类蔬菜多数生长在温暖季节，而第一季度通常是春季或初夏，气温适宜，这类蔬菜适宜生长在冬无严寒、夏无酷暑的环境。同时，第一季度也是农产品供应季节的开始，辣椒类和水生根茎类蔬菜的生长周期可能在这个时候达到成熟，供应量相对较大，从而提供了足够的货源来满足市场需求。所以此时的销售量较高。

• **销售时间段效应**

由于蔬菜的保鲜期较短，且只能当日售卖，所以对一天内各个时间段的销量进行分析能帮助了解蔬菜销量的分布规律。

于是根据一天早午晚三个时间段对各品类的销售量进行统计：

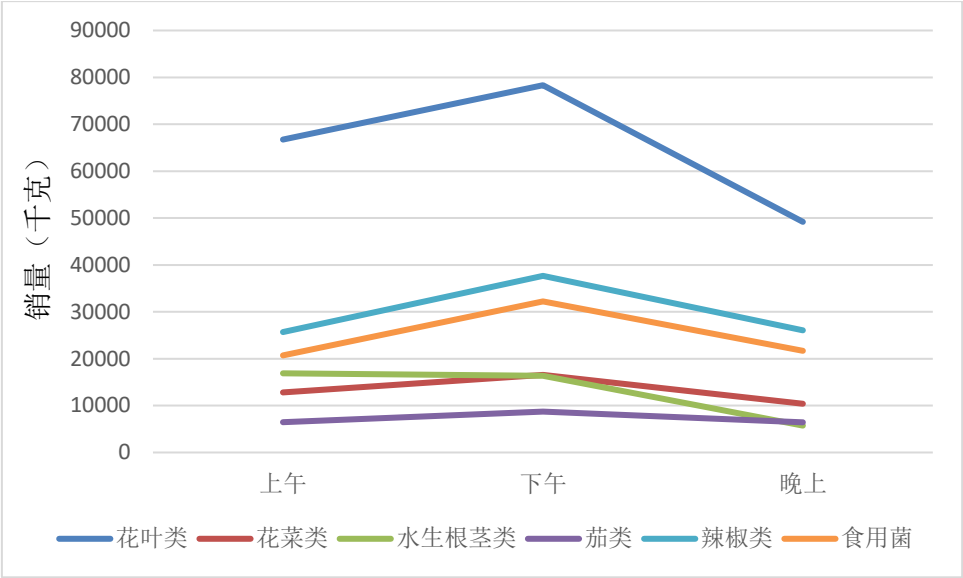

图 3：各品类一天内销量折线图

根据一天三个时间段对各品类销量进行分析，发现各品类蔬菜基本都呈现“中间高，两边低”的趋势，也就是说各蔬菜品类在下午时销量高，晚上销量较低，可能是因为许多人在下午或傍晚才有时间去购物，所以下午人流量较多，购买力也相对更高，而晚上蔬菜不太新鲜，人们购买意愿降低。

其中受时间段影响最大的是花叶类，这是因为花叶类蔬菜主要部分通常是叶片，而叶片的鲜嫩度和质量会随时间流逝而变化。随着时间的推移，叶片可能会逐渐失去水分、变黄或变软，影响了蔬菜的品质。

2. 蔬菜各单品销售量

对每个单品的销售量求和并进行排序，得到簇状条形图如下：

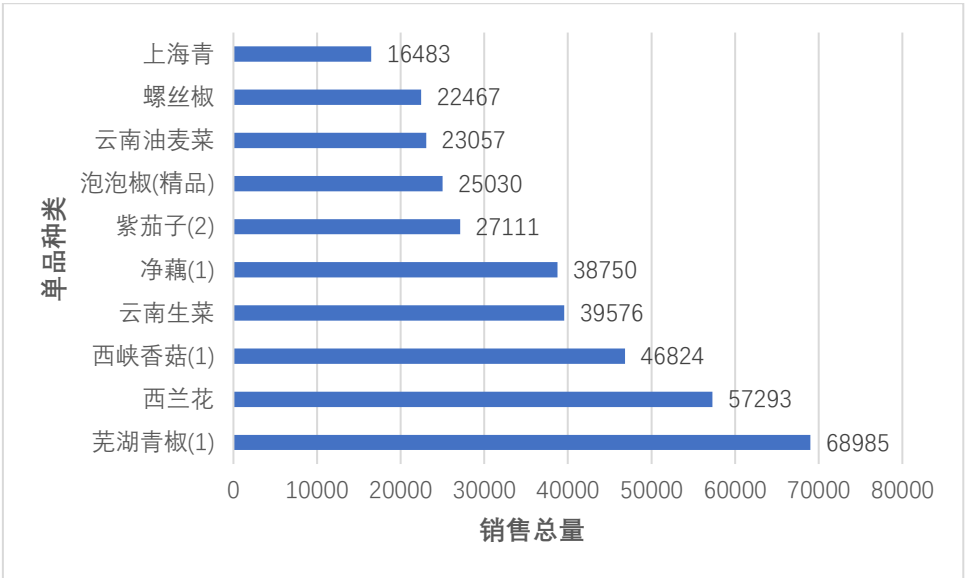

图 4：蔬菜单品销售量 TOP10

上图仅选取了前 10 个单品的销量排名，每个单品详细销售量见附录表格。

由上图可知，按单品的销售量来看，芜湖青椒（1）在各单品销售量中排名第一，其

次是西兰花、西峡香菇（1）。这表明居民在购买蔬菜类商品时对芜湖青椒（1）的需求非常大，分析这种数据分布可以帮助商超更好地了解消费者的需求，调整产品组合和供应链管理，以更好地满足消费者的需求同时提高商超的销售额。

利用 Python 求得各单品蔬菜销量的相关统计指标进行分析，进一步查看其分布规律，选取部分数据如下表所示：

表 1：部分单品销量统计指标表

单品编码	单品名称	总量	均值	最小值	最大值	25%	75%
102900011001561	莲蓬(个)	492	4.20122	1	10	3	5
102900005115960	大白菜	15066	1.228864	0.051	3.182	0.859	1.531
102900005115823	上海青	16483	0.44873	0.04	1.159	0.314	0.553
102900005116714	西兰花	57293	0.448284	0.022	0.995	0.339	0.513
102900011006948	外地茼蒿	1538	0.444099	0.156	0.864	0.372	0.508
102900005115984	云南油麦菜	23057	0.438134	0.067	1.032	0.326	0.523
102900011008164	奶白菜	13300	0.43183	0.025	1.181	0.291	0.547
102900011016701	芜湖青椒(1)	68985	0.39207	0.017	1.055	0.265	0.485
102900005116899	净藕(1)	38750	0.668279	0.057	2.023	0.407	0.854
102900005116530	西峡香菇(1)	46824	0.241992	0.015	0.668	0.161	0.301

1）均值分析：由表格可以看出，莲蓬和大白菜是各单品中均值较高的蔬菜类商品，忽略小部分量纲差异，可以说明莲蓬和大白菜具有较高的市场需求。在同一品类或量纲中考虑，如果有单品均值过于低，此时商超需要根据实际情况决定是否需要调整补货策略。

2）稳定性分析：根据最大值和最小值，可以了解单品的一个销售范围，范围的大与小表明销量的差异和稳定程度。同时，还可以通过第一四分位数和第三四分位数的接近程度来判断销量的稳定性。从上表中选取的部分数据来看，西峡香菇（1）和外地茼蒿的销量都比较稳定，对此类商品，商超可以考虑稳定的进货方式；反之则可以根据商品特点采用灵活的进货方式。

5.1.3 相互关系

1.蔬菜各品类销售量

接下来，采用相关性分析来探究蔬菜不同品类之间的相互关系。

1）正态性检验

首先将整理后的数据导入 SPSS，进行正态分布检验，导出表格如下：

表 2：6 品类蔬菜销量数据的正态分布检验表

正态性检验						
	Kolmogorov-Smirnov			Shapiro－Wilk		
	统计	自由度	显著性	统计	自由度	显著性
花叶类	.076	1041	.000	.842	1041	.000
花菜类	.090	1041	.000	.916	1041	.000
水生根茎类	.125	1041	.000	.817	1041	.000
茄类	.094	1041	.000	.900	1041	.000
辣椒类	.137	1041	.000	.778	1041	.000
食用菌	.116	1041	.000	.789	1041	.000

a. 里利氏显著性修正

若样本量小于 50，则根据 Shapiro－Wilk 检验的结果进行判定；若数据量非常大，

则根据 Kolmogorov-Smirnov（简称 K-S）检验的结果进行判定。本问中数据为大样本，于是以 K-S 检验为依据。

由上表，显著性（P 值）小于 0.05，即拒绝原假设，样本不满足正态分布，所以这里采用 Spearman 相关性分析，而不能用 Pearson 相关性分析。

2）Spearman 相关系数

两组随机变量 $X=(x_1,x_2,\cdots,x_n)$ 和 $Y=(y_1,y_2,\cdots,y_n)$ 之间的 Spearman（等级）相关系数可用如下公式计算得到：

$$r_s=\frac{\sum_{i=1}^{n}(p_i-\bar{p})(q_i-\bar{q})}{\sqrt{\sum_{i=1}^{n}(p_i-\bar{q})^2}\sqrt{\sum_{i=1}^{n}(q_i-\bar{q})^2}}$$

其中，r_s 为等级相关系数，p_i 和 q_i 分别为 x_i 和 y_i 的排名，若变量中有数值相等，该数值对应的排名为这几个值对应排名的平均值。若变量间的等级差值可以计算，则上述公式可简化为：

$$r_s=1-\frac{6\sum_{i=1}^{n}d_i^2}{n(n^2-1)}$$

上述公式中，r_s 为等级相关系数，d_i 为随机变量 X 、Y 中对应数据 x_i 和 y_i 的排名的差值，同为升序或者降序，n 为样本容量。若 $0<r_s\le 1$，则变量间呈正相关；若 $-1\le r_s<0$，则变量间呈负相关；当 $r_s=0$，则表示不相关。

3）相关性分析

计算出相关系数导出热力图如下：

	花叶类	花菜类	水生根茎类	茄类	辣椒类	食用菌
花叶类	1.000	0.645	0.453	0.320	0.612	0.607
花菜类	0.645	1.000	0.399	0.240	0.420	0.458
水生根茎类	0.453	0.399	1.000	-0.182	0.332	0.603
茄类	0.320	0.240	-0.182	1.000	0.151	-0.086
辣椒类	0.612	0.420	0.332	0.151	1.000	0.546

食用菌	0.607	0.458	0.603	-0.086	0.546	1.000

图 5：6 品类蔬菜相关系数热力图

·由上方热力图可知，花叶类销量与花菜类、辣椒类和食用菌的销量之间均有较高的正相关关系，食用菌销量与水生根茎类和辣椒类的销量之间均有较高的相关关系，说明当花菜类、辣椒类和食用菌的销量增加时，花叶类销量也可能增加；而例如辣椒类与茄类之间的相关系数为 0.151，则相关关系偏低，说明它们的销售模式可能是独立的。

·相关性强的品类之间可以考虑进行联合营销活动或合作供应链策略，以提高销售和效率；相关性较弱的品类可以进行独立的市场定位和推广活动，以增加曝光度和市场份额。

2.蔬菜各单品销售量

与前面分析蔬菜各品类的方法相同，仍然采用 Spearman 相关系数进行相关性的分析，来帮助我们探究哪些单品的销售趋势是相关的。

本文通过 Python 计算出了各个单品之间的相关系数，得到相关系数矩阵表（详见附录），并截取部分强相关系数表格如下（详见附录）：

表 3：蔬菜各单品强相关系数截取表

	单品 1	单品 2	相关系数
28055	红橡叶	绿牛油	1
27991	红珊瑚(粗叶)	绿牛油	1
27990	红珊瑚(粗叶)	红橡叶	1
22995	云南生菜(份)	云南油麦菜(份)	0.932007
22881	枝江红菜苔(份)	小青菜(2)	0.894907
6069	金针菇(1)	杏鲍菇(1)	0.892009
24892	小米椒(份)	小皱皮(份)	0.878716
6103	金针菇(1)	青梗散花	0.874198
26661	螺丝椒(份)	姜蒜小米椒组合装(小份)	0.864408
4037	大白菜	金针菇(1)	0.8574

·上表为蔬菜各单品强相关系数 TOP10 表格，从表中可以看到红橡叶、绿牛油和红珊瑚（粗叶）之间的相关系数为 1，这意味着它们之间存在完全的正相关关系。也就是说它们的变化总是同时发生，且变化的幅度和方向都完全一致。通过查阅资料，这也许是因为这几种蔬菜作为轻食沙拉的常见食材，所以消费者会同时购买。

·再以金针菇（1）和杏鲍菇（1）为例，金针菇和杏鲍菇都属于食用菌类别，且在口感和风味上相似，都有着白色丝状的菌肉和独特的香味。由于它们的相似性，消费者在购买时往往会同时选择这两种蘑菇品类，使得它们之间的销量出现较高的相关性。

此外，金针菇和杏鲍菇在市场上的定位和宣传也可能起到了一定的促进作用。商家可能会将它们放在一起销售、打造套餐或利用它们的共同特点进行联合营销，从而进一步增加了它们之间的销售相关性。就像在北方菜席上经常可以看到的凉菜“菌菇拼盘”，就通常同时含有金针菇和杏鲍菇这两种食用菌。

5.2 问题 2

5.2.1 数据预处理

Step1)将进货价格并入表格：以单品编码和销售日期为标度来合并数据，先固定单品编码，根据销售日期将附件 3 中的进货价格（批发价格）并入，选取一周内与当前表

格中销售日期最为接近的进货价格并入表格。

Step2)将损耗率和和分类名称并入表格：依据单品编码将对应损耗率和分类名称并入表格。

Step3)异常值处理：完成上述合并步骤后，将批发价格 20 倍以上的销售价格认作是异常数据，并将其剔除。

5.2.2 销售总量与成本加成定价关系分析

1.相关关系散点图

成本加成定价是假设销售者根据生产和销售成本所决定的销售价格，所以要求解销售总量与成本加成定价的关系就转换成销售总量与进货成本和销售价格这两个主要方面的关系。

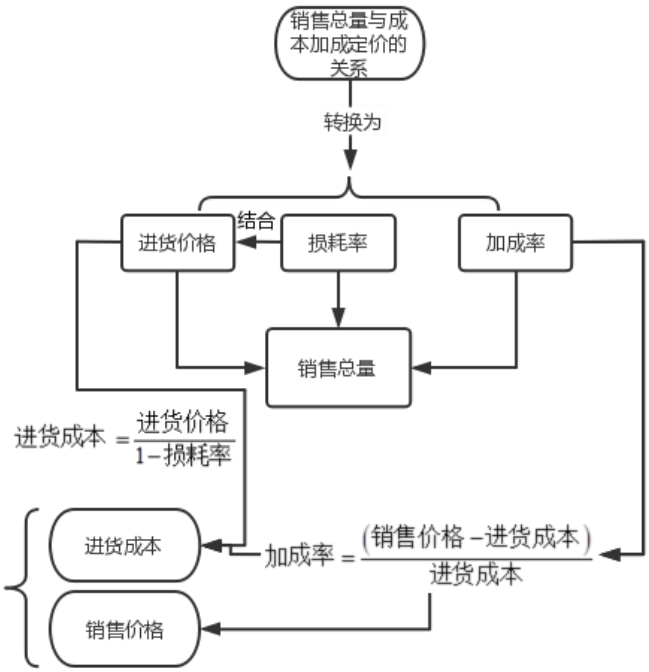

图 6：关系转换流程图

接着做出各品类销售总量与进货成本、销售价格的散点图：

图 7：各品类销售散点图

以花叶类和花菜类销量为例：花叶类销售总量集中分布在 0-400kg 之间，平均销售单价集中分布在 4-8 元/kg，进货成本集中分布在 2-6 元/kg；花菜类销售总量集中分布在 20-60kg 之间，平均销售单价集中分布在 5-15 元/kg 之间，进货成本集中分布在 4-10 元/kg。

2. 回归模型的建立

观察散点图能够得到 6 品类间变量之间的线性或非线性关系，对于线性关系采用线性回归模型（Linear Regression），而对于非线性关系采用多项式回归（Polynomial

Regression）。

• 线性回归模型

多元线性回归方法是通过多个自变量 x 来预测因变量 y 和探究 x 与 y 之间的关系，其中认为 y 与 x 之间存在如下函数关系：

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_i x_i + \varepsilon$$

式中，β_i 为多元回归系数，i 为自变量数量；ε 为残差项，通常认为其服从正态分布。

通过对问题的观测，收集一定数量的样本观测值，即可探究自变量和因变量之间的关系，通过求解回归系数 β_i 分析。

• 多项式回归模型

研究一个因变量和一个或多个自变量间的多项式的回归分析方法，称为多项式回归。本问中所探究的自变量是平均销售单价和进货成本，所以是二元的。从数据的散点图来观察，存在一个“弯”，所以考虑用二次多项式。

二元二次多项式方程为：

$$\hat{y} = a_0 + a_1 x + a_2 x + a_3 x_1^3 + a_4 x_2^{\ 3} + a_5 x_1 x_2$$

3.回归模型的求解

解出部分回归模型如下（其余详见附录）：

1）食用菌：$y = 69.2379 - 15.1867 * $平均销售单价$-1.9259 * $进货成本

--R^2 分数为：0.1249

2）水生根茎类：$y = 35.9101 - 8.7202 * $平均销售单价$-0.0603 * $进货成本

--R^2 分数为：0.1147

一般来说，如果样本量非常庞大，那么这个 R^2 就不会很大。且在该任务场景下没有其他的模型，本题中所选的这个模型在能提高相应任务的效率的同时结果也能被接受。所以根据实际情况，尽管 R^2 较小，但这个回归模型仍能使用。

以上述两个回归模型为例，食用菌的平均销售单价每增加一个单位，它的销售量就平均减少 15.1867 个单位，进货成本每增加一个单位，销售量就平均减少 1.9259 个单位，且平均销售单价对销量的影响程度更大。

水生根茎类蔬菜的平均销售单价每增加一个单位，它的销售量就平均减少 8.7202 个单位，进货成本每增加一个单位，销售量就平均减少 0.0603 个单位，且平均销售单价对销量的影响程度更大。

5.2.3 收益最大化问题的求解

1.模型的建立

第二小问要求给出各蔬菜品类未来一周的日补货总量和定价策略，也就是说对总销售量（每天售出的数量其实也是我们补货的总量）和进货成本做一个预测。本文采用差分整合移动平均自回归模型（ARIMA）预测出未来一周的销售量和进货成本。

• 平稳性检验

时间序列的平稳性是合理进行时间序列分析和预测的重要保证，因此在建模之前，待分析的时间序列必须满足平稳性条件，非平稳时间序列可通过差分法使之平稳化并进行平稳性检验。

平稳性检验结果如下：

表 4：平稳性检验结果输出表

模型统计						
		模型拟合度统计	杨-博克斯 Q(18)			
模型	预测变量数	平稳 R 方	统计	DF	显著性	离群值数
花叶类-模型_1	0	0.825	35.720	14	0.061	19
花菜类-模型_2	0	0.666	21.208	14	0.096	16
水生根茎类-模型_3	0	0.677	26.723	13	0.074	10
茄类-模型_4	0	0.796	18.106	15	0.257	19
辣椒类-模型_5	0	0.651	33.558	13	0.051	13
食用菌-模型_6	0	0.687	16.943	11	0.110	18

1）平稳 R 方越接近于 1，表明模型拟合越好，观察上表，平稳 R 方均在 0.6 以上，所以各个模型拟合效果都比较好。

2）对于白噪声测试，显著性 P 值大于 0.05，则说明时间序列平稳，表中模型均通过平稳性检验。

• ARIMA 算法原理

ARIMA 模型，即差分整合移动平均自回归模型，其分为自回归模型 (AR)、移动平均模型(MA)和自回归移动平均模型(ARMA)。

AR 模型表现为观测值 X_t 与其滞后 p 阶观测值的线性组合加上随机误差项，即:

$$\boldsymbol{X}_t = \varphi_1 \boldsymbol{X}_{t-1} + \varphi_2 \boldsymbol{X}_{t-2} + \cdots + \varphi_p \boldsymbol{X}_{t-p} + \alpha_t$$

上式中，X_t 为零均值平稳序列，α_t 为随机误差项，φ 为模型回归系数。AR 模型通常简记为 AR(p)。

MA 模型表示观测值 X_t 与先前 $t-1$、$t-2$、…、$t-q$ 个时刻进入系统的 q 个随机误差项 α_t、α_{t-1}、…、α_{t-q} 的线性组合，即：

$$\boldsymbol{X}_t = \alpha_t - \theta_1 \alpha_{t-1} - \theta_2 \alpha_{t-2} - \cdots - \theta_q \alpha_{t-q}$$

上式中，θ 为模型回归系数。MA 模型简记为 MA(q)。

ARMA 模型的观测值 X_t 不仅与前 p 个时刻的自身观测值相关，而且还与在其之前的时刻进入系统的 q 个随机误差有一定的依存关系，即:

$$\boldsymbol{X}_t = \varphi_1 \boldsymbol{X}_{t-1} + \varphi_2 \boldsymbol{X}_{t-2} + \cdots + \varphi_p \boldsymbol{X}_{t-p} + \alpha_t$$
$$-\theta_1 \alpha_{t-1} - \theta_2 \alpha_{t-2} - \cdots - \theta_q \alpha_{t-q}$$

由上式可知，ARMA(p，q)模型实际上就是 AR(p)模型与 MA（q）模型的一个组合。

时间序列在通过 d 阶差分平稳化后，先建立一个 ARMA 模型，让模型参数估计结果适应平稳化之前的数据。通过此过程建立的模型称为整合的 ARMA 模型，即 ARIMA(p，d，q)模型。

2.模型的求解

Step1)对模型求解后**预测**出未来一周的**销售量和进货成本**，预测值如下：

表 5：未来一周销售量预测值

花叶类	花菜类	水生根茎类	茄类	辣椒类	食用菌
183.2035	30.46915	25.426253	30.70296	114.8498	71.75271
186.1068	27.68025	22.13958	31.01206	111.375	67.66507

123.8933	16.18327	13.232093	18.93738	80.18253	48.34614
141.3536	15.40191	14.954264	16.33766	78.31667	52.30741
136.8572	16.62524	15.514376	16.92528	82.29286	57.19929
132.9592	18.06216	17.569701	16.76847	87.89206	53.30933
148.9687	19.81328	20.701671	19.90674	95.06323	60.9028

表 6：未来一周进货成本预测值

花叶类	花菜类	水生根茎类	茄类	辣椒类	食用菌
3.560209	8.722404	15.148968	5.331869	5.106393	8.222438
3.553467	8.720151	15.142693	5.316718	5.114137	7.894035
3.56648	8.718364	15.152396	5.291037	5.123034	7.894035
3.547412	8.719436	15.162105	5.306861	5.128069	7.894035
3.536644	8.719462	15.17182	5.307172	5.145876	7.894035
3.522389	8.719985	15.181542	5.278613	5.1462	7.894035
3.510805	8.719254	15.191269	5.299974	5.15345	7.894035

Step2)计算加成率

$$加成率=\frac{(销售价格-进货成本)}{进货成本}$$

优化得到加成率如下：

表 7：加成率矩阵

花叶类	花菜类	水生根茎类	茄类	辣椒类	食用菌类
0.5	0.4	0.4	0.45	0.52	0.45
0.5	0.4	0.4	0.45	0.52	0.45
0.4	0.35	0.35	0.5	0.48	0.43
0.4	0.35	0.35	0.5	0.48	0.43
0.4	0.35	0.35	0.5	0.48	0.43
0.4	0.35	0.35	0.5	0.48	0.43
0.4	0.35	0.4	0.5	0.52	0.45

Step3)计算定价

$$价格=单位成本\times(1+加成率)$$

得到定价矩阵如下：

表 8：定价矩阵表

花叶类	花菜类	水生根茎类	茄类	辣椒类	食用菌类
5.340314	12.21137	21.2085555	7.73121	7.761717	11.92253
5.330201	12.20821	21.1997699	7.709241	7.773488	11.44635
4.993072	11.76979	20.4557341	7.936556	7.58209	11.28847
4.966377	11.77124	20.4688413	7.960292	7.589543	11.28847
4.951302	11.77127	20.481957	7.960759	7.615897	11.28847
4.931344	11.77198	20.4950811	7.91792	7.616375	11.28847
4.915127	11.77099	21.267777	7.949961	7.833244	11.44635

Step4)解得最大收益

依据成本加成定价相关算法以及结合实际情况，求得**最大收益**为6317.493487050731元。

5.3 问题 3

5.3.1 数据预处理

根据题目条件，筛选出2023年6月24-30日的可售品种，根据上述数据选取一定品类进行下一步求解。根据第二问的合并的总数据数据表，根据单品编码和销售日期（天）进行分组聚合，销量列进行求和处理，商品单价和销售价格和损耗率这三列综合求出加成率，求解公式如式

$$进货成本 = \frac{进货价格}{1 - 损耗率}$$

$$加成率 = \frac{销售价格 - 进货成本}{进货成本}$$

5.3.2 使用随机森林预测销售量和进货成本

由于x_i与p_i之间的拟合优度(R^2)小于期望值，直接使用非线性规划模型进行商超收益的最大值求解可能不准确。因此采用首先使用随机森林算法根据 2023 年上半年的数据进行预测，将预测出来的销售量和销售价格进行存储，以便于最后的分析。

5.3.3 建立非线性规划模型

Step1)**确定**目标函数：根据题目所给条件，给出规划模型如下：

$$maxW_{总} = \sum_{i=1}^{49} c_i w_i x_i$$

其中，$W_{总}$表示商超的收益，c_i表示第i个单品当天是否出售，c_i的含义如下所示，w_i代表第i个单品的利润率，x_i代表第i个单品的当天销售总量。

$$c_i = \begin{cases} 0, 第\ i\ 个单品当天没有出售 \\ 1, 第\ i\ 个单品当天进行出售 \end{cases}$$

Step2)**确定约束条件**

综合题目中给出的各种约束，再结合样本数据自身的特点，给出下面的约束条件。

$$st \begin{cases} 0 < p_i < 300, \\ 2.5 < x_i < 500, \\ p_i = (1 + w_i)q_i, \\ x_i = k_i p_i + b_i, \\ q_i > 0 \\ w_i \in \mathbb{R} \\ w_i < 100 \\ c_i \in \{0,1\} \\ i = 1,2,3,\ldots,49, \\ 27 \le \sum_{i=1}^{49} C_i \ge 33. \end{cases}$$

其中，p_i代表第i个单品当天的销售价格，x_i代表第i个单品的当天销售总量，q_i代表第i个单品当天的进货成本，k_i，b_i分别代表回归方程的两个常数，i代表单品的索引。

Step3)**进行规划方程的求解**

将目标函数和约束导入到 python 中进行求解，使用 scipy 库求解出来的回归结果，与使用随机森林模型预测的结果进行综合比对，最后得出 7 月 1 日使得商超收益最大的单品补货量和定价策略。具体如下表所示：

表 9：规划方程预测值

单品名称	预测销量	加成率

苋菜	7.30464	0.207917
竹叶菜	11.88589	0.269591
菜心	2.28129	0.11819
木耳菜	5.5792	0.607921
紫茄子(2)	9.9566	0.50757
娃娃菜	9.41	0.383013
红薯尖	5.43379	0.385723
奶白菜	8.3912	0.364028
白玉菇(袋)	1	0.227594
芜湖青椒(1)	14.3292	0.456541
云南生菜(份)	33.68	0.296886
云南油麦菜(份)	24.51	0.361001
菠菜(份)	10.83	0.260275
鲜木耳(份)	3.26	0.814361
小米椒(份)	24.78	1.416605
虫草花(份)	1.94	0.255157
小皱皮(份)	9.72	0.573702
青线椒(份)	1.11	0.321746
螺丝椒(份)	12.22	0.531334
姜蒜小米椒组合装(小份)	7.61	0.721636
紫茄子(1)	1.515	0.87818
双孢菇(盒)	8.89	0.456199
青红杭椒组合装(份)	3.05	0.636127
蟹味菇与白玉菇双拼(盒)	1.91	0.154731
木耳菜(份)	1	1.10662
金针菇(盒)	17.68	0.326756
海鲜菇(包)	7.13	0.317153

5.4 问题 4

为了更好地制定蔬菜商品的补货和定价决策，商超可以考虑采集以下相关数据，以便更好地解决问题，数据以及相应的意见和理由如下：

1）气候和季节数据

对于问题 1 蔬菜销量问题的分析，各种蔬菜适宜的生长温度、湿度等气候条件以及不同蔬菜的生长生产季节都极大地影响着蔬菜的需求和供应，进一步影响着蔬菜的销量。所以，掌握了解不同蔬菜对应的气候和季节数据，能够帮助商超清楚蔬菜销量的分布规律以便预测销售和供应需求。

2）库存数据

掌握当前的实时库存剩余量，有利于商家合理灵活地进行补货与定价的决策，避免库存过度挤压或者长时间空缺。

3）市场行业数据

收集市场行业数据，例如同行业竞争对手的数据、市场上的产品数和行业行情的政策变化等数据，可以使商超更好更精准地制定定价和补货策略，并及时做出调整，避免市场机制的滞后性带来的损害。

4）促销活动和节假日数据

节假日和促销活动会刺激需求并影响商品销量，同时，商超可以根据节假日和促销活动的情况灵活调整库存和定价与补货，使商超能够实现利益的最大化。因此，了解和

掌握促销活动以及节假日的数据非常必要。

5）地域范围和风俗习惯数据

·地域范围：不同地区的条件、农业发展水平等因素会影响着当地蔬菜的产量和品种，也就影响着蔬菜的供应情况，了解它可以帮助商超更好地选择合适的采购渠道和供应商，确保供应的新鲜和稳定性。

·风俗习惯：不同文化和地域的人们有着不同的饮食习惯和偏好。了解当地消费者相关的偏好以及避讳，就能及时调整商品的种类和数量，更好地满足需求侧。

6）消费者行为数据

·了解消费者需求：消费者行为数据可以提供消费者对蔬菜类商品的需求和偏好。通过分析购买习惯、消费者反馈等数据，商超可以了解消费者对于不同种类蔬菜的偏好、购买频率以及重要的购买决策因素，帮助商超更准确地预测和满足消费者需求，以更好地促进销量。

·制定定价策略：消费者行为数据可以揭示消费者对于价格的敏感度和支付意愿。商超可以通过分析消费者购买决策中的价格因素，确定适当的定价策略。

六、模型的评价、改进与推广

1.描述性统计模型：

·优点：可分析的维度数据比较多，对统一数据从不同维度进行分析，可以分析多种特征。

·缺点：仅适用于连续变量，不能完全客观的反映问题，分析样本单一。

2.Spearman 相关性分析模型：

由于本文数据量较大，且 p 值小于 0.05，不满足正态分布，所以使用 Spearman 而不使用 Perason。

·优点：适用范围广，所选的实验数据不来自正态分布的总体也适用。

·缺点：精确度低。

3.回归模型

问题二求取各蔬菜品类的销售总量与成本加成定价的关系，采用线性回归与多项式回归模型来寻求两者关系。

·优点：既可以准确表明自变量与因变量的关系，也可以表现多个自变量对因变量的关系。

·缺点：算法较为低级。

4.时间序列模型

·优点：拟合效果好，对模型参数有动态确定能力。

·缺点：只是适合短期预测，不能反映内在关系。

5.随机森林预测：

·优点：随机森林能够处理高维数据和大量特征，并具有较高的预测准确性。它能够通过多个决策树的集成降低过拟合的风险，并且对于噪声和异常值具有一定的鲁棒性。

·缺点：虽然随机森林可以一定程度上降低过拟合的风险，但在某些情况下，仍然可能出现过度拟合的问题，特别是当训练样本不平衡或噪声较大时。

6.非线性规划：

·优点：非线性规划模型具有更强的模型表达能力和灵活性，可以处理复杂的实际问题，并可以产生更精确的结果。

·缺点：求解非线性规划问题的复杂性较高，可能存在多个局部最优解，并且对初始值较为敏感。

七、参考文献

[1]贠涛,张金倩楠,李姗姗等.基于斯皮尔曼系数和多层感知机的专家评审行为分析评法探究[J].科技通报,2022,38(05):107-112.DOI:10.13774/j.cnki.kjtb.2022.05.019.
[2]刘建明,冯吉昌.武威食用菌产业发展现状分析及对策建议[J].甘肃农业,2023(08):32-35.DOI:10.15979/j.cnki.cn62-1104/f.2023.08.013.
[3]韩俊华,干胜道.成本加成定价法评介[J].财会月刊,2012(22):74-75.DOI:10.19641/j.cnki.42-1290/f.2012.22.034.
[4]付业伟,张文侃,徐世濠等.基于多元线性回归分析方法的临水基坑防渗降水措施研究[J].水电与新能源,2023,37(08):13-17.DOI:10.13622/j.cnki.cn42-1800/tv.1671-3354.2023.08.004.
[5]高龙.基于 ARIMA 的小批量物料生产需求预测模型研究[J].现代信息科技,2023,7(15):97-101.DOI:10.19850/j.cnki.2096-4706.2023.15.021.
[6]卢亚杰. 我国超市优质生鲜蔬菜动态定价问题研究
[D].北京交通大学,2010.