学校食堂蔬菜价格季节性与长期趋势分析报告

摘要

在这份报告中，我旨在通过建立数学模型，分析我校食堂主要蔬菜价格的时间序列特性，重点是识别它们的季节性波动规律和长期变化趋势。我使用了从学校官网获取的 `vegetable_prices_1.csv` 文件，数据覆盖时间从 2021 年 5 月到 2025 年 3 月。在对数据进行清洗、取对数以及构建时间趋势和傅里叶季节项等特征后，我首先尝试了 OLS 回归模型。然而，分析显示 OLS 模型的残差存在强烈的自相关性，表明这个简单的模型不足以描述数据。为了解决这个问题，我转而采用了带有外生变量的 ARIMA 模型。通过结合 ADF 单位根检验、观察 ACF/PACF 图以及运行一个自动搜索程序（在 $p \le 2, d \le 1, q \le 2$ 范围内寻找 AIC 最小值），我为分析的 9 种蔬菜分别确定了 ARIMA(p,d,q) 阶数。模型验证结果令人鼓舞：对于其中的 7 种蔬菜（胡萝卜、青椒、西红柿、红椒、莴笋、小白菜、绿豆芽、土豆），最终选定的 ARIMA 模型成功消除了残差的自相关性（Ljung-Box 检验 $p > 0.05$），这为我们分析趋势和季节性提供了更可靠的基础。不过，为黄豆芽和生姜找到的模型仍然存在残差自相关问题，说明这两者需要进一步的模型调整。我的分析结果表明，在考虑了时间序列的动态特性后，大多数蔬菜的长期线性趋势并不显著，但季节性模式（年度和/或半年度）普遍存在，并且具体模式因蔬菜品种而异。一个普遍存在的问题是，即使自相关问题解决了，多数模型的残差仍然显著偏离正态分布（特别是呈现高峰度），这提示价格数据中可能包含不少异常波动或跳跃。

目录

1. 引言
    1.1. 目标
    1.2. 数据来源与初步观察
2. 数据准备工作
    2.1. 数据清洗步骤
    2.2. 数据变换与特征构建
3. 我的建模方法：时间序列分析
    3.1. 第一步：OLS 模型的尝试与反思
    3.2. 改进方案：采用 ARIMA 与外生变量
    3.3. 确定模型细节：阶数选择策略
4. 我如何验证模型
    4.1. 检查残差自相关 (Ljung-Box)
    4.2. 检查残差正态性 (Jarque-Bera)
    4.3. 检查异方差性
    4.4. 利用图形进行诊断

5. 建模结果与分析
   5.1. 最终模型选择汇总
   5.2. 各种蔬菜模型的详细结果解读
6. 讨论与思考
   6.1. 主要发现总结
   6.2. 模型的局限性反思
   6.3. 可能的优化方向
7. 结论

---

1. 引言

1.1. 目标

利用拿到的价格数据，建立一个合适的数学模型（主要是时间序列模型），来回答这几个问题：

- 各种蔬菜的价格是不是有明显的季节性？是每年一个周期还是半年一个周期？强度怎么样？
- 从长期来看，这些蔬菜的价格是倾向于上涨还是下跌？
- 我建的模型靠不靠谱？有哪些地方可能还需要改进？

1.2. 数据来源与初步观察

我用的数据来自学校官网公布的 `vegetable_prices_1.csv` 文件。这个文件记录了从 **2021年5月11日** 到 **2025年3月31日** 期间，多种蔬菜的每日价格。主要信息有蔬菜名 (`name`)、价格 (`price`) 和日期 (`date`)。（还有一个 `period` 列，仅用于推出date列。）

在处理完数据后（详情见后），我总共有 8 万多条有效的价格记录。为了分析得更细致，我挑选了其中记录比较多、时间跨度足够长（超过两年）的 10 种常见蔬菜作为重点分析对象。

2. 数据准备工作

为了让数据能用于时间序列建模，我做了下面这些准备：

2.1. 数据清洗步骤

1. **加载数据**: 用 Pandas 库把 CSV 数据读进来，中间顺便解决了中文编码可能遇到的问题。
2. **处理无效值**:
   - 把 `date` 列转成标准的日期格式，读不了的就当缺失值处理。
   - 把 `price` 列转成数字，转不了的也当缺失值。
   - 删掉了日期或价格缺失的行。
   - 确保所有价格都是正数，把小于等于 0 的记录去掉了。
3. **处理重复值**: 有些蔬菜同一天有好几个价格，我把它们取了平均值，保证每天每种蔬菜只有一个价格记录。
4. **排序**: 按蔬菜名和日期给数据排了个序。

2.2. 数据变换与特征构建

1. **对数变换**: 价格数据波动往往比较大，而且波动的幅度可能和价格本身有关。为了让数据更“稳定”，更容易满足模型假设，我决定对价格 `price` 取自然对数：

$$\log\_price_t = \ln(price_t)$$

   后面的分析都是基于这个 `log_price`。
2. **时间趋势变量**: 为了看价格有没有长期线性变化的趋势，我创建了一个变量 `time_numeric` (记作 `t`)，就是从数据开始那天算起，过了多少天：

$$t = (date_t - start\_date).days$$

3. **季节性傅里叶项**: 蔬菜价格很可能有季节性，我用了傅里叶项来捕捉这种固定的周期。假设主要的周期是每年 (P ≈ 365.25 天)，我用了前两阶谐波（能同时捕捉年度和半年度周期）：
   - **年度周期 (k=1)**:

$$\cos 1_t = \cos(2\pi t/P)$$

$$\sin 1_t = \sin(2\pi t/P)$$

   - **半年度周期 (k=2)**:

$$\cos 2_t = \cos(4\pi t/P)$$

$$\sin 2_t = \sin(4\pi t/P)$$

   这些 cos 和 sin 项会作为模型的解释变量（外生变量）。
4. **时间索引设置**: 在用 ARIMA 模型分析时，我把日期设成了数据的索引，这对时间序列模型很重要。

3. 我的建模方法：时间序列分析

考虑到价格数据前后是有关联的，我主要用了时间序列建模的方法。

3.1. 第一步：OLS 模型的尝试与反思

我最开始想得比较简单，就用普通的 OLS 回归试了一下，模型大概是这样：

$$\log\_price_t = \beta_0 + \beta_1 t + \beta_2 \cos 1_t + \beta_3 \sin 1_t + \beta_4 \cos 2_t + \beta_5 \sin 2_t + \epsilon_t$$

这里的 `ε_t` 是误差。这个模型能大致看看有没有长期趋势 (`β_1`) 和固定的季节性 (其他 `β`)。但跑完发现，几乎所有蔬菜模型的**残差 (`ε_t`) 都存在非常严重的自相关**（Durbin-Watson 值很低，Ljung-Box 检验 p 值接近 0）。这意味着模型忽略了数据自身的时间依赖性，比如今天的价格受昨天价格影响的特性，这使得 OLS 的结果（尤其是 p 值）变得不可靠。

3.2. 改进方案：采用 ARIMA 与外生变量

为了解决自相关问题，我决定采用更适合时间序列的 **ARIMA 模型，并把趋势项和季节项作为外生变量 (Exogenous Variables)** 加进去。这种模型通常也叫 REGARIMA 或 ARIMAX。

它的思路是把 `log_price` (记作 `y_t`) 分成两部分：一部分是由外生变量 `X_t` (就是常数、时间趋势 `t` 和那四个 cos/sin 项）解释的，另一部分是误差项 `η_t`，但这个误差项本身不是随机的，而是遵循一个 ARIMA(p, d, q) 过程：

$$y_t = X_t'\beta + \eta_t$$

这里的误差 `η_t` 的动态结构由 ARIMA(p, d, q) 描述，用滞后算子 `B` (即 `B^k z_t = z_{t-k}`) 可以写成：

$$\Phi(B)(1-B)^d\eta_t = \Theta(B)\alpha_t$$

简单解释一下：

- `p`, `d`, `q` 分别是模型的 AR（自回归）、I（差分）、MA（移动平均）阶数。
- `(1-B)^d` 表示对 `η_t` 做 `d` 阶差分，目的是让序列变平稳。
- `Φ(B)` 代表 AR 部分，表示当前误差受过去 `p` 期误差的影响。
- `Θ(B)` 代表 MA 部分，表示当前误差受过去 `q` 期随机冲击 (`α_t`) 的影响。
- `α_t` 是我们最终希望得到的**白噪声**残差，也就是随机的、没有自相关的部分。

这样一来，模型就能同时估计趋势/季节性的影响 (β)，并把误差中的时间依赖性吸收掉（通过 `ϕ` 和 `θ` 参数）。

3.3. 确定模型细节：阶数选择策略

给每种蔬菜找到合适的 (p, d, q) 阶数是关键。我是这么做的：

1. **确定 `d` (差分阶数)**: 我先用 **ADF 单位根检验**检验每种蔬菜的 `log_price` 序列是否平稳。如果检验 p 值很大（>0.05），说明可能不平稳，需要差分，也许 d=1；如果 p 值很小（<=0.05），说明可能平稳，d=0 或许就行。
2. **初步判断 `p`, `q` (AR/MA 阶数)**: 我画了原始（或差分后）序列的 **ACF 和 PACF 图**（保存在 `img_diag` 目录了），看看自相关和偏自相关系数是怎么衰减或者在第几阶后"断掉"（截尾）的，这能给我一些关于 p 和 q 可能取值的线索。
3. **自动搜索 `p`, `q`, `d`**: 手动定阶比较麻烦也容易出错，所以我写了个程序来自动尝试不同的组合。它会在一个预设范围内（例如我采取的： `p` 最大 2, `d` 最大 1, `q` 最大 2）跑很多个 ARIMA 模型，然后计算每个模型的 **AIC** 值。AIC 是一个权衡模型拟合好坏和模型复杂度的指标，**AIC 值越小通常认为模型越好**。最终自动选择了那个 AIC 最小的 (p, d, q) 组合作为最终模型的阶数。
    - 当然，这个自动搜索也有**局限**：范围是我设定的，不一定包含了真正的最优解；而且 AIC 最小也不能 100% 保证残差就完全是白噪声，具体效果得依照后续的检验检验。

4. 我如何验证模型

选好并拟合了最终的 ARIMA 模型后，就要进行模型检验了，主要是看它的残差 (`α̂_t`) 是不是真的像我所希望的白噪声，以及其他假设满不满足。

4.1. 检查残差自相关 (Ljung-Box)

- **目的**: 这是最重要的检查！看模型是不是真的把时间序列里的自相关性都提取干净了。
- **方法**: 用 **Ljung-Box Q 检验。**它的原假设 (H0) 是：残差序列前面 m 个滞后项的自相关系数都是 0。计算公式是：

$$Q = n(n+2)\sum_{k=1}^{m}\frac{\hat{\rho}_k^2}{n-k}$$

- **判断**: 如果这个检验的 **p 值大于 0.05**，我就放心了，说明没理由认为残差还有自相关，模型在这方面是合格的。如果 p 值小于 0.05，那模型就还得调整。

- **图形辅助**: 同时我也会看残差的 ACF 和 PACF 图，理想情况是几乎所有滞后项的柱子都在蓝色置信区间里面（除了 lag 0）。

4.2. 检查残差正态性 (Jarque-Bera)

- **目的**: 看看残差是不是符合正态分布。如果符合，一些统计推断（比如预测区间）会更可靠。
- **方法**: 用 **Jarque-Bera (JB) 检验。**它的原假设 (H0) 是：残差是正态分布的。它是根据残差的偏度 (S) 和峰度 (K) 算出来的：

$$JB = \frac{n}{6}\left(S^2 + \frac{(K-3)^2}{4}\right)$$

- **判断**: 如果 **p 值小于 0.05**，我就得承认残差和正态分布长得不像。
- **图形辅助**: Q-Q 图（看点在不在直线上）和直方图（看像不像钟形）能更直观地展示。

4.3. 检查异方差性

- **目的**: 看看残差的波动幅度（方差）是不是随着时间变化。标准的 ARIMA 模型假设方差是恒定的。
- **方法**: 模型摘要里有个 `Heteroskedasticity (H)` 检验结果和对应的 `Prob(H)` (p 值)。原假设 (H0) 是同方差。
- **判断**: 如果 **p 值小于 0.05**，说明可能存在异方差性。

4.4. 利用图形进行诊断

除了上面的检验，我还为每个最终模型生成了一套诊断图（保存在 `img_arima_auto_v2` 目录），包括：

- 实际值 vs. 拟合值图：看看模型跟实际数据贴合得怎么样。
- 残差 vs. 时间图：找找有没有奇怪的模式或者异常点，看看波动是不是稳定。
- 残差 ACF/PACF 图：这是检查自相关性最直观的图。
- 残差 Q-Q 图和直方图：直观判断正态性。

5. 建模结果与分析

5.1. 最终模型选择汇总

根据自动搜索程序（在 p<=2, d<=1, q<=2 范围内基于 AIC 选择），我为 9 种蔬菜选定的最终 ARIMA 阶数以及关键的残差自相关检验结果如下：

蔬菜名称	ADF p 值	(初步建议 d)	自动选择阶数 (p,d,q)	Ljung-Box p 值	残差自相关
胡萝卜	0.0003	d=0	(1, 0, 1)	0.9306	**已解决**
青椒	0.0115	d=0	(2, 0, 1)	0.5853	**已解决**
西红柿	0.0009	d=0	(1, 0, 2)	0.6337	**已解决**
黄豆芽	0.0335	d=0	(1, 0, 1)	0.0004	**未解决**
红椒	0.0024	d=0	(1, 0, 1)	0.2963	**已解决**
莴笋	0.0005	d=0	(2, 0, 1)	0.5432	**已解决**
小白菜	0.0009	d=0	(1, 0, 1)	0.2713	**已解决**

蔬菜名称	ADF p 值	(初步建议 d)	自动选择阶数 (p,d,q)	Ljung-Box p 值	残差自相关
生姜	0.6761	d=1	(2, 0, 2)	0.0000	**未解决**
绿豆芽	0.0476	d=0	(2, 0, 0)	0.1940	**已解决**
土豆	0.0087	d=0	(2, 0, 1)	0.4644	**已解决**

5.2. 各种蔬菜模型的详细结果解读

基于上面选出的模型（对诊断通过的 7 种蔬菜，结果更可靠），我总结了每种蔬菜的价格特性：

- **胡萝卜 (Carrot) - 可靠模型 ARIMA(1, 0, 1)**
    - 残差诊断通过了自相关检验。
    - 结果显示，胡萝卜价格**没有显著的长期线性趋势或固定的季节模式。**价格的波动主要是由其自身的持续性（AR(1)≈0.97）和短期调整（MA(1)≈-0.28）驱动。
    - 但残差不满足正态分布（峰度特别高）。
- **青椒 (Green Pepper) - 可靠模型 ARIMA(2, 0, 1)**
    - 残差诊断通过了自相关检验。
    - **没有显著的长期线性趋势。**
    - **季节性方面**: 年度 `sin1` 和半年度 `sin2` 项显著，说明青椒价格有其独特的季节性规律。
    - ARIMA 部分显示了价格动态（AR(1)和MA(1)显著）。
    - 残差不满足正态分布（峰度高），且可能存在异方差。
- **西红柿 (Tomato) - 可靠模型 ARIMA(1, 0, 2)**
    - 残差诊断通过了自相关检验。
    - **没有显著的长期线性趋势。**
    - **季节性**: 年度 `cos1`/`sin1` 和半年度 `sin2` 项都显著，表明西红柿价格受年度和半年度周期的共同影响，季节性很强。
    - ARIMA 部分的 AR(1), MA(1), MA(2) 项都显著，说明其时间依赖结构比较复杂。
    - 残差不满足正态分布（峰度高，负偏），且可能存在异方差。
- **黄豆芽 (Soybean Sprout) - 模型 ARIMA(1, 0, 1) 不充分**
    - **模型失败**: 残差仍然有显著的自相关性。因此，基于这个模型的所有结论都**不可靠。**
    - 初步迹象（系数不显著）可能暗示缺乏趋势和季节性，但这需要在找到合适模型后才能确认。残差极度非正态。**需要进一步研究。**
- **红椒 (Red Pepper) - 可靠模型 ARIMA(1, 0, 1)**
    - 残差诊断通过了自相关检验。
    - **没有显著的长期线性趋势。**
    - **季节性**: 年度 `cos1` 和半年度 `cos2`/`sin2` 项显著，显示了年度和半年度结合的影响模式。
    - ARIMA 部分的 AR(1) (系数≈0.98) 和 MA(1) (系数≈-0.17) 项显著。
    - 残差不满足正态分布（峰度高），且可能存在异方差。
- **莴笋 (Asparagus Lettuce / Celtuce) - 可靠模型 ARIMA(2, 0, 1)**

    - 残差诊断通过了自相关检验。
    - **没有显著的长期线性趋势。**
    - **季节性**: 年度 `cos1`/`sin1` 和 半年度 `cos2` 项都显著，特别是半年度 `cos2` 项系数很大，表明半年周期对莴笋价格影响非常大。
    - ARIMA 部分的 AR(1), AR(2), MA(1) 项都显著。
    - 残差不满足正态分布（峰度高）。

- **小白菜 (Small Bok Choy) - 可靠模型 ARIMA(1, 0, 1)**
    - 残差诊断通过了自相关检验。
    - **没有显著的长期线性趋势** (p≈0.12，不显著)。
    - **季节性**: 半年度 `cos2` 项**极其显著且影响巨大**，这是小白菜价格最主要的季节特征，年度项不显著。
    - ARIMA 部分的 AR(1) (系数≈0.95) 和 MA(1) (系数≈-0.25) 项显著。
    - 残差不满足正态分布（峰度极高）。

- **生姜 (Ginger) - 模型 ARIMA(2, 0, 2) 不充分**
    - **模型失败**: 残差仍然有显著的自相关性。而且，ADF 检验强烈建议 d=1（差分），但自动搜索（基于 AIC）却选了 d=0 的模型，这本身就存在矛盾。模型结果**不可靠**。
    - 之前的 OLS 模型显示有强劲的上升趋势，但在这个不可靠的 ARIMA 模型中趋势项不显著。这更说明需要重新建模，**强烈建议尝试 d=1**。残差也严重非正态。**需要进一步研究。**

- **绿豆芽 (Mung Bean Sprout) - 可靠模型 ARIMA(2, 0, 0)**
    - 残差诊断通过了自相关检验。
    - 结果显示，绿豆芽价格**没有显著的长期线性趋势或固定的季节模式。**
    - 价格动态主要由自身的 AR(1) (系数≈0.84) 和 AR(2) (系数≈0.12) 项决定。
    - 残差严重非正态（峰度极高）。

- **土豆 (Potato) - 可靠模型 ARIMA(2, 0, 1)**
    - 残差诊断通过了自相关检验。
    - 结果显示，土豆价格**没有显著的长期线性趋势或固定的季节模式。**
    - 价格动态主要由显著的 AR(1), AR(2), MA(1) 项决定。
    - 残差不满足正态分布（峰度高，负偏度），且可能存在异方差性。

6. 讨论与思考

6.1. 主要发现总结

这次建模过程让我对这些蔬菜的价格行为有了更深入的了解：

1. **ARIMA 模型确实更靠谱**: 相比简单的 OLS，ARIMA 模型能更好地处理价格数据中普遍存在的时间依赖性（自相关），这对得出可靠结论至关重要。我们成功为 7 种蔬菜找到了残差不相关的模型。
2. **长期趋势不明显**: 一个有趣的发现是，当模型充分考虑了价格自身的动态后，之前 OLS 模型可能显示的长期线性趋势（无论是上升还是下降）大多变得不再显著。这可能意味着价格的长期变化更像是随机游走或者受到非线性因素影响，而非简单的直线式增长或减少（生姜可能是个例外，但它的模型还不完善）。

3. **季节性各有千秋**: 不同蔬菜确实有不同的“脾气”。有的（像西红柿、青椒、红椒、莴笋）表现出比较复杂的年度和半年度结合的季节模式；有的（像小白菜）则有一个非常突出的半年周期；还有的（像胡萝卜、绿豆芽、土豆）在这个模型框架下，似乎没有固定的、可以用傅里叶项捕捉的季节性规律。
4. **价格粘性强**: 大部分模型的 AR(1) 系数都很大且接近 1，说明价格有很强的“惯性”或“记忆”，昨天的价格对今天的价格影响非常大。
5. **“问题”蔬菜**: 黄豆芽和生姜的模型始终不太理想，特别是残差自相关问题没解决。生姜可能需要差分处理（d=1），而黄豆芽可能需要更高阶的 AR/MA 项或者根本不适合这个模型。

6.2. 模型的局限性反思

在这次建模中，我也意识到了一些局限：

1. **自动搜索不是万能的**: 我设置的搜索范围 (p<=2, d<=1, q<=2) 是有限的，主要是为了提高效率。对于黄豆芽和生姜，可能真正的最优阶数超出了这个范围。
2. **季节性处理方式**: 我用的傅里叶项假设季节模式每年都一样。但现实中，某年的春节特别早或晚，或者天气异常，都可能让季节模式发生变化。更复杂的模型也许能捕捉这种时变性。
3. **外部信息局限性**: 模型没有包含外部信息，比如极端天气、政策调整、疫情影响等，这些都可能是导致残差（尤其是那些极端值）的原因，也限制了模型的解释力。
4. **残差正态性和异方差**: 这两个问题在很多蔬菜模型中都存在。虽然 ARIMA 对此有一定包容性，但如果要做很精确的预测或者风险分析，这些问题还是需要关注的，比如用 GARCH 模型处理波动性。

6.3. 可能的优化方向

基于此次分析，我觉得可以从这几个方面进行优化：

1. **深挖“问题”蔬菜**: 对黄豆芽和生姜，需要花更多时间进行模型识别。可以尝试更大的 p, q 搜索范围，或者引入 SARIMA 模型（如果残差图显示出季节性滞后相关）。特别是生姜，一定要试试 d=1 的模型。
2. **处理残差问题**: 对普遍存在的非正态（尤其是高峰度）问题，可以研究一下是不是由少数几个极端异常值引起的，尝试识别和处理它们。对于异方差，可以考虑结合 GARCH 类模型。
3. **引入更多解释变量**: 如果能获取到相关数据，比如天气数据（温度、降水）、重大节假日虚拟变量、甚至是化肥农药等成本信息，把它们加入模型作为外生变量，可能会大大提高模型的解释能力和准确性。
4. **探索非线性**: 价格变化不一定是线性的，未来也可以试试看一些非线性时间序列模型。

7. 结论

通过这次系统的建模分析，我对学校食堂 9 种主要蔬菜的价格变化规律有了更清晰的认识。我采用了带有外生变量的 ARIMA 模型，并通过自动搜索确定了适合大部分蔬菜的模型阶数，成功解决了 OLS 模型中存在的严重自相关问题，为其中的 7 种蔬菜建立了统计上更可靠的模型。

我的主要结论是：在控制了时间序列的内在动态后，大多数蔬菜价格并未显示出显著的长期线性趋势，但普遍存在显著的季节性波动，其模式（年度、半年度或组合）因品种而异。例如，小白菜价格呈现非常强的半年周期。然而，黄豆芽和生姜的价格动态更为复杂，当前的模型设定未能充分捕捉其特性。此外，即使模型较好地处理了自相关，残差的非正态性（尤其是高峰度/厚尾）问题依然普遍，这提示蔬菜价格容易受到各种未包含在模型中的因素或随机冲击的影响，导致价格出现跳跃。

总的来说，这次分析利用 ARIMA 模型为理解蔬菜价格提供了有效的数学框架和定量的洞察，同时也指出了模型存在的局限以及未来可以进一步研究和优化的方向。