Basic Principles of the ARIMA Model

The ARIMA model consists of three components:

AR (Autoregressive): Uses past data to predict current values.
For example, today's sales may be correlated with sales from the previous few days.

I (Differencing): Transforms non-stationary data (those with trends or seasonality) into stationary data by differencing the data.
For example, taking a first-order difference can eliminate the trend in an upward-trending time series.

MA (Moving Average): Uses an error term from previous periods to correct the predicted value.
For example, a discrepancy between yesterday's predicted value and the actual value may affect today's forecast.

ARIMA(p, d, q):
p: The order of the autoregressive function (how many past observations are used).
d: The number of differencing steps (the number of times the data is stationary).
q: The order of the moving average function (how many past errors are used).

Modeling Steps：

1.Visualization and Stationarity Testing

- Plot the time series to observe any trend or seasonality.
- Use the Augmented Dickey-Fuller (ADF) test to determine whether the data is stationary.

2.Differencing

- If it is not stationary, perform difference analysis on the data until it becomes stationary.

3.Model Order (Determine p and q)

- Use the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots to select appropriate p and q.

4.Model Fitting
5.Model Diagnostics
6.Forecasting

This is a graph of average daily prices for the Fruit & Vegetables category, with 2,378,674 raw transaction records aggregated into 731 days of daily data (approximately 2 years, from January 2024 to January 2026).

Time Series Trend Analysis

1. Overall Level

The average daily price is concentrated between $5.04 and $5.14, with a relatively small fluctuation range, indicating overall price stability in this category.

2. Cyclical Characteristics

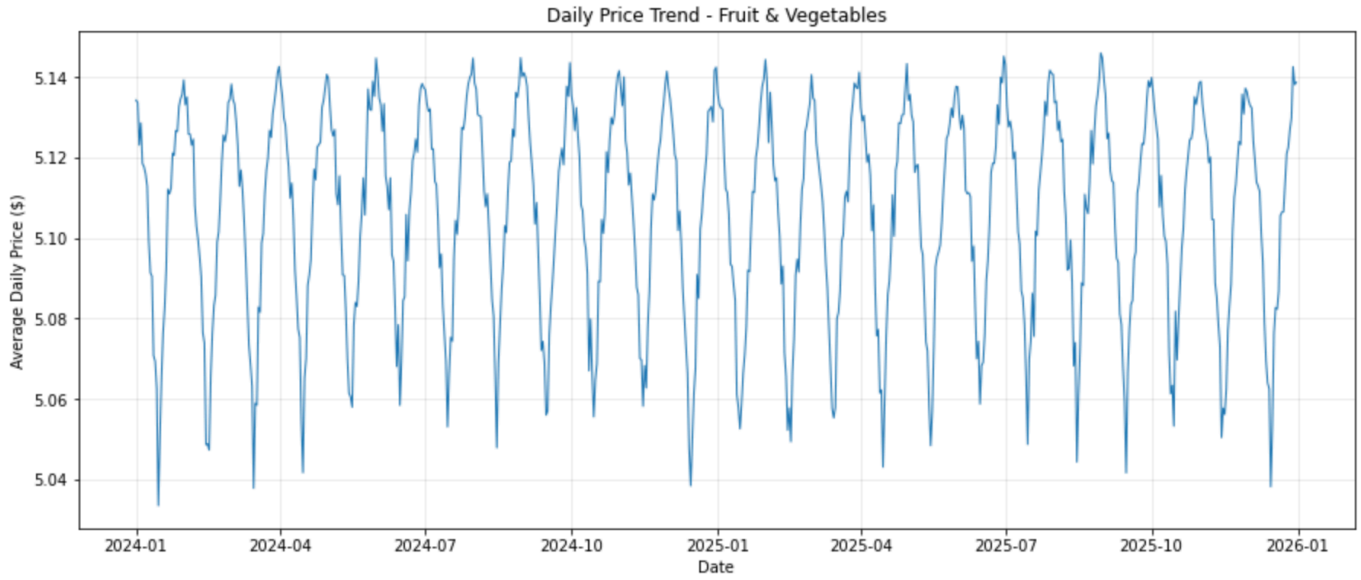
The chart clearly shows regular ups and downs, exhibiting a roughly monthly cycle.

3. Short-Term Fluctuations

Local prices may experience rapid declines and rebounds, potentially due to short-term supply and demand shocks, holiday promotions, or seasonal factors.

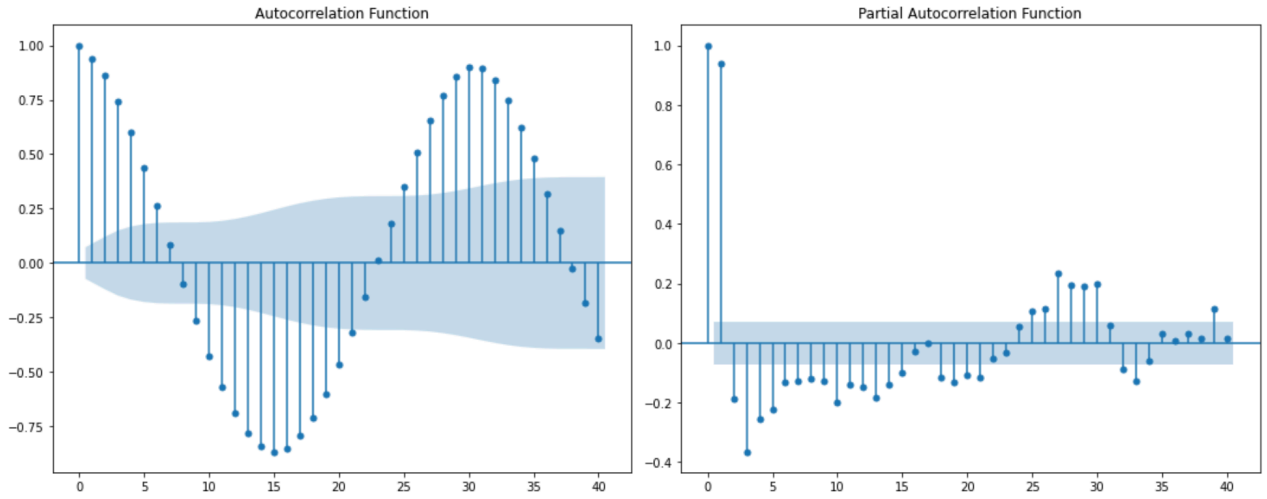
4. Trend

There is no clear upward or downward trend; prices remain within a narrow range.




```
Missing values: daily_price    0
dtype: int64
```

```
Test Statistic                  -1.196817e+01
p-value                          3.955338e-22
#Lags Used                       2.000000e+01
Number of Observations Used      7.100000e+02
Critical Value (1%)             -3.439594e+00
Critical Value (5%)             -2.865619e+00
Critical Value (10%)            -2.568942e+00
dtype: float64
```

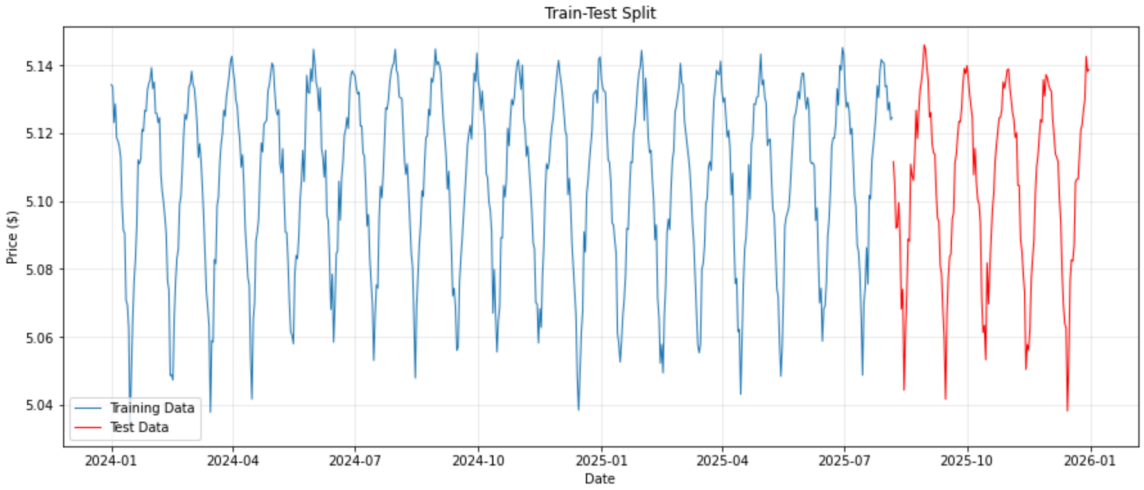


1. Stationary Test (ADF Test)
- Test Statistic = -11.97 (much less than the critical values of -3.43, -2.86, and -2.57).
- p-value ≈ 3.96e-22 < 0.05.

2. ACF
- The ACF plot shows cyclical fluctuations (approximately 12–13 days per cycle).
- This indicates significant seasonality/cyclicality in the series, typically a monthly cycle effect.

3. PACF
- The PACF plot has significant peaks at lag=1 and lag=2, followed by a rapid decay.

```
Train period: 2024-01-01 00:00:00 to 2025-08-06 00:00:00
Test period: 2025-08-07 00:00:00 to 2025-12-31 00:00:00
Train size: 584, Test size: 147
```

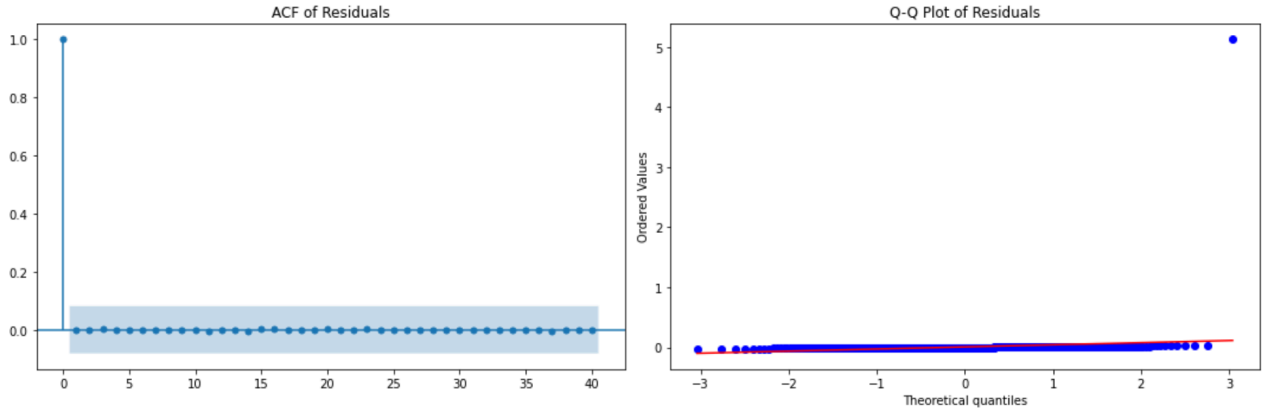


Training Set (80%): 01.01.2024 → 06.08.2025, a total of 584 days.
Test Set (20%): 07.08.2025 → 31.12.2025, a total of 147 days.

```
Model Summary:
                               SARIMAX Results
==============================================================================
Dep. Variable:            daily_price   No. Observations:                  584
Model:                 ARIMA(2, 1, 0)   Log Likelihood                1964.355
Date:                Fri, 12 Sep 2025   AIC                          -3922.711
Time:                        15:26:27   BIC                          -3909.606
Sample:                    01-01-2024   HQIC                         -3917.603
                         - 08-06-2025
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.1129      0.033      3.407      0.001       0.048       0.178
ar.L2          0.3172      0.034      9.268      0.000       0.250       0.384
sigma2       6.93e-05   3.57e-06     19.410      0.000    6.23e-05    7.63e-05
===================================================================================
Ljung-Box (L1) (Q):                   1.46   Jarque-Bera (JB):                27.35
Prob(Q):                              0.23   Prob(JB):                         0.00
Heteroskedasticity (H):               0.84   Skew:                             0.31
Prob(H) (two-sided):                  0.23   Kurtosis:                         3.86
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

 Model Diagnostics:
Residuals mean: 0.0088
Residuals std: 0.2126
```

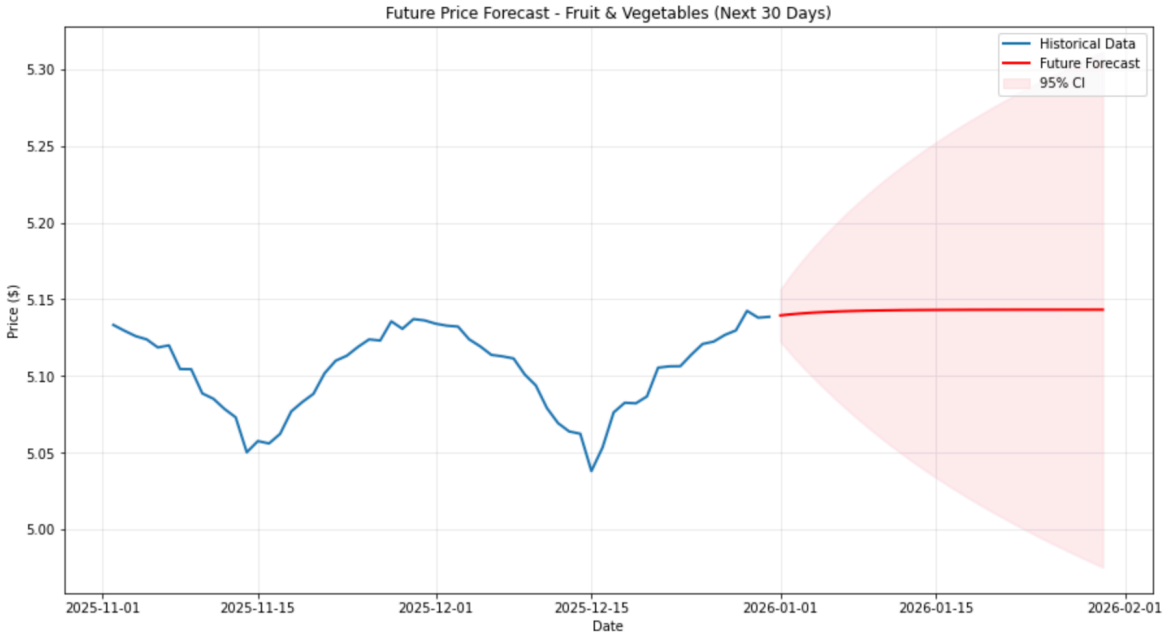


Model Fitting Results：
Model: ARIMA(2,1,0)
AIC = -3922.71, BIC = -3909.61 （The low AIC values indicate a good model fit.）

AR Coefficients:
AR(1) = 0.113 (Significant, p=0.001)
AR(2) = 0.317 (Highly Significant, p<0.001)
This indicates that the current price is primarily influenced by price changes over the previous two days.

ACF of Residuals: Most of the data fall within the confidence interval.

Q-Q Plot: There are slight deviations in the tails, but the overall distribution is close to normal.



Parameter Optimization
Order (1, 1, 0): MAE=$0.03, AIC=-3863.0, BIC=-3854.3
Order (2, 1, 0): MAE=$0.02, AIC=-3922.7, BIC=-3909.6
Order (3, 1, 0): MAE=$0.02, AIC=-3935.4, BIC=-3917.9
Order (0, 1, 1): MAE=$0.03, AIC=-3856.8, BIC=-3848.1
Order (1, 1, 1): MAE=$0.02, AIC=-3898.3, BIC=-3885.2
Order (2, 1, 1): MAE=$0.02, AIC=-3930.2, BIC=-3912.7

The (3,1,0) and (2,1,1) models have lower AIC values (indicating a better fit).However, they are more complex (requiring more parameters to estimate, making them prone to overfitting).

The (1,1,1) model also has a very small MAE ($0.02), and its prediction accuracy is comparable to that of the more complex models.

Forecast Results:
The 30-day forecast shows the price will remain around $5.13-5.14