FAO Food Price Index Data Preprocessing Documentation

Dataset Overview

The FAO Food Price Index dataset contains time-series data tracking food prices across different categories including:

- Food Price Index (overall)
- Meat
- Dairy
- Cereals
- Oils
- Sugar

The data spans multiple years with monthly recordings of price indices.

Preprocessing Operations

1. Missing Value Analysis and Treatment

Operation: `df.isnull().sum()` and forward fill method **Purpose:**

- To determine whether the time series data has any gaps
- To guarantee dataset continuity for precise analysis
- To preserve the integrity of the data without adding fictitious values Impact of

Output:

- Whole dataset with no values missing
- maintained trends by filling up any gaps with the most recent value available.

2. Date Parsing and Feature Extraction

Operation: Converting 'Date' column to datetime and extracting components **Purpose:**

- To make time-based analysis and operations possible
- To make time-series analysis easier at various levels of detail
- In order to facilitate seasonal analysis

Impact of Output:

- 'Year' and 'Month' columns were added.
- made it simpler to organize and filter by time components.
- Enhanced capacity to identify seasonal trends

3. Data Normalization

Operation: MinMaxScaler application on price columns **Purpose:**

- To bring all features to a comparable scale (0-1 range)
- To prevent larger-scale features from dominating the analysis
- To improve the performance of any subsequent machine learning algorithms **Output Impact:**
- Created normalized versions of all price columns
- Maintained original data while adding scaled versions
- Enabled fair comparison between different price components

4. Outlier Detection and Treatment

Operation: IQR (Interquartile Range) method **Purpose:**

- To identify extreme values that might skew analysis
- To maintain data integrity while handling unusual price spikes
- To provide cleaned versions of the data for sensitive analyses **Output Impact:**
- Created cleaned versions of price columns
- Identified and handled extreme price movements
- Preserved original data while providing cleaned alternatives

5. Derived Feature Creation

Operation: Calculating component ratios and monthly averages **Purpose:**

- To understand the relative contribution of each component
- To identify structural changes in price composition
- To enable analysis of price component relationships **Output Impact:**
- Added ratio columns for each component
- Created monthly average statistics
- Enhanced dataset with derived analytical features

6. Duplicate Removal

Operation: `drop_duplicates()` **Purpose:**

- To ensure data quality
- To prevent double-counting in statistical analyses

- To maintain data integrity **Output Impact:**
- Removed any potential duplicate entries
- Ensured each time point has unique values

7. Time Series Smoothing

Operation: Rolling averages with 3-month window **Purpose:**

- To reduce noise in the time series
- To highlight underlying trends
- To facilitate trend analysis **Output Impact:**
- Added smoothed versions of each price series
- Created clearer trend visualization possibilities
- Reduced impact of monthly volatility

8. Data Transformation

Operation: Log transformation **Purpose:**

- To handle skewness in price distributions
- To make relationships more linear
- To reduce the impact of extreme values **Output Impact:**
- Added log-transformed versions of price columns
- Improved distribution characteristics
- Enhanced ability to detect multiplicative patterns

9. Change Calculations

Operation: Month-over-month percentage changes **Purpose:**

- To track price momentum
- To identify significant price movements
- To enable growth rate analysis **Output Impact:**
- Added monthly change columns
- Created indicators of price dynamics
- Enabled volatility analysis

10. Index Structure Enhancement

Operation: Creation of YearMonth index **Purpose:**

- To improve time series functionality
- To enable easier period-based analysis
- To enhance data organization **Output Impact:**
- Restructured dataset with proper time series index

- Improved time-based querying capabilities
- Enhanced data structure for time series analysis

Final Dataset Structure

The preprocessing resulted in an enhanced dataset containing:

- Original price columns
- Normalized versions of all price components
- Cleaned (outlier-free) versions of price components
- Component ratios and derived features
- Various transformations (log, moving averages)
- Time-based features and changes
- Properly structured time series index

Data Quality Assurance

- No missing values in the final dataset
- All derived features properly calculated and verified
- Time series continuity maintained
- Original data preserved alongside transformations
- Clear naming conventions for all new features

This preprocessing pipeline ensures the dataset is ready for various types of analyses, from simple trend visualization to complex machine learning applications.