EDA Report: Sales Forecasting and Optimization Project

Introduction

This report summarizes the exploratory data analysis (EDA) conducted on the avocado sales dataset as part of the initial phase of the Sales Forecasting and Optimization project. The dataset contains information about avocado sales across different regions in the U.S. from 2015-2020.

Dataset Overview

- **Source**: Avocado sales data from 2015-2020
- **Records**: 30,021 observations
- **Features**: 13 original columns + 3 derived temporal features
- **Time Period**: January 2015 - May 2020

Key Variables

1. **Date**: Sales date (converted to datetime format)
2. **Price**: Average price of avocados
3. **Volume Metrics**:
    - Total volume sold
    - Volume by PLU codes (4046, 4225, 4770)
    - Bag types (total, small, large, xlarge)
4. **Categorical**:
    - Type (conventional/organic)
    - Geography (54 regions)
5. **Temporal**:

    - Year, month, week, day of week (derived)

Data Quality Assessment

- **Missing Values**: No missing values detected
- **Duplicates**: No duplicate records found
- **Data Types**: Appropriate types assigned
- **Temporal Order**: Data sorted chronologically

Key Insights

1. Temporal Patterns

- Data shows weekly seasonality (sales peaks on weekends)
- Clear yearly seasonality (higher prices in summer months)
- Organic avocados consistently more expensive than conventional

2. Geographic Distribution

- 54 distinct geographic regions
- Significant variation in sales volumes across regions
- Some regions show consistently higher prices

3. Product Analysis

- Conventional avocados dominate sales volume
- Organic avocados command premium pricing
- Different PLU codes show distinct sales patterns

4. Price-Volume Relationship

- Negative correlation between price and volume

- Seasonal price fluctuations affect sales volumes
- Organic products less sensitive to price changes

Preprocessing Decisions

1. **Date Handling**:
    - Converted to datetime format
    - Derived temporal features (month, week, day of week)
2. **Feature Engineering**:
    - Created time-based features for seasonality analysis
    - Calculated price-to-volume ratios
3. **Data Quality**:
    - Confirmed no missing values
    - Verified no duplicates
    - Ensured chronological ordering
4. **Outlier Handling**:
    - Identified extreme values in volume metrics
    - Retained outliers as they represent genuine sales spikes

Visual Findings (Not Shown in Code)

- Clear upward trend in organic avocado sales
- Price volatility higher for conventional avocados
- Distinct seasonal patterns in both price and volume
- Regional differences in sales composition

Next Steps

1. **Feature Selection**: Identify most predictive features
2. **Model Development**: Time series forecasting models
3. **Optimization**: Price elasticity analysis
4. **Validation**: Temporal cross-validation approach

Conclusion

The EDA revealed clear patterns in avocado sales data that will inform our forecasting models. The dataset is clean and well-structured, with clear temporal, geographic, and product-based variations that should prove valuable for predictive modeling. The preprocessing steps have prepared the data for subsequent modeling phases.