Avocado Price and Volume Analysis

Milestone Overview

This is data cleaning, analysis, and visualization of avocado price and sales data. The dataset contains information about avocado prices, volumes, types (conventional/organic), regions, and temporal patterns across multiple years in the United States.

Data Cleaning Process

- **Initial Data Inspection**

The dataset contains 30,021 entries with 22 columns including:

- Date, Average price, Total volume, PLU codes (4046, 4225, 4770), Total bags, Type (conventional/organic), Year, Geography (region)

- Various derived features (month, season, holiday flags, promotion effects)

- **Cleaning Steps Performed**

1. Handling Missing Values:

- The column 'month_change' had 21 missing values

- Missing values were filled with the mean of the column

2. Data Type Conversion:

- Date column converted to datetime format

- Categorical variables properly encoded

3. Feature Engineering:

- Derived temporal features (month, season, weekday)

- Created interaction terms (volume_price_interaction)

- Calculated monthly average volume

4. Outlier Treatment:

- Visual inspection of price and volume distributions

- No extreme outliers requiring removal were found

5. Consistency Checks:

- Verified price and volume relationships

- Checked for logical inconsistencies in date ranges and regions

Challenges Encountered

1. Temporal Aggregation:

- Data was recorded at different geographic levels

- Needed to ensure consistent time series analysis

2. Volume Measurement:

- Multiple volume columns (total volume, specific PLU volumes, total bags)

- Required careful aggregation for analysis

3. Seasonal Patterns:

- Significant seasonal variation in both price and volume

- Needed decomposition for proper analysis

Key Insights from Analysis

- **Price Trends**

1. Organic vs. Conventional:

- Organic avocados consistently command higher prices (average $1.79 vs $1.22 in sample rows)

- Price premium has remained stable over time

2. Seasonal Patterns:

- Prices typically peak in winter months

- Lowest prices observed during summer harvest seasons

3. Regional Variations:

- West Coast regions show highest average prices

- Southeast and Midwest have most competitive pricing

Volume Analysis

1. Sales Growth:

- Steady year-over-year increase in total volume

- Particularly strong growth in organic segment

2. Package Type Distribution:

- PLU 4046 (small/medium Hass) dominates conventional sales

- Organic sales show more balanced distribution across PLUs

3. Promotional Impact:

- Promotions show clear volume spikes

- Limited effect on organic avocado demand

Advanced Visualizations

- **Time Series Decomposition**

The notebook includes seasonal decomposition of time series (STL) for:

- Average price trends

- Total volume sold

- Regional comparisons

Components visualized:

- Observed data

- Trend component

- Seasonal component

- Residuals

- **Interactive Dashboard Elements**

1. **Price Trends Dashboard:**

- Time slider for date range selection

- Region and type filters

- Moving average controls

2. **Volume Analysis Dashboard:**

- Stacked area chart of volume by PLU type

- Year-over-year comparison tool

- Geographic heatmap of sales distribution

3. Price-Volume Relationship:

- Scatter plot with regression line

- Dynamic quadrant analysis

- Correlation by region and type

Conclusion

The cleaned dataset and analysis provide valuable insights into avocado market dynamics. The interactive visualizations enable exploration of complex relationships between price, volume, geography, and time. This analysis forms a strong foundation for demand forecasting and strategic decision-making in the avocado industry.