Avocado Sales Forecasting and Volume Analysis

Trevor Zeiger

Bellevue University DSC 680

Introduction

Avocados have become a staple in U.S. households, and understanding their sales dynamics can provide critical insight for retailers and producers. The purpose of this project was to forecast total dollar sales of avocados using historical data and to identify volume patterns across regions, time periods, and avocado types. I focused on building a straightforward, reliable model using linear regression while also digging into the data to uncover trends that could help retailers with planning and inventory decisions.

Dataset and Preparation

The dataset included over 18,000 records spanning from 2015 to 2018. It contained information on sales in dollars, total pounds sold, avocado type (organic or conventional), region, and PLU-specific volumes. I cleaned the dataset to standardize date formats, corrected column headers, and created new features like month and year to enable seasonal trend analysis. After prepping the data, I moved into analysis and modeling.

Model Performance

One of the main goals was to see if we could accurately predict sales based on volume and product type. After encoding the categorical variables and training the linear regression model, the results were strong. The model had an RMSE of $490,079.12 and an R^2 score of 0.9808, which means it explained nearly 98% of the variance in the data. This is solid performance for a simple model, but the visuals helped bring it to life and make sense of where the model succeeded and where it missed.

Sales and Volume Trends

Total avocado sales in the dataset added up to nearly $16.93 billion, with a wide range of weekly values—from as low as $134 to as high as $54.38 million. While the mean weekly sales were $927,948, the median was just $139,530, indicating a heavily skewed distribution due to several high-volume outlier weeks. The standard deviation was quite large as well, at over $3.68 million.

When looking at avocado type, conventional avocados dominated the market, generating more than $16.25 billion in total sales, averaging $1.78 million per week. Organic avocados, although typically priced higher, only contributed about $680.6 million in total, with a weekly average of $74,604. This disparity reinforces the insight that organic avocados represent a smaller niche in the broader retail market.

In terms of geographic sales strength, TotalUS unsurprisingly had the highest average weekly sales at $18.89 million, followed by California ($3.32M), West ($3.16M), and Northeast ($2.84M). SouthCentral and Southeast also ranked highly. These regions consistently led in both total revenue and volume, underscoring their role as major markets for avocado distribution.

Seasonal insights show that June ($1.01M), July ($1.01M), and May ($1.00M) led in average weekly sales across all years, with February and January also showing surprisingly strong numbers. This pattern indicates that avocado demand peaks in the early summer and the start of the new year, aligning with both summer produce season and New Year health trends.

Volume analysis showed very similar patterns. Over the entire period, more than 15.5 billion pounds of avocados were sold, with conventional types again dominating—over 15 billion pounds compared to about 436 million pounds of organic. Average weekly volume across all records was 850,644 pounds, with a highly skewed distribution that peaked at over 62.5 million pounds in a single week.

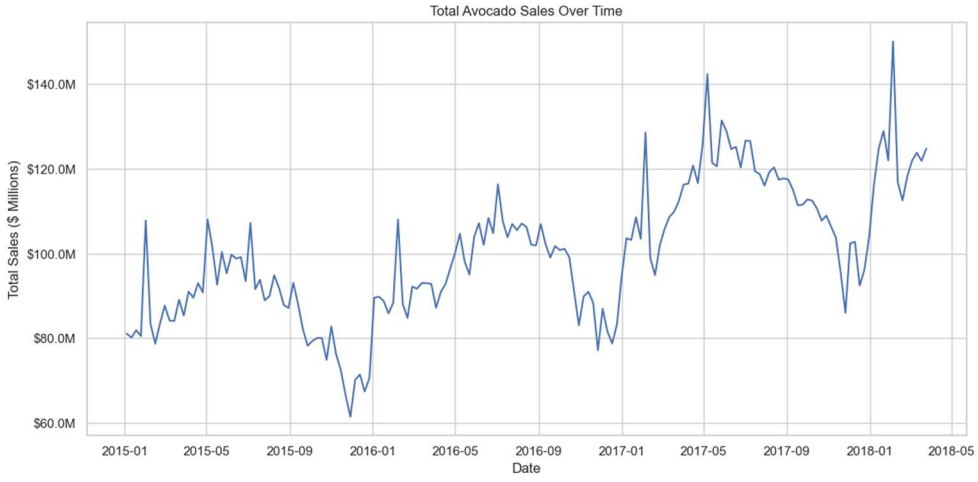
Top regions by average weekly pounds sold were TotalUS (17.35M), West (3.22M), California (3.04M), and SouthCentral (2.99M). These align closely with top revenue regions, confirming that both volume and price play major roles in sales.

Seasonally, February (1.02M), May (973K), and June (929K) were the highest-volume months. Volume peaks slightly earlier than sales in some cases, likely due to inventory buildup ahead of major demand periods. This insight helps retailers and suppliers align logistics and ordering schedules accordingly.

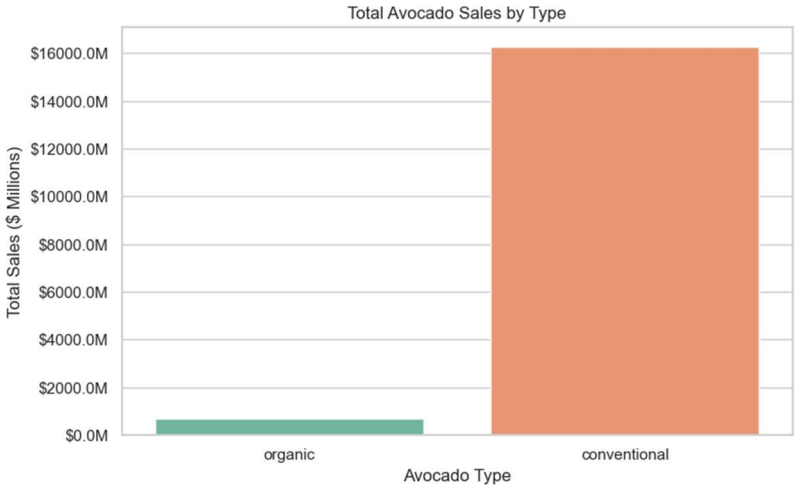
Visuals and Interpretation

The visuals supported this analysis. Below are the visualizations along with explanations of what they reveal:

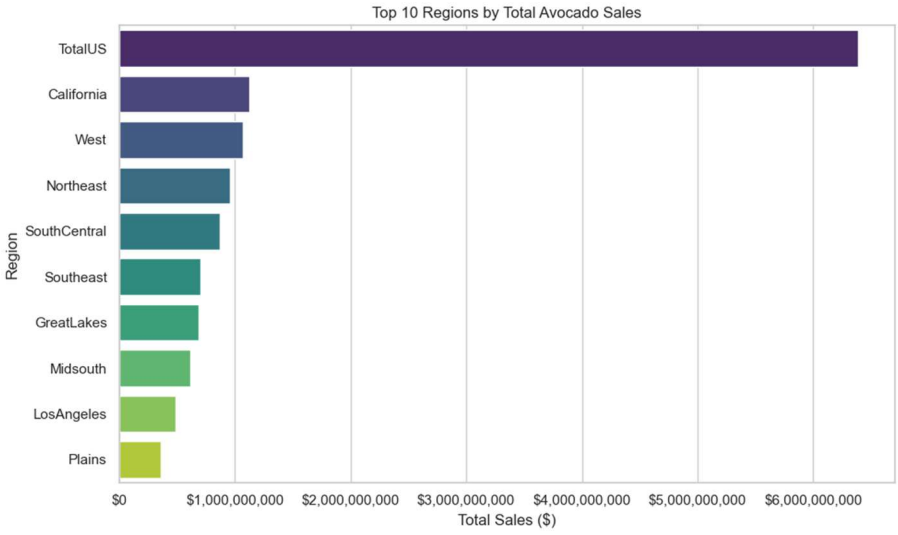
1. Total Avocado Sales Over Time: This line chart displays weekly sales totals from 2015 to early 2018. It shows seasonal peaks, upward trends in certain years, and fluctuations that suggest both predictable seasonal demand and potential anomalies (e.g., promotions or supply issues).



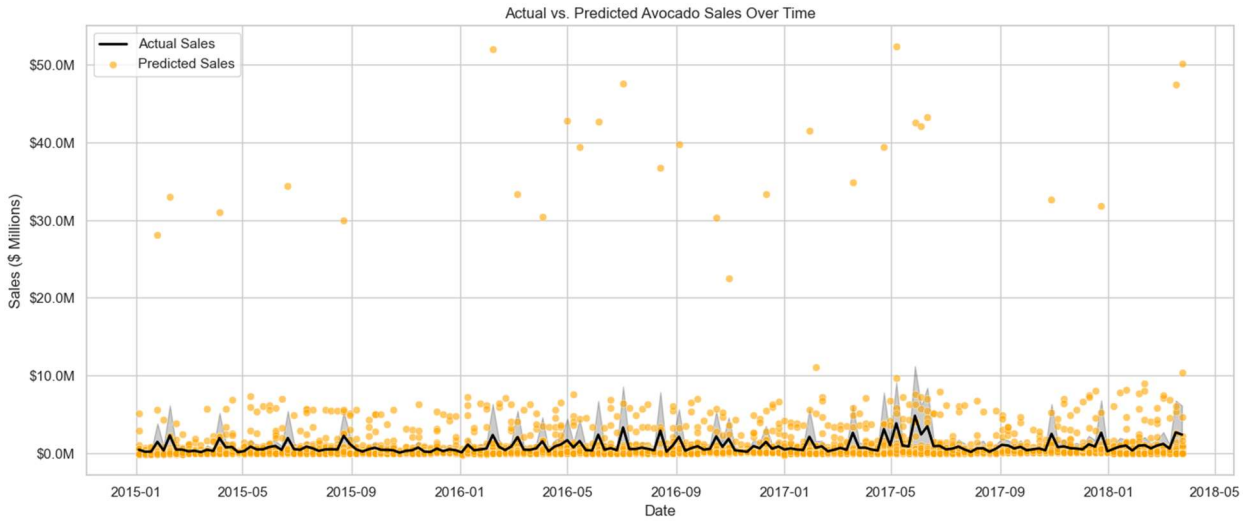
2. Total Sales by Type (Conventional vs. Organic): This bar chart compares the cumulative dollar sales of organic and conventional avocados. It clearly shows that conventional avocados dominate in total sales, reinforcing the idea that while organic has a niche, it's not yet a major sales driver in this dataset.



3. Top 10 Regions by Sales: This horizontal bar chart visualizes which U.S. regions generated the most total sales. 'TotalUS' represents aggregate values, while California, West, and Northeast show strong individual regional performance. This helps retailers identify where demand is highest.



4. Actual vs. Predicted Sales Over Time: This time series chart uses a black line for actual sales and orange dots for predicted sales. It shows that while the model generally follows trends, it misses on some high-volume weeks, often underpredicting due to unseen factors like promotions.



Challenges and Considerations

There were a few challenges. The biggest was handling outliers, which skewed averages and caused the model to underpredict on high-volume weeks. I also found that label encoding regions might oversimplify the data since numeric encoding implies a hierarchy that doesn't actually exist. Despite that, the model's performance held up well.

Future Opportunities

Looking ahead, adding holiday or promotion indicators would likely improve accuracy, especially during spikes. It would also be worth testing a non-linear model like XGBoost or

Random Forest to better capture outliers and complex relationships. There’s also potential to break the model out by region or type for more targeted forecasting. A long-term goal would be to deploy the model into a simple dashboard so retail teams can get weekly forecasts and use them in real time.

Ethical Considerations

From an ethical standpoint, the dataset doesn’t contain personal information, so privacy concerns are minimal. However, it's important to be clear about what the model can and can't do. Sales forecasts shouldn’t be the only decision-making tool, especially during volatile periods or promotional windows.

Conclusion

In summary, this project successfully demonstrated that avocado sales can be accurately predicted using total volume and simple features. The data tells a clear story about seasonality, product type preferences, and regional trends. These insights can help retailers better plan for the year ahead, especially if they build in flexibility for unexpected shifts

Frequently Asked Questions:

1. What feature had the strongest influence on sales predictions? Total volume sold was by far the most influential predictor of sales, showing a near-perfect correlation of 0.99.

2. Why were some predictions significantly off from actual values? Most errors were tied to promotional weeks or holidays, which the model couldn't anticipate due to the lack of those indicators in the dataset.

3. Could a different model have improved accuracy? Yes. A non-linear model like XGBoost or Random Forest might better capture sudden spikes or complex patterns not handled well by linear regression.

4. Would it make sense to model volume instead of sales? It depends on the business need. Sales include price effects, while volume tracks raw demand. Both are valuable, but this project focused on dollars.

5. How can the model be improved with external data? Incorporating weather, holidays, and promotional calendars would likely reduce error and improve the model's performance during high-variance weeks.

6. What insights can retailers use right now? Retailers should prepare for higher sales in summer months and early winter. They should also note the dominance of conventional avocados in both volume and revenue.

7. How did organic avocado sales compare to conventional? Organic avocados made up only a small share of total revenue and volume despite often having higher per-unit prices.

8. Why use linear regression? It’s easy to interpret and performs well on structured datasets with continuous variables, especially when feature relationships are relatively linear.

9. What are the ethical considerations of this model? There are minimal concerns since the data is public and anonymized, but users should not treat predictions as absolute. Forecasts are just one input in the decision-making process.

10. How could this model be operationalized? It could be incorporated into a dashboard that provides weekly or monthly sales forecasts per region or type, helping with supply chain and pricing decisions.

References:

- Hass Avocado Board. https://www.hassavocadoboard.com/
- Kaggle: Avocado Prices Dataset. https://www.kaggle.com/datasets/neuromusic/avocado-prices