White Paper on Mastering Avocado Pricing

A Strategic Approach to Regional, Seasonal and Predictive Insights

By
Keneilwe Rangwaga

Abstract

This white paper examines the evolving dynamics of the avocado market from 2015 to 2023, focusing on pricing, sales trends, and consumer preferences. Utilizing data provided by the Hass Avocado Board, the analysis highlights the significant impact of regional demand, seasonality, and market dynamics on avocado prices and sales volumes. Through exploratory data analysis and predictive modelling, employing techniques such as Linear Regression, Random Forest, and XGBoost, the study identifies key patterns and trends influencing the avocado market. The findings indicate that the Random Forest model offers the most accurate predictions, providing stakeholders with actionable insights for optimizing pricing strategies and enhancing supply chain efficiency. The paper concludes with recommendations for future research to explore external factors affecting avocado sales, contributing to a more comprehensive understanding of this dynamic market.

Introduction

Background of Study

Avocados have transcended their status as a mere culinary trend to become a significant player in the global produce market. Their popularity has soared in recent years, evidenced by the ubiquity of avocado-based dishes, from guacamole at social gatherings to avocado toast gracing breakfast menus worldwide. This growing demand has created complexities in the avocado market, particularly concerning pricing and sales volume. Various factors, including regional preferences, seasonal fluctuations, and broader market dynamics, have notably influenced how avocados are priced and sold across different regions and times of the year.

To navigate this intricate landscape, this project aims to dissect the pricing and sales patterns of avocados between 2015 and 2023, leveraging comprehensive data from the Hass Avocado Board. By

examining the interplay between geography, seasonality, and economic factors, this analysis seeks to unveil the underlying trends that shape the avocado market. The insights derived from this study will prove invaluable for stakeholders within the avocado industry, enabling them to optimize pricing strategies, anticipate market trends, and enhance overall supply chain efficiency. Ultimately, a clearer understanding of the avocado market can lead to more informed decision-making, benefiting producers, retailers, and consumers alike.

Data Collection and Description

The dataset used in this analysis was initially downloaded from the Hass Avocado Board website in May 2018 and compiled into a single CSV file. It tracks avocado sales across different U.S. regions, product types (conventional vs. organic), and various sizes, over time. According to the Hass Avocado Board, the data reflects weekly retail scan data for national retail volume and price. This retail scan data is aggregated from multiple outlets, including grocery, mass, club, and drug stores, providing a comprehensive view of the avocado market.

[Kaggle dataset] (https://www.kaggle.com/datasets/vakhariapujan/avocado-prices-and-sales-volume-2015-2023/data)

The dataset contains 53,415 entries and 12 columns, covering a period from 2015 to 2023. It includes the following key features:

- **Date**: The date of the observation (466 unique dates).
- **AveragePrice**: The average price of avocados.
- **TotalVolume**: The total volume of avocados sold.
- **plu4046, plu4225, plu4770**: Product lookup codes (PLU) for small, medium, and large avocados, respectively.
- **TotalBags, SmallBags, LargeBags, XLargeBags**: The number of avocados sold in different packaging types.
- **Type**: Indicates whether the avocado is “conventional” or “organic.”
- **Region**: The region of the sale (60 unique regions).

This dataset provides a comprehensive view of avocado sales patterns and pricing dynamics across the U.S. over nearly a decade, making it an ideal source for the analysis of regional and seasonal trends.

Data Cleaning and Preparation

Before proceeding with the analysis, the dataset was cleaned and prepared to ensure it was suitable for modelling. The data cleaning process involved:

- **Handling Missing Values**: Any missing or incomplete data entries were either imputed or removed, depending on their impact on the analysis.
- **Standardizing Text Data**: Text-based columns, such as the Type and Region, were standardized to ensure consistency throughout the dataset.

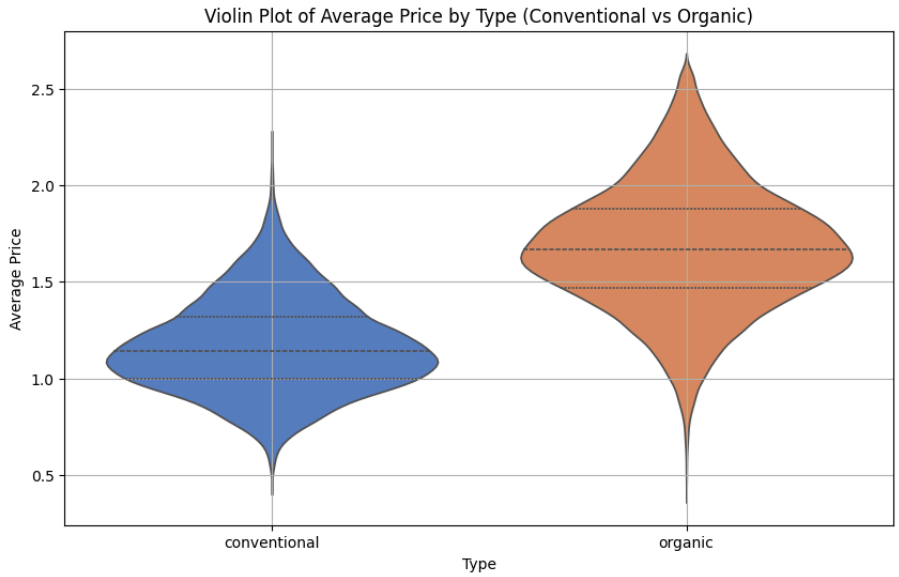
- **Detecting and Removing Outliers**: Outliers in pricing and sales volumes were identified and removed where appropriate, to avoid skewing the results.
- **Feature Engineering**: Additional columns were created to enhance the dataset. This included converting the Date column into Year and Month columns and calculating Price Per Unit Volume to better understand how pricing relates to sales volumes.

This data preparation process ensured that the dataset was well-structured and ready for analysis, enabling the exploration of trends and the development of predictive models.

Exploratory Data Analysis (EDA)

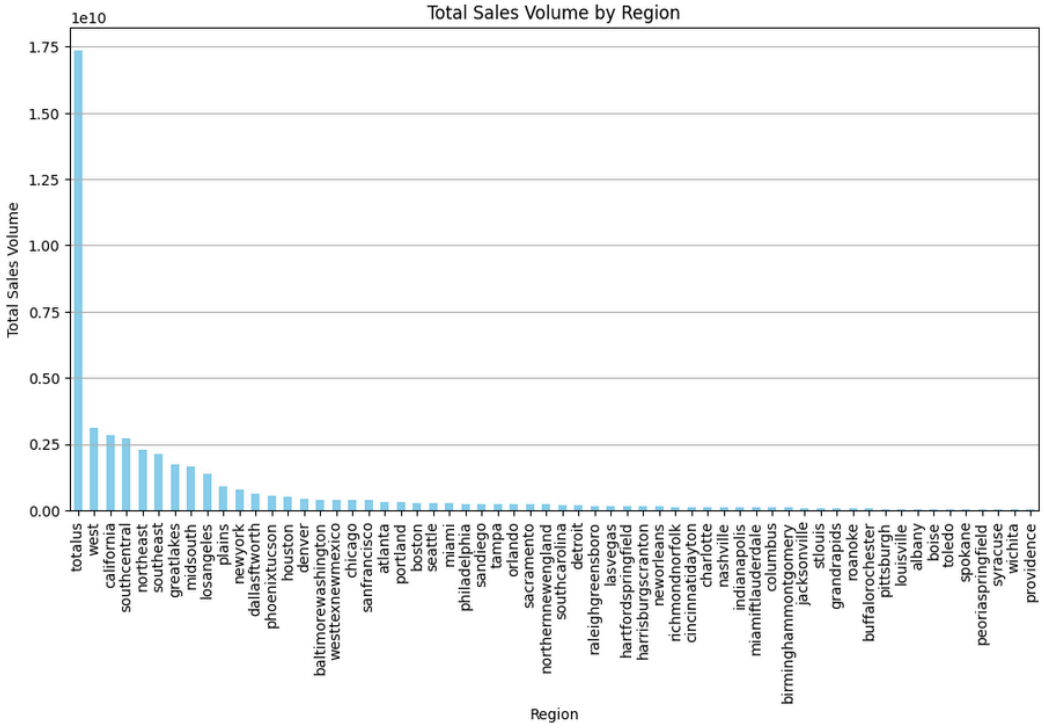
The EDA focused on understanding consumer preferences, market trends, and seasonal changes across different regions.

1. Distribution of Avocado Prices and Sales Volume



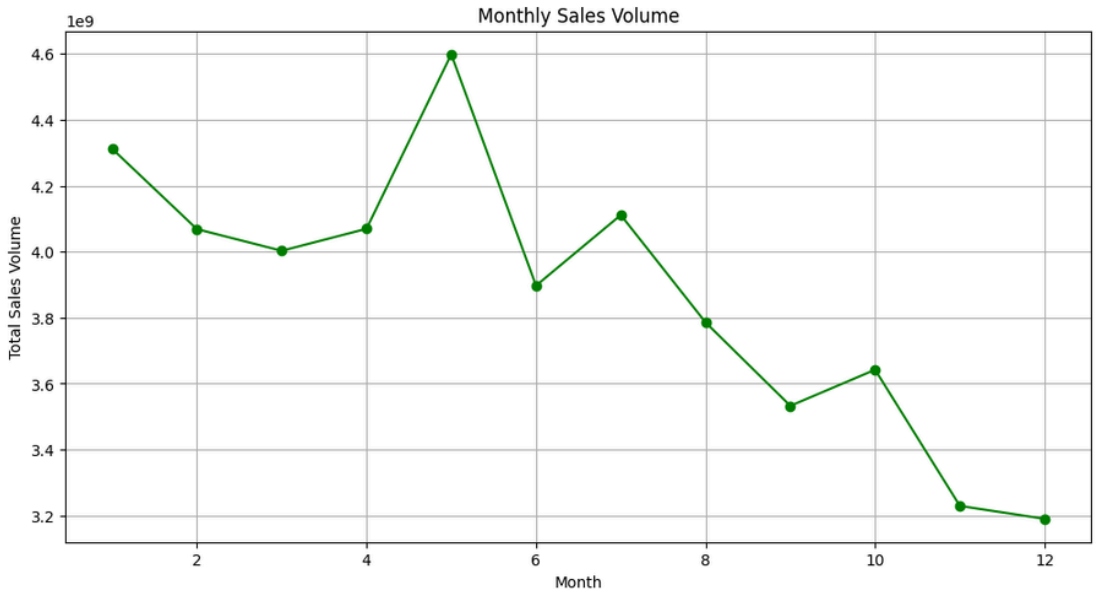
The distribution of avocado prices, specifically between conventional and organic varieties.

2. Regional Analysis



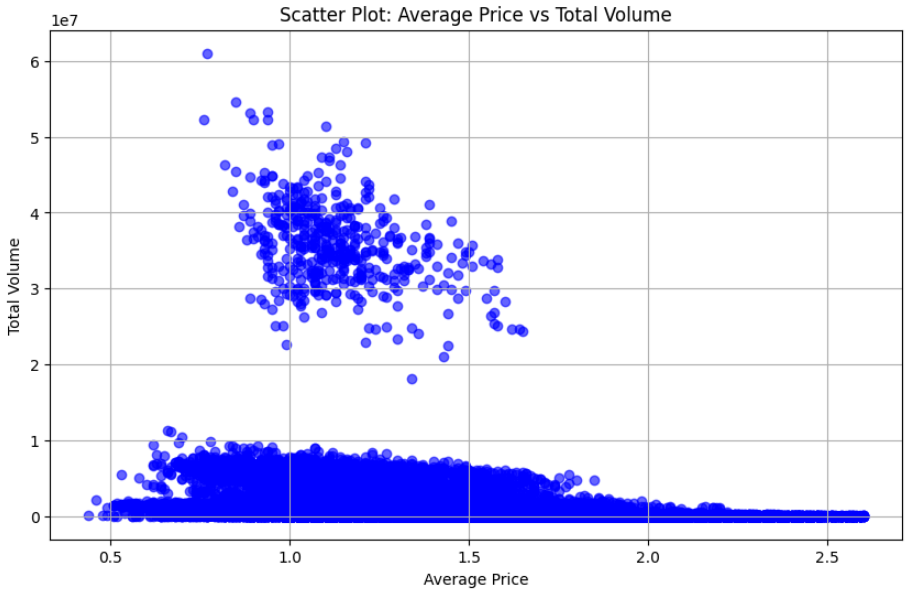
The bar chart reveals a significant disparity in sales volume among different regions, highlighting the need for targeted strategies to address market differences.

3. Seasonal Trends



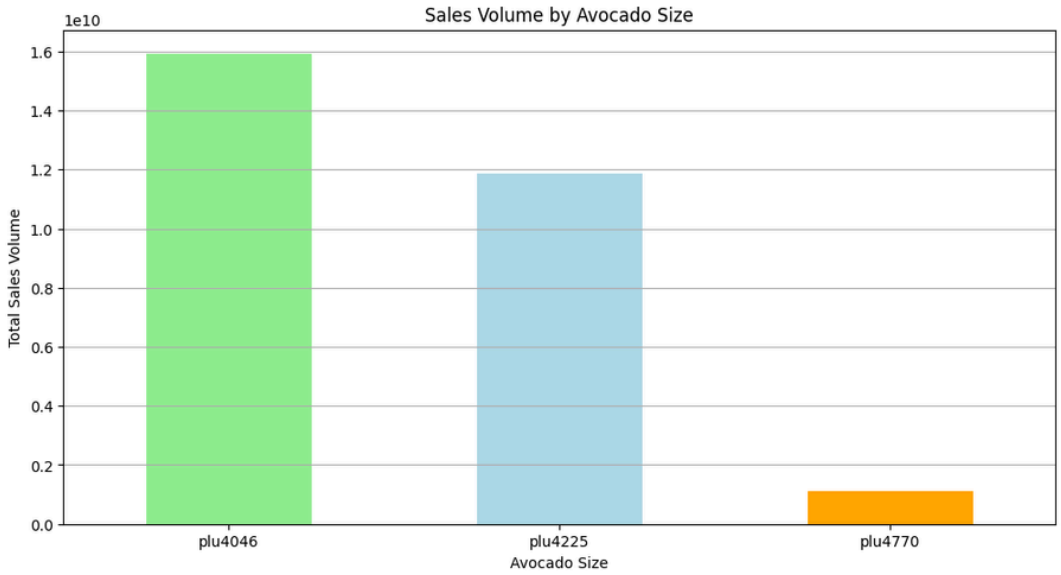
The line graph shows that monthly sales volume fluctuates significantly throughout the year.

4. Price Sensitivity



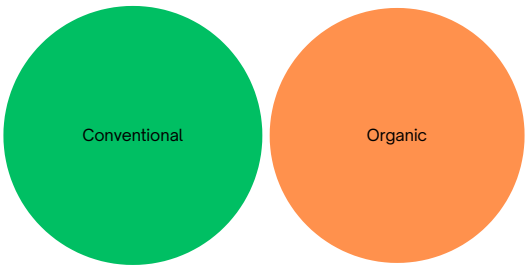
The scatter plot indicates a negative relationship between average price and total volume, suggesting that the product or service is sensitive to price changes.

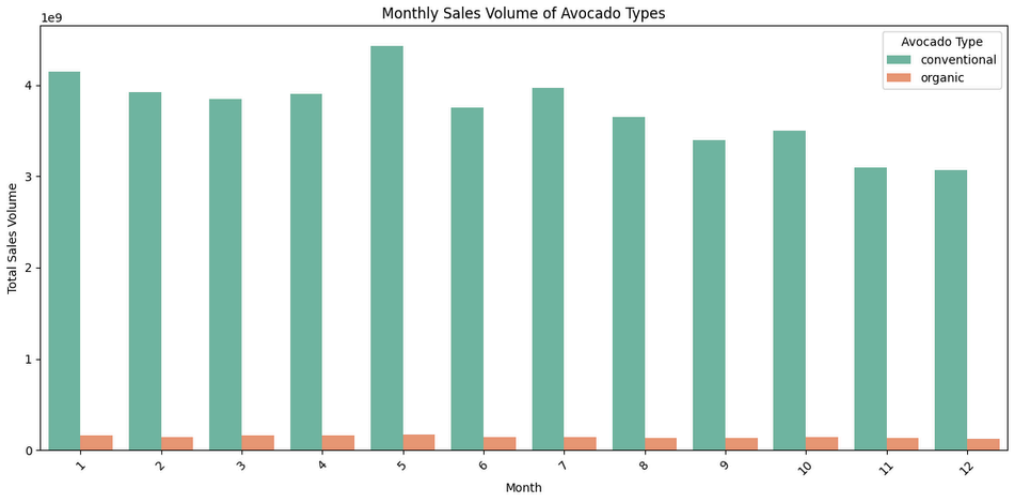
5. Product Packaging Preferences



The bar chart reveals that the sales volume for the avocado size plu4046 is significantly higher than that of plu4225 and plu4770, indicating a strong consumer preference for this size.

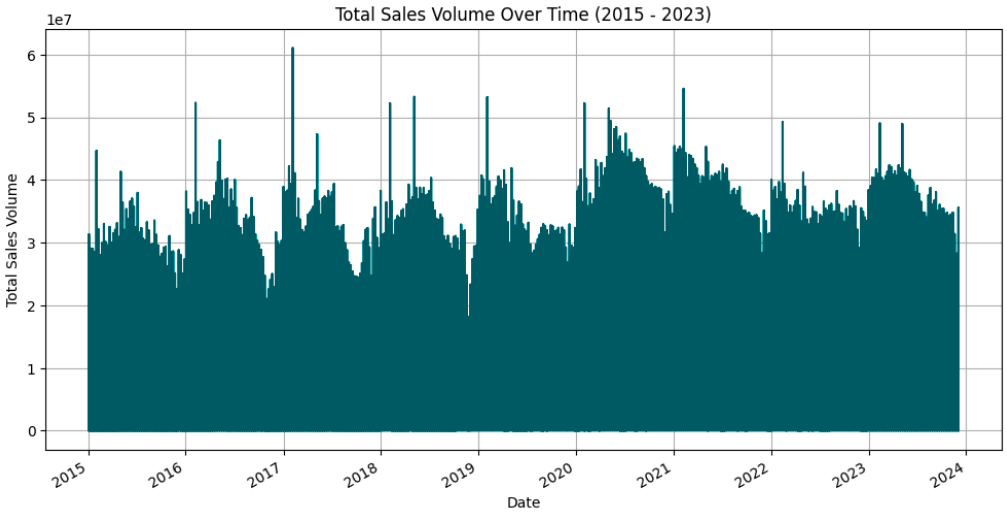
6. Consumer Preferences





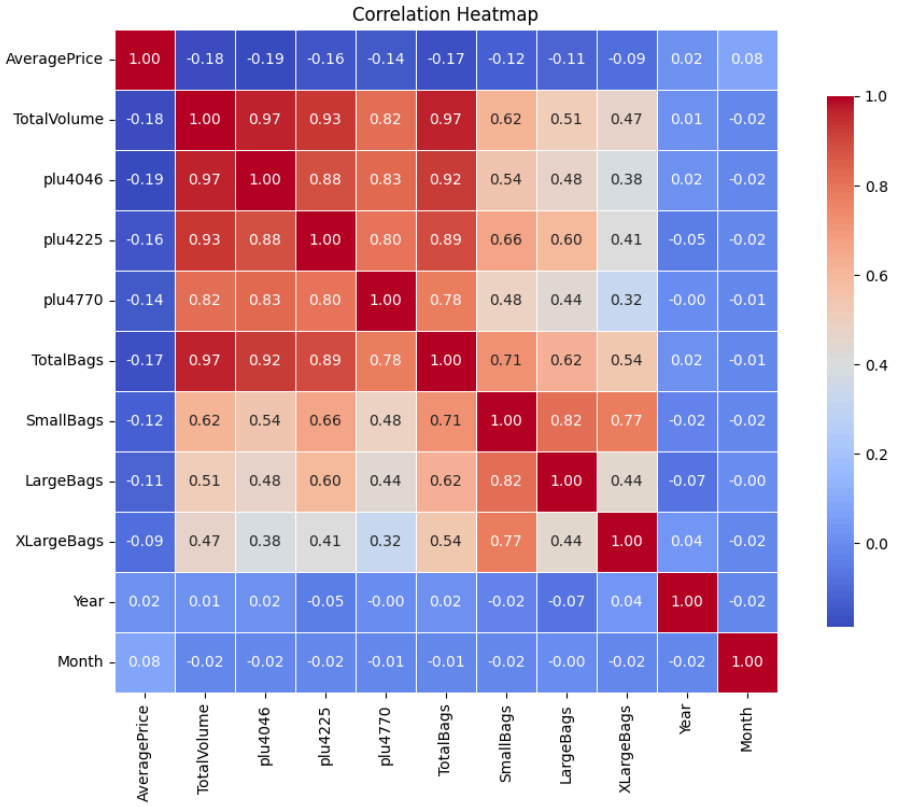
The charts above answer this question, which type is sold more and when do the sell more? They demonstrate that conventional and organic products have nearly equal market shares. This indicates diverse consumer preferences and significant growth potential for both segments.

7. Long-Term Trends



The total sales volume has generally increased over the past nine years, despite significant fluctuations.

7. Correlation Analysis



The correlation heatmap offers valuable insights into how different variables in the dataset relate to each other. The strong positive correlations between volume and size variables, along with the negative correlation between price and volume, suggest that businesses need to focus on their pricing and supply chain strategies.

Modelling Approaches

To predict avocado sales volumes, three different machine learning models were tested: **Linear Regression**, **Random Forest**, and **XGBoost**. The performance of each model was evaluated using Root Mean Squared Error (RMSE).

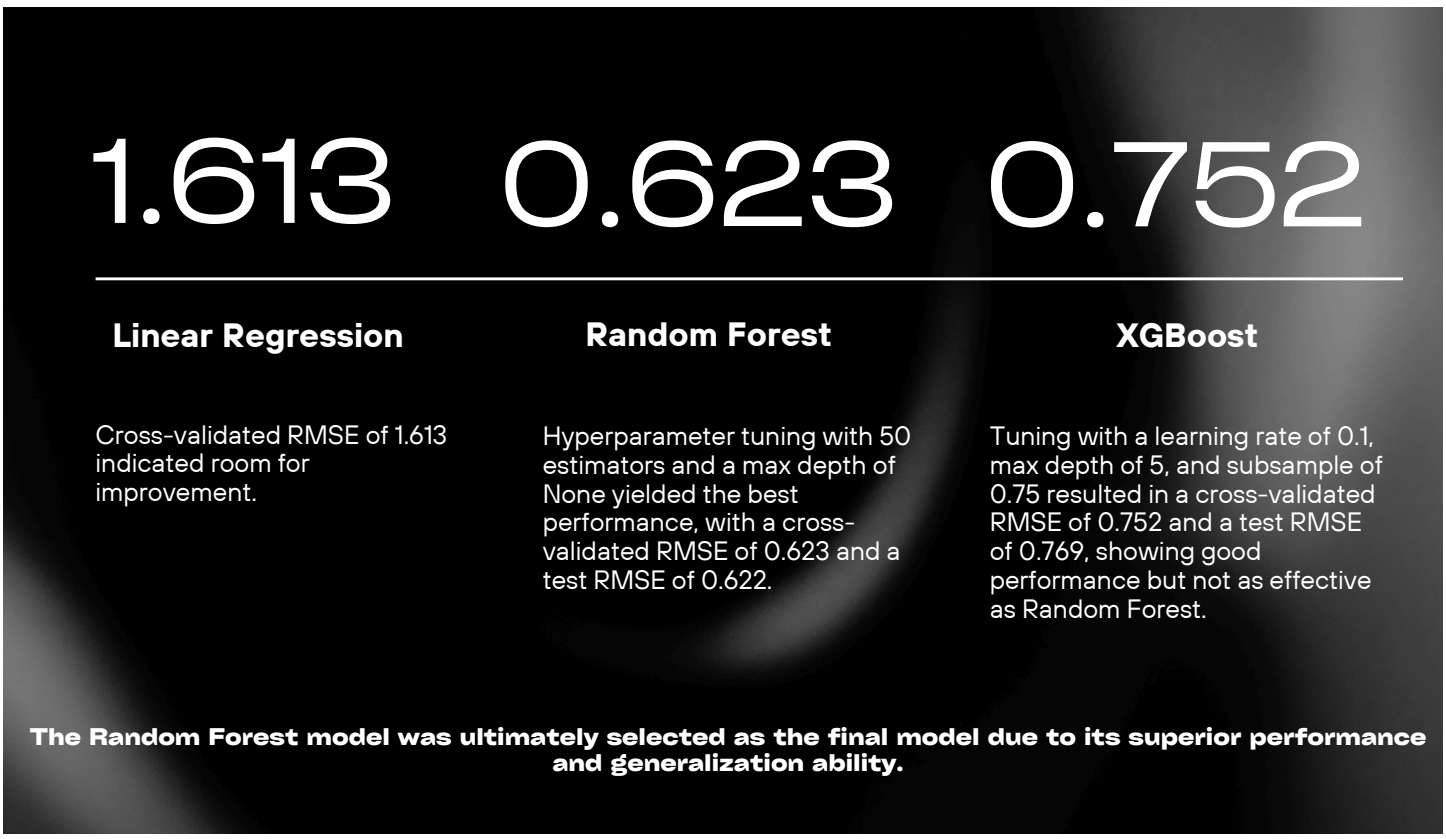
- **Linear Regression**: The baseline model achieved an RMSE of 1.59, providing reasonable accuracy but showing significant deviations from the actual sales volumes.
- **Random Forest**: This model significantly improved prediction accuracy, with an RMSE of 0.327, and produced predictions that closely followed actual sales trends.
- **XGBoost**: XGBoost achieved an RMSE of 0.402, demonstrating strong predictive performance but falling slightly behind Random Forest in terms of accuracy.

The Random Forest model stood out as the best-performing model due to its ability to capture complex relationships within the data, providing the most accurate predictions of avocado sales volumes.

Model Tuning and Validation

To further refine the models and improve their predictive capabilities, hyperparameter tuning and cross-validation were performed.

Results



Key Findings

The analysis uncovered several important insights into the avocado market:

1.	The U.S. market demonstrates strong avocado sales, particularly during warm-weather months and holiday seasons.
2.	Consumer preferences lean toward conventional avocados and smaller bag sizes, indicating opportunities for targeted marketing.
3.	Price sensitivity was weak, suggesting that small price increases are unlikely to cause significant drops in sales.
4.	Long-term sales patterns remained stable, with predictable fluctuations across seasons and years, offering valuable information for forecasting.

Conclusion

This project provided a comprehensive analysis of avocado pricing and sales trends in the U.S. market from 2015 to 2023. By employing EDA and predictive modeling techniques, key insights were derived about consumer preferences, seasonal trends, and regional differences. The Random Forest model proved to be the most effective at predicting sales volumes, thanks to its ability to handle complex data relationships.

Recommendations for Stakeholders

- **Adjust Marketing Strategies**: Tailor marketing efforts to align with seasonal peaks and regional preferences.
- **Pricing Strategies**: Implement competitive pricing to mitigate the effects of price sensitivity among consumers.
- **Product Offerings**: Focus on packaging options that meet consumer preferences, such as smaller bags and conventional avocados.

Next Steps

Moving forward, the model can be further optimized by exploring additional features and integrating alternative data sources, such as economic conditions and environmental factors. Future projects could also investigate the impact of these external factors on avocado sales, enriching the analysis and improving forecasting accuracy.

Resources

- [Kaggle dataset] (https://www.kaggle.com/datasets/vakhariapujan/avocado-prices-and-sales-volume-2015-2023/data)
- [UK Climate summaries] (https://www.metoffice.gov.uk/research/climate/maps-and-data/summaries/index)
- [NYC Data Science Academy] (https://nycdatascience.com/blog/student-works/exploring-avocado-data-and-building-predictive-models/)

GitHub- https://github.com/keneilweRangw/my_project

Email-patricia001105@gmail.com