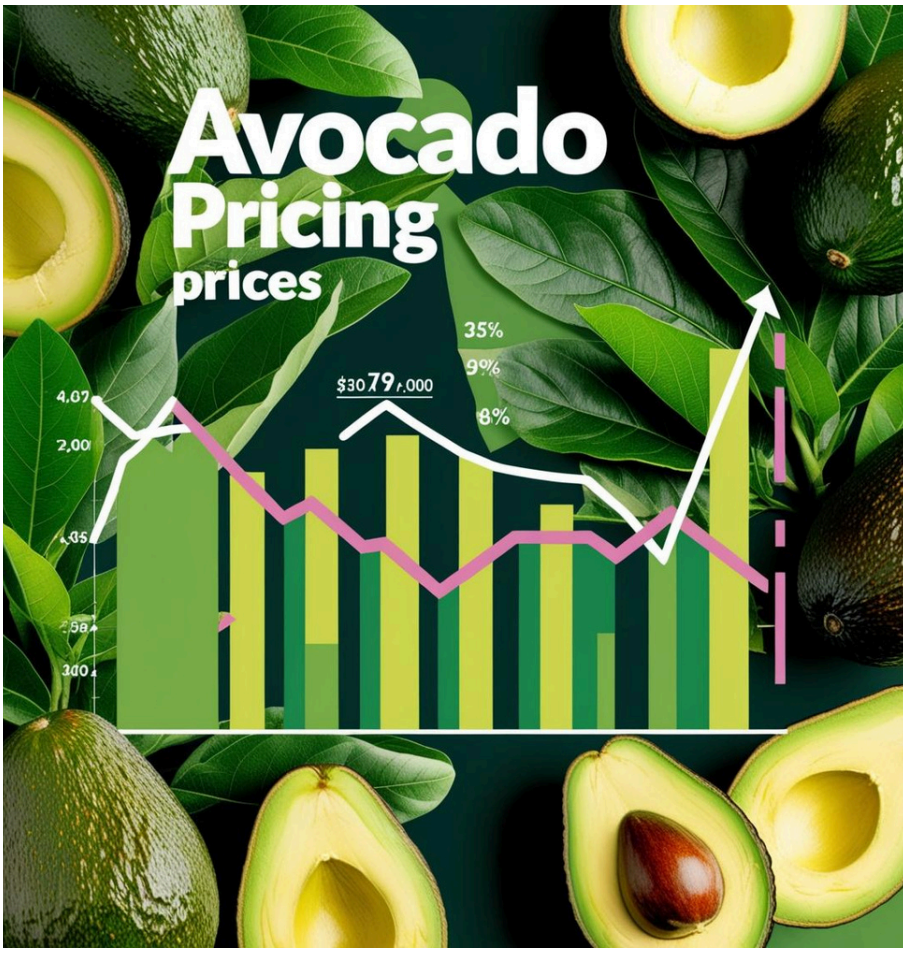
White Paper on Mastering Avocado Pricing

A Strategic Approach to Regional, Seasonal and Predictive Insights

By
Keneilwe Rangwaga



Abstract

This paper examines the evolution of the avocado market between 2015 and 2023, focusing on pricing, sales trends, and consumer preferences. Using data from the Hass Avocado Board, the analysis highlights how factors such as regional demand, seasonal changes, and overall market dynamics influence pricing and sales volumes. Machine learning models, including Random Forest and XGBoost, were employed to predict sales volumes. The Random Forest model demonstrated superior performance, providing actionable insights for optimizing pricing strategies, improving supply chains, and forecasting future trends.

Background of Study

Avocados have become a popular food item worldwide, from guacamole at parties to avocado toast at breakfast. As demand grows, the market has become more complex, with factors like regional preferences, seasonal trends, and prices affecting sales.

This study analyses avocado data from 2015 to 2023 to identify trends and offer actionable insights. The goal is to help stakeholders, farmers, retailers, and marketers improve pricing strategies, predict sales, and manage supply chains better.

Data Collection and Description

The dataset, sourced from the Hass Avocado Board, contains 53,415 records spanning 2015 to 2023. The data tracks weekly avocado sales across various U.S. regions and product types.

Key Features:

- **Date**: Weekly observations.
- **AveragePrice**: The average price of avocados.
- **TotalVolume**: The total volume of avocados sold.
- **Product Type**: Conventional or organic.
- **Region**: Sales data from 60 unique regions.
- **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.

Data Cleaning and Preparation

To ensure the dataset was suitable for analysis, several preparation steps were taken:

1. **Handling Missing Data**: Missing values were imputed or removed as necessary.
2. **Standardizing Text Data**: Columns such as “Type” and “Region” were standardized for consistency.
3. **Removing Outliers**: Extreme values in pricing and sales were identified and excluded to avoid skewing results.
4. **Feature Engineering**: Additional columns, such as Year, Month, and Price Per Unit Volume, were created to uncover deeper insights.

Key Insights from Data Analysis

1. **Seasonal Trends**

- Sales increase during warmer months and holidays. Supported by: *Figure 3 (Seasonal Trends)*, which shows significant fluctuations in monthly sales volumes throughout the year.

2. **Regional Differences**

- Sales vary by region, with some areas selling significantly more. Supported by: *Figure 2 (Regional Analysis)*, which highlights disparities in sales volume across regions, with the West, California, and South Central leading in sales.

3. **Price Sensitivity**

- Higher prices slightly reduce sales, but the impact is small. Supported by: *Figure 4 (Price Sensitivity)*, where a scatter plot demonstrates a negative relationship between average price and total sales volume.

4. **Consumer Preferences**

- Smaller avocados (PLU 4046) are the most popular, and conventional avocados sell more than organic ones. Supported by:
  - *Figure 5 (Product Packaging Preferences)*, which shows PLU 4046 significantly outperforms other sizes in sales.
  - *Figure 6 (Consumer Preferences)*, which demonstrates that both conventional and organic products hold nearly equal market shares.

5. **Long-Term Growth**

- Sales have steadily grown over nine years, despite some fluctuations. Supported by: *Figure 7 (Long-Term Trends)*, which illustrates a consistent increase in total sales volume over the past nine years.

Predictive Modeling

Three machine learning models were employed to predict avocado sales volumes:

1. **Linear Regression**:
   - Served as the baseline model with limited accuracy (RMSE: 1.613).
2. **Random Forest**:
   - Delivered the most accurate predictions (RMSE: 0.623), effectively capturing non-linear relationships.
3. **XGBoost**:
   - Achieved strong performance (RMSE: 0.752), though slightly less accurate than Random Forest.

Model Tuning and Validation

Hyperparameter tuning and cross-validation were applied to improve model performance. The Random Forest model, optimized with 50 estimators and no maximum depth, provided the best results.

Why Random Forest Worked Best:

The Random Forest model handles complex relationships in the data, making it better at predicting sales than simpler models.

NB//Have a look at Appendix B for the results

Recommendations

- **Adjust Marketing Strategies**: Tailor marketing efforts to align with seasonal peaks and regional preferences and focus on peak sales seasons and regions with high demand.
- **Pricing Strategies**: Small price increases won't hurt sales much, so adjust prices strategically and implement competitive pricing to mitigate the effects of price sensitivity among consumers.
- **Product Offerings**: Focus more on packaging options that meet consumer preferences, such as smaller bags and conventional avocados.

Next Steps

To improve the analysis, future work can:

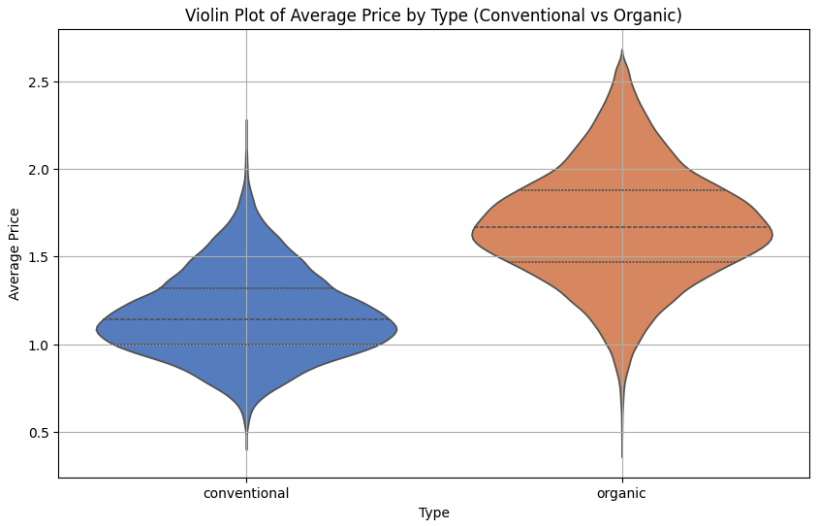
1. Add more data, like climate and economic factors, to make predictions more accurate.
2. Expand the study to include international markets.
3. Investigate how global issues, like climate change, affect avocado sales.

Conclusion

This study provides insights into how avocados are sold and priced across the U.S. from 2015 to 2023. The Random Forest model stands out as the best tool for predicting sales, helping stakeholders make better decisions. By understanding consumer preferences, regional differences, and seasonal trends, producers and retailers can improve their strategies and succeed in the avocado market.

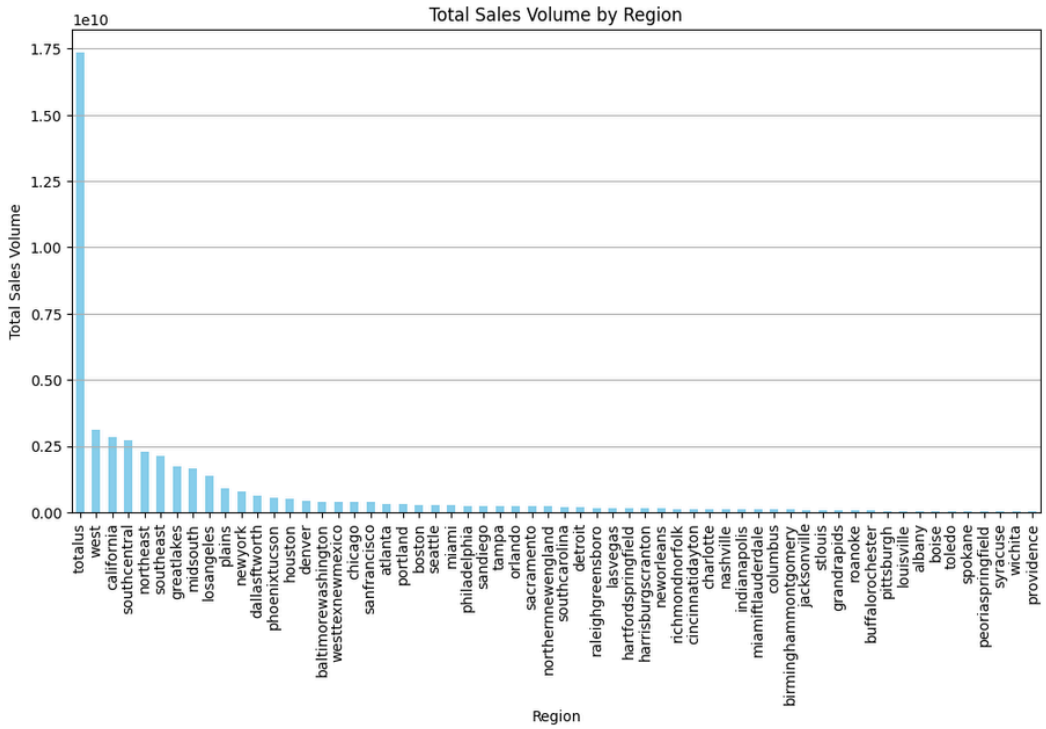
Appendix A- Supplementary charts

Figure 1: Distribution of Avocado Prices and Sales Volume



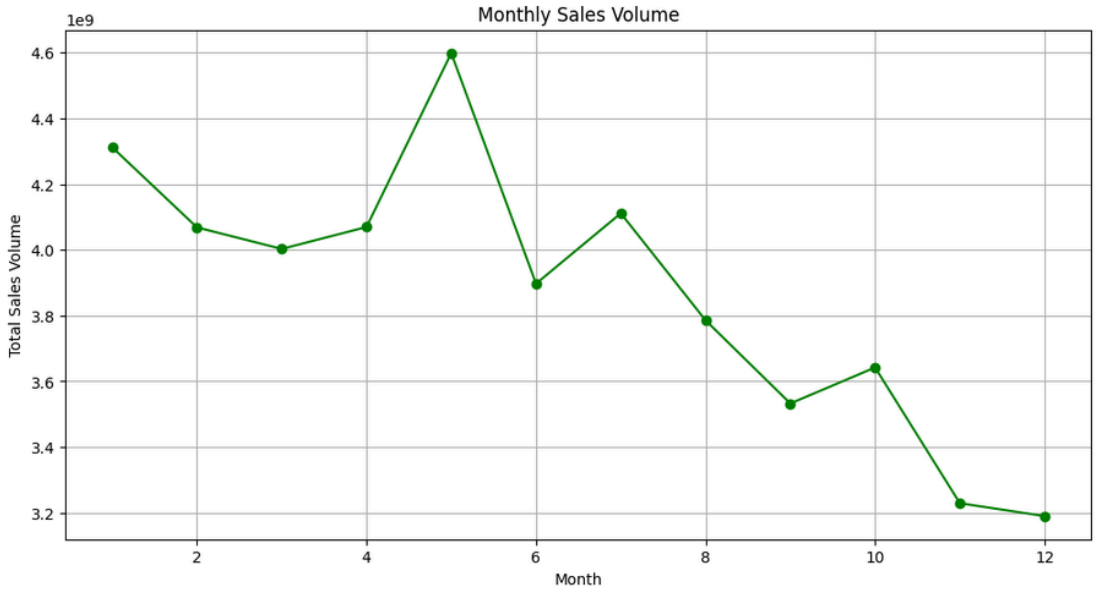
The distribution of avocado prices, specifically between conventional and organic varieties. Even though there is no big difference, this reveals that the organic type is the most preferred.

Figure 2: Regional Analysis



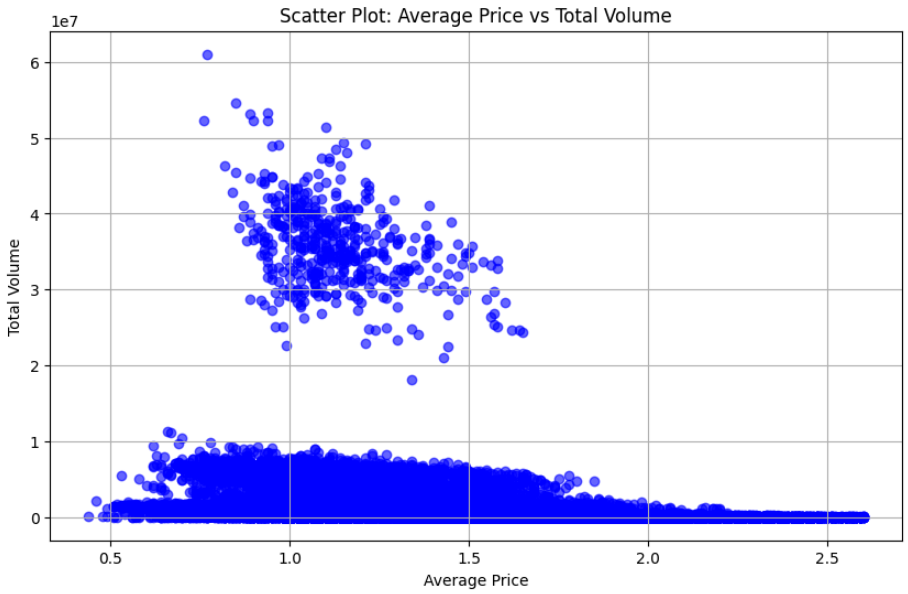
The bar chart reveals a significant disparity in sales volume among different regions, highlighting the need for targeted strategies to address market differences. The West, California and South Central are the top three countries with high sales while Providence, Wichita and Syracuse are the top three lowest

Figure 3: Seasonal Trends



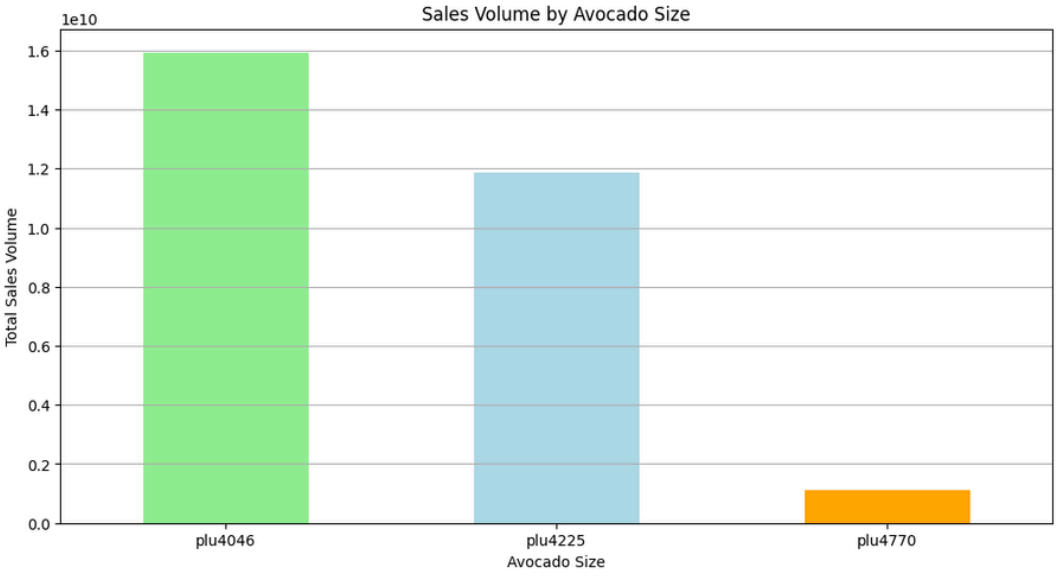
The line graph shows that monthly sales volume fluctuates significantly throughout the year with May being the highest.

Figure 4: Price Sensitivity



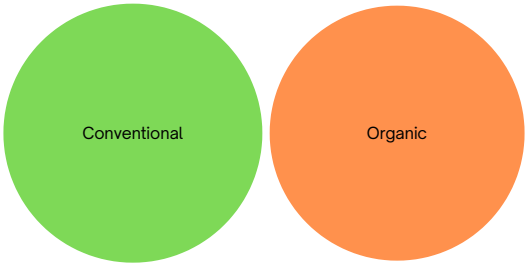
The scatter plot indicates a negative relationship between average price and total volume, suggesting that the product or service is sensitive to price changes.

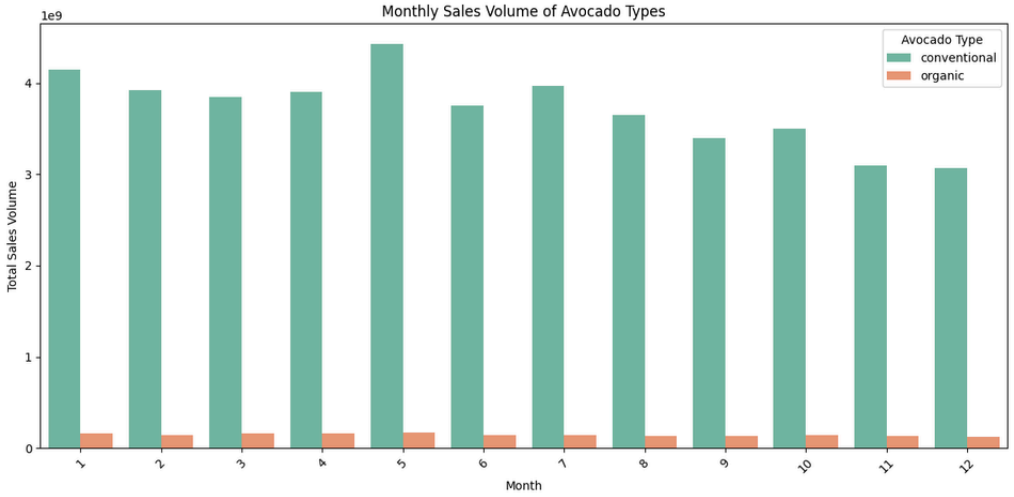
Figure 5: Product Packaging Preferences



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.

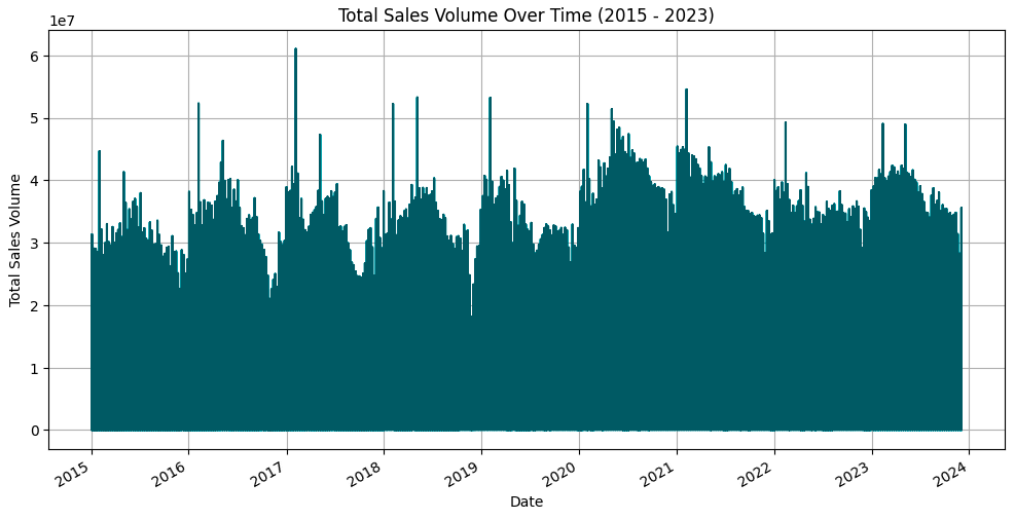
Figure 6: Consumer Preferences





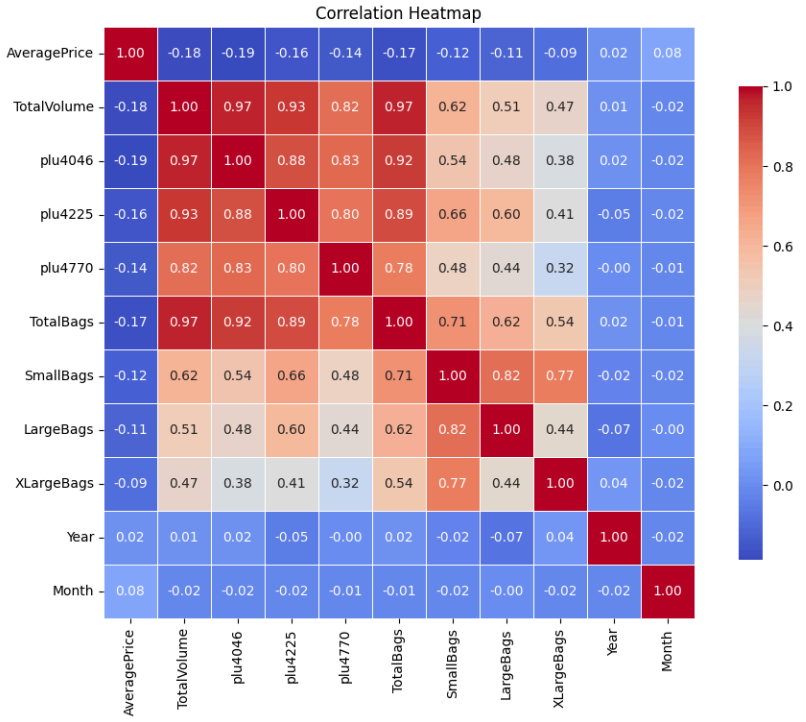
The charts above answer this question, "which type is sold more and when do the sell more?" They demonstrate that conventional and organic products have nearly equal market shares. This indicates diverse consumer preferences and significant growth potential for both segments.

Figure 7: Long-Term Trends



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

Figure 8: 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.

Appendix B- Tuning Results

1.613	0.623	0.752
Linear Regression	**Random Forest**	**XGBoost**
Cross-validated RMSE of 1.613 indicated room for improvement.	Hyperparameter tuning with 50 estimators and a max depth of None yielded the best performance, with a cross-validated RMSE of 0.623 and a test RMSE of 0.622.	Tuning with a learning rate of 0.1, max depth of 5, and subsample of 0.75 resulted in a cross-validated RMSE of 0.752 and a test RMSE of 0.769, showing good performance but not as effective as Random Forest.

The Random Forest model was ultimately selected as the final model due to its superior performance and generalization ability.

Appendix C-Resources

- Hass Avocado Dataset
- NYC Data Science Academy
- UK Climate Summaries

Author: Keneilwe Rangwaga

GitHub: Project Repository

Email: patricia001105@gmail.com