Modeling Price Change of Seasonal Fruits-Flavored Food Products: A Predictive Analysis*

The Price of Strawberry and Banana-Flavored Products will Increase after November, with Strawberry Flavored Products Rising More in Price

Yiyi Feng

December 1, 2024

This study develops a model to predict price changes for seasonal fruit-flavored products during in-season and off-season periods. It forecasts price increases for strawberry and banana-flavored products after November, with a larger rise for strawberry. Vendor and month are key factors in price trends, but the complexity of price changes calls for further investigation. Due to the complex dynamics of seasonal fruit-flavored product price changes, improved data collection is essential for more accurate forecasting.

Table of contents

*Code and data are available at: https://github.com/kqlqkqlqF/Modeling-Price-Fluctuations-of-Seasonal-Fruits-in-Food-Products.git.

1 Introduction

As a fruit enthusiast, I've noticed seasonal fruits show significant price changes between in-season and off-season periods. This raises the question: do seasonal fruit-flavored products follow similar patterns, and what drives these changes? What are the key factors driving these price changes? This study uses Canadian grocery price data from eight major vendors to build a predictive model analyzing seasonal fruit-flavored product pricing and identifying key factors influencing price dynamics.

Our estimand focuses on the price changes in seasonal fruit-flavored products during in-season and off-season periods, focusing on strawberries and bananas due to their popularity as summer

and winter fruits (Watson 2019) (Reeves 2022). These choices ensure a sufficiently large sample size for analysis. Other seasonal fruits, such as watermelons, were excluded due to their relatively smaller consumer base in North America compared to Asia.

We used a linear model to estimate price changes for strawberry and banana-flavored products, including predictors like vendor, month, rainfall, food category, and their interactions. The goal was to understand how these factors affect price fluctuations and seasonal pricing patterns.

Our model predicts small price increases for both banana and strawberry-flavored products, with average monthly increases of 0.0085 and 0.0103, respectively. Monthly trends show stability in July, slight decreases in August and September, and increases in October and November. The most significant factors influencing price changes are vendor, month, and the interaction between month and food category, emphasizing the important roles of vendors, timing, and product category in price fluctuations.

The remainder of this paper is structured as follows: Section 2 provides an overview of the dataset, details of the parameters, outcome and predictor variables, and the packages used during processing. Section 3 explains the modeling approach, and best model selection, justifying the choice of predictors and outlining the methods used to forecast the price change of banana and strawberry-flavored products. Section 4 presents the findings, including a summary of the predicted price change for each flavored product, and figures demonstrating the price change distribution across month and food categories. In Section 5, we discuss the implications of these results, the limitations of our analysis, and what can we do next to improve our model. Additional methodological details and diagnostics are included in the appendix.

2 Data

2.1 Overview

We used R (R Core Team 2023) to analyze Canadian grocery price data. The dataset, from Jacob Filipp's Project Hammer (Filipp 2024), tracks price changes from eight vendors: Voila, T&T, Loblaws, No Frills, Metro, Galleria, Walmart, and Save-On-Foods, from February 28, 2024 to November 26, 2024. To examine factors influencing the price of seasonal fruit-flavored foods, we included data on vendor, rainfall, month, and food category, along with rainfall data from the "About Rain Gauge Locations and Precipitation" dataset for opendataToronto (Shelter and Services 2021).

Several R packages were essential for our data manipulation, modeling, and visualization efforts. dplyr facilitated data transformation and summarization (Wickham et al. 2023), while modelsummary enhanced the presentation of model outputs (Arel-Bundock 2023). kableExtra improved data presentation with customizable tables (Yu 2023), and testthat ensured code reliability (Wickham and RStudio 2020). Lubridate and stringr simplified date-time handling and

string manipulation (Grolemund and Wickham 2020) (Hadley Wickham 2020). For efficient data storage, tibble and arrow enhanced our workflow with memory-efficient data frames and fast data reading/writing (Müller and Wickham 2022) (Richardson et al. 2024). Caret and randomForest supported machine learning, and rstanarm provided Bayesian regression tools (Kuhn et al. 2023) (Liaw and Wiener 2023) (Goodrich et al. 2022). Finally, ggplot2 powered visualizations, and knitr streamlined reproducible reporting (Kassambara 2023) (Xie 2023). This robust package ecosystem aligned with best practices, as noted in (Alexander 2023).

2.2 Measurement

This section outlines how we transformed the raw Canadian grocery price data into a structured dataset for analysis. We focused on seasonal fruit-flavored foods, selecting bananas and strawberries as representatives. Bananas were chosen for their popularity in winter, and strawberries for spring. These fruits have a larger consumer base in Canada compared to other seasonal fruits like watermelon and pomegranate, making data collection for model building more feasible.

The original Canadian grocery price data was collected by Jacob Filipp through screen-scraping the website interfaces of eight vendors and compiled into the **Project Hammer** dataset. This dataset contains two tables: **Hammer 4 Raw** and **Hammer 4 Product**. **Hammer 4 Raw** includes the scraping time, product name, single-item price, unit price (e.g., price per 100g), past prices, additional information (e.g., availability or discounts), and a unique product ID. **Hammer 4 Product** provides detailed information, such as brand, vendor, sales unit, and product detail page links. Since the data comes from website scraping, it contains many gaps and ambiguities, making cleaning difficult. Details about data collection and cleaning are provided in Appendix A.

We considered sources such as Statistics Canada, the Retail Council of Canada, and Open Data Portals, but after reviewing their datasets, we found they did not meet our study's requirements. These datasets focus on broad food categories, not specific flavored products like strawberries or bananas. Additionally, they lack the detailed time scale and vendor-level data necessary for tracking monthly price changes, which are crucial for modeling seasonal price shifts. As a result, we determined these sources were not suitable for our research.

To better analyze price changes for banana and strawberry-flavored foods, we added data from the **"About Rain Gauge Locations and Precipitation"** dataset from Open Data Toronto (Shelter and Services 2021). This dataset includes rain gauge locations across Toronto and recorded precipitation. Rainfall was included as a predictor since consumer demand for seasonal fruits often correlates with weather conditions, such as a preference for watermelon and strawberries during dry summers (Watson 2019). The dataset only uses rainfall data because the original grocery price data lacks sales location details.

This structured dataset enables us to analyze trends in seasonal fruit-flavored food prices over time. The goal is to examine price trends for these foods during in-season and off-season sales and identify key influencing factors.

2.3 Outcome Variables

2.3.1 Change of Monthly Averaged Price for Banana and Strawberry Flavored Product

Figure 1 shows the monthly average price distribution of banana and strawberry-flavored products. Most products are priced between 0 and 10 dollars, with few exceeding this range and prices above 20 dollars being rare. This indicates that the majority of sales fall within the lower price range. The price distributions of both flavors are similar, though strawberry-flavored products have a slightly smaller share in the 0 to 10-dollar range compared to banana-flavored products.

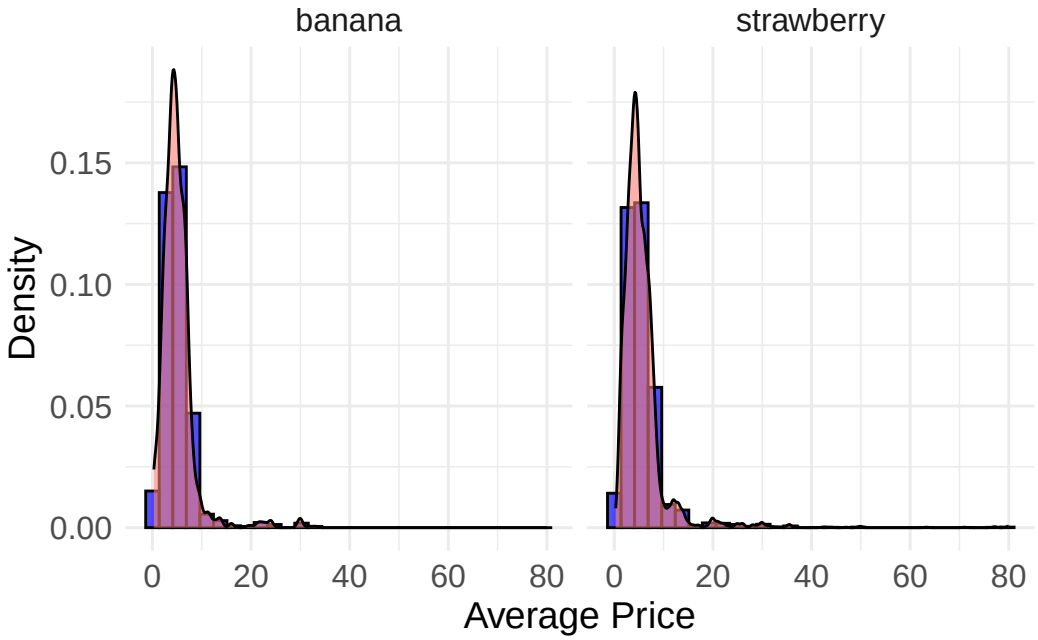

Figure 1: Distribution of Monthly Averaged Price for Banana and Strawberry Flavored Product

Figure 2 shows the monthly price changes for banana and strawberry-flavored products. Strawberry-flavored products exhibit greater fluctuations, ranging within ± 2.5 dollars, compared to banana-flavored products, which vary within ± 1.5 dollars. Neither flavor shows a clear overall trend of increasing or decreasing prices across the months.

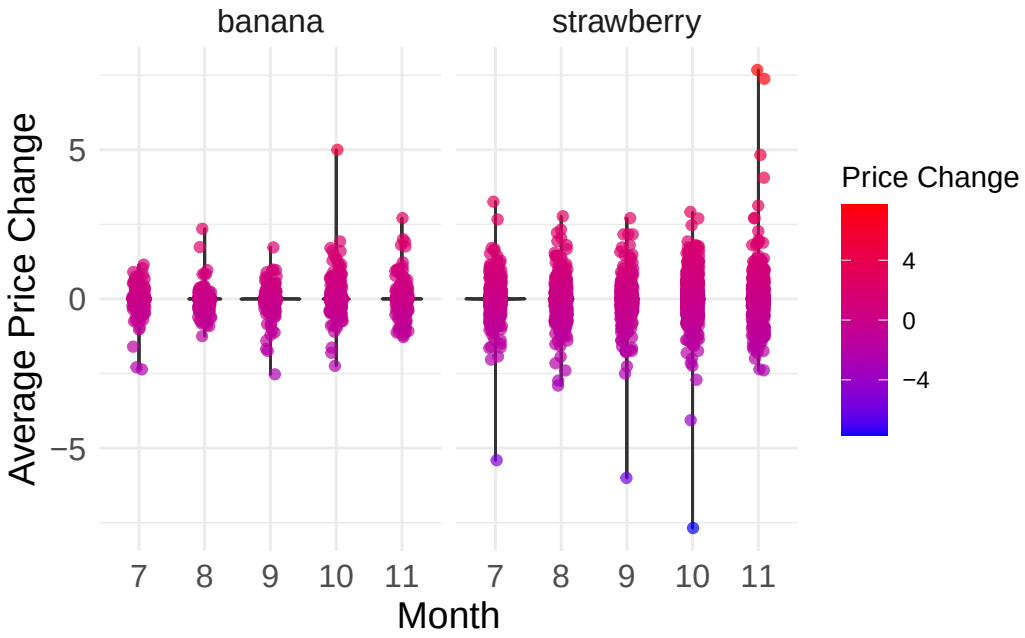

Figure 2: Distribution of Monthly Averaged Price Change for Banana and Strawberry Flavored Product from July 2024 to Novermber 2024

This outcome challenges our assumption. We expected banana-flavored product prices to rise between July and November due to their popularity in fall and winter and strawberry-flavored prices to decrease during this period as a summer flavor.

2.4 Predictor Variables

2.4.1 Summary of Predictor Variables

- **Vendor:** Includes Canada's eight major suppliers: Voila, T&T, Loblaws, No Frills, Metro, Galleria, Walmart, and Save-On-Foods.
- **Average Rainfall Per Month:** Monthly average rainfall for Toronto from July to November 2024, measured in millimeters, based on the "About Rain Gauge Locations and Precipitation" dataset.
- **Food Category:** Banana and strawberry-flavored products are divided into five categories: beverage, yogurt, flavored tea, solid snack, and fruit. "Beverage" includes all drinks except tea and yogurt-based beverages, "flavored tea" covers tea drinks and flavored tea bags, and "fruit" includes various forms of banana or strawberry fruit.

- **Month:** Indicates the month when the price data for each product was collected, covering July to November.

2.4.2 Vendor

Based on Figure 3, strawberry-flavored products are significantly more common than banana-flavored ones, with the dataset showing 3-4 times as many strawberry-flavored items. This suggests that strawberry flavors appeal to a larger audience. Save-On-Foods has the smallest share for both flavors, which may reflect a narrower focus on fruit-flavored products or a smaller overall selection of food types.

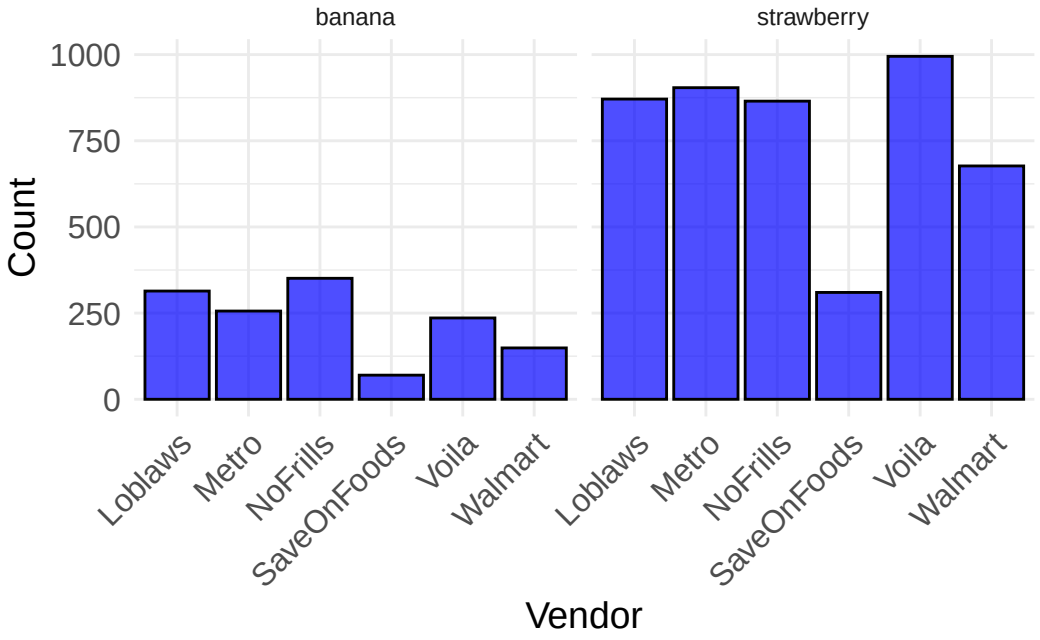

Figure 3: Distribution of the Counts of Banana and Strawberry Flavored Food Offered in Each Vendor

2.4.3 Average Rainfall Per Month

Figure 4 displays the average monthly rainfall in Toronto, measured in millimeters. Rainfall decreases steadily from July to October, with October having the lowest levels, followed by a rebound in November. Rainfall was included as a predictor due to its strong correlation with time and seasonal changes, making it a potentially valuable factor in the model.

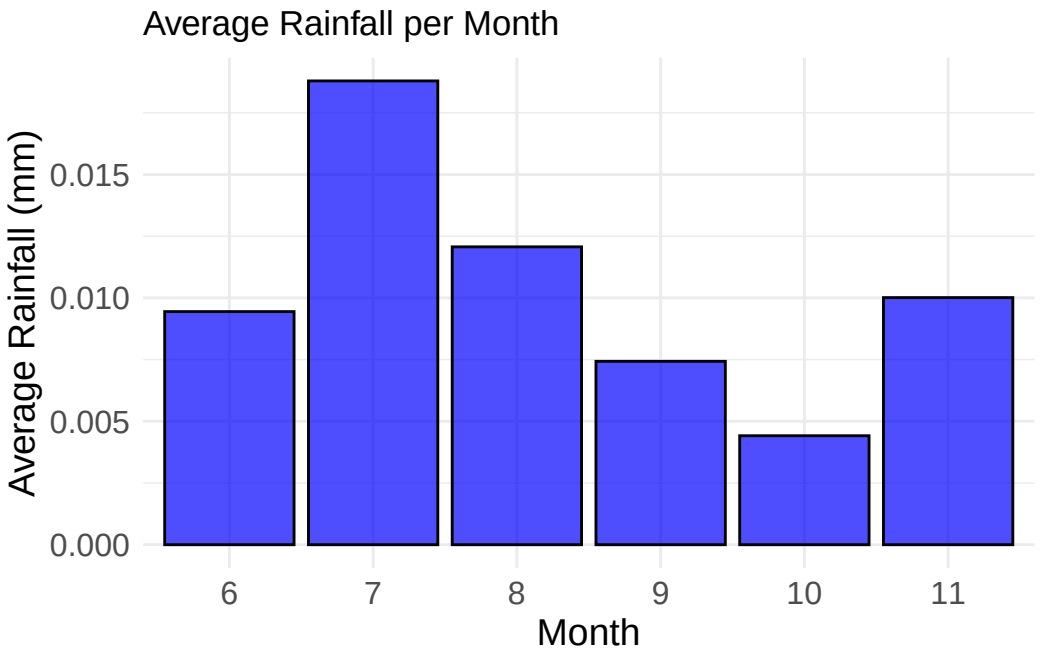

Figure 4: Average Rainfall in Toronto from July 2024 to November 2024

2.4.4 Food Category

In Figure 5, we can see that solid snacks have the highest overall prices, while flavored tea has the lowest overall prices. Additionally, the price distribution for solid snacks, fruit, and beverages is relatively dispersed, whereas yogurt and flavored tea show more tightly clustered price distributions. We believe this is because flavored tea and yogurt are more concentrated categories compared to the others.

2.4.5 Month

Figure 6 shows the distribution of the Monthly average price for Banana and strawberry-flavored products over time. It is clear that strawberry-flavored products are priced higher than banana-flavored products in all months. However, neither flavor shows a noticeable increasing or decreasing trend in price distribution across the months.

3 Model

The goal of this section is to build a predictive model for price changes in banana and strawberry-flavored products. The main challenge is to create a model that captures price

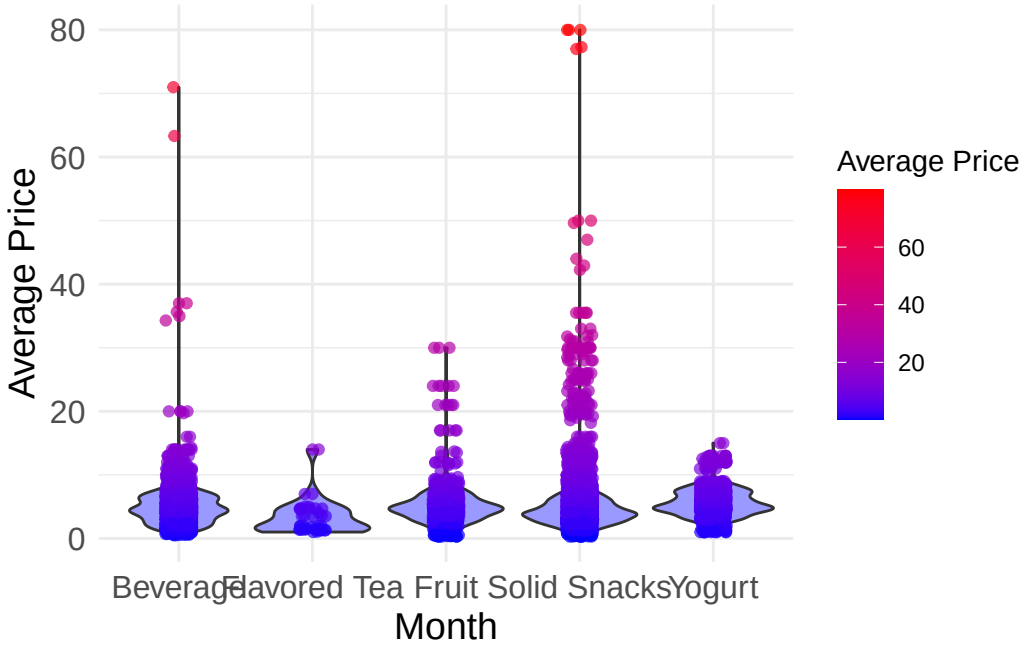

Figure 5: Distribution of Monthly Averaged Price for Banana and Strawberry Flavored Product within Category

fluctuations accurately, despite the limited sample size and temporal scope of the data. To address this, we evaluated different model specifications to identify the best approach for achieving reliable predictions.

We decided to model the monthly average unit price changes for banana and strawberry-flavored products, instead of predicting more specific metrics like price per 100g. This choice was made due to limitations in the dataset, which includes various measurement units (e.g., per 100g, per 100 ml, per kg, per lb) and the fact that many products lack detailed measurement information. While this approach may slightly reduce the model's performance, it allowed us to focus on unit prices.

We included several predictors in the model, such as Vendor, Food Category, Average Rainfall, and Month, and considered potential interactions among these variables. For example, rainfall and month are often closely correlated, and food categories might interact with both month and vendor. We developed a series of linear models and systematically compared them, testing different combinations of interaction terms to identify the model that best predicts price changes.

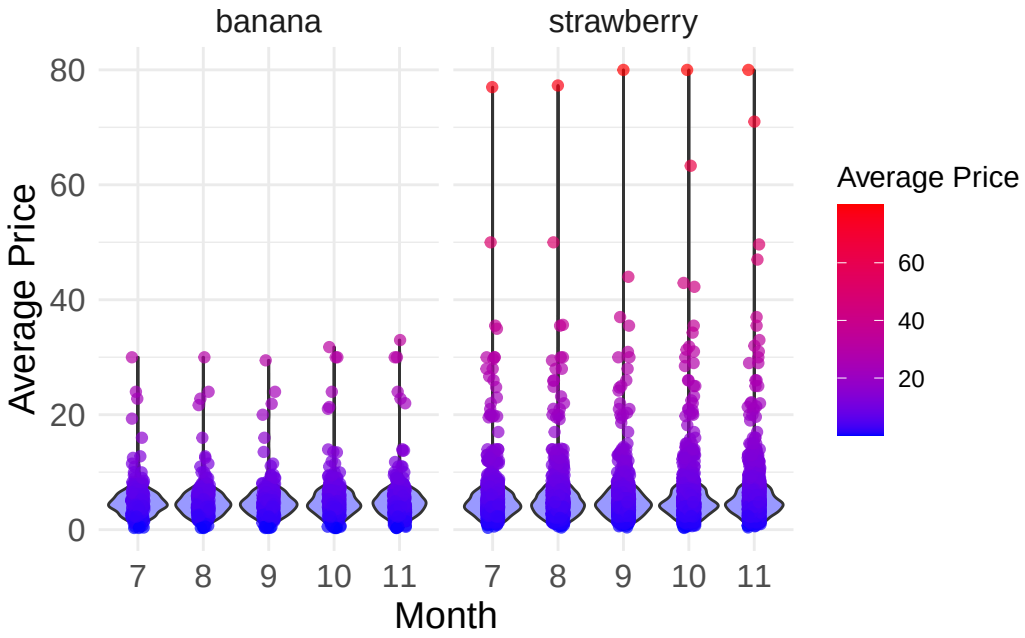

Figure 6: Distribution of Monthly Averaged Price for Banana and Strawberry Flavored Product from July 2024 to Novermber 2024

3.1 Model Set-up

Our goal is to model the price changes of strawberry and banana-flavored products using Vendor, Food Category, Average Rainfall, and Month as predictors. The model includes interaction terms to capture how combinations of these factors influence price changes, offering insights into the key factors that drive the pricing of seasonal fruit-flavored products.

$$\begin{aligned}\text{Pct}_i = \beta_0 &+ \beta_1 \cdot \text{Vendor}_i + \beta_2 \cdot \text{Category}_i + \beta_3 \cdot \text{Month}_i + \beta_4 \cdot \text{AverageRainfall}_i \\ &+ \beta_5 \cdot (\text{Vendor}_i \times \text{Category}_i) + \beta_6 \cdot (\text{Category}_i \times \text{Month}_i) \\ &+ \beta_7 \cdot (\text{Month}_i \times \text{AverageRainfall}_i) + \beta_8 \cdot (\text{Category}_i \times \text{AverageRainfall}_i) + \epsilon_i\end{aligned}$$

Where

- y_i : The percentage of support for candidate in poll i.

- β_0: Intercept term, representing the predicted `Average price of the Product` when all independent variables are 0.
- β_1: Main effect of `Vendor`, capturing the influence of the candidate.
- β_2: Main effect of `Category`, reflecting the influence of how recent the poll is on `Average price of the Product`.
- β_3: Main effect of `Month`, indicating the impact of different states on `Average price of the Product`.
- β_4: Main effect of `Average Rainfall`, indicating the impact of different states on `Average price of the Product`.
- β_5: Interaction effect between `Vendor` and `Category`, capturing the combined influence of the candidate and state.
- β_6: Interaction effect between `Category` and `Month`, reflecting the joint impact of recency and state on `pct`.
- β_7: Interaction effect between `Month` and `Average rainfall`, representing the combined influence of the candidate and recency of the poll.
- β_8: Interaction effect between `Category` and `Average rainfall`, representing the combined influence of the candidate and recency of the poll.
- ϵ_i: The error term, assumed to follow a normal distribution with mean 0.

3.1.1 Model Interpretation

This regression model predicts price changes for strawberry and banana-flavored products by incorporating both main effects and interaction terms. The intercept represents the baseline price change when all predictors are zero. Main effects, such as vendor, average rainfall, food category, and month, capture the independent influence of each factor on price. For example, strawberry-flavored products, typically associated with summer, may see price drops as the months move into winter.

To capture more complex relationships, the model includes two-way interactions, such as vendor and category, category and month, and month and average rainfall. These interactions help explain how factors influence each other. For instance, some vendors may focus on specific products, like beverages, while the category-month interaction reflects seasonal demand shifts, such as reduced beverage demand in Canada's colder months. The rainfall-month interaction shows their direct relationship, with June being the wettest month and November the driest in Canada.

An error term is included to account for unexplained variability, improving the model's predictive accuracy. By considering both individual and interactive effects, the model provides a comprehensive approach to forecasting price changes for these products.

3.2 Model Justification

Table 1: Linear Model Summary with Included Variables and Interactions

Model	Variables	R^2	Adjusted R^2	AIC	BIC	RMSE
Model 1	Vendor, Category, Month, Average Rainfall	0.0038	0.0017	8970.228	9070.715	0.5098
Model 2	Vendor, Category × Month, Average Rainfall	0.0089	0.0041	8971.611	9179.285	0.5085
Model 3	Vendor, Category, Month × Average Rainfall	0.0038	0.0017	8970.228	9070.715	0.5098
Model 4	Vendor, Month, Category × Average Rainfall	0.0043	0.0015	8975.297	9102.582	0.5097
Model 5	Vendor × Category, Month, Average Rainfall	0.0082	0.0029	8981.806	9209.578	0.5087

Table 1 summarizes the performance metrics for five models, each with different interactions, while the first model contains no interaction.

Model 1, which includes basic predictors (Vendor, Category, Month, and Average Rainfall), exhibits poor predictive performance with an R^2 of 0.0038 and an RMSE of 0.5098, indicating limited explanatory power. Model 2 improves slightly by incorporating an interaction between Category and Month, increasing R^2 to 0.00890 and lowering RMSE to 0.5085, suggesting minor gains in prediction accuracy. Model 3, which includes an interaction between Month and Average Rainfall, yields no improvement over Model 1, with identical R^2 (0.00383) and RMSE (0.5098), highlighting the limited value of this interaction.

Model 4 adds an interaction between Category and Average Rainfall, resulting in a marginally higher R^2 of 0.0043 but with negligible impact on RMSE (0.5097). Finally, Model 5 introduces an interaction between Vendor and Category, leading to an R^2 of 0.00821 and an RMSE of 0.5087, showing slight improvement but still limited overall explanatory power.

In conclusion, none of the models provide strong predictive accuracy, but Model 2 achieves the best balance among them with the highest R^2 (0.0089) and lowest RMSE (0.50853). However, the overall low R^2 values indicate that these predictors and interactions capture only a small portion of the variation in price changes.

3.3 Optimization

```
tree_data <- model_data

tree_data$month_category <- with(tree_data, interaction(month, category))

interaction_terms <- model.matrix(~month * category - 1, data=tree_data)
tree_data <- cbind(tree_data, interaction_terms)
main_effects <- tree_data[c("vendor", "avg_rainfall")]
X <- cbind(main_effects, interaction_terms)
y <- tree_data$price_change

set.seed(333)
trainIndex <- createDataPartition(y, p = 0.8, list = FALSE)
train_data <- X[trainIndex,]
test_data <- X[-trainIndex,]
train_y <- y[trainIndex]
test_y <- y[-trainIndex]

tuneGrid <- expand.grid(mtry = c(2, 5, 10))

set.seed(333)
rf_model <- train(
  x = train_data,
  y = train_y,
  method = "rf",
  tuneGrid = tuneGrid,
  trControl = trainControl(method = "cv", number = 10),
  metric = "RMSE",
  ntree = 200
)
```

Linear models performed poorly in predicting price changes for banana and strawberry-flavored products, highlighting their limitations in handling interaction terms and nonlinear relationships. To address this, we propose using Random Forest, a model well-suited for capturing complex interactions and nonlinearity. Unlike linear models, which require manual specification of interaction terms and struggle with nonlinear effects, Random Forest automatically detects and incorporates these relationships through its decision tree structure, providing a more robust approach for this analysis.

Through Table 2, we can compare the evaluation metrics and variable importance for the Random Forest model and the best-performing linear model (Model 2).

Table 2: Table of Model Evaluation Metrics and Variable Importance by Random Forest

(a)

	Model Evaluation Metrics
Metric	Value
RMSE	0.5076
Number of Trees	200.0000
Best mtry	5.0000

(b)

	Variable Importance
Variable	Overall
vendor	100.00
month	87.00
month:category	59.32
category	26.36
average rainfall	23.65

Model 2, which includes basic predictors (Vendor, Category, Month, and Average Rainfall) along with an interaction between Category and Month, achieves an RMSE of 0.5085. The Random Forest model, using 200 trees, slightly improves on this with an RMSE of 0.5076. While the reduction in error is minor, the Random Forest model provides a better overall fit to the data. The Random Forest model demonstrates a stronger ability to handle complex relationships and interactions that are challenging for linear models. It automatically incorporates key predictors like Vendor, Month, Month: Category, and Average Rainfall while effectively capturing intricate interactions between variables. Variable importance analysis identifies 'Vendor' as the most influential predictor, followed by 'Month' and 'Month: Category.'

In conclusion, while the Random Forest model only slightly reduces the RMSE compared to the linear model, it offers greater predictive power by accounting for the data's underlying complexities and interactions more effectively.

4 Result

Table 3: Summary Statistics of Predicted Change in Average Price of Strawberry and Banana Flavored Product by Linear Model Two

Flavor	Avg Change	Median Change	Min Change	Max Change	SD of Change	Sample Size
banana	0.0091	0.0181	-0.109	0.1125	0.0441	1376
strawberry	0.0100	0.0207	-0.109	0.1125	0.0493	4622

Table 3 summarizes the price change statistics for banana and strawberry flavored products predicted by linear model two.

Banana-flavored products have an average price change of 0.0091, with a median change of 0.0091 across 1,376 products. Strawberry-flavored products show a slightly higher average change of 0.0100, with a median of 0.0207 across 4,622 products. For both of the flavors, their price changes range from -0.109 to 0.1125. However, for banana flavored products, their sample size is roughly only a quarter of strawberry product sample size. Overall, strawberry-flavored products exhibit slightly higher average price changes and more variation compared to banana-flavored products.

Table 4: Summary Statistics of Predicted Change in Average Price of Strawberry and Banana Flavored Product by Random Forest

Flavor	Average Predicted Price Change
banana	0.0085
strawberry	0.0103

Table 4 summarizes the price change statistics for banana and strawberry-flavored products predicted by a random forest model.

Banana-flavored products have an average price change of 0.0085, while strawberry-flavored products show a slightly higher average change of 0.0103. This result means that for random forest estimation, the price of both banana and strawberry-flavored products are gonna increase, but strawberry-flavored products get a slightly higher rise in price.

[fig-lmone] illustrates the predicted monthly price changes for strawberry and banana-flavored products. The price trends show no significant differences between the two flavors, with nearly stable prices in July, a slight dip in August and September, and minor increases in October and November. October saw the highest increase, while August experienced the largest decrease. However, the average price changes remained within ± 0.1 units, indicating limited fluctuation.

[fig-lmtwo], like [fig-lmone], uses predictions from Linear Model 2. However, while [fig-lmone] shows predicted price changes across months, [fig-lmtwo] present their distribution across food categories. Similar to the monthly distribution, there is no significant difference between strawberry and banana-flavored products across categories, except for flavored tea. Banana-flavored tea shows a downward price trend with fewer products compared to strawberry-flavored tea,

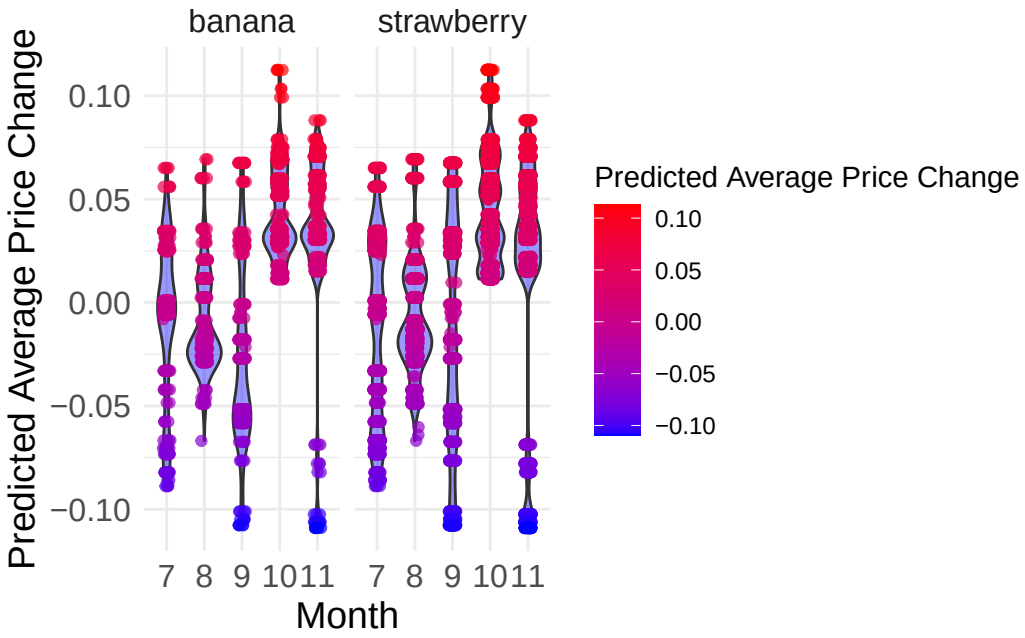

Figure 7: Distribution of Linear Model Two Predicted Monthly Price Changes for Banana and Strawberry Products by Month

which exhibits rising prices. For beverage, fruit, and solid snack categories, both flavors demonstrate increasing price trends, while yogurt prices remain largely stable.

Predictions generated by the random forest model were shown in Figure 9 and Figure 10.

Figure 9 illustrates the predicted price changes for banana and strawberry-flavored products across months. While the differences are not pronounced, there are notable variations in their distributions. In July, August, and September, prices for both flavors decline, with banana-flavored products experiencing a steeper drop. By October, prices for both flavors rise, with a sharper increase for banana-flavored products. In November, price increases continue but at a slower pace, with some products, particularly banana-flavored ones, showing significant price reductions. Overall, the price changes largely fall within a range of ± 0.1 units.

Figure 10 presents the predicted price changes for banana and strawberry-flavored products across food categories. In both of the flavors, prices for beverages, fruit, and solid snacks show an overall increase, while yogurt remains relatively stable. For flavored tea, banana-flavored products exhibit minimal price changes, whereas strawberry-flavored products see a noticeable price increase.

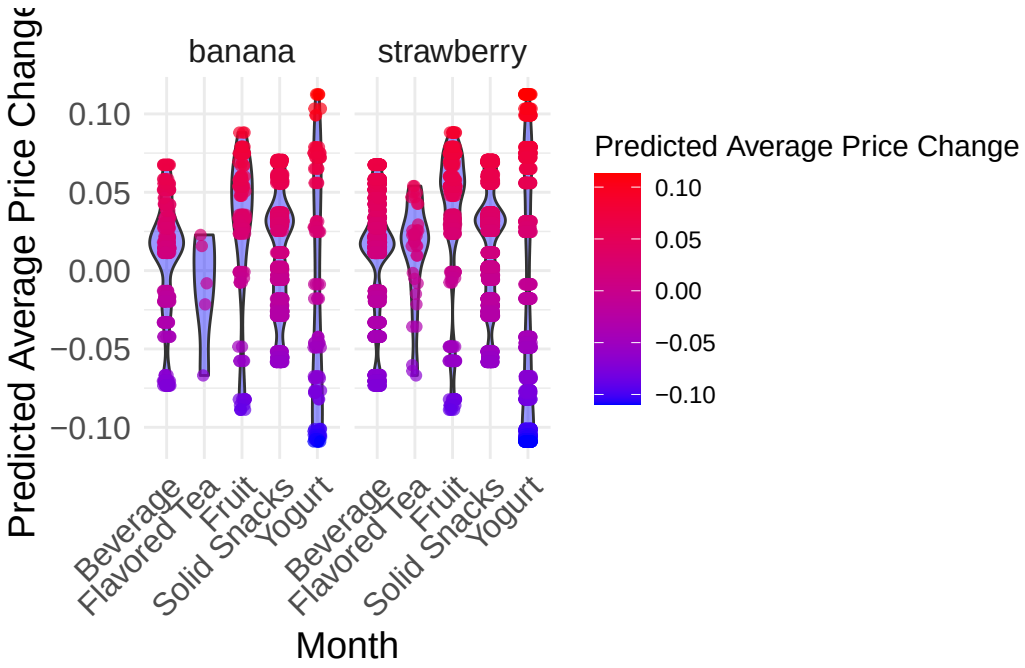

Figure 8: Distribution of Linear Model Two Predicted Monthly Price Changes for Banana and Strawberry Products by Category

5 Discussion

5.1 Summery of Findings

In our study, we analyzed price changes for banana and strawberry-flavored products using linear and Random Forest models. The linear model showed average price changes of 0.0091 dollars for banana-flavored products and 0.0100 dollars for strawberry-flavored ones. Similarly, the Random Forest model reported 0.0085 for bananas and 0.0103 for strawberries. Both models revealed small price increases from July to November, with stable prices in July, slight declines in August and September, and increases in October and November, peaking in October. Price changes for both flavors stayed within ± 0.1 units. Across food categories, beverages, fruits, and solid snacks showed price increases, yogurt remained stable, and flavored tea showed divergent trends: banana-flavored tea prices decreased, while strawberry-flavored tea prices increased. Overall, both models indicated similar trends, with strawberry-flavored products experiencing slightly higher price increases than banana-flavored ones.

This result was unexpected, as we initially assumed that bananas would be more popular in winter and strawberries in summer, and hence banana products would increase more in price. However, the model predicts a greater price increase for strawberry-flavored products compared to banana-flavored ones. This suggests that while seasonal fruits are more popular

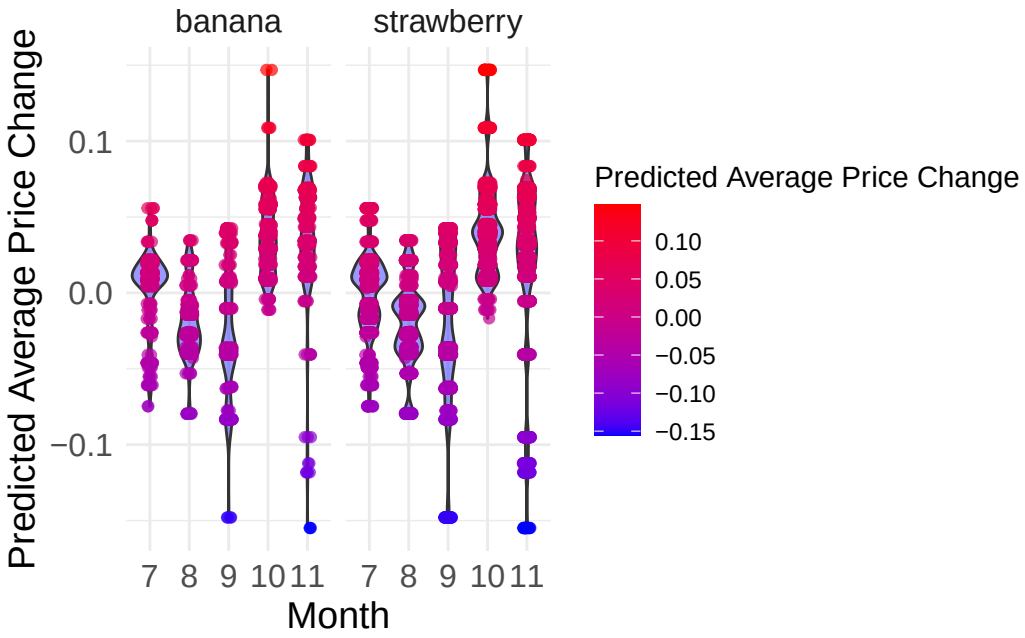

Figure 9: Distribution of Random Forest Model Predicted Monthly Price Changes for Banana and Strawberry Products by Month

during their peak seasons, their flavored products may not follow the same pattern. For instance, in winter, consumers might prefer strawberry-flavored products to recall the taste of strawberries in summer, rather than choosing banana-flavored ones.

Among all predictors, vendor and month are the most influential, while food category and average rainfall have less impact. The interaction between month and food category ranks after month. The dominance of vendors is due to the sales strategies each vendor employs, such as deciding which products to sell and how to adjust prices across different periods and categories. This makes vendors the most important contributor to price changes, as vendors directly control pricing. The next most important predictors are month and its interaction with category, which reflects seasonal variations in consumer behavior and preferences for different food types. The interaction term provides deeper insights into these temporal consumption patterns. Although food category and average rainfall contribute less, both have importance values exceeding 20, demonstrating their limited but indispensable role in the model.

5.2 Limitation

The model's limited performance underscores a key challenge in this study: the quality of the original dataset, which affected its accuracy.

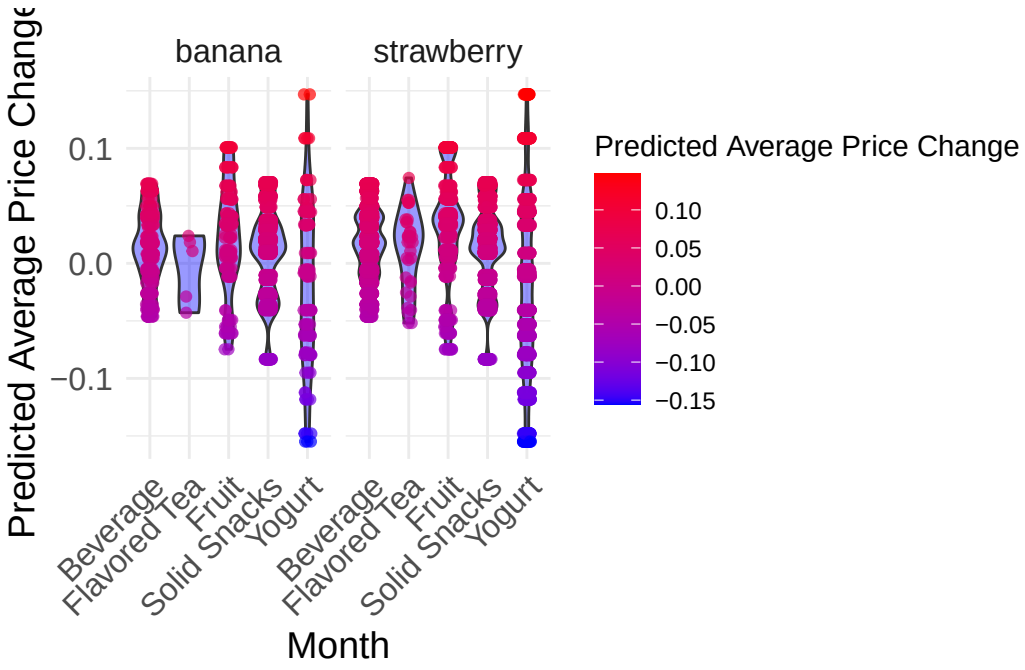

Figure 10: Distribution of Random Forest Model Predicted Monthly Price Changes for Banana and Strawberry Products by Category

First, while data collection from vendor websites began on February 28, 2024, as noted by the dataset's creator, only a small number of products were recorded until large-scale scraping began on July 10. Strawberry- and banana-flavored products were absent from the early data, with their earliest prices recorded on June 11. This nearly four-month gap reduced the dataset's size and temporal coverage, complicating predictions.

Second, the dataset's variability in product categories and inconsistent labeling across vendors led to data heterogeneity. For instance, units of measurement were not standardized. Ideally, predictions should have used consistent units, such as price per 100g of bananas or 100ml of strawberry beverages. However, some products only had unit prices, while others used different units—for example, yogurt was priced per 100g, per 100ml, or box. This forced us to rely on less precise unit prices. A 300ml beverage and a 2L beverage, for instance, might have similar per-100ml prices but very different in unit price.

Third, the dataset had an uneven distribution of products between the two flavors. Strawberry-flavored products were more prevalent, likely due to their broader popularity or the limited representation of banana-flavored products. This led to denser data distributions for strawberry products, which could introduce bias in predictions for banana products. In certain categories, such as flavored tea, the scarcity of banana-flavored products was especially pronounced, further worsening prediction biases.

Lastly, our assumptions may not fully align with real-world dynamics. We expected seasonal fruits to have higher demand and prices during their peak seasons, with lower demand off-season. However, this may apply more to the fruits themselves than seasonal fruit-flavored products, which could be influenced by other factors beyond seasonality. For instance, off-season sales might not result in price decreases but could lead to price increases as vendors adjust for lower demand. In conclusion, food price changes are influenced by multiple factors, making this a complex issue. We underestimated the challenges of modeling price dynamics accurately.

5.3 Future Study

Future research should address the limitations of this study to improve the accuracy of price change models for seasonal fruit-flavored products. First, data collection methods need to be improved. Future efforts should ensure consistent and comprehensive data coverage for all target product categories and flavors from the start, ideally spanning 2–3 years to enhance reliability. Second, standardizing product measurement units is essential. Converting prices to uniform units, such as per 100 grams or 100 milliliters, would reduce data heterogeneity and enable better comparisons. Efforts should also focus on unifying units within each food category, resolving inconsistencies through manual data supplementation, or improved web-scraping techniques. For example, fruit prices could be standardized per 100 grams, while beverages could be per 100 milliliters. Products without detailed unit information, like "boxed fruits" or "cases of strawberry juice," could be supplemented using manual research or advanced scraping methods. Finally, incorporating additional factors such as regional variations and promotional activities could provide a clearer understanding of the drivers behind price changes. We can also collect consumers' consumption preferences for seasonal fruits and their flavored products through questionnaires. With these improvements, future studies could develop more robust models to better forecast the dynamics of seasonal fruit-flavored product prices.

To build on the existing work with linear models and Random Forest, future studies could explore alternative machine learning algorithms that better capture the complexities of flavored product price dynamics. Gradient Boosting Machines (GBMs), such as XGBoost or LightGBM, could be particularly effective, as they excel in handling structured data, capturing non-linear relationships, and minimizing overfitting. Additionally, Support Vector Machines (SVMs) could be applied, particularly for classification tasks, such as predicting the likelihood of price increases or decreases based on features like flavor and category.

Appendix a.

5.1 Data Retrieval

We obtained the raw data from Jacob Filipp's Project Hammer website, available at: https://jacobfilipp.com/hammer/. To access the data, scroll down the page until you see the heading "Get Full Data in 2 CSV Files". Click on the link labeled "Zipped CSVs with full price data" to download the raw data. After extracting the files, you will obtain two CSV files: hammer-4-product.csv and hammer-4-raw.csv. Next, place these two files in the project's data/01-raw_data folder, and you'll be able to run the code without issues.

5.2 Data Cleaning

In the data cleaning process for the banana and strawberry product datasets, we first removed any rows in the product_name or concatted columns where both "banana" and "strawberry" appeared together, to avoid mixed category products. After filtering out these mixed-category rows, the dataset was split into two separate datasets: one for banana-related products and the other for strawberry-related products. Each dataset was further categorized based on keywords found in the product_name and concatted columns, such as "beverages," "yogurt," "flavored tea," and "solid snacks," or defaulted to "fruit." We first categorized products containing the keyword "yogurt" into the yogurt category, then those containing the keyword "tea" into the flavored tea category. The flavored tea category includes both flavored tea beverages and solid tea bags, which cannot be distinguished by units alone. Next, we filtered for products with keywords like "drink," "lemonade," "juice," "sparkling," or "beverage" and categorized them as beverages. Products with the "organic" keyword were categorized as fruit. For the remaining products, we classified those with units in ml or L as beverages and those with units in kg or g as solid snacks. Any remaining products were classified as fruit. Although a few products were misclassified during this process, the overall categorization was accurate. We then removed unnecessary columns such as concatted, detail_url, sku, and upc to simplify the dataset.

Next, we filtered the original product dataset to retain only the products found in the cleaned banana and strawberry datasets. We removed rows with missing price_per_unit values and converted the nowtime column into the correct date format. Other unnecessary columns, such as old_price and other non-essential data, were also removed. Then, the data was sorted by nowtime for each product, and missing dates were handled by generating a complete date sequence for each product_id. Since web scraping can sometimes result in missing or duplicated data, missing current_price values were filled in with the latest available price for each product.

For each product, we calculated the average monthly price from June to November and merged this average price back into the banana and strawberry datasets. Rows without an average

price during these months were then deleted. The cleaned banana and strawberry datasets were saved as Parquet files for further analysis. This systematic approach ensured proper data filtering, handled missing values and prepared the final datasets for analysis.

Appendix b.

5.1 Idealized Methodology for Analyzing Seasonal Price Changes of Fruit-Flavored Products

We aim to understand price fluctuations of seasonal fruit-flavored products during their in-season and off-season periods. By combining consumer surveys and retail data, we will identify trends in purchasing behavior and retailer pricing. With a $200,000 budget, we will use stratified sampling, incentivized recruitment, and strict data validation to ensure reliable results. The findings will be further enhanced with secondary data, such as retail sales reports and online marketplace trends.

5.2 Budget Allocation

We assume we will have a budget of $200,000, which will be used for data collection, effective sampling, and comprehensive analysis. The proposed budget breakdown is as follows:

The survey platform costs $30,000 Subscription fees for Qualtrics or Google Forms for dynamic and customizable surveys.

Respondent incentives: $40,000 Gift cards to encourage participation among 2,000 respondents.

Recruitment and staffing: $60,000 Staffing costs for survey distribution and observational data collection.

Observational data collection: $40,000 Travel and logistical costs for monitoring prices across various retail locations.

Data analysis tools: $20,000 Statistical software licenses and tools for data cleaning and analysis.

Marketing and promotion: $6,000 Social media campaigns to boost survey reach.

Contingency fund: $4,000 Reserved for unforeseen expenses.

5.3 Sampling Approach

We will use stratified sampling to ensure representation across demographics and consumer behavior. The target population includes regular consumers of fruit-flavored snacks, drinks, and desserts, focusing on factors like age, income, location, and purchasing habits, such as how often they buy seasonal fruit-flavored products. We will survey 2,000 respondents, providing a 95% confidence level and a $\pm 2\%$ margin of error. Recruitment will use online platforms, social media, and community networks, targeting frequent buyers (at least once a month), seasonal buyers, and price-sensitive shoppers.

We will also collect observational data from 30 retailers, including supermarkets, convenience stores, and online marketplaces, across urban, suburban, and rural areas. We will track fruit-flavored products like strawberry, banana, and mixed-fruit items over a year, covering both in-season and off-season periods to capture pricing and availability trends.

5.4 Respondent Recruitment and Data Validation

To gather a representative and diverse dataset, we will use a multi-mode recruitment strategy. Online surveys will be distributed via platforms like Google Forms or Qualtrics, targeting food enthusiasts and regular grocery shoppers through social media and email lists. A questionnaire is available at the following link: https://forms.gle/gGy5CFKF33MCRm5ZA. In-store recruitment will involve flyers and QR codes in participating retail stores to directly engage shoppers. To boost participation, each respondent will receive a $20 digital gift card upon completing the survey. Data validation will include automated checks for incomplete or inconsistent responses, and manual review of open-ended answers to identify irrelevant or duplicate entries.

Observational data will provide real-world insights into price trends and product availability. Weekly price monitoring will track the prices of targeted products, while stock availability will be checked to identify stockouts or limited supply during off-seasons. Observers will also record promotional campaigns and discounts to account for pricing influences.

5.5 Analysis and Modeling

Survey and observational data will be combined and analyzed using statistical methods to identify trends. Weighted analysis will adjust survey responses to reflect the demographics of the target population. Observational data will track seasonal price changes and compare them across different regions. Predictive modeling, including regression analysis, will identify factors that affect consumer willingness to pay for seasonal fruit-flavored products during in-season and off-season periods. Secondary data from retail sales reports and online platforms like Amazon and Walmart will be incorporated to enhance the analysis and provide a broader market view.

5.6 Expected Outcomes

The study aims to collect and model consumer behavior and market trends. It will identify differences in purchasing behavior and price sensitivity between in-season and off-season periods, as well as factors affecting willingness to pay for seasonal fruit-flavored products. Market trends will reveal seasonal pricing patterns, regional and demographic influences, and retailer strategies for pricing and promotions during off-seasons. These findings will help businesses and policymakers align their strategies with consumer preferences.

References

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

Arel-Bundock, Vincent. 2023. *Modelsummary: Summary Tables and Plots for Statistical Models and Data*. https://CRAN.R-project.org/package=modelsummary.

Filipp, Jacob. 2024. "Project Hammer." https://jacobfilipp.com/hammer/.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "rstanarm: Bayesian applied regression modeling via Stan." https://mc-stan.org/rstanarm/.

Grolemund, Garrett, and Hadley Wickham. 2020. *Lubridate: Make Dealing with Dates a Little Easier*. https://CRAN.R-project.org/package=lubridate.

Hadley Wickham, Romain François. 2020. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://stringr.tidyverse.org.

Kassambara, Alboukadel. 2023. *Ggpubr: 'Ggplot2' Based Publication Ready Plots*. https://rpkgs.datanovia.com/ggpubr/.

Kuhn, Max et al. 2023. *Caret: Classification and Regression Training*. https://CRAN.R-project.org/package=caret.

Liaw, Andy, and Matt Wiener. 2023. *randomForest: Breiman and Cutler's Random Forests for Classification and Regression*. https://CRAN.R-project.org/package=randomForest.

Müller, Kirill, and Hadley Wickham. 2022. *Tibble: Simple Data Frames*. https://CRAN.R-project.org/package=tibble.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Reeves, Joy. 2022. "When Are Bananas in Season?" https://seasonalcornucopia.com/48/banana-tree-cycle-and-season/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Shelter, Toronto, and Support Services. 2021. *About Daily Shelter and Overnight Service Occupancy & Capacity*. https://open.toronto.ca/dataset/daily-shelter-overnight-service-occupancy-capacity/.

Watson, Molly. 2019. "Your Guide to Seasonal Fruits and Vegetables." https://www.thespruceeats.com/guide-to-seasonal-fruits-and-vegetables-2216387.

Wickham, Hadley et al. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, and RStudio. 2020. *Testthat: Unit Testing for r*. https://CRAN.R-project.org/package=testthat.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Yu, Hui. 2023. *kableExtra: Construct Complex Tables for LaTeX and HTML*. https://cran.r-project.org/package=kableExtra.