Analyzing Factors Affecting the Potato Prices in Canadian Grocery Markets*

Vendor Type, Historical Pricing, and Sales Month Play Key Roles in Determining Potato Prices through Bayesian Analysis

Fangning zhang

December 2, 2024

In this study, we analyze the factors influencing potato prices in Canadian grocery markets by employing Bayesian regression modeling. The analysis focuses on how variables such as vendor type, historical pricing, and the sales month affect current pricing trends for yellow and white potatoes. Our findings indicate that vendor type and historical pricing significantly impact potato prices, with distinct variations based on sales month. These insights enhance our understanding of potato pricing dynamics, offering valuable information to consumers, retailers, and policymakers.

Table of contents

*Code and data are available at: https://github.com/FangningZhang81/Canadian-Grocery-Price

1 Introduction

The potato is one of the most widely consumed vegetables globally, serving as a vital source of starch and alcohol. It has become an essential staple food in many parts of the world, particularly in regions with predominantly white populations(Ritchie and Dustan 1940). In Canada, potatoes are not only an important dietary component for millions of people but also play a significant role in the grocery market. They are versatile in cooking and contribute substantially to the local economy, with Canadian farmers producing millions of tonnes annually(Khakbazan et al. 2015). However, the price of potatoes fluctuates significantly throughout the year, influenced by a variety of factors such as vendor pricing strategies, seasonality, historical price trends, yield fluctuations, and shifts in consumer demand. These price variations have far-reaching implications for producers, consumers and retailers. To gain a more scientific and intuitive understanding of these price fluctuations, it is essential to quantify the factors that influence them. This paper explores how factors such as vendor, historical pricing, and

the month of sale affect the current price of potatoes, using Bayesian regression model to provide insights into the dynamics of potato pricing in the Canadian market.

In this paper, to investigate the factors influencing the current price of potatoes in Canadian grocery markets, we analyzed data Canadian grocery price data(Filipp 2024) from June 2024 to November 2024 cross various grocery chains, including, Loblaws, Metro, NoFrills, SaveOnFoods and Walmart. Our estimand is the influence of these factors on the current prices of potatoes, taking into account variations between vendors, old prices, and the month of sale.

The results from the Bayesian model indicate that all three factors—vendor, old pricing, and sales month—significantly influence the current price of potatoes. Old price plays a key role on the current price, with previous prices strongly influencing the current price. The results also show that higher-priced potatoes have greater price fluctuations, while lower-priced potatoes have little difference between the current and old prices. Vendors influence the current price a lot, with some vendors showing consistently higher or lower prices than others. Seasonal variations, as reflected in the sales month, have a more moderate but still important impact, with potato prices generally increasing during certain months.

Understanding the factors driving fluctuations in potato prices is essential for producers, consumers, and retailers. For consumers, particularly low-income households, awareness of pricing trends aids in budgeting and making informed purchasing decisions. When the price increases, consumers may shift to alternative staple foods or frozen potato products. Retailers can use these insights to optimize pricing strategies and inventory management. Furthermore, policymakers and agricultural producers can also benefit from the findings, as they are crucial for the market dynamics and potato supply chain.

The remainder of this paper is structured as follows. Section 2 provides an overview of the dataset used in the analysis, measurement relating to the dataset, cleaning data process and some details about variables of interest. Section 3 contains the Bayesian regression model used to analyze the factors influencing the current price. Section 4 contains data result visualizations and also the results of the Bayesian regression model. Section 5 discusses what was done, implications, limitations and next steps. In the appendix, Section A and Section B includes more details for the models and Section C we focused on some aspect of surveys, sampling and observational data.

In this project, We used the R programming language(R Core Team 2023) and several R packages for data processing, analysis, and visualization. Specifically, tidyverse(Hadley Wickham and the tidyverse team 2023), arrow(Neal Richardson and Apache Arrow contributors 2023), dplyr(Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller 2023), tidyr(Hadley Wickham, Lionel Henry, and other contributors 2023), janitor(Sam Firke 2023), rstanarm(Goodrich et al. 2024), and here(Müller 2020) were used for data processing, cleaning, date and time operations, project path management, and efficient data storage and reading. knitr(Yihui Xie 2023), patchwork(Thomas Lin Pedersen 2023), bayesplot(Michael B. Goodrich and Gabry 2023) and ggplot2(Wickham 2016) were used for creating dynamic reports, beautifying table outputs, data visualization, and arranging

multiple charts. testthat(Hadley Wickham and others 2023) was used for writing and executing unit tests. styler(Müller and Walthert 2024) is used for final code formatting.

2 Data

2.1 Overview

The Canadian Grocery Price Data (Filipp 2024) used in this analysis was sourced from the Project Hammer website and consists of two datasets: raw sales data and product specific information. Project Hammer aims to drive more competition and reduce collusion in the Canadian grocery sector. Therefore, a database of historical grocery prices from top grocers' websites is compiled for academic and business analysis. Specifically, information of the products and data of the prices are gathered from a screen-scrape of website UI - that is why some information is missing. The database includes sales data from eight vendors: Voila, T&T, Loblaws, No Frills, Metro, Galleria, Walmart, and Save-On-Foods. The available dates range from February 28, 2024, to the latest data load at November 28, 2024.

The dataset used in this paper was downloaded on November 22, 2024; data released after this date was not considered anywhere in this paper.

2.2 Measurement

The price of potatoes fluctuates over time, influenced by a variety of factors such as vendor pricing strategies, seasonality, historical price trends, yield fluctuations, and shifts in consumer demand. These fluctuations influence purchasing decisions of consumers, pricing strategies of retailers and also the decisions for policymakers and agricultural producers. To understand these fluctuations, it is essential to quantify the factors that influence them.

To transition from real-world sales events to a structured dataset, we began by recording each transaction. The information of each product was recorded as a row in the raw dataset, which included details like the product name and brand, prices, and the vendor information. This research focuses on a single product: potatoes. Therefore, `product_name` was used to identify the specific types of potatoes (yellow and white), allowing us to focus on these items exclusively. The `current_price` and `old_price` fields represent the financial aspect of these sales, capturing both the price at the time of the transaction and the historical price for comparison.The `month` variable was derived from the transaction timestamp (`nowtime`) to provide a temporal context for analysis, allowing us to observe trends and patterns over different periods.

By filtering and cleaning the raw data, the dataset reflects the aspects of potato sales that are of interest for this study. This process of measurement transforms abstract sales activities

into quantifiable data points that can be used for statistical analysis, providing insights into vendor behavior and price dynamics over time.

2.3 Cleaning Process

Initially, two raw datasets were downloaded from the Project Hammer website. These datasets were merged to create a comprehensive dataset containing all relevant columns, including `nowtime`, `vendor`, `product_id`, `product_name`, `brand`, `units`, `current_price`, `old_price`, `price_per_unit`, and `other`.

The cleaning process involved several key steps to refine the dataset for analysis. First, we selected only the columns relevant to our study shown in Table 1. The month of each transaction was extracted from the `nowtime` to provide temporal context for the sales data.

Table 1: Summary of Selected Columns

Column	Description
nowtime	Timestamp indicating when the data was gathered
vendor	One of the 7 grocery vendors
current_price	Price at time of extract
old_price	An 'old' struck-out price, indicating a previous sale price
product_name	Product name, may include brand and units

Next, we filtered the dataset to include only the products of interest (yellow and white potatoes) by searching for these keywords in the `product_name` column, ensuring our analysis was focused specifically on these items. Additionally, any entries with missing values (NA) were removed to maintain data quality. Finally, the `nowtime` column was dropped after extracting the month, as it was no longer necessary for the analysis. Table 2 gives an preview of the cleaned dataset we will use in the following sections.

Table 2: Preview of Cleaned Data on Potato Sales

vendor	current_price	old_price	product_name	month
Loblaws	4.99	5.99	Small White Potatoes	10
Loblaws	4.99	5.99	Small White Potatoes	10
Loblaws	4.99	5.99	Small White Potatoes	10
Loblaws	4.99	5.99	Small White Potatoes	10
Loblaws	4.50	5.99	Small White Potatoes	10
Loblaws	4.50	5.99	Small White Potatoes	10

These steps resulted in a cleaned dataset containing the essential information needed for the analysis of potato sales trends.

2.4 Variables of Interest

The cleaned dataset contains 685 rows, representing individual products sale, along with vendor information and price details. The summary statistics of the cleaned data indicate that the mean of current price is 2.811, and the mean of old price is 3.785. In Figure 1, the first plot shows a positive linear relationship between old and current potato prices, while the second and third plots indicate how vendor type and month influence average current prices, respectively, with significant price variations between different vendors and months. Figure 1 and Table 3 show that the cleaned dataset only contains five vendors instead of the eight in the raw dataset, and the month is from June to November. This is expected because data collection from February 28 to July 10 focused on a smaller set of products by the description of the raw dataset. After July, more products were added, and some data may be missing for certain vendors or days when extraction failed.

Table 3: Overview of Analysis Data

current_price	old_price	month
Min. :0.940	Min. :1.490	Min. : 6.000
1st Qu.:2.000	1st Qu.:2.490	1st Qu.: 8.000
Median :2.290	Median :2.490	Median :10.000
Mean :2.811	Mean :3.785	Mean : 9.064
3rd Qu.:3.990	3rd Qu.:5.990	3rd Qu.:10.000
Max. :7.990	Max. :8.990	Max. :11.000

2.5 Other dataset

A dataset on monthly average retail prices for selected products (Government of Canada 2024) was identified but ultimately not utilized. The dataset only provides average prices, which limits its ability to capture price fluctuations over time. Furthermore, it lacks vendor-specific information, which is essential for analyzing the impact of different vendors on product pricing.

3 Model

The goal of our analysis was to understand the relationship between current potato prices and influencing factors, specifically `month`, `old_price`, and `vendor`. To accomplish this, we

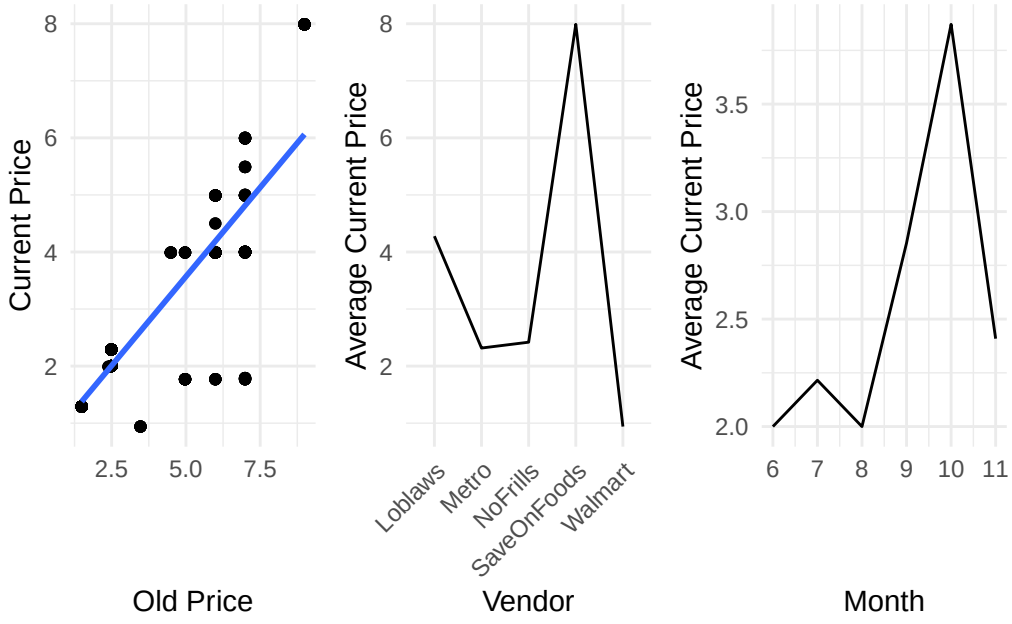

Figure 1: Factors Influence Current Prices: Comparison Across Old Pricing, Vendor, and Month

developed a Bayesian regression model to capture the underlying dynamics that contribute to changes in pricing.

Our variables of interest will be the predictor variables for the model. The Bayesian model uses the following predictors:

- `month`: Represents the month during which the data was collected. The values range from June to November (i.e., months 6 to 11).
- `old_price`: Represents the previous price of the product, providing insight into historical pricing trends. It is used to understand how past prices influence current pricing.
- `vendor`: Represents the specific grocery vendor selling the product. The dataset includes five vendors, including Loblaws, Metro, No Frills,Save-On-Foods and Walmart.

3.1 Model set-up

To predict the current price of a product, we assume a linear relationship between the outcome variable (`current_price`) and our predictor variables (`month`, `old_price`, and `vendor`). We define our model as a Bayesian regression model as follows:

$$
\begin{aligned}
y_i &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \beta_0 + \beta_1 \times \text{month}_i + \beta_2 \times \text{old\_price}_i + \beta_3 \times \text{vendor}_i \\
\beta_0 &\sim \text{Normal}(0, 2.5) \\
\beta_1 &\sim \text{Normal}(0, 2.5) \\
\beta_2 &\sim \text{Normal}(0, 2.5) \\
\beta_3 &\sim \text{Normal}(0, 2.5) \\
\sigma &\sim \text{Exponential}(1)
\end{aligned}
$$

In this model, y_i represents the current price for product i, modeled as a normal distribution with mean μ_i and standard deviation σ. The predictor variables month_i, old_price_i, and vendor_i represent the month of data collection, the old price of the product, and the vendor, respectively. The coefficients β_0, β_1, β_2, and β_3 are assigned normal priors with a mean of 0 and a standard deviation of 2.5, while σ is assigned an exponential prior with rate 1. Choosing a relatively loose prior allows the coefficients to fluctuate within a moderate range while avoiding overly restrictive assumptions. A normal distribution with a standard deviation of 2.5 can capture most plausible true effects but avoids assigning excessively high probabilities to very large coefficients. This takes into account that the influence of predictors (such as month, old price, and vendor) on the response variable (current price) is generally limited or moderate.

The Bayesian model offers advantages, such as the ability to incorporate prior information and quantify parameter uncertainty more comprehensively. We ran the model using the `rstanarm` package (Goodrich et al. 2024) in R (R Core Team 2023). Model diagnostics, including trace plots, Rhat values, and posterior predictive checks, can be found in the Appendix (Section B).

3.2 Model Justification

The multiple linear regression model using Ordinary Least Squares, with the same predictor variables as the Bayesian model, was also considered as an alternative during the model-building process. It exhibited strong explanatory power, as reflected by high R-squared and adjusted R-squared values, and most predictor variables were statistically significant, enabling a clear interpretation of their effects on `current_price`. Furthermore, the conclusions about the significance of the variables were consistent between the linear regression and Bayesian models, enhancing the reliability of the findings. Details of the linear regression model can be found in the Appendix (Section A).

However, despite its strengths in simplicity and computational efficiency, the linear regression model has limitations in quantifying parameter uncertainty and incorporating prior knowledge. In contrast, the Bayesian model offers greater flexibility by integrating prior distributions and

providing posterior distributions for parameters, which better capture the uncertainty and variability in the data. Given that the cleaned dataset contains only around 700 observations, the Bayesian approach is more robust and suitable for handling smaller datasets and scenarios with potential noise or variability. These considerations ultimately led to the selection of the Bayesian model, relying on Markov Chain Monte Carlo (MCMC) sampling for estimation, for this analysis.

The model has certain underlying assumptions, potential limitations, and scenarios in which it may not be appropriate. The model assumes a linear relationship between the predictors and the target variable. If the relationship is non-linear, the model may yield biased results without additional interaction terms. Furthermore, the model assumes that the residuals, which represent the differences between observed and predicted prices, follow a normal distribution and have constant variance. Any deviations from this assumption can lead to inaccurate inferences, such as changes in variance over time or among vendors. The model also assumes that observations are independent, which may not always be true in practice, as pricing data can exhibit temporal dependencies (such as trends or seasonality) or spatial correlations (such as vendor location effects) that are not accounted for. The model assumes the effect of a vendor is constant across time and products. This might oversimplify the dynamics of vendor-specific pricing strategies. More limitations are discussed in the Section 5.3.

4 Results

In this section, we visualized our data through graphs and tables as well as present the results from our model.

4.1 Data Results

Figure 2 illustrates the distribution of old and current prices of potatoes, with point sizes indicating the frequency of occurrence. A red dashed line represents where the old and current prices are equal (i.e., the line x = y). All points fall below the x = y line, which suggests that the current price of potatoes is lower than their old price. This observation aligns with expectations, as only discounted products have an "old price." Additionally, points representing lower prices are closer to the red line, indicating minimal changes between old and new prices for lower-priced potatoes. The new prices of lower-priced potatoes are similar to the old prices.

The size of each point reflects the frequency of a particular combination of old and current prices. Larger points are seen at lower price values, particularly where old prices are around 2.5, suggesting that lower prices occurred more frequently.

Figure 3 shows the trend of total products, while Figure 4 illustrates the box plots of current and old prices across different months. The total number of products is closely related to the

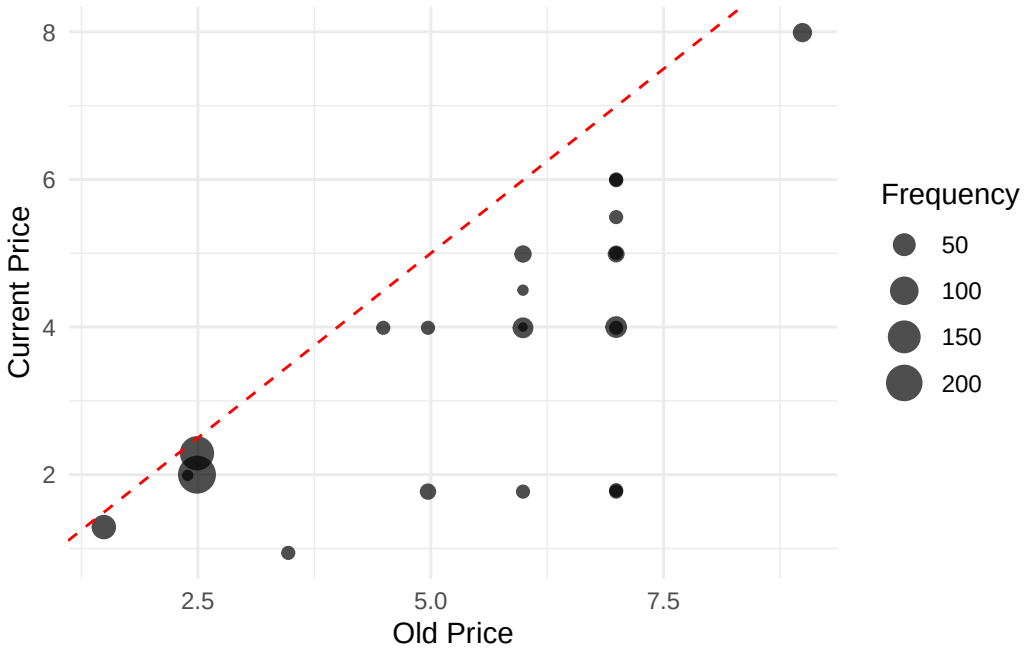

Figure 2: Distribution of Current vs. Old Price of Potatoes (Point Size by Frequency)

price range and fluctuations during each period. As shown in Figure 3, the total number of products starts increasing in September, reaches its peak in October, and then slightly declines in November. From June to August, both current and old prices remained relatively stable, with minimal variability and a few outliers. From September onward, both prices exhibited significant fluctuations, with increased median values and widened price ranges, along with noticeable outliers. In October, the median values reached their highest point, followed by a decrease in November, but overall variability remained substantial.

Figure 5 illustrates the average current and old prices of potatoes across different vendors. A notable common characteristic between the plots is that both show significant variation in prices among vendors, with peaks and troughs at similar vendor locations. Specifically, both average current and old prices reach a maximum at "SaveOnFoods," while the average prices for "Metro," "NoFrills," and "Walmart" are relatively lower.

In summary, the analysis indicates that current potato prices are lower than old prices, with smaller differences observed for lower-priced potatoes, as shown in Figure 2. Figure 3 and Figure 4 show a rise in the number of products from September to October, along with increased price variability. Average prices differ significantly among vendors, with "SaveOn-Foods" having the highest prices, while "Metro," "NoFrills," and "Walmart" are relatively lower, as illustrated in Figure 5.

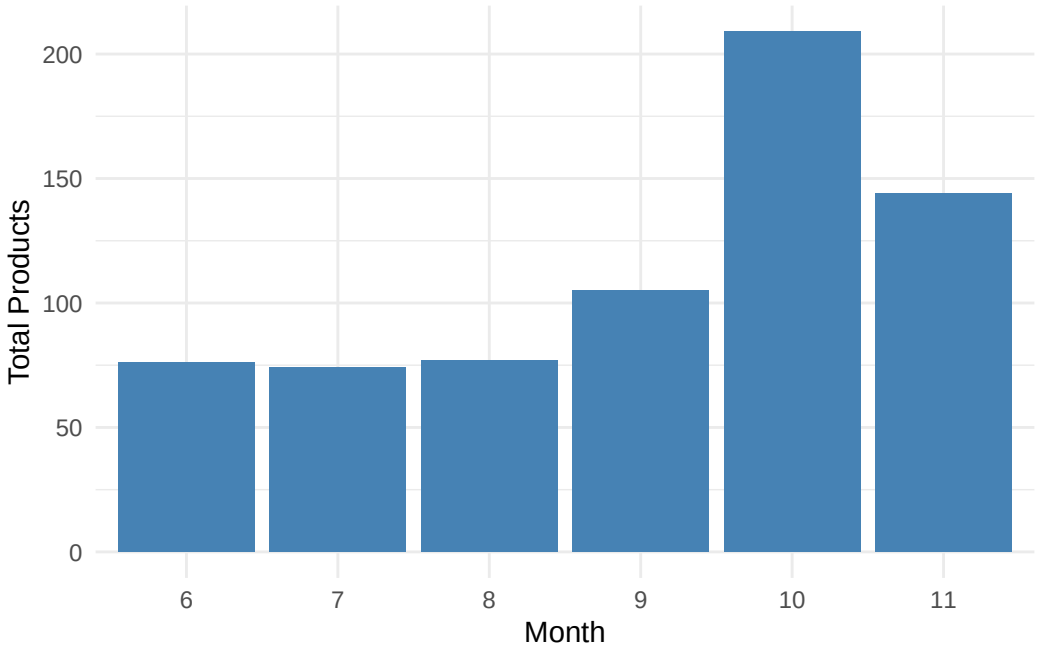

Figure 3: Total Number of Products by Month

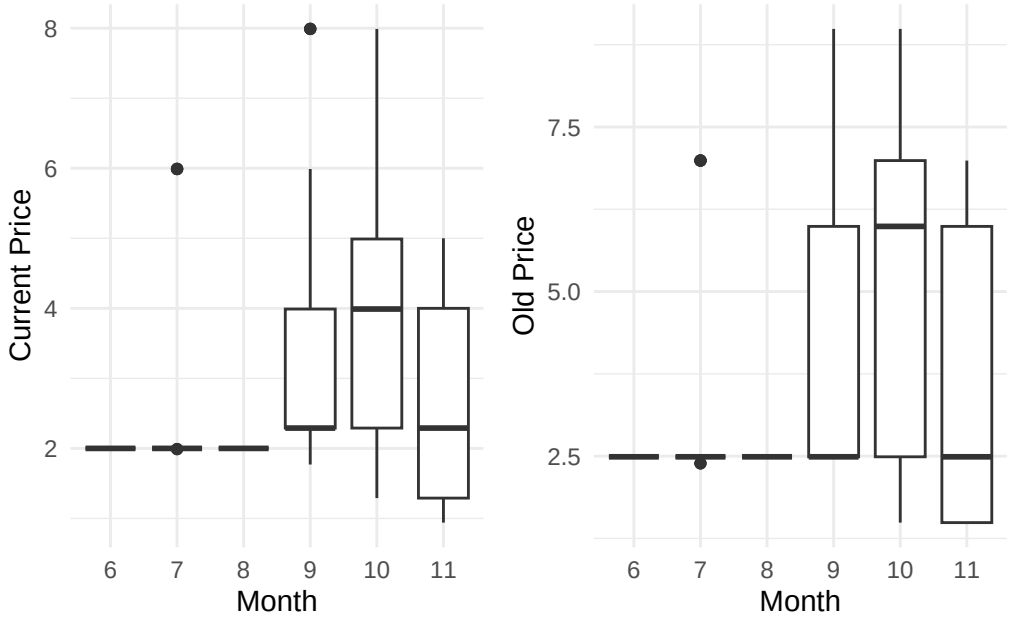

Figure 4: Distribution of Current and Old Prices of Potatoes by Month

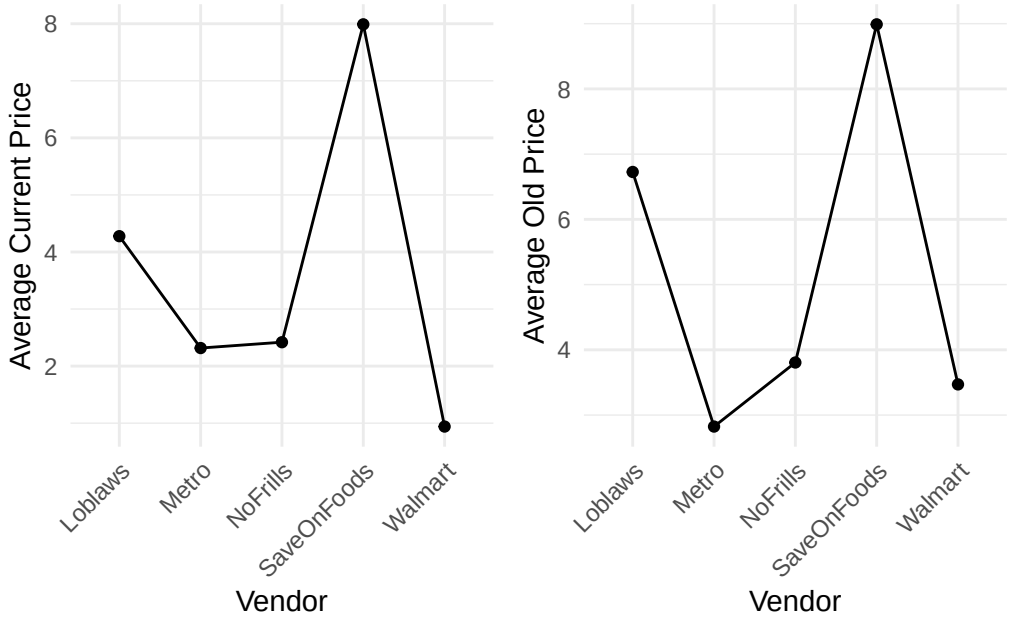

Figure 5: Distribution of Average Current and Old Prices of Potatoes by Vendor

4.2 Model Results

Table 6 and Figure 6 present the results of our Bayesian regression model that explores the relationship between current price and various predictors including `month`, `old price`, and `vendor`. The parameter estimates, calculated using the mean of the posterior distributions, along with mean absolute deviation (MAD) and additional model metrics are summarized in Table 6, while Figure 6 provides a graphical representation of the 90% credible intervals to facilitate understanding of the uncertainty around each summary of Coefficients

The intercept is estimated at 0.31, with an MAD of 0.16, suggesting that the base value of current price (when all other predictors are at their reference levels) is positive. The coefficients for `month7`, `month8`, and `month9` are 0.08, 0.00, and 0.05 respectively, indicating minor or negligible changes in current price depending on the month, especially since the 90% credible intervals for `month8` and `month9` include zero, implying a lack of statistical significance.

The predictor `old_price` has an estimated coefficient of 0.54 with an MAD of 0.02, which suggests a strong and significant positive relationship with current price. This means that a higher old price is positively associated with a higher current price, and the narrow credible interval suggests this relationship is consistent.

Vendor categories also show varied impacts on `current_price`. For instance, `vendorMetro` has a coefficient of 0.35, while `vendorNoFrills` has a negative effect of -0.34. The strongest posi-

tive effect is observed for `vendorSaveOnFoods` with a coefficient of 2.56, while `vendorWalmart` has a negative coefficient of −1.75. The coefficient for `vendorSaveOnFoods` stands out with a high positive estimate and a narrow credible interval, suggesting a strong and reliable positive effect on `current_price`. Conversely, `vendorWalmart` exhibits a significantly negative effect. The credible intervals for these predictors (as shown in Figure 6) do not include zero, which implies that these effects are statistically significant.

In summary, the results suggest that `old_price` and `vendor` type are significant predictors of `current_price`, while certain months do not significantly impact the price.

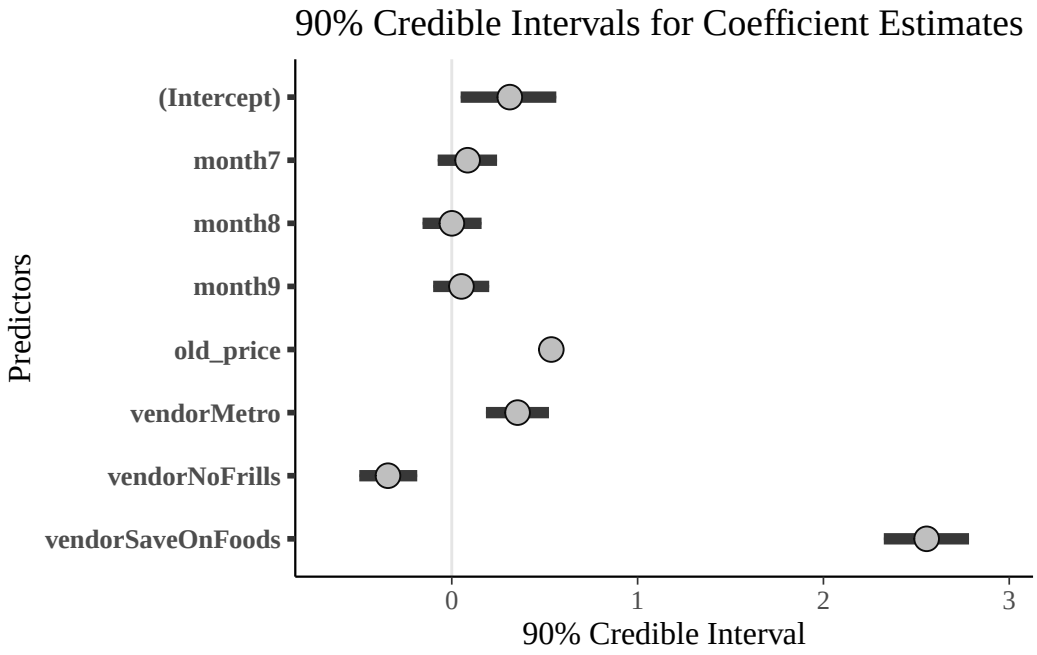

Figure 6

5 Discussion

5.1 What is Done and Implications

This paper investigates the factors influencing the current price of potatoes in Canadian grocery markets. Using a Bayesian regression model, we analyze how vendor type, historical pricing, and seasonal variations contribute to price fluctuations. The study leverages cleaned data from multiple grocery chains, covering both old and current prices of potatoes across different months. The cleaning process involved selecting variables of interest and specific products (white and yellow potatoes), removing missing values, and extracting the month from the date

to provide temporal context. After cleaning, the data was tested for consistency, visualized, and then used to build the models. We first built a linear regression model, followed by a more complex Bayesian regression model. After modeling, we reviewed the model summary and conducted checks and diagnostics for model validity. We decided to use the Bayesian regression model for our research. The analysis aims to provide a clearer understanding of the dynamics that affect potato pricing, with implications for consumers, retailers, and policymakers.

For consumers, especially those in low-income households, understanding these pricing dynamics can inform purchasing decisions and help manage household budgets. By knowing when prices are likely to be lower, consumers can plan their purchases to save money, which is particularly beneficial for those with limited financial resources. Retailers can use this information to optimize inventory management and pricing strategies, potentially increasing competitiveness. Specifically, understanding the seasonal and vendor-driven price fluctuations allows retailers to adjust their stock levels and promotional offers strategically, reducing waste and improving profitability. By anticipating price changes, retailers can also better align their pricing with consumer expectations, enhancing customer satisfaction and loyalty. Agricultural producers and policymakers can benefit from these findings by aligning production and distribution strategies with seasonal trends to stabilize markets and address supply-demand mismatches. For producers, insights into seasonal demand can help with planning harvest times and storage to maximize profits. Policymakers can use this information to design interventions, such as subsidies or support programs, during times of surplus or scarcity to maintain stable prices and support both producers and consumers. Furthermore, these findings can help guide agricultural policies aimed at improving the efficiency of the supply chain and mitigating the effects of price volatility on the agricultural sector.

5.2 Factors Influencing Potato Prices

The analysis illustrates vendor types significantly impact current potato prices. Through the Bayesian regression model, we found that the effect of vendors on pricing is significant. Figure 5 also shows that some vendors consistently have higher prices compared to others, for both old and current prices. This pattern aligns with market principles, as different vendors adopt distinct market strategies. Some vendors may position themselves as premium brands, justifying higher prices, while others, like Walmart, pursue a cost leadership strategy to attract a larger customer base. Additionally, vendors targeting premium customers may prefer to sell high-quality, well-packaged, or organic potatoes, which also contributes to higher prices.

Old prices and sale months also significantly influence current prices from our Section 4. Old prices exhibit a strong positive correlation with current prices, which is reasonable, as the current price is a discounted old price. Figure 2 indicates smaller changes between old and new prices for lower-priced potatoes. The new prices of lower-priced potatoes are similar to the old prices. Higher-priced products typically offer greater profit margins, which allows for more substantial discounts. Additionally, higher-priced products are more likely to experience slower sales, necessitating larger discounts to attract consumers.

Agricultural products are often influenced by seasonal factors, and potatoes in this research as well. The findings indicate an increase in prices during certain months, such as October, likely due to increased demand or seasonal supply constraints. However, the absence of significant effects for months like July and August highlights the variability in seasonality's impact, which may depend on regional harvesting cycles or promotional strategies.

5.3 Limitations and Weaknesses

The study's primary limitation is its reliance on a relatively small clean dataset, which only includes aroundc 700 rows. The data only includes five vendors and only covered the months from June to November. This limited dataset may not fully capture the diverse pricing behaviors across all vendors and throughout the entire year, thereby restricting the generalizability of the findings across broader markets or over longer periods. The analysis only spans a limited timeframe of six months, which may overlook longer-term trends and seasonality effects that influence potato pricing throughout the entire year.

Additionally, during the cleaning the variable `old_price`, products without an old price were filtered out, meaning that non-discounted products were excluded from the analysis. This exclusion limits the ability to fully understand the factors that influence the current price of all potatoes, as it only focuses on products with a discount history.

Another key limitation is the scope of the potato types analyzed. The study focuses only on yellow and white potatoes, which limits its ability to accurately reflect the pricing dynamics of potatoes more broadly. However, a broader classification of "potatoes" would include diverse products such as processed potato items, which could compromise the analysis' accuracy in this dataset.

Furthermore, the dataset lacks important variables such as transportation costs, potato quality, regional market differences, and consumer preferences, which could play significant roles in determining potato prices. The absence of these variables may lead to an incomplete understanding of the pricing dynamics.

5.4 Next Steps

To address the identified limitations and enhance the reliability of the analysis, future research should enhance several key areas. Firstly, expanding the dataset is crucial. Collecting data from more vendors over an extended period of several years. Including all four seasons will allow better accounting of temporal trends and seasonality that affect pricing behaviors. Including data on non-discounted products will enable a more comprehensive analysis of price dynamics of all kinds of potatoes. Besides, future research could benefit from a more sophisticated classification system to better capture the pricing dynamics of a wider variety of potato types while maintaining analytical rigor.

Incorporating additional predictors into the model could provide a more comprehensive understanding of the factors influencing potato prices. Potential variables include transportation costs, quality grades of potatoes, regional market differences, and consumer preferences. These additions would help mitigate the current model's oversimplification and provide a more complete view of the factors driving pricing decisions. However, removing insignificant variables, such as some months, could enhance model simplicity.

Using some advanced modeling techniques could enhance model accuracy. Exploring potential non-linear relationships, models such as generalized additive models (GAMs) or machine learning techniques could be employed to capture the complexities of the vendor-specific pricing dynamics.

Appendix

A Linear model details

Table 4: Summary of Linear Regression Coefficients and Significance Levels

Term	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.307	0.156	1.967	0.050
month7	0.086	0.096	0.897	0.370
month8	0.000	0.095	0.000	1.000
month9	0.052	0.092	0.572	0.568
month10	0.338	0.089	3.820	0.000
month11	0.517	0.097	5.350	0.000
old_price	0.536	0.019	28.496	0.000
vendorMetro	0.358	0.105	3.411	0.001
vendorNoFrills	-0.339	0.095	-3.560	0.000
vendorSaveOnFoods	2.557	0.135	18.982	0.000
vendorWalmart	-1.745	0.239	-7.287	0.000

B Bayesian model details

The Bayesian regression model summaries are shown in Table 6.

B.1 Posterior predictive check

In Figure 7 we implement a posterior predictive check. The close alignment between the observed and predictive distributions suggests that the model captures the data distribution well, though slight deviations may indicate areas for refinement.

B.2 Diagnostics

Figure 8 is a trace plot. It shows the sampling paths of four MCMC chains for each parameter in the model. The chains appear to mix well and remain stationary, suggesting good convergence and no obvious issues with sampling.

Figure 9 is a Rhat plot. It shows that all parameters have Rhat values close to 1, indicating good convergence across the MCMC chains. There is no evidence of non-convergence or poor mixing for any of the parameters in the model.

Table 5: Linear regression model of current price based on month, old price, and vendor

	Linear regression model for current price
(Intercept)	0.31
	(0.05)
month7	0.09
	(0.37)
month8	0.00
	(1.00)
month9	0.05
	(0.57)
month10	0.34
	(<0.01)
month11	0.52
	(<0.01)
old_price	0.54
	(<0.01)
vendorMetro	0.36
	(<0.01)
vendorNoFrills	−0.34
	(<0.01)
vendorSaveOnFoods	2.56
	(<0.01)
vendorWalmart	−1.74
	(<0.01)
Num.Obs.	685
R2	0.858
R2 Adj.	0.856
AIC	1232.4
BIC	1286.8
Log.Lik.	−604.200
RMSE	0.58

Table 6: Bayesian regression model of current price based on month, old price, and vendor

	Bayesian regression model for current price
(Intercept)	0.31
	(0.16)
month7	0.08
	(0.10)
month8	0.00
	(0.09)
month9	0.05
	(0.09)
month10	0.34
	(0.09)
month11	0.52
	(0.10)
old_price	0.54
	(0.02)
vendorMetro	0.35
	(0.11)
vendorNoFrills	−0.34
	(0.09)
vendorSaveOnFoods	2.56
	(0.14)
vendorWalmart	−1.75
	(0.24)
Num.Obs.	685
R2	0.856
R2 Adj.	0.854
Log.Lik.	−607.781
ELPD	−618.4
ELPD s.e.	35.4
LOOIC	1236.8
LOOIC s.e.	70.7
WAIC	1236.8
RMSE	0.58

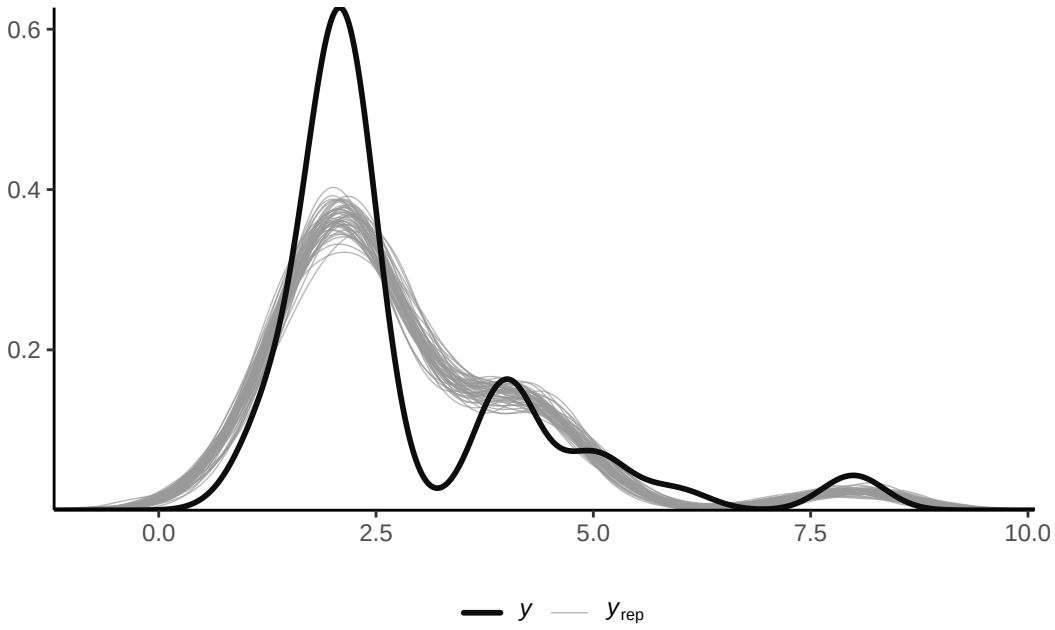

Figure 7: Posterior predictive check

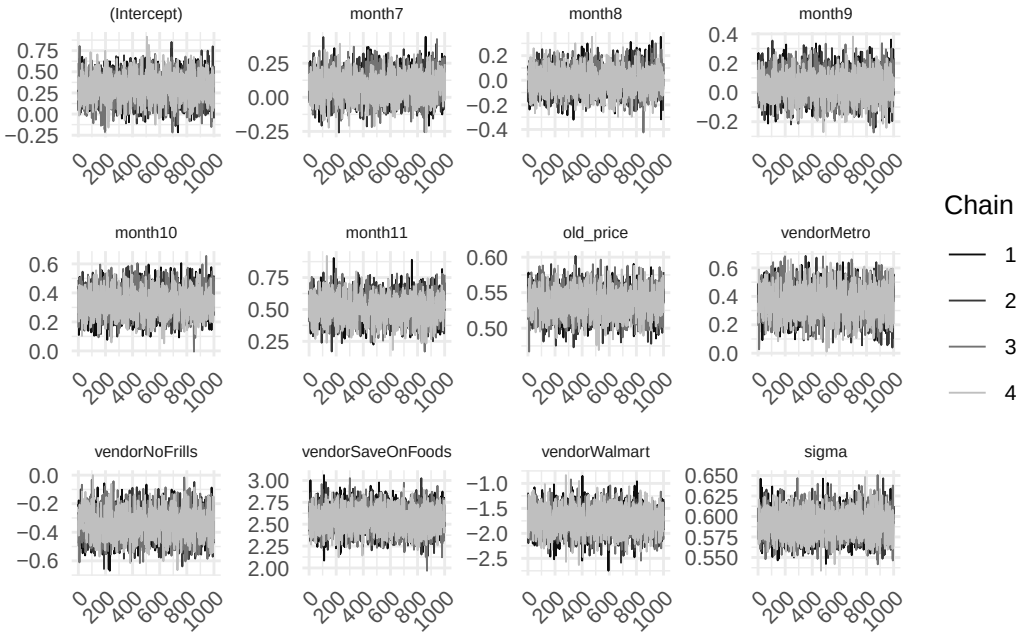

Figure 8: Trace plot

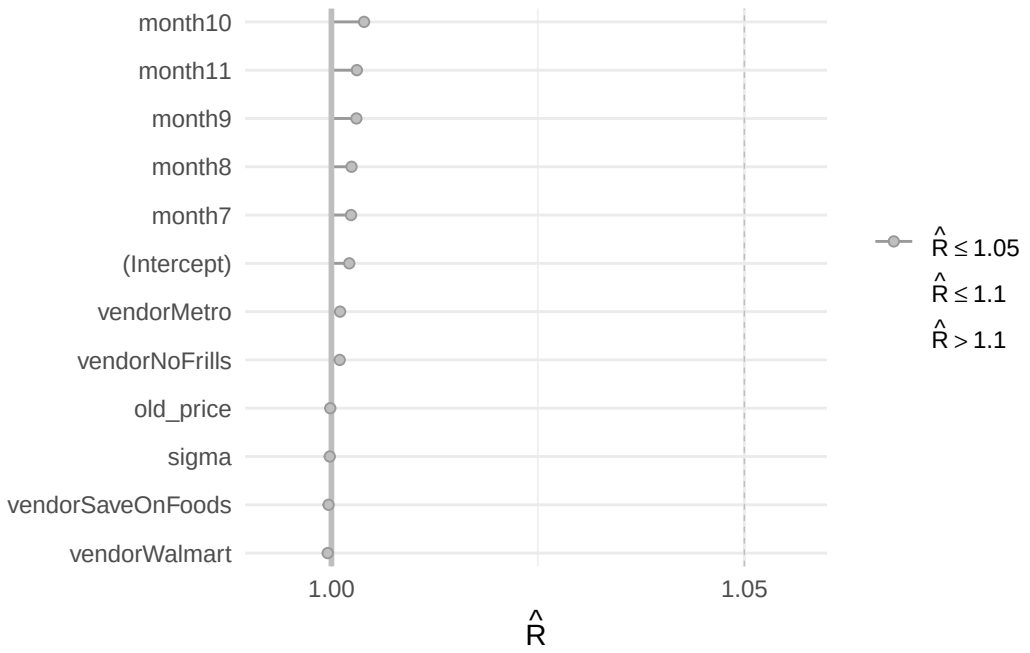

Figure 9: Rhat plot

Figure 10 compares the posterior distributions (left) with the prior distributions (right) for each parameter in the model. The posterior estimates differ significantly from the priors, indicating that the data provides substantial information to update the model parameters.

C Enhancing Observational Data with Survey and Sampling Methods

In this appendix, we explore how the research question addressed in this paper could be enhanced by integrating survey and sampling methodologies with the observational data used. By utilizing these methods, we could gain a more nuanced understanding of how one variable directly influences another which is causal relationships and overcome some of the limitations inherent in observational data.

C.1 Observational Data Limitations and the Role of Surveys

Observational data provides insights into correlations but often falls short when it comes to establishing causal relationships. For instance, in the context of evaluating how certain factors influence potato prices in Canada, there may be confounding variables, such as changes in

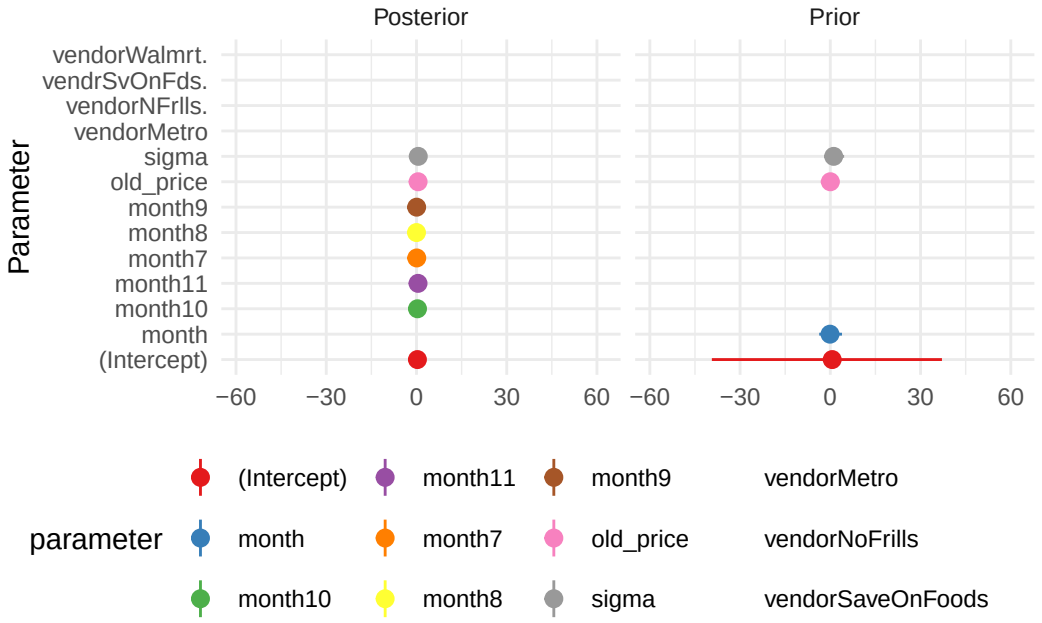

Figure 10: posterior and prior

weather or policies that affect both supply and pricing. These variables can introduce biases, making it challenging to isolate the effect of individual predictors on potato pricing. This is where an augmented survey approach could be beneficial.

To address these limitations, we could design a survey targeted at stakeholders in the potato supply chain, such as farmers, vendors, and consumers. The survey would include questions designed to capture key qualitative factors that might be influencing price variations but are not easily captured in the existing observational data. For example, we could ask farmers about their perceptions of cost trends for inputs like fertilizer or transportation, and vendors about their experiences with consumer demand during certain periods.

C.2 Designing the Hypothetical Survey and Sampling Approach

The proposed survey would utilize stratified random sampling to ensure representation across different segments of the potato supply chain. Stratification variables might include the geographical region, scale of operation (e.g., small vs. large farms), and type of vendor (e.g., grocery chain vs. farmers' market). By segmenting the sample in this way, we can reduce sampling error and ensure that we are capturing variability across different contexts within the supply chain.

Additionally, we would aim for a sample size that is statistically powered to detect meaningful differences between strata, such as variations in production costs, market challenges, or consumer demand trends across different geographical regions or scales of operation. This would involve using simulation methods to estimate the minimum required sample size for key analyses. Simulations could also be used to explore the potential impact of nonresponse bias. By running simulations under different nonresponse scenarios, we could better understand how such bias might affect our results and design appropriate weighting adjustments to compensate.

C.3 Survey Details

The survey would contain the following sections:

1. **Introduction**: An overview of the purpose of the survey, how the information collected will be used, and contact details for questions or concerns.
2. **General Information**: Questions about the respondent's role in the supply chain, including their type of business (e.g., farmer, vendor) and the scale of their operation.
    - What is your role in the potato supply chain? (e.g., farmer, vendor)
    - How large is your operation? (e.g., small farm, large farm, grocery chain)
3. **Production Costs**: Specific questions for farmers regarding costs of inputs such as fertilizer, labor, and transportation.
    - What are your average monthly costs for fertilizer?
    - How have labor costs changed for your farm in the past year?
    - What are the main challenges you face regarding transportation costs?
4. **Demand and Sales Trends**: Questions for vendors about consumer demand trends, challenges faced, and perceived influences on pricing.
    - How would you describe consumer demand for potatoes over the past six months? (e.g., increasing, stable, decreasing)
    - What factors do you think have influenced consumer demand recently?
    - Have you faced any challenges in maintaining stock levels? If so, what were they?
5. **Market Challenges**: Open-ended questions allowing respondents to provide qualitative insights into challenges they are experiencing in the market.
    - What are the biggest challenges you are currently facing in the potato market?
    - Are there any specific external factors (e.g., weather, government policies) that have impacted your business?
6. **Concluding Section**: A thank-you message and information about how respondents can access the survey results if they are interested.
    - Thank you for your time and valuable insights. If you are interested in receiving a summary of the survey results, please provide your contact information below.

C.4 Linking Survey Data to Observational Data

The survey data would be linked with the observational dataset using key identifiers, such as the vendor or farm ID. This linkage would allow us to augment the existing dataset with new variables derived from the survey responses. For example, the inclusion of self-reported data on production costs or perceived consumer preferences would provide additional context to our analysis of potato pricing trends.

The augmented dataset would then be used in a series of regression models to assess whether the new variables improve model fit or explain additional variance in potato prices. Specifically, we could explore interaction effects between observed variables (e.g., monthly average price) and survey-derived variables (e.g., self-reported demand trends) to determine whether changes in specific factors, such as demand trends or production costs, have a direct impact on pricing. This approach helps establish causal links between influencing factors and price changes, rather than merely identifying correlations.

C.5 Simulation and Validation

To validate our survey approach and ensure robustness, we could conduct simulations to assess the reliability of survey-based estimates. This might involve generating synthetic datasets under different assumptions about survey response distributions and comparing the results to those from the observational data alone. Such simulations can provide insight into how survey augmentation could enhance causal inference. Additionally, we could use cross-validation techniques to test the predictive performance of models augmented with survey data, or conduct sensitivity analyses to understand how changes in key assumptions or survey responses affect the final results. We can also compare survey-derived estimates with existing benchmarks or known external data sources to evaluate their accuracy and consistency.

References

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

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2024. *Rstanarm: Bayesian Applied Regression Modeling via Stan*. https://mc-stan.org/rstanarm.

Government of Canada. 2024. *Monthly Average Retail Prices for Selected Products*. https://open.canada.ca/data/en/dataset/8015bcc6-401d-4927-a447-bb35d5dfcc91/resource/5ab003e5-fea3-40b8-8c10-8ac8988bfc93.

Hadley Wickham and others. 2023. *testthat: Unit Testing for R*. https://CRAN.R-project.org/package=testthat.

Hadley Wickham and the tidyverse team. 2023. *tidyverse: Easily Install and Load the 'Tidyverse'*. https://CRAN.R-project.org/package=tidyverse.

Hadley Wickham, Lionel Henry, and other contributors. 2023. *tidyr: Tidy Messy Data*. https://CRAN.R-project.org/package=tidyr.

Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. 2023. *dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Khakbazan, M, R Henry, R Mohr, R Peters, S Fillmore, V Rodd, and A Mills. 2015. "Economics of Organically Managed and Conventional Potato Production Systems in Atlantic Canada." *Canadian Journal of Plant Science* 95 (1): 161–74.

Michael B. Goodrich, Ben S. McFadden, and Jonah Gabry. 2023. *Bayesplot: Plotting for Bayesian Models*. https://cran.r-project.org/package=bayesplot.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Müller, Kirill, and Lorenz Walthert. 2024. *Styler: Non-Invasive Pretty Printing of r Code*. https://CRAN.R-project.org/package=styler.

Neal Richardson and Apache Arrow contributors. 2023. *arrow: Integration to 'Apache Arrow'*. https://CRAN.R-project.org/package=arrow.

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

Ritchie, Thomas Frederick, and G Gordon Dustan. 1940. *The Potato in Canada*. Division of Horticulture, Canada Department of Agriculture.

Sam Firke. 2023. *janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Thomas Lin Pedersen. 2023. *patchwork: The Composer of Plots*. https://CRAN.R-project.org/package=patchwork.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Yihui Xie. 2023. *knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://CRAN.R-project.org/package=knitr.