Influences on Egg Pricing: The Role of Brands and Vendors in Canadian Retail*

Seasonal Trends and Strategic Impacts Revealed

Caichen sun

December 3, 2024

This study examines the factors influencing egg prices across major Canadian grocery stores, using a linear regression model to analyze transaction data over an eight-month period. It was found that brand recognition and vendor strategies significantly affect pricing, with premium brands consistently commanding higher prices and seasonal trends impacting vendor pricing adjustments. These findings underscore the importance of strategic brand management and the need for responsive pricing strategies in the retail sector. The insights provided by this research help consumers and policymakers understand how pricing decisions are made in the retail market, highlighting the influence of brand and vendor strategies on everyday consumer costs..

1 Introduction

In today's economic environment, characterized by global market volatility and fluctuating supply chains, understanding the pricing dynamics of everyday commodities is crucial. This study examines the determinants of egg prices in the Canadian retail market, a critical segment due to its essential nature and economic sensitivity. Despite extensive research on retail pricing, there remains a significant gap regarding the combined effects of brand influence, vendor strategies, and seasonal variations within Canada. This paper addresses this gap by employing a linear regression model to analyze data from Jacob Filipp's publicly accessible grocery dataset (Filipp 2024), which includes detailed transaction records over an eight-month period.

The analysis reveals that brand recognition has a significant impact on pricing, with premium brands maintaining higher prices. Furthermore, vendor strategies and seasonal trends critically influence price adjustments in response to market and supply conditions. These findings offer

*Code and data are available at: https://github.com/RohanAlexander/starter_folder.

nuanced insights into market responsiveness and are crucial for retailers aiming to optimize pricing strategies and for policymakers focused on consumer protection.

The paper is structured to provide a detailed account of the methodology and data sources in the subsequent section, followed by a comprehensive analysis of the findings. The conclusion synthesizes these insights, discusses potential study limitations, and suggests directions for future research. This approach not only fills a significant gap in the existing literature but also enhances our understanding of economic factors influencing retail pricing in a volatile market.

2 Data

2.1 Overview

The data used in this analysis comes from Jacob Filipp's publicly available groceries dataset (Filipp 2024). The analysis utilizes the statistical programming language R (R Core Team 2024) and several key packages from the R ecosystem. For data manipulation and transformation, the analysis employs tidyverse (Wickham et al. 2019), dplyr (Wickham et al. 2024), and arrow (Richardson et al. 2024). Date handling is managed through the lubridate (Spinu, Grolemund, and Wickham 2024) package, while data visualization is accomplished using ggplot2 (Wickham 2016) and gridExtra (Auguie 2017). Data import functionality is provided by readr (Wickham, Hester, and Bryan 2024), and testing is implemented using testthat (Wickham 2024).

The dataset captures grocery purchasing information spanning from February 28, 2024 to November 13, 2024, providing approximately 8.5 months of transaction data. This contemporary dataset allows for analysis of recent consumer purchasing patterns and price trends in the grocery retail sector.

2.2 Measurement

The Project Hammer data collection process employs a systematic approach to monitor pricing trends in the Canadian grocery market and organize the information into a well-structured dataset. The data is predominantly gathered through automated web scraping from the online platforms of major grocery retailers, capturing real-time price variations for a wide range of products. Vendors include notable names like Voilà, T&T, Loblaws, No Frills, Metro, Galleria, Walmart, and Save-On-Foods. Each dataset entry represents a timestamped record of the listed price, marked as "nowtime," along with details such as the vendor, product ID, product name, brand, unit of measurement, and both current and historical prices. This approach adds a temporal dimension to the data, enabling the study of pricing patterns, marketing strategies, and competitive dynamics across various regions and timeframes. The automated nature of the process ensures regular updates and consistent data quality, creating a reliable resource for analyzing market trends and informing practical decisions.

2.3 Clean Data

The cleaned dataset was prepared using a series of data manipulation steps implemented in R, leveraging libraries such as tidyverse (Wickham et al. 2019), dplyr (Wickham et al. 2024), arrow (Richardson et al. 2024), and lubridate (Spinu, Grolemund, and Wickham 2024). Initially, raw data were sourced from two CSV files: hammer-4-raw.csv and hammer-4-product.csv. These datasets were combined using a left join operation on the product_id and id fields, resulting in a dataset that included various product attributes such as current and historical prices, product name, brand, and vendor.

To refine the dataset for analysis, products were filtered to include only those with names containing "Eggs" or "Egg", excluding any products associated with a wide range of non-relevant terms (e.g., "Chocolate", "Candy", "Toy"). Further, the dataset was limited to products sold by specific vendors—Loblaws, T&T, and NoFrills—known for their relevance in the analysis context.

The dataset was then cleansed to ensure that both the current_price and old_price fields contained non-missing numeric values. Additionally, a new variable month was created from the nowtime timestamp to facilitate temporal analysis.

The final cleaned dataset consists of selected attributes (year_month, current and old prices, product name, brand, and vendor) and is stored in both CSV and Parquet formats (Richardson et al. 2024) to support subsequent analytical procedures. This data preparation phase ensures that the dataset is optimally structured for detailed examination of price trends and product availability over time.

2.4 Outcome variables

current_price: This variable captures the current market price of eggs at the time of data extraction. It is crucial for assessing the immediate pricing environment and evaluating how prices are positioned relative to historical prices. Analyzing this variable helps understand the current economic factors influencing egg prices.

The graph depicted in (Figure 1) illustrates the monthly trends in the average current price from February to November 2024, reflecting the dynamics within the market. Initially, there is a noticeable decline in average prices from February, starting at approximately 6.2 units, to a low point in June at just above 5.6 units. This downward trend suggests a possible seasonal influence or a response to market oversupply. From July onwards, there is a gradual recovery in prices, which stabilizes through August and early September. Notably, a sharp increase occurs from September to November, where prices soar to over 6.4 units. This sudden rise may be indicative of increased demand or reduced supply, highlighting the importance of further investigation into market conditions and external factors such as economic policies or global events that could influence these fluctuations.

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```

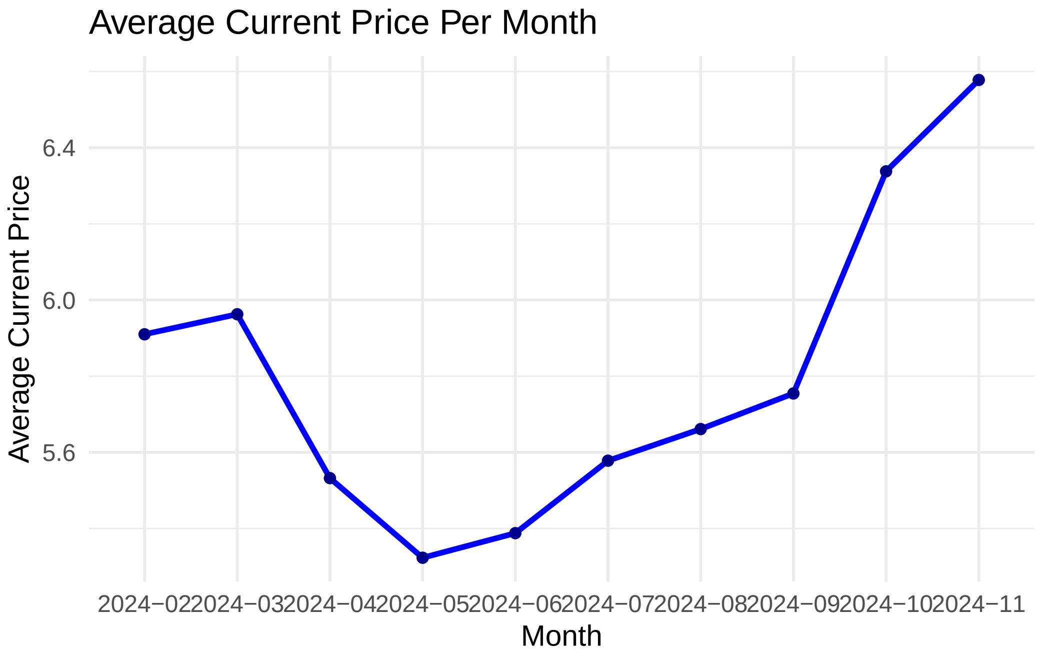

Figure 1: This graph displays the average current price per month from February to November 2024, illustrating significant fluctuations throughout the year.

2.5 Predictor variables

year_month: Derived from the nowtime timestamp, this variable is formatted as "YYYY-MM" and captures the specific month and year when each data record was collected, specifically within the year 2024. This segmentation allows for a concentrated analysis of price trends and market behaviors over the course of a single year, highlighting how seasonal factors, marketing campaigns, and other temporal events influence pricing and consumer decisions during different months of 2024. The extraction of year_month from nowtime ensures precision in time-related analyses, making it possible to trace and understand the nuances of market dynamics and pricing strategies within a defined annual scope. This approach is particularly useful for identifying periods of high fluctuation or stability in prices, facilitating targeted business strategies and economic forecasting.

vendor: This variable represents the grocery store where the data was collected, specifically Loblaws, Metro, Walmart, and NoFrills. These vendors were chosen from the original dataset, which included a broader range of supermarkets, to focus on stores with significant market presence and diverse customer bases. As a predictor variable, vendor is critical for analyzing how pricing strategies vary across different retailers within the same market. Each vendor may have unique pricing models, promotional strategies, and supply chain dynamics, which can significantly influence the current price of egg products. By including this variable, the analysis can explore the extent to which vendor-specific factors contribute to price variations and consumer purchasing behavior.

brand: The brand associated with the egg products, which can vary from premium to store brands. Analyzing this variable helps understand brand influence on pricing and consumer preference, particularly across different market segments.

old_price: Represents the price before any current updates, allowing for analysis of historical pricing trends and their influence on current prices. This variable can help identify patterns in pricing strategies, such as frequent discounts or price hikes, and their effects on consumer demand.

product_name: Includes the specific names given to egg products, such as descriptors indicating whether the eggs are whole, liquid, or processed in other forms. This variable is key to differentiating product types within the category and analyzing pricing by product type.

The graph presented in Figure (Figure 2) illustrates the trends in the average old price per month from February to November 2024. This chart shows a volatile pattern, beginning with a slight increase followed by a dip in April, after which the prices gradually recover and escalate significantly from July onwards, peaking in October. The fluctuations suggest a dynamic market behavior that might be driven by seasonal demand or supply changes, promotional activities, or macroeconomic conditions."

Figure Figure 3 showcases a comprehensive view of the monthly average price differences for the year 2024, using dual visualization techniques to highlight the data's dynamics. The line graph in the upper section captures the fluctuating nature of the price differences, starting with a notable decrease from February to April. This period may reflect a strategic adjustment in pricing policies or a market reaction to external economic pressures. The prices stabilize somewhat in the middle months, indicating a period of market correction or adaptation to consumer behaviors.

As the year progresses, the data reveals a dramatic spike in price differences beginning in October, peaking sharply in November. This upward trend could be associated with increased consumer activity during the holiday season, or possibly supply constraints driving prices upward. The corresponding bar graph below serves to emphasize these fluctuations by presenting the data in a segmented format, which allows for a clear comparison of month-to-month changes in price differences. Each bar graphically represents the magnitude of change, making the relative stability of the summer months and the volatile movements in the later months particularly evident.

Both graphical representations underscore the significant volatility within the market across 2024, suggesting that pricing strategies were likely influenced by a combination of economic

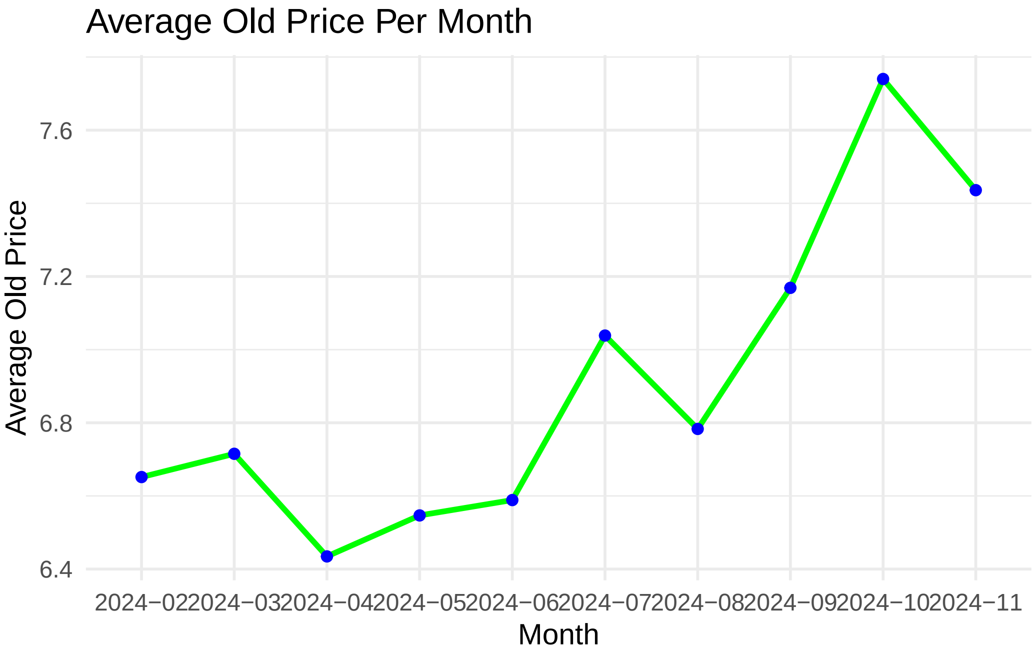

Figure 2: The plot demonstrates the monthly progression of average old prices throughout 2024, characterized by notable volatility and a sharp rise in the latter half of the year, emphasizing the instability and reactive nature of the market.

conditions, consumer demand, and possibly promotional campaigns. The dual-graph format effectively conveys the temporal trends and highlights the significant shifts in market behavior, providing a visually compelling narrative of how average price differences evolved throughout the year.

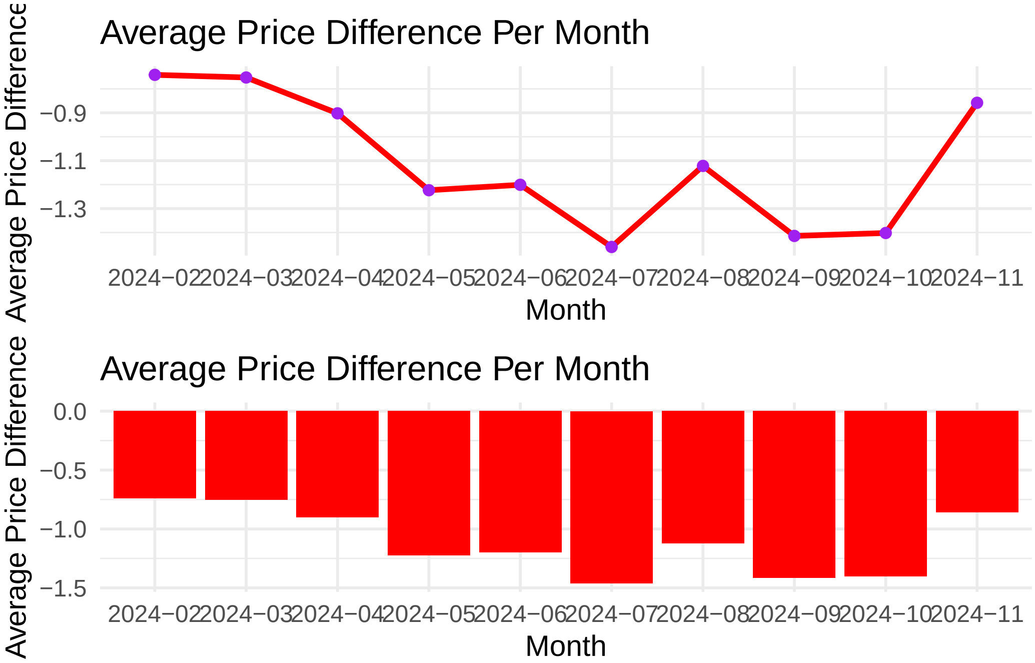

Figure 3: Figure Figure 3: This composite graph illustrates the monthly average price differences throughout 2024, featuring both a line graph and a bar graph for enhanced clarity on trend and magnitude analysis. The line graph above highlights the temporal fluctuations, with notable decreases in early months followed by stabilization and a sharp increase towards the end of the year. The bar graph below complements this by visually emphasizing the month-to-month variance, clearly depicting the periods of relative stability and significant change. Together, these visualizations underscore the dynamic nature of pricing strategies in response to varying market conditions and external economic factors.

The scatter plot in (Figure 4) provides a detailed visualization of the relationship between old prices and current prices, marked by a notable positive correlation. The blue points represent individual data entries, dispersed across the plot, indicating the variety in how much old prices have influenced the current prices. The red line, representing a linear model fit, further underscores this relationship by showing a consistent upward trajectory, implying that higher old prices are typically associated with higher current prices.

A more detailed observation of the plot reveals that while the general trend suggests a linear relationship, there are variations that could be indicative of other influencing factors such as market anomalies, supply and demand changes, or different pricing strategies applied during specific periods. The concentration of data points around the line suggests a strong correlation, yet the spread above and below the line could also hint at occasional significant deviations from the expected pricing pattern.

```
`geom_smooth()` using formula = 'y ~ x'
```

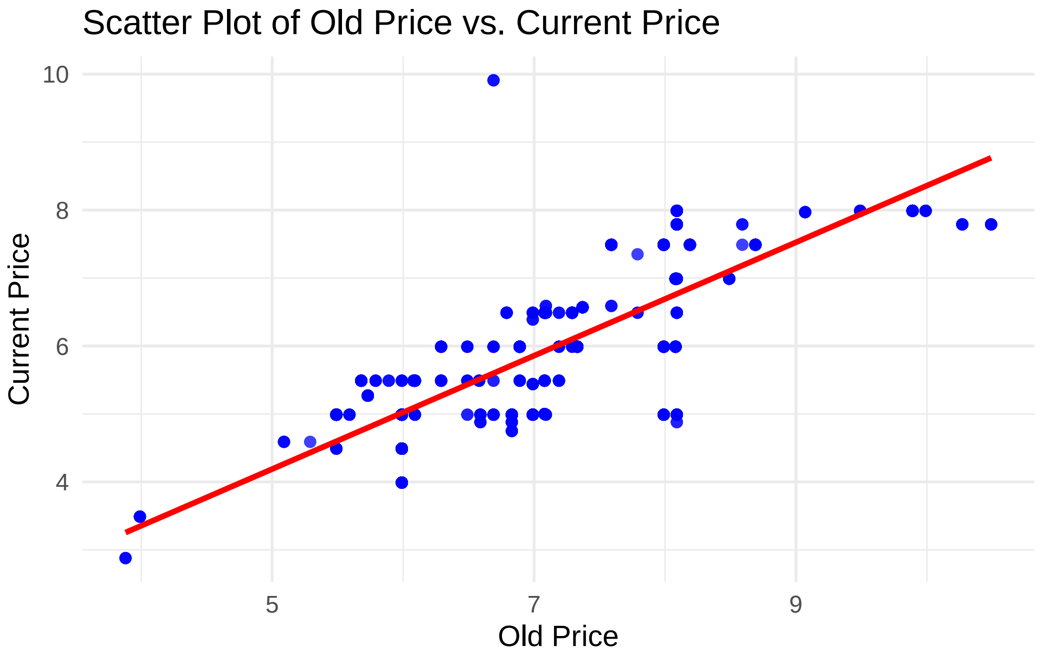

Figure 4: Figure Figure 4: This scatter plot displays the relationship between old prices and current prices, with a red line representing a linear fit that suggests a positive correlation. The distribution of points indicates that higher old prices are generally associated with higher current prices, reflecting potentially consistent market behavior or pricing strategies.

The density plot depicted in (Figure 5) visually compares the distribution of old and current prices. The plot is marked by two distinct, overlapping density curves: the blue curve for old prices and the red curve for current prices.

The blue curve peaks sharply around the price point of 6, suggesting a high concentration of older prices within a lower range. This peak diminishes as the price increases, reflecting a limited occurrence of higher old prices. Conversely, the red curve, representing current prices, shows its peak around the price of 8, indicating that the most frequent current prices are higher than the most frequent old prices. The broader base and multiple peaks of the red curve also suggest a greater variability in current prices compared to the past.

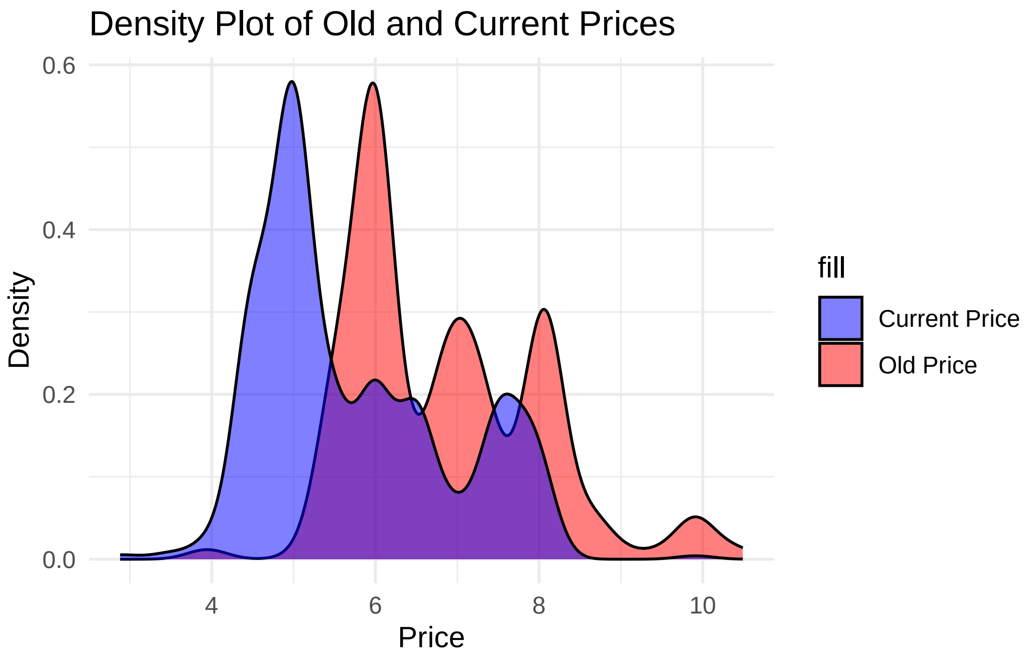

Figure 5: Figure Figure 5: This density plot displays the distributions of old and current prices. The blue curve represents old prices with a concentrated peak at lower values, while the red curve for current prices shows a higher peak and greater dispersion. The comparison highlights shifts towards higher price ranges and increased variability in recent pricing strategies.

The intersection of the two curves represents price points that are consistent between the two time periods, highlighting some stability in certain pricing segments. The observed shifts in the peaks and the broader spread in the current prices suggest an overall trend toward price increases and greater variability in pricing strategies over time.

The boxplot depicted in Figure Figure 6 provides a detailed comparative analysis of the distribution of current prices by different vendors: Loblaws, Metro, NoFrills, and Walmart. The visualization employs different colors to differentiate each vendor, which helps in quickly assessing the pricing landscape as outlined by their respective market positioning.

Detailed Analysis:

Loblaws (Red): The boxplot for Loblaws indicates a median price slightly above 6 dollars, with the interquartile range (IQR) extending from just under 5 dollars to just over 7 dollars. This suggests a moderate price spread. The presence of outliers below 4 dollars and near $8 points to some exceptional pricing instances that are significantly lower or higher than the typical range. This could imply promotional deals or premium product offerings that diverge from their standard price points. Metro (Green): Metro displays a slightly lower median price than Loblaws, hovering around 6 dollars. The IQR is tighter, ranging from about 5 dollars to 7 dollars, with fewer outliers, suggesting a more consistent pricing strategy with less variation in product pricing. This could indicate a stable market strategy aimed at maintaining a steady customer base through consistent pricing. NoFrills (Cyan): With the lowest median price just below 6 dollars, and a very narrow IQR from 5 dollars to around 6.5 dollars, NoFrills positions itself as the most budget-friendly option among the compared vendors. The limited range and fewer outliers signify a focused approach to pricing, potentially appealing to cost-conscious consumers looking for consistent low prices. Walmart (Purple): Walmart shows the highest median price at close to 7 dollars and a wider IQR that stretches from around 5 dollars to 8 dollars. This broad range and outliers on both ends suggest a diverse product offering that spans from more affordable to premium options. Walmart's pricing strategy appears to accommodate a wide array of consumer demographics, from budget shoppers to those willing to pay more for higher-end products.

The bar chart in Figure Figure 7 illustrates the average current prices of various egg brands, showing significant price variation across different brands. The chart is organized in ascending order of price from left to right, offering a visual comparison of brand pricing strategies within the market.

Detailed Analysis:

Lowest Priced Brands: The bar for "No Name" is the shortest, indicating it has the lowest average price among the brands represented. This suggests that "No Name" is positioned as a budget-friendly option in the market.

Middle Range Brands: Brands like "Great Value," "Burnbrae Farms," "Gray Ridge," and "PC Blue Menu" exhibit mid-range prices. These brands likely balance between affordability and perceived quality, targeting middle-income consumers.

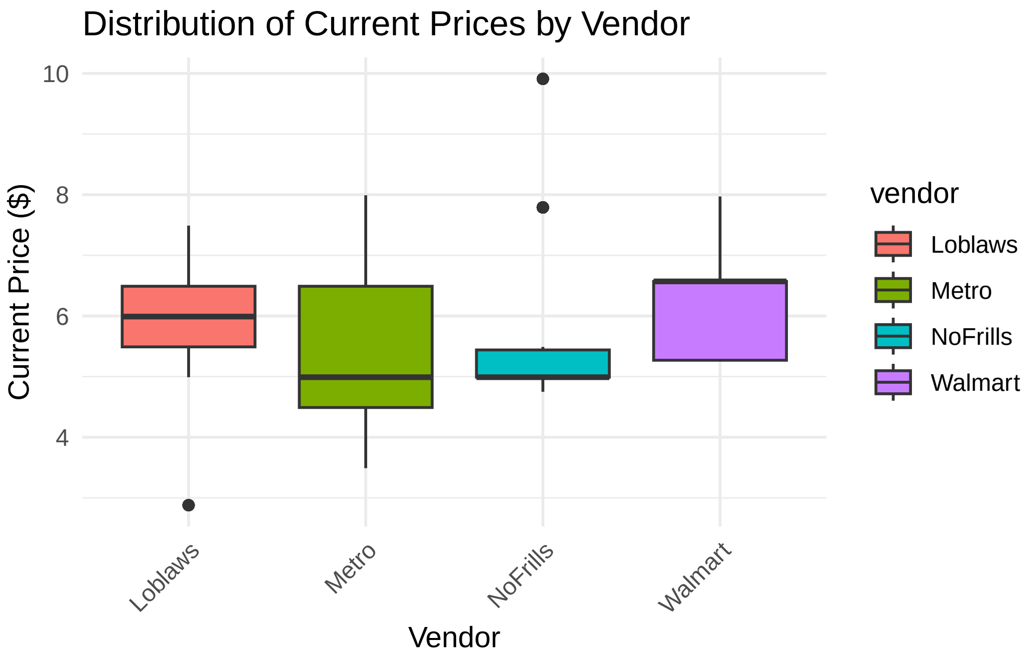

Figure 6: Figure Figure 6: This boxplot visualizes the current price distribution across four major vendors: Loblaws, Metro, NoFrills, and Walmart. Each vendor's price range, median, and variability are distinctly displayed, illustrating differences in pricing strategies and market positioning. The presence of outliers in certain vendors underscores occasional price extremes that deviate from the typical range.

Highest Priced Brands: "Selection" and "GoldEgg" appear at the higher end of the pricing spectrum, as indicated by their taller bars. These brands might be positioned as premium options, possibly offering organic or free-range eggs, which typically command higher prices. Variability and Market Positioning:

The chart shows a wide range of pricing strategies, from economical to premium, reflecting the diverse consumer base and varied consumer preferences in the egg market. Each brand's position on the chart directly relates to its market segmentation strategy, aiming at specific demographic groups based on their spending ability and preference for quality. The use of color for each brand enhances visual differentiation and helps to quickly associate each bar with its respective brand. This is particularly useful in distinguishing between brands that might offer similar products but at different price points.

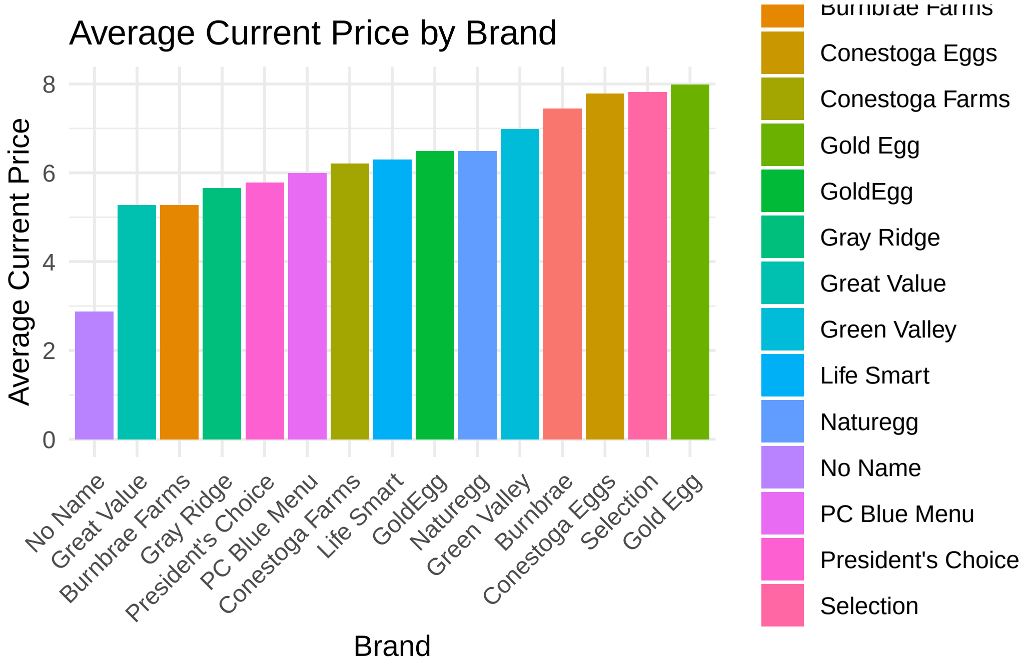

Figure 7: Figure Figure 7: This bar chart displays the average current prices of various egg brands, arranged in ascending order. It highlights the broad range of pricing strategies from budget to premium within the egg market. The visualization aids in understanding how each brand positions itself in terms of price relative to perceived quality and market segmentation.

3 Model

The goal of our modelling strategy is twofold. Firstly,...

Here we briefly describe the Bayesian analysis model used to investigate… Background details and diagnostics are included in **?@sec-model-details**.

3.1 Model set-up

The objective of this analysis is to employ a **linear regression model** in R to predict the current price of egg products (`current_price`) based on historical pricing data and contextual factors. The outcome variable, `current_price`, is continuous and represents the most recent selling price of egg products across different vendors.

3.2 Mathematical Representation

The model is mathematically represented as:

$$current\_price_i = \beta_0 + \beta_1 \cdot old\_price_i + \sum(\beta_{brand_j} \cdot brand_{ij}) + \sum(\beta_{vendor_k} \cdot vendor_{ik}) + \sum(\beta_{month_l} \cdot month_{il}) + \epsilon_i$$

Where:

- $current\_price_i$: The observed current price of egg products for the i-th entry in the dataset.
- old_price_i: The previous recorded price of the product, reflecting historical price trends.
- $\beta_{brand_j} \cdot brand_{ij}$: Represents the categorical effect of product brands, capturing unique pricing strategies or brand reputation influences.
- $\beta_{vendor_k} \cdot vendor_{ik}$: Accounts for pricing differences between vendors (e.g., Walmart, Loblaws), representing the unique market dynamics of each store.
- $\beta_{month_l} \cdot month_{il}$: Captures monthly variations in pricing, reflecting potential seasonal promotions or market demand shifts.
- ϵ_i: The error term, assumed to follow a normal distribution with mean 0 and constant variance, accounting for unexplained variability.

This setup allows us to explore how historical prices, brand, vendor, and month collectively influence current prices, providing a basis for strategic pricing and market analysis.

3.3 Model Justification

A linear regression model was chosen for this analysis to predict the current price of egg products (`current_price`) based on historical prices and contextual factors. This approach balances interpretability and efficiency, making it well-suited for analyzing pricing data where relationships among predictors (e.g., old price, brand, vendor, and month) can be reasonably approximated as linear.

The decision to use linear regression stems from its ability to provide clear insights into the contribution of individual predictors. Historical price (`old_price`) is a strong determinant of current price, as trends in pricing are often sustained over time. Additionally, categorical variables such as `brand`, `vendor`, and `month` capture market dynamics, including differences in branding strategies, vendor pricing policies, and potential seasonal variations.

While alternative modeling approaches, such as machine learning techniques or Bayesian regression, could also be considered, linear regression offers distinct advantages in this context:

Interpretability: The model's coefficients directly quantify the impact of each predictor on the outcome variable, facilitating straightforward conclusions about pricing influences.

Efficiency: Linear regression is computationally efficient and performs well with datasets that are adequately sized for the number of predictors.

Simplicity: With limited predictors, linear regression avoids overfitting and provides robust estimates without the need for extensive tuning or computational overhead.

Machine learning techniques, such as decision trees or neural networks, were not selected due to their potential complexity and reduced interpretability. While these models excel at capturing non-linear relationships, they often require larger datasets to generalize effectively and risk overfitting with smaller data. Bayesian regression, on the other hand, introduces flexibility through prior distributions but adds complexity that may not be necessary for this dataset, where strong prior knowledge or uncertainty quantification is less critical.

Linear regression also aligns with the goal of understanding the underlying factors driving egg product pricing. By focusing on the relationships between historical prices, brand effects, vendor strategies, and monthly variations, the model sheds light on actionable insights for stakeholders in the retail market.

This choice ensures a balance between statistical rigor and practical application, making it a reliable and interpretable tool for price prediction in the egg product market.

3.4 Assumptions and Limitations

Linear regression makes several key assumptions that must be considered when interpreting the results. First, it assumes a linear relationship between the predictors (e.g., old price, brand, vendor, and month) and the outcome variable (current price). While this assumption provides simplicity and interpretability, it may not fully capture more complex, non-linear relationships in the data. If significant non-linear trends exist, the model's predictions could be biased or imprecise.

Another critical assumption is that the residuals (the differences between observed and predicted prices) are independent and identically distributed (i.i.d.) with constant variance (homoscedasticity). Violations of this assumption, such as heteroscedasticity (where variance changes with predictor levels) or autocorrelation (where residuals are correlated), could indicate model misspecification or the need for transformation or additional variables. Diagnostic checks such as residual plots are necessary to validate these assumptions.

The model also assumes that all relevant predictors affecting the current price have been included. Factors such as unrecorded promotions, sudden market disruptions, or broader economic changes may not be reflected in the data, potentially leading to omitted variable bias. Additionally, the dataset's granularity may limit the ability to capture more subtle or localized pricing influences.

Outliers pose another limitation. Extreme values in `current_price` or `old_price` can exert undue influence on the regression coefficients, potentially skewing the results. While linear regression is sensitive to outliers, robust diagnostics and preprocessing steps, such as trimming or winsorizing, can mitigate this issue.

The model presumes that relationships between predictors and the outcome variable remain consistent across the dataset. However, if subgroups exist, such as differences in pricing strategies among premium and budget brands or regional vendor variations, the model may fail to account for these dynamics without additional interaction terms or hierarchical structures.

Lastly, the model's generalizability depends on the representativeness of the data. If the dataset does not adequately reflect the diversity of brands, vendors, or seasonal conditions, predictions may not extend well to unseen or new contexts. For example, shifts in consumer behavior or external market trends could limit the model's applicability over time.

While linear regression provides an efficient and interpretable framework for predicting current prices, these assumptions and limitations emphasize the importance of robust data preparation, diagnostic validation, and cautious interpretation of the results. Proper attention to these factors will ensure the model's findings are both meaningful and actionable.

3.5 Model Validation

To validate the linear regression model, several diagnostic and evaluation steps were performed to ensure its reliability and effectiveness:

- **Training and Testing Split**: The dataset was divided into training (80%) and testing (20%) subsets. The model was trained on the training set and its predictive performance was evaluated on the test set. This split helps assess how well the model generalizes to new, unseen data.
- **Root Mean Square Error (RMSE)**: RMSE was calculated on both the training and testing sets to measure the average prediction error. A small difference between the two RMSE values indicates that the model is not overfitting and generalizes well to new data.

- **Residual Analysis**: Residuals (differences between observed and predicted values) were plotted against predicted values. The absence of patterns in the residuals suggests that the assumptions of linearity and homoscedasticity (constant variance of residuals) hold. Additionally, residual histograms and Q-Q plots were used to check for normality of residuals.
- **Adjusted R-Squared**: The adjusted R-squared was used to evaluate how well the predictors collectively explain the variability in the current price, while accounting for the number of predictors in the model.
- **Multicollinearity Check**: Variance Inflation Factor (VIF) was calculated for all predictors to identify multicollinearity. Predictors with VIF values above 5 were flagged for further review to ensure the stability of coefficient estimates.

3.5.1 Limitations of Validation

While the above steps provide a robust framework for validating the model, it is important to note that linear regression has inherent limitations. For example, the assumption of linearity may not fully capture complex relationships, and the presence of outliers can disproportionately influence the model. Furthermore, the representativeness of the training data plays a critical role in determining the accuracy of predictions when applied to new data.

Overall, these validation techniques ensure that the linear regression model is robust, interpretable, and performs well within the constraints of the dataset.

4 Results

The results of the linear regression model in Table 1 highlight significant brand-specific effects on current prices. Brands such as Gold Egg demonstrate the highest positive effect (estimate = 1.42, $p < 0.001$), reflecting their premium positioning in the market. In contrast, budget brands like No Name (estimate = -0.91, p = 0.0017) and PC Blue Menu (estimate = -1.01, $p < 0.001$) are associated with significantly lower prices, appealing to cost-conscious consumers. Some brands, including Great Value and Green Valley, show coefficients that are not statistically significant, indicating negligible pricing differences compared to the baseline brand.

This analysis, illustrated in @ref(tbl-sum), emphasizes the role of brand differentiation in pricing strategies. Premium brands maintain higher pricing, likely due to perceived quality or marketing strategies, whereas budget brands cater to price-sensitive market segments. The findings suggest that pricing strategies should consider brand positioning and consumer perception to optimize revenue and market share.

```
`geom_smooth()` using formula = 'y ~ x'
```

Table 1

Table 2: Model Coefficients Summary

term	estimate	std.error	statistic	p.value
(Intercept)	1.6000522	0.2603957	6.1446948	0.0000000
old__price	0.7491045	0.0219144	34.1832327	0.0000000
year__month2024-03	-0.1027761	0.1582203	-0.6495761	0.5160554
year__month2024-04	-0.3252499	0.1586061	-2.0506776	0.0404541
year__month2024-05	-0.7395212	0.1578963	-4.6835892	0.0000030
year__month2024-06	-0.6121613	0.1589920	-3.8502654	0.0001224
year__month2024-07	-0.7178133	0.1706016	-4.2075423	0.0000272
year__month2024-08	-0.5365126	0.1609820	-3.3327495	0.0008787
year__month2024-09	-0.7924422	0.1622764	-4.8832867	0.0000011
year__month2024-10	-0.4988621	0.1646735	-3.0294019	0.0024880
year__month2024-11	-0.0663153	0.1639819	-0.4044063	0.6859658
brandBurnbrae Farms	-0.2856025	0.0892776	-3.1990381	0.0014047
brandConestoga Eggs	-0.5143151	0.1827258	-2.8146831	0.0049398
brandConestoga Farms	0.1023573	0.1030699	0.9930864	0.3208117
brandGold Egg	1.4185396	0.1293942	10.9629326	0.0000000
brandGoldEgg	0.0974026	0.1380819	0.7053971	0.4806615
brandGray Ridge	-0.3637725	0.1456298	-2.4979258	0.0125878
brandGreat Value	0.0482701	0.1912856	0.2523456	0.8008050
brandGreen Valley	0.2297679	0.2012578	1.1416597	0.2537592
brandLife Smart	-0.1770192	0.0869451	-2.0359883	0.0419091
brandNaturegg	0.0942114	0.2812991	0.3349154	0.7377310
brandNo Name	-0.9087643	0.2890025	-3.1444856	0.0016931
brandPC Blue Menu	-1.0093582	0.2648614	-3.8108914	0.0001435
brandPresident's Choice	-0.4819220	0.1150557	-4.1885976	0.0000295
brandSelection	-0.0248473	0.1077938	-0.2305074	0.8177258
vendorMetro	-0.4012132	0.0434509	-9.2337085	0.0000000
vendorNoFrills	-0.5725795	0.0748629	-7.6483707	0.0000000
vendorWalmart	0.0688303	0.1234702	0.5574651	0.5772844

Scatter plot Figure 8 showing Actual vs Predicted values. Points should align along the red dashed line (y = x) if predictions are perfect. The blue line shows the smoothed trend of the relationship, and the gray band represents the 95% confidence interval.

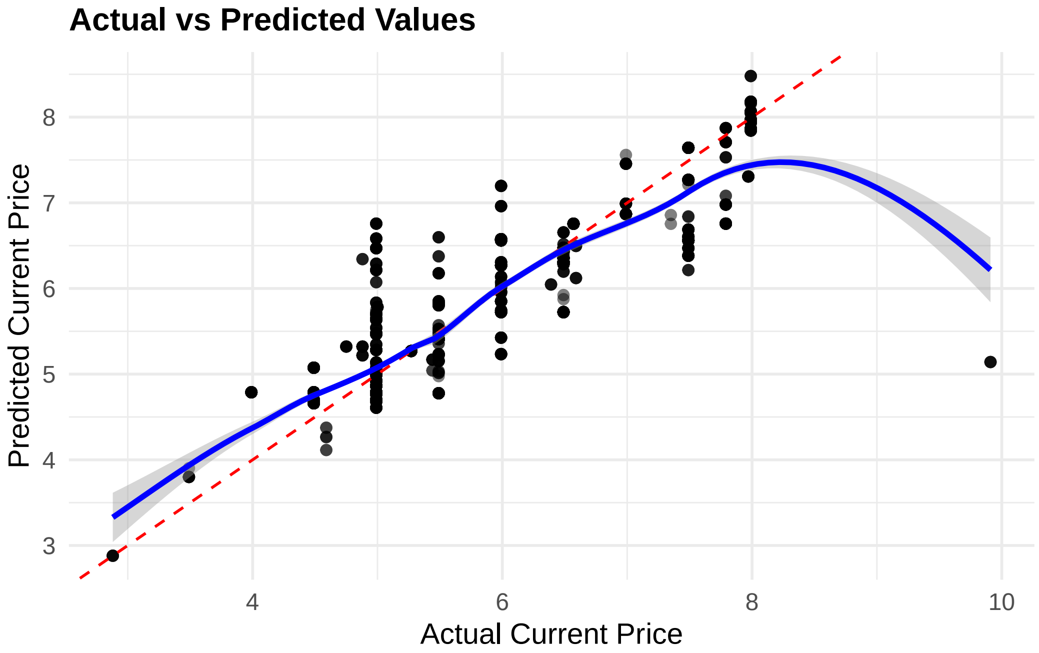

Figure 8: Scatter plot Figure 8 showing Actual vs Predicted values. Points should align along the red dashed line (y = x) if predictions are perfect. The blue line shows the smoothed trend of the relationship, and the gray band represents the 95% confidence interval.

The scatter plot in Figure 8 compares the actual and predicted current prices. Ideally, the points should align along the red dashed line, representing perfect predictions (y = x). The majority of points cluster near this line, indicating that the model captures the overall relationship between actual and predicted prices well. However, deviations from the line suggest instances where the model's predictions either underestimate or overestimate the actual prices.

The blue smoothed line, with its 95% confidence interval shaded in gray, highlights the trend in the relationship. While the model generally performs well in the mid-range of actual prices, a slight deviation from linearity is observed at the higher price ranges, where the blue line dips below the red line. This suggests that the model may overpredict prices at higher ranges or that nonlinear effects not captured by the linear model are influencing the results.

Overall, the plot underscores the effectiveness of the model in predicting current prices while highlighting areas where improvements, such as incorporating nonlinear relationships or additional predictors, could enhance performance. The residual patterns in higher price ranges suggest further investigation may be needed to account for these discrepancies.

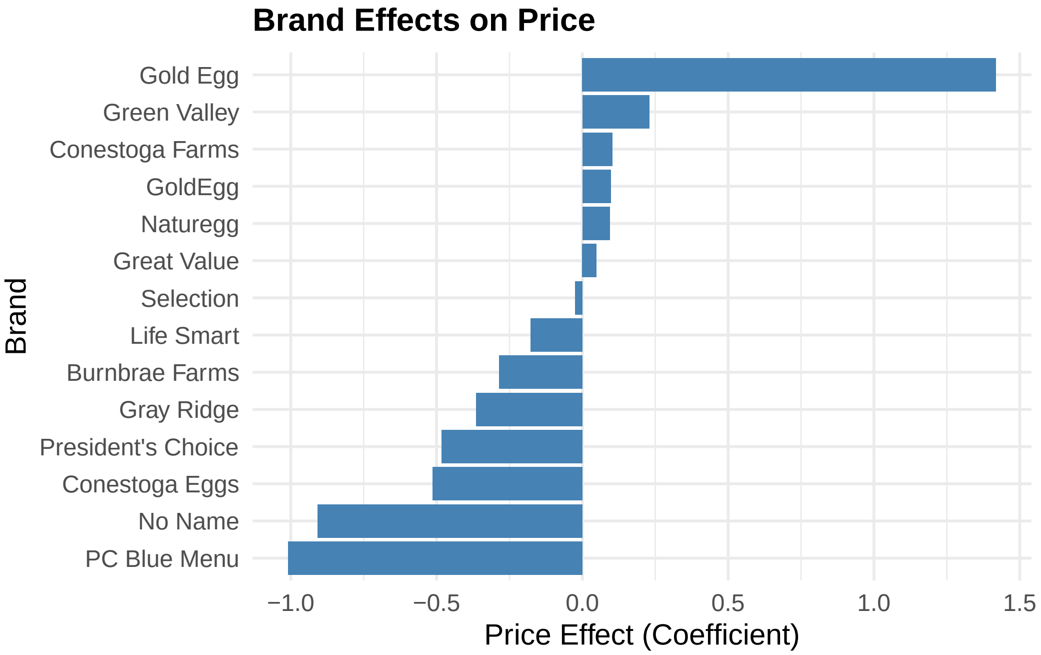

Figure 9: Brand effects on price showing estimated coefficients from the linear model. Positive values indicate brands associated with higher prices, while negative values indicate lower prices relative to the baseline brand.

The bar chart in (Figure 9) visualizes the influence of different brands on current prices based on the estimated coefficients from the linear regression model. Positive coefficients indicate that a brand is associated with higher prices compared to the baseline, while negative coefficients indicate lower prices.

Gold Egg stands out as the most premium brand, with a substantial positive effect on price (coefficient ~ 1.5). This suggests that products under this brand command significantly higher prices, likely due to strong brand reputation, higher quality perception, or market positioning. Green Valley and Conestoga Farms also show slight positive effects, indicating their positioning as mid-to-premium range brands.

In contrast, brands like No Name and PC Blue Menu exhibit the largest negative effects, with coefficients around -0.9 and -1.0, respectively. These brands likely target price-sensitive consumers and adopt a low-cost pricing strategy. Other brands, such as Gray Ridge and Burnbrae Farms, show moderate negative effects, suggesting their pricing falls below the baseline but remains less aggressive than purely budget brands.

Overall, the chart demonstrates a clear segmentation of pricing strategies among brands. Premium brands leverage higher perceived value to justify elevated prices, while budget brands cater to a cost-conscious demographic. These insights highlight the importance of branding in shaping pricing strategies and market positioning.

```
`geom_smooth()` using formula = 'y ~ x'
```

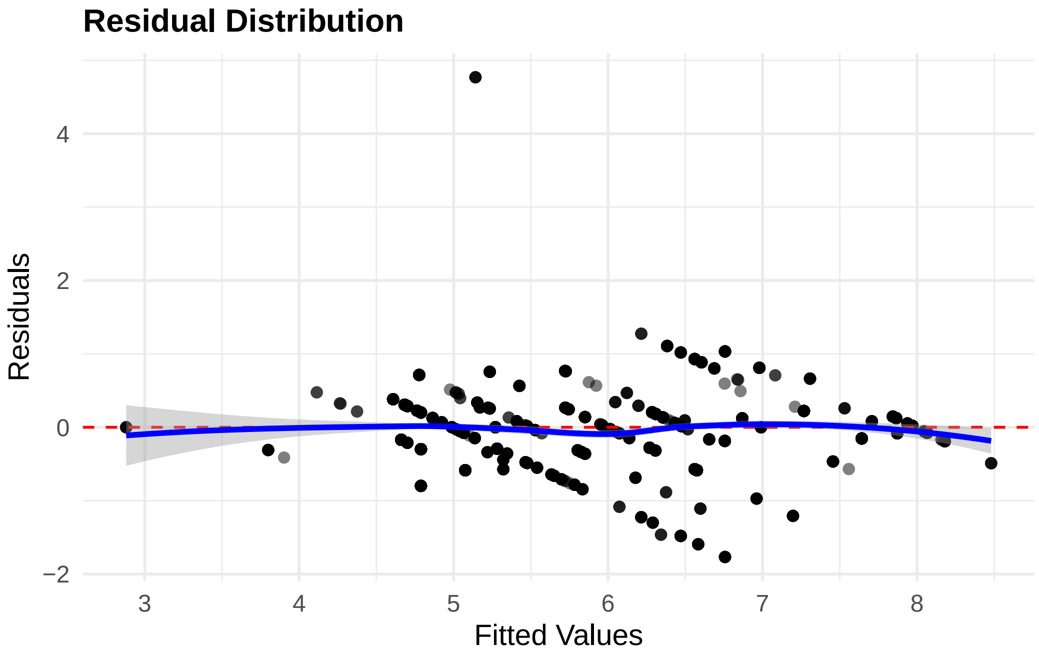

Figure 10: Residual plot showing model diagnostics. Points should be randomly scattered around the zero line (red dashed) with no clear pattern. The blue line shows the smoothed trend of residuals, and the gray band represents the 95% confidence interval.

The residual plot in (Figure 10) assesses the model's performance by examining the residuals (differences between observed and predicted values) against the fitted values. Ideally, the points should be randomly scattered around the horizontal zero line (red dashed), indicating that the model captures the relationship without systematic bias.

The plot shows no apparent strong patterns or trends in the residuals, suggesting that the assumption of linearity is reasonably upheld. However, there is slight variability at the extremes of fitted values, where some points deviate more substantially. This could indicate heteroscedasticity or potential model limitations in accurately capturing variability at lower and higher fitted values.

The blue smoothed line provides a trend of the residuals, staying close to the zero line across most of the range, with minimal curvature. This indicates that while the model performs well overall, slight deviations at the boundaries could suggest areas for improvement. The gray confidence band confirms that most residuals lie within expected bounds, reinforcing the reliability of the model for most data points.

In conclusion, the residuals' random distribution supports the validity of the model assumptions, with minor deviations at the edges warranting further examination to enhance predictive accuracy in extreme cases.

5 Discussion

5.1 Summary of Research

This paper investigates the dynamics of egg pricing across major Canadian grocery retailers, employing a linear regression model to analyze the influence of brand, vendor, and temporal factors on pricing. By focusing on the interactions between these variables, the study provides insights into the pricing strategies adopted by different brands and their impact on market prices. This approach allows us to quantify the effects of marketing and brand positioning within a competitive retail environment.

5.2 Insights into Consumer Behavior

One significant insight from this research is the impact of brand perception on consumer purchasing decisions. Premium brands like "Gold Egg" command higher prices, which reflects a consumer willingness to pay more for perceived quality. This finding underscores the importance of brand management in retail strategies and offers evidence of how brand strength can influence pricing power in the market.

5.3 Insights into Market Dynamics

Another key insight is the role of vendor strategies and seasonal variations in shaping pricing trends. The study highlights how vendors adjust prices in response to market demands and seasonal peaks, such as holiday periods when egg consumption typically increases. Understanding these patterns helps retailers and policymakers predict consumer behavior and adjust supply chains accordingly, optimizing both profitability and consumer satisfaction.

5.4 Limitations of the Study

While the research provides valuable insights, it is not without limitations. The reliance on linear regression may oversimplify the complex interactions within the market, potentially missing non-linear relationships or interactive effects between brands and market conditions. Additionally, the model assumes that all significant predictors are included, which may not account for unobserved variables such as local market disruptions or specific promotional activities.

5.5 Future Research Directions

Future studies should consider more complex models that can handle non-linear interactions and higher-dimensional data to better capture the intricacies of market behavior. Exploring models like mixed-effects models or advanced machine learning techniques could provide deeper insights into the predictive factors of pricing. Furthermore, expanding the dataset to include more granular regional and demographic data could refine the accuracy of the findings and enhance the generalizability of the results.

5.6 Conclusion

This paper advances our understanding of the factors influencing egg prices in Canadian retail markets, revealing significant effects of brand and vendor strategies on pricing. It highlights the necessity for robust brand management and strategic pricing to navigate the competitive retail landscape. The findings from this study not only inform retail management strategies but also offer a foundation for future research aimed at exploring deeper market mechanisms and refining economic models in the retail sector.

Appendix

.1 Egg Price Prediction Methodology

Target Population and Sample Design The target population for this study includes egg prices from Canada's four major retailers—Loblaws, Metro, Walmart, and NoFrills—covering three key categories: organic eggs, free-range eggs, and conventional eggs. These categories represent the primary purchasing behaviors of consumers and capture price variations based on production methods (Agriculture and Canada 2023). To ensure comprehensive data coverage, both offline and online data sources are utilized. Offline, the sampling frame includes stores located in urban, suburban, and rural areas to account for regional pricing differences. Online, the sampling frame consists of product pages from the retailers' websites, capturing promotional activities and regional pricing adjustments displayed on digital platforms (Brown 2020).

For each retailer, a systematic sampling method will be used, with stratification by geographic region (e.g., Western, Central, and Eastern Canada) and store size (e.g., small, medium, large). Approximately 50 stores will be selected for each retailer, resulting in a total sample size of around 200 stores (Canada 2023). Additionally, considering that prices fluctuate over time, data collection will occur weekly over a three-month period. This frequency will allow for capturing short-term price variations caused by seasonal demand, supply chain disruptions, and promotional activities (Kumar 2022).

Data Collection Methods Staff members will conduct weekly visits to designated stores to collect price data for various egg categories, including organic, free-range, and conventional eggs. Data will be recorded at different times of the day to capture price discrepancies caused by demand or promotional timing. Collected information includes current prices, promotional prices, and unit prices. To ensure consistency, all data collectors will use a standardized template for recording observations (Handbook 2021).

Additionally, web scraping tools will be employed to extract pricing information from the retailers' websites, including egg prices, product descriptions, specifications, and current promotional details (Brown 2020). These online data points need to be updated daily, especially during holidays or other periods with frequent changes in promotional activities. Finally, online or phone surveys may be conducted with store managers to understand current promotion schedules, regional pricing strategies, and inventory levels, compensating for dynamic information that observational data cannot directly capture (Tukey 1977).

Data Validation To ensure the quality and reliability of the data, the following validation steps will be implemented:

Cross-Validation of Data Sources By comparing offline and online data for consistency, discrepancies will be investigated to select the most reliable and accurate data. Outlier Detection and Handling Statistical methods, such as z-scores, will be applied to identify outliers in price data. These outliers will be manually reviewed through field checks or ignored in cases where they are likely due to temporary factors like holiday discounts (Tukey 1977).

Weighting Adjustments

Data will be weighted based on store size, regional consumer demographics, and purchasing power. Weighting standards will be based on relevant data from Statistics Canada to ensure that the sample accurately reflects national consumer behavior (Canada 2023). Seasonal Adjustments Historical price trends will be analyzed to account for seasonal variations. For example, during holidays like Easter or Thanksgiving, increased demand may lead to temporary price spikes. Advantages and Limitations Combining offline observations, online scraping, and survey data allows this methodology to capture a multi-dimensional view of egg prices, including regional, categorical, and promotional timing differences (Brown 2020). Weekly data collection provides detailed insights into short-term trends, while stratified sampling ensures representation across Canada's diverse regions. Automated web scraping reduces the time and labor required for data collection. However, regional disparities in tax policies, transportation costs, and consumer behavior may create challenges for ensuring balanced data. Collecting offline data requires significant human resources, making it difficult to cover all stores, particularly those in remote areas. Additionally, online data may lack accuracy and may not fully reflect actual inventory and prices in physical stores. Promotional information may also vary by region, introducing further variability (Brown 2020).

Conclusion This methodology integrates observational data, surveys, and online scraping techniques to establish a robust and detailed framework for predicting egg prices. By employing systematic sampling, rigorous data validation, and weighted adjustments, it provides accurate and reliable insights into price trends (Agriculture and Canada 2023). Despite challenges such as regional disparities, this methodology ensures actionable and detailed findings, with the potential for application in forecasting prices for other products.

suvery link https://forms.gle/eQ7YW8sKcjm4qGyMA

Copy of survey: Egg Consumption Behavior Survey

Hello! We are conducting a survey about egg consumption behavior. Your responses will provide valuable data for our research. The survey will take approximately 5–8 minutes to complete. All information will be kept condential and used solely for research purposes.

If you have any questions or would like to follow up on the survey, please contact us: Email: eggsurvey@researchteam.com Phone: 6476256706

1.What is your age? A.Under 18 B.18–29 years C.30–39 years D.40–49 years E.50 years and above

2.What is your gender? A. Male B.Female C..Other:

3.Where do you live? A.Urban area B.Suburban area C.Rural area

4.What is your approximate household monthly expenditure? A.Less than $2000 B.$2000-$4000 C.$4000-$6000 D.More than $6000

5.On average, how many eggs do you or your household purchase per week? A.Less than 6 B.6–12 C.13–24 D.More than 24

6.Which types of eggs do you prefer to purchase? (Select all that apply) Check all that apply. A.Organic eggs B.Free-range eggs C.Regular eggs D.Other

7.Where do you usually purchase eggs? (Select all that apply)

A.Supermarkets (e.g., Loblaws, Walmart) B.Farmers' markets C. Local grocery stores D.Online platforms

8.Rank the following factors in terms of importance when choosing an egg brand or type (1 = most important, 5 = least important) :Price 1 2 3 4 5

9. Rank the following factors in terms of importance when choosing an egg brand or type (1 = most important, 5 = least important) :Origin 1 2 3 4 5

10. Rank the following factors in terms of importance when choosing an egg brand or type (1 = most important, 5 = least important) :Organic/free-range certification 1 2 3 4 5

11. Rank the following factors in terms of importance when choosing an egg brand or type (1 = most important, 5 = least important) :Taste and quality 1 2 3 4 5

12. Where and at what price did you last purchase eggs? description:

13. If egg prices increased significantly (e.g., by more than 20%), how would your purchasing behavior change?

description:

14.Do you feel that current egg prices are reasonable? 1 2 3 4 5

15. Do you have any expectations regarding the quality or production standards of eggs?

1 2 3 4 5

Thank you for taking the time to complete this survey. Your insights are invaluable to our research. If you have any questions, need further assistance, or would like to share additional feedback, please feel free to contact us at: Email: eggsurvey@researchteam.com Phone: 6476256706 We deeply appreciate your participation and support!

References

Agriculture, and Agri-Food Canada. 2023. *Annual Agricultural Market Trends Report*. Ottawa: Canadian Agricultural Outlook.

Auguie, Baptiste. 2017. *gridExtra: Miscellaneous Functions for "Grid" Graphics*. https://CRAN.R-project.org/package=gridExtra.

Brown, T. 2020. *Data Scraping Techniques for Retail Analytics*. New York: TechPress.

Canada, Statistics. 2023. "Consumer Behavior and Expenditure Trends." https://www.statcan.gc.ca.

Filipp, Jacob. 2024. "Groceries Dataset." https://jacobfilipp.com/hammer/.

Handbook, Field Methods. 2021. *Research Methods Publishing*.

Kumar, A. et al. 2022. "Short-Term Price Fluctuations and Supply Chain Disruptions."

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Jonathan Keane, and Apache Arrow Community. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Spinu, Vitalie, Garrett Grolemund, and Hadley Wickham. 2024. *Lubridate: Make Dealing with Dates a Little Easier*. https://CRAN.R-project.org/package=lubridate.

Tukey, J. W. 1977. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

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

———. 2024. *Testthat: Unit Testing for r*. https://CRAN.R-project.org/package=testthat.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2024. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.