Competition Analysis in Canadian Grocery Prices

Prepared for:
Greg Baker
Tayaba Abbasi
Aryan Mikaeili
Shahrzad Mirzaei

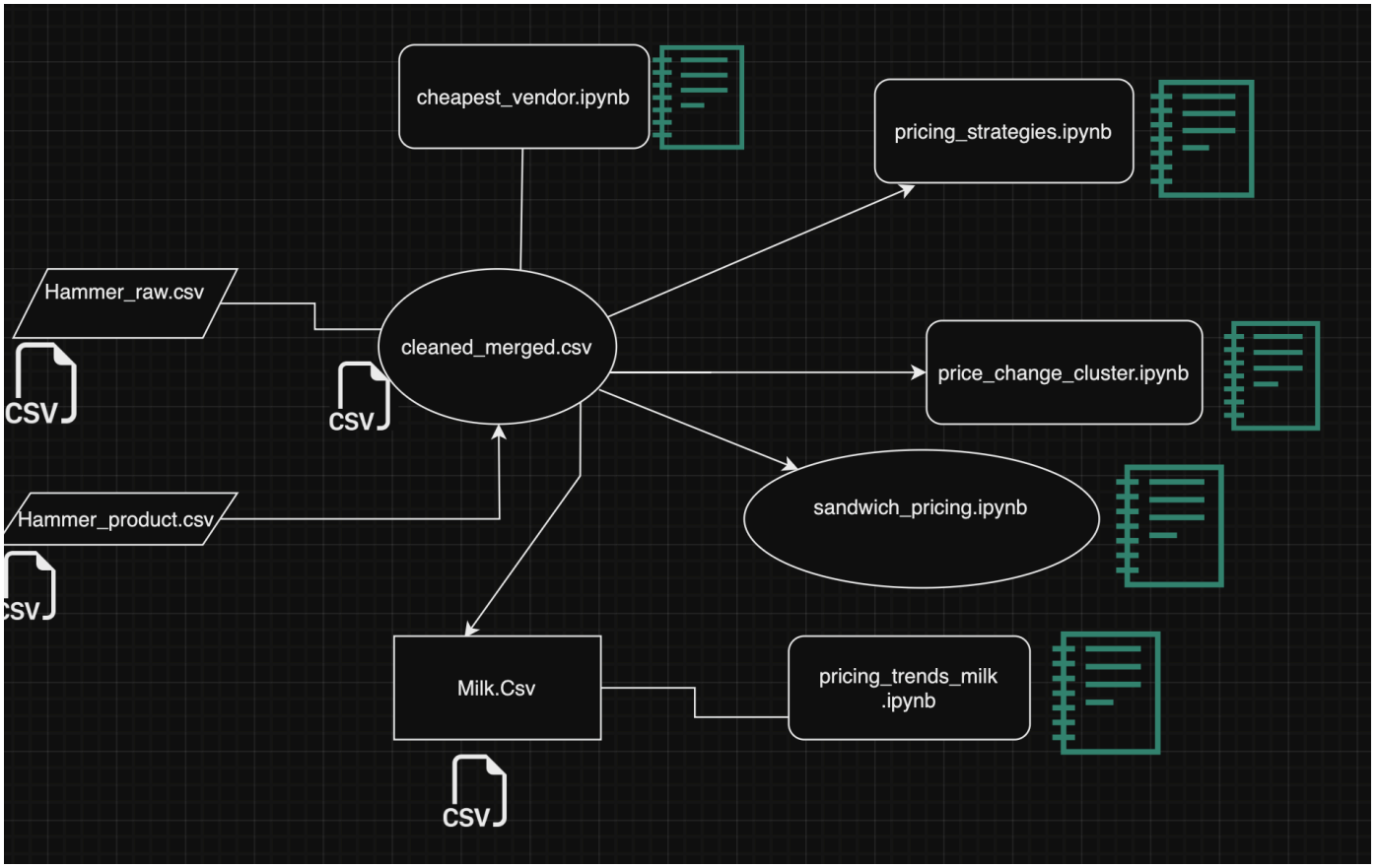
Prepared by:
William Desa, wdd1
Chitransh Motwani, cma115
Le hue anh Nguyen, lhn7

CMPT 353
December 05, 2024

Introduction

Our project seeks to address the common question, “Which grocery store offers the lowest prices?” by analyzing this issue across various scenarios. We examine the price relationships of eight different vendors to identify the most cost-effective option under specific circumstances. For example, we explore questions such as “Which vendor is the best choice for purchasing milk products?” and “Which vendor provides the most affordable options for making a sandwich?” Additionally, we tackle broader inquiries, including “Which vendor has the lowest average prices overall?” and “What are the pricing trends for each vendor over time?” By considering these diverse perspectives, our analysis provides a comprehensive and informed approach to selecting the most economical vendor.

Data Pipeline



Data Process

The **data process** involved cleaning, validating, and preparing datasets from Jacob Filipp's Hammer Project for analysis. The two files used—raw grocery price data and product metadata—underwent a series of preprocessing steps to ensure consistency and readiness for subsequent exploration.

1. Loading Data

- Datasets were loaded from CSV files using **Pandas**.

2. Cleaning the Raw Data File

- **Column Renaming:** Adjusted column names for clarity and usability, including nowtime, price, and product_id.
- **Timestamp Conversion:** The nowtime column was converted to datetime format to enable time-based analysis.
- **Price Data Cleanup:**
    - Converted price values to numeric, with non-numeric entries coerced to NaN.
    - Filled missing old_price values with zeros.
    - Extracted numeric values from the price_per_unit column, discarding any non-numeric characters.

3. Cleaning the Product Metadata File

- Renamed columns for consistency with the raw data.
- Converted product IDs to numeric types to facilitate merging with the raw data.

4. Merging Datasets

- Unified the raw data with product metadata by merging on product_id and id.
- Filtered out rows with missing product_name to retain only valid entries.
- Removed duplicate columns introduced during the merging process.
- Performed deduplication to eliminate repeated entries.

5. Validation of Data

- Conducted validation checks:
    - Ensured all price values were non-negative.
    - Addressed missing or inconsistent entries in critical columns such as product_name.

6. Output

- The processed dataset was exported as a new CSV file, ready for further analysis and modeling.

Known Issues Resolved

- **Duplicate Price Entries:** Identified and consolidated duplicate records for products with multiple entries on the same day.
- **Handling Missing Data:** Addressed missing values through filtering and imputation, ensuring no critical columns (e.g., product_name) remained empty.
- **Data Type Inconsistencies:** Enforced consistent data types for numerical and categorical values to avoid processing errors.

Summary of the Data Process

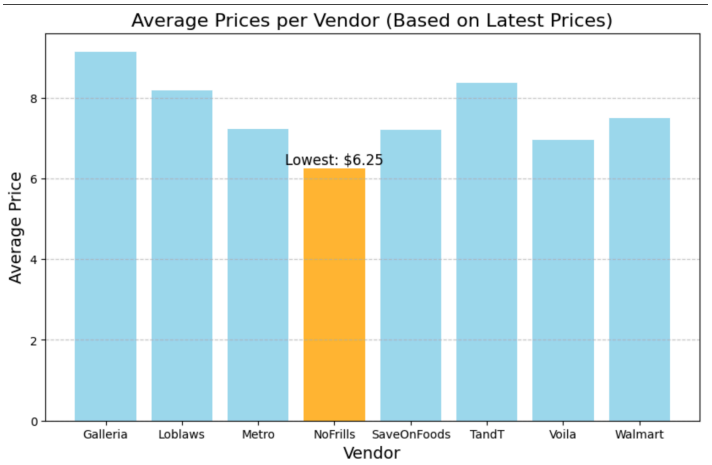
The preprocessing workflow combined two datasets from Jacob Filipp's Hammer Project, ensuring they were clean, accurate, and integrated. By resolving issues like duplicate entries, missing data, and inconsistencies, the unified dataset became suitable for advanced analysis and further exploration.

Techniques Used

In our analysis of grocery vendor pricing strategies, we employed several techniques to derive insights and inform our recommendations:

1. **Data Preprocessing Framework:** We established a data preprocessing pipeline that ensured the dataset's cleanliness and accuracy. This included loading raw data, renaming columns for clarity, and converting timestamps to datetime formats. My systematic approach resulted in a clean dataset ready for analysis.
2. **Data Cleaning and Transformation:** We executed comprehensive data cleaning, converting prices to numeric types and addressing non-numeric values by coercing them to NaN. This meticulous attention to detail ensured our analysis was based on accurate data.
3. **Merging and Deduplication:** We successfully merged the cleaned raw data with product metadata, creating a comprehensive dataset. This process involved filtering out rows with missing product names and deduplicating entries, ensuring our analysis was based on a consistent data foundation.
4. **Time Series Analysis:** We visualized average milk prices over time for each vendor to identify trends and fluctuations. This helped us observe how pricing strategies evolved over different periods.
5. **Granger Causality Tests:** These tests were used to assess interdependencies between vendors, allowing us to determine whether pricing changes in one vendor could predict changes in another, thereby revealing leader-follower relationships in pricing dynamics.
6. **Net Price Change Calculations:** We quantified overall price changes and their percentages for each vendor, capturing strategic pricing behaviors and providing a clearer picture of how vendors adjusted their prices over time.
7. **Correlation Analysis:** We explored the relationships between vendor prices, which helped us understand how pricing decisions among vendors might influence each other.
8. **Visualization:** Time series plots and correlation matrices were created to enhance data interpretation. These visual aids illustrated key findings and made complex relationships easier to understand.
9. **Feature Engineering:** A feature vector consisting of average price change and percentage price increase was assembled to represent each vendor numerically. This step was critical for subsequent modeling and analysis.
10. **Feature Scaling:** We standardized the features using the StandardScaler to ensure all variables had equal importance in the analysis, preventing bias due to differing scales.
11. **KMeans Clustering:** Vendors were grouped into three clusters based on their pricing behaviors. Each cluster represented a distinct pattern of price changes and frequencies of price increases, facilitating targeted insights into competitive strategies.
12. **Predictive Modeling:** We developed a Random Forest Regressor model to predict discounts based on vendor information. This involved selecting relevant features and optimizing model parameters, allowing us to derive insights into discount prediction.

Lowest Price Vendor: NoFrills has the lowest average price at $6.25, as highlighted in orange. This makes it the most cost-effective option for consumers seeking affordability.



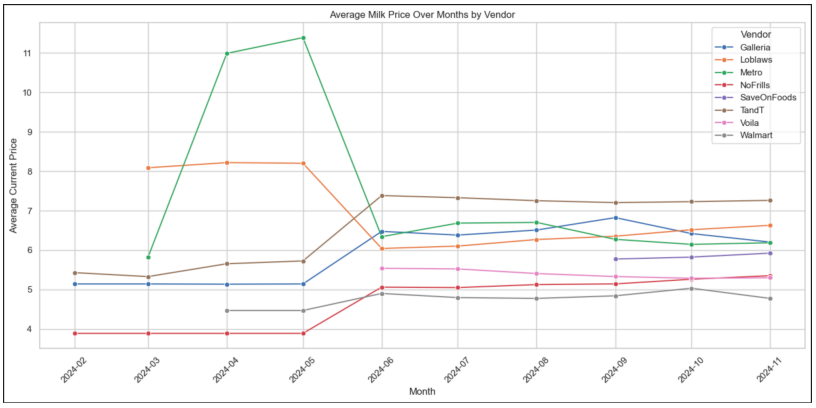
Price Distribution: Other vendors, such as Galleria, Loblaws, and T&T, have significantly higher average prices, indicating they cater to a different market segment or adopt different pricing strategies. SaveOnFoods, Walmart, and Voila have intermediate pricing, making them moderately competitive.

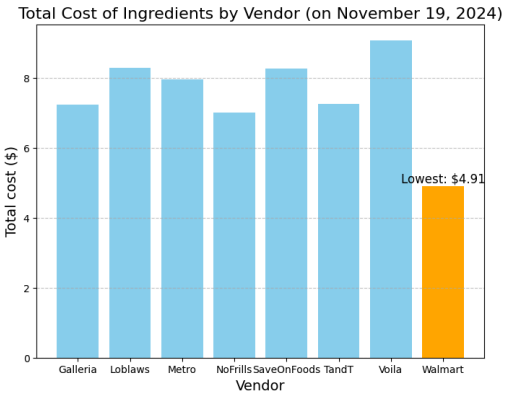
Market Positioning: NoFrills stands out as the price leader, which may reflect a strategic focus on low-cost offerings. Vendors like Galleria and Loblaws, with higher average prices, might emphasize quality, branding, or premium products, explaining their higher cost.

Consumer Implications: Consumers focused on affordability are likely to prioritize NoFrills. For those prioritizing other factors, such as product variety or perceived quality, the higher-priced vendors may still hold appeal.

Note: While No Frills is on average the cheapest vendor it could still be proven that choosing No Frills could be a more expensive place to shop at depending on the product you are buying

The average price over months by Vendor using milk as an example:





Vendors exhibited diverse pricing behaviors: while some, like Walmart and SaveOnFoods, maintained stable prices, others, such as Metro, experienced significant mid-year drops. NoFrills showed a steady increase, suggesting a growth-oriented pricing strategy.

Another example is the lowest price customers need to pay to buy ingredients to make sandwich:

Walmart is shown to offer customers the most appealing prices to buy white bread, ham and lettuce, which are the most basic ingredients to make a sandwich. Other vendors' offers are pretty similar, fluctuating from $7 to $8.5.

Granger causality analysis revealed structured interdependencies:

- **Loblaws** influenced **Metro** and **T&T prices**, acting as a leader in specific scenarios.
- **Metro** demonstrated influence on Loblaws, NoFrills, and T&T, highlighting its central role in price dynamics.
- **T&T** exhibited bidirectional relationships, influencing and being influenced by both Metro and Loblaws.

These relationships suggest a competitive and strategic pricing landscape, where some vendors take the lead while others follow market cues.

For Consumers:

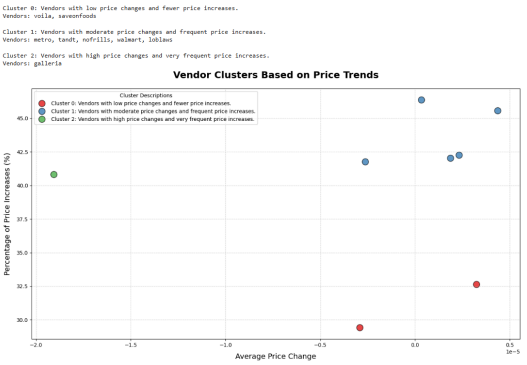
- **Stability Seekers:** Walmart and SaveOnFoods are ideal choices for those seeking predictable pricing trends.
- **Cost-Conscious Buyers:** Loblaws and Voila offer competitive pricing with net decreases, making them attractive for budget-oriented shopping.
- **Aggressive Pricing:** While NoFrills and T&T demonstrate consistent price increases, these may reflect higher-quality offerings or aggressive market positioning.

For Businesses:

- **Leverage Causality Insights:** Vendors like Metro and Loblaws should monitor their pricing influence on competitors to optimize market strategies.
- **Price Leadership Opportunities:** T&T and NoFrills can capitalize on their price-setting roles to reinforce market dominance
- **Adapt to Competition:** Vendors with declining or stable prices, such as Loblaws and SaveOnFoods, can use competitive pricing as a strategy to attract a larger consumer base.

Vendors Clustered Based On Price Trends.

This clustering utilized Kmeans clustering to group each vendor based on three categories: those with low price changes and fewer increases in price, Vendors with moderate price changes and frequent price increases, Vendors with high price changes and very frequent price increases, this was done to understand the price trends that each vendor seemed to follow.



This helps us come to the conclusion that ideally the Vendor's with low price changes and fewer price increases may be the vendors that a consumer would want to purchase their products at. This clustering takes into account all products that each vendor shares in common.

Cluster-Level Analysis:

- Stable Pricing (Cluster 0): Vendors like voila and SaveOnFoods demonstrate stability, making them ideal for consumers seeking predictable pricing.
- Active Adjustments (Cluster 1): Vendors in this cluster exhibit dynamic pricing strategies, possibly reflecting competitive pressures or market responsiveness.
- Volatile Pricing (Cluster 2): Galleria stands out for its frequent and significant price changes, which could be indicative of aggressive or reactionary pricing strategies.

Consumer Implications:

- Consumers valuing stability may prefer vendors in Cluster 0.
- Those seeking competitive pricing options should monitor vendors in Cluster 1 for potential discounts or promotional opportunities.

Business Implications:

- Vendors in Cluster 1 might monitor competitors' behavior to refine their pricing strategies.
- Understanding Galleria's behavior in Cluster 2 could help competitors predict volatile market changes.

To be able to evaluate how pricing strategies vary among different vendors. We developed a model using **Random Forest Regressor** trained to predict discounts based on vendor information. The R-squared value indicates that the model explains only about 7% of the variance in discount data, suggesting that other factors not included in the model may significantly influence discount levels.

Top Vendors with Average Discounts

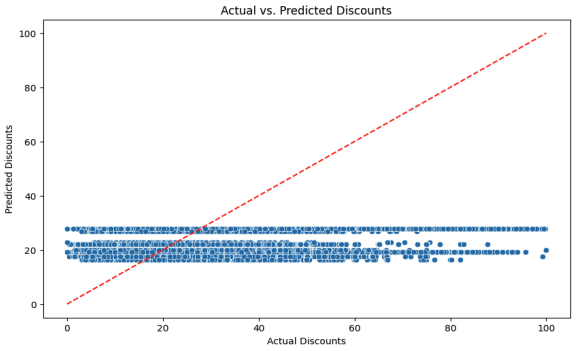
The analysis revealed the top three vendors with the highest average discounts:

This indicates that Walmart and T&T offer the most significant average discounts, which could attract price-sensitive customers but may also affect availability.

Vendor	Average Discount
Walmart	27.89
T&T	26.89
Galleria	22.83

Prediction Results

The model resulted in significant discrepancies between predicted and actual discounts, indicating areas where the model may struggle, particularly with larger discounts.



Insights and Recommendations

Pricing Strategy Analysis

1. **Prediction Accuracy**:
   Loblaws shows the best prediction accuracy with the lowest Mean Absolute Error, indicating a stable pricing strategy that relies on competitive pricing rather than extreme discounts.
2. **Segmented Strategies**:
   Vendors exhibit distinct pricing approaches. A new vendor should consider starting with competitive prices akin to Loblaws and Save-On-Foods while offering promotional discounts to gauge customer interest.
3. **Competitive Pricing Landscape**:
   Walmart and T&T provide substantial average discounts, suggesting aggressive pricing. A balanced strategy aligning with these competitors, coupled with unique value propositions, would attract customers seeking quality and affordability.

Consumer Insights

1. **Cost Efficiency**:
   No Frills is the most affordable option for cost-conscious shoppers, but consumers should compare specific product prices.
2. **Stability Seekers**:
   Walmart and Save-On-Foods cater to those seeking predictable pricing, fostering trust and loyalty.
3. **Opportunities for Competitive Buyers**:
   Loblaws and Voila may offer budget-friendly options, despite recent pricing declines.

Strategic Business Implications

1. **Market Leadership**:
   Metro and Loblaws influence overall pricing strategies in the market. Monitoring their pricing is crucial for competitiveness.
2. **Clustering Insights**:
   Vendors in Cluster 0, such as Voila and Save-On-Foods, demonstrate pricing stability, while those in Cluster 2, like Galleria, show volatility, reflecting aggressive pricing tactics.

Limitations

1. **Data Gaps:** There were instances of missing or insufficient data in the raw dataset, particularly for certain vendors. This limitation affected the reliability of our analyses, particularly in the Granger causality tests, where incomplete data may have hindered our ability to accurately assess interdependencies and pricing dynamics across all vendors.
2. **Non-Uniform Data Formats:** The initial dataset contained various formatting issues, such as inconsistent date formats and non-standardized price representations. Although the preprocessing pipeline addressed many of these issues, some discrepancies may have persisted, impacting the accuracy of the analysis.
3. **Handling of Non-Numeric Values:** Coercing non-numeric values to NaN during data cleaning may have resulted in the loss of valuable information. In retrospect, a more nuanced approach to handling these values could have preserved additional insights.
4. **Dependency on Initial Data Quality:** The quality of our analysis heavily depended on the initial dataset. If the raw data contained inaccuracies or inconsistencies, these issues would carry through to the final analysis, potentially leading to erroneous conclusions.
5. **Category Specificity:** Our primary analysis focused on milk pricing, which may not fully reflect broader vendor pricing behaviors across different product categories. This specificity might overlook important trends and strategies that are evident in other grocery items, limiting the generalizability of our findings.
6. **Time Constraints:** Due to time limitations, we were unable to conduct a deeper exploration of additional products and vendors. Expanding our analysis to include a wider range of items could provide a more comprehensive understanding of pricing strategies and vendor competitiveness in the market.
7. **Focus on Averages:** By concentrating on average prices, our analysis may obscure variations within product categories. For example, a vendor might offer significantly lower prices on specific product lines but appear more expensive overall when averages are considered.
8. **Limited Data Capture:**
   The dataset may not capture all regional or seasonal variations that could influence pricing strategies. Ignoring these factors could lead to incomplete insights.
9. **Interpretation of Clusters:**
   The clusters formed in our analysis are based solely on pricing trends, which means we could not consider other critical factors, such as product quality, customer satisfaction, or brand loyalty.
10. **Modeling Assumptions:**
    The assumptions underlying our statistical models may not fully hold true in all cases. For instance, the assumption of linear relationships in some analyses may not reflect the complexity of real-world pricing dynamics.

Future Considerations

If we had more time and resources, the following steps could be taken to enhance our analysis:

- **Expand the Dataset:** Include more vendors and product categories to capture a broader spectrum of pricing behaviors and dynamics.
- **Incorporate Qualitative Data:** Gather insights from vendors or market experts to complement quantitative data and provide a more nuanced understanding of pricing strategies.
- **Explore Seasonal and Regional Variations:** Analyze how prices fluctuate based on seasonal trends or regional differences to understand the factors influencing vendor pricing more comprehensively.
- **Utilize More Features for Clustering:** Incorporate additional features beyond price to better capture vendor behavior and competitive strategies, such as customer demographics or promotional activities.

Project Experience:

William Desa (wdd1): Grocery Pricing Trends Analysis

In this group project, I contributed significantly to the data analysis and machine learning components. I led data cleaning and preprocessing to ensure consistent and reliable datasets, performed trend analysis using correlation and Granger causality tests to uncover vendor pricing patterns, and implemented K-Means Clustering to group vendors based on pricing dynamics. I also created clear, insightful visualizations with Seaborn and Matplotlib and summarized actionable consumer insights to guide informed purchasing decisions. My technical expertise and collaborative efforts helped deliver impactful and practical results.

Key Accomplishments:

1. **Data Cleaning & Preprocessing**
    - Processed and cleaned raw data from multiple CSV files, including `Hammer.raw` and `Hammer.Product`.
    - Ensured proper formatting of critical columns such as `current_price` and `nowtime` for accurate analysis.
    - Removed invalid and missing data, creating a consistent and high-quality dataset for downstream tasks.
2. **Trend Analysis**
    - Conducted statistical tests, including correlation analysis and Granger causality tests, to analyze vendor reactions to price changes for the product "Milk."
    - Explored vendor interactions to identify whether they led, followed, or reacted to pricing changes.
    - Created time-series visualizations to track average price trends over months, offering insights into vendor pricing behavior.
3. **Vendor Comparison**
    - Aggregated pricing data to calculate and compare average prices for each vendor, identifying those with the lowest costs.
    - Used Seaborn and Matplotlib to develop bar charts, histograms, and time-series plots, making vendor performance comparisons clear and intuitive.
4. **Cluster Analysis**
    - Applied K-Means Clustering to segment vendors into three distinct clusters based on their pricing behaviors:
        - **Cluster 0:** Vendors with low price changes and fewer price increases (e.g., Voila, Save-On-Foods).
        - **Cluster 1:** Vendors with moderate price changes and frequent price increases (e.g., Metro, Walmart, Loblaws).
        - **Cluster 2:** Vendors with high price changes and frequent price increases (e.g., Galleria).
    - Visualized these clusters with scatter plots, differentiating groups by color for easy interpretation.
5. **Consumer Insights**
    - Summarized findings to identify vendors that consistently offer the lowest prices and those with stable versus fluctuating pricing patterns.
    - Delivered actionable insights to help consumers make informed purchasing decisions based on vendor-specific pricing trends and dynamics.

Results:

Successfully identified vendor-specific pricing trends, highlighted the cheapest and most stable vendors, and demonstrated expertise in applying data engineering, statistical analysis, and machine learning to solve real-world problems effectively.

Le hue anh Nguyen (lhn7):

In this project, I used Python and data visualization tools to analyze grocery prices. I decided to analyze the price to buy ingredients (white bread, ham, and lettuce) to make sandwiches since I want to see if the vendor that offers the lowest price overall will be the one that has the lowest total price to buy simplest ingredients to make a sandwich.

Individual contributions:

- The first step is to clean the data. We need to merge 2 csv files to create a whole complete file. Then, I drop all the NaN values or prices that are $0. I also make sure that all the prices and dates are correctly formatted.
- I choose bread, ham and lettuce as they provide humans with necessary nutritions to survive.
- Then I sorted the data, retrieving only products that are bread, ham or lettuce. Calculate the sum of the cheapest bread, ham and lettuce to see the minimum price that people need to pay to buy a sandwich's ingredients.
- After getting the final data, I plot a simple bar chart.
- The result is pretty surprising. Walmart provides people with the most appealing price to make a sandwich, however, from our analysis above, it shows that NoFrills offers the best price on average for customers.
- Besides, I also try to figure out if vendors hike the price before the sale so that the sale price is just the normal one. However, it’s quite hard to figure it out. I did find this happens for some products, they raised the price 1.5 times and then lowered the price back to normal. However, the percentage of this happening is not large enough to conclude that they hike the price before the sale start to deceive customers.

Chitransh Motwani (cma115)

As part of the team project, I conducted an in-depth analysis of pricing strategies among various grocery vendors to identify pricing strategies. This part of the project aimed to evaluate how different pricing strategies impact competitiveness and market penetration. My contributions included developing a predictive model, analyzing feature importance, and providing actionable insights based on the results, with a particular focus on the data preprocessing phase that ensured the integrity and quality of the project.

Individual Contributions

- **Data Preprocessing Framework:** I established a data preprocessing pipeline that was crucial in ensuring the dataset's cleanliness and accuracy. This included loading raw grocery price data and product metadata from CSV files, renaming columns for clarity, and converting timestamps to datetime formats. My systematic approach resulted in a clean and unified dataset ready for the project.
- **Data Cleaning and Transformation:** I executed comprehensive data cleaning, converting prices to numeric types and handling non-numeric values by coercing them to NaN. This meticulous attention to detail not only ensured that our analysis was based on accurate data but also prepared the dataset for effective model training.
- **Error Handling and Validation:** I integrated an error logging mechanism to capture issues during data loading and cleaning, enabling timely diagnosis and resolution of data integrity problems. I conducted validation checks to ensure non-negative prices and addressed inconsistencies, significantly enhancing the dataset's reliability for analysis.
- **Merging and Deduplication:** I successfully merged the cleaned raw data with product metadata, creating a comprehensive dataset. This process involved filtering out rows with missing product names and deduplicating entries, ensuring that the analysis was based on a consistent and accurate data foundation.
- **Exploratory Data Analysis (EDA):** I conducted exploratory data analysis to uncover vendor-specific pricing patterns and identify key metrics. This involved visualizing price distributions, calculating average current and old prices, and analyzing discount percentages across different vendors. I compiled a vendor summary to assess dominance in various price segments. This analysis showed that some vendors, like NoFrills and Save-On-Foods, excelled in low-price segments, while others, such as TandT and Walmart, effectively combined discount strategies with competitive average pricing. This comprehensive assessment provided a foundation for understanding market dynamics.
- **Model Development:** I worked on creating a Random Forest Regressor model to predict discounts based on vendor information. This involved selecting relevant features, optimizing model parameters, and evaluating performance metrics.
- **Feature Importance Analysis:** I examined the importance of various vendors in predicting discounts, with Walmart emerging as the most influential vendor with an importance score of 48%. This analysis underscored the significant impact of Walmart's pricing strategies on market dynamics.
- **Residual Analysis:** I conducted a detailed comparison of predicted versus actual discounts, identifying significant discrepancies and areas for model improvement.
- **Vendor-Specific Performance Insights:** I calculated the Mean Absolute Error (MAE) for each vendor, providing insights into prediction accuracy.

This project not only enhanced my analytical and data modeling skills but also provided valuable experience in collaborative research and strategic decision-making. The insights gained from this analysis can significantly impact how a new vendor positions itself in the market, contributing to its long-term success.