Grocery Price Prediction and Comparative Analysis in NYC

Introduction to the Problem and Data

Problem Statement

New York City (NYC) is known for its complex and diverse grocery retail landscape. Grocery prices vary significantly based on factors such as location, store type, and product category. For consumers, this variability leads to uncertainty about where to shop to save money. For retailers, it presents an opportunity to align their pricing strategies with consumer demand and competitive trends.

This project addresses two key objectives:

1. Phase 1: Analyze price variations across multiple grocery stores in NYC to identify trends, disparities, and price drivers.
2. Phase 2: Merge the datasets to perform a focused price comparison between Whole Foods and other retail stores in NYC. This includes identifying overpriced, fairly priced, and underpriced items.

Significance of the Project

- For Consumers: Helps identify where they can shop for affordable groceries.
- For Retailers: Highlights opportunities for competitive pricing.
- For Policymakers: Identifies neighborhoods with disproportionately high food costs, which can influence policies to address food insecurity.

Context: NYC Grocery Landscape

NYC has one of the highest costs of living in the United States, with grocery costs contributing significantly to household expenses. Whole Foods, known for its premium offerings and organic products, often charges higher prices than competitors. However, such price gaps can disproportionately impact lower-income neighborhoods, where affordability is a critical concern.

Data Overview

This project utilized two datasets scraped manually from grocery store websites:

1. NYC Grocery Prices Dataset:
    - Size: 250 rows
    - Features: Store, Neighborhood, Item, Price
    - Purpose: Analyze price trends across stores and neighborhoods.
2. Whole Foods Dataset:
    - Size: 1657 rows
    - Features: Product, Regular, Sale, Prime, Category, and Discounts
    - Purpose: Compare Whole Foods prices with other NYC grocery stores.

The datasets were merged based on item descriptions to enable direct price comparisons in the second base.

Exploratory Data Analysis (EDA)

Phase 1: Price Analysis Across NYC Grocery Stores

The first phase focused on understanding how grocery prices vary across stores, neighborhoods, and items.

1. Descriptive Statistics

Statistic	Price
Count	250
Mean	$3.12
Min	$0.31
Max	$10.35

Key Insight: Most grocery items fall within an affordable price range, but higher-priced items like specialty meats or organic products push the maximum price upward.

2. Price Distribution

- A histogram revealed that 75% of items cost between $1 and $6.
- Interpretation: Affordable products dominate the market, catering to budget-conscious consumers. Outliers reflect premium or specialty items.

3. Average Prices by Neighborhood

Neighborhood	Average Price ($)
Manhattan	4.26
Brooklyn	3.55
Queens	2.94
Bronx	2.59
Staten Island	2.24

Insight:

- Manhattan had the highest average prices, aligning with its affluent demographic.
- Staten Island reported the lowest prices, reflecting cost-conscious consumer behavior.

4. Prices by Store

- Key Food showed the highest average prices ($3.26), followed closely by Whole Foods and Trader Joe's.
- C-Town had the lowest prices, likely targeting budget-conscious consumers.

Interpretation:
Stores' pricing strategies vary based on their target audience, product quality, and brand positioning.

Phase 2: Focused Whole Foods Price Comparison

In this phase, the two datasets were merged to directly compare Whole Foods prices with other NYC stores.

Initial Analysis

- Whole Foods consistently had higher prices across most items.
- A new feature, Price Difference, was calculated: Price Difference=Whole Foods Price−Other Store Price\text{Price Difference} = \text{Whole Foods Price} - \text{Other Store Price}Price Difference=Whole Foods Price−Other Store Price

Price Discrepancies by Category

Product Category	Average Difference ($)
Dairy	1.50
Pantry Essentials	1.30
Beverages	1.10

- Observation: Whole Foods prices were 40–50% higher in these categories.
- Context: Dairy and pantry staples are essential purchases, making price gaps particularly impactful for cost-conscious consumers.

Modeling

Goal

The primary goal of this modeling task was to predict grocery prices based on features like Store, Neighborhood, and Item. Given that price is a continuous numerical variable, regression models were appropriate.

Models Tested

We tested four models to predict grocery prices:

1. Linear Regression
2. Decision Tree Regressor
3. K-Nearest Neighbors (KNN)
4. Random Forest Regressor

Evaluation Metrics

To evaluate the performance of the regression models, the following metrics were used:

1. Mean Squared Error (MSE)
    - Definition: MSE measures the average squared difference between the predicted and actual values.
    - Why MSE?:
        - Since the task involves predicting grocery prices (a continuous value), MSE is appropriate because it penalizes large prediction errors more heavily than smaller ones.
        - This ensures that the model focuses on reducing large discrepancies, which are critical when predicting prices that directly impact consumer decisions.
2. R^2 Score (Coefficient of Determination)
    - Definition: R^2 indicates the proportion of variance in the target variable (price) explained by the model. It ranges from 0 to 1, where 1 means perfect prediction.
    - Why R^2?:
        - R^2 provides an intuitive understanding of how well the model fits the data.
        - A high R^2 value means the model successfully captures the relationships between the features (Neighborhood, Item, Store) and the target variable (price).

Why are these metrics appropriate for the task?

- Given that grocery prices are continuous and not binary or categorical, metrics like accuracy, precision, or recall (used in classification tasks) are not suitable.
- Instead, MSE ensures the model minimizes prediction errors, while R^2 offers insight into the model's explanatory power. These metrics align perfectly with the goal of predicting prices as accurately as possible.

Pipeline and Implementation

1. Data Preprocessing:
    - OneHotEncoding was applied to categorical variables (Store, Neighborhood, Item).
    - Data was split into 80% Training and 20% Testing sets to ensure robust evaluation.

2. Model Results:

Model	MSE	R² Score
Linear Regression	0.6578	0.8465
Random Forest	0.6731	0.8428
K-Nearest Neighbors	1.3296	0.6896
Decision Tree	0.9894	0.7690

Best Performing Model: Linear Regression

- MSE: 0.6578
- R² Score: 0.8465

Why Linear Regression?

- Linear Regression performed the best in terms of both MSE and R², suggesting that grocery prices can be effectively modeled using simple linear relationships between features.
- While Random Forest performed comparably, its higher complexity makes it less interpretable.

Feature Importance (Random Forest)

To further understand the key drivers of grocery prices, feature importance was extracted from the

Random Forest model:

R	Feature	Importance (%)
1	Item_Cheese (Pound)	26.7%
2	Neighborhood_Manhattan	14.8%
3	Item_Bananas (Pound)	14.7%
4	Item_Chicken (Pound)	14.5%
5	Item_Milk (Gallon)	8.3%

Key Insight:

- Item Type and Neighborhood have the strongest impact on price predictions, which aligns with expectations since prices depend heavily on the type of grocery item and its location in NYC.

Feature Importance (Random Forest)

The top predictors were:

1. Item Type (e.g., Cheese, Milk)
2. Neighborhood (Manhattan had the strongest influence).
3. Store (Whole Foods and Key Food).

Interpretation of Results

Phase 1: Key Findings

- Prices vary significantly across neighborhoods and stores.
- Item Type and Neighborhood are key drivers of grocery prices.

Phase 2: Whole Foods Comparison

- Whole Foods prices were 40–50% higher across essential categories.
- 60% of items were categorized as Overpriced.

Limitations

1. Small Data Sample: Only 250 rows were available for NYC grocery prices.
2. Manual Data Collection: Risk of errors and missing data.
3. No Temporal Data: Static pricing data does not account for seasonal promotions.
4. Simplified Item Matching: String matching may cause minor inaccuracies.
5. Unseen Factors: Product quality, demand, and brand loyalty were not included.

Next Steps and Recommendations

1. Expand Data Collection:
    - Include more stores, neighborhoods, and items.
2. Incorporate Promotions: Analyze time-series data to capture discounts and price trends.
3. Interactive Dashboard: Build a tool for consumers to compare grocery prices.
4. Feature Engineering: Add product quality and consumer demand metrics for better models.

Conclusion and Summary

This project provided a two-phase analysis of grocery prices in NYC and a focused comparison with Whole Foods. The findings reveal significant price disparities driven by neighborhood, store, and product type. While Whole Foods' premium positioning justifies its higher prices, opportunities exist to optimize pricing strategies and improve affordability for essential goods.

Summary:

- Recapped the findings from Phase 1 and Phase 2:
    - Price variability across NYC stores and neighborhoods.
    - Whole Foods' higher prices and price discrepancies.
- Highlighted the Linear Regression model's performance in predicting prices.