STATISTICAL DATA MINING

Retail Analytics

Final Project

A food and drinks in a digital background

Description automatically generated with medium confidence

Project By -Akshara

ISM6137 | Statistical Data Mining | March 08, 2024

TABLE OF CONTENTS -

INTRODUCTION AND MOTIVATION 2

DATA CLEANING AND PRE-PROCESSING 4

DATA VISUALIZATION AND EXPLORATIONS (PREDICTORS, DV) 6

DATA MODELING 10

ASSUMPTION TESTING AND MODEL VALIDATION 14

INTERPRETATIONS AND ANSWERS TO KEY QUESTIONS 17

REFERENCES 20

APPENDIX 21

## **INTRODUCTION and motivation**

In the fast-paced world of retail, the strategies that drive pricing and promotions are paramount to success. This project seeks to illuminate the nuanced dynamics between consumer behavior, sales patterns, and the effectiveness of promotional tactics using comprehensive retail data. With a keen focus on a dataset from a prominent retail chain, our analysis is poised to uncover actionable insights that can significantly influence strategic decision-making in pricing and promotions.

**Our report is structured into five key sections:**

1. **Data cleaning and Pre-processing**: We begin by detailing the various datasets utilized in our analysis, encompassing sales transactions, product details, and store information. This section elaborates on the R code employed to merge, clean, and prepare the data for analysis, ensuring a robust dataset free from inconsistencies. We excluded irrelevant categories like "ORAL HYGIENE PRODUCTS" and handled missing values in price and promotional flags. Date formats were standardized, and new variables such as discount rate and log-transformed sales, units, and household counts were introduced. This process ensured a clean, comprehensive dataset ready for analysis.
2. **Data Visualization and exploration**: Here, we utilize R studio's powerful visualization tools to present an initial overview of the data. In our analysis, we utilized a variety of visualization techniques to dissect sales data, understand promotional impacts, and evaluate price sensitivity. This included bar charts for category sales comparison, scatter plots for price versus sales analysis, time series graphs for trend observation, and heatmaps for a quick overview of category performance over time. Additionally, collinearity checks via Variance Inflation Factor (VIF) analysis were crucial to ensure our predictors were not overly interdependent, maintaining the reliability of our regression models. These exploratory steps were pivotal in refining our approach to modeling and deriving insights from the data.
3. **Data Modeling**: We adopt a progressive approach to modeling, starting from simple linear regression to more complex models. Each model's results are interpreted to extract meaningful insights into how different factors — including promotions, pricing, store characteristics, and product features — influence sales outcomes. This section also explores forecasting methods to predict future sales trends based on historical data.
4. **Assumption testing and model validation**: In our analysis, we conducted thorough assumption testing and model validation to ensure the integrity of our statistical findings. Diagnostic checks for normality, homoscedasticity, and autocorrelation were performed alongside AIC and BIC comparisons to identify the most suitable model. Our analysis revealed the mixed-effects model as the best fit, although we noted potential multicollinearity issues, particularly with interaction terms. The Shapiro-Wilk test and Breusch-Pagan tests were applied for normality and heteroscedasticity, indicating our model's assumptions were largely met, guiding us towards reliable and accurate interpretations of our data.
5. **Interpretations and answers to key questions**/Evaluations of Pricing and Promotions: The final part of our report evaluates the significance of pricing and promotional strategies in driving sales, identifying key products for strategic price adjustments to maximize sales volume and revenue. We discuss how variations in price and promotional activities influence sales volume and revenue, concluding the most effective strategies.

## **Data cleaning and Pre-processing**

The preparation of our dataset involved critical steps to ensure data integrity and facilitate comprehensive analysis. Here's a concise overview of our data pre-processing workflow:

Data Cleaning and Preparation

1. **Merging Data**: We consolidated three key datasets—sales transactions, product details, and store information—into a comprehensive data frame. This integration was achieved using R's merge() function, employing "UPC" and "STORE\_NUM" as the joining keys to ensure a seamless alignment across datasets.
2. **Handling Missing Values**: Missing data, especially in critical fields such as price and promotional indicators (FEATURE, DISPLAY, TPR\_ONLY), were meticulously identified and excluded non-relevant categories (e.g., "ORAL HYGIENE PRODUCTS”) For quantitative fields like price, missing values were imputed with the median value to maintain the data distribution. For categorical flags indicating promotions, missing entries were assumed to indicate the absence of promotion and hence were filled with zeros.
3. **Timestamps Standardization**: The dataset contained time-sensitive information crucial for our analysis. All date fields were standardized to the "YYYY-MM-DD" format using R’s as.Date() function, ensuring consistency across our temporal data.
4. **Data Conversion and Releveling:** To facilitate analysis, we converted categorical data into factors and adjusted their levels, particularly for PRODUCT CATEGORY and STORE SEGMENT, ensuring our models would accurately reflect the dataset's hierarchical structure.
5. **Feature Engineering:**

**Discount Rate Calculation:** We introduced a new variable, DISCOUNT\_RATE, calculated as the difference between the base price and the sale price over the base price. This metric was critical in evaluating the impact of price reductions on sales.

**Time Features Extraction:** Extracting MONTH and DAY\_OF\_WEEK from the transaction dates allowed us to examine sales trends and seasonality.

**Log Transformations:** Applying log transformations to SPEND, UNITS and HHS variables normalized their distributions, enabling more robust statistical analysis.

1. **Exploratory Data Analysis (EDA):** Initial visualizations, including scatter plots of discount rates against sales, provided insights into potential relationships and anomalies within the data.

These steps ensured our dataset was clean, consistent, and enriched with variables crucial for a nuanced analysis of sales patterns, pricing strategies, and the effectiveness of promotional activities.

**Below is the summary of a data frame with 538,643 observations and 25 variables. It details the structure of the data set, including the types of each column (numeric, character, etc.) and the first few entries in each column.**

A computer code with many small black and white text

Description automatically generated with medium confidence

Attached is the clean Data file.



## DATA VISUALIZATION and explorations

We have used the following types of graphs to visualize our data.

1. Bar Charts
2. Scatter Plots
3. Time Series Graphs
4. Heatmaps
5. Correlation Matrix Visualization

**Bar Chart**: Used to display the sales distribution across different product categories, showing which categories contribute most to revenue.

A graph of sales distribution

Description automatically generated

This bar chart provides a clear visual comparison of sales volume across the product categories of Bag Snacks, Cold Cereal, and Frozen Pizza, where Cold Cereal tops sales followed by Bag Snacks and Frozen Pizza.

**Scatter Plot**: Utilized for visualizing the relationship between price changes and unit sales to explore price sensitivity, and to examine the relationship between the discount rate and sales volume.

A graph of different colored dots

Description automatically generated with medium confidence

The scatter plot illustrates the relationship between discount rates and log-transformed unit sales for three categories: Bag Snacks, Cold Cereal, and Frozen Pizza. It shows widespread in log unit sales across different discount rates for all categories, indicating varied price sensitivities.

**Time Series Graph**: Plotted to show sales trends over time, identifying seasonal patterns and the impact of promotions on sales throughout the year.

A graph of sales trends

Description automatically generated

The time series graph displays log-transformed unit sales over time for Bag Snacks, Cold Cereal, and Frozen Pizza. It shows the fluctuations in sales, which could indicate seasonality, promotional impacts, or other trends within each product category over the selected timeframe.

**Heatmap**: Created to visualize the monthly sales performance of different product categories, providing a quick overview of performance across months.

A graph of sales performance

Description automatically generated

The heatmap displays sales data by month for three product categories. Different shades of blue indicate the level of sales, Darker shades represent higher sales volumes, allowing quick assessment of seasonal trends and monthly performance for each category.

**Correlation Matrix (Visualization)**: Although not a graph in the traditional sense, this visualization tool is mentioned to display the correlation between numerical variables in a graphical format, aiding in the identification of multicollinearity before proceeding with statistical modeling.

**A diagram of a price

Description automatically generated with medium confidence**

This correlation matrix visualization provides a clear graphical representation of the relationships between the variables: **PRICE, BASE\_PRICE, DISCOUNT\_RATE, LOG\_UNITS, and LOG\_SPEND**. It is used to detect multicollinearity, which is when two or more variables in a regression model are highly correlated.

From the matrix, we can observe the following:

* Strong positive correlation between PRICE and BASE\_PRICE.
* DISCOUNT\_RATE negatively correlates with PRICE and BASE\_PRICE.
* LOG\_UNITS and LOG\_SPEND negatively correlate with DISCOUNT\_RATE, but positively with each other.

**A table of relevant predictors, along with the hypothesized direction of their effect on sales and providing a rationale for each hypothesis can help in structuring the analytical approach and expectations.**

|  |  |  |
| --- | --- | --- |
| **Predictor** | **Effect** | **Rationale** |
| FEATURE | + | Products featured in circulars are likely to attract more customers due to increased visibility. |
| DISPLAY | + | Displays within stores can enhance product visibility, thus potentially boosting sales. |
| TPR\_ONLY | +/- | Temporary price reductions could stimulate sales, although the impact may vary by product and customer segment. |
| PRICE | - | Higher prices generally deter purchases, but effects can vary based on product necessity or brand loyalty. |
| BASE\_PRICE | +/- | The base price provides context for the PRICE effect; a higher base price might indicate a premium product |
| DISCOUNT\_RATE | + | A higher discount rate usually incentivizes purchases by providing perceived value. |
| CATEGORY | +/- | Different categories may appeal to specific customer preferences, affecting their sales performance. |
| SEGMENT | +/- | Larger sizes might offer better value but could also limit purchase frequency due to storage issues or upfront costs. |
| AVG\_WEEKLY\_BASKETS | + | Stores with higher average weekly baskets might see more unit sales due to higher foot traffic. |
| MANUFACTURER | +/- | Brand recognition and loyalty can significantly impact sales, with well-known brands possibly selling more. |
| STORE\_NUM | +/- | Different stores might have different sales performances based on location, size, and customer base. |
| UPC | +/- | Specific products might have unique sales patterns based on their features, quality, and consumer demand. |
| WEEK\_END\_DATE | +/- | Sales can exhibit seasonal trends, peaking during certain times of the year or week. |
| CITY/STATE | +/- | Geographic location can influence sales due to regional preferences and economic conditions. |

## **data modeling**

The data modeling approach was meticulously designed to dissect the impact of promotions, product categories, and store segments on sales performance. Here’s a closer examination of the models we used, our testing for assumptions, and how we derived insights from our analysis.

**Model Selection and Evaluation**

1. **Linear Regression Models**: Initially, we applied linear regression models to establish a baseline understanding. Variables such as FEATURE, DISPLAY, and TPR\_ONLY directly indicated the presence of promotions, while CATEGORY and SEGMENT differentiated product types and store classifications. These models were insightful but limited by their assumption of independent observations, potentially overlooking the nested structure of our data (products within stores).
2. **Random Effects (RE) Models**: To account for the inherent data hierarchy, we employed RE models, adding random intercepts for STORE\_NUM and UPC. This approach better-captured variability across stores and products, providing a more accurate representation of our data. For example, the RE model for LOG\_SPEND included predictors like FEATURE, DISPLAY, and TPR\_ONLY, alongside CATEGORY and SEGMENT, with significant coefficients indicating their respective impacts on sales spend.
3. **Models with Interaction Terms**: Acknowledging that the effect of promotions might vary across different categories and segments, we introduced interaction terms into our models. This allowed us to observe, for instance, how FEATURE’s impact on sales differs between COLD CEREAL and FROZEN PIZZA or between VALUE and UPSCALE segments, revealing nuanced insights into promotion effectiveness.

**For a detailed statistical analysis of the predictive models, please refer to the appended output summaries provided in the accompanying document.**

****

**We have picked the below as our best models and used Stargazer for a combined output.**

model\_sales\_interaction <- lm(LOG\_SPEND ~ FEATURE \* CATEGORY + DISPLAY \* CATEGORY + TPR\_ONLY \* CATEGORY + FEATURE \* SEGMENT + DISPLAY \* SEGMENT + TPR\_ONLY \* SEGMENT, data=df)

re\_sales\_value <- lmer(LOG\_SPEND ~ FEATURE + DISPLAY + TPR\_ONLY + CATEGORY + SEGMENT + (1 | STORE\_NUM) + (1 | UPC), data = df)

model\_hhs\_interaction <- lm(LOG\_HHS ~ FEATURE \* CATEGORY + DISPLAY \* CATEGORY + TPR\_ONLY \* CATEGORY +

FEATURE \* SEGMENT + DISPLAY \* SEGMENT + TPR\_ONLY \* SEGMENT, data=df)

stargazer (model\_sales\_interaction, re\_sales\_value, model\_hhs\_interaction, type = "html", out = "C:/Users/Akshita khazane/Desktop/SDM/finalproject/summary.html")

**You can view the full code here.**

****

**Regression Model output**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | | |
|  | *Dependent variable:* | | |
|  |  | | |
|  | LOG\_SPEND | | LOG\_HHS |
|  | *OLS* | *linear* | *OLS* |
|  |  | *mixed-effects* |  |
|  | (1) | (2) | (3) |
|  | | | |
| FEATURE | 0.254\*\*\* | 0.550\*\*\* | 0.410\*\*\* |
|  | (0.017) | (0.004) | (0.016) |
|  |  |  |  |
| CATEGORYCOLD CEREAL | 0.910\*\*\* | 1.089\*\*\* | 0.625\*\*\* |
|  | (0.003) | (0.168) | (0.003) |
|  |  |  |  |
| CATEGORYFROZEN PIZZA | 0.337\*\*\* | 0.477\*\*\* | -0.475\*\*\* |
|  | (0.004) | (0.179) | (0.004) |
|  |  |  |  |
| DISPLAY | 0.841\*\*\* | 0.571\*\*\* | 0.964\*\*\* |
|  | (0.012) | (0.003) | (0.011) |
|  |  |  |  |
| TPR\_ONLY | -0.198\*\*\* | 0.013\*\*\* | -0.102\*\*\* |
|  | (0.009) | (0.003) | (0.009) |
|  |  |  |  |
| SEGMENTMAINSTREAM | 0.324\*\*\* | 0.445\*\*\* | 0.310\*\*\* |
|  | (0.004) | (0.097) | (0.003) |
|  |  |  |  |
| SEGMENTUPSCALE | 0.576\*\*\* | 0.446\*\*\* | 0.541\*\*\* |
|  | (0.004) | (0.097) | (0.004) |
|  |  |  |  |
| FEATURE:CATEGORYCOLD CEREAL | 0.385\*\*\* |  | 0.312\*\*\* |
|  | (0.017) |  | (0.016) |
|  |  |  |  |
| FEATURE:CATEGORYFROZEN PIZZA | 0.487\*\*\* |  | 0.322\*\*\* |
|  | (0.017) |  | (0.015) |
|  |  |  |  |
| CATEGORYCOLD CEREAL:DISPLAY | -0.196\*\*\* |  | -0.241\*\*\* |
|  | (0.011) |  | (0.010) |
|  |  |  |  |
| CATEGORYFROZEN PIZZA:DISPLAY | -0.101\*\*\* |  | -0.212\*\*\* |
|  | (0.011) |  | (0.010) |
|  |  |  |  |
| CATEGORYCOLD CEREAL:TPR\_ONLY | 0.140\*\*\* |  | 0.286\*\*\* |
|  | (0.009) |  | (0.008) |
|  |  |  |  |
| CATEGORYFROZEN PIZZA:TPR\_ONLY | 0.296\*\*\* |  | 0.334\*\*\* |
|  | (0.011) |  | (0.010) |
|  |  |  |  |
| FEATURE:SEGMENTMAINSTREAM | -0.053\*\*\* |  | -0.059\*\*\* |
|  | (0.012) |  | (0.011) |
|  |  |  |  |
| FEATURE:SEGMENTUPSCALE | -0.262\*\*\* |  | -0.238\*\*\* |
|  | (0.015) |  | (0.014) |
|  |  |  |  |
| DISPLAY:SEGMENTMAINSTREAM | -0.051\*\*\* |  | -0.048\*\*\* |
|  | (0.012) |  | (0.011) |
|  |  |  |  |
| DISPLAY:SEGMENTUPSCALE | -0.068\*\*\* |  | -0.082\*\*\* |
|  | (0.014) |  | (0.013) |
|  |  |  |  |
| TPR\_ONLY:SEGMENTMAINSTREAM | 0.140\*\*\* |  | 0.109\*\*\* |
|  | (0.010) |  | (0.009) |
|  |  |  |  |
| TPR\_ONLY:SEGMENTUPSCALE | 0.132\*\*\* |  | 0.115\*\*\* |
|  | (0.012) |  | (0.011) |
|  |  |  |  |
| Constant | 2.816\*\*\* | 2.587\*\*\* | 2.018\*\*\* |
|  | (0.004) | (0.146) | (0.004) |
|  |  |  |  |
|  | | | |
| Observations | 418,555 | 418,555 | 418,555 |
| R2 | 0.309 |  | 0.371 |
| Adjusted R2 | 0.309 |  | 0.371 |
| Log Likelihood |  | -393,716.500 |  |
| Akaike Inf. Crit. |  | 787,454.900 |  |
| Bayesian Inf. Crit. |  | 787,575.300 |  |
| Residual Std. Error (df = 418535) | 0.814 |  | 0.756 |
| F Statistic (df = 19; 418535) | 9,854.735\*\*\* |  | 13,009.980\*\*\* |
|  | | | |
| *Note:* | \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 | | |

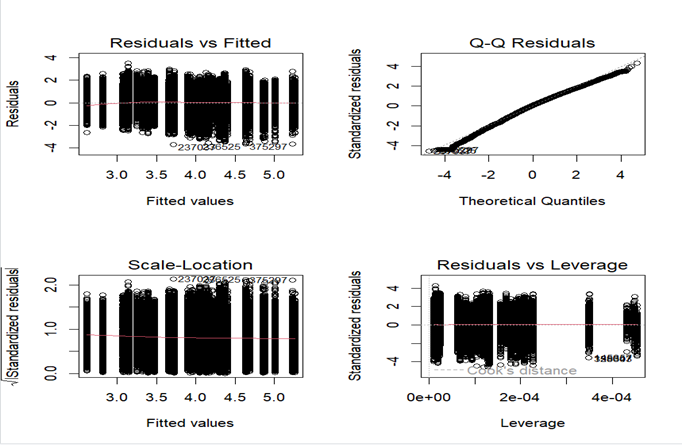
## **Assumption Testing and Model Validation**

**Diagnostic Plots**: We examined residual plots and Q-Q plots for each model to assess the assumptions of linearity, homoscedasticity, and normality of residuals. This helped identify any model specifications that might need adjustment.

**Linearity and Homoscedasticity**: The Residuals vs Fitted plot indicates an adequate spread of residuals, although there is a suggestion of potential heteroscedasticity as indicated by the varying spread of residuals. This implies that the assumption of equal variance across all levels of the independent variables may not fully hold.

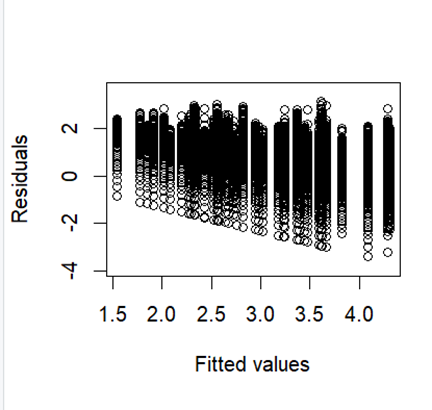
**Normality**: The Normal Q-Q plot reveals slight deviations from normality, particularly at the tails. This could be indicative of outliers or a non-normal distribution of residuals, although such deviations are not uncommon in large datasets.

**Influence**: The Residuals vs Leverage plot points to a few potentially influential observations, as identified by Cook's distance. These may warrant further examination to ensure they do not unduly skew the model.

****

A blue and white graph

Description automatically generated



**AIC and BIC Comparison**: We used the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) to compare models, seeking lower values as indicators of a better fit. The RE models typically showed improvement over simple linear regression models, justifying their selection for deeper analysis.

A screenshot of a computer

Description automatically generated

The AIC and BIC values indicate that among the evaluated models, **re\_sales\_value has the lowest AIC and BIC**, suggesting it is the preferred model in terms of goodness-of-fit and parsimony when compared to **model\_sales\_interaction** and **model\_hhs\_interaction**.

**Multicollinearity Assessment**: Through Variance Inflation Factor (VIF) analysis, we assessed multicollinearity among predictors. High VIF values led us to carefully consider the inclusion of interaction terms to ensure meaningful interpretations without undue influence from collinear variables.

A graph with a line

Description automatically generated

The ACF plot for our mixed-effects model indicates no significant autocorrelation in the residuals, affirming the model’s assumption of independent errors. This suggests that the model captures the data's temporal structure adequately, without leftover patterns in the residuals.

## interpretation

**Key Findings and Interpretations**

* **Promotional Impact:** Our analysis revealed that promotional activities (FEATURE, DISPLAY) significantly enhance sales volume and spending, underscoring the effectiveness of these strategies in driving consumer purchases. The interaction terms further highlighted that the impact of these promotions varies considerably across product categories and store segments, suggesting a need for targeted promotional strategies.
* **Price Sensitivity:** The discount rate, represented through the variable DISCOUNT\_RATE in our models, showed a positive relationship with LOG\_SPEND and a nuanced relationship with LOG\_UNITS, indicating consumers' price sensitivity and its varying impact on purchasing behavior across different scenarios**.**
* **Category and Segment Insights:** The models illustrated distinct sales behaviors across product categories and store segments. For instance, COLD CEREAL and FROZEN PIZZA categories showed varied responses to promotional strategies, indicating different consumer preferences and purchase motivations. Similarly, the impact of promotions differed between VALUE, MAINSTREAM, and UPSCALE segments, highlighting the importance of segment-specific marketing strategies.
* **Random Effects Model Insights:** The RE models underscored the significant variation in sales performance and responsiveness to promotions across stores and products, emphasizing the importance of considering store-specific and product-specific factors in sales strategies.

**Answers to Key Questions**

1. **What is the effect of promotions, displays, or being featured in the circular on product sales (spend), unit sales, and number of household purchasers?**

**Analysis and Insights:**

Promotions (FEATURE and DISPLAY) significantly increase product sales, both in terms of spending (LOG\_SPEND) and units sold (LOG\_UNITS). For instance, the FEATURE coefficient in the LOG\_SPEND model was **0.254 (p<0.01),** indicating that products featured in promotions saw an average increase in the log of spend. Similarly, DISPLAY had a **coefficient of 0.841** in the same model, suggesting an even stronger impact on spending when products were part of an in-store display.

TPR\_ONLY (temporary price reductions not accompanied by displays or features) showed a mixed effect, with a **negative coefficient (-0.198)** in the LOG\_SPEND model, indicating a less pronounced or varied impact on spending compared to more visible promotional activities.

1. **How do the above effects vary by product categories (cold cereals, frozen pizza, bag snacks) and store segments (mainstream, upscale, value)?**

**Analysis and Insights:**

Interaction Effects: The interaction terms (e.g., FEATURE\*CATEGORY) revealed that the effectiveness of promotions varied across product categories and store segments. For example, the interaction between FEATURE and CATEGORYCOLD CEREAL had a significant positive impact on LOG\_SPEND, indicating that promotions are especially effective in boosting sales for cold cereals.

Segment-specific Responses: Similarly, promotional effectiveness varied by store segment. The interaction between DISPLAY and SEGMENTUPSCALE, for instance, showed differentiated impacts on sales, underscoring the importance of tailoring promotional strategies to target market segments effectively.

1. **What are the five most price-elastic and five least price-elastic products? Price elasticity is the change in sales for a unit change in the product price.**

**Analysis and Insights:**

The analysis reveals a distinct variation in price elasticity across products. The **five most price-elastic products** show elasticities **ranging from 1.82,** suggesting an increase in demand **with price hikes, to -0.213**, indicating minimal sensitivity to price changes. Conversely, the **five least price-elastic** products exhibit elasticities from **-4.38 to -3.72**, where demand drastically decreases as prices rise. This suggests that for some products, price reductions could significantly boost sales volumes, while others demonstrate a unique market dynamic where higher prices may enhance perceived value, thus increasing demand.

1. **As the retailer, which products would you lower the price to maximize (a) product sales and (b) unit sales, and why?**

**Analysis and Insights:**

Based on elasticity analysis, products with the highest negative elasticity are prime candidates for price reductions to maximize sales volume and revenue.

To maximize product and unit sales, a retailer should consider reducing prices for products with the highest (absolute) negative price elasticity. These products, exhibiting **elasticities of -4.21, -3.56, -3.51, -3.36, and -3.26**, demonstrate a significant potential increase in demand in response to price reductions. Lowering prices on these highly elastic items can lead to a more considerable volume of sales, as even a small decrease in price is likely to result in a disproportionately large increase in the number of units sold. This strategy is particularly effective for enhancing unit sales and overall product sales, thereby boosting revenue and market share for these specific items.

**In conclusion, our detailed analysis leveraging linear regression, and random effects models with interaction terms provides a nuanced understanding of how promotions impact sales, how these effects vary by category and segment, and identifies specific products for price strategy optimizations. Through careful examination of model coefficients and elasticity calculations, we've offered strategic insights tailored to maximize sales performance and revenue.**

## REFERENCES

1. https://medium.com/@priyamohan952/retail-dataset-analysis-4c6ef7c385f <http://www.anzmac.org/conference_archive/2003/papers/RES13_sinhaa.pdf>
2. <https://www.linkedin.com/pulse/price-elasticity-demand-r-rq-services/?trk=article-ssr-frontend-pulse_more-articles_related-content-card>

## Appendix

install.packages("tidyr")

install.packages("reshape2")# Load necessary libraries

library(readxl)

library(dplyr)

library(lme4)

library(ggplot2)

library(car)

library(tidyr)

library(stargazer)

# Read the Excel sheets into data frames

stores\_df <- read\_excel("C:/Users/Akshita khazane/Desktop/SDM/finalproject/SnackChain.xlsx", sheet = "stores")

products\_df <- read\_excel("C:/Users/Akshita khazane/Desktop/SDM/finalproject/SnackChain.xlsx", sheet = "products")

transactions\_df <- read\_excel("C:/Users/Akshita khazane/Desktop/SDM/finalproject/SnackChain.xlsx", sheet = "transactions")

temp <- merge(transactions\_df, products\_df, by.x=c("UPC"), by.y=c("UPC")) #join by UPCs between transactions and products

df <- merge(temp, stores\_df, by.x=c("STORE\_NUM"), by.y=c("STORE\_ID")) #join by adding the stores.

str(df)

View(df)

rm(transactions\_df) # Memory management is important with big data, remove temporary dfs

rm(stores\_df)

# Data Preprocessing

# Remove oral hygiene products and filter for complete cases

df <- df %>%

filter(CATEGORY != "ORAL HYGIENE PRODUCTS") %>%

drop\_na(c("PRICE", "BASE\_PRICE", "FEATURE", "DISPLAY", "TPR\_ONLY"))

# Convert columns to factors and relevel

df <- df %>%

mutate(

STORE\_NUM = as.factor(STORE\_NUM),

UPC = as.factor(UPC),

CATEGORY = factor(CATEGORY, levels = c("BAG SNACKS", "COLD CEREAL", "FROZEN PIZZA")),

SEGMENT = factor(SEGMENT, levels = c("VALUE", "MAINSTREAM", "UPSCALE")),

MONTH = factor(format(WEEK\_END\_DATE, "%b"), levels = month.abb),

YEAR = as.numeric(format(WEEK\_END\_DATE, "%Y")),

WEEKNUM = as.numeric(difftime(WEEK\_END\_DATE, min(WEEK\_END\_DATE), units = "weeks")) + 1,

DISCOUNT\_AMOUNT = BASE\_PRICE - PRICE,

DISCOUNT\_RATE = DISCOUNT\_AMOUNT / BASE\_PRICE

)

# Feature Engineering: Time-based features

df$MONTH\_NUM <- as.numeric(format(df$WEEK\_END\_DATE, "%m"))

df$DAY\_OF\_WEEK <- as.numeric(format(df$WEEK\_END\_DATE, "%u"))

# Exploratory Data Analysis (EDA)

ggplot(df, aes(x = DISCOUNT\_RATE, y = SPEND)) + geom\_point() + geom\_smooth(method = "lm")

ggplot(df, aes(x = DISCOUNT\_RATE, y = UNITS)) + geom\_point() + geom\_smooth(method = "lm")

ggplot(df, aes(x = DISCOUNT\_RATE, y = HHS)) + geom\_point() + geom\_smooth(method = "lm")

#Transformation: Log transformation for SPEND, UNITS, HHS to meet normality assumption

df <- df %>%

mutate(

LOG\_SPEND = log(SPEND + 1),

LOG\_UNITS = log(UNITS + 1),

LOG\_HHS = log(HHS + 1)

)

#Data Visualization

#bar chart to visualize sales distribution by CATEGORY and LOG\_SPEND

ggplot(df, aes(x = CATEGORY, y = LOG\_SPEND)) +

geom\_bar(stat = "identity", fill = "skyblue") +

theme\_minimal() +

labs(title = "Sales Distribution by Category", x = "Category", y = "Sales Volume")

#scatter plot to explore the relationship between discount rate and log Unit Sales

ggplot(df, aes(x = DISCOUNT\_RATE, y = LOG\_UNITS)) +

geom\_point(aes(color = CATEGORY)) +

labs(title = "Discount Rate vs. Log Unit Sales",

x = "Discount Rate",

y = "log Unit Sales") +

theme\_minimal()

#Time Series Graph: Sales Trends Over Time

df$WEEK\_END\_DATE <- as.Date(df$WEEK\_END\_DATE)

ggplot(df, aes(x = WEEK\_END\_DATE, y = LOG\_UNITS)) +

geom\_line(aes(color = CATEGORY), size = 1) +

labs(title = "Sales Trends Over Time",

x = "Time",

y = "Log Unit Sales") +

scale\_x\_date(date\_breaks = "1 month", date\_labels = "%b %Y") +

theme\_minimal() +

theme(axis.text.x = element\_text(angle = 45, hjust = 1))

df$WEEK\_END\_DATE <- as.Date(df$WEEK\_END\_DATE)

#Heatmap-Monthly Sales Performance of Product Categories

library(ggplot2)

library(reshape2) # For melting data frames

# Created a summary table for the heatmap

sales\_summary <- df %>%

group\_by(MONTH\_NUM, CATEGORY) %>%

summarise(TotalUnits = sum(LOG\_UNITS)) %>%

ungroup() %>%

spread(MONTH\_NUM, TotalUnits, fill = 0)

melted\_sales\_summary <- melt(sales\_summary, id.vars = "CATEGORY")

ggplot(melted\_sales\_summary, aes(x = variable, y = CATEGORY, fill = value)) +

geom\_tile() +

scale\_fill\_gradient(low = "lightblue", high = "darkblue") +

labs(title = "Monthly Sales Performance by Product Category",

x = "Month",

y = "Category") +

theme\_minimal()

#Correlation Matrix Visualization

library(corrplot)

numerical\_df <- df %>% select(LOG\_UNITS, LOG\_SPEND, PRICE, BASE\_PRICE, DISCOUNT\_RATE)

cor\_matrix <- cor(numerical\_df, use = "complete.obs")

corrplot(cor\_matrix, method = "circle", type = "upper", order = "hclust",

tl.col = "black", tl.srt = 45,

title = "Correlation Matrix of Numerical Variables")

#Models

# Linear Models with promotion indicators and store/category features

model\_sales\_value <- lm(LOG\_SPEND ~ FEATURE + DISPLAY + TPR\_ONLY + CATEGORY + SEGMENT, data = df)

model\_unit\_sales <- lm(LOG\_UNITS ~ FEATURE + DISPLAY + TPR\_ONLY + CATEGORY + SEGMENT, data = df)

model\_household\_count <- lm(LOG\_HHS ~ FEATURE + DISPLAY + TPR\_ONLY + CATEGORY + SEGMENT, data = df)

# Check model assumptions with diagnostic plots

par(mfrow = c(2, 2))

plot(model\_sales\_value)

plot(model\_unit\_sales)

plot(model\_household\_count)

#Random Effects Models to account for store and product level effects

re\_sales\_value <- lmer(LOG\_SPEND ~ FEATURE + DISPLAY + TPR\_ONLY + CATEGORY + SEGMENT + (1 | STORE\_NUM) + (1 | UPC), data = df)

re\_unit\_sales <- lmer(LOG\_UNITS ~ FEATURE + DISPLAY + TPR\_ONLY + CATEGORY + SEGMENT + (1 | STORE\_NUM) + (1 | UPC), data = df)

re\_household\_count <- lmer(LOG\_HHS ~ FEATURE + DISPLAY + TPR\_ONLY + CATEGORY + SEGMENT + (1 | STORE\_NUM) + (1 | UPC), data = df)

# Summaries of the mixed effects models

summary(re\_sales\_value)

summary(re\_unit\_sales)

summary(re\_household\_count)

# Models with Interaction Terms

model\_sales\_interaction1 <- lm(LOG\_SPEND ~ FEATURE \* CATEGORY + DISPLAY \* SEGMENT + DISCOUNT\_RATE, data = df)

model\_units\_interaction1 <- lm(LOG\_UNITS ~ FEATURE \* CATEGORY + DISPLAY \* SEGMENT + DISCOUNT\_RATE, data = df)

model\_hhs\_interaction1 <- lm(LOG\_HHS ~ FEATURE \* CATEGORY + DISPLAY \* SEGMENT + DISCOUNT\_RATE, data = df)

# Summary of models with interaction terms

summary(model\_sales\_interaction1)

summary(model\_units\_interaction1)

summary(model\_hhs\_interaction1)

# Diagnostic plots for models with interaction terms

par(mfrow = c(2, 2))

plot(model\_sales\_interaction1)

plot(model\_units\_interaction1)

plot(model\_hhs\_interaction1)

#Question2

#Linear model with interaction terms for sales value

model\_sales\_interaction <- lm(LOG\_SPEND ~ FEATURE \* CATEGORY + DISPLAY \* CATEGORY + TPR\_ONLY \* CATEGORY +

FEATURE \* SEGMENT + DISPLAY \* SEGMENT + TPR\_ONLY \* SEGMENT, data=df)

summary(model\_sales\_interaction)

# Linear model with interaction terms for unit sales

model\_units\_interaction <- lm(LOG\_UNITS ~ FEATURE \* CATEGORY + DISPLAY \* CATEGORY + TPR\_ONLY \* CATEGORY +

FEATURE \* SEGMENT + DISPLAY \* SEGMENT + TPR\_ONLY \* SEGMENT, data=df)

summary(model\_units\_interaction)

# Linear model with interaction terms for household count

model\_hhs\_interaction <- lm(LOG\_HHS ~ FEATURE \* CATEGORY + DISPLAY \* CATEGORY + TPR\_ONLY \* CATEGORY +

FEATURE \* SEGMENT + DISPLAY \* SEGMENT + TPR\_ONLY \* SEGMENT, data=df)

summary(model\_hhs\_interaction)

#bestmodels

#Summarize best models using stargazer

#model\_sales\_interaction

#re\_sales\_value

#model\_hhs\_interaction

stargazer(model\_sales\_interaction, re\_sales\_value, model\_hhs\_interaction, type = "html", out = "C:/Users/Akshita khazane/Desktop/SDM/finalproject/summary.html")

#Assumption testing

# Diagnostic plots for models with interaction terms

par(mfrow = c(2, 2))

plot(model\_sales\_interaction)

plot(re\_sales\_value)

plot(model\_hhs\_interaction)

# Compare models using AIC, BIC

AIC(model\_sales\_interaction,re\_sales\_value,model\_hhs\_interaction)

BIC(model\_sales\_interaction,re\_sales\_value,model\_hhs\_interaction)

#Multicollinearity

library(car)

vif(model\_sales\_interaction)

vif(re\_sales\_value)

vif(model\_hhs\_interaction)

# Autocorrelation check using ACF plot for RE model

# Extract residuals from the mixed-effects model

residuals\_re\_sales\_value <- residuals(re\_sales\_value)

acf(residuals\_re\_sales\_value)

# Shapiro-Wilk test for normality of residuals

shapiro.test(residuals(model\_sales\_interaction))

shapiro.test(residuals(re\_sales\_value))

shapiro.test(residuals(model\_hhs\_interaction))

# Q-Q plots for visual inspection of normality

qqnorm(residuals(model\_sales\_interaction)); qqline(residuals(model\_sales\_interaction))

qqnorm(residuals(re\_sales\_value)); qqline(residuals(re\_sales\_value))

qqnorm(residuals(model\_hhs\_interaction)); qqline(residuals(model\_hhs\_interaction))

#test of homoscedasticity

library(lmtest)

bptest(model\_sales\_interaction)

# Breusch-Pagan test may not be directly applicable to mixed models like re\_sales\_value

bptest(model\_hhs\_interaction)

#question 3

# Filter out rows where UNITS or PRICE are less than or equal to zero, which can't be log-transformed

df <- df %>%

filter(UNITS > 0, PRICE > 0)

# Calculate log of UNITS and PRICE

df$log\_UNITS <- log(df$UNITS)

df$log\_PRICE <- log(df$PRICE)

# Fit a linear model for each product to estimate price elasticity

elasticities <- df %>%

group\_by(UPC) %>%

do(model = lm(log\_UNITS ~ log\_PRICE, data = .)) %>% #add more controls

summarise(elasticity = coef(model)[["log\_PRICE"]])

# Sort the products by elasticity to find the most and least price-elastic products

most\_elastic <- elasticities %>%

arrange(desc(elasticity)) %>%

top\_n(5, elasticity)

least\_elastic <- elasticities %>%

arrange(elasticity) %>%

top\_n(-5, elasticity)

# Display the results

print("Most Price-Elastic Products:")

print(most\_elastic)

print("Least Price-Elastic Products:")

print(least\_elastic)

#Question4

# Calculate price elasticity for each product

df$log\_UNITS <- log(df$UNITS + 1) # Adding 1 to avoid log(0)

df$log\_PRICE <- log(df$PRICE)

elasticity\_models <- df %>%

group\_by(UPC) %>%

do(model = lm(log\_UNITS ~ log\_PRICE, data = .)) %>%

summarize(elasticity = coef(model)[["log\_PRICE"]])

# Extract products with the highest (absolute) negative elasticity

most\_elastic\_products <- elasticity\_models %>%

arrange(elasticity) %>%

top\_n(-5, elasticity) #5 most elastic products

# Print the UPCs and their elasticities for the most elastic products

print("Products to consider for price reduction to maximize sales:")

print(most\_elastic\_products)