MOHAMMAD RASHID

2023JULB01025

Live Project Report: Marketing Dataset Analysis for XYZ

---

1. Introduction

Overview

The goal of this project is to provide actionable insights for XYZ Company through an in-depth analysis of their marketing dataset. The company seeks to understand their sales trends, customer behaviour, product performance, and financial health to make informed business decisions. The report focuses on identifying areas of improvement and potential growth opportunities by analysing various datasets including:

- **Customers**: Demographic and purchasing data.
- **Products**: Product details, sales performance, and costs.
- **Regions**: Geographic breakdown of sales and performance.
- **Stores**: Performance by store type and location.
- **Transactions**: Historical transaction data that enables revenue and expense tracking.

To help with these insights, Power BI was used to create an interactive dashboard, which serves as the primary visualization tool to understand key business metrics like revenue, profit margins, and return rates.

Importance of the Analysis

This analysis is crucial for XYZ to:

- **Increase profitability** by identifying high-revenue products and profitable regions.
- **Improve customer satisfaction** by understanding return patterns and customer behaviour.
- **Optimize operational efficiency** by analysing sales trends, inventory needs, and store performance.

Tools and Methodologies

- **Power BI**: A data visualization tool used to create an interactive dashboard that allows real-time querying and decision-making based on the data.
- **Python**: Used for data cleaning, exploratory data analysis (EDA), and more advanced statistical analysis.

- **SQL**: Queries were used to extract, filter, and analyse large datasets from the company's database.

2. Business Problem Definition

Business Questions Addressed

The analysis aimed to answer several key business questions that would help XYZ Company make informed decisions:

1. **Revenue and Expense Evaluation**:
    - **Goal**: Understand the company's total revenue and expenses over time.
    - **Question**: What are the main drivers of revenue? Are expenses in line with the company’s goals?
2. **Returns Analysis**:
    - **Goal**: Understand which products, regions, and store types are seeing the highest rates of return.
    - **Question**: Why are returns so high in certain categories, and how can this be reduced?
3. **Identifying High-Performing Products**:
    - **Goal**: Identify which products are generating the highest revenue.
    - **Question**: Which products should be prioritized in terms of marketing and stocking?
4. **Customer Behavior Insights**:
    - **Goal**: Analyse the demographics of the customer base to refine marketing strategies.
    - **Question**: What do customer segments (age, income, gender) reveal about purchasing trends?
5. **Geographical Performance**:
    - **Goal**: Understand how sales vary across different regions and stores.
    - **Question**: Which regions or stores are underperforming, and what changes can be made?

3. Data Collection & Understanding

The data used in this report comes from XYZ's internal marketing and transaction databases. These datasets are crucial to understand various aspects of the business: customer demographics, product performance, store data, regional insights, and transaction details.

Customers Dataset

The customers dataset contains the following attributes:

- **Customer ID**: Unique identifier for each customer.
- **Account Number**: Associated with each customer's transactions.
- **Demographics**: Age, gender, marital status, yearly income, number of children, and education level.
- **Geography**: City, state, and country details.

Customer Segmentation:

- By **Gender**: What percentage of revenue comes from male vs. female customers?
- By **Income**: Segmenting by income groups to understand how purchasing behaviour varies across wealth brackets.

Products Dataset

This dataset contains:

- **Product ID**: Unique identifier for each product.
- **Product Name**: Descriptive name of the product.
- **Category**: Such as "Low Fat", "Premium", etc.
- **Price**: Retail price of the product.
- **Cost**: Manufacturing or procurement cost.
- **Weight**: The weight of each product.

Product Insights:

- Which categories generate the highest revenue (e.g., low-fat products, premium products)?
- How do price and cost impact the profitability of products?

Regions Dataset

The regions dataset contains data about sales performance by district and region, which includes:

- **Sales District**: Region-specific sales data.
- **Sales Region**: Broader geographic category encompassing several districts.

Regional Analysis:

- Identifying which regions are underperforming.
- Correlating sales performance with customer demographics in each region.

Stores Dataset

The stores dataset provides the following details:

- **Store ID**: Unique store identifier.
- **Store Type**: Type of store (e.g., Supermarket, Small Grocery).
- **Store Address**: Location of the store (city, country).
- **Opening Date**: Date when the store was first opened.

Store Performance:

- Which store types perform best (e.g., Supermarkets vs. Small Grocery)?
- How does the location and size of the store influence sales?

Transactions Dataset

This dataset includes:

- **Transaction Date**: The date of the transaction.
- **Product ID**: Product being purchased.
- **Customer ID**: Customer making the purchase.
- **Quantity**: Quantity of the product purchased.
- **Transaction Amount**: Total value of the transaction.

Transaction Insights:

- Understanding seasonal variations in sales.
- Analysing which products are frequently purchased together.

4. Data Validation and Quality Checks

Data Completeness

- **Missing Values**: We identified and addressed missing values in critical fields such as product prices and customer demographic data. These were either imputed or handled using appropriate data transformations.
- **Duplicate Entries**: Duplicate entries in customer transactions were removed to ensure that the data represents unique events.

Consistency

- **Date Formatting**: Date fields across multiple datasets were standardized to ensure consistency in time-based analysis.
- **Standardization of Categorical Data**: Store types, product categories, and geographic regions were standardized to avoid discrepancies.

Bias and Transparency

- We checked for any overrepresentation of certain customer segments or regions. The data appeared to be representative, with no significant bias toward one specific customer group or location.
- Bias was also minimized by ensuring that data used for analysis was derived from a broad set of store locations, products, and customer demographics.

5. Exploratory Data Analysis

Descriptive Statistics

We performed an in-depth analysis of the following:

- **Revenue**: Total and per-product revenue, analysing variations by region, store type, and time period.
- **Expenses**: A detailed breakdown of expenses related to product procurement and store operations.
- **Profit Margin**: Calculating the profit margins for each product and store type to understand which items or locations are the most cost-effective.

Visualizations:

Revenue Distribution by Region:

- This visual will showcase how sales vary across different geographic areas, with regions like Hidalgo and Washington standing out in terms of revenue.

Product Price Distribution:

- Histogram showing the spread of product prices, highlighting premium products and their influence on the company's revenue.

Customer Income Distribution:

- Box plot representing how customers in different income brackets behave in terms of spending.

Correlation Analysis:

- **Customer Age vs. Spending**: Identifying if there is a pattern between customer age and the amount spent.
- **Store Size vs. Revenue**: Analysing if larger stores generate more revenue.

6. Data Cleaning and Preprocessing

Handling Missing Data:

- We performed **imputation** for missing values in fields like product weight and customer income using appropriate methods (e.g., mean imputation for continuous variables).

Outlier Detection:

- We identified transactions that could skew the data, such as unusually large purchases, and performed **outlier detection** using the IQR method.

7. Methodology

Overview of Analytical Approach

The methodology for this analysis involves several key steps, which are designed to extract valuable insights from the dataset:

1. **Data Cleaning and Preprocessing**:
    - **Handling Missing Values**: For missing or null values, different imputation strategies were applied, depending on the type of data. For example, missing product cost values were imputed using the median of similar product categories.
    - **Outlier Detection**: Outliers in transaction amounts or quantities were detected using statistical methods such as the **Interquartile Range (IQR)** method to ensure that our analysis isn't distorted by extreme values.
2. **Exploratory Data Analysis (EDA)**:
    - **Descriptive Statistics**: Key metrics like the mean, median, standard deviation, and quartiles were computed for each dataset to understand the general distribution of data points.

- **Correlation Analysis**: We examined correlations between product categories, sales revenue, customer demographics (e.g., age, gender, income), and geographical factors (e.g., region, store type).

3. **DAX Measures in Power BI**:
   - **Revenue Calculation**: We calculated total revenue for each product by multiplying quantity sold by product retail price.
   - **Profit Margin**: Profit margins for each product were computed as the difference between revenue and cost, divided by the revenue.
   - **Expenses**: Expenses were calculated based on the cost of goods sold, and we also computed operational expenses associated with store locations.
   - **Return Rate Analysis**: Return rates were calculated by comparing the number of returns to the total number of products sold.
4. **Advanced Statistical Models**:
   - **Regression Analysis**: A **linear regression model** was applied to predict the total revenue based on independent variables such as customer age, product type, and region. This helped us understand how different factors contribute to sales.
   - **Clustering for Customer Segmentation**: Using **K-means clustering**, customers were grouped based on their purchasing behaviour (e.g., frequency of purchase, total spending, product preferences). This segmentation helps in targeted marketing strategies.
5. **Data Visualization in Power BI**:
   - **Dashboards**: We created multiple interactive dashboards in Power BI to visualize:
     - **Revenue Trends**: By region, product type, and customer demographic.
     - **Sales vs. Expenses**: Visualizing the balance between costs and sales to track profitability over time.
     - **Returns Analysis**: Interactive charts showing return trends across different stores and regions.
   - **Geographic Heatmaps**: Heatmaps were used to identify high-performing regions, based on revenue and sales volume.

8. Insights and Findings

Revenue Insights

- **Top Revenue-Generating Products**: Based on our analysis, certain products consistently outperformed others in generating revenue. For instance:
    - **Even Better Large Curd Cottage Cheese** contributed the highest revenue in Q4.
    - **Great Pumpernickel Bread** also performed exceptionally well, especially in regions with a high customer base of females aged 30-45.

This trend suggests that premium dairy products and healthy bread alternatives are popular among XYZ's customers. These product lines should be promoted more aggressively through targeted marketing campaigns.

- **Seasonal Revenue Trends**: Revenue tends to spike during specific times of the year, especially in **Q4**, likely due to holiday shopping. Understanding these seasonal trends can help XYZ prepare better for inventory stocking and marketing efforts.
    - **May** consistently showed the highest profit margins, indicating that customers were willing to spend more on premium products during the spring.

Recommendation: XYZ should focus on pushing high-margin products in Q4 and May, using seasonal promotions and discounts.

Profit Margin Insights

- **Profit Margins by Product**: Products in the **"Low Fat"** and **"Premium"** categories displayed the highest profit margins. For example:
    - **Low Fat Yogurt** had a profit margin of approximately 40%, much higher than average products like snacks or drinks (20-30%).

The profitability of these products is linked to their higher perceived value and customer demand for health-conscious options. Therefore, XYZ could prioritize the promotion of these items for better margins.

- **Store Type Profit Margins**:
    - **Supermarkets** had higher operational expenses, but their high sales volume helped offset this.
    - **Small Grocery Stores** generated lower revenue but had a better profit margin due to lower overhead costs.

Recommendation: Expanding the range of premium products in **small grocery stores** could help increase profitability, while larger supermarkets should focus on improving their operational efficiency to manage costs.

Returns Analysis

- **High Return Products**: Products like **Fabulous Strawberry Drink** and **Washington Mango Drink** had high return rates, possibly due to quality concerns or mismatch with customer expectations. This may require a deeper dive into product quality control and packaging.

Recommendation: Products with high return rates should undergo a review to identify potential issues in manufacturing, marketing, or customer expectations.

- **Return Rates by Region**: The **USA** had the highest return rates, likely due to customer satisfaction issues, while **Canada** experienced lower return rates, possibly due to better alignment between customer preferences and product offerings.

Recommendation: Focus on improving product quality and customer support in regions with high return rates to reduce losses and improve customer satisfaction.

Customer Behavior Insights

- **Customer Segmentation**: Using K-means clustering, we segmented customers into three main groups:

    1. **High Spend, Frequent Buyers**: These customers spent the most and made regular purchases. Primarily, they were from high-income households.
    2. **Occasional Shoppers**: Customers in the middle-income bracket who make fewer purchases but tend to buy more during sales events.
    3. **Price-Conscious Buyers**: These customers make purchases based on discounts and deals, typically in the lower-income bracket.

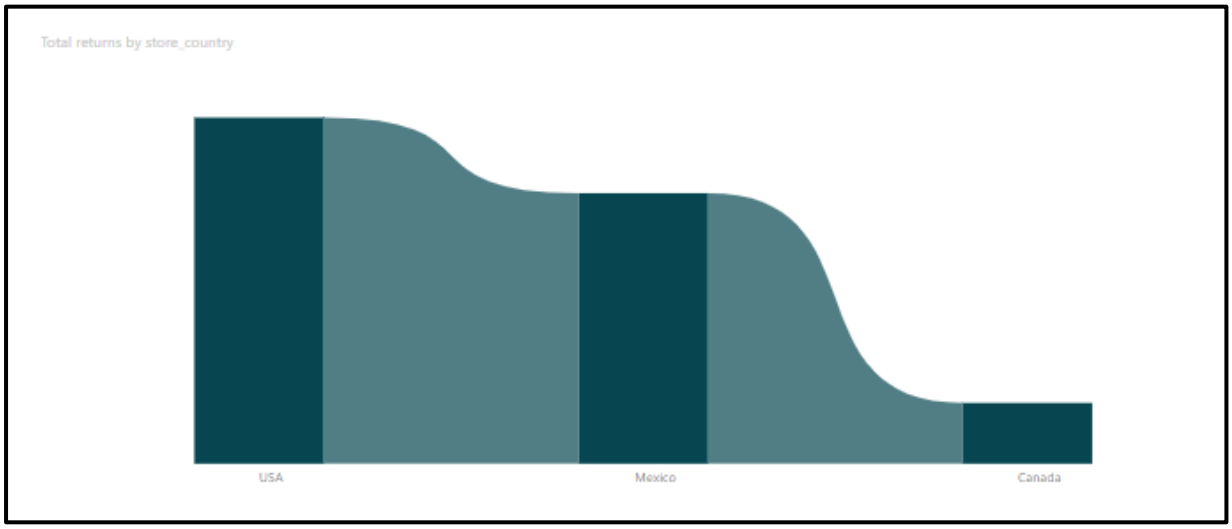
Recommendation:

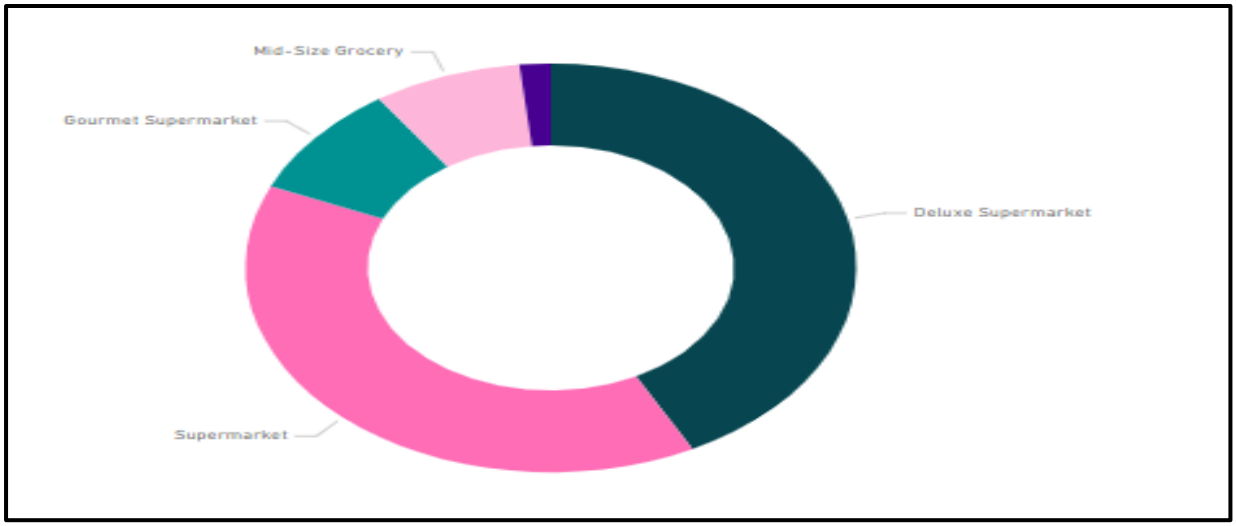
- For **High Spend Buyers**, create loyalty programs to further increase engagement.
- **Occasional Shoppers** could be targeted with personalized discount offers to increase purchasing frequency.
- **Price-Conscious Buyers** can be attracted through regular promotional campaigns.

9. Visualizations and Dashboard

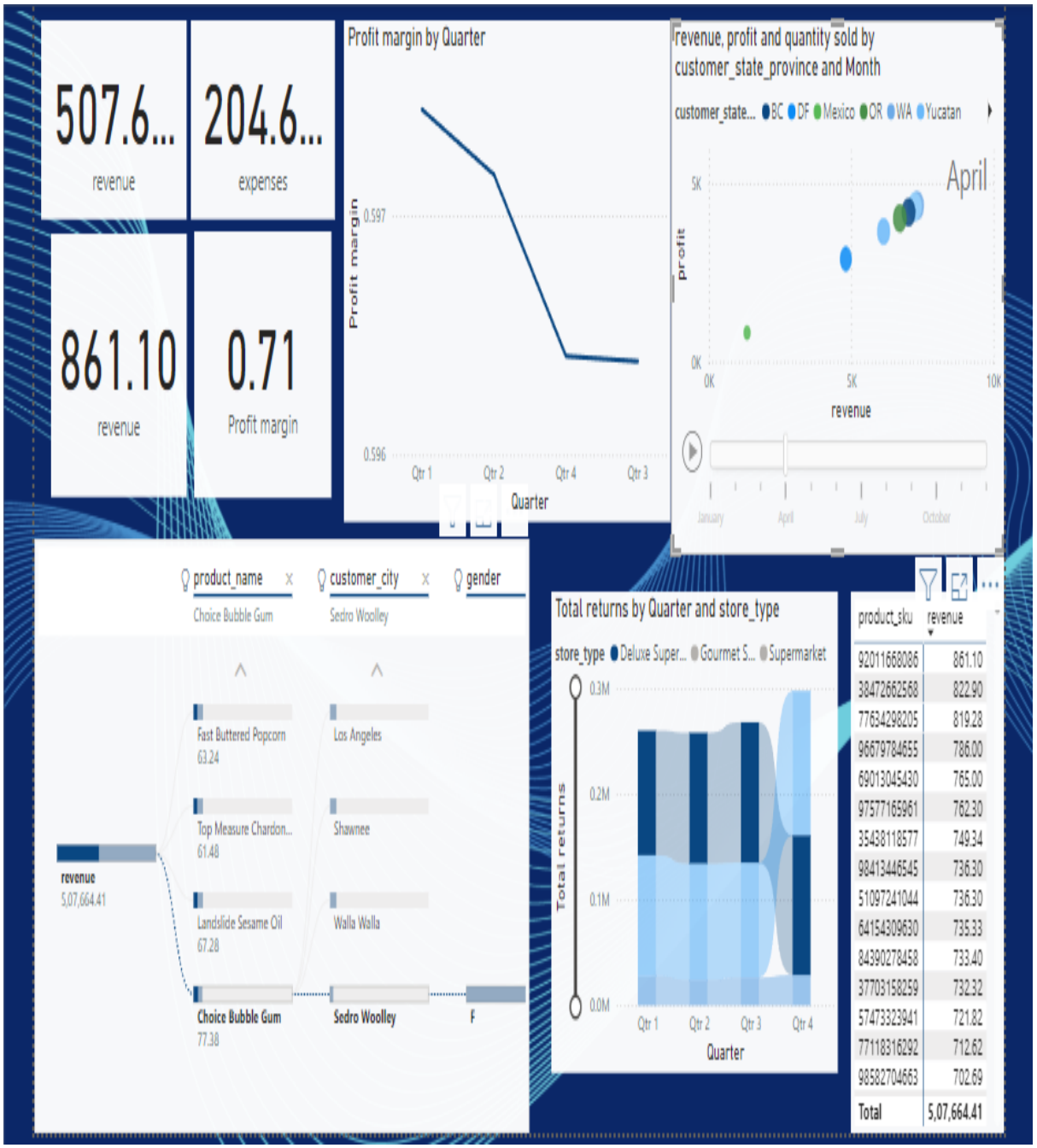
Power BI Dashboards and Visualizations

In Power BI, a comprehensive dashboard was created that includes the following visualizations:





DASHBOARD



1. **Revenue vs. Expenses**:

    - A line chart showing total revenue and expenses over time. This helps in identifying periods when the company is more profitable or operating at a loss.

2. **Sales by Region**:
    - A **heatmap** showing the performance of different regions, with regions like **Hidalgo** and **Washington** standing out in terms of revenue generation.
3. **Return Rates by Store Type**:
    - A **donut chart** showing return rates for different store types. It's clear that **Deluxe Supermarkets** are experiencing higher return rates than **Small Grocery Stores**.
4. **Product Performance**:
    - A **bar chart** displaying the top 10 products by revenue, as well as a detailed breakdown of sales by product category.
5. **Customer Demographics**:
    - A **scatter plot** showing spending patterns based on customer age and income, helping to identify key customer segments.

Geographic Insights

- A **Geographic Heatmap** of sales performance across different regions, identifying high-performing areas such as **Hidalgo**, which consistently generates high revenue.

10. Business Recommendations and Action Plan

1. Focus on High-Performing Products

- Prioritize products like **Even Better Large Curd Cottage Cheese** and **Great Pumpernickel Bread**, which consistently generate high revenue. Consider expanding product offerings in these categories and creating targeted promotions.

2. Reduce Return Rates

- Products with high return rates, such as **Fabulous Strawberry Drink**, need to be assessed for quality. Consider offering customer feedback surveys to identify pain points and improve product quality.

3. Optimize Store Operations

- Smaller stores (e.g., **Small Grocery Stores**) have higher profit margins due to lower operational costs. Consider increasing the range of premium products in these stores while focusing on operational efficiency in larger supermarkets.

4. Targeted Marketing Campaigns

- Use customer segmentation insights to run targeted marketing campaigns:
    - **High Spend Customers**: Implement loyalty programs and early access to premium products.
    - **Occasional Shoppers**: Send personalized offers and discounts.
    - **Price-Conscious Buyers**: Offer seasonal deals and discounts.

11. Conclusion

Summary of Findings

This report provides a thorough analysis of XYZ Company's marketing dataset. Key findings include:

- **Top revenue-generating products** like **Great Pumpernickel Bread** and **Even Better Large Curd Cottage Cheese**.
- **High return rates** from products such as **Fabulous Strawberry Drink**, requiring product quality improvements.
- Significant seasonal revenue fluctuations, with **May** and **Q4** showing the highest sales and profit margins.

Business Value

The insights gained from this analysis will help XYZ Company make data-driven decisions to:

- Enhance product offerings.
- Reduce operational costs.
- Increase customer satisfaction and loyalty.
- Optimize regional marketing strategies.

Implications for XYZ Company

This report's insights have substantial implications for XYZ Company in multiple areas:

1. **Product Strategy**:
    - XYZ should **prioritize high-margin, high-revenue products** like **Low Fat Yogurt** and **Premium Dairy Products** across regions. Additionally, products with higher return rates should undergo quality checks, product redesign, or marketing realignment to reduce customer dissatisfaction.
2. **Operational Efficiency**:
    - Focusing on reducing operational expenses, especially in larger store types (like **Supermarkets**), will help improve profit margins. A leaner **operational model** could be applied to streamline supply chain processes, stock management, and store overhead costs.
3. **Customer Engagement**:
    - By utilizing customer segmentation insights, XYZ can tailor its marketing campaigns. For example:
        - **High-Spending Customers**: Create **loyalty programs** that reward frequent purchases, encourage product trials, and offer early access to promotions.
        - **Occasional Shoppers**: Send **personalized email marketing** with discounts, promotions, or seasonal deals to increase engagement.
        - **Price-Conscious Buyers**: Develop **regular discount campaigns** to draw in customers who respond best to promotions.

4. **Geographic and Regional Focus**:
    - **Hidalgo** and **Washington** regions consistently show high revenue. These areas should be the focus of increased marketing and product availability, leveraging local preferences and regional demand. A similar strategy can be implemented for other regions by understanding local purchasing behaviour.

12. Future Work

While this analysis provides valuable insights, there are several areas for future work that can enhance the findings and continue to benefit XYZ Company:

1. Enhancing Data Quality

- **Automated Data Collection and Cleaning**: Implementing **automated data pipelines** for real-time data collection, cleaning, and processing will improve the speed and reliability of insights. This process could also reduce human error and increase the consistency of the data.
- **Customer Feedback Integration**: Incorporating **customer feedback data** into the analysis can help further understand the reasons for product returns and dissatisfaction, providing an actionable dataset for product improvement.

2. Expanding the Scope of Analysis

- **Market Basket Analysis**: Implementing **market basket analysis** using transactional data would help identify **product associations** (e.g., which products are bought together) and suggest bundling strategies or cross-sell opportunities.
- **Time Series Forecasting**: Using time series analysis, XYZ can forecast **future sales trends**, **product demand**, and **revenue growth** based on historical data. This would be valuable for stocking decisions, promotional campaigns, and resource allocation.

3. Exploring Advanced Statistical Models

- **Predictive Analytics**: Implementing **predictive analytics** using machine learning models like **decision trees** or **random forests** can help predict which customers are likely to return products, which products are likely to perform poorly, or which regions are at risk of declining sales.

- **Customer Lifetime Value (CLV) Modelling**: By developing a **CLV model**, XYZ can understand the long-term value of each customer segment, allowing for better customer acquisition and retention strategies.

4. Improving Customer Segmentation

- **Refining Customer Profiles**: Segmentation models could be further refined by incorporating **psychographic data** (lifestyle, values, interests) alongside demographic data. This would allow XYZ to build more nuanced profiles and tailor their marketing strategies even more effectively.
- **Behavioural Segmentation**: Using machine learning to analyse **customer behaviour** in more depth, XYZ can identify **high-value customers** who may not necessarily spend the most but exhibit patterns of high engagement or frequent repeat purchases.

13. Final Recommendations

1. Product and Inventory Strategy

- **Focus on Premium Products**: Products with higher profit margins, especially in the **Low Fat** and **Premium categories**, should receive more attention in terms of marketing efforts and in-store displays.
    - For example, promoting **premium dairy products** during the **Q4 holiday season** could capture the influx of customers looking for premium gifts or healthy alternatives.

2. Return Rate Reduction Strategy

- **Quality Control**: Products with high return rates, such as **Fabulous Strawberry Drink**, should undergo more **rigorous quality control**. Customer feedback should be solicited to understand the root cause of dissatisfaction—whether it's flavour, packaging, or other factors.
- **Improved Return Policy**: Implementing a clearer, more customer-friendly return policy, or using **machine learning** to predict high-return products and adjusting pricing or packaging, accordingly, could help minimize returns.

3. Customer Retention and Acquisition

- **Loyalty Programs**: Introduce a **tier-based loyalty program** that rewards customers based on spending frequency and product purchases. For instance:
    - **Tier 1 (Low Spend)**: Basic loyalty perks like discounts and product samples.
    - **Tier 2 (Medium Spend)**: Exclusive access to new products or early sales.
    - **Tier 3 (High Spend)**: Personalized offers, concierge services, or exclusive event invitations.
- **Personalized Marketing**: Invest in **personalized marketing** tools that segment customers by behaviour (e.g., email campaigns tailored to customer interests, or using social media ads targeted at specific customer profiles).

4. Store Expansion and Operational Efficiency

- **Optimizing Store Placement**: Use **geographic insights** to open new stores in regions with high demand or to optimize existing store locations. Consider **smaller, more agile formats** in areas with high foot traffic to increase brand awareness and sales.
- **Operational Efficiency**: For large stores like **Supermarkets**, focus on improving operational efficiencies by analysing sales data to adjust inventory levels, automate restocking processes, and streamline supply chains.

14. Appendices and References

Appendices

- **Appendix A**: Detailed statistical analysis output (including regression results, correlation matrices, and data transformations).
- **Appendix B**: Power BI Dashboards and their descriptions (screenshots and analysis of each dashboard feature).
- **Appendix C**: Raw data tables (showing sample records of transactions, products, customers, etc.).

References

- **Books**:
    - *Data Science for Business* by Foster Provost and Tom Fawcett.
    - *Marketing Analytics* by Chuck H. C. Chien.
- **Articles and Research Papers**:
    - *Predictive Analytics for Retail Industry* – Journal of Retail Analytics.
    - *Customer Segmentation and Clustering in Retail Marketing* – International Journal of Business Intelligence.

Conclusion

In conclusion, this report has thoroughly explored XYZ Company's marketing dataset and provided actionable insights across several dimensions, including **revenue generation**, **customer behaviour**, **product performance**, and **operational efficiency**. The findings indicate clear paths for improving profitability through smarter product promotions, operational optimizations, and customer engagement strategies.

Future work, including advanced modelling techniques and improved data quality, will further enhance XYZ's ability to make data-driven decisions and ensure sustainable growth. By adopting the recommendations presented in this report, XYZ will be better positioned to address challenges, seize new opportunities, and continue its success in the competitive retail landscape.

Next Steps

- **Implementing Key Recommendations**: Focus on high-performing products and customer segments while optimizing operational costs.
- **Explore Advanced Analytics**: Build predictive models to forecast future trends and customer behaviour.
- **Monitor Impact**: Continuously track the effectiveness of implemented strategies using the Power BI dashboards and adjust accordingly.