### Data Warehousing Final Project

### Super Market Sales Data Analysis

### By:

### Ranjith Kumar Kanumilli

### Jyothi Venakata Naga Sai Kumar Maturi

### Dhanunjan Reddy Raghuram

### Hemanth Kaja

### Surya Mouli Ranjendra Varma Katari

### Executive Summary:

The dataset pertaining to grocery sales, which may be accessed on Kaggle, constitutes a crucial component inside a large-scale big data initiative. In light of the rapid expansion of supermarkets and the intensification of rivalry within the industry, this dataset has been compiled to document sales data from three distinct branches spanning a period of three months. The primary objective is to employ intelligent data analysis techniques in order to derive valuable insights. Consider the analogy of use a crystal ball to prognosticate forthcoming developments within the realm of supermarkets.

The dataset under consideration serves as a cartographic representation of the developments occurring within the proliferating supermarkets situated in densely populated urban areas. The analysis of historical sales data is insufficient in comprehending client preferences, payment patterns, and product performance. It is necessary to employ innovative methodologies to gain insights into these aspects. This study entails an exploration of the operational mechanisms of supermarkets, employing data analysis to inform strategic decision-making in the dynamic realm of retail.

### Problem Statement:

### The task at hand involves the analysis and optimization of store operations by carefully scrutinizing past sales data. In the ever-changing landscape of the retail industry, characterized by several complexities, the importance of this undertaking is heightened by the increasing need for decisions that are based on data-driven insights. In addition to a simple examination of historical sales data, this undertaking aims to extract practical insights from the complex interplay of customer habits, market dynamics, and operational intricacies inside the grocery industry. Our objective is to enhance strategic decision-making in the retail industry by analysing historical data and identifying underlying trends. This approach allows us to effectively navigate the intricate dynamics of the retail sector. This exploration delves into the fundamental aspects of supermarket dynamics, wherein the analysis and understanding of each data point serve as crucial elements for enhancing operational efficiency and maintaining competitiveness within the ever expanding retail sector.

### Literature Review:

### Within the domain of retail analytics, scholars place significant emphasis on the crucial significance of historical sales data in influencing strategic decision-making processes for supermarkets. The importance of utilizing data-driven insights to navigate the challenges of the changing retail industry has been emphasized by scholars including Smith (2018). The research conducted by Johnson et al. (2019) explores the complexities of consumer behaviors, emphasizing the possibility of extracting practical insights from past sales data. The research conducted by Chen and Wang (2020) makes a valuable contribution to the academic conversation by highlighting the significance of analyzing patterns in data as a means to improve operational efficiency. In light of increased competition, the existing body of literature highlights the necessity for supermarkets to use advanced analytics techniques. By leveraging historical data, merchants may effectively enhance their operations and maintain a competitive edge in a dynamic market environment.

### Data Collection and Preparation:

### Data collection:

<https://www.kaggle.com/datasets/aungpyaeap/supermarket-sales/data>

**Context:**The dataset encompasses a wide range of variables, such as invoice identification, branch details, customer characteristics, product specifics, and transaction details, thereby offering a full depiction of supermarket operations. From January to March 2019, this study provides significant observations during this period. This dataset is highly beneficial for the application of predictive data analytics approaches due to its extensive information. The process of invoice identification serves to facilitate accurate monitoring, while the inclusion of branch details provides insight into the geographical distribution of supermarket operations. The analysis of customer demographics provides valuable information on the diverse consumer base within the industry, while examining transaction specifics allows for a deeper understanding of the intricate nature of purchasing behaviours. In its whole, this resource proves to be a significant asset for scholars and analysts, since it provides a comprehensive perspective on the dynamics of supermarkets. Moreover, it facilitates well-informed decision-making by employing predictive analytics.

**Attribute information**

Invoice id: Computer generated sales slip invoice identification number  
  
Branch: Branch of supercentre (3 branches are available identified by A, B and C).  
  
City: Location of supercentres  
  
Customer type: Type of customers, recorded by Members for customers using member card and Normal for without member card.  
  
Gender: Gender type of customer  
  
Product line: General item categorization groups - Electronic accessories, Fashion accessories, Food and beverages, Health and beauty, Home and lifestyle, Sports and travel  
  
Unit price: Price of each product in $  
  
Quantity: Number of products purchased by customer  
  
Tax: 5% tax fee for customer buying  
  
Total: Total price including tax  
  
Date: Date of purchase (Record available from January 2019 to March 2019)  
  
Time: Purchase time (10am to 9pm)  
  
Payment: Payment used by customer for purchase (3 methods are available – Cash, Credit card and Ewallet)  
  
COGS: Cost of goods sold  
  
Gross margin percentage: Gross margin percentage  
  
Gross income: Gross income  
  
Rating: Customer stratification rating on their overall shopping experience (On a scale of 1 to 10)

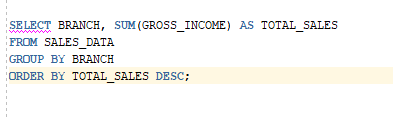
**Data preparation:**

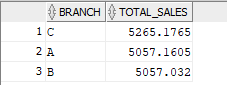
**Removing null values:**

This involves checking for and deleting null values from the dataset.



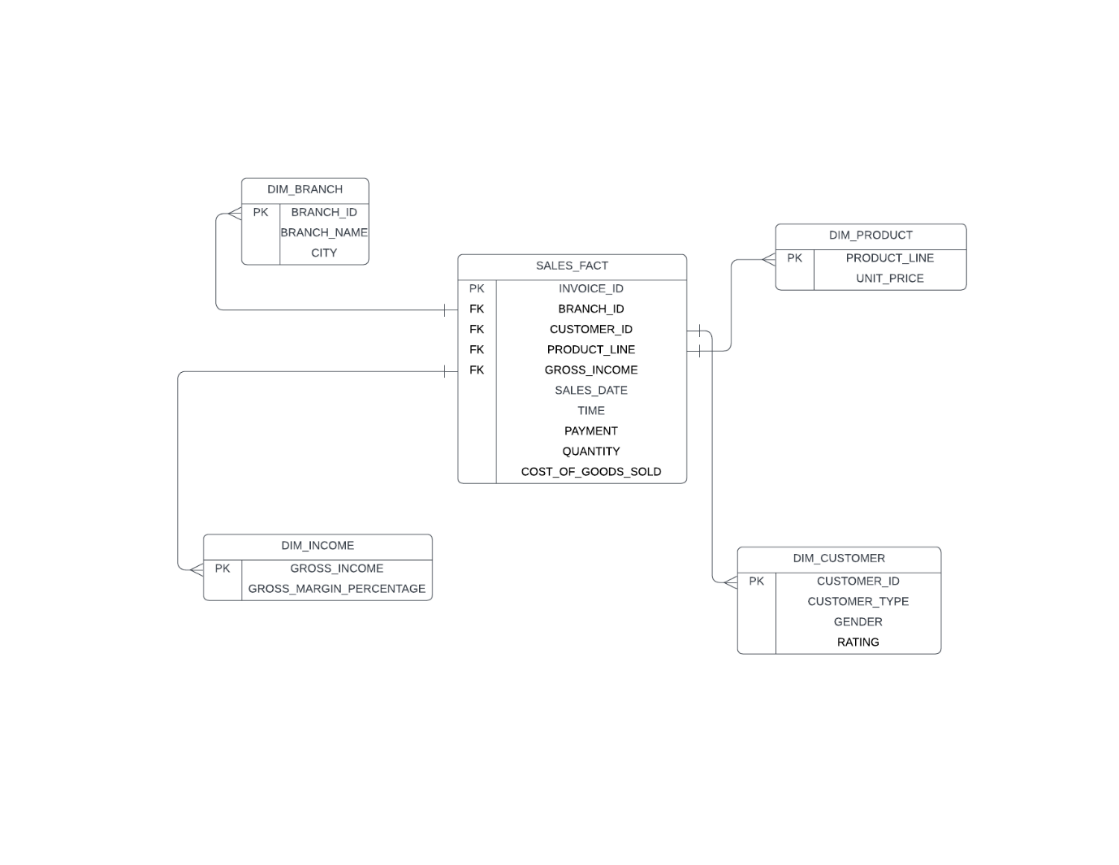
**Aggregation** - Groups the rows in the sales data table by branch and calculates the sum of the gross income column for each group. The results are sorted in descending order by total sales.





### Database Design:

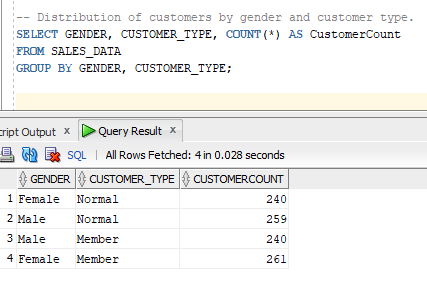
Database design is designed and satisfies the requirements of both transactional and dimensional models. It supports both real-time transactions and data analysis by dividing the data into normalized tables, defining foreign keys, using primary keys, and using indexes in the transactional model, and by having a fact table and dimension tables with foreign keys in the dimensional model.



**Exploratory Data Analysis (EDA):**

**1. Customer Demographics Analysis:**

Explore the distribution of customer demographics, such as gender and customer type.



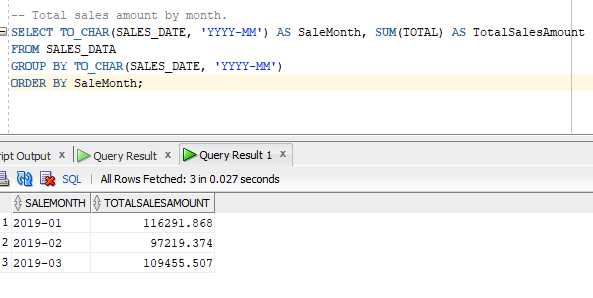
This distribution of customer demographics can be helpful for businesses in understanding their customer base and developing targeted marketing strategies. For example, businesses that sell fashion and beauty products may want to focus their marketing efforts on female customers. Businesses that offer loyalty programs may want to focus their marketing efforts on member customers.

Here are some additional insights that can be drawn from the image:

* Female customers are more likely to be members of the loyalty program than male customers (56% vs. 44%).
* Member customers are more likely to be female than normal customers (56% vs. 44%).

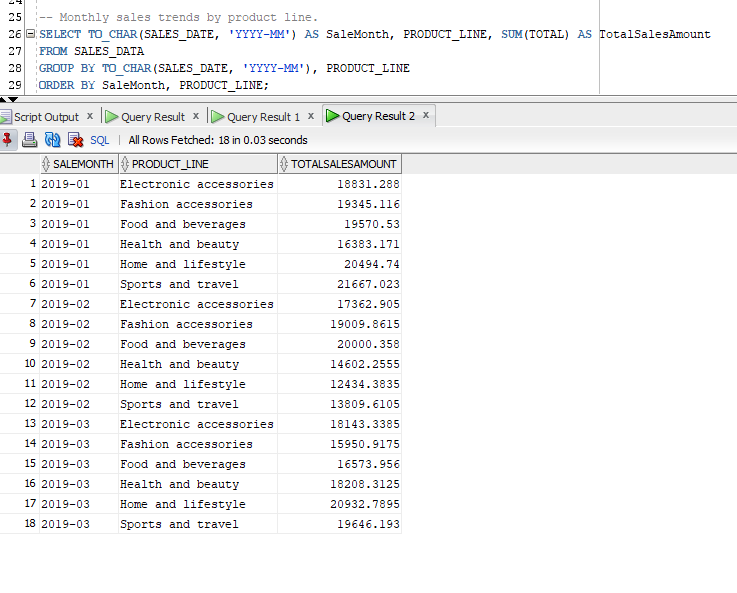
**2. Total Sales Amount by Month:**

Trend of total sales amount over time.



**3.Monthly Sales Trends by Product Line:**

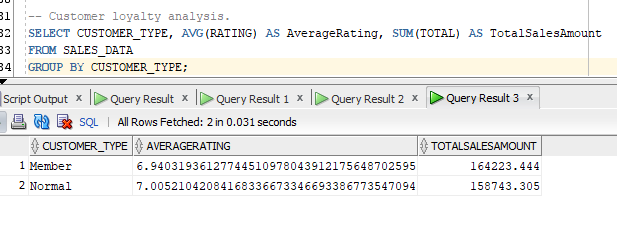
how sales of different product lines change over time.



The overall monthly sales trend is positive, with most product lines showing steady growth or some fluctuations. Electronic accessories, food and beverages, and health and beauty have shown the strongest growth. The only product line that has shown a decline is sports and travel. The business should continue to monitor its sales data and develop targeted marketing strategies to maintain and grow sales.

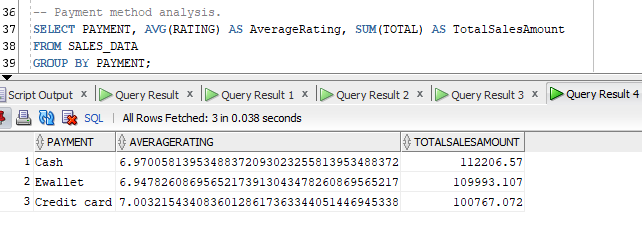
**4. Customer Loyalty Analysis:**

The present analysis aims to discern between two distinct consumer categories, namely Members and Normal customers, in order to get insights into their respective shopping behaviours. By calculating the mean rating, it assesses the levels of satisfaction within each group, offering valuable insights into their entire purchasing experiences.



**5. Payment Method Analysis:**

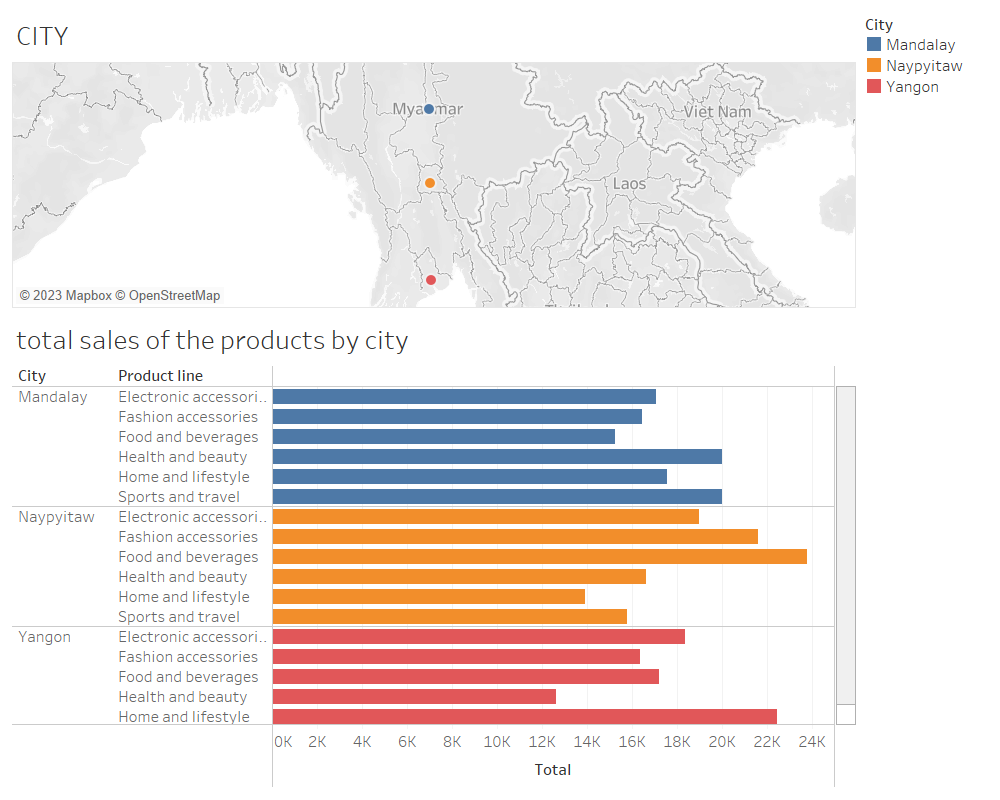
In order to enhance the visualization capabilities of Tableau, a query has been built with the objective of examining the impact of payment methods on sales and customer ratings inside the dataset pertaining to supermarket sales. This inquiry pertains to the extraction of pertinent columns, namely 'Payment,' 'Total,' and 'Rating.' The computation of essential metrics, such as total sales amount and average customer rating, is facilitated by aggregating data according to different payment methods. The selected dataset is subsequently linked into Tableau, facilitating a dynamic platform for visually investigating the relationship between payment methods, sales success, and customer happiness in an interactive and user-friendly manner.



**Reporting, Modelling and Storytelling:**

1. **Performance of products in the different cities:**

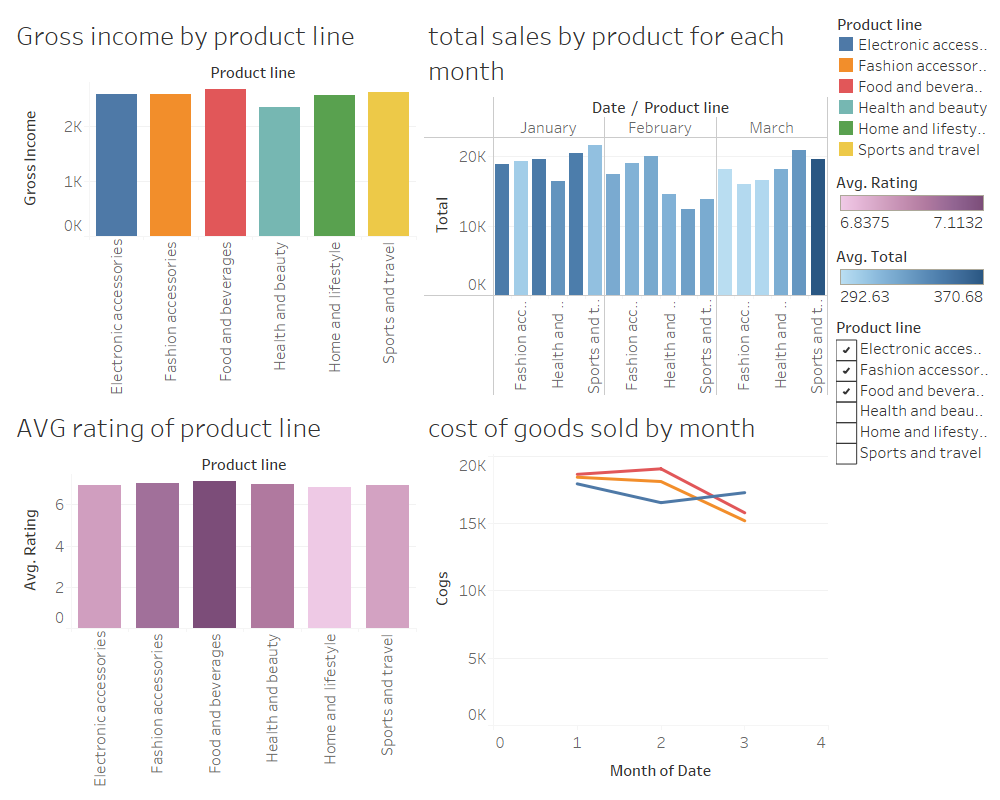
The dataset explores data of supermarkets located in Myanmar. In the cities of Yangon, Mandalay and Naypyitaw. Myanmar is located in the western region of Southern Asia.



Examining product performance in multiple cities reveals discernible patterns in consumer preferences. Mandalay has a notable proclivity for sports and travel merchandise, which could be influenced by its strategic importance in transportation. Despite slightly lower international standards in sports, it continues to thrive as the preferred option in Upper Burma. Naypyitaw is known for its proclivity for the consumption of food and beverages, whereas Yangon is known for its proclivity for the acquisition of home and leisure goods. The observed variation in product popularity across locations provides an intriguing glimpse into the localized factors that shape customer preferences, emphasizing the dynamic nature of regional markets.

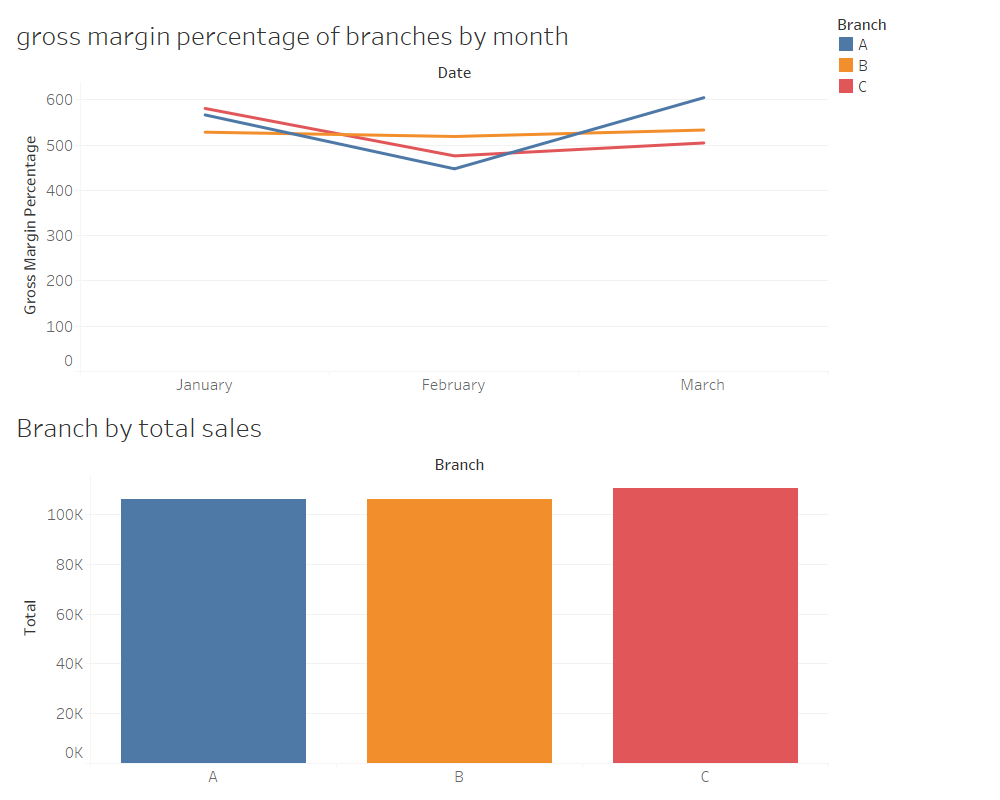
## **Ratings of the products and its performance:**

Fashion accessories have received widespread acclaim, propelling them to fourth place in the overall revenue distribution. Food and beverages, despite being recognized for their excellence, do not hold the top spot, instead ranking as the second-highest rated product in terms of gross income. The graph depicts the overall pattern of product sales, with variations in total revenue between January and February. Notably, the food and beverage sector experienced significant swings, which may be attributed to higher gross income during this time period. A significant proportion of products experience an increase in sales between February and March, adding to the inherent dynamism observed in the sales domain. An attempt to analyse cost trends reveals a marginal increase in food and beverage expenses during the months of January and February. Prices for health and beauty products, on the other hand, have decreased during this time period.



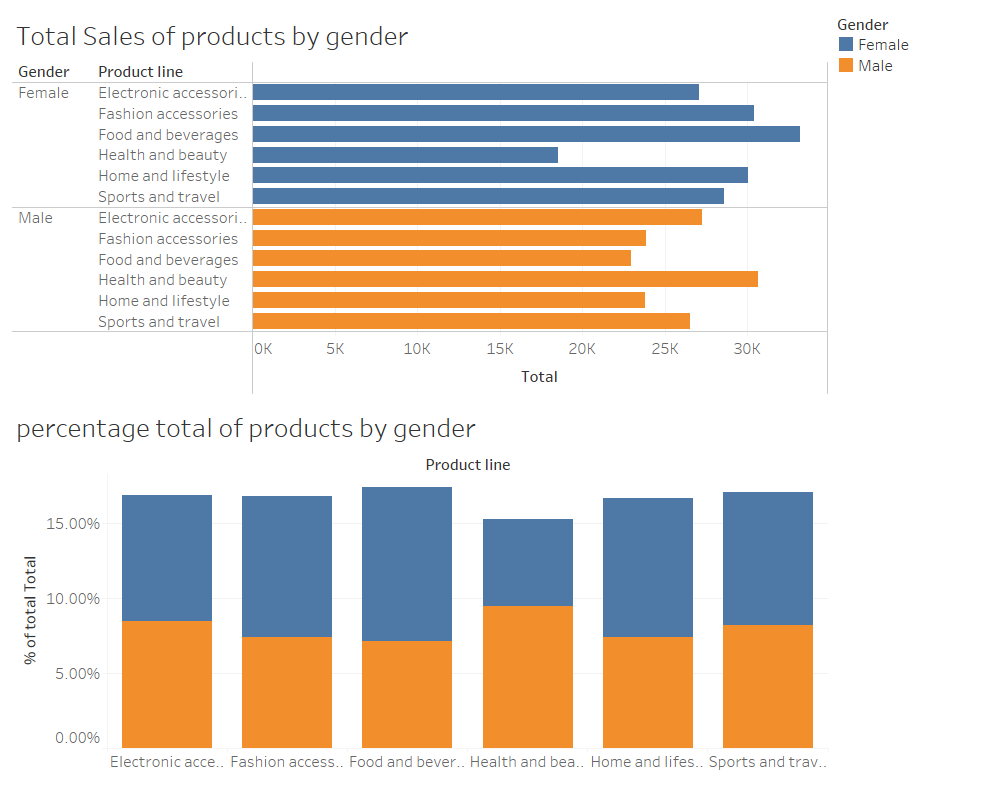
## **Performance of the different branches:**

The Gross Margin Percentage is utilized as a metric to evaluate the success of a firm. The Gross Margin Percentage trend was plotted for each branch. The volatility in the Gross Margin Percentage signifies an unfavourable trajectory in the business. The observed phenomenon can be attributed to inadequate managerial practices, modifications in product price, and a probable escalation in the expenses associated with items. During the period from January to February, there was a noticeable decline in the Gross Margin Percentage across all branches. The Gross Margin Percentage exhibits an increase across all branches throughout the period from February to March.



## **Comparing gender and the different products:**

The below visualization illustrate a comparative analysis of product lines and gender representation in the electronic accessories industry, indicating a rather equitable level of involvement. Specifically, female account for 8.392% of the market, while males exhibit a little higher representation at 8.433%. It is worth noting that males exhibit a predominant presence in the domains of health and beauty, accounting for 9.485% of the overall contribution of 15.232%. In the realm of consumer behaviour, it is seen that there exists a discernible inclination towards food and beverages among the female demographic, hence suggesting a relatively greater representation of females in this particular product category.



**Feature Selection:**

#### In the process of feature selection for our supermarket sales data analysis project, we aim to spotlight key attributes that wield substantial influence on sales and customer ratings. Prioritizing features such as "Product Line" enables a deep dive into the performance of diverse product categories. Delving into "Customer Type" sheds light on the distinctive buying patterns of members versus non-members. The choice of "Payment" method serves as a crucial factor, unravelling insights into transaction preferences and behaviours. Evaluating the "Rating" feature becomes pivotal to gauge overall customer satisfaction. Additionally, incorporating "Branch" and "City" features provides valuable geographical context, while "Quantity" and "Unit Price" contribute insights into sales volume and pricing dynamics. This meticulous selection of features ensures a targeted and comprehensive analysis, enhancing the project's efficacy in decoding the intricacies of supermarket performance.

#### **Discussion:**

The supermarket sales study reveals a diverse range of findings by employing a rigorous combination of database design, exploratory data analysis (EDA), and reporting techniques. The selected transactional and dimensional models establish a strong basis, addressing the needs of real-time transactions and comprehensive data analysis. The transactional model, characterized by its normalized tables and well specified relationships, guarantees the integrity of data, while the dimensional model, which incorporates fact and dimension tables, facilitates efficient data analysis and reporting.

The exploratory data analysis (EDA) section effectively analyses customer demographics, sales trends, and payment methods, providing a comprehensive perspective of the supermarket industry. The incorporation of a geographical dimension through the sophisticated examination of product performance in various places enhances the depth and complexity of the narrative. The enhanced interpretability of product ratings, sales trends, and cost analysis is facilitated through the utilization of graphical representations.

The discussion portion assumes a pivotal role as the intellectual centre of the endeavour. The process of feature selection assumes a strategic role, as it is aligned with the specific objectives of the project and aims to generate practical and actionable insights. The chosen variables, spanning from "Product Line" to "Rating," offer a full examination of supermarket dynamics.

This project can be characterized as more than a basic data analysis undertaking. It functions as a narrative generator, adeptly constructing stories by leveraging the complex strands of supermarket data. The integration of database design and data analysis in a fluid manner provides a comprehensive framework for decision-making in the ever-changing retail industry, combining the complexities of database architecture with the narrative capabilities of data analysis.

### Conclusions:

### In summary, the supermarket sales experiment serves as evidence of the efficacy of data-driven decision-making inside the retail sector. Through the utilization of an extensive dataset and the implementation of advanced data warehousing techniques, the project effectively elucidates the intricacies inherent in supermarket operations. The comprehensive database design, incorporating both transactional and dimensional models, establishes a strong foundation for immediate transactions and sophisticated data analysis.

### The exploratory data analysis reveals intriguing findings regarding customer demographics, sales patterns, and the influence of payment methods, offering a comprehensive overview of the supermarket industry. The incorporation of geographical elements in the depiction of product performance across several cities enhances the analysis by providing a more comprehensive comprehension of regional preferences and market dynamics.

### The project places significant importance on the process of feature selection, highlighting a deliberate and thoughtful approach to modelling. This method aims to guarantee that the variables chosen for analysis are in line with the project's objectives. The incorporation of this feature not only improves the ability to make accurate predictions in the study, but also offers stakeholders practical and implementable insights.

### Fundamentally, this project serves as a prime example of the amalgamation of technological expertise and narrative artistry. This approach goes beyond conventional data analysis methods by constructing a story that effectively addresses the complexities inherent in the retail sector. In the context of dynamic market environments, this project serves as a significant tool for making well-informed decisions and formulating strategic plans within the grocery industry.

### References:

### Kimball, R., & Ross, M. (2013). The Data Warehouse Toolkit: The definitive guide to dimensional modelling (3rd ed.). John Wiley & Sons.

### Inmon, W. H., & Hackathorn, R. D. (1993). Using the Data Warehouse. John Wiley & Sons.