**Project Title: Comprehensive Business Insights from Supermarket Sales: A Data-Driven Analysis Using Power BI and SQL**

# **Presented by: Bodhisattwa Guha**

# **Date: 04-Sep-2024**

# Introduction:

## Project Overview:

* This project involves analysing a dataset containing supermarket sales data. The goal is to identify sales trends, understand customer purchasing behaviour, and provide actionable insights that can help optimize sales strategies.

## Project Objective:

* The objective of this project is to analyse supermarket sales data to uncover key patterns and trends that can guide in data driven decision making in inventory management, promotions and customer engagement.

# Problem Statement:

## Background:

* Retails businesses especially supermarkets need to continuously optimize their sales strategies to remain competitive. Understanding customer purchasing patterns and identifying trends in sales data are crucial for making informed business decision.

## Specific Problems:

* This analysis will address common questions such as: Which product lines generate the highest revenue? What are the peak sales hours? How do customer demographics influence purchasing behaviours?

# Understanding the Data Set:

## Dataset Overview:

* The dataset has been downloaded from Kaggle. It contains sales data from a supermarket chain, spanning over three months and involving three different branches(each city has specific branch). Fields are described as below:

Invoice id: Computer generated sales slip invoice identification number  
  
Branch: Branch of supermarket chain (3 branches are available identified by A, B and C).  
  
City: Location of supermarket chains  
  
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 E-wallet)  
  
COGS: Cost of goods sold  
  
Gross margin percentage: Gross income / Total  
  
Gross income: Total - COGS  
  
Rating: Customer stratification rating on their overall shopping experience (On a scale of 1 to 10).

## Initial Observation:

* During the initial exploration it has been observed that gross margin percentage is same across all the rows, indicating standardized profit margin.
* During initial data exploration, it was also observed that date formats vary across rows, requiring standardization. Correct and update dates to YYYY-MM-DD format in a new column to ensure consistency.

# ETL (Extract, Transform, Load):

## Extract:

* The supermarket\_sales CSV file was uploaded to a MySQL database to facilitate ETL operations. This table contains detailed records of sales transactions from a supermarket chain.

## Transform:

* This phase involved cleaning and transforming the data to ensure it was suitable for analysis.

1. Removing Unnecessary Columns:
   * The gross\_margin\_percentage column was removed because it contained the same value for every row, which added no value to the analysis.
   * SQL Command:

ALTER TABLE supermarket\_sales DROP COLUMN gross\_margin\_percentage;

1. Checking for Duplicates:
   * The dataset was checked for duplicate entries based on the Invoice\_ID to ensure the uniqueness and integrity of the data.
   * SQL Command:

SELECT Invoice\_ID, COUNT(Invoice\_ID) FROM supermarket\_sales GROUP BY Invoice\_ID HAVING COUNT(Invoice\_ID) > 1;

1. Date format consistency:
   * The dataset was found to have inconsistent date formats. A check was conducted to identify these variations, which is crucial for standardizing the dates.
   * SQL Command:  
     SELECT CASE WHEN date LIKE '\_/\_\_/\_\_\_\_' THEN 'M/DD/YYYY' WHEN date LIKE '\_\_-\_\_-\_\_\_\_' THEN 'MM/DD/YYYY' ELSE 'unknown format' END AS date\_format, COUNT(\*) AS format\_count FROM supermarket\_sales GROUP BY date\_format;
2. Standardizing Date Format:
   * A new column formatted\_date was added to store all dates in the standardized YYYY-MM-DD format, ensuring consistency.
   * SQL Command:  
     ALTER TABLE supermarket\_sales ADD COLUMN formatted\_date DATE; UPDATE supermarket\_sales SET formatted\_date = CASE WHEN date LIKE '\_/\_\_/\_\_\_\_' THEN STR\_TO\_DATE(date, '%m/%d/%Y') WHEN date LIKE '\_\_-\_\_-\_\_\_\_' THEN STR\_TO\_DATE(date, '%m-%d-%Y') ELSE NULL -- or use a default date if necessary END;

## Load:

* The transformed data, now with a standardized date format and encoded categorical variables, was loaded back into the supermarket\_sales table, ready for further analysis and visualization.

# Exploratory Data Analysis:

## Statistical Analysis:

* Total Sales Revenue by Product Line: We calculated the total sales revenue for each product line to understand which product lines contribute the most to revenue.
  + Query: SELECT product\_line, SUM(total) FROM supermarket\_sales GROUP BY product\_line;
  + Insight: The **Food and beverages** product line leads with the highest revenue at $56,144.84, while **Health and beauty** trails with the lowest at $49,193.74. The revenue gap between the highest and lowest lines is $6,951.10, and the **Sports and travel** line is just $1,022.01 behind **Food and beverages**, showing a close competition among the top performers.
* Average Product Rating by Product Line: We analyzed the average rating for each product line to gauge customer satisfaction across different product categories.
  + Query: SELECT product\_line, AVG(Rating) FROM supermarket\_sales GROUP BY Product\_line;
  + Insight: The **Food and beverages** line has the highest average rating at 7.11, indicating strong customer satisfaction. In contrast, **Home and lifestyle** has the lowest average rating at 6.84. Most product lines have ratings close to each other, with only minor differences.
* Sales Distribution by Customer Type (Member vs Normal): We explored how sales are distributed between member and non-member customers.
  + Query: SELECT customer\_type, COUNT(\*) AS num\_sales, SUM(Total) AS total\_rev, (SUM(Total) / (SELECT SUM(Total) FROM supermarket\_sales) \* 100) AS pct\_total\_revenue FROM supermarket\_sales GROUP BY customer\_type;
  + Insight: Overall, both customer types are nearly balanced in their contribution to total revenue, with Members slightly leading.
* Most Popular Payment Method: We identified the most frequently used payment method.
  + Query: WITH cte\_payment AS (SELECT Payment, COUNT(Payment) AS num\_transaction, DENSE\_RANK() OVER(ORDER BY COUNT(Payment) DESC) AS ranking FROM supermarket\_sales GROUP BY Payment) SELECT \* FROM cte\_payment WHERE ranking = 1;
  + Insight: The most popular payment method is **Ewallet**, with 345 transactions, making it the top choice among customers. **Cash** follows closely with 344 transactions, and **Credit Card** ranks third with 311 transactions.
* Revenue by Gender and Customer Type: We analyzed how revenue varies based on gender and customer type (Member vs Normal).
  + Query: SELECT Customer\_type, Gender, SUM(Total) FROM supermarket\_sales GROUP BY Customer\_type, Gender;
  + Insight: Overall, **Female Members** contribute the highest revenue at $88,146.94, while **Male Members** generate the lowest revenue at $76,076.50. Among the other groups, **Normal Females** contribute $79,735.98, and **Normal Males** contribute $79,007.32.
* Top 3 Busiest Days of the Week: We found the top three busiest days of the week in terms of total sales.
  + Query: SELECT dayname(formatted\_date) AS week\_day, SUM(Total) AS total\_rev FROM supermarket\_sales GROUP BY dayname(formatted\_date) ORDER BY SUM(Total) DESC LIMIT 3;
  + Insight: The top 3 busiest days of the week in terms of total sales are **Saturday** with $56,120.81, **Tuesday** with $51,482.25, and **Thursday** with $45,349.25. Saturday leads by a significant margin, making it the highest revenue day. Tuesday follows as a strong performer, reflecting a significant mid-week boost. Thursday ranks third, indicating a solid end-of-week performance. This analysis highlights that weekends and early weekdays are key for maximizing sales, with Saturday being the most dominant day.
* Average Unit Price by Product Line: We calculated the average unit price for each product line to understand pricing patterns.
  + Query: SELECT Product\_line, AVG(Unit\_price) FROM supermarket\_sales GROUP BY Product\_line;
  + Insight: **Fashion accessories** have the highest average unit price at $57.15, suggesting a premium positioning, while **Electronic accessories** have the lowest at $53.55. This indicates a pricing strategy where premium lines like **Fashion accessories** may focus on higher-value sales, while lower-priced lines like **Electronic accessories** aim for volume.
* Relationship Between Quantity Sold and Total Revenue: We explored how quantity sold correlates with total revenue.
  + Query: SELECT Quantity, SUM(Total) AS Total\_revenue FROM supermarket\_sales GROUP BY Quantity ORDER BY Quantity;
  + Insight: There is a strong positive correlation between the **quantity sold** and **total revenue**, as indicated by the high Pearson correlation coefficient. This means that as the quantity of items sold increases, total revenue increases significantly. This confirms the relationship, showing that higher quantities generally correspond to higher revenue, suggesting that selling more units is strongly associated with **increased revenue**.
* Relationship Between Quantity Sold and Total Revenue: We explored how quantity sold correlates with total revenue.
  + Query: SELECT monthname(formatted\_date) AS month\_name, SUM(Total) AS sales\_amount FROM supermarket\_sales GROUP BY month\_name;
  + Insight: The month-wise sales analysis reveals that **January** had the highest sales, with a total revenue of **$116,291.87**, indicating strong performance at the beginning of the year. However, **February** experienced a dip in sales, generating **$97,219.37**, potentially due to seasonal factors or lower demand. Sales picked up again in **March**, with a total of **$109,455.51**, suggesting a recovery in customer spending after the February slump. This pattern may highlight seasonal trends that businesses could leverage for better planning and inventory management.

# Data Visualization using Power BI:

## Data Preparation and Integration:

* **Data Integration:**
  + After the ETL operation done on the dataset in MySQL Power BI was connected to the MySQL database for data integration.
* **Columns Added:**
  + Week\_Number = WEEKNUM('project\_db. supermarket\_sales'[formatted\_date],11).
  + Hours = HOUR('project\_db supermarket\_sales'[Time]).
  + Day\_number = WEEKDAY('project\_db supermarket\_sales'[formatted\_date],1).
  + Day\_name = FORMAT('project\_db supermarket\_sales'[formatted\_date], "dddd").
  + Dayoftheweek = CONCATENATE('project\_db supermarket\_sales'[Day\_number], CONCATENATE(" ", 'project\_db supermarket\_sales'[day\_name])).
* **Columns Deleted:**
  + date : Removed due to mixed formats; formatted\_date used for analysis.
  + Branch: Removed as each city represented a specific branch, thus redundant.
  + Tax: Removed as it had no impact on data analysis (uniform 5% across transactions).
  + Gross\_income: Deleted since a measure for profit was created, calculated as **Revenue - COGs**, making the gross\_income column unnecessary.
* **Measures Created:**
  + Avg Transaction Value = DIVIDE([Revenue], [No of Transactions], 0).
  + COGs = SUM('project\_db supermarket\_sales'[cogs]).
  + No of Transactions = COUNT('project\_db supermarket\_sales'[Invoice\_ID]).
  + Profit =[Revenue]-[COGs].
  + Revenue = SUM('project\_db supermarket\_sales'[Total]).

## Data Preparation and Integration:

* **Card Visuals:**
  + Total Revenue: Displays total revenue.
  + Total Profit: Displays total profit earned.
  + Avg Transaction Value: Represents the average value of the transaction.
  + No of Transactions: Counts the number of transactions.
* **Pie Charts:**
  + Gender Based Revenue Distribution: Illustrates revenue distribution across male and female customers.
  + City Based Revenue Distribution: Shows revenue distribution across 3 different cities- Naypyitaw, Yangon, Mandalay.
  + Payment Mode Based Revenue Distribution: Displays revenue distribution across 3 payment modes- Ewallet, Cash and Credit card.
* **Stacked Bar Chart:**
  + Revenue Product Wise: Represents revenue generated by different product line.
* **Waterfall Chart:**
  + Monthly Profit Flow by Product Line: Displays the flow of profit month over month, broken down by product lines. This chart helps in visualizing how different product lines contribute to monthly profits.
  + Design: Category- Month(formatted\_date), Breakdown- Product\_line, Y axis- Profit.
* **Scatter Plot:**
  + This scatter plot examines the relationship between the number of units sold and the revenue generated. It helps to identify trends or patterns, such as whether higher sales volume correlates with higher revenue.
  + Design: X axis-Quantity/Units Sold, Y axis: Revenue.
* **Heatmap or Matrix:**
  + Revenue Heatmap: Weekday vs Hours: This heatmap helps identify which hours of the day and which days of the week generate the most revenue. Darker or lighter shades represent the intensity of revenue for the respective hour and weekday.
  + Design: Rows- Dayoftheweek, Columns- Hours, Values- Revenue.
* **Stacked Column Chart:**
  + Profit by Gender and Customer Type: This chart compares profits generated by different genders, further segmented by customer type (e.g., Member vs Non-member). It helps to evaluate which groups are more profitable.
  + Design: X axis- Gender, Y axis- Profit, Legends- Customer\_type.
* **Funnel Chart:**
  + No of Transactions Funnel Analysis by Product Line: A funnel chart shows the number of transactions across different product lines. This visual helps analyze which product lines have the highest number of transactions, providing insights into product performance.
  + Design: Category- Product\_line, Values- No of Transactions.
* **Line Chart:**
  + Revenue Trendline & Forecast: The line chart shows the trend of revenue over time, with projections for future performance. It helps in tracking revenue patterns and forecasting future growth.
  + Design: X axis- Week\_number, Y axis- Revenue.

# Insights and Recommendations:

## Insights:

* **Revenue and Profit Overview:**
  + Total revenue stands at \*322.97K, while profit is \*\*15.38K. This indicates a profit margin of around \*\*4.76%.
  + Average transaction value is consistent at \*322.97\* with a total of \*1000 transactions\*. This suggests that revenue per transaction is stable, but increasing transaction volume could boost profits.
* **Product Line Performance:**
  + The top-performing product lines in terms of revenue are \*Food and beverages, followed by \*\*Sports and travel. However, \*\*Health and beauty\* lags behind in both revenue and number of transactions.
  + For maximizing profits, focusing on promoting \*Fashion accessories\* and \*Home and lifestyle\* (which have a better profit flow) might be beneficial.
* **Revenue by City and Payment Mode:**
  + Cities contribute almost equally to revenue, with \*Naypyitaw\* leading slightly at \*110.57K\*.
  + In terms of payment mode, \*Cash\* is the most popular method, bringing in \*112.21K\* (34.74%). This suggests that promoting digital payment options could further boost sales.
  + Consider running location-specific promotions in \*Yangon\* and \*Mandalay\* to balance sales across cities.
* **Gender-based Insights:**
  + Gender distribution shows a slight advantage for \*female customers\* in terms of both profit and revenue. Additionally, \*members\* tend to generate slightly more profit than normal customers.
  + A loyalty or membership program targeted specifically at female shoppers could enhance engagement and drive revenue.
* **Revenue Heatmap – Weekdays and Hours:**
  + Peak revenue times are between *7 PM and 1 PM*, especially on **Tuesday** and **Saturday**, with significant spikes at ***7 PM***.
  + Revenue dips after ***4 PM*** on most days but spikes again around ***7 PM***, indicating a pattern of high evening sales.
  + **Sunday** continues to underperform compared to other days, with no major revenue spikes throughout the day.
  + To improve underperforming time slots, consider offering time-limited promotions during low-traffic hours on **Sunday afternoons** or **early afternoon hours (12 PM to 2 PM)** on weekdays like **Monday** and **Thursday.**
* **Revenue Trendline and Forecast:**
  + The forecast shows that there is no significant up-trend in revenue in upcoming weeks, after reaching a peak earlier..
  + This could be a signal to launch marketing campaigns or promotions to maintain momentum and counteract the decline.
* **Number of Transactions by Product Line:**
  + The product lines with the highest number of transactions are \*Fashion accessories\* and \*Food and beverages, with \*\*178\* and \*174\* transactions respectively. The least number of transactions comes from \*Health and beauty\*, suggesting that targeted marketing might be necessary for that segment.

## Recommendations:

* **Boost Sales in Underperforming Categories**: Increase promotional efforts for Health and beauty products to boost their visibility and appeal. Consider strategies such as targeted ads, special discounts, or bundling with popular items like Fashion accessories to drive transactions in this category.
* **Leverage Peak Revenue Times:** Maximize marketing and promotional activities during peak revenue periods between 7 PM and 1 PM on Tuesdays and Saturdays. This can include special offers, flash sales, or limited-time discounts to capitalize on high customer traffic.
* **Promote Digital Payments:** Given that Cash remains the most popular payment method, implement incentives to encourage the use of digital payment options such as discounts or rewards for using E-wallets or credit cards. This can help increase overall revenue and streamline transactions.
* **Expand City-Specific Promotions:** Run targeted promotions in Yangon and Mandalay to balance revenue distribution across cities. This could involve city-specific discounts, events, or partnerships to boost local engagement.
* **Enhance Customer Loyalty Programs:** Develop and promote a loyalty program aimed at female customers and existing members. Offer exclusive benefits or rewards to drive repeat purchases and increase profit margins.
* **Revamp Sunday Sales Strategy:** Implement time-limited promotions or discounts during the low-traffic hours on Sunday afternoons to improve performance on this traditionally slower day. Similarly, consider targeting early afternoon hours on Monday and Thursday to boost revenue during these periods.
* **Act on Revenue Trends:** Launch marketing campaigns or special promotions to counteract the anticipated decline in revenue. Use insights from the forecast to plan timely interventions that can sustain or boost revenue levels.

## Future work:

* Suggest areas for further analysis or additional questions to explore.
* Future analysis could explore customer satisfaction data to correlate with sales performance and identify areas for improvement.

# References:

* Data Source: Kaggle. (2024). Supermarket Sales Dataset. Retrieved from <https://www.kaggle.com/datasets/aungpyaeap/supermarket-sales>
* Tools Used: MySQL Workbench 8.0; Microsoft Power BI.

# Appendix:

* **Complete SQL Script:**

create database project\_db;

use project\_db;

desc supermarket\_sales;

/\* Check whether the supermarket\_sales dataset has been imported sucessfully or not \*/

select \* from supermarket\_sales;

/\* Modifying the column names where it contains ' ' for easier code writing. Replacing ' ' with '\_' \*/

alter table supermarket\_sales

change `Invoice ID` `Invoice\_ID` text;

alter table supermarket\_sales

change `Customer type` `Customer\_type` text;

alter table supermarket\_sales

change `Product line` `Product\_line` text;

alter table supermarket\_sales

change `gross margin percentage` `gross\_margin\_percentage` double;

alter table supermarket\_sales

change `gross income` `gross\_income` double;

alter table supermarket\_sales

change `Tax 5%` `Tax\_5%` double;

alter table supermarket\_sales

change `Unit price` Unit\_price double;

/\*DATA CLEANING & TRANSFORMATION (ETL)/\*

/\* Deleting the gross\_margin\_percentage column as it is same for every row \*/

alter table supermarket\_sales

drop column gross\_margin\_percentage;

/\* Checking for duplicates and removing if required \*/

select Invoice\_ID, count(Invoice\_ID) from supermarket\_sales

group by Invoice\_ID having count(Invoice\_ID) > 1;

/\* Checking for different date formats in the dataset \*/

select

case

when date like '\_/\_\_/\_\_\_\_' then 'M/DD/YYYY'

when date like '\_\_-\_\_-\_\_\_\_' then 'MM/DD/YYYY'

else 'unknown format'

end as date\_format,

count(\*) as format\_count

from supermarket\_sales group by date\_format;

/\* Adding a new column for date as formatted\_date which will hold evry date in YYYY-MM-DD format \*/

alter table supermarket\_sales

add column formatted\_date date;

/\* Updating the formatted\_date column \*/

update supermarket\_sales

set formatted\_date =

case

when date like '\_/\_\_/\_\_\_\_' then

STR\_TO\_DATE(date, '%m/%d/%Y')

when date like '\_\_-\_\_-\_\_\_\_' then

STR\_TO\_DATE(date, '%m-%d-%Y')

else

NULL -- or use a default date if necessary

end;

/\* converting categorical value to numeric value for analysis \*/

/\* Adding a new column as customer\_type\_code \*/

alter table supermarket\_sales

add customer\_type\_code int;

/\* EXPLORATORY DATA ANALYSIS \*/

/\* Find out total sales revenue by product line \*/

select product\_line, sum(total) from supermarket\_sales

group by product\_line;

/\* Find out the avg rating of products by product line \*/

select product\_line, avg(Rating) from supermarket\_sales

group by Product\_line;

/\* What is the sales distribution across different customer types (Member vs Normal) \*/

select customer\_type, count(\*) as num\_sales, sum(Total) as total\_rev,

(sum(Total)/(select sum(Total) from supermarket\_sales) \* 100) as pct\_total\_revenue

from supermarket\_sales

group by customer\_type;

/\* What is the most popular payment method \*/

with cte\_payment as

(

select Payment, count(Payment) as num\_transaction,

dense\_rank() over(order by count(Payment) desc) as ranking

from supermarket\_sales

group by Payment

)

select \* from cte\_payment where ranking = 1;

/\* How does revenue vary by gender and customer type \*/

select Customer\_type, Gender, sum(Total)

from supermarket\_sales

group by Customer\_type, Gender;

/\* What are the top 3 busiest days of the week in terms of total sales \*/

select dayname(formatted\_date) as week\_day, sum(Total) as total\_rev

from supermarket\_sales

group by dayname(formatted\_date)

order by sum(Total) desc limit 3;

/\* What is the average unit price by product line \*/

select Product\_line, avg(Unit\_price) from supermarket\_sales

group by Product\_line;

/\* What is the relationship between quantity sold and total revenue \*/

select Quantity, sum(Total) as Total\_revenue

from supermarket\_sales group by Quantity order by Quantity;

/\* Find out the total monthly sales amounnt \*/

select monthname(formatted\_date) as month\_name, sum(Total) as total\_revenue

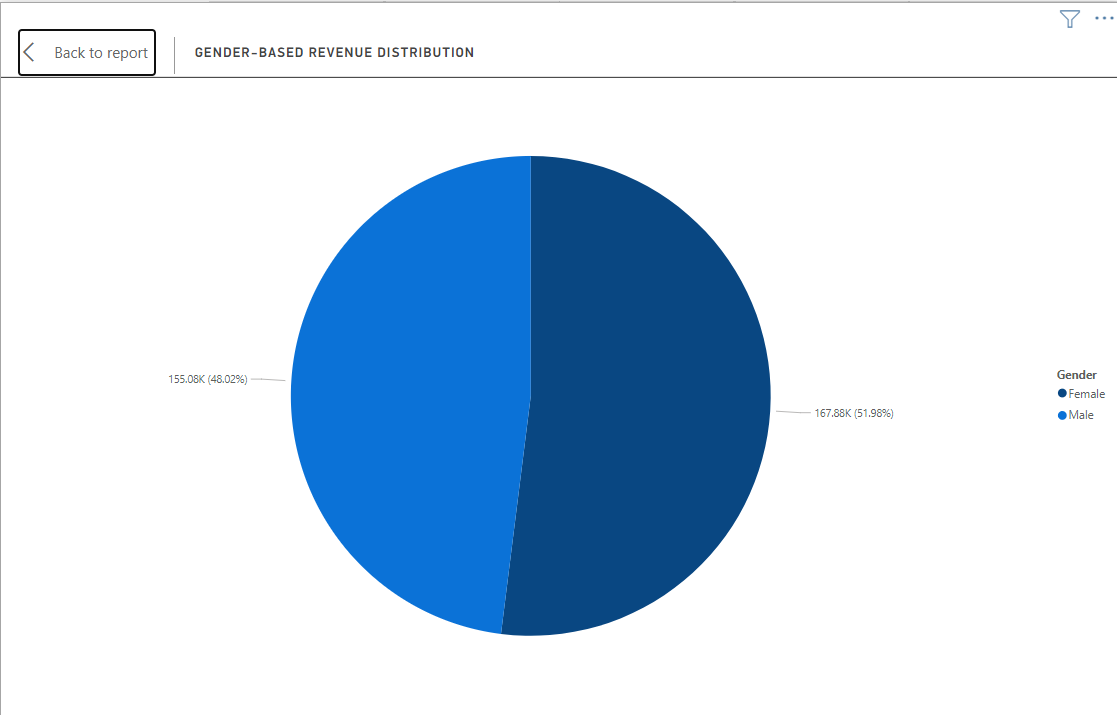
from supermarket\_sales group by month\_name;

* **Charts of Power BI:**

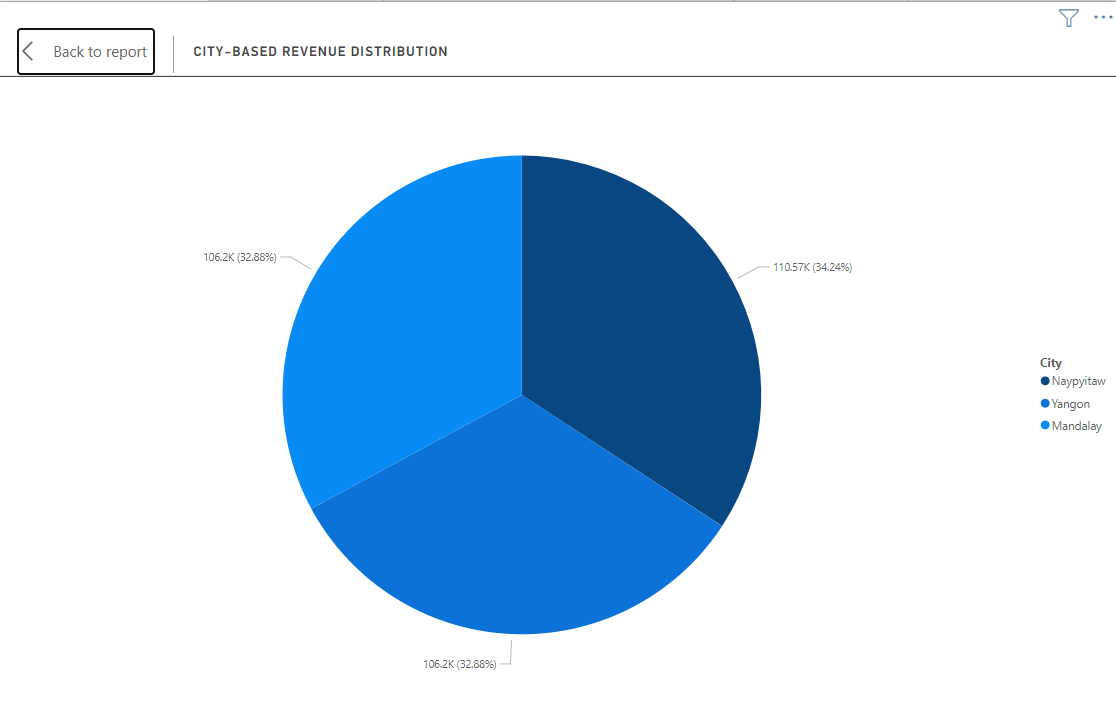
**Card Visuals:**

****

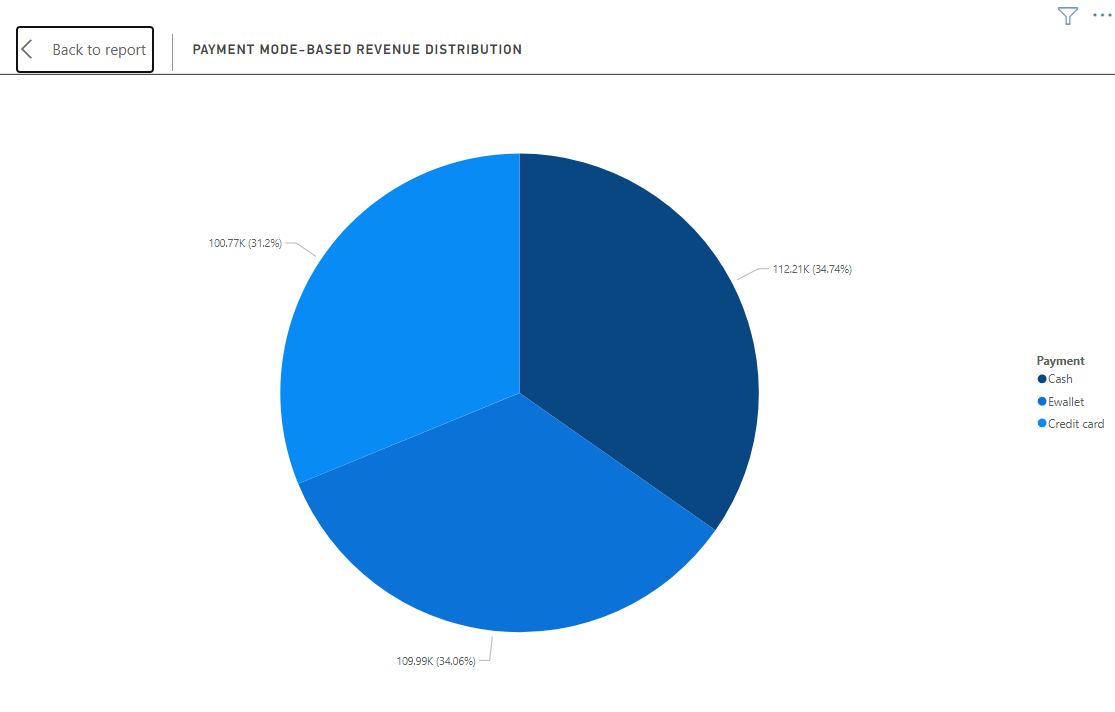
**Pie Chart - Gender Based Revenue Distribution:**

****

**Pie Chart – City Based Revenue Distribution:**

****

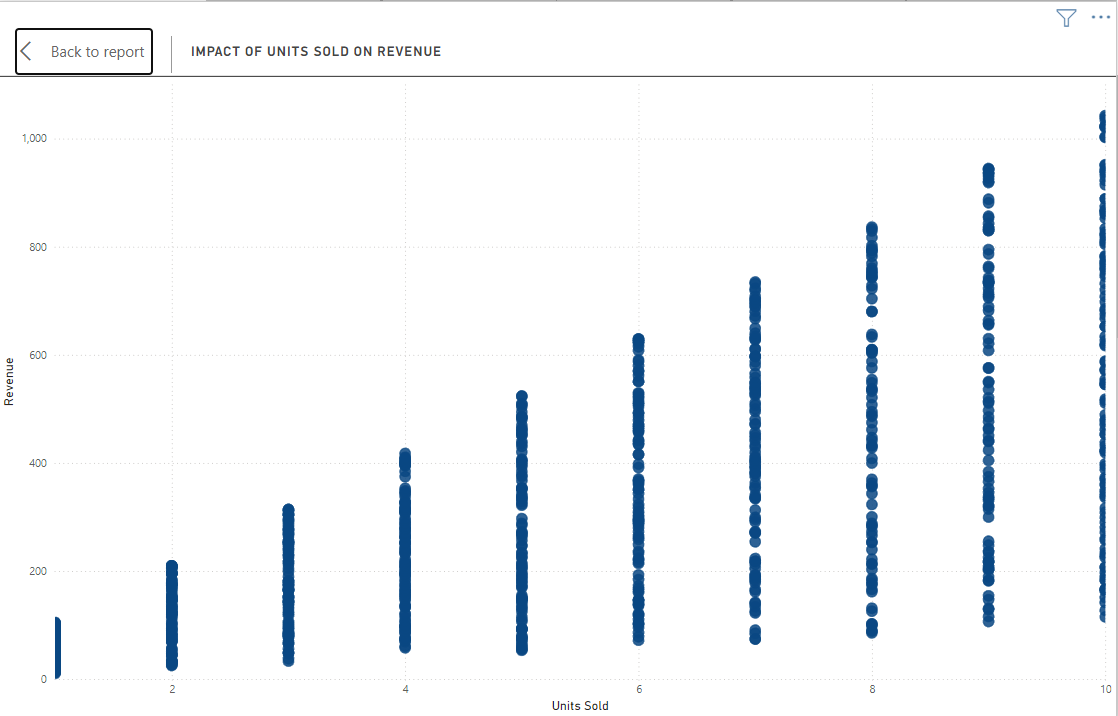
**Pie Chart - Payment Mode Based Revenue Distribution:**

****

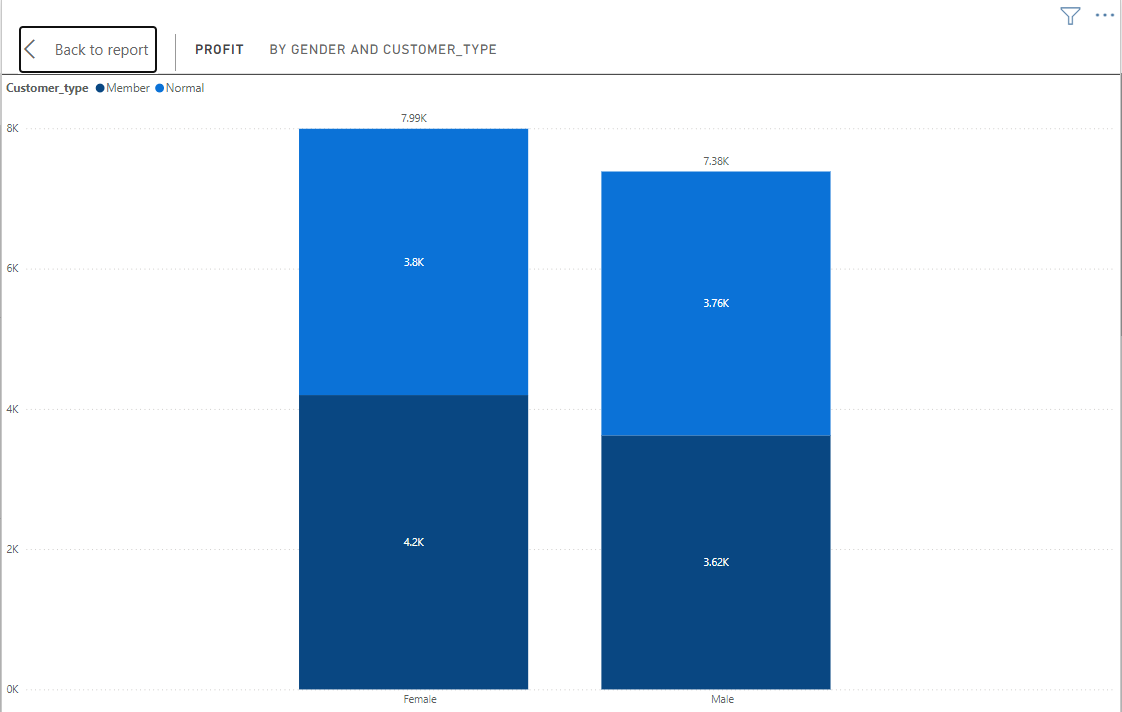
**Stacked Bar Chart – Revenue Product Wise**

****

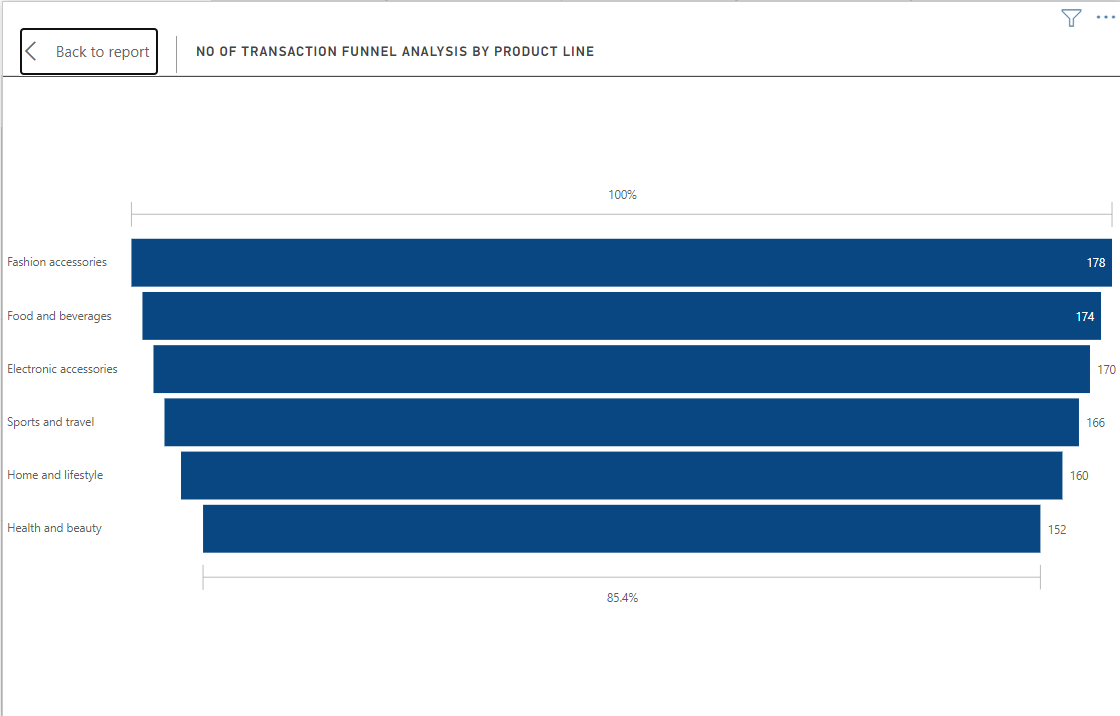
**Scatter plot – Impacts of Units sold on Revenue**

****

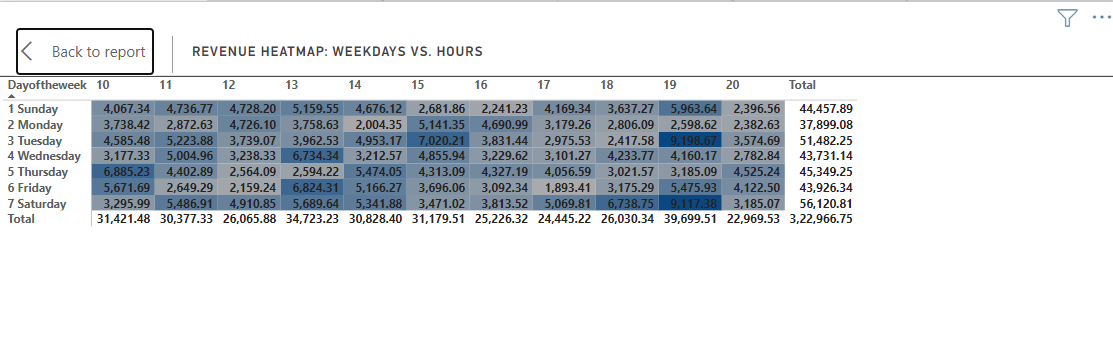
**Stacked Column Chart – Profit by Gender and Customer Type**

****

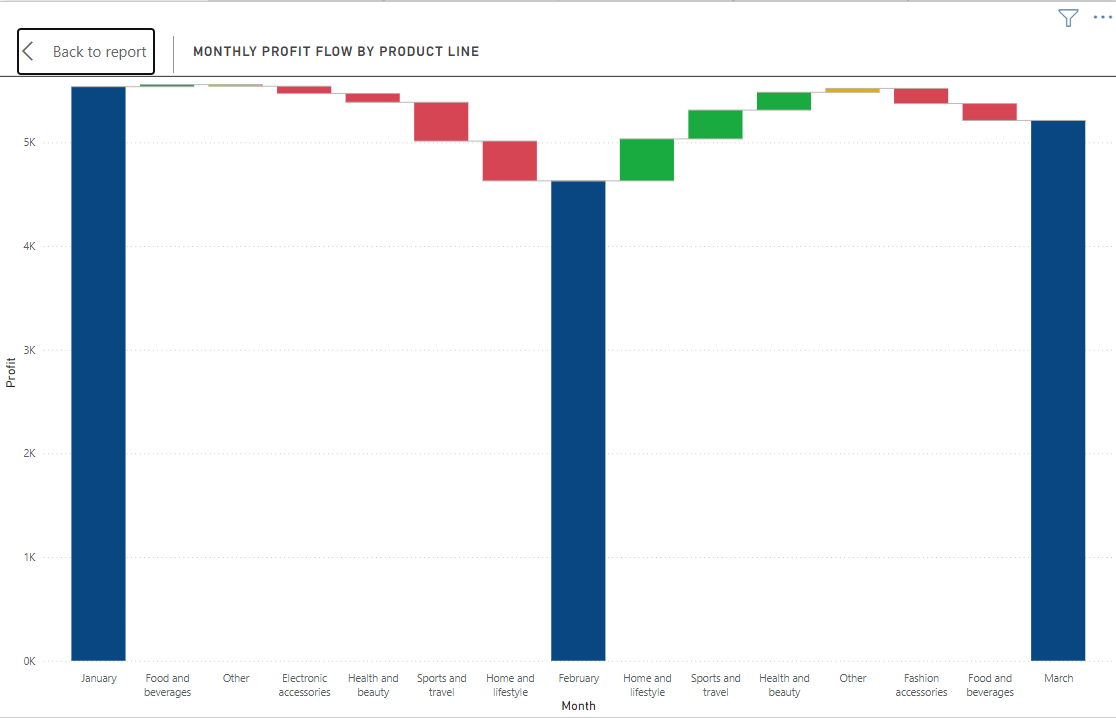
**Funnel Chart – No of Transaction Funnel Analysis by Product Line**

****

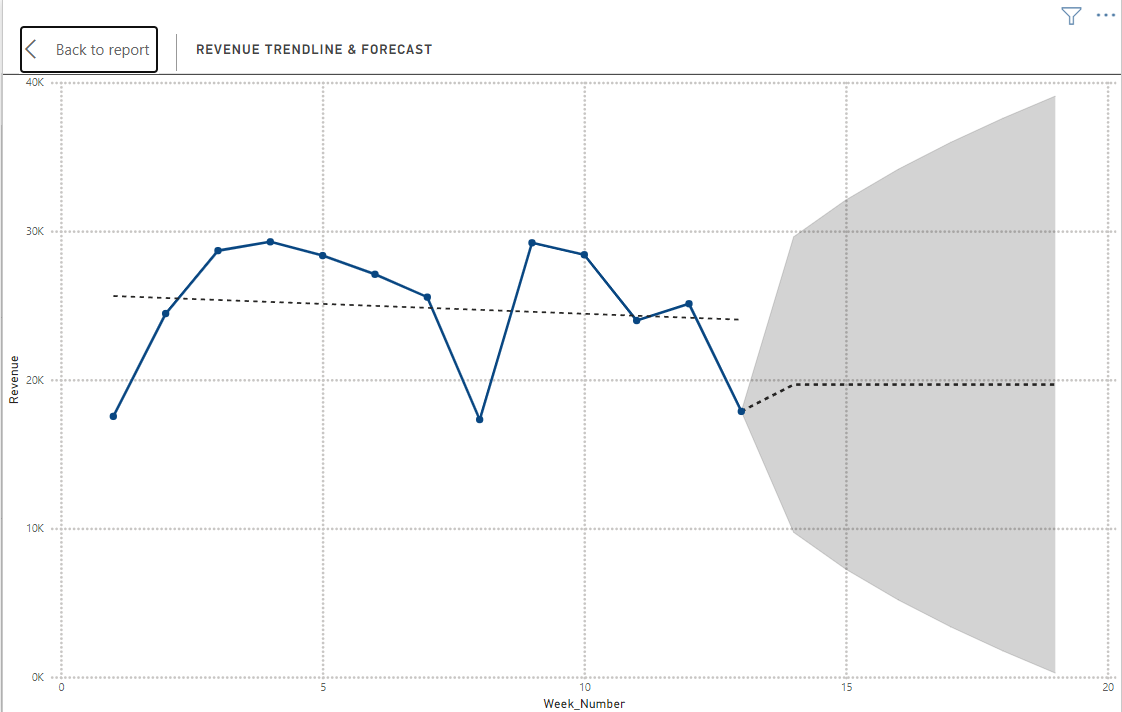
**Heatmap or Matrix – Revenue Heatmap: Weekdays vs Hours**

****

**Waterfall Chart – Monthly Profit Flow by Product Line**

****

**Line Chart – Revenue Trendline & Forecast**

****