Walmart

Walmart Sales Data Analysis

About

This project aims to analyse Walmart sales data to identify the top-performing branches and products, examine sales trends across various products, and understand customer behaviour. The goal is to explore ways to improve and optimize sales strategies for better performance.

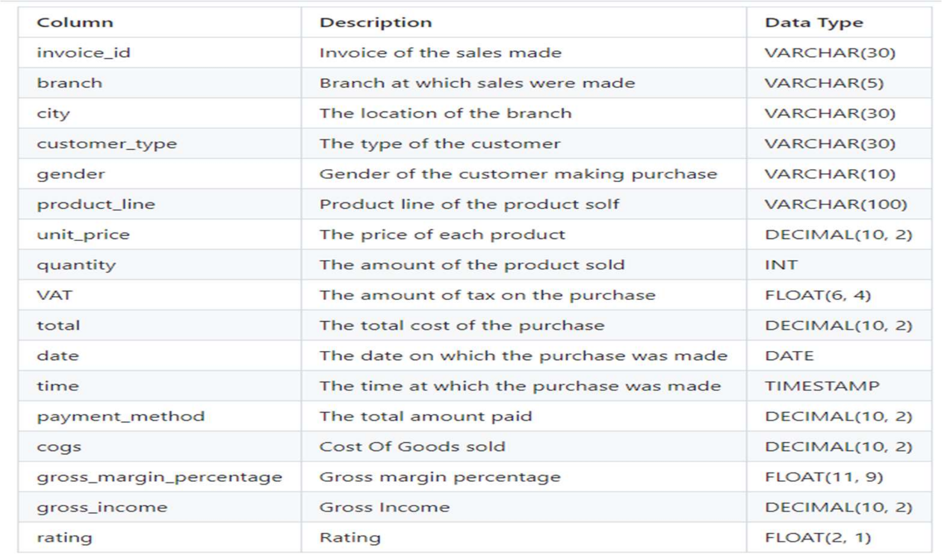
In this recruiting competition, job seekers are provided with historical sales data from 45 Walmart stores across various regions. Each store has multiple departments, and participants are tasked with forecasting sales for each department in every store. Adding to the complexity, the dataset includes selected holiday markdown events, which are known to influence sales. However, predicting which departments will be affected and to what extent poses a significant challenge.

Purposes of the Project

The primary goal of this project is to analyse Walmart's sales data to gain insights into the various factors influencing sales across different branches.

About Data

The dataset used in this project was sourced from the Kaggle Walmart Sales Forecasting Competition. It includes sales transactions from three Walmart branches, located in Mandalay, Yangon, and Naypyitaw. The dataset comprises 17 columns and 1,000 rows.



Column	Description	Data Type
invoice_id	Invoice of the sales made	VARCHAR(30)
branch	Branch at which sales were made	VARCHAR(5)
city	The location of the branch	VARCHAR(30)
customer_type	The type of the customer	VARCHAR(30)
gender	Gender of the customer making purchase	VARCHAR(10)
product_line	Product line of the product solf	VARCHAR(100)
unit_price	The price of each product	DECIMAL(10, 2)
quantity	The amount of the product sold	INT
VAT	The amount of tax on the purchase	FLOAT(6, 4)
total	The total cost of the purchase	DECIMAL(10, 2)
date	The date on which the purchase was made	DATE
time	The time at which the purchase was made	TIMESTAMP
payment_method	The total amount paid	DECIMAL(10, 2)
cogs	Cost Of Goods sold	DECIMAL(10, 2)
gross_margin_percentage	Gross margin percentage	FLOAT(11, 9)
gross_income	Gross Income	DECIMAL(10, 2)
rating	Rating	FLOAT(2, 1)

Analysis List

1. Product Analysis

Analyse the data to gain insights into the various product lines, identify the top-performing product lines, and determine which product lines require improvement

2. Sales Analysis

This analysis aims to examine sales trends across different products. The results will help evaluate the effectiveness of current sales strategies and identify necessary adjustments to boost sales further.

3. Customer Analysis

This analysis seeks to identify various customer segments, their purchasing patterns, and the profitability associated with each segment.

Approach Used

1. **Data Wrangling**: This initial step involves inspecting the data to identify any NULL or missing values, followed by employing data replacement methods to address these gaps.

   1. Establish a Database: Create the necessary tables and populate them with data.

   2. Identify Columns with NULL Values: Select the columns that may contain NULL values. However, since all fields were defined with NOT NULL constraints during table creation, there are no NULL values present in our database.

   3. Data Integrity: As a result of the NOT NULL constraints, any potential NULL values have been effectively filtered out.

2. **Feature Engineering:** This process allows us to create new columns by transforming and combining existing ones.

   1. Introduce a new column called time_of_day to categorize sales into Morning, Afternoon, and Evening. This will provide insights into which part of the day sees the highest sales volume.
   2. Create a new column named day_name that captures the day of the week on which each transaction occurred (Mon, Tue, Wed, Thu, Fri). This will help identify which day of the week each branch experiences the highest activity.
   3. Add a column titled month_name that indicates the month in which each transaction took place (Jan, Feb, Mar). This will assist in determining which month generates the most sales and profit.

3. **Exploratory Data Analysis (EDA):** The exploratory data analysis is conducted to address the outlined questions and objectives of this project.
4. **Conclusion:**

Business Questions to Answer

Generic Question

1. How many unique cities does the data have?

   This query calculates the count of distinct city names present in the sales **table**. The result indicates that there are **exactly 3 unique cities** in the dataset. This insight helps in understanding the geographical spread of the sales data and can be crucial for region-specific analysis.

2. In which city is each branch?

   This query retrieves unique combinations of city and branch from the sales table, ensuring that each branch is listed only once along with its corresponding city. The results are as follows:

- Branch A is located in Yangon.
- Branch B is located in Mandalay.
- Branch C is located in Naypyitaw.

Product

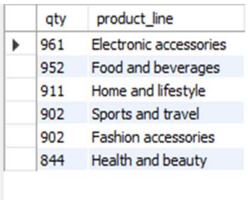
1. How many unique product lines does the data have?

   The query results indicate that there are **6 unique product lines** in the dataset. This diversity in product lines could be beneficial for analyzing sales performance across different categories.

2. What is the most common payment method?

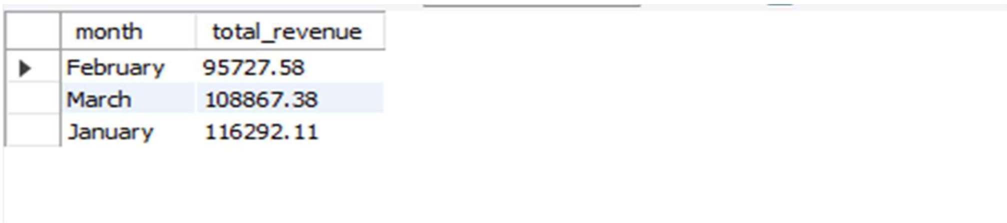
   After analyzing the sales data, the most common payment method is found to be Cash, with a **total of 344 transactions.** This indicates that cash remains a preferred choice for transactions within the dataset examined.

3. What is the most selling product line?



qty	product_line
961	Electronic accessories
952	Food and beverages
911	Home and lifestyle
902	Sports and travel
902	Fashion accessories
844	Health and beauty

4. What is the total revenue by month?



	month	total_revenue
▶	February	95727.58
	March	108867.38
	January	116292.11

The results indicate that **January** had the **highest revenue**, **followed by March and then February**. This information can be useful for understanding seasonal trends and planning future sales strategies effectively.

5. What month had the largest COGS?

January had the **highest COGS**, suggesting it was the month with the most significant amount of goods sold or possibly higher costs associated with those goods.

6. What product line had the largest revenue?

The product line with the **largest revenue** is **"Food and Beverages,"** generating a **total of $56,144.96**. This indicates a strong consumer preference for food and beverage products, suggesting that this category might be the most crucial in terms of sales focus and inventory management.

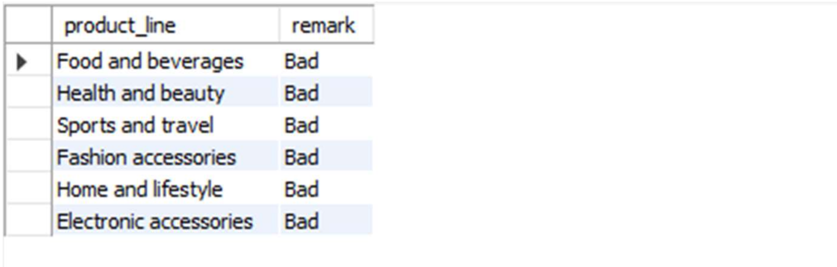
7. What is the city with the largest revenue?

Naypyitaw emerged as the city with the largest revenue, generating a **total of $110,490.93**. This indicates that Naypyitaw might have the highest sales activity or more expensive items sold compared to the other cities.

8. What product line had the largest VAT?

The **"Home and Lifestyle"** product line has the highest **average VAT at 16.0303**, indicating it generally has higher tax rates compared to other categories.

9. Fetch each product line and add a column to those product line showing "Good", "Bad". Good if its greater than average sales

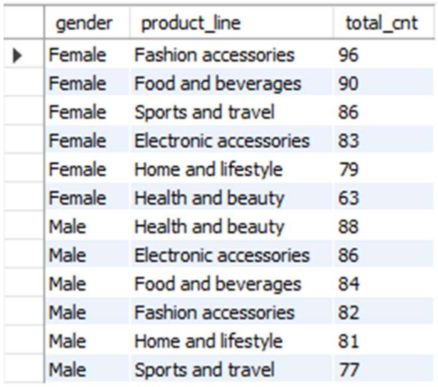


	product_line	remark
▶	Food and beverages	Bad
	Health and beauty	Bad
	Sports and travel	Bad
	Fashion accessories	Bad
	Home and lifestyle	Bad
	Electronic accessories	Bad

10. Which branch sold more products than average product sold?

    All branches exceeded the average quantity sold, demonstrating robust performance across the board. This indicates that each branch is contributing positively to the overall sales volume, with **Branch A leading in quantity sold**.

11. What is the most common product line by gender?



gender	product_line	total_cnt
Female	Fashion accessories	96
Female	Food and beverages	90
Female	Sports and travel	86
Female	Electronic accessories	83
Female	Home and lifestyle	79
Female	Health and beauty	63
Male	Health and beauty	88
Male	Electronic accessories	86
Male	Food and beverages	84
Male	Fashion accessories	82
Male	Home and lifestyle	81
Male	Sports and travel	77

    These insights highlight distinct preferences between genders, with females favouring fashion accessories most and males leading purchases in electronic accessories.

12. What is the average rating of each product line?

    The **"Food and Beverages" category** leads with the **highest average rating** of **7.11**, indicating strong customer satisfaction. In contrast, **"Home and Lifestyle"** receives the **lowest average rating**, which may signal areas for improvement or a focus on further customer feedback collection.

Sales

1. Number of sales made in each time of the day per weekday?

    Evenings are generally the peak sales period across most weekdays, suggesting that promotions or increased shopping activity in the evening could be beneficial.

    Afternoon sales are consistently high but become the **highest on Wednesday and Friday**, which may indicate different shopping habits on these days.

    Morning sales are the lowest across all days, with slight increases on Tuesdays and Thursdays. Morning promotions or special offers could potentially boost sales during these lower periods.

2. Which of the customer types brings the most revenue?

   Members contribute more to the total revenue compared to non-member customers. This suggests that members, possibly due to loyalty programs or other incentives, tend to spend more, highlighting the effectiveness of membership programs in driving higher sales volumes.

3. Which city has the largest tax percent/ VAT (**Value Added Tax**)?

Naypyitaw	16.09
Mandalay	15.13
Yangon	14.87

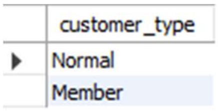
4. Which customer type pays the most in VAT?

Normal	15.09805040
Member	15.61457214

   Members contribute more to the total revenue compared to non-member customers. This suggests that members, possibly due to loyalty programs or other incentives, tend to spend more, highlighting the effectiveness of membership programs in driving higher sales volumes.

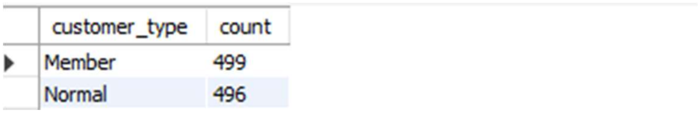
Customer

1. How many unique customer types does the data have?



customer_type
Normal
Member

2. How many unique payment methods does the data have?



customer_type	count
Member	499
Normal	496

Revenue and Profit Calculations

- $ COGS = unitsPrice * quantity $
- $ VAT = 5% * COGS $
- VAT is added to the COGS and this is what is billed to the customer.
- $ total(gross_sales) = VAT + COGS $
- $ grossProfit(grossIncome) = total(gross_sales) - COGS $

Gross Margin is gross profit expressed in percentage of the total(gross profit/revenue)

$ \text{Gross Margin} = \frac{\text{gross income}}{\text{total revenue}} $

Example with the first row in our DB:

Data given:

- $ \text{Unite Price} = 45.79 $
- $ \text{Quantity} = 7 $

$ COGS = 45.79 * 7 = 320.53 $

$ \text{VAT} = 5% * COGS\= 5% 320.53 = 16.0265 $

$ total = VAT + COGS\= 16.0265 + 320.53 = 336.5565

$ \text{Gross Margin Percentage} = \frac{\text{gross income}}{\text{total revenue}}\=\frac{16.0265}{336.5565} = 0.047619\\approx 4.7619% $

Recommendations:

Based on the data analysis and the SQL queries, here are some recommendations for Walmart's sales strategy:

- **Targeted Marketing Strategies:** Develop targeted marketing campaigns for gender-specific products. Increase marketing efforts for fashion accessories aimed at females and electronic accessories aimed at males.

Enhance Food and Beverage Experience: Given the high customer satisfaction in the Food and Beverages category, consider expanding these offerings or promoting this line more aggressively to capitalize on its popularity.

- **Promotional Timing:** Implement targeted promotions during peak sales times, especially in the evenings and on weekends, to maximize customer turnout and sales. Consider morning promotions or special offers to boost sales during these lower periods, especially on Tuesdays and Thursdays when there are slight increases in sales.
- **Customer Loyalty Programs:** Since there is a good balance of members and non-members, enhance the loyalty program to convert more normal customers into members. Offer exclusive

deals or loyalty points for purchases during off-peak hours to balance the sales distribution throughout the day.

- **Inventory and Pricing Strategy:** Reassess the pricing and inventory strategy for the "Home and Lifestyle" category due to its lower customer ratings. Collect more detailed customer feedback to understand the dissatisfaction and adjust product offerings accordingly.

Action

Its incredible how powerful SQL can be! In this project, I challenged myself to utilize only simple functions, operators, and clauses to get my answers:

Aggregate Functions: SUM, AVG, MAX, MIN, COUNT

Operators: AND, NOT, OR, NULL, LIKE, DESC

Clauses: WHERE, GROUP BY, ORDER BY, LIMIT, AS, WITH

Thank you so much for reading my project. Suggestions are welcome. If you have any questions, feel free to comment below connect with me on LinkedIn or check out my portfolio to see my other projects.

Stay tuned for my next SQL project, where I'll explore more advanced SQL capabilities to examine data.