**Walmart Sales Data Analysis**

**Project Purpose**:

This project endeavors to delve into the Walmart Sales dataset with the overarching objective of comprehending the performance trends of different branches and products. The project entails analyzing the sales patterns of diverse products and decoding the intricate behavioral patterns exhibited by customers. The core aspiration revolves around identifying opportunities for enhancing and optimizing sales strategies.

**Key Functions and Calculations:**

Database Setup:

To lay the foundation for the analysis, a MySQL database named "walmart" is created. The data is then organized into a "sales" table, capturing critical attributes such as branch, customer type, gender, product line, and more.

Exploratory Data Analysis (EDA)

Branch Performance Analysis:

Total revenue, cost of goods sold, transaction count, and item quantity sold are evaluated for each branch, offering insights into the financial performance of different branches.

Product Analysis:

Various aspects related to product lines are analyzed, including the count of unique product lines, the most common payment method, best-selling product lines, and total revenue, cost of goods sold, and profit by month.

Sales Analysis:

The queries delve into sales dynamics, exploring the most popular day and time combinations for transactions, customer types' revenue contribution, cities with the highest Value Added Tax (VAT), and customer types' VAT payments.

Customer Analysis:

Queries are crafted to analyze customer-centric aspects, including customer type distribution and corresponding sales, the distribution of payment methods, gender distribution, gender distribution per branch, peak rating times, and the day of the week with the best average ratings.

Tableau Visualization -> [Clic here for the Dashboard](https://public.tableau.com/app/profile/fernando.molina2632/viz/WalmartSalesAnalysis_16918952370360/WalmartSalesAnalysis) (or copy and paste the address: *https://public.tableau.com/app/profile/fernando.molina2632/viz/WalmartSalesAnalysis\_16918952370360/WalmartSalesAnalysis* )

The results of this comprehensive analysis were exported into Tableau to create a dynamic visualization, an interactive dashboard. This dashboard effectively translates the extracted insights into visually engaging representations. It now stands as tool, allowing stakeholders to gain a holistic view of the analyzed data.

Through the adept execution of functions and calculations in MySQL, combined with Tableau's strengths, this project has successfully attained comprehensive insights into Walmart's sales data. This achievement significantly contributes to informed decision-making processes aimed at optimizing sales strategies

**About the Data**

The dataset was obtained from the [Kaggle Walmart Sales Forecasting Competition] (https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting). This dataset contains sales transactions from three different branches of Walmart, respectively located in Mandalay, Yangon and Naypyitaw. The data contains 17 columns and 1000 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 | VARCHAR(10) |
| product\_line | Product line of the product sold | VARCHAR(100) |
| unit\_price | The price of each product | DECIMAL(10, 2) |
| quantity | The amount of the product sold | INTEGER |
| 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 | Time at which the purchase was made | TIMESTAMP |
| payment\_method | The type of payment used | VARCHAR(30) |
| 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) |

**Data Preparation Stage**

1. Data Wrangling: Initiating the process, this initial step involves meticulously inspecting the data to identify any occurrences of **NULL** or missing values. In cases of such discrepancies, appropriate data replacement techniques are employed to address and rectify these gaps..

* Create a Database in MySQL
* Create a Table with the columns defined in the data file
* Identify columns containing null values. The database is designed with a meticulous approach; during table creation, each field is assigned a NOT NULL constraint, effectively eliminating the presence of null values.

2. Feature Data engineering: This phase will help the analysis by generating some new columns from the existing ones.

* Introduce an additional column named "time\_of\_day" designed to provide a perspective on sales distribution across *Morning*, *Afternoon*, and *Evening* periods. This addition facilitates the exploration of the peak sales periods within a day.
* Introduce an additional column named “day\_name” designed to house the extracted weekdays corresponding to each transaction occurrence (Mon, Tue, Wed, Thur, Fri, Sat, Sun). This addition facilitates the analysis of weekly trends, aiding in identifying the busiest days for each branch.
* Add a new column named `month\_name` that contains the extracted months of the year on which the given transaction took place (Jan, Feb, Mar). Help determine which month of the year has the most sales and profit.

**Exploratory Data Analysis**

Exploratory data analysis serves as the cornerstone for addressing the objectives and queries outlined in this project.

* **Product Analysis:**

Thoroughly examine the dataset to gain insights into the diverse product lines. Identify the top-performing product lines as well as those that exhibit room for improvement.

* **Sales Analysis:**

The focal point of this analysis is to unravel the sales trends of products. The findings will serve as a yardstick to evaluate the efficacy of various sales strategies employed by the business and to identify necessary adjustments that can be made to enhance sales.

* **Customer Analysis:**

This segment of analysis is aimed at unveiling distinct customer segments, discerning their purchasing trends, and assessing the profitability associated with each customer segment.Exploratory data analysis is done to answer the listed questions and aims of this project.

**Conclusion:**

**Exploratory Data Analysis**

***Preliminary Exploration***

Conducting preliminary data analysis involves an assessment of the sales data across the branches and their respective locations. This evaluation includes aggregating key metrics such as total revenue, total cost of goods sold, transaction count, and the quantity of items sold for each individual branch.

* There are 3 branches in the dataset: A, B, C and they are located in the cities of Yangon, Mandalay and Naypyitaw, respectively. The branch with the most sales is Branch C with a total of $110,490.78 total revenue.

***Product Analysis***

1. How many unique product lines does the data have?
   * 6 unique product lines
2. What are the payment methods used and what is transaction count for each one?
   * There are 3 payment methods: Cash, e-Wallet and Credit card and the most used method is Cash with a total of 344 transactions, then e-wallet with 342 transactions and credit cards with 309 transactions
3. What is the most selling product line?
   * The product line with most sales is the Fashion Accessories with 174 transactions.
4. What is the total revenue, cost of good solds and Profit by Month?
   * The data contains information of the months of January, February and March.

|  |  |  |  |
| --- | --- | --- | --- |
| month | total\_revenue | total\_cost | profit |
| January | $116,291.87 | $110,754.16 | $5,537.71 |
| March | $108,867.15 | $103,683.00 | $5,184.15 |
| February | $95,727.38 | $91,168.93 | $4,558.45 |

1. What product line had the largest revenue?
   * Food and Beverages with a total revenue of $56,144.84
2. What product line had the largest VAT?
   * Home and Lifestyle with an average VAT of 16.03%
3. Retrieve every product line and introduce an additional column for each line indicating its performance status, denoted as "Good" or "Bad". A product line is marked as "Good" if its sales surpass the average sales.
   * 
4. What is the most common product line by gender?
   * Fashion and accessories shows the highest preferences for female compared to male.
5. What is the average rating of each product line?
   * 

***Sales Analysis***

1. What is the day with most products sold made in each time of the day?
   * The day with the highest product sales happens on Fridays, the most products were sold with 319 items in the afternoon, 296 items in the evening, and 140 items in the morning.
2. Among the different customer types, which one generates the most revenue?
   * The "Members" customer type contributes the highest share of sales revenue, accounting for 50.99% of total revenue, with only a slight difference from regular customers, who bring in 49.01% of the revenue.
3. Which city has the largest tax percent/ VAT (Value Added Tax) rate?
   * Among the cities, Naypyitaw holds the distinction of having the largest Value Added Tax (VAT) rate, with an average of 16.09%.
4. Which customer type pays the highest amount in VAT?
   * The “Members” customer type tends to pay a slightly higher average Value Added Tax (VAT) of 15.61%, while regular customers pay an average of 15.10% in VAT.

***Customer Analysis***

1. How many unique customer types and what is the sales distribution?
   * 
2. How many unique payment methods and what is the sales distribution?
   * 
3. What is the predominant gender among the customers?
   * Among the customers, there are 497 females, contributing to a total revenue of $166.4 thousand. Additionally, there are 498 male customers generating a total revenue of $154.5 thousand.
4. What is the gender distribution per branch?
   * 
5. During which time of the day do customers provide the most ratings for each branch?
   * For Branch A, the highest average rating is observed during the afternoons, reaching 7.19. In the mornings, the average rating is 7.01, and in the evenings, it's 6.87.
6. Which day of the week receives the highest average ratings?
   * Mondays stand out with the best average ratings, achieving a score of 7.13.

**ANALYSIS CODE - MYSQL**

**CREATE** DATABASE **IF** NOT **EXISTS** walmart;

**CREATE** **TABLE** **IF** NOT **EXISTS** sales (

invoice\_id VARCHAR(30) not NULL **PRIMARY** **KEY**,

branch VARCHAR(5) NOT NULL,

city VARCHAR (30) NOT NULL,

customer\_type VARCHAR(30) NOT NULL,

gender VARCHAR(10) NOT NULL,

product\_line VARCHAR(100) NOT NULL,

unit\_price DECIMAL(10,2) NOT NULL,

quantity INT NOT NULL,

vat FLOAT(6,4) NOT NULL,

total DECIMAL(12,4) NOT NULL,

date DATETIME NOT NULL,

time TIME NOT NULL,

payment\_method VARCHAR(15) NOT NULL,

cogs DECIMAL(10,2) NOT NULL,

gross\_margin\_pct FLOAT(11,9) NOT NULL,

gross\_income DECIMAL(12,4) NOT NULL,

rating FLOAT(2,1)

);

------------------------------------------------------------------------

------------------- Feature Engineering -------------------------------------------------------------------------------------------------------

------------------ Assign time\_of\_day to sales table ---------------

**SELECT**

time,

(**CASE**

**WHEN** time **BETWEEN** "00:00:00" AND "12:00:00" **THEN** "Morning"

**WHEN** time **BETWEEN** "12:01:00" AND "16:00:00" **THEN** "Afternoon"

**ELSE** "Evening"

**END**

) **AS** time\_of\_day

**FROM** sales;

**ALTER** **TABLE** sales **ADD** **COLUMN** time\_of\_day VARCHAR(20);

**UPDATE** sales

**SET** time\_of\_day = (**CASE**

**WHEN** time **BETWEEN** "00:00:00" AND "12:00:00" **THEN** "Morning"

**WHEN** time **BETWEEN** "12:01:00" AND "16:00:00" **THEN** "Afternoon"

**ELSE** "Evening"

**END**

)

------------------ Assign day\_name to sales table ------------------

**SELECT**

date,

DAYNAME(date) **AS** day\_name

**FROM** sales;

**ALTER** **TABLE** sales **ADD** **COLUMN** day\_name VARCHAR(15);

**UPDATE** sales

**SET** day\_name = DAYNAME(date);

------------------ Assign month\_name to sales table --------------

**SELECT**

date,

MONTHNAME(date) **AS** month\_name

**FROM** sales;

**ALTER** **TABLE** sales **ADD** **COLUMN** month\_name VARCHAR(15);

**UPDATE** sales

**SET** month\_name = MONTHNAME(date);

-----------------------------------------------------------

------------- Exploratory Data Analysis -------------------

-----------------------------------------------------------

-- -- First evaluate the sales data with the objective of assessing financial performance for the different branches: total revenue,total cost of goods sold, transaction number and quantity sold.

**SELECT**

branch,

city,

ROUND(SUM(total),2) **AS** total\_revenue,

ROUND(SUM(cogs),2) **AS** total\_cost,

COUNT(invoice\_id) **AS** transaction\_number,

SUM(quantity) **AS** items\_quantity\_sold

**FROM** sales

**GROUP** **BY** city, branch

**ORDER** **BY** total\_revenue **DESC**;

-----------------------------------------------------------

------------------- PRODUCT ANALYSIS ---------------------

-----------------------------------------------------------

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

**SELECT**

COUNT(**DISTINCT** product\_line)

**FROM** sales;

-- 2. What is the most common payment method? ----

**SELECT**

payment\_method,

COUNT(payment\_method) **as** transactions\_per\_method

**FROM** sales

**GROUP** **BY** payment\_method

**ORDER** **BY** transactions\_per\_method **DESC**;

-- 3. What is the most selling product line? ----

**SELECT**

product\_line,

COUNT(product\_line) **as** count

**FROM** sales

**GROUP** **BY** product\_line

**ORDER** **BY** count **DESC**;

-- 4. What is the total revenue, cost of good solds and Profit by Month----

**SELECT**

month\_name **AS** **month**,

ROUND(SUM(total),2) **AS** total\_revenue,

ROUND(SUM(cogs),2) **AS** total\_cost,

ROUND(SUM(total),2) - ROUND(SUM(cogs),2) **AS** profit

**FROM** sales

**GROUP** **BY** month\_name

**ORDER** **BY** Total\_revenue **DESC**;

-- 5. What product line had the largest revenue? ----

**SELECT**

product\_line,

ROUND(SUM(total),2) **AS** total\_revenue

**FROM** sales

**GROUP** **BY** product\_line

**ORDER** **BY** total\_revenue **DESC**;

-- 6. What product line had the largest VAT?

**SELECT**

product\_line,

ROUND(AVG(vat),2) **AS** avg\_tax

**FROM** sales

**GROUP** **BY** product\_line

**ORDER** **BY** avg\_tax **DESC**;

-- 7. Retrieve every product line and introduce an additional column for each line indicating its performance status,

-- denoted as "Good" or "Bad". A product line is marked as "Good" if its sales surpass the average sales.

**SELECT**

product.\*,

**CASE**

**WHEN** product.sales > avg\_sales.average **THEN** 'Good'

**ELSE** 'Bad'

**END** **AS** sales\_evaluation

**FROM** (

**SELECT**

product\_line,

SUM(total) **AS** sales

**FROM** sales

**GROUP** **BY** product\_line

) product

**JOIN** (

**SELECT**

AVG(sales) **AS** average

**FROM** (

**SELECT**

product\_line,

SUM(total) **AS** sales

**FROM** sales

**GROUP** **BY** product\_line

) avg\_sales

) avg\_sales

**ORDER** **BY** product.product\_line;

-- 8. Which branch sold more products than average product sold?

**SELECT**

branch,

SUM(quantity) **AS** quantity

**FROM** sales

**GROUP** **BY** branch

**HAVING** SUM(quantity) > (**SELECT** AVG(quantity) **FROM** sales);

-- 9. What is the most common product line by gender?

**SELECT**

gender,

product\_line,

COUNT(gender) **AS** total\_cnt

**FROM** sales

**GROUP** **BY** gender, product\_line

**ORDER** **BY** total\_cnt **DESC**;

-- 10. What is the average rating of each product line?

**SELECT**

product\_line,

ROUND(AVG(rating),2) **AS** Avg\_rating

**FROM** sales

**GROUP** **BY** product\_line

**ORDER** **BY** Avg\_rating **DESC**;

-----------------------------------------------------------

--------------------- SALES ANALYSIS ---------------------

-----------------------------------------------------------

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

**SELECT**

day\_name,

time\_of\_day,

SUM(quantity) **AS** Quantity

**FROM** sales

**GROUP** **BY** day\_name, time\_of\_day

**ORDER** **BY** Quantity **DESC**;

-- 2. Which of the customer types brings the most revenue? What is the participation for each segment?

**SELECT**

customer\_type,

ROUND(SUM(total),2) **AS** total\_revenue,

ROUND((SUM(total) / (**SELECT** SUM(total) **FROM** sales)) \* 100, 2) **AS** segment\_part\_pct

**FROM** sales

**GROUP** **BY** customer\_type

**ORDER** **BY** total\_revenue **DESC**;

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

**SELECT**

city,

ROUND(AVG(vat),2) **as** Average\_Tax\_pct

**FROM** sales

**GROUP** **BY** city

**ORDER** **BY** Average\_Tax\_pct **DESC**;

-- 4. Which customer type pays the most in VAT?

**SELECT**

customer\_type,

ROUND(AVG(vat),2) **AS** avg\_vat,

ROUND(SUM(vat),2) **AS** total\_vat

**FROM**

sales

**GROUP** **BY** customer\_type

**ORDER** **BY** total\_vat **DESC**;

-----------------------------------------------------------

------------------- CUSTOMER ANALYSIS --------------------

-----------------------------------------------------------

-- 1. How many unique customer types and what is the sales distribution?

**SELECT**

customer\_type,

COUNT(\*),

ROUND(SUM(total),2) **AS** total\_revenue

**FROM**

sales

**GROUP** **BY** customer\_type

**ORDER** **BY** total\_revenue;

-- 2. How many unique payment methods and what is the sales distribution?

**SELECT**

payment\_method,

COUNT(\*),

ROUND(SUM(total),2) **AS** total\_revenue

**FROM**

sales

**GROUP** **BY** payment\_method

**ORDER** **BY** total\_revenue **DESC**;

-- 3. What is the gender of most of the customers?

**SELECT**

gender,

COUNT(\*),

ROUND(SUM(total),2) **AS** total\_revenue

**FROM**

sales

**GROUP** **BY** gender

**ORDER** **BY** total\_revenue **DESC**;

-- 4. What is the gender distribution per branch?

**SELECT**

branch,

gender,

COUNT(\*) **AS** count gender,

ROUND((COUNT(\*) / SUM(COUNT(\*)) OVER (PARTITION **BY** branch)) \* 100, 2) **AS** gender percentage

**FROM** sales

**GROUP** **BY** branch, gender

**ORDER** **BY** branch, gender, count\_gender;

-- 5. Which time of the day do customers give most ratings?

**SELECT**

branch,

time\_of\_day,

ROUND(AVG(RATING),2) **AS** avg\_rating

**FROM** sales

**GROUP** **BY** branch, time\_of\_day

**ORDER** **BY** branch, avg\_rating **DESC**;

-- 6. Which day fo the week has the best avg ratings?

**SELECT**

day\_name,

ROUND(AVG(rating),2) **AS** avg\_rating

**FROM** sales

**GROUP** **BY** day\_name

**ORDER** **BY** avg\_rating **DESC**;