**Walmart Sales Data Analysis**

**About**

This project aims to explore the Walmart Sales data to understand top performing branches and products, sales trend of of different products, customer behaviour. The aims is to study how sales strategies can be improved and optimized.

"In this recruiting competition, job-seekers are provided with historical sales data for 45 Walmart stores located in different regions. Each store contains many departments, and participants must project the sales for each department in each store. To add to the challenge, selected holiday markdown events are included in the dataset. These markdowns are known to affect sales, but it is challenging to predict which departments are affected and the extent of the impact”.

**Purposes Of The Project**

The major aim of this project is to gain insight into the sales data of Walmart to understand the different factors that affect sales of the different branches.

**About Data**

This dataset contains sales transactions from a 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 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

Conduct analysis on the data to understand the different product lines, the products lines performing best and the product lines that need to be improved.

1. Sales Analysis

This analysis aims to answer the question of the sales trends of product. The result of this can help use measure the effectiveness of each sales strategy the business applies and what modificatoins are needed to gain more sales.

1. Customer Analysis

This analysis aims to uncover the different customers segments, purchase trends and the profitability of each customer segment.

**Approach Used**

1. **Data Wrangling:** This is the first step where inspection of data is done to make sure **NULL** values and missing values are detected and data replacement methods are used to replace, missing or **NULL** values.
2. Build a database
3. Create table and insert the data.
4. Select columns with null values in them. There are no null values in our database as in creating the tables, we set **NOT NULL** for each field, hence null values are filtered out.
5. **Feature Engineering:** This will help use generate some new columns from existing ones.
   1. Add a new column named time\_of\_day to give insight of sales in the Morning, Afternoon and Evening. This will help answer the question on which part of the day most sales are made.
   2. Add a new column named day\_name that contains the extracted days of the week on which the given transaction took place (Mon, Tue, Wed, Thur, Fri). This will help answer the question on which week of the day each branch is busiest.
   3. 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.
6. **Exploratory Data Analysis (EDA):** Exploratory data analysis is done to answer the listed questions and aims of this project.
7. **Conclusion:**

**Business Questions To Answer**

**Generic Question**

1. How many unique cities does the data have?
2. In which city is each branch?

**Product**

1. How many unique product lines does the data have?
2. What is the most common payment method?
3. What is the most selling product line?
4. What is the total revenue by month?
5. What month had the largest COGS?
6. What product line had the largest revenue?
7. What is the city with the largest revenue?
8. What product line had the largest VAT?
9. Fetch each product line and add a column to those product line showing "Good", "Bad". Good if its greater than average sales
10. Which branch sold more products than average product sold?
11. What is the most common product line by gender?
12. What is the average rating of each product line?

**Sales**

1. Number of sales made in each time of the day per weekday
2. Which of the customer types brings the most revenue?
3. Which city has the largest tax percent/ VAT (**Value Added Tax**)?
4. Which customer type pays the most in VAT?

**Customer**

1. How many unique customer types does the data have?
2. How many unique payment methods does the data have?
3. What is the most common customer type?
4. Which customer type buys the most?
5. What is the gender of most of the customers?
6. What is the gender distribution per branch?
7. Which time of the day do customers give most ratings?
8. Which time of the day do customers give most ratings per branch?
9. Which day fo the week has the best avg ratings?
10. Which day of the week has the best average ratings per branch?

**Revenue And Profit Calculations**

$ COGS = unitsPrice \* quantity $

$ VAT = 5% \* COGS $

 is added to the  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 =

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

**Code**

**# Creating Table**

CREATE TABLE [WalmartSales]

(

[Invoice ID] VARCHAR(512),

[Branch] VARCHAR(512),

[City] VARCHAR(512),

[Customer type] VARCHAR(512),

[Gender] VARCHAR(512),

[Product line] VARCHAR(512),

[Unit price] FLOAT,

[Quantity] INT,

[Tax 5%] FLOAT,

[Total] FLOAT,

[Date] DATE,

[Time] TIME,

[Payment] VARCHAR(512),

[cogs] FLOAT,

[gross margin percentage] FLOAT,

[gross income] FLOAT,

[Rating] FLOAT

);

**# INSERT VALUES INTO TABLE (Sample)**

INSERT INTO [WalmartSales] ([Invoice ID], [Branch], [City], [Customer type], [Gender], [Product line], [Unit price], [Quantity], [Tax 5%], [Total], [Date], [Time], [Payment], [cogs], [gross margin percentage], [gross income], [Rating]) VALUES ('750-67-8428', 'A', 'Yangon', 'Member', 'Female', 'Health and beauty', '74.69', '7', '26.1415', '548.9715', '2019-01-05', '13:08:00', 'Ewallet', '522.83', '4.761904762', '26.1415', '9.1');

INSERT INTO [WalmartSales] ([Invoice ID], [Branch], [City], [Customer type], [Gender], [Product line], [Unit price], [Quantity], [Tax 5%], [Total], [Date], [Time], [Payment], [cogs], [gross margin percentage], [gross income], [Rating]) VALUES ('226-31-3081', 'C', 'Naypyitaw', 'Normal', 'Female', 'Electronic accessories', '15.28', '5', '3.82', '80.22', '2019-03-08', '10:29:00', 'Cash', '76.4', '4.761904762', '3.82', '9.6');

INSERT INTO [WalmartSales] ([Invoice ID], [Branch], [City], [Customer type], [Gender], [Product line], [Unit price], [Quantity], [Tax 5%], [Total], [Date], [Time], [Payment], [cogs], [gross margin percentage], [gross income], [Rating]) VALUES ('631-41-3108', 'A', 'Yangon', 'Normal', 'Male', 'Home and lifestyle', '46.33', '7', '16.2155', '340.5255', '2019-03-03', '13:23:00', 'Credit card', '324.31', '4.761904762', '16.2155', '7.4');

INSERT INTO [WalmartSales] ([Invoice ID], [Branch], [City], [Customer type], [Gender], [Product line], [Unit price], [Quantity], [Tax 5%], [Total], [Date], [Time], [Payment], [cogs], [gross margin percentage], [gross income], [Rating]) VALUES ('123-19-1176', 'A', 'Yangon', 'Member', 'Male', 'Health and beauty', '58.22', '8', '23.288', '489.048', '2019-01-27', '20:33:00', 'Ewallet', '465.76', '4.761904762', '23.288', '8.4');

INSERT INTO [WalmartSales] ([Invoice ID], [Branch], [City], [Customer type], [Gender], [Product line], [Unit price], [Quantity], [Tax 5%], [Total], [Date], [Time], [Payment], [cogs], [gross margin percentage], [gross income], [Rating]) VALUES ('373-73-7910', 'A', 'Yangon', 'Normal', 'Male', 'Sports and travel', '86.31', '7', '30.2085', '634.3785', '2019-02-08', '10:37:00', 'Ewallet', '604.17', '4.761904762', '30.2085', '5.3');

INSERT INTO [WalmartSales] ([Invoice ID], [Branch], [City], [Customer type], [Gender], [Product line], [Unit price], [Quantity], [Tax 5%], [Total], [Date], [Time], [Payment], [cogs], [gross margin percentage], [gross income], [Rating]) VALUES ('699-14-3026', 'C', 'Naypyitaw', 'Normal', 'Male', 'Electronic accessories', '85.39', '7', '29.8865', '627.6165', '2019-03-25', '18:30:00', 'Ewallet', '597.73', '4.761904762', '29.8865', '4.1');

INSERT INTO [WalmartSales] ([Invoice ID], [Branch], [City], [Customer type], [Gender], [Product line], [Unit price], [Quantity], [Tax 5%], [Total], [Date], [Time], [Payment], [cogs], [gross margin percentage], [gross income], [Rating]) VALUES ('355-53-5943', 'A', 'Yangon', 'Member', 'Female', 'Electronic accessories', '68.84', '6', '20.652', '433.692', '2019-02-25', '14:36:00', 'Ewallet', '413.04', '4.761904762', '20.652', '5.8');

INSERT INTO [WalmartSales] ([Invoice ID], [Branch], [City], [Customer type], [Gender], [Product line], [Unit price], [Quantity], [Tax 5%], [Total], [Date], [Time], [Payment], [cogs], [gross margin percentage], [gross income], [Rating]) VALUES ('315-22-5665', 'C', 'Naypyitaw', 'Normal', 'Female', 'Home and lifestyle', '73.56', '10', '36.78', '772.38', '2019-02-24', '11:38:00', 'Ewallet', '735.6', '4.761904762', '36.78', '8');

--SQL CODING FOR DATA ANALYSIS BASED ON QUESTIONS ABOVE :

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\_date

FROM WalmartSales;

-- Adding new column : Time\_of\_date with Type Date VARCHAR (20)

BEGIN TRANSACTION;

BEGIN TRY

ALTER TABLE WalmartSales

ADD Time\_of\_date VARCHAR(20);

COMMIT TRANSACTION;

PRINT ('Data is being processing')

END TRY

BEGIN CATCH

ROLLBACK TRANSACTION;

PRINT('Data is cancel')

END CATCH;

--Input Value to Column Time\_of\_date with Morning, Afternoon or Evening.

BEGIN TRANSACTION;

BEGIN TRY

UPDATE WalmartSales

SET Time\_of\_date = (

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

)

COMMIT TRANSACTION;

PRINT ('Data is being Processing')

END TRY

BEGIN CATCH

ROLLBACK TRANSACTION;

PRINT ('Data is cancel')

END CATCH;

-- day\_name

SELECT

Date,

DATENAME(WEEKDAY,Date) AS day\_name

FROM WalmartSales;

-- Adding new column : day\_name with Type Date VARCHAR (10)

BEGIN TRANSACTION;

BEGIN TRY

ALTER TABLE WalmartSales

ADD day\_name VARCHAR (10);

COMMIT TRANSACTION;

PRINT('Data is being processing')

END TRY

BEGIN CATCH

ROLLBACK TRANSACTION;

PRINT('Data is cancel')

END CATCH;

-- Input value to day\_name with The name of the day:

BEGIN TRANSACTION;

BEGIN TRY

UPDATE Walmartsales

SET day\_name = DATENAME(WEEKDAY,Date)

COMMIT TRANSACTION;

PRINT('Data is being processing')

END TRY

BEGIN CATCH

ROLLBACK TRANSACTION;

PRINT('Data is cancel')

END CATCH;

-- month\_name

SELECT

Date,

DATENAME(MONTH,Date) AS day\_name

FROM WalmartSales;

-- Adding new column : month\_name with Type Date VARCHAR (10)

BEGIN TRANSACTION;

BEGIN TRY

ALTER TABLE WalmartSales

ADD month\_name VARCHAR (10)

COMMIT TRANSACTION;

PRINT ('Data is being processing')

END TRY

BEGIN CATCH

ROLLBACK TRANSACTION;

PRINT ('Data is cancel')

END CATCH;

-- Input value to month\_name with The name of the month:

BEGIN TRANSACTION;

BEGIN TRY

UPDATE WalmartSales

SET month\_name = DATENAME(MONTH,Date)

COMMIT TRANSACTION;

PRINT('Data is being processing')

END TRY

BEGIN CATCH

ROLLBACK TRANSACTION;

PRINT('Data is cancel')

END CATCH;

-- GENERIC QUESTION --

--1. How many unique cities does the data have?

SELECT

DISTINCT (City) AS [Total Unique Cities]

FROM Walmartsales;

--2. In which city is each branch?

SELECT

DISTINCT (City),

Branch

FROM Walmartsales

ORDER BY Branch;

-- Product --

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

SELECT

DISTINCT ([Product line]) AS [Unique Product Lines]

FROM Walmartsales;

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

SELECT TOP 1

Payment,

COUNT(\*) AS [Total Common Payment Method]

FROM Walmartsales

GROUP BY Payment

ORDER BY [Total Common Payment Method] DESC;

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

SELECT TOP 1

[Product line],

COUNT(\*) AS [Most Selling Product Line]

FROM Walmartsales

GROUP BY [Product line]

ORDER BY [Most Selling Product Line] DESC;

--4. What is the total revenue by month?

SELECT

month\_name,

ROUND(SUM(Total),2) AS [Total Revenue]

FROM Walmartsales

GROUP BY month\_name

ORDER BY [Total Revenue] DESC;

--5. What month had the largest COGS?

SELECT TOP 1

month\_name,

ROUND(SUM(cogs),2) AS [Total COGS]

FROM Walmartsales

GROUP BY month\_name

ORDER BY [Total COGS] DESC;

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

SELECT

[Product line],

ROUND(SUM(Total),2) AS [Total Revenue Product]

FROM Walmartsales

GROUP BY [Product line]

ORDER BY [Total Revenue Product] DESC;

--7. What is the city with the largest revenue?

SELECT TOP 1

Branch,

City,

ROUND(SUM(Total),1) AS [Total Revenue]

FROM Walmartsales

GROUP BY Branch,City

ORDER BY [Total Revenue] DESC;

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

SELECT --TOP 1

[Product line],

AVG([Tax 5%]) AS [Average Tax]

FROM Walmartsales

GROUP BY [Product line]

ORDER BY [Average Tax] DESC;

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

WITH AverageSales AS (

SELECT

[Product line],

AVG([Total]) AS AvgSales

FROM

Walmartsales

GROUP BY

[Product line]

)

SELECT

W.\*,

CASE

WHEN W.[Total] > A.AvgSales THEN 'Good'

ELSE 'Bad'

END AS SalesCategory

FROM

Walmartsales W

JOIN

AverageSales A ON W.[Product line] = A.[Product line];

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

SELECT

Branch,

SUM(Quantity) AS [Total Quantity]

FROM Walmartsales

GROUP BY Branch

HAVING SUM(Quantity) > (SELECT AVG(Quantity) FROM Walmartsales);

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

SELECT

[Product line],

gender,

COUNT(gender) AS [Most Common Product By Gender]

FROM Walmartsales

GROUP BY [Product line],gender

ORDER BY [Most Common Product By Gender] DESC;

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

SELECT

[Product line],

ROUND(AVG(Rating),2) AS [Average Rating]

FROM Walmartsales

GROUP BY [Product line]

ORDER BY [Average Rating] DESC;

-- SALES --

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

SELECT

Time\_of\_date,

COUNT(\*) AS [Total Sales Daily]

FROM Walmartsales

WHERE day\_name = 'Sunday'

GROUP BY time\_of\_date;

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

SELECT

[Customer type],

ROUND(SUM(Total),2) AS [Total Revenues]

FROM Walmartsales

GROUP BY [Customer type]

ORDER BY [Total Revenues] DESC;

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

SELECT

City,

ROUND(AVG([Tax 5%]),2) AS [Average Tax]

FROM Walmartsales

GROUP BY City

ORDER BY [Average Tax] DESC;

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

SELECT

[Customer type],

ROUND(AVG([Tax 5%]),2) AS [Average Tax]

FROM Walmartsales

GROUP BY [Customer type]

ORDER BY [Average Tax] DESC;

-- CUSTOMER --

--1. How many unique customer types does the data have? 2

SELECT

DISTINCT [Customer type]

FROM Walmartsales;

--2. How many unique payment methods does the data have? 3

SELECT

DISTINCT [Payment]

FROM Walmartsales;

--3. What is the most common customer type? Member

SELECT

[Customer type],

COUNT(\*) AS [Total Common Customer Type]

FROM Walmartsales

GROUP BY [Customer type]

ORDER BY [Total Common Customer Type] DESC;

--4. Which customer type buys the most? Member

SELECT

[Customer type],

COUNT(\*) AS [Total Customer Type Buys]

FROM Walmartsales

GROUP BY [Customer type]

ORDER BY [Total Customer Type Buys] DESC;

--5. What is the gender of most of the customers? Female

SELECT

Gender,

COUNT(\*) AS [Total Gender]

FROM Walmartsales

GROUP BY Gender

ORDER BY [Total Gender] DESC;

--6. What is the gender distribution per branch?

SELECT

Branch,

Gender,

COUNT(\*) AS [Total Gender]

FROM Walmartsales

--WHERE Branch = 'A'

GROUP BY Branch,Gender

ORDER BY Branch,[Total Gender] DESC;

--7. Which time of the day do customers give most ratings? Afternoon

SELECT

Time\_of\_date,

ROUND(AVG(Rating),2) AS [Average Rating]

FROM Walmartsales

GROUP BY Time\_of\_date

ORDER BY [Average Rating] DESC;

--8. Which time of the day do customers give most ratings per branch?

SELECT

Branch,

Time\_of\_date,

ROUND(AVG(Rating),2) AS [Average Rating]

FROM Walmartsales

GROUP BY Branch,Time\_of\_date

ORDER BY Branch,[Average Rating] DESC;

--9. Which day of the week has the best avg ratings?

SELECT

day\_name,

ROUND(AVG(Rating),2) AS [Average Rating]

FROM Walmartsales

GROUP BY day\_name

ORDER BY [Average Rating] DESC;

--10. Which day of the week has the best average ratings per branch?

SELECT

Branch,

day\_name,

ROUND(AVG(Rating),2) AS [Average Rating]

FROM Walmartsales

GROUP BY Branch,day\_name

ORDER BY Branch,[Average Rating] DESC;