**Client: SQL Capstone Project @Odinschool**

**Contents:**

* Project Overview & Scope
* Business Problem
* Technical Stacks
* Data Collection and Understanding
* System Requirements
* Data Preprocessing
* Analysis List
* Exploratory Data Analysis
* Challenges
* Recommendations

1. **Project Overview :**

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

1. **Business Problem:**

The dataset addresses several key business problems aimed at improving overall business performance. These include:

1. **Sales Optimization:** Understanding and improving sales performance across branches and product lines to increase revenue.
2. **Customer Segmentation and Engagement:** Tailoring marketing efforts and enhancing customer engagement through effective segmentation and targeted strategies.
3. **Product Management:** Assessing product performance and optimizing offerings based on customer preferences and market trends.
4. **Operational Efficiency:** Improving inventory management and streamlining operational processes to reduce costs and improve service quality.
5. **Marketing Effectiveness:** Evaluating the impact of marketing campaigns and optimizing marketing strategies to maximize return on investment.
6. **Profitability Analysis:** Analyzing profitability across different segments to identify opportunities for cost reduction and revenue enhancement.
7. **Customer Satisfaction:** Measuring and enhancing customer satisfaction levels to foster loyalty and improve overall customer experience.
8. **Market Trends and Competition:** Monitoring market trends and competitor activities to stay competitive and capitalize on emerging opportunities.

By addressing these business problems and leveraging insights from the dataset, our organization can make informed decisions to drive growth, enhance competitiveness, and improve overall business performance.

1. **Technical Stocks:** MYSQL workbench,Jupyter notebook, matplotlib, seaborn, python, numpy, pandas
2. **Data Collection and Understanding**

**About Data:**

This dataset contains sales transactions from three different branches of Amazon, 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 sold | 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 | Varchar(20) |
| 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) |

1. **System Requirements**

--Operating System: Windows 10

--System Type: 64 bit Operating System,x64-based processor

--Ram:8GB

--Processor: 11th Gen Intel(R) Core(TM) i3-1115G4 @ 3.00GHz 3.00 GHz

**6. Data Preprocessing:**

**1. Data Wrangling:**

1.1 Build a database: Set up a database where we will store our data.

create database AmazonDB;

use AmazonDB;

1.2 Create a table and insert the data: Define a table schema that matches the structure of our dataset and insert the data into the table.

--Here I have imported Amazon data in to the database AmaoznDB and retrieving the data by

select \* from amazon;

1.3 Select columns with null values: Verify if there are any null values in our dataset and handle them appropriately.

Here,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.

**2. Feature Engineering:**

Feature engineering is the process of creating new features or variables from existing ones to improve the performance of machine learning models or gain insights from the data. In the context of our analysis, feature engineering serves several important purposes:

By extracting additional information such as time of day, day of the week, and month of the year from the existing columns (time and date) for understanding sales patterns by time of day (morning, afternoon, evening) and different timescales (daily, weekly, monthly) can help identify peak hours for sales, while analyzing sales trends by day of the week and month of the year can reveal recurring patterns and seasonal variations.

/\*1**.Add a new column named timeofday 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.\*/

ALTER TABLE amazon

ADD COLUMN timeofday VARCHAR(20);

UPDATE amazon

SET timeofday =

CASE

WHEN HOUR(time) >= 0 AND HOUR(time) < 12 THEN 'Morning'

WHEN HOUR(time) >= 12 AND HOUR(time) < 18 THEN 'Afternoon'

ELSE 'Evening'

END;

/\* **2.Add a new column named dayname 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.\*/

ALTER TABLE amazon

ADD COLUMN dayname VARCHAR(20);

UPDATE amazon SET dayname = dayname(date);

**/\*3.Add a new column named monthname that contains the extracted months of the year on which the given transaction took place (Jan, Feb, Mar).**

It helps to determine which month of the year has the most sales and profit.\*/

ALTER TABLE amazon

ADD COLUMN monthname varchar(20);

UPDATE amazon

set monthname=monthname(date);

1. **Analysis List**

**Product Analysis:**

To conduct a product analysis on the dataset to understand the different product lines, identify the best-performing product lines, and determine areas for improvement.

* **Identify Product Lines:**

--Count the distinct product lines in the dataset.

SELECT COUNT(DISTINCT `product line`) AS distinct\_product\_lines FROM amazon;

--Determine the list of unique product lines available.

SELECT DISTINCT `product line` FROM amazon;

* **Analyze Sales Performance:**

--Calculate the total sales for each product line.

SELECT `product line`, SUM(total) AS total\_sales

FROM amazon GROUP BY `product line` ORDER BY total\_sales DESC;

--Determine which product lines generate the highest sales revenue.

SELECT `product line`, SUM(total) AS total\_sales

FROM amazon GROUP BY `product line` ORDER BY total\_sales DESC LIMIT 6;

--Evaluate Product Line Performance:

--Calculate average sales or revenue for each product line to assess performance.

SELECT `product line`, AVG(total) AS average\_sales FROM amazon GROUP BY `product line` ORDER BY average\_sales DESC;

--Identify product lines with above-average performance as well as those with below-average performance.

-- Identify product lines with above-average performance\*/

SELECT `product line`, AVG(total) AS average\_sales

FROM amazon GROUP BY `product line`

HAVING AVG(total) > (SELECT AVG(total) FROM amazon);

**--Identify Opportunities for Improvement:**

--Review product lines with below-average performance to understand reasons behind their lower sales.

-- Identify product lines with below-average performance\*/

SELECT `product line`, AVG(total) AS average\_sales

FROM amazon GROUP BY `product line`

HAVING AVG(total) < (SELECT AVG(total) FROM amazon);

**Sales Analysis:**

This analysis aims to answer the question of the sales trends of product.

The result of this can help to measure the effectiveness of each sales strategy the business applies and what modifications are needed to gain more sales.

To conduct a sales analysis aiming to understand sales trends of products, measure the effectiveness of sales strategies, and identify areas for improvement, we can follow these steps:

**Analyze Overall Sales Trends:**

--Calculate total sales over time to understand overall sales trends.

-- Calculate total sales over time\*/

SELECT date, SUM(total) AS total\_sales

FROM amazon GROUP BY date ORDER BY date;

--Visualize the trend of total sales over time (e.g., monthly or quarterly sales trends).

(see in PPT)

**Evaluate Sales Performance by Product Line:**

--Calculate total sales for each product line to identify top-selling product lines.

-- Calculate total sales for each product line

SELECT `product line`, SUM(total) AS total\_sales

FROM amazon GROUP BY `product line` ORDER BY total\_sales DESC;

--Assess Sales Performance by Branch:

Calculate total sales for each branch to identify top-performing branches. -- Calculate total sales for each branch

SELECT branch, SUM(total) AS total\_sales FROM amazon GROUP BY branch

ORDER BY total\_sales DESC;

**Identify Seasonal Trends:**

--Analyze sales data to identify seasonal trends or patterns in sales.

-- we can aggregate sales data by month or quarter and visualize the trend.

--Determine peak seasons or months with highest sales and off-peak seasons.

select sum(total) as total\_sales,monthname from amazon group by monthname order by total\_sales desc;

--identify In which city was the highest revenue recorded

select sum(Total) as highest\_revenue,city from amazon group by city order by highest\_revenue desc limit 1;

**Evaluate Sales by Customer Type:**

--Calculate total sales for each customer type to understand their contribution to overall sales.

--Analyze sales trends for different customer types to identify key customer segments.

-- Calculate total sales for each customer type

SELECT `customer type`, SUM(total) AS total\_sales

FROM amazon

GROUP BY `customer type`

ORDER BY total\_sales DESC;

**Assess Sales by Payment Method:**

--Calculate total sales for each payment method to understand payment preferences.

--Analyze trends in sales by payment method to identify popular payment methods.

-- Calculate total sales for each payment method\*/

SELECT payment, SUM(total) AS total\_sales

FROM amazon

GROUP BY payment

ORDER BY total\_sales DESC;

**Customer Analysis**

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

To conduct a customer analysis aiming to uncover different customer segments, purchase trends, and profitability of each customer segment, we can follow these steps:

**Segment Customers:**

--Identify different customer segments based on their characteristics (e.g., customer type, gender, location).

--Segment customers based on their purchasing behavior (e.g., frequency of purchases, total purchase amount).

-- Segment customers based on customer type\*/

SELECT `customer type`, COUNT(\*) AS customer\_count

FROM amazon

GROUP BY `customer type`;

-- Segment customers based on gender

SELECT gender, COUNT(\*) AS customer\_count

FROM amazon

GROUP BY gender;

**Analyze Purchase Trends:**

--Calculate total sales for each customer segment to understand their contribution to overall sales.

--Analyze purchase trends over time for each customer segment to identify patterns or changes in behavior.

-- Calculate total sales for each customer segment\*/

SELECT `customer type`, SUM(total) AS total\_sales

FROM amazon

GROUP BY `customer type`

ORDER BY total\_sales DESC;.

**Assess Profitability:**

--Calculate gross margin or profit for each customer segment to understand their profitability.

--Analyze profitability trends over time to identify segments with high or low profitability.

-- Calculate gross margin or profit for each customer segment\*/

SELECT `customer type`, SUM(`gross income`) AS total\_profit

FROM amazon

GROUP BY `customer type`

ORDER BY total\_profit DESC;

**Identify High-Value Customers:**

--Identify high-value customers based on their total purchase amount or frequency of purchases.

--Analyze the characteristics and purchasing behavior of high-value customers to understand their importance to the business.

-- Identify high-value customers based on total purchase amount\*/

SELECT `invoice id`, SUM(total) AS total\_purchase\_amount

FROM amazon

GROUP BY `invoice id`

ORDER BY total\_purchase\_amount DESC

LIMIT 10;

**Evaluate Customer Retention:**

--Analyze customer retention rates to understand how well the business is retaining customers over time.

--Identify strategies to improve customer retention and loyalty.

-- Analyze customer retention rates over time

-- You can calculate the percentage of customers who make repeat purchases in each time period.

**8. Exploratory Data Analysis (EDA):**

Exploratory data analysis is done to answer the listed questions and aims of this project.

**Business Questions To Answer:**

/\*What is the count of distinct cities in the dataset?\*/

select count(distinct(city)) from amazon;

/\*For each branch, what is the corresponding city?\*/

select city ,branch from amazon group by branch,city;

/\*What is the count of distinct product lines in the dataset?\*/

SELECT count(Distinct `product line`) from amazon;

/\*Which payment method occurs most frequently?\*/

select count(payment),payment from amazon group by payment;

/\*Which product line has the highest sales?\*/

select sum(total),`product line` from amazon group by `product line` order by sum(total) desc limit 1;

/\*How much revenue is generated each month?\*/

select sum(total),monthname from amazon group by monthname;

/\*In which month did the cost of goods sold reach its peak?\*/

select sum(cogs),monthname from amazon group by monthname order by sum(cogs) desc limit 1;

/\*Which product line generated the highest revenue?\*/

select sum(total),`product line` from amazon group by `product line` order by sum(total) desc limit 1;

/\*In which city was the highest revenue recorded?\*/

select sum(Total) as highest\_revenue,city from amazon group by city order by highest\_revenue desc limit 1;

/\*Which product line incurred the highest Value Added Tax?\*/

/\*For each product line, add a column indicating "Good" if its sales are above average, otherwise "Bad."\*/

SELECT

\*,

CASE

WHEN tot\_sales > avg\_sales THEN 'Good'

ELSE 'Bad'

END AS sales\_quality

FROM (

SELECT

`product line`,

SUM(total) AS tot\_sales,

AVG(sum(total)) over() AS avg\_sales

FROM amazon

GROUP BY `product line`

) AS subquery;

/\*Identify the branch that exceeded the average number of products sold.\*/

SELECT

branch,

COUNT(\*) AS num\_products\_sold

FROM

amazon

GROUP BY

branch

HAVING

COUNT(\*) > (SELECT AVG(num\_products\_sold) FROM

(SELECT COUNT(\*) AS num\_products\_sold FROM amazon GROUP BY branch) AS subquery);

/\*Which product line is most frequently associated with each gender?\*/

select count(\*) as frequency,`product line` ,gender from amazon group by `product line`,gender order by frequency desc;

/\*Calculate the average rating for each product line.\*/

select avg(rating) as avg\_rating,`product line` from amazon group by `product line`;

/\*Count the sales occurrences for each time of day on every weekday.\*/

select count(\*) as sales\_occurances,dayname,timeofday from amazon group by timeofday,dayname order by dayname,timeofday;

/\*Identify the customer type contributing the highest revenue.\*/

select sum(total) as highest\_revenue,`customer type` from amazon group by `customer type` order by highest\_revenue desc limit 1;

/\*Determine the city with the highest VAT percentage.\*/

select city,(sum(`Tax 5%`)/sum(total))\*100 as vat\_percentage from amazon group by city order by vat\_percentage desc limit 1;

/\*Identify the customer type with the highest VAT payments.\*/

select `customer type`, sum(`Tax 5%`) as highest\_vat\_payment from amazon group by `customer type` order by highest\_vat\_payment desc limit 1;

/\*What is the count of distinct customer types in the dataset?\*/

select count(distinct(`customer type`)) as dist\_cust\_type from amazon;

/\*What is the count of distinct payment methods in the dataset?\*/

select count(distinct(payment)) as dist\_pay\_type from amazon;

/\*Which customer type occurs most frequently?\*/

select count(\*) as freq,`customer type` from amazon group by `customer type` order by freq desc ;

/\*Identify the customer type with the highest purchase frequency.\*/

select `customer type` ,sum(total) from amazon group by `customer type`

/\*Determine the predominant gender among customers.\*/

select gender,count(\*) as predominant from amazon group by gender order by predominant desc limit 1;

/\*Examine the distribution of genders within each branch.\*/

select branch,gender,count(\*) from amazon group by branch,gender order by branch ;

/\*Identify the time of day when customers provide the most ratings.\*/

select timeofday ,sum(rating) as ratings from amazon group by timeofday order by ratings desc limit 1;

/\*Determine the time of day with the highest customer ratings for each branch.\*/

select timeofday ,branch,sum(rating) as ratings from amazon group by timeofday,branch order by ratings desc limit 1;

/\*Identify the day of the week with the highest average ratings.\*/

select dayname ,avg(rating) as ratings from amazon group by dayname order by ratings desc limit 1;

/\*Determine the day of the week with the highest average ratings for each branch.\*/

select branch,dayname,avg(rating) as ratings from amazon group by dayname,branch order by ratings desc limit 1;

1. **Results:**

* This dataset contains sales transactions from three different branches A, B and C of Amazon, respectively located in Yangon and Mandalay Naypyitaw and having 5 product lines namely Food and Beverages, sports and travel, home and lifestyle, Electronic accessories, Health and Beauty and Fashion accessories
* This dataset contains 1000 rows and 17 columns. Added 3 more columns to identify trends over time like day wise, week wise and monthly wise.
* There are two types of customers namely amazon members and normal customers.
* There are 3 payment methods namely Ewallet,cash and credit card.
* This data is generated during the period of feb and mar of 2019.
* This data set contains 2 types of genders males and females.
* From the total customers, amazon membership customers are more than amazon normal customers.
* In this,when we compare total sales with each product line
  + Food and Beverages generated highest sales and health and beauty generated lowest sales .
  + Avg sales of each product line, highest avg sales generated at home and life syle and lowest avg sales generated at Fashion accessories
  + The product lines Health and Beauty, sports & travel and Home & life style are performed above average performance and remaining done are below avg performance.
  + The product line “ Fashion accessories” is most frequently associated with each gender.
* When we compare product line with ratings,
  + ‘Food and beverages’ has highest average rating and Home and life style has lowest rating.
  + Highest ratings occurred during afternoon if we calculate with all branches.
  + Highest ratings occurred during afternoon with branch A ,if we calculate branch wise.
  + Highest avg ratings occurred on Monday if we calculate with all branches.
  + Highest ratings occurred during on monday with branch B ,if we calculate branch wise.
* When we compare revenue over time,
  + Highest revenue generated in jan month and lowest revenue generated in feb 2019.
  + cost of goods sold reach its peak in the jan month and low in the month of feb 2019
  + sales occurrences are more in the afternnon on every weekday especially in Saturday afternoon.
  + Highest sales generated on march 9th , 2019 , lowest on feb 13th,2019.
  + Highest sales generated on jan, and lowest sales generated on feb.
* When we compare revenue with city,
  + Highest revenue generated at city Naypyitaw and lowest revenue generated at city mandalay.
* When we compare cutomertype with revenue,
  + customer type “members” occurs most frequently and Amazon members are generating highest revenue than normal customers. And total profit also high with the amazon members.
  + Females are more than males in amazon customers
  + Amazon members contributed more sales than normal customers
  + With the invoice id 860-79-0874 purchased more than other invoice id’s.
* When we compare sales with respective branches,
  + branch c generated highest sales and low as branch B.
* .when we compare sales with payments,
  + Sales are more by cash payments than ewallet and credit card.

1. **Recommendations:**

* Additional analysis and strategies for improvement based on identified product lines.
* Explore potential strategies to improve sales for these product lines, such as marketing campaigns, promotions, or product improvements.
* Launch targeted marketing campaigns to promote top-selling products or product lines.
* Optimize pricing strategies for products with below-average sales performance to increase competitiveness.
* Enhance customer engagement through personalized recommendations or loyalty programs based on customer segmentation.
* Improve inventory management processes to ensure adequate stock levels for high-demand products during peak seasons.
* Invest in training and development programs to enhance sales team performance and customer service.
* Expand product offerings or introduce new product lines based on emerging trends or customer preferences identified in the analysis.