SQL - Capstone Project

(AMAZON SALES ANALYSIS)

Name-- Sumit Baviskar.

Date-- 11 March 2024

Purposes Of This Capstone Project

The major aim of this project is to gain insight into the sales data of Amazon to understand the different factors that affect sales of the different branches. Gather the information to give solution the client question and requirement that improve sales so they can take data-driven decision.

Program used—MySql .

Understanding the 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 995 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	TIME
payment_method	The total amount paid	VARCHAR(100)
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)

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.

1.1 **Build a database**
- created a base named” Amazon_data ”.

1.2 **Create a table and insert the data**
- create a new table name “Amazon”.

1.3 **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.**
-- checked for not null .there are no null values in our dataset.

2. **Feature Engineering:** This will help us generate some new columns from existing ones.

2.1) **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.**
-- creating new column and updating data accordingly with Morning , Afternoon, Evening category in the time_of_day columns by using case statement . Category of time_of_day with Morning(190 rows),Afternoon(525 rows),Evening(280 rows),

2.2) **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.**
-- Added a new column as ”DayName” at the last of the table using Dayname with Date function the field is filled. There was 5 distinct values Monday, Tuesday, Wednesday, Thursday, and Friday.

2.3) **Add a new column named monthname that contains the extracted months of the year on which the given transaction took place (Jan, Feb, Mar). Help determine which month of the year has the most sales and profit**.

-- using the MONTHNAME is used with date to get month from date and get the column filled with all values. There are January, February and March.

Business Questions To Answer:

1) What is the count of distinct cities in the dataset?

-- Count of distinct city is 3 (Yangon, Naypyitaw , Mandalay)

2) For each branch, what is the corresponding city?

--Each branch has only one city to each branch . Branch A- Yangon, Branch C-Naypyitaw , Branch B-Mandalay respectively

3) What is the count of distinct product lines in the dataset?

--There are 6 distinct product line count (Health and Beauty, Electronics accessories, Sport and travel, Home and lifestyle, Food and Beverages, Fashion accessories)

4) Which payment method occurs most frequently?

--Most payment method is Cash followed by E-wallet , Credit cash.

5) Which product line has the highest sales?

-- Highest sales is done by Food and Beverages followed by Fashion accessories, Sport and travel, Home and lifestyle, Electronics accessories and Health and Beauty has lowest sales.

6) How much revenue is generated each month?

--Highest revenue is earned in January Followed by March and February.

7) In which month did the cost of goods sold reach its peak?

-- January is the highest revenue month followed by March and February. This analysis can be get by summing cogs(Cost Of Goods sold) by grouping the month. There are only 3 months January, February and March.

8) Which product line generated the highest revenue?
We have calculate summed the unit price and quantity get revenue and order by Revenue . As Food and beverages has generate the highest revenue followed by fashion accessories ,Sport & travel ,home & life style , Electronics & accessories and Health & beauty.

9) In which city was the highest revenue recorded?
As mentioned above, we calculate revenue by adding gross income and cost of sold goods has arranged in descending to get city with highest revenue which is Naypyitaw Followed by Yangon ,Mandalay.

10) Which product line incurred the highest Value Added Tax?
As the vat is given which is grouped by product line get us insight that the Food & beverages Followed by accessories Sport & travel, Home & travel Electronics accessories, and Health & beauty.

11) Which product line is most frequently associated with each gender?
As there is different preferences to product line for **female** which is Fashion accessories , Food & beverages , Sport & travel, Electronics accessories, Home & lifestyle and Health & beauty,
Male had preferences to Health & beauty followed by Electronics accessories, Food & beverages, Fashion accessories , Home & lifestyle and Sport & travel.

12) Calculate the average rating for each product line.
--The highest average rating for product line is Food & beverages followed by Fashion accessories, Health & beauty, Electronics accessories , Sport & travel and Home & lifestyle.

13) Count the sales occurrences for each time of day on every weekday.
-- The highest sales occurrences at Saturday(Evening) followed by Tuesday(Evening),Wednesday(Afternoon).

14) Identify the customer type contributing the highest revenue.
-- there are only two customer type as member and normal but as per revenue member customer type generate more revenue than normal customer type around 4% more than normal member type.

15) Determine the city with the highest VAT percentage.
-- As the cities is arranged in the highest VAT are Naypyitaw after that Yangon and Mandalay. Yangon and Mandalay had the similar percentage but the value of Yangon is higher then Mandalay.

16) Identify the customer type with the highest VAT payments.
--There is a small difference between only 2 customer type which is normal and member type . Member has sightly higher value which is around 300 more than normal customer type VAT value.

17) What is the count of distinct customer types in the dataset?
--There are only 2 customer type Member and Normal customer type.

18) What is the count of distinct payment methods in the dataset?
--There is only 3 payment method which is E-wallet, cash and credit card

19) Which customer type occurs most frequently?
--there are normal and member customer type but there are 3 members customer type than normal customer type.

20) Identify the customer type with the highest purchase frequency.
--there are normal and member customer type but there are 3 members customer type than normal customer type.

21) Determine the predominant gender among customers.
-- In member customer type, there are female are predominant than male and in normal customer type, there are male are predominant than female.

22) Examine the distribution of genders within each branch.
-- gender distribution among branch A has male has 179 members and Female 160 members, branch B has male has 169 members and Female 160 members, and So branch A has Female has 177 members and Female 150 members

23) Identify the time of day when customers provide the most ratings.

-- As per data, Afternoon has highest average rating has 7.02 rating followed by Morning and Evening

24) Determine the time of day with the highest customer ratings for each branch.
-- For every branch there is a time of day in which the highest rating is given 9.9 rating at each time of day for each branch has given 9.9 rating ones in every time of day.

25) Identify the day of the week with the highest average ratings.
Monday has highest average rating 7.13 rating followed by Friday(7.06), Tuesday(7.00) , Sunday(6.99) , Saturday(6.90) ,Thursday(6.89) and Wednesday(6.76).

26) Determine the day of the week with the highest average ratings for each branch.
-- For **branch A** , Friday(7.31) as highest average rating followed by Monday(7.10) ,Sunday, Tuesday, Thursday, Wednesday and Saturday.
For **Branch B**, Monday (7.27) as highest average rating followed by Tuesday, Sunday, Thursday, Saturday, Friday, Wednesday .
For **Branch C,** Saturday(7.23) as highest average rating followed by Friday, Wednesday, Monday , Sunday, Tuesday, and Thursday .

Analysis List --

3. 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.

Product Line – there are 6 distinct product line-

- Health and Beauty
- Electronics accessories
- Sport and travel
- Home and lifestyle
- Food and Beverages
- Fashion accessories

He product line should be test on average quantity produced ,average rating, revenue and cogs(cost of goods sold)

Average quantity produced by each product line – so there is highest product line is home and lifestyle, Electronics accessories, Health & beauty, Sport and travel , Food and beverages and Fashion accessories

By rating if we rank product line home and lifestyle , Sports and travel, Electronics and accessories, health and beauty ,Fashion and accessories and Food & beverages .

According to revenue, Food and beverages followed by Fashion accessories ,sports and travel , Food & beverages . ,home & lifestyle ,Electronics accessories and Health & beauty.

If we use criteria of good sold by each product line Food and beverages , Fashion accessories, Sports and travel, home and travel , home & lifestyle, Electronics & accessories and Health & beauty.

As per the above criteria we can see, the product line like Electronics and accessories, home and lifestyle, Food and beverages are good performers product line . By the health &beauty and home& lifestyle are low performing product line and should used some strategy to improve the revenue , rating and unit sold .

4. Sales Analysis--

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

Sales analysis can be done by revenue and gross income generated by each month . So as per data we can analyze and get insight that monthly revenue generated is highest at January followed March and February and if we arranged by criteria of total gross income in January followed March and February

So Sales in February is reduce to certain extend so with sales strategy to improve sales in February and March.

5. Customer Analysis--

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

Customer Analysis can be done by order behavior by average order value, count of order by each customer . There are member customer type has more average order values than normal customer type.

Femal has more average order value than men which is around 150 more than men in each order placed by men. As more deep analysis Member female has highest average order value followed by normal female customer type, Member male and lowest male normal customer type .

We have to make improvement in male category sales and specially in normal customer type.