Walmart SQL Project Vivek Chauhan

```
#we are working with csv files, here are the steps to import csv files in sql
# 1) create a new database in mysql
# 2) Go to import section & select the file you want to upload
# 3) Turn on this setting at the bottom part "The first line of the file contains the table column names (if this is unchecked, the first line will become part of the data)"
```

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. The dataset was obtained from the Kaggle Walmart Sales Forecasting Competition.

Purposes Of The Project :-

The major aim of thie 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 :-

The dataset was obtained from the Kaggle Walmart Sales Forecasting Competition. 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:

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.

2) 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.

3) 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.

Build a database :-

Create table and insert the data.

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.

2) 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.

3) Exploratory Data Analysis (EDA): Exploratory data analysis is done to answer the listed questions and aims of this project.

Business Questions To Answer :-

Generic Question :-

How many unique cities does the data have?

In which city is each branch?

Product :-

How many unique product lines does the data have?

What is the most common payment method?

What is the most selling product line?

What is the total revenue by month?

What month had the largest COGS?

What product line had the largest revenue?

What is the city with the largest revenue?

What product line had the largest VAT?

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

Which branch sold more products than average product sold?

What is the most common product line by gender?

What is the average rating of each product line?

Sales :-

Number of sales made in each time of the day per weekday

Which of the customer types brings the most revenue?

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

Which customer type pays the most in VAT?

Customer :-

How many unique customer types does the data have?

How many unique payment methods does the data have?

What is the most common customer type?

Which customer type buys the most?

What is the gender of most of the customers?

What is the gender distribution per branch?

Which time of the day do customers give most ratings?

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

Which day fo the week has the best avg ratings?

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

Revenue And Profit Calculations :-

$ COGS = unitsPrice * quantity $

$ VAT = 5% * COGS $

V

A

T

is added to the

C

O

G

S

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 $

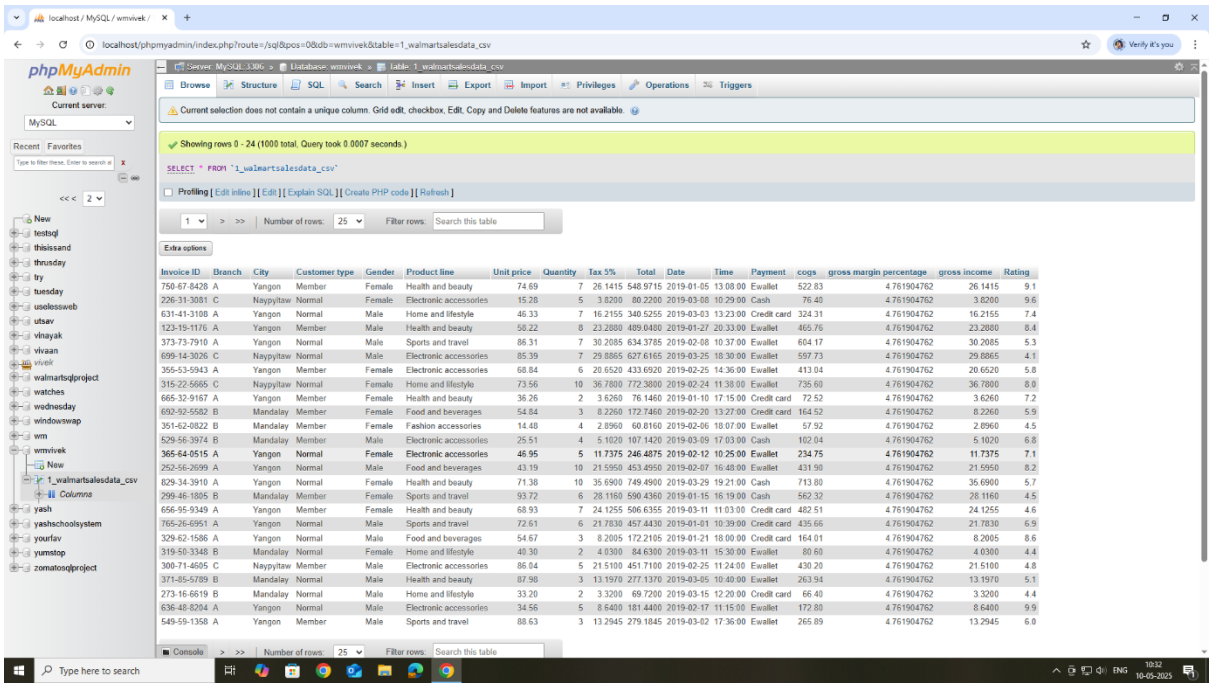
$ total = VAT + COGS\= 16.0265 + 320.53 =

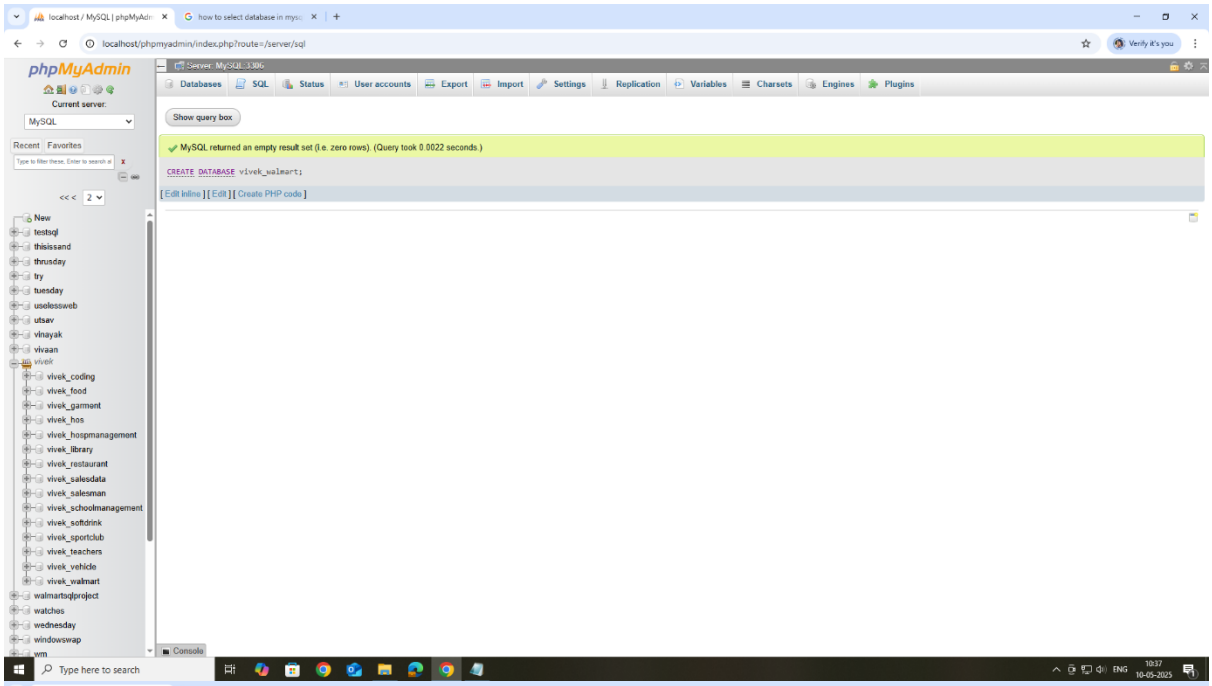
336.5565

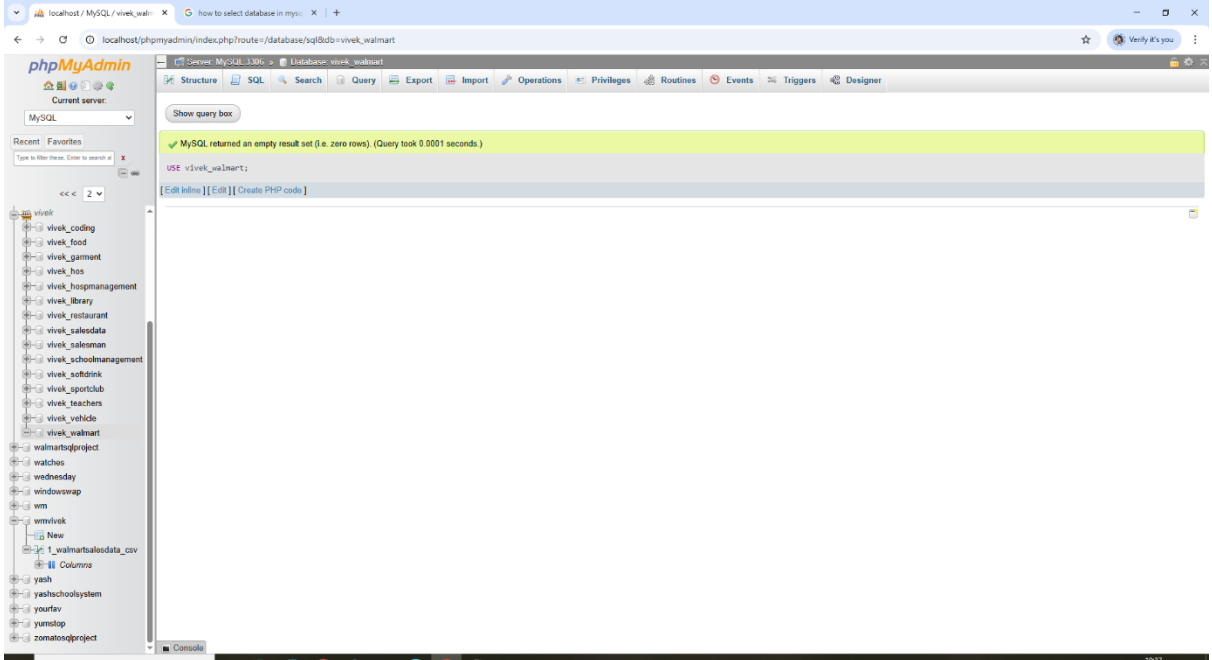
$ \text{Gross Margin Percentage} = \frac{\text{gross income}}{\text{total revenue}}\=\frac{16.0265}{336.5565} = 0.047619\\approx 4.7619% $

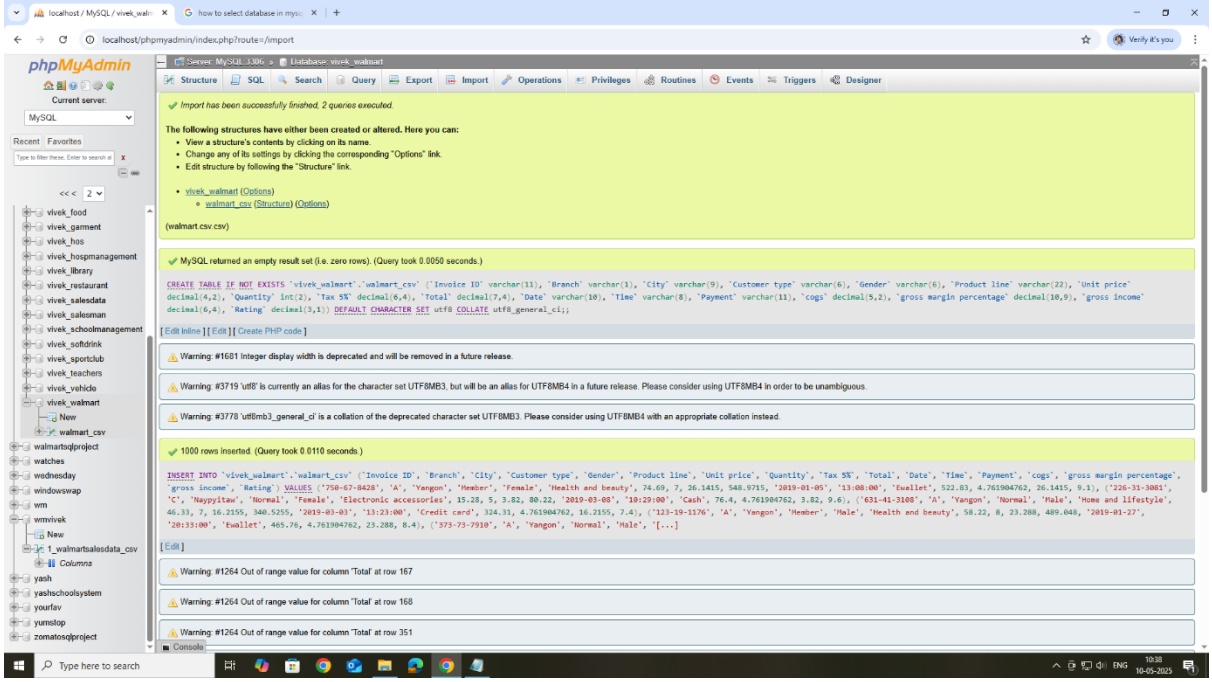
→Project Starts from here :-

Import the .csv format file in phpMyAdmin then after your project Query is start.




phpMyAdmin
Server: MySQL:3306
Databases SQL Status User accounts Export Import Settings Replication Variables Charsets Engines Plugins
Show query box
MySQL returned an empty result set (i.e. zero rows). (Query took 0.0022 seconds.)
CREATE DATABASE vivek_walmart;
[Edit inline] [Edit] [Create PHP code]


phpMyAdmin
Server: MySQL:3306 » Database: vivek_walmart
Structure SQL Search Query Export Import Operations Privileges Routines Events Triggers Designer
Show query box
MySQL returned an empty result set (i.e. zero rows). (Query took 0.0001 seconds.)
USE vivek_walmart;
[Edit inline] [Edit] [Create PHP code]


phpMyAdmin
Server: MySQL:3306 » Database: vivek_walmart
Structure SQL Search Query Export Import Operations Privileges Routines Events Triggers Designer
Import has been successfully finished, 2 queries executed.
The following structures have either been created or altered. Here you can:
View a structure's contents by clicking on its name.
Change any of its settings by clicking the corresponding "Options" link.
Edit structure by following the "Structure" link.
vivek_walmart (Options)
walmart_csv (Structure) (Options)
(walmart.csv.csv)
MySQL returned an empty result set (i.e. zero rows). (Query took 0.0050 seconds.)
[Edit inline] [Edit] [Create PHP code]
Warning: #1681 Integer display width is deprecated and will be removed in a future release.
Warning: #3719 'utf8' is currently an alias for the character set UTF8MB3, but will be an alias for UTF8MB4 in a future release. Please consider using UTF8MB4 in order to be unambiguous.
Warning: #3778 'utf8mb3_general_ci' is a collation of the deprecated character set UTF8MB3. Please consider using UTF8MB4 with an appropriate collation instead.
1000 rows inserted. (Query took 0.0110 seconds.)
[Edit]
Warning: #1264 Out of range value for column 'Total' at row 167
Warning: #1264 Out of range value for column 'Total' at row 168
Warning: #1264 Out of range value for column 'Total' at row 351

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.

→With this we get the data classified into time of day

→Create the table first to apply the query.

→Your Query is :-

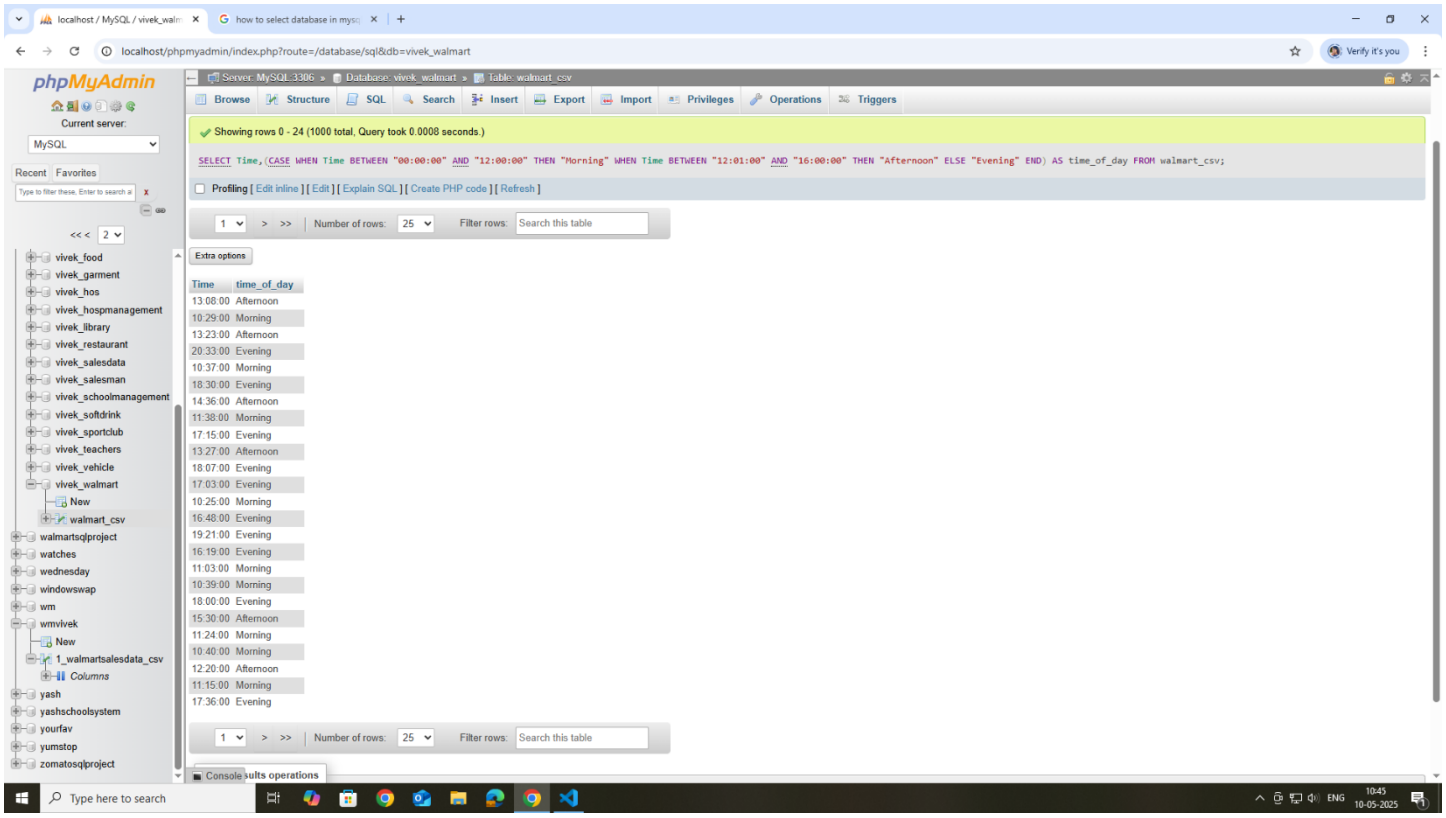
ALTER TABLE walmart_csv ADD COLUMN time_of_day varchar(20);

→Second Step is :-

UPDATE walmart_csv 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);

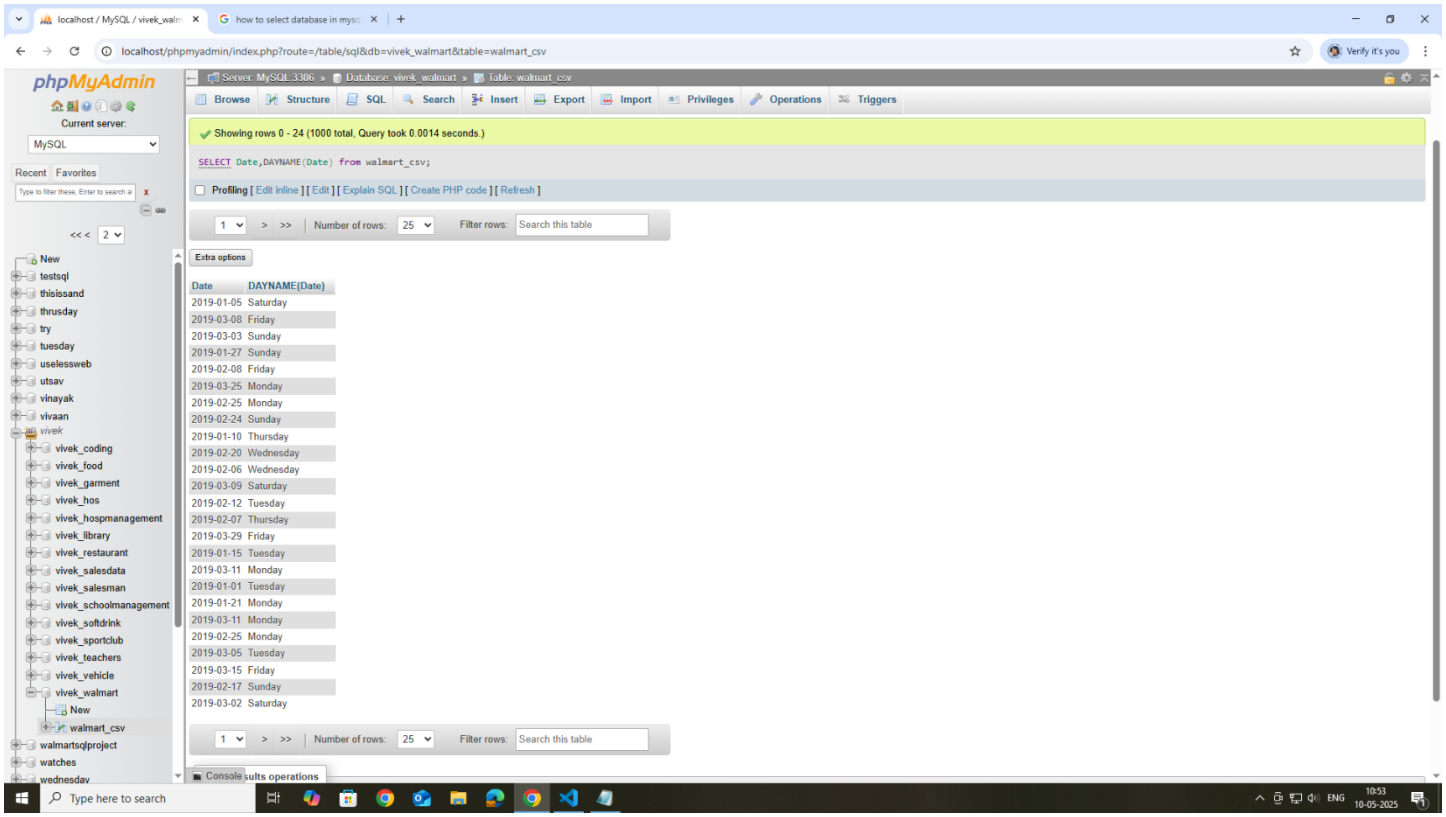
→Third Step is :-

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 walmart.csv;



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.

from this we had bring the name of the days from the date column



now we want to add new column in our dataset named day_name

→Your Query is :-

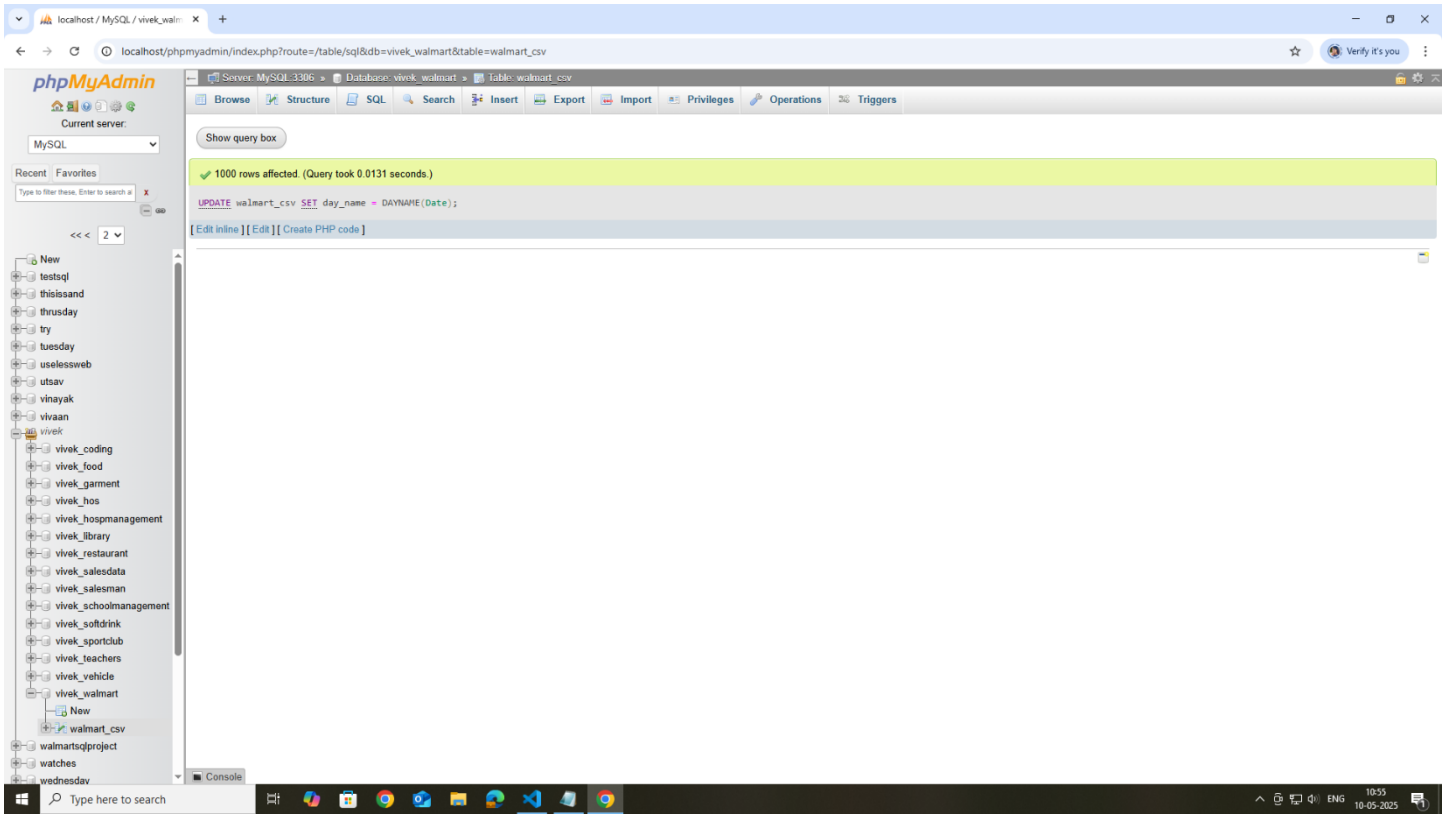
ALTER TABLE walmart_csv ADD COLUMN day_name varchar(10);

this query will update the column in the data set

→Second Query is :-

UPDATE walmart_csv SET day_name = DAYNAME(Date);

if you check a column had been added at last named day_name with all the data inserted

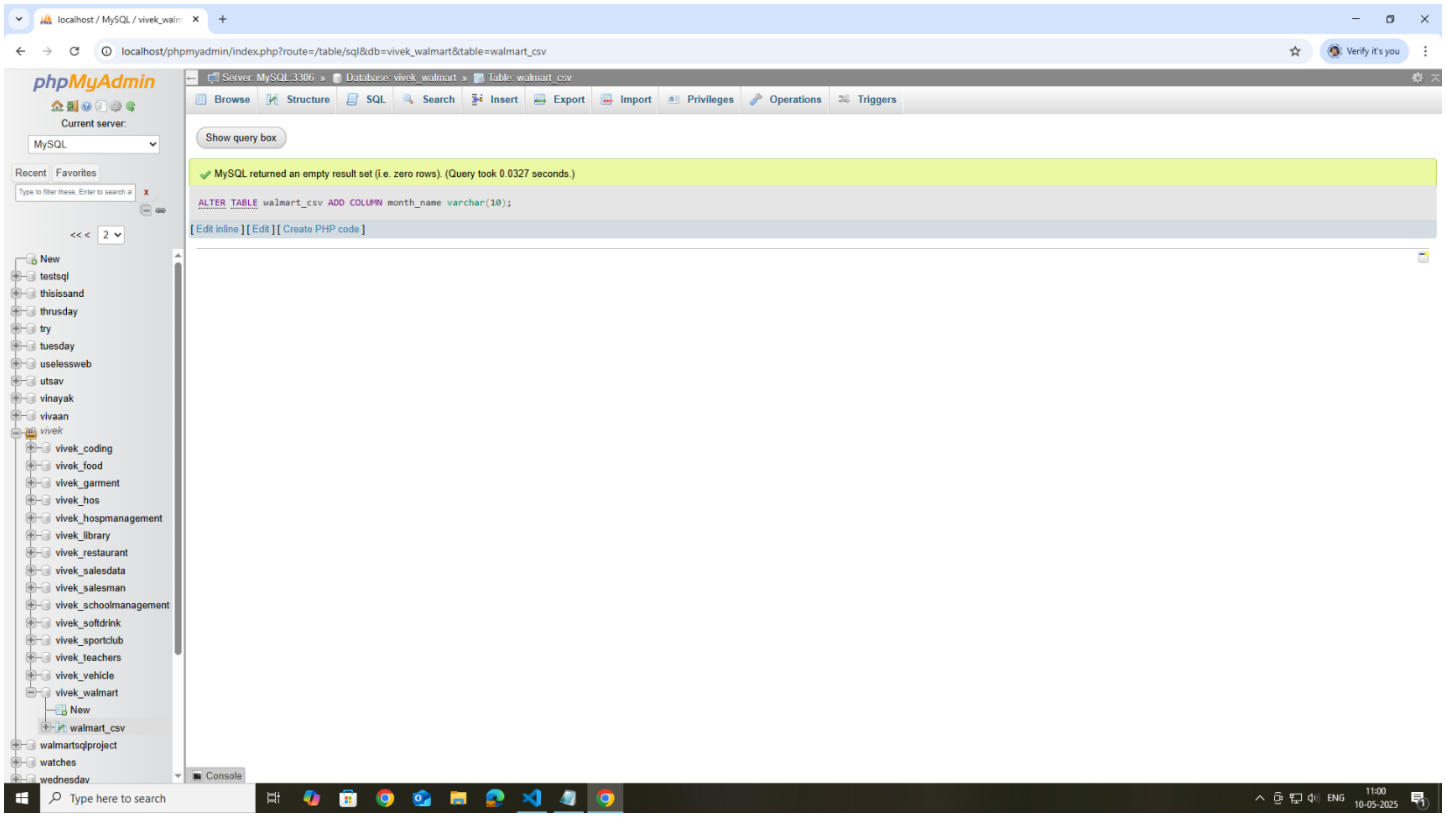


3) Add a new column named month_name that contains the extracted months of the year on which the given transaction took place (Jan, Feb, Mar). Help determine which month of the year has the most sales and profit.

now we need to add another new column named month_name in our dataset

→Your Query is :-

ALTER TABLE walmart_csv ADD COLUMN month_name varchar(10);

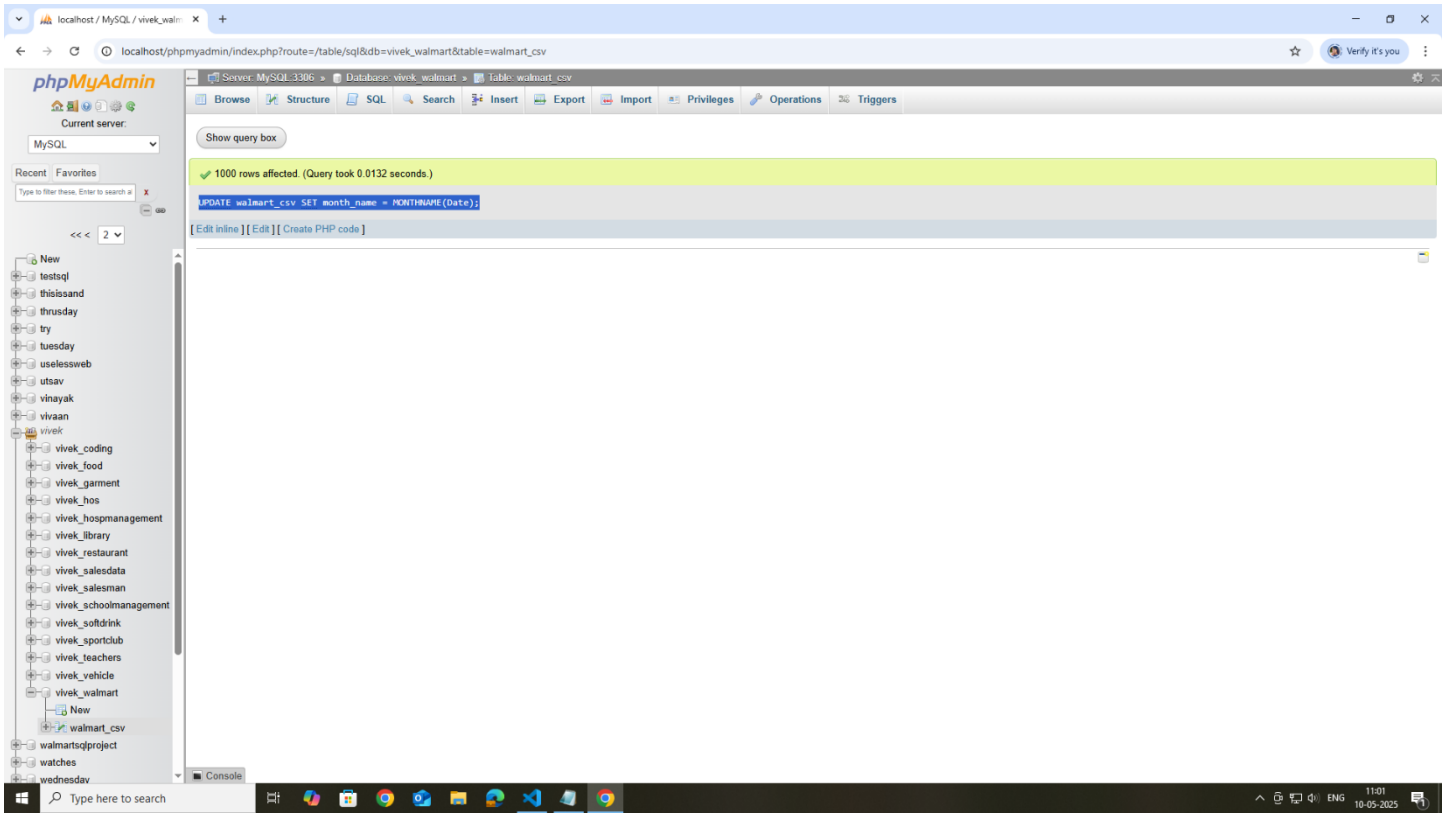


this will add a new column in our dataset, now we need to add data into the column

now this will add data in our dataset

→Second Query is :-

UPDATE walmart_csv SET month_name = MONTHNAME(Date);

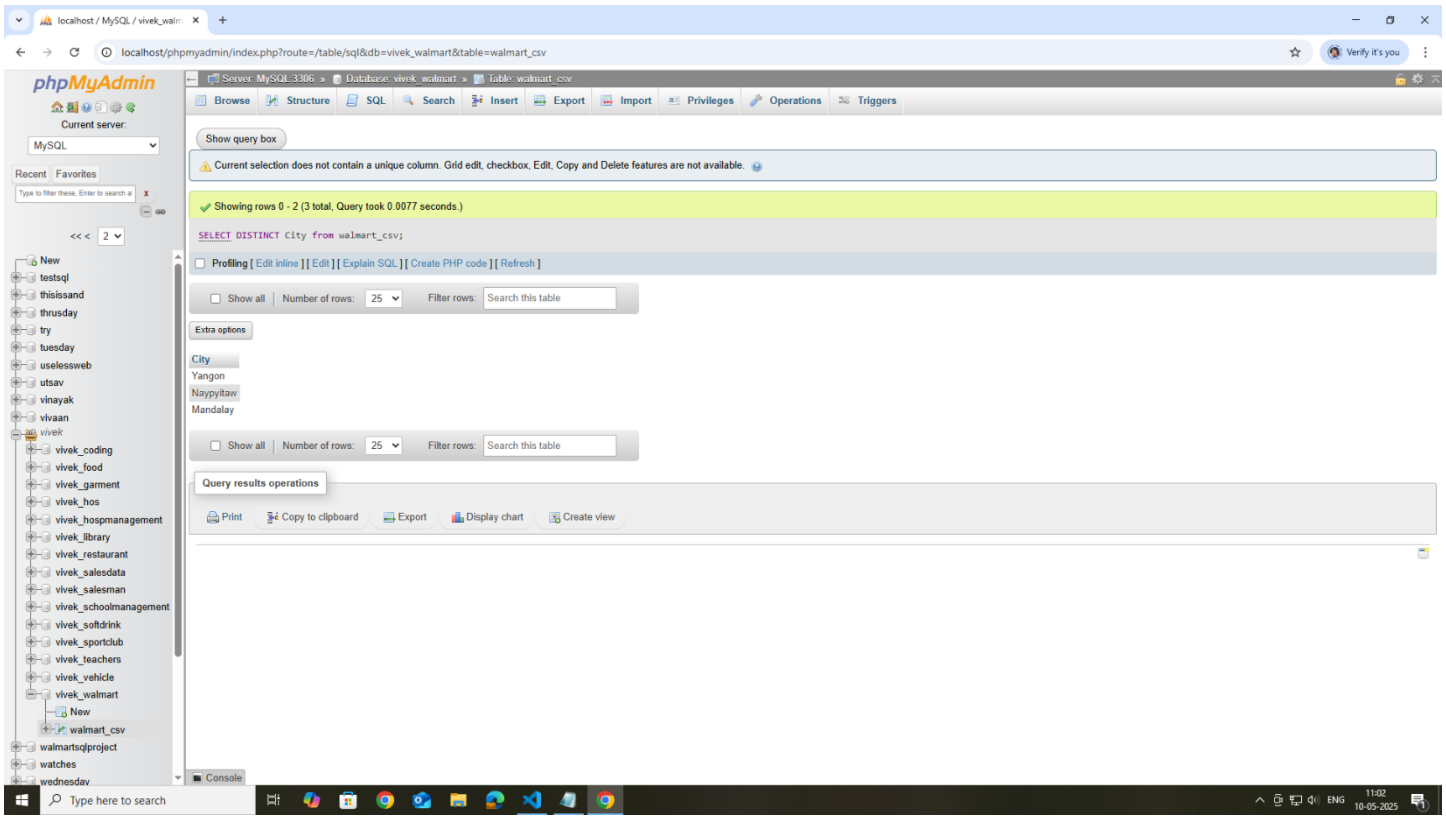


4) How many unique cities does the data have?

→Your Query is :-

SELECT DISTINCT City from walmart_csv;

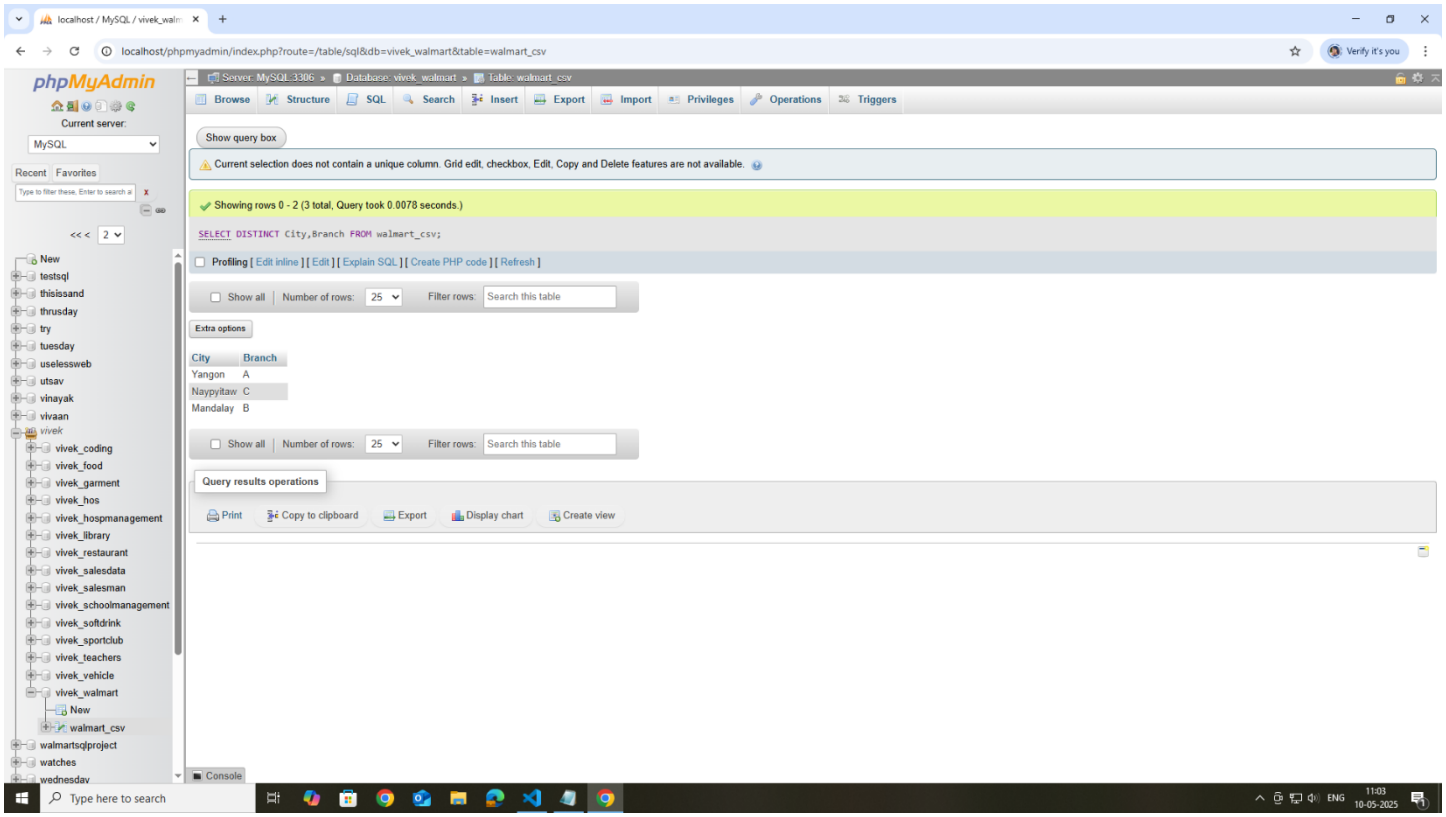
we have total 3 different cities in our dataset



5) In which city is each brand ?

→Your Query is :-

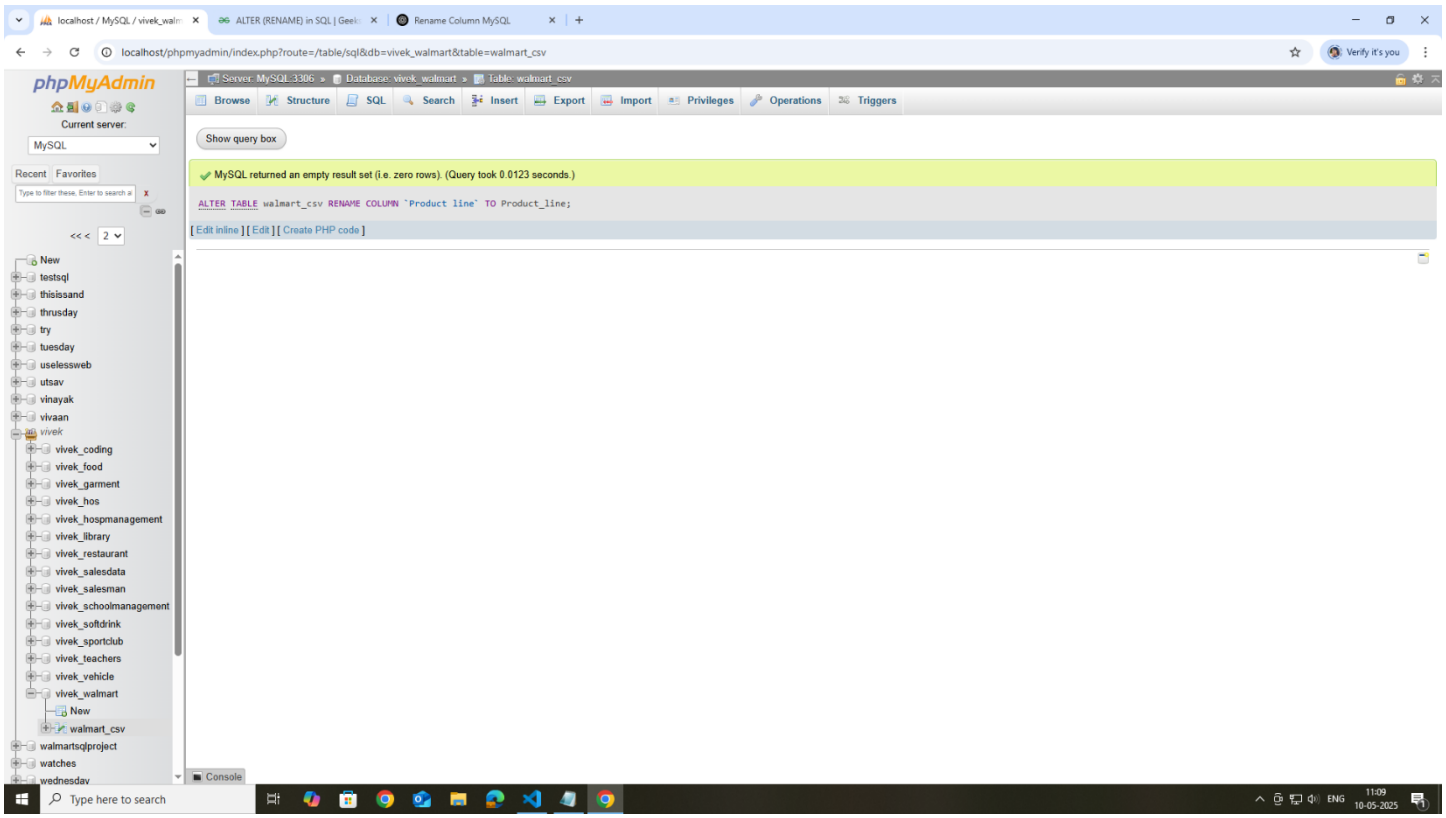
SELECT DISTINCT City,Branch FROM walmart_csv;



6) How many unique product lines does the data have ?

→Your Query is :-

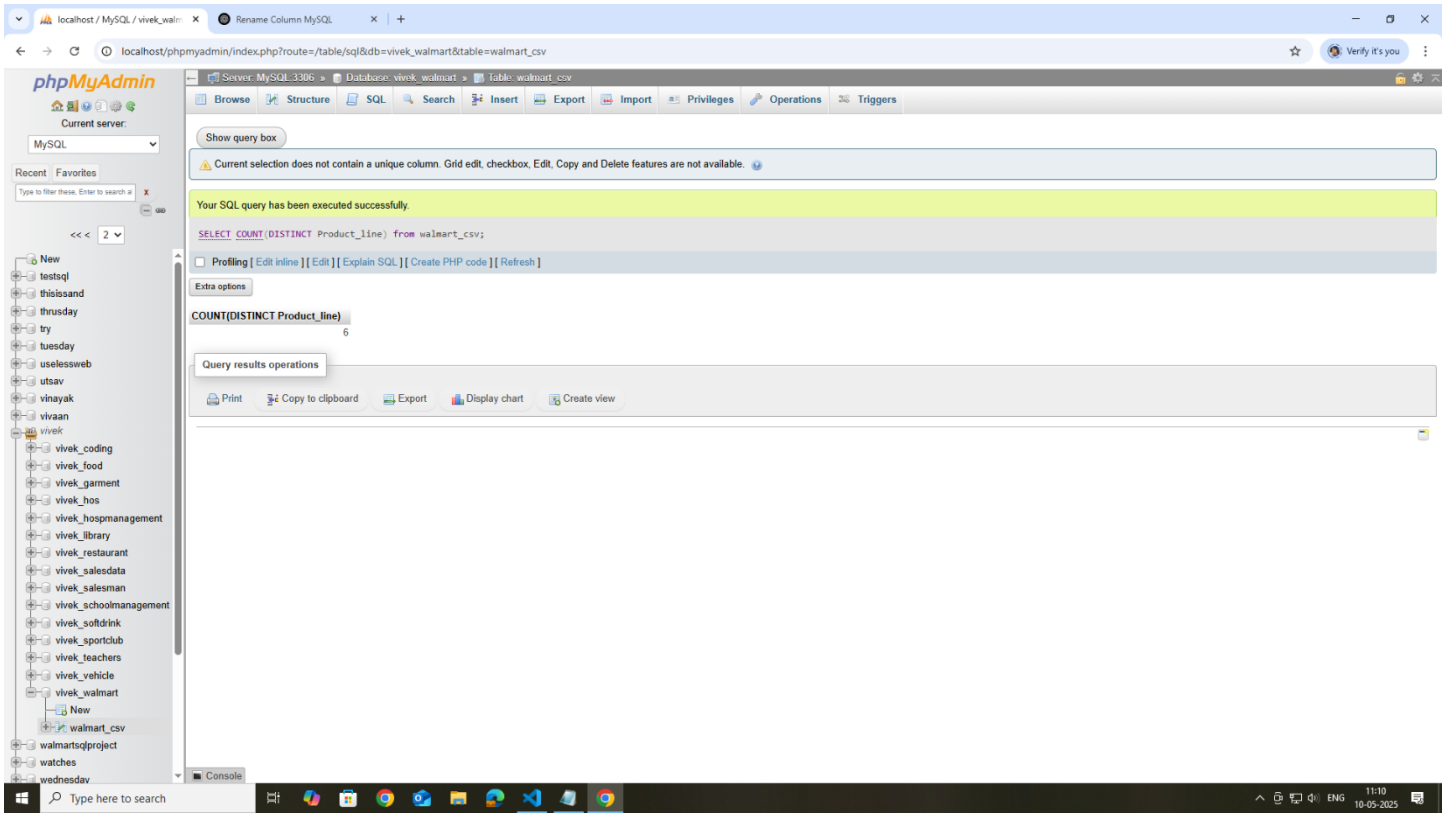
ALTER TABLE walmart_csv RENAME COLUMN `Product line` TO Product_line;



→Second Query is :-

SELECT COUNT(DISTINCT Product_line) from walmart_csv;

the business have total 6 unique product lines

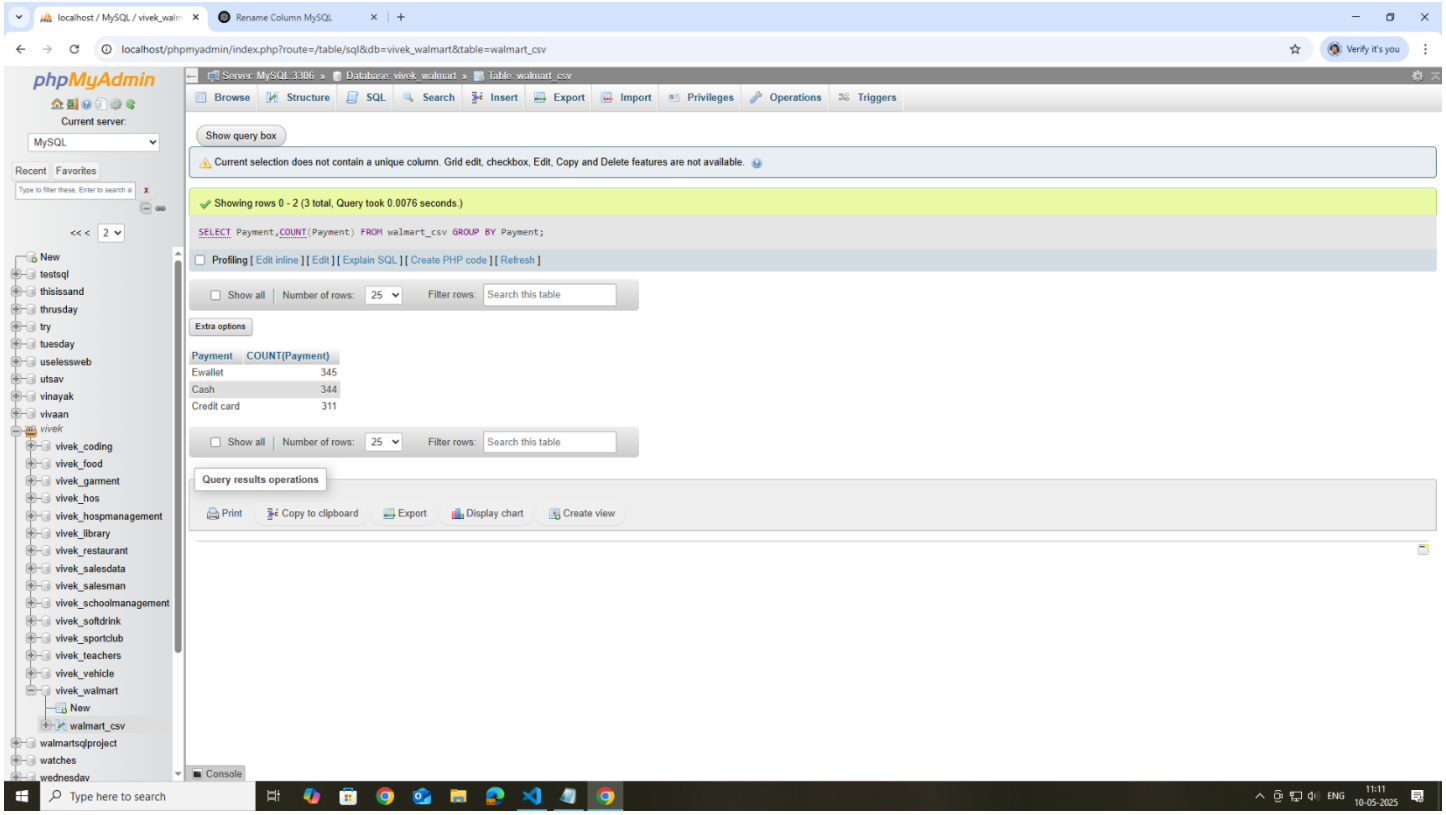


7) What is the most common payment method ?

as we can see that e wallet & cash are the most common payment method

→Your Query is :-

```
SELECT Payment,COUNT(Payment) FROM walmart_csv GROUP BY Payment;
```

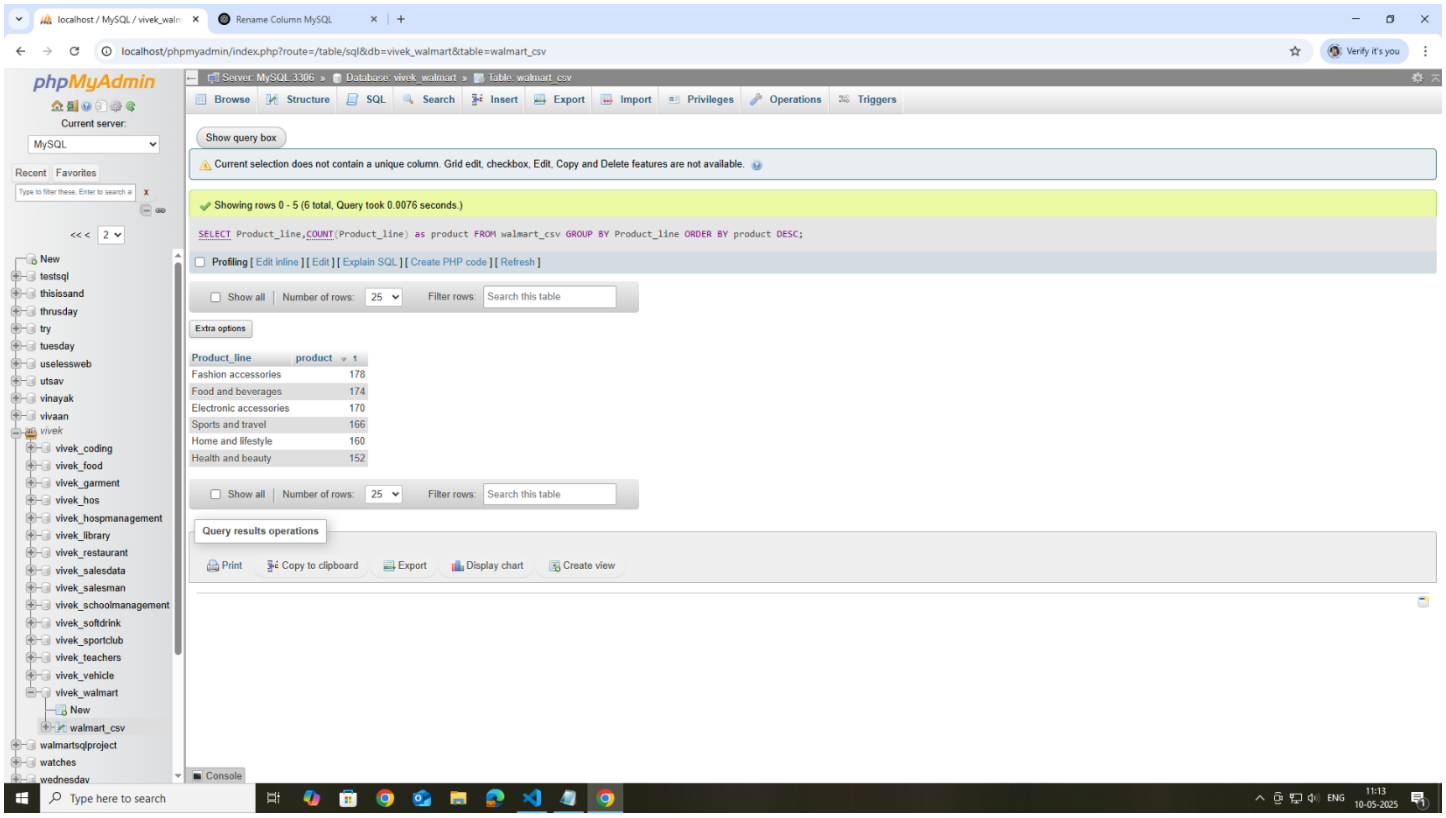


8) What is the most selling product line ?

As we can see that Fashion accessories & Food beverages are the most selling product line

→Your Query is :-

SELECT Product_line,COUNT(Product_line) as product FROM walmart_csv GROUP BY Product_line ORDER BY product DESC;

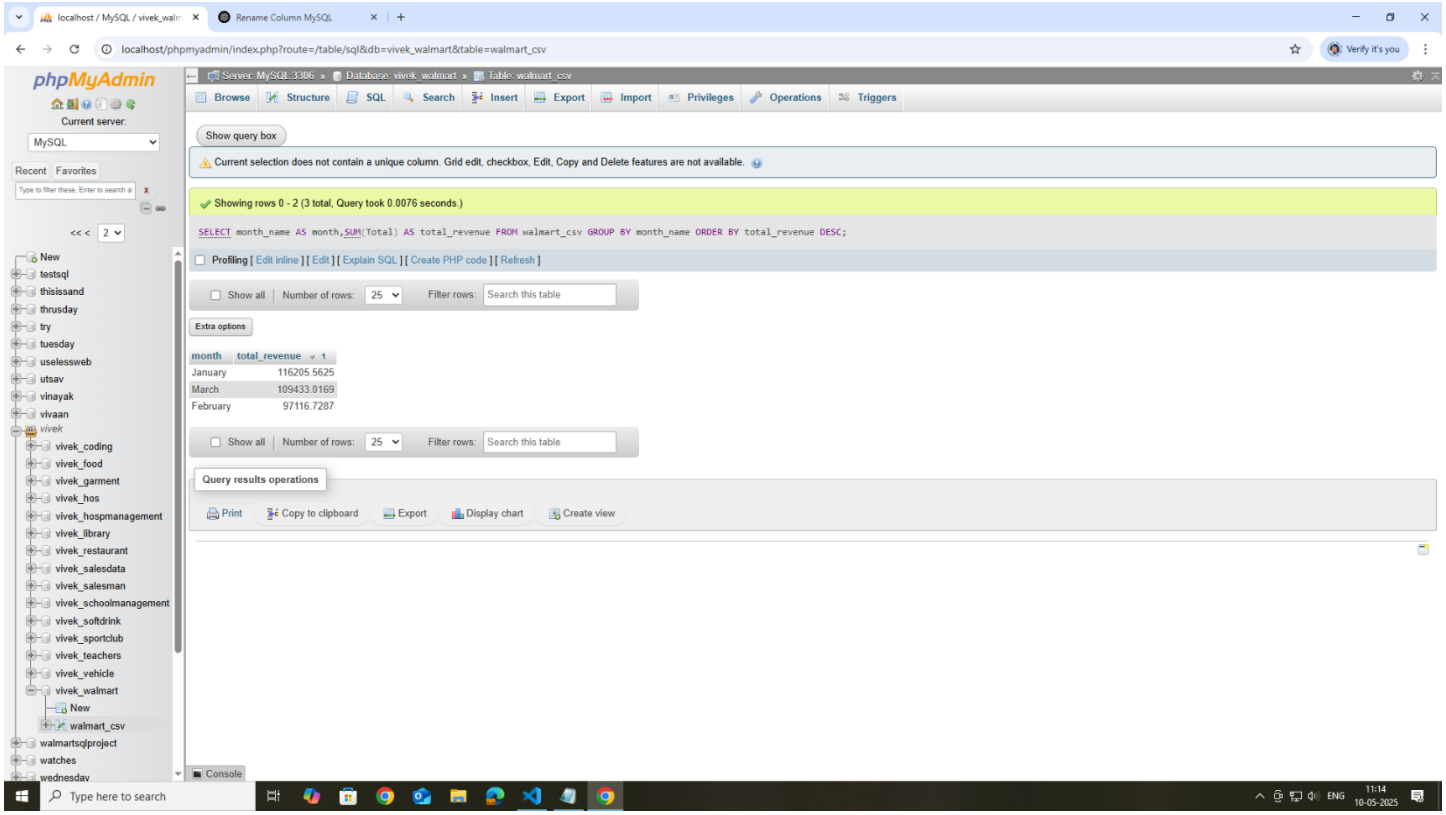


9) What is the total revenue by the month ?

as we can see that janauary had the highest revenue

→Your Query is :-

SELECT month_name AS month,SUM(Total) AS total_revenue FROM walmart_csv GROUP BY month_name ORDER BY total_revenue DESC;

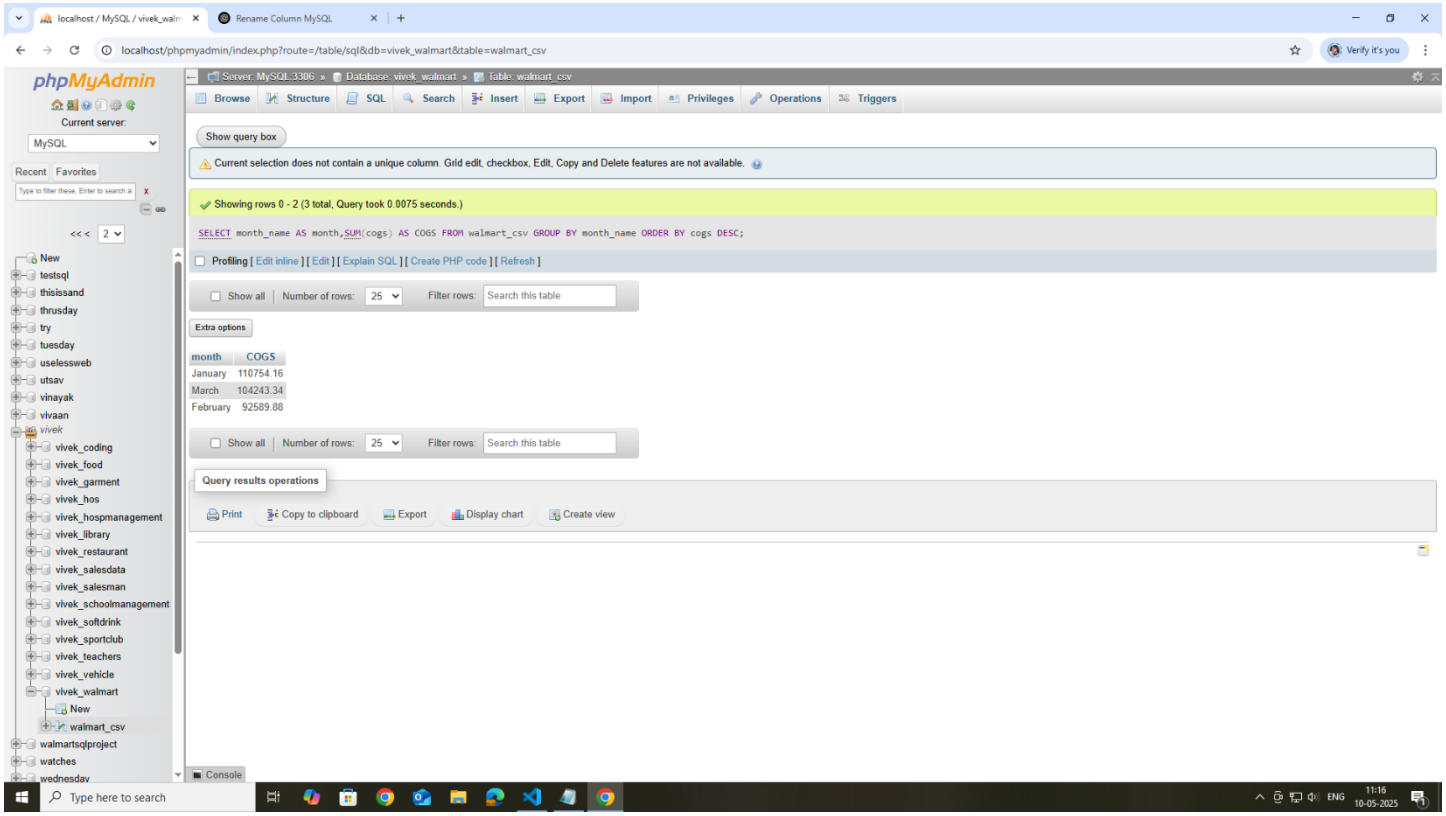


10) What month had the largest COGS ?

As you can see that january month had the highest cogs

→Your Query is :-

SELECT month_name AS month,SUM(cogs) AS COGS FROM walmart_csv GROUP BY month_name ORDER BY cogs DESC;

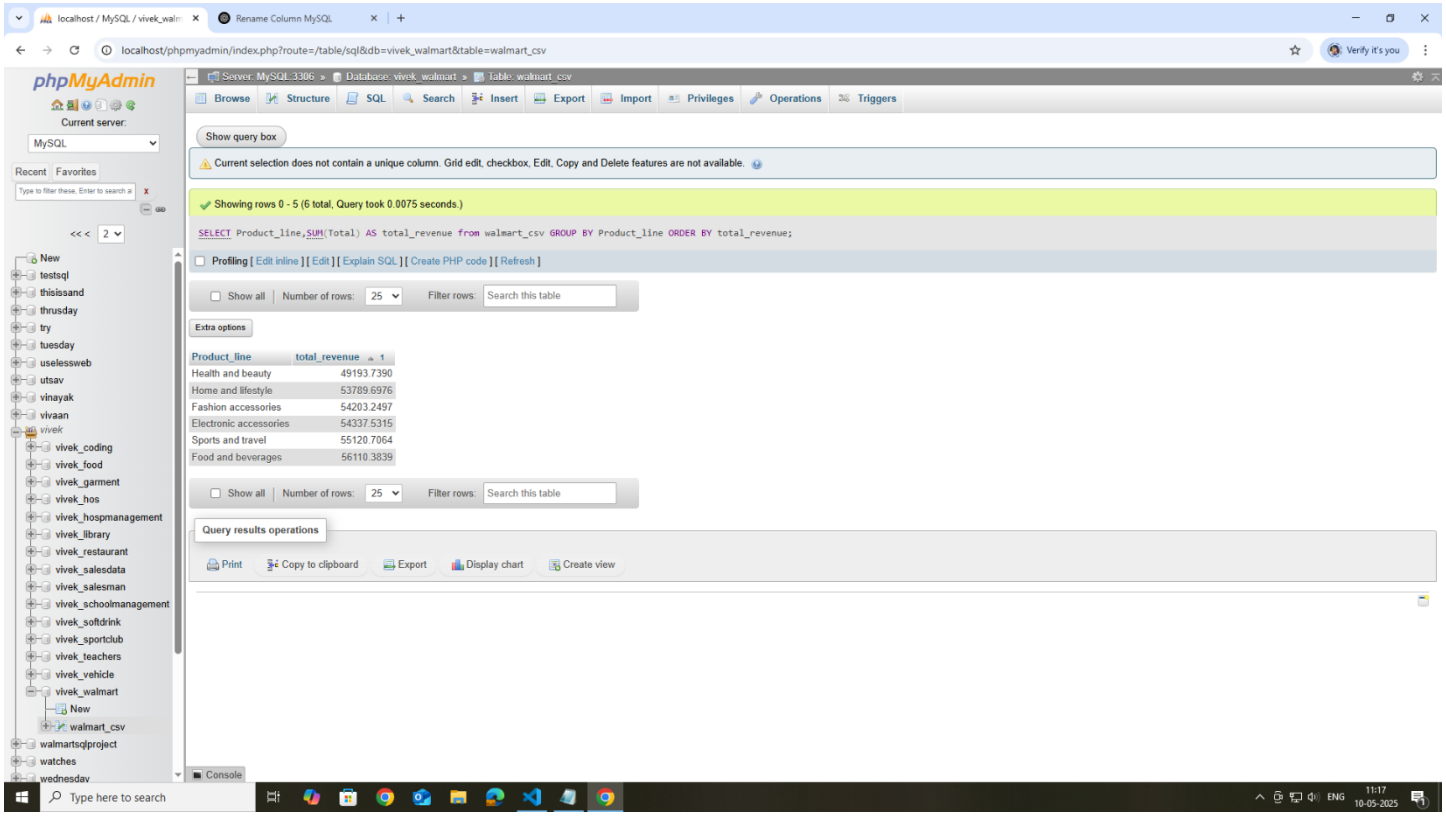


11) What product line had the largest revenue ?

As from the data we can say that food & beverages had the highest revenue.

→Your Query is :-

SELECT Product_line,SUM(Total) AS total_revenue from walmart_csv GROUP BY Product_line ORDER BY total_revenue;

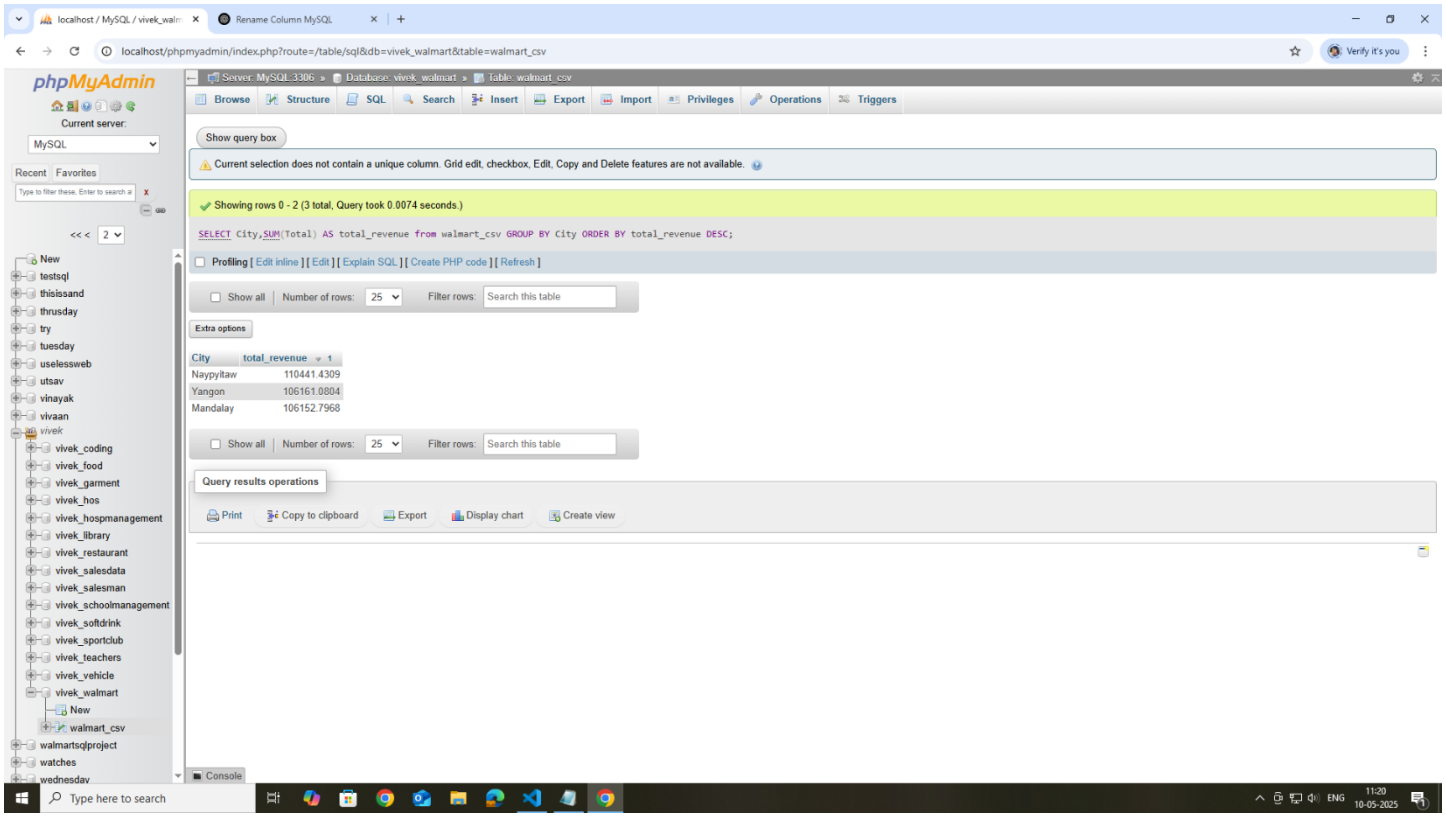


12) What is the city with the largest revenue ?

As we can see that city naypyitaw had the largest revenue.

→Your Query is :-

SELECT City,SUM(Total) AS total_revenue from walmart_csv GROUP BY City ORDER BY total_revenue DESC;

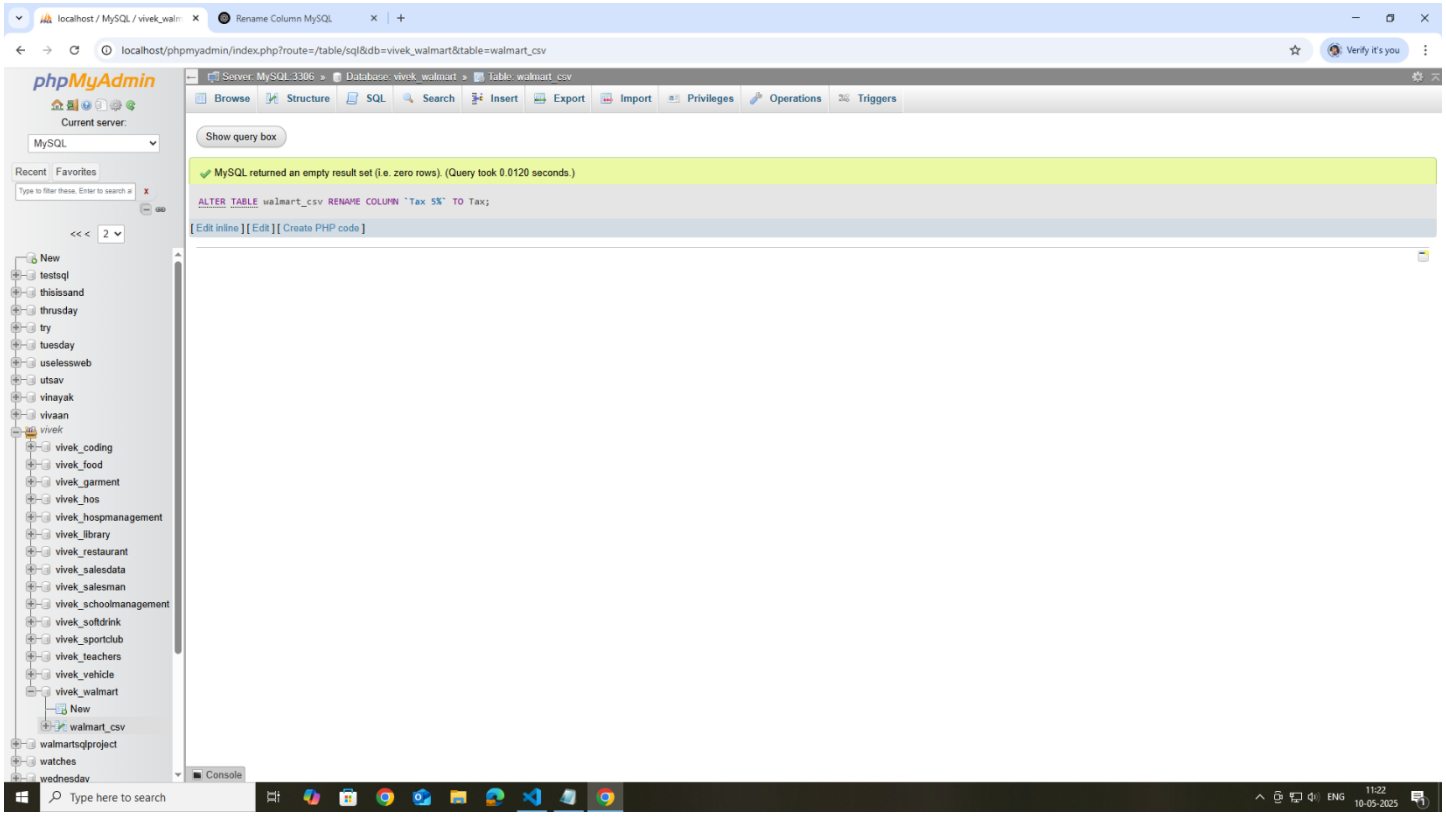


13) What product lines had the largest VAT (TAX) ?

As we can see that food & beverages had the largest tax

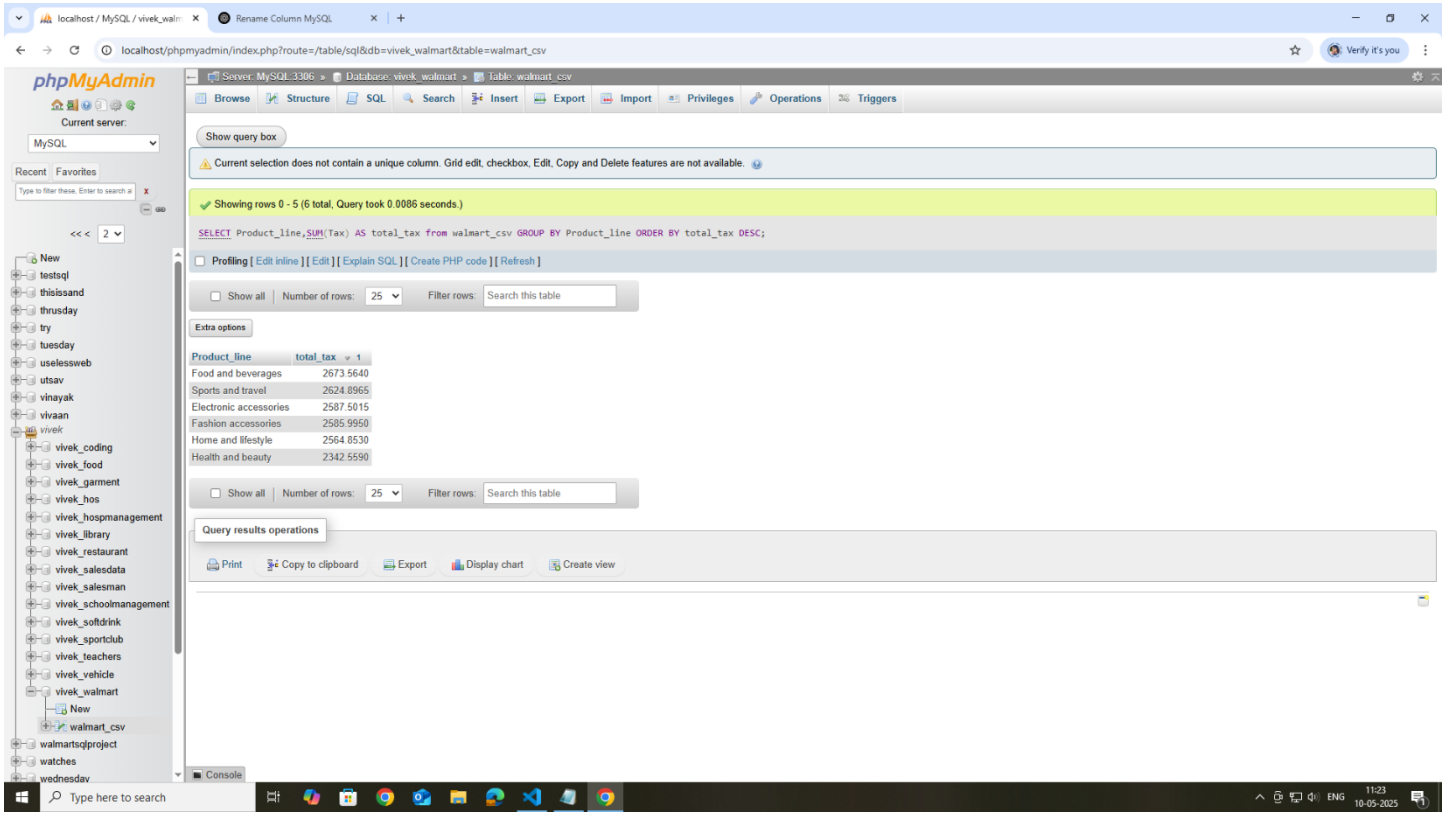
→Your Query is :-

ALTER TABLE walmart_csv RENAME COLUMN `Tax 5%` TO Tax;



→Second Query is :-

SELECT Product_line,SUM(Tax) AS total_tax from walmart_csv GROUP BY Product_line ORDER BY total_tax DESC;

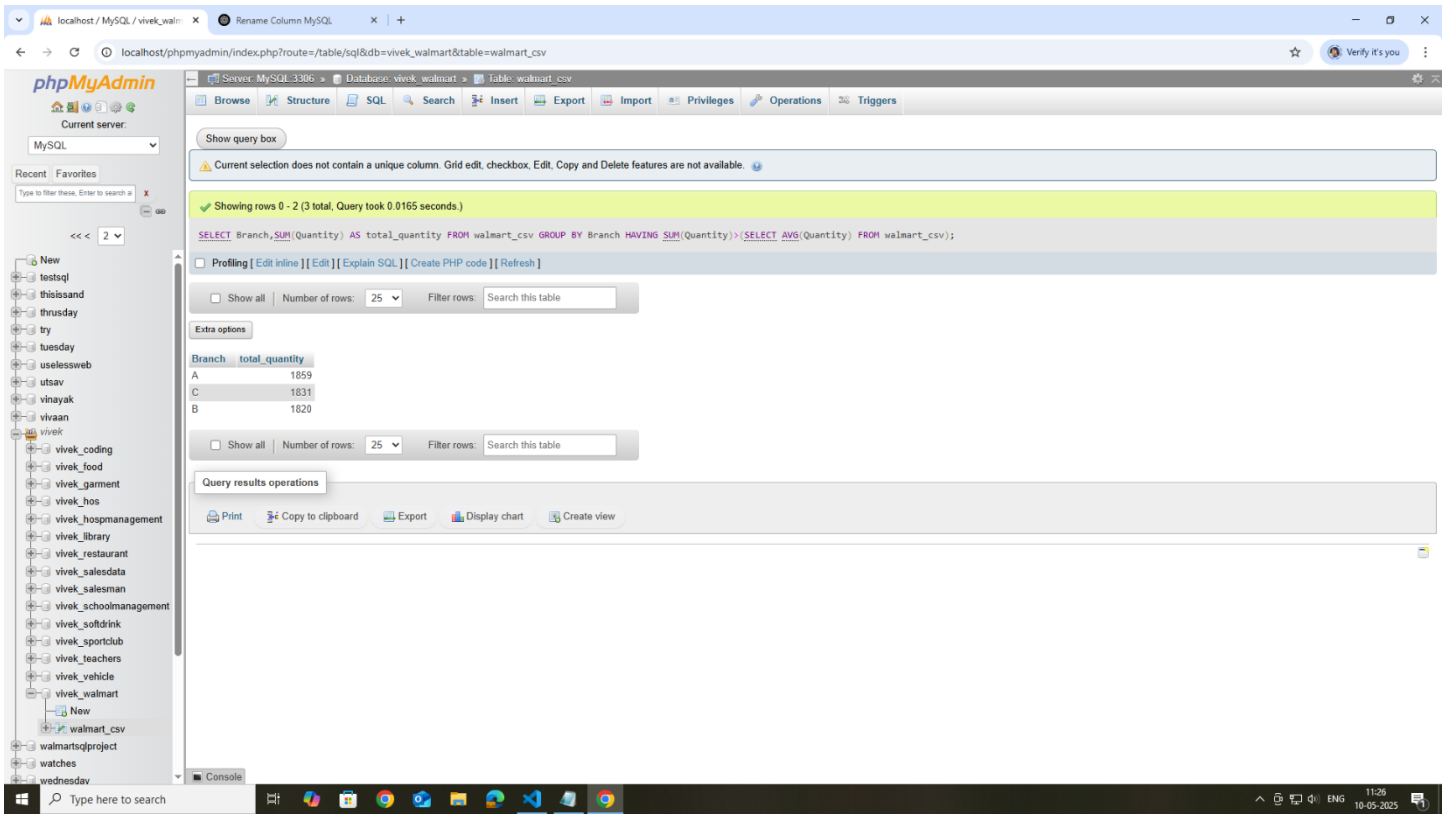


14) Which branch sold more products than average product sold ?

As we can see that branch A had sold more products than average product sold

→Your Query is :-

SELECT Branch,SUM(Quantity) AS total_quantity FROM walmart_csv GROUP BY Branch HAVING SUM(Quantity)>(SELECT AVG(Quantity) FROM walmart_csv);

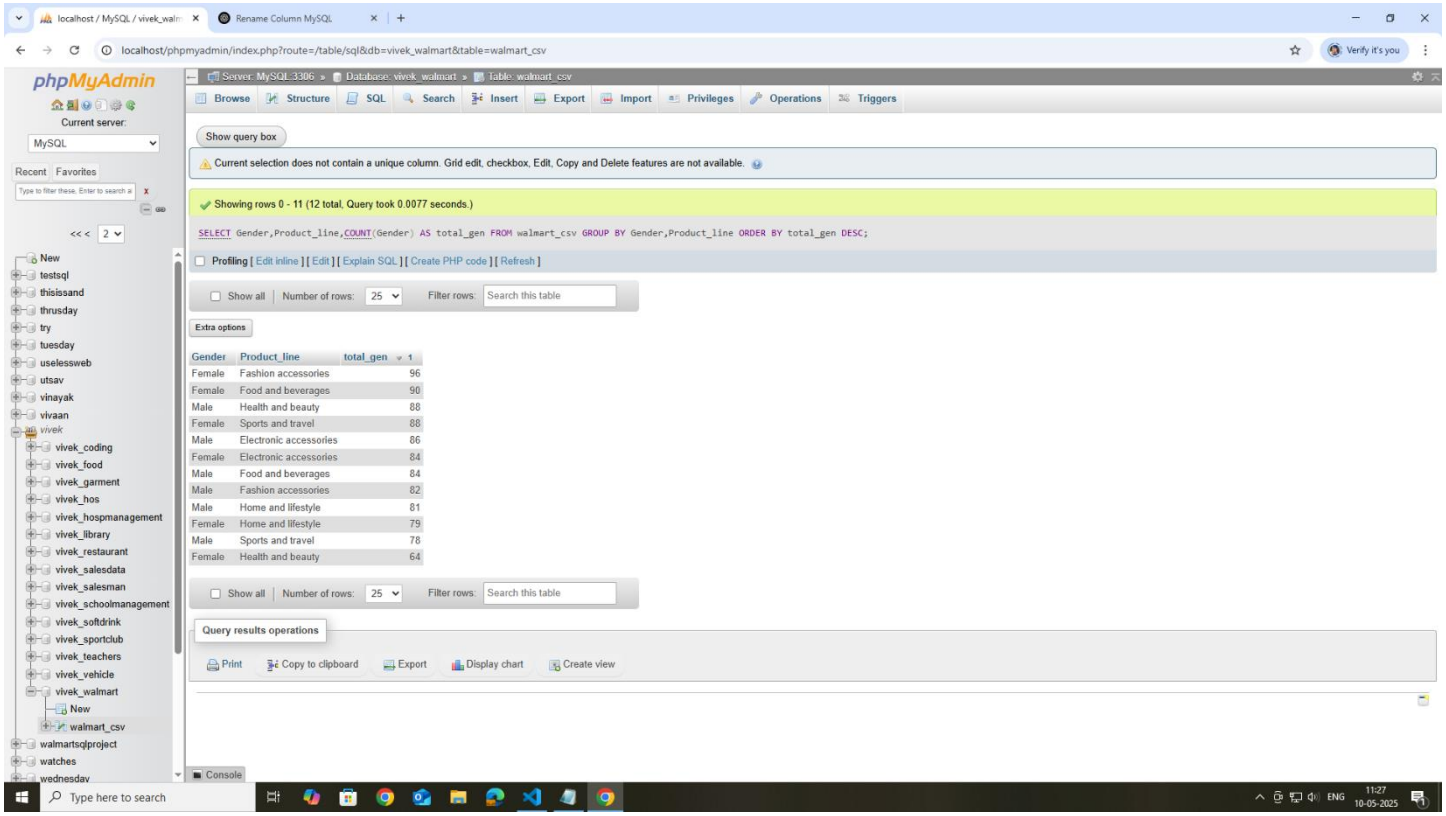


15) What is the most common product line by gender ?

As we can see that Gender female & product line fashion accessories have the highest count

→Your Query is :-

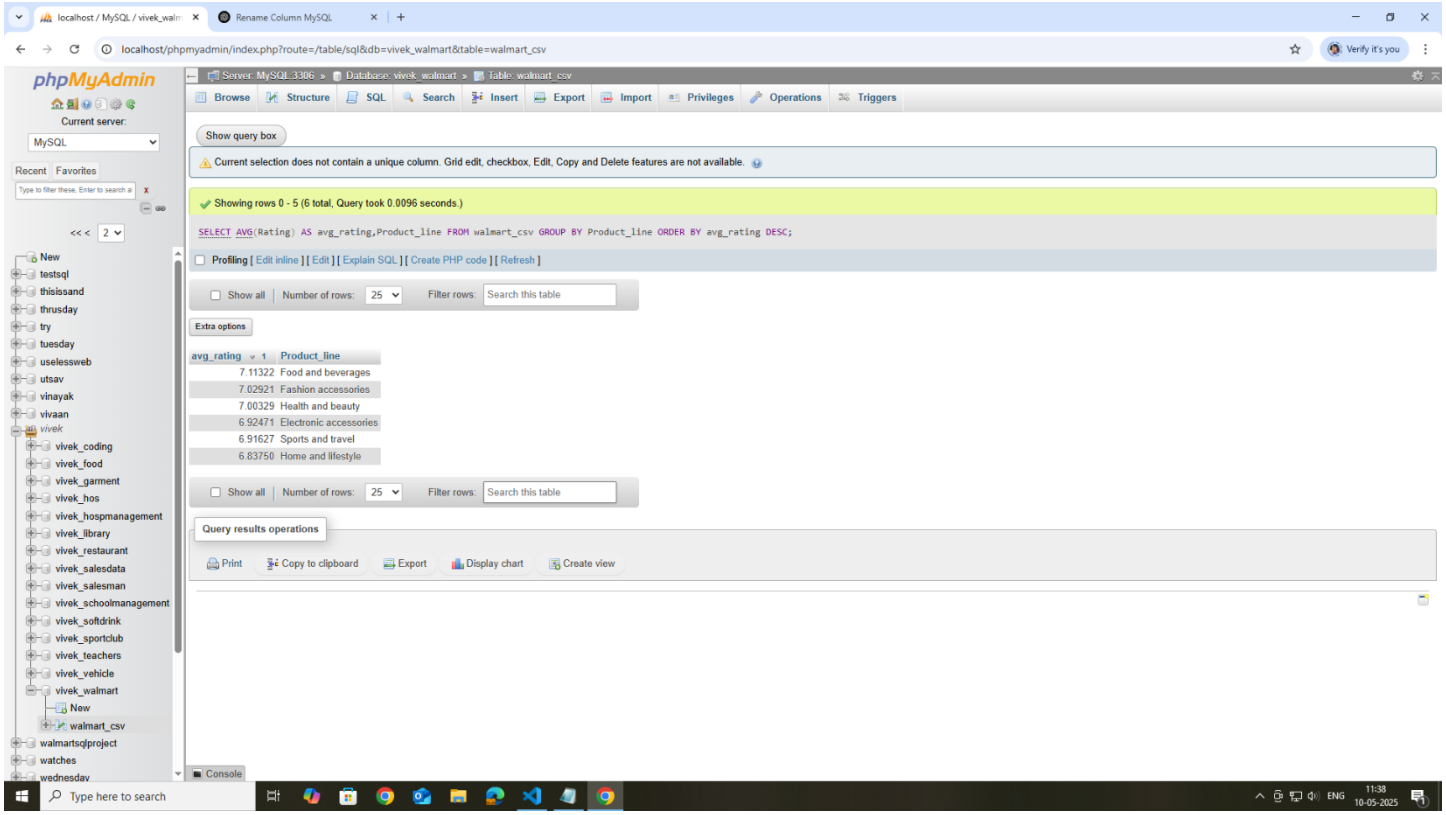
SELECT Gender,Product_line,COUNT(Gender) AS total_gen FROM walmart_csv GROUP BY Gender,Product_line ORDER BY total_gen DESC;



16) What is the average rating of each product line ?

→Your Query is :-

SELECT AVG(Rating) AS avg_rating,Product_line FROM walmart_csv GROUP BY Product_line ORDER BY avg_rating DESC;

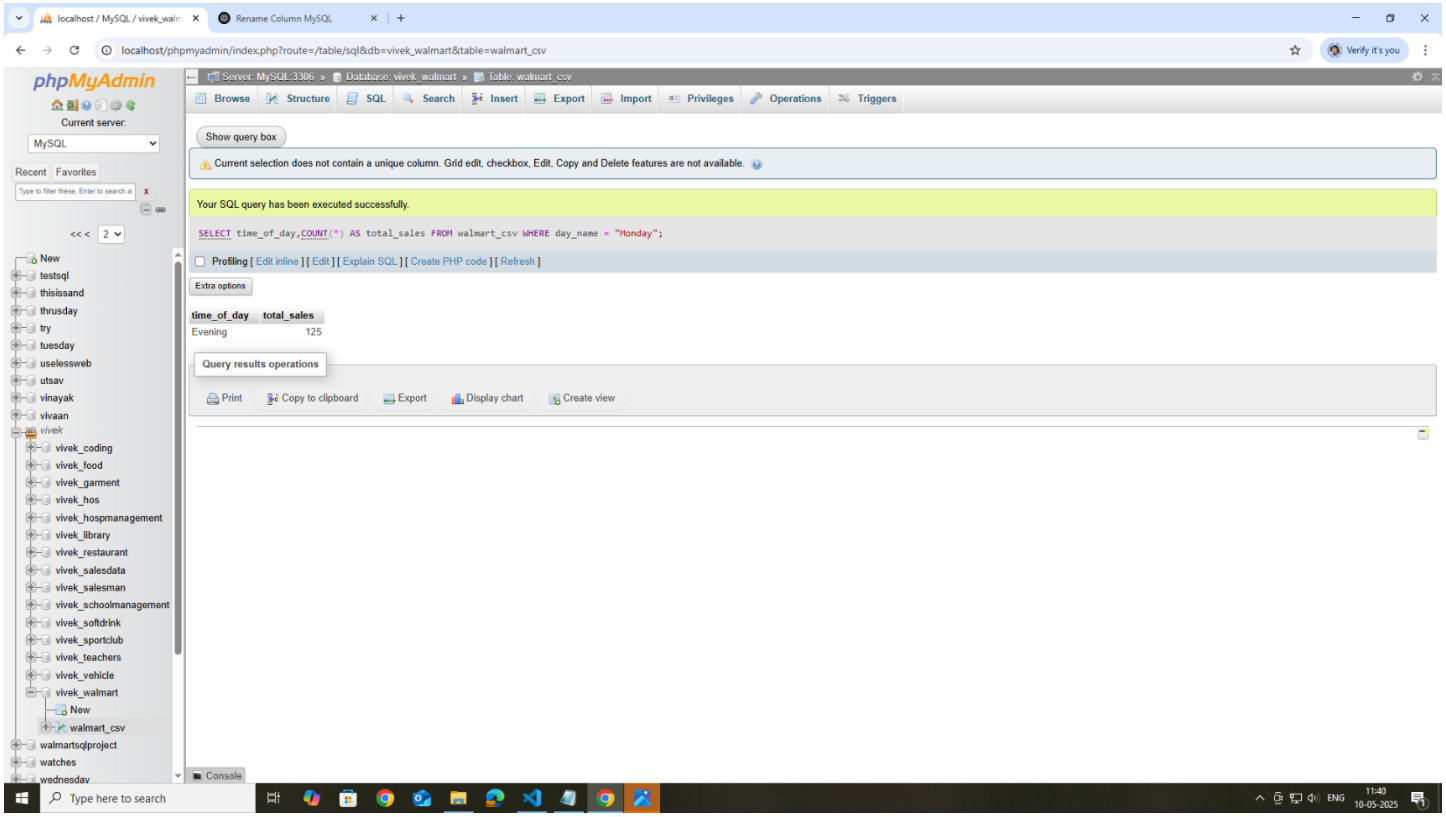


17) Number of sales made in each time of the day per weekday ?

we had counted for monday, you can change the day as per your requirement

→Your Query is :-

SELECT time_of_day,COUNT(*) AS total_sales FROM walmart_csv WHERE day_name = "Monday"

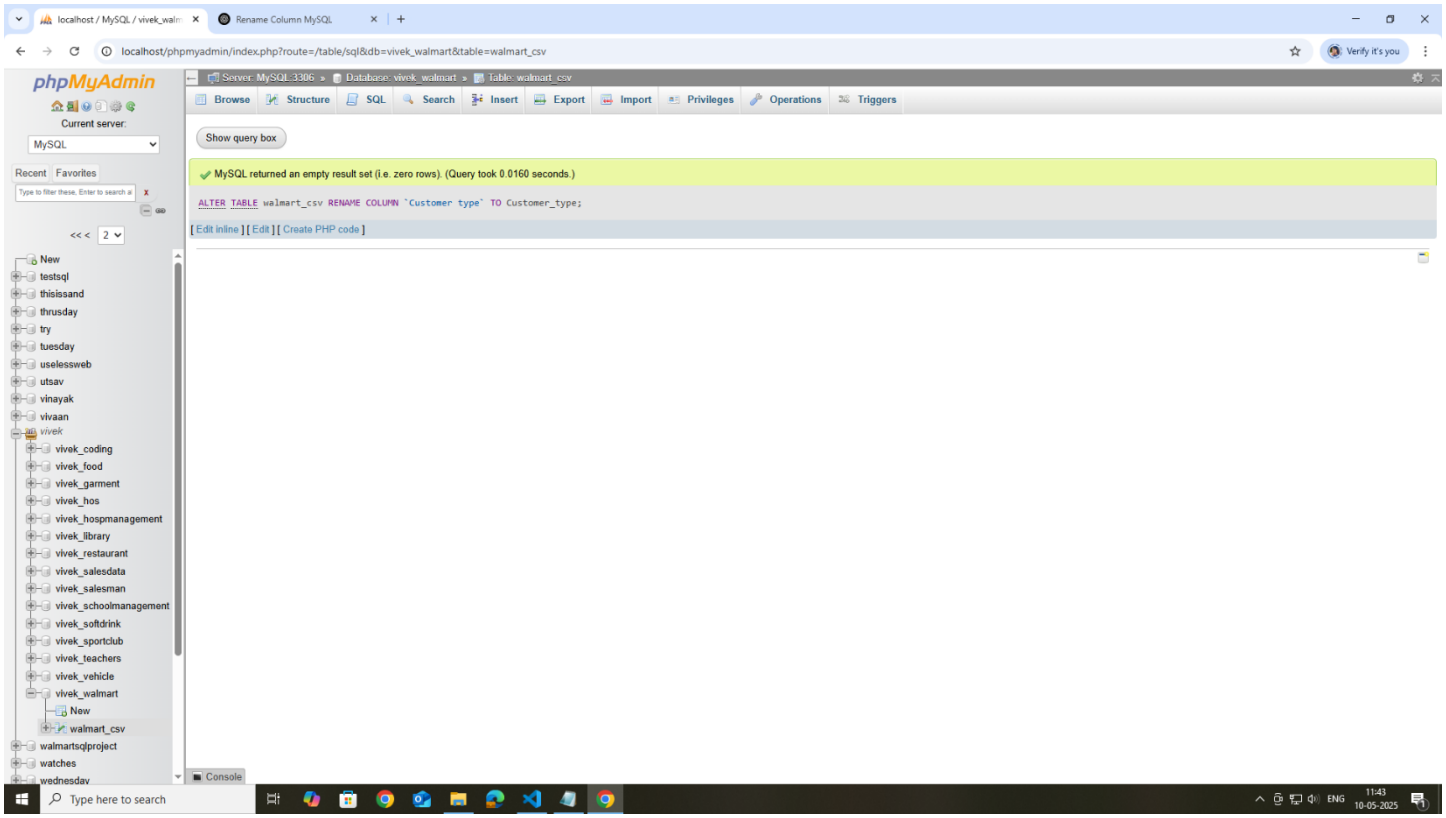


18) Which of the customer types brings the most revenue ?

As we can see that customer type member brings the most revenue

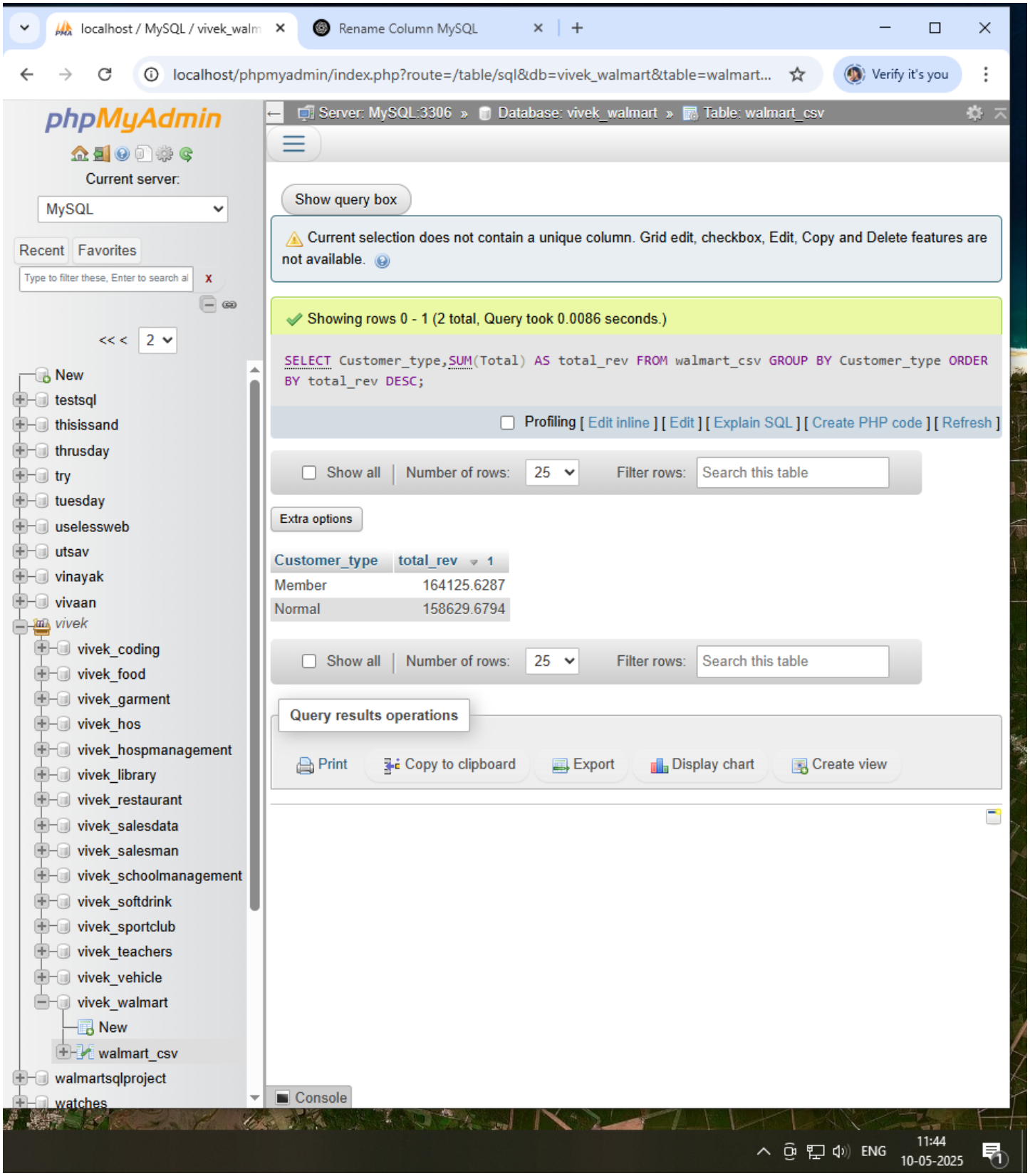
→Your Query is :-

ALTER TABLE walmart_csv RENAME COLUMN `Customer type` TO Customer_type;



→Second Query is :-

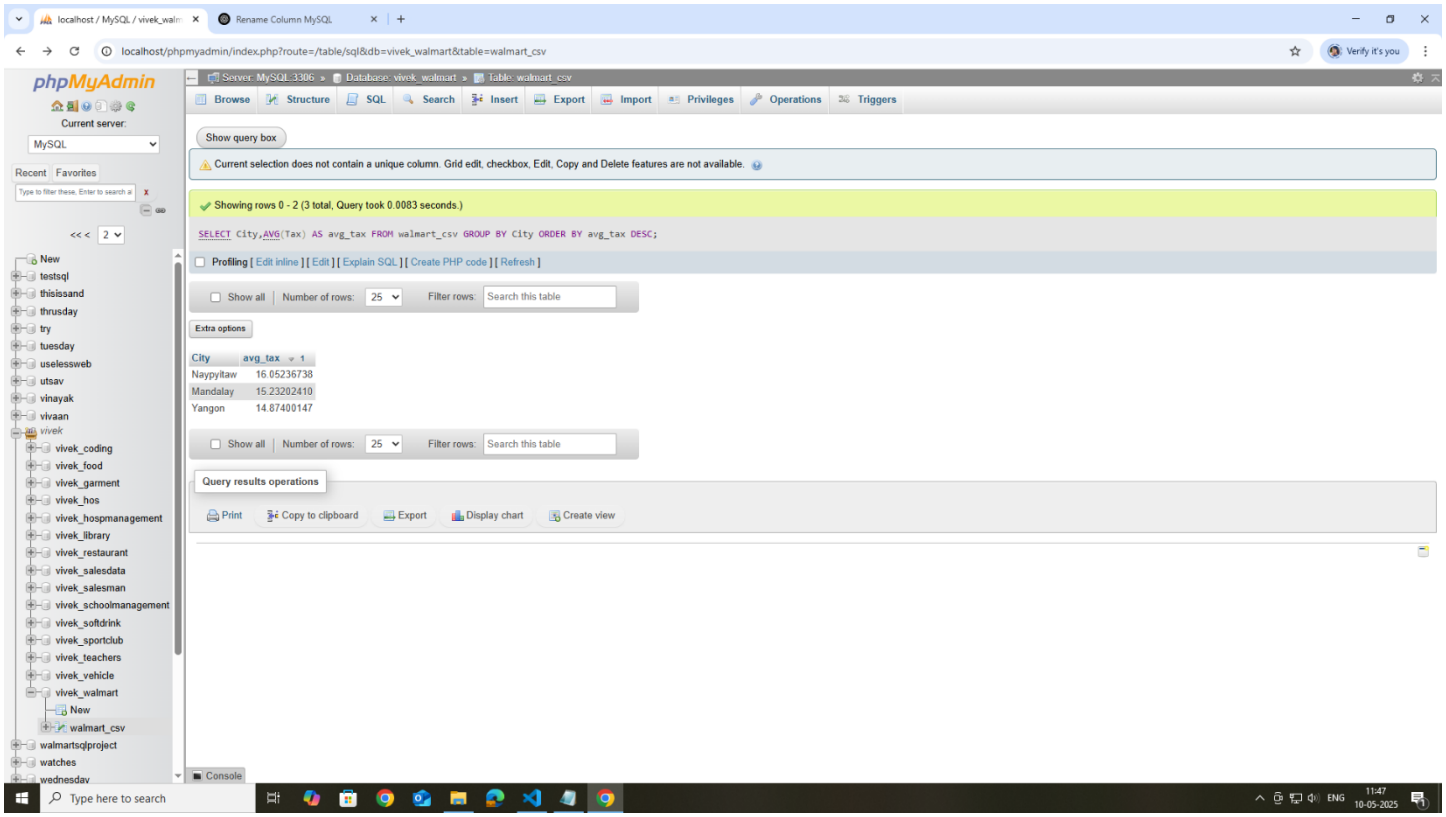
SELECT Customer_type,SUM(Total) AS total_rev FROM walmart_csv GROUP BY Customer type ORDER BY total_rev DESC;



19) Which city has the largest tax/vat percent ?

→Your Query is :-

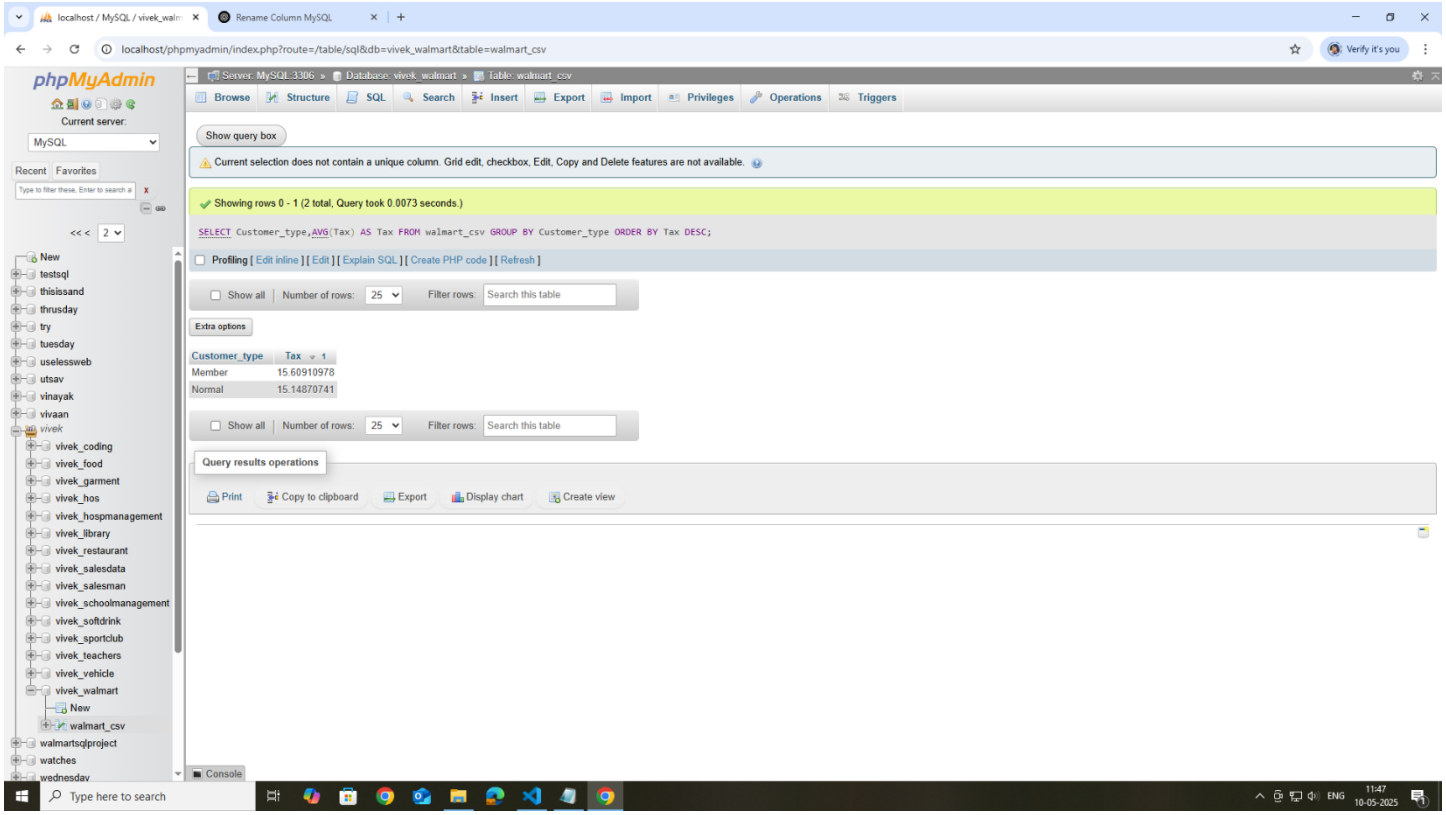
SELECT City,AVG(Tax) AS avg_tax FROM walmart_csv GROUP BY City ORDER BY avg_tax DESC;



20) Which customer type pays the most tax/vat ?

→Your Query is :-

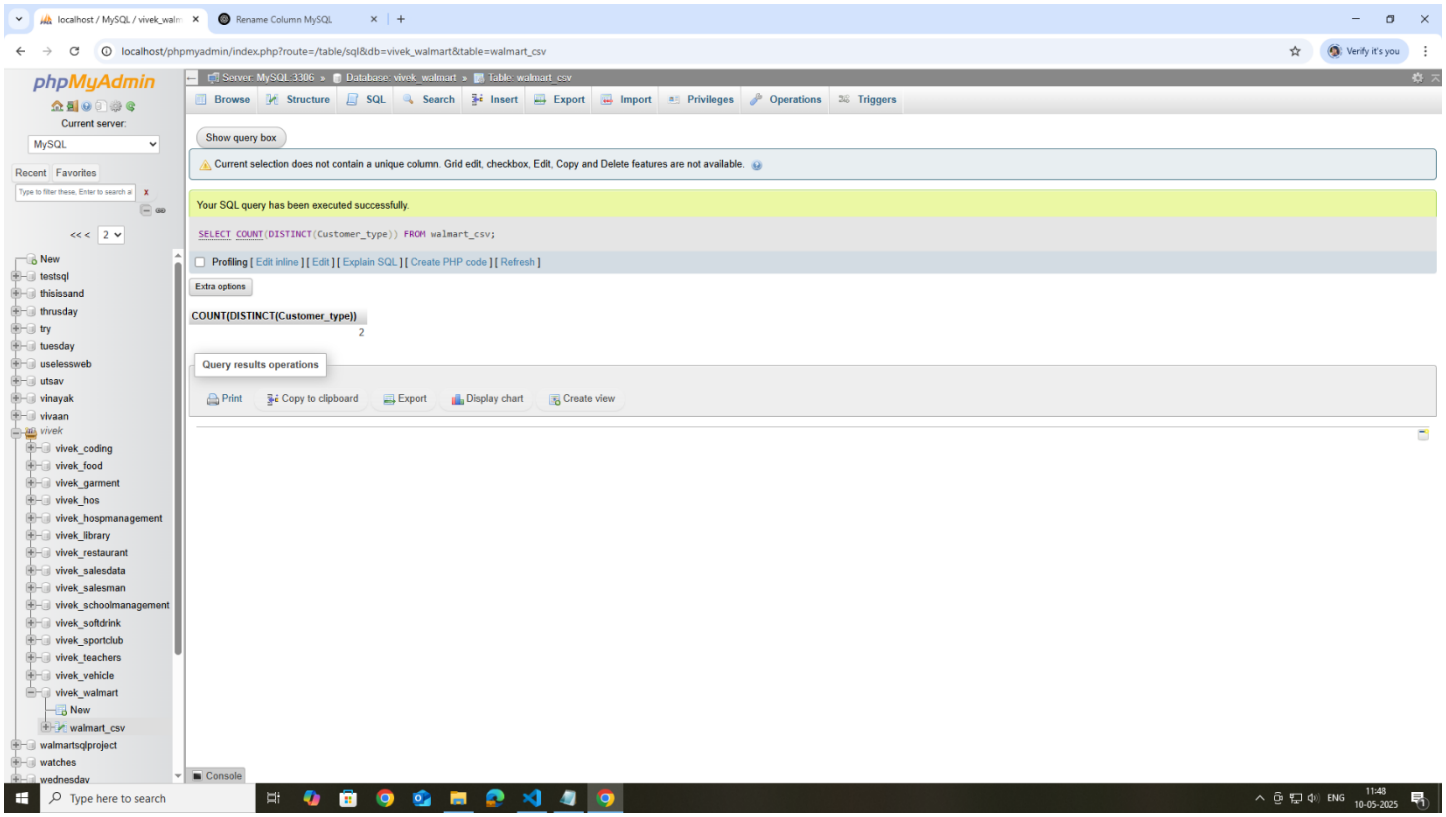
SELECT Customer_type,AVG(Tax) AS Tax FROM walmart_csv GROUP BY Customer_type ORDER BY Tax DESC;



21) How many unique customer types does the data have?

→Your Query is :-

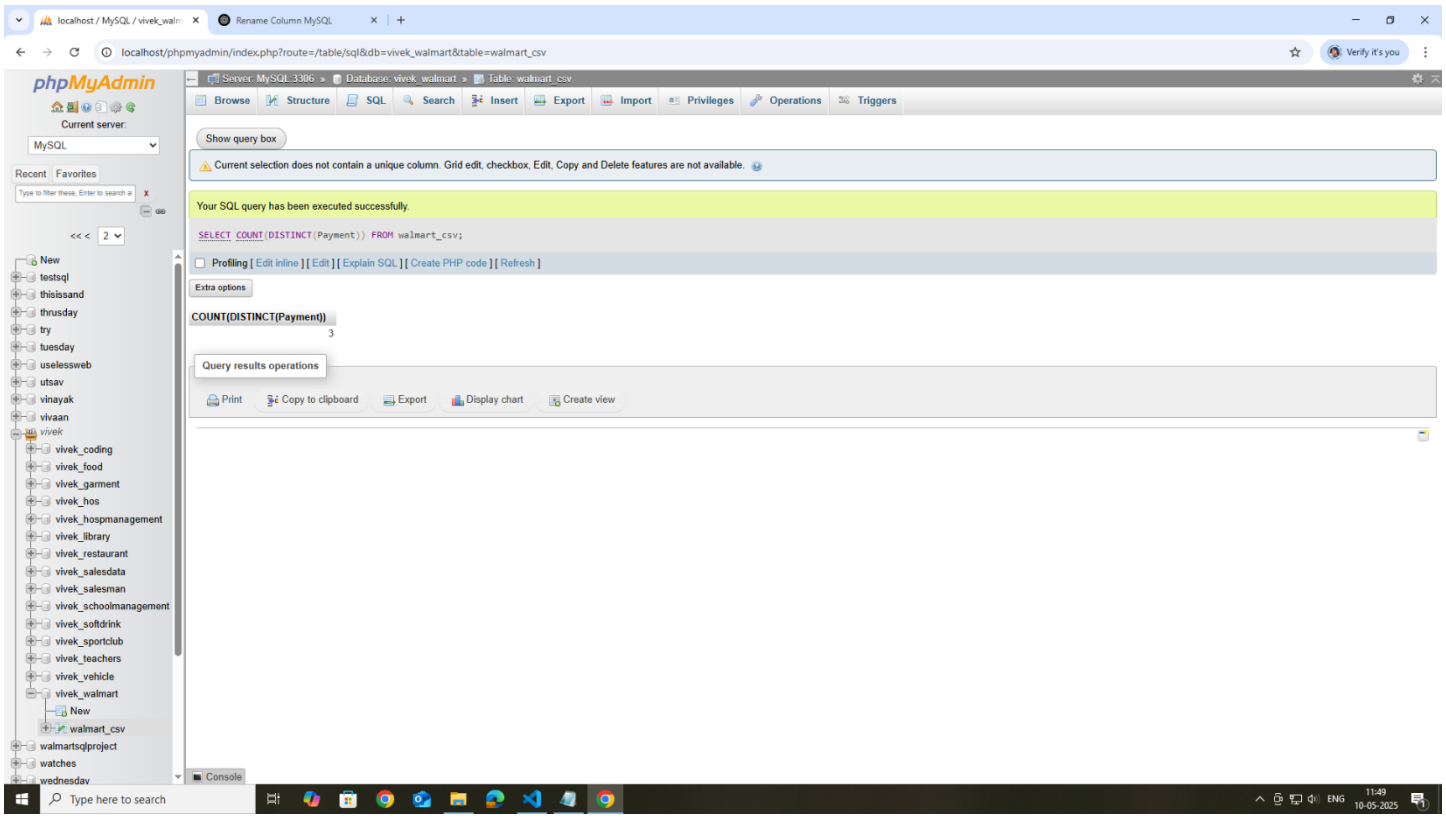
SELECT COUNT(DISTINCT(Customer_type)) FROM walmart_csv;



22) How many unique payment methods does the data have?

→Your Query is :-

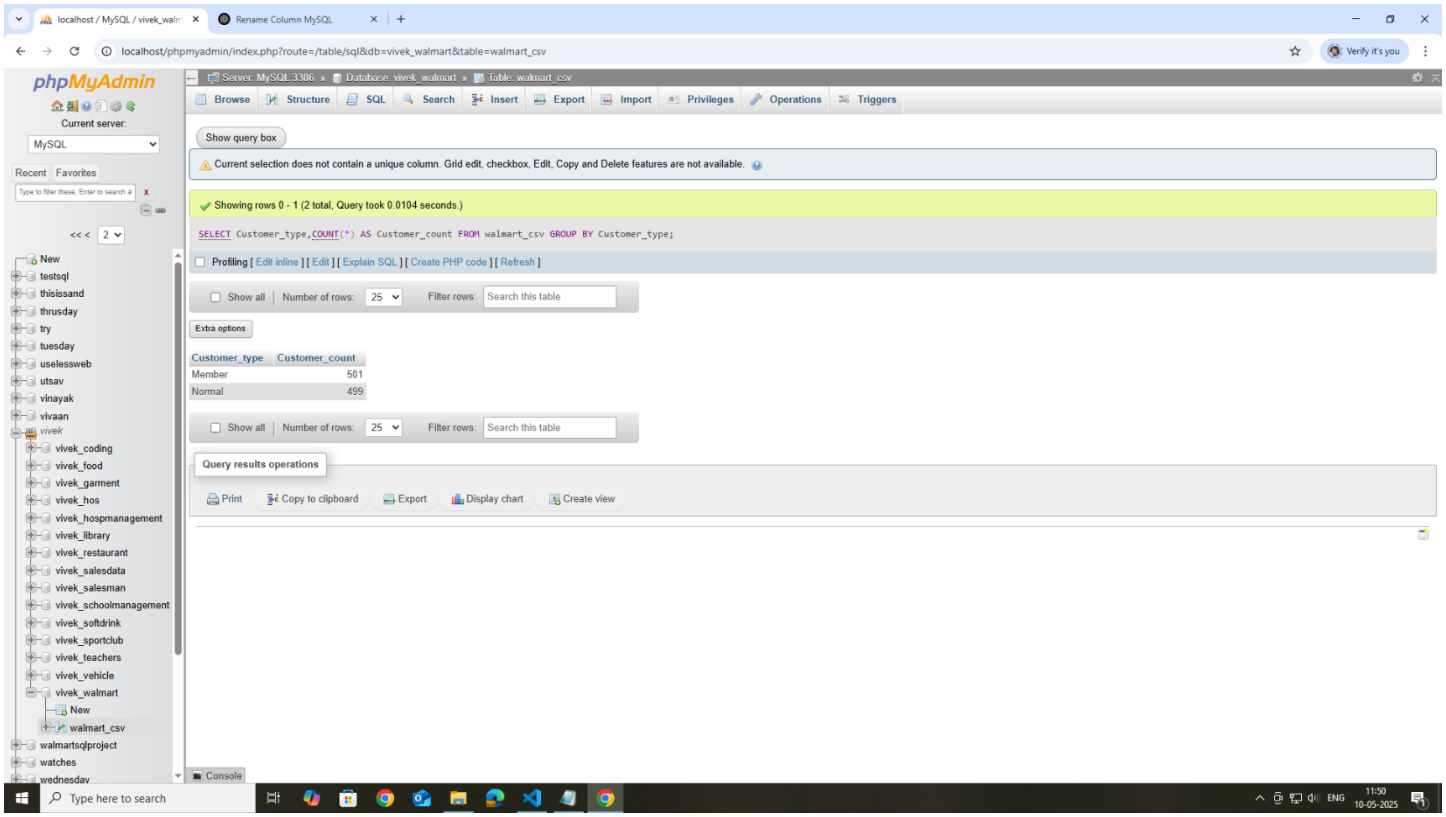
SELECT COUNT(DISTINCT(Payment)) FROM walmart_csv;



23) What is the most common customer type?

→Your Query is :-

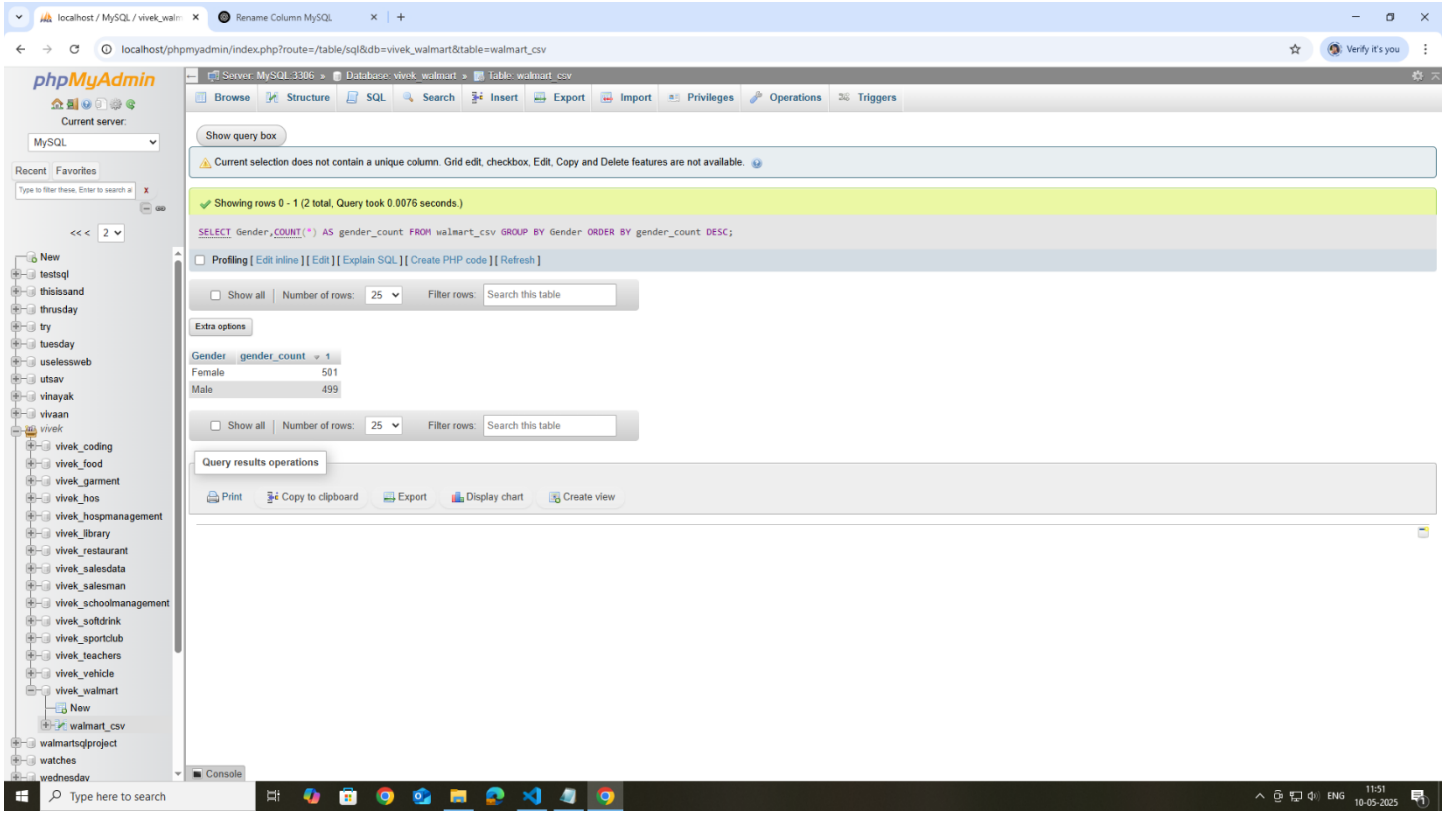
SELECT Customer_type,COUNT(*) AS Customer_count FROM walmart_csv GROUP BY Customer_type;



24) What is the gender of most of the customers?

→Your Query is :-

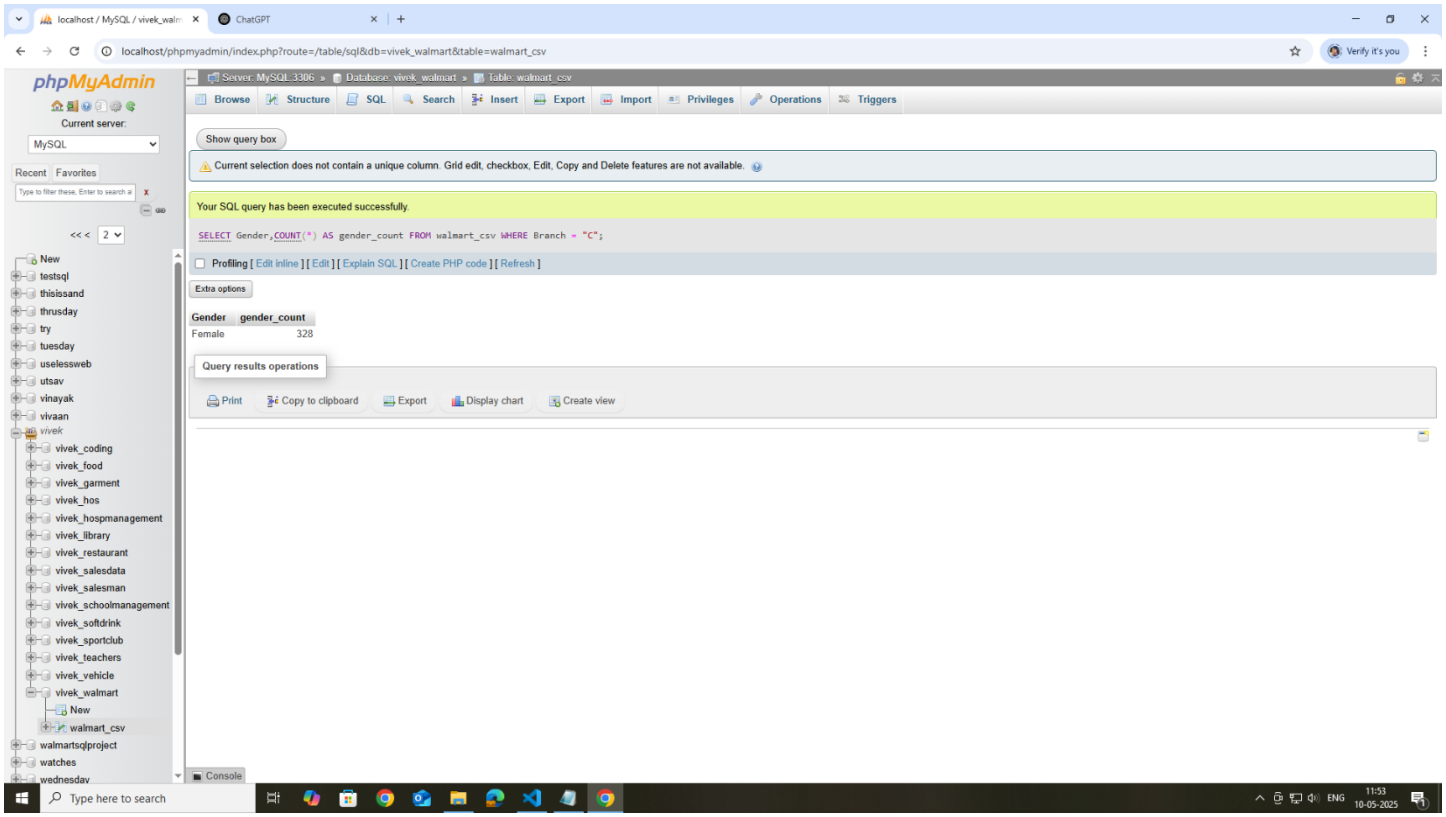
SELECT Gender,COUNT(*) AS gender_count FROM walmart_csv GROUP BY Gender ORDER BY gender_count DESC;



25) What is the gender distribution per branch?

→Your Query is :-

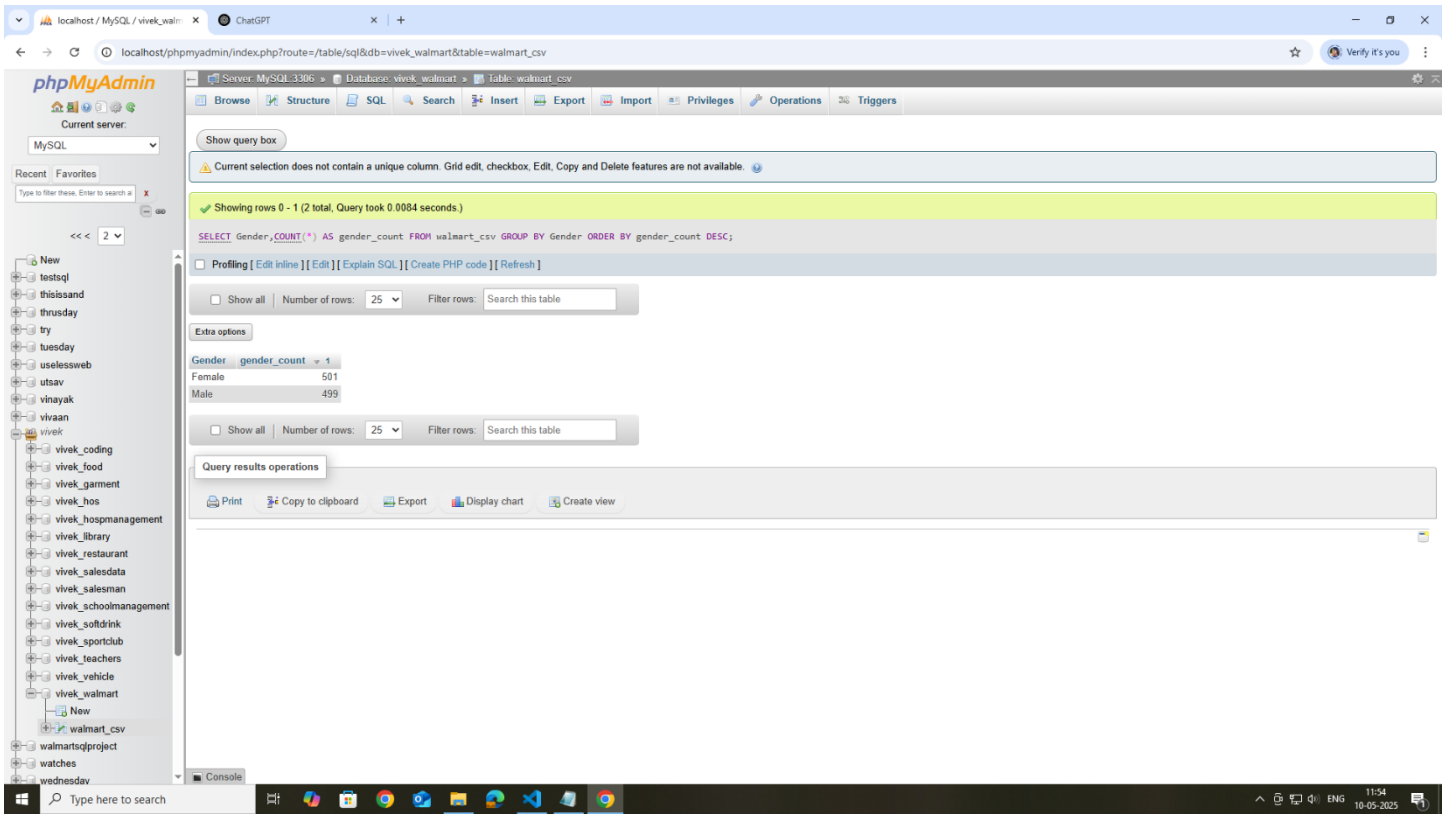
SELECT Gender,COUNT(*) AS gender_count FROM walmart_csv WHERE Branch = "C";



26) Which time of the day do customers give most ratings?

→Your Query is :-

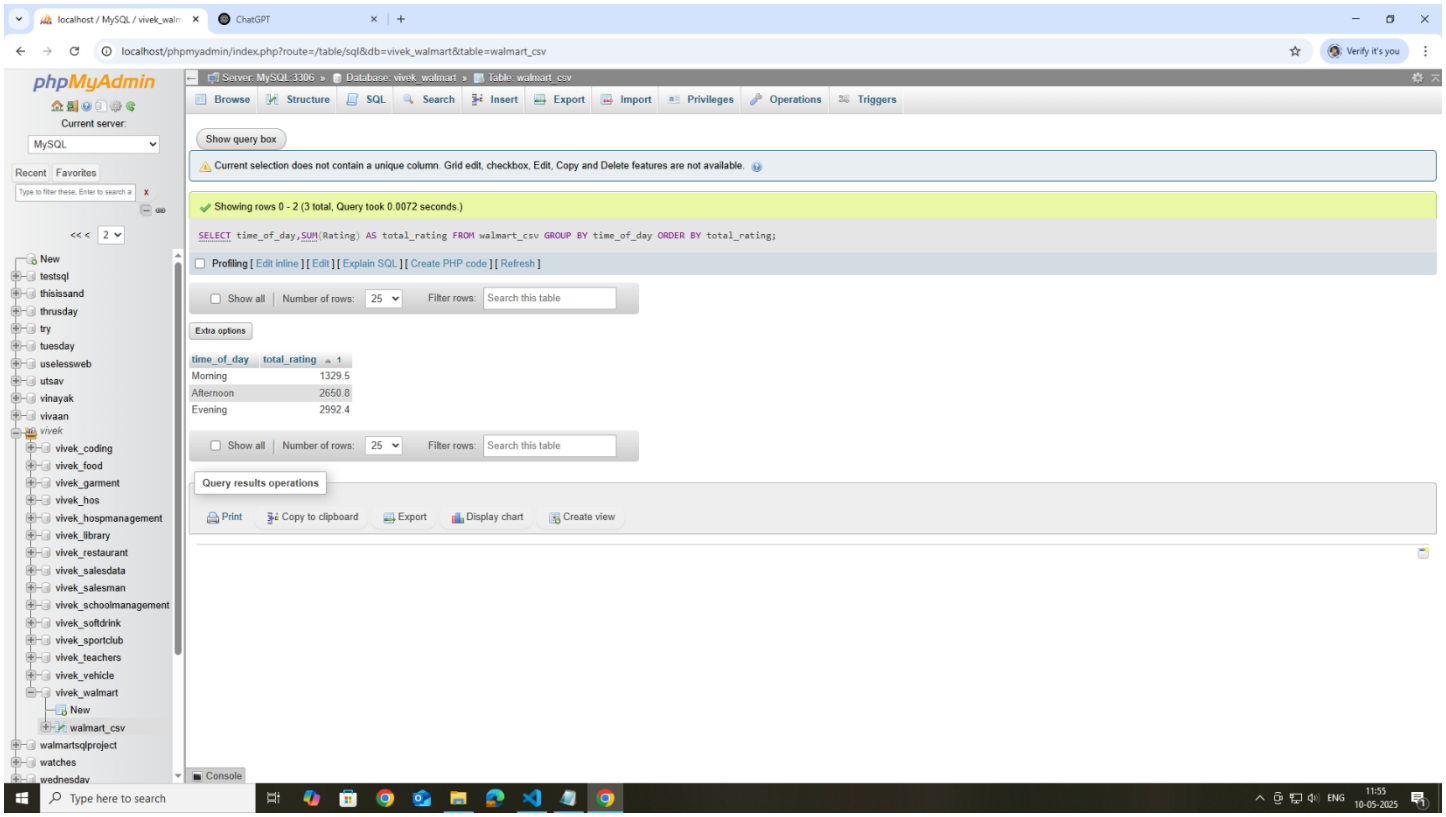
SELECT Gender,COUNT(*) AS gender_count FROM walmart_csv GROUP BY Gender ORDER BY gender_count DESC;



27) Which time of the day do customers give most ratings per branch?

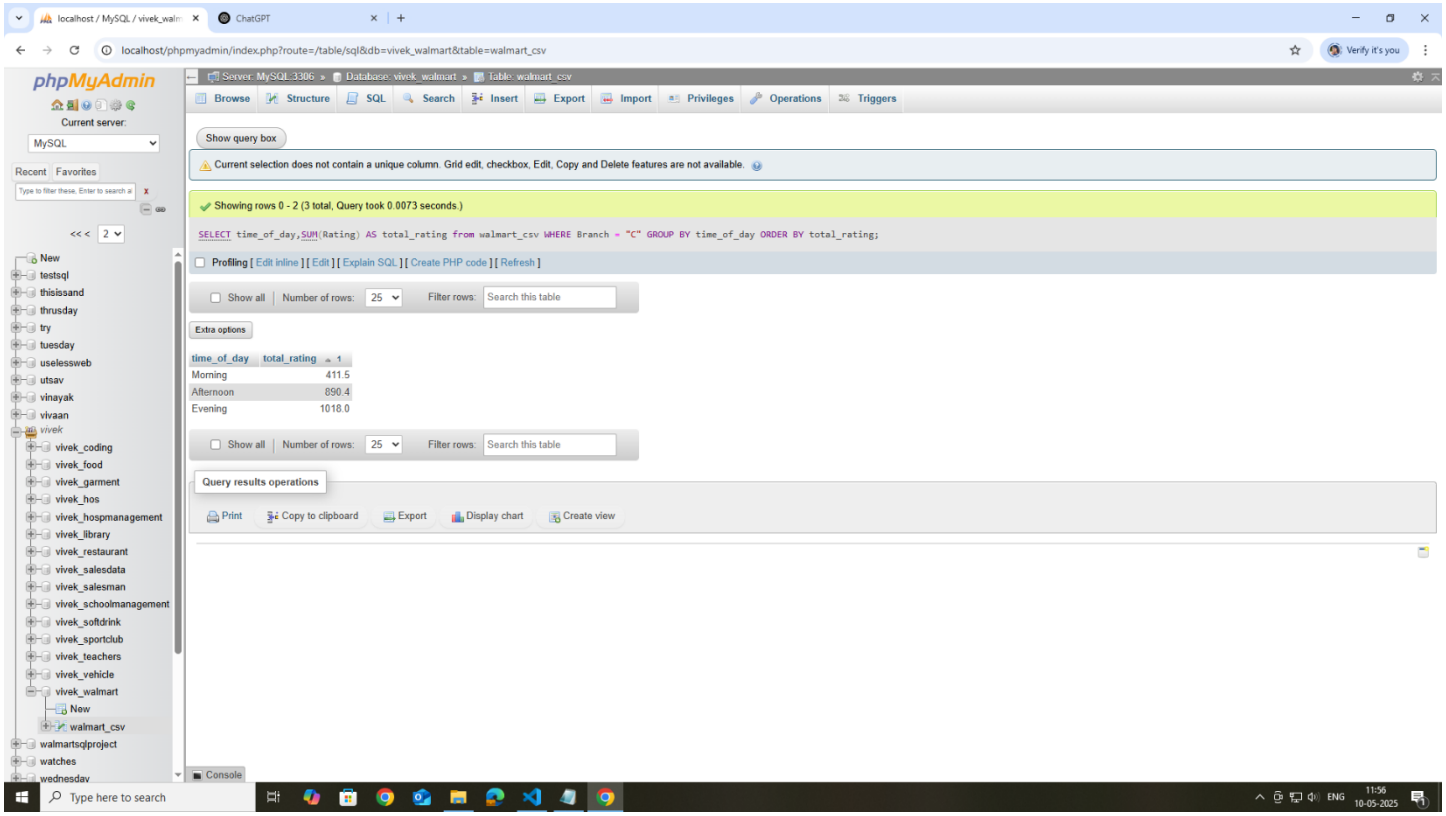
→Your Query is :-

SELECT time_of_day,SUM(Rating) AS total_rating FROM walmart_csv GROUP BY time_of_day ORDER BY total_rating;



→Second Query is :-

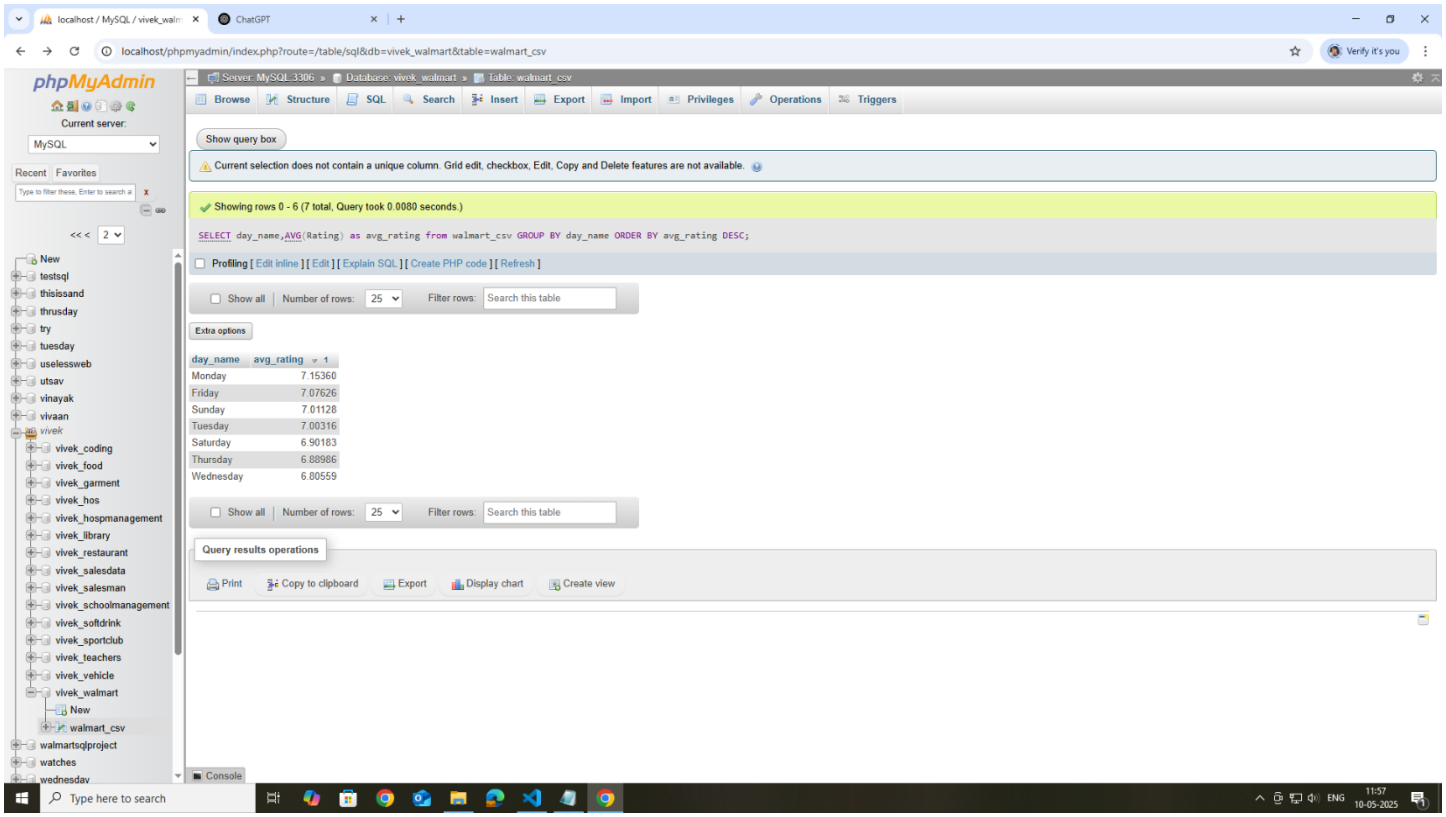
SELECT time_of_day,SUM(Rating) AS total_rating from walmart_csv WHERE Branch = "C" GROUP BY time_of_day ORDER BY total_rating;



28) Which day of the week has the best avg ratings?

→Your Query is :-

SELECT day_name,AVG(Rating) as avg_rating from walmart_csv GROUP BY day_name ORDER BY avg_rating DESC;



→Second Query is :-

SELECT day_name,AVG(Rating) AS avg_rating from walmart_csv WHERE Branch = "C" GROUP BY day_name ORDER BY avg_rating DESC;