BUAN 6320 - Database Foundations for Business Analytics

SQL PROJECT - SUPERMARKET SALES DATA

Group Number: 7
Members :

Laxmi Mounika Kalidindi - laxmimounika.kalidindi@utdallas.edu
Sai Sravya Bhupathiraju - saisravya.bhupathiraju@utdallas.edu
Siddhartha Duggirala - siddhartha.duggirala@utdallas.edu
Shaik Mohammed Sohail - shaik.mohammedsohail@utdallas.edu

Under the Supervision of

Prof. Ravishankar Narayan

Group Activity

Activity	Member 1	Member 2	Member 3	Member 4
Prepared Data Model and Created Physical DB	✓	✓	✓	✓
Loaded Data into Database	✓	✓	✓	✓
Wrote SQL Queries	✓	✓		
Prepared Mongo Database			✓	✓
Loaded data into MongoDB			✓	✓
Wrote Mongo Queries	✓	✓		
Prepared Report	✓	✓	✓	✓
Reviewed Report	✓	✓	✓	✓

Contents

Relational Data Model

Assumptions About Data Entities and Relationships

From the given supermarket sales dataset, the data has 1000 customers. Since the demographic information is not specified and the uniqueness of the customer cannot be obtained with customer type and gender, we are assuming that there are 1000 unique customers and 1000 unique orders.

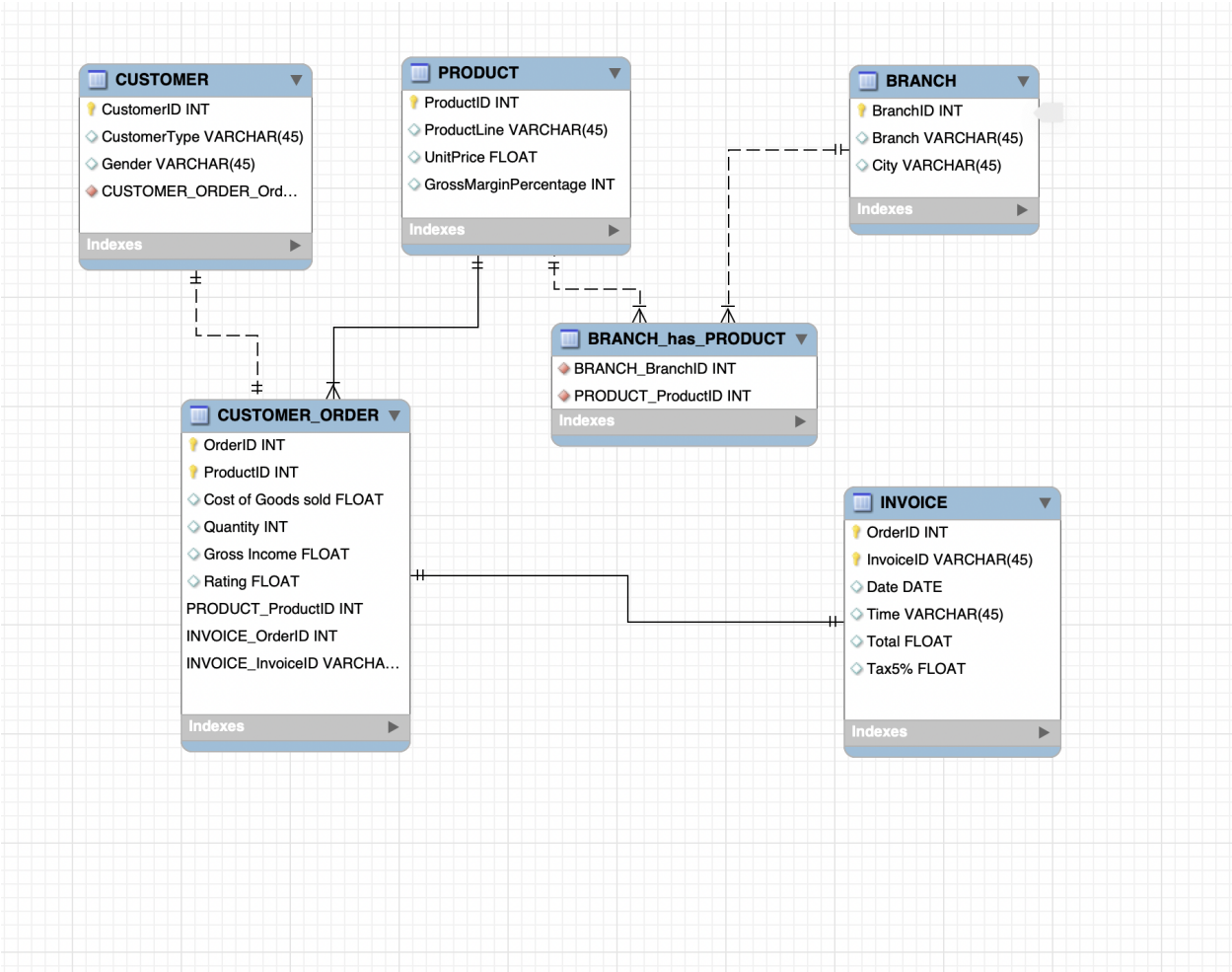
- Every customer has one and only one customer order. Each customer order is associated with one and only one customer. There is a 1-1 Non-Identifying Relationship(primary key of the customer is not a part of the primary key of a Customer order)
- Every customer order has one and only one product. Each product is present in many customer orders. There is a 1-M Identifying Relationship. (primary key of product is a part of the primary key of a customer order)
- Every product is available in many branches. Each branch is associated with many products. There is an M-M Non-Identifying Relationship. (primary key of the branch is not a part of the primary key of products)
- Every customer order has one and only one invoice. Each invoice is associated with one and only one customer order. There is 1-1 Identifying Relationship. (primary key of a customer order is a part of the primary key of invoice).

Data Model is in Third Normalized form(3 NF).

1. All the tables have primary keys with unique and non null values.
2. All the columns have unique clear names with atomic values.
3. All the non-key columns must be functionally dependent on the entire primary key.
    a. **Customer Table:** All the non key values(Gender, Customer type) are functionally dependent on the customer-id which is the primary key.
    b. **Customer order Table:** All the non key values (Tax, Cogs, Quantity, Gross Income, Rating) are functionally dependent only the composite primary (Order id, Product id) key
    c. **Invoice Table:** All the non key values (Date, Time, Total, Payment type) are functionally dependent only on the composite primary (Invoice id, OrderID) key
    d. **Branch Table:** All the non key values (Branch, City) are functionally dependent only the primary (Branch id) key
    e. **Product Table:** All the non key values (Product line, unit price, Gross Margin) is functionally dependent only the primary (product id) key

4. There is no transitive dependency. No non-key is functionally dependent on another non-key column in all the tables. Furthermore, there are no calculated columns that are dependent on other nonkey columns.

Entity-Relationship Diagram



Physical MySQL Database

Assumptions

Dimensional Tables(Master Data)

Branch

From the data, we assumed that Customer orders are placed at different branches of the SuperMarket. So created Branch table to describe Branches.

BranchID - Unique Branch ID to identify different branches(*Autogenerated data for insertion of unique records)*

Branch - Branch code *(from the given data)*

City - City where the branch is located *(from the given data*)

Product

We have assumed that customers choose from the available products list based on the corresponding Product Line and its Unit Price and gross margin percentage obtained based on the products available at Supermarket.

Product ID - Unique ProductID to identify different products in Supermarket(*Autogenerated data for insertion of unique records)*

Product Line - Under which category of goods the product falls into(*from the given data)*

UnitPrice - The price of each unit of Product.This is actually stored in the database.It does not change frequently for each order but supermarkets could change at times if there are any circumstances to change that unit price. So, it comes under dimensional tables.*(from the given data)*

Gross Margin Percentage - For this data, gross margin percentage is unchanging for all orders.So, if any new product is introduced by the Supermarket, we could insert new data for Gross Margin Percentage(*from the given data)*

Customer

There is no CustomerID given in the data. We have assumed that for any order to be placed, we should map the order to each different customer and could assign the Total price for each unique customer based on the order he made

Customer ID - Unique CustomerID to identify different customers(*Auto-generated data for insertion of unique records)*

Customer Type - Whether the customer holds any membership with the Supermarket or a normal member. This could be used to determine any discounts or offers for members. *(From the given data*)
Gender - Gender of the Customer (*From the given data*)

Customer-Order

There is no OrderID in the given data, so we have assumed to include unique OrderID for each order made and for the corresponding invoice generated for that particular customer.
OrderID - Unique OrderID to identify different orders made by Customers in different branches of supermarket(*Auto generated data for insertion of unique records)*
ProductID - Unique ProductID pulled from Product table when customer selects a Product in his order*(from the auto generated data in Product table)*
Quantity - Quantity of each product purchased by Customer in any particular order*(from the given data)*
Cost of Goods Sold - Total cost of goods purchased*(from the given data)*
Gross Income - Income generated for the supermarket for each order made by any customer*(from the given data)*
Rating - Rating given by the customer on his satisfaction of that particular order and products in the order*(from the given data)*

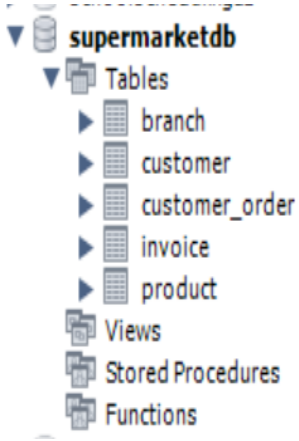
Invoice

An Invoice is generated with each customer order which has a Unique InvoiceID.Ideally, an invoice can have multiple orders. However, based on the data given, as the Invoice is unique for all the records, we are assuming each invoice contains only one order.
Invoice ID - Unique Invoice ID to identify different Invoices(*from the given data)*
Order ID - Unique OrderID to identify different orders made by Customers in different branches of the supermarket(*from the auto-generated data in the customer order table)*
Date&Time - The Date & Time at which the order is placed(*from the given data)*
Total - Total bill generated and amount to be paid by the Customer including the tax on his purchase*(from the given data)*
Tax5% - Amount calculated for a total cost of goods purchased by taking constant 5% of the tax for all the goods*(from the given data*

Payment type - Type of payment used by the customer to place the order*(from the given data)*

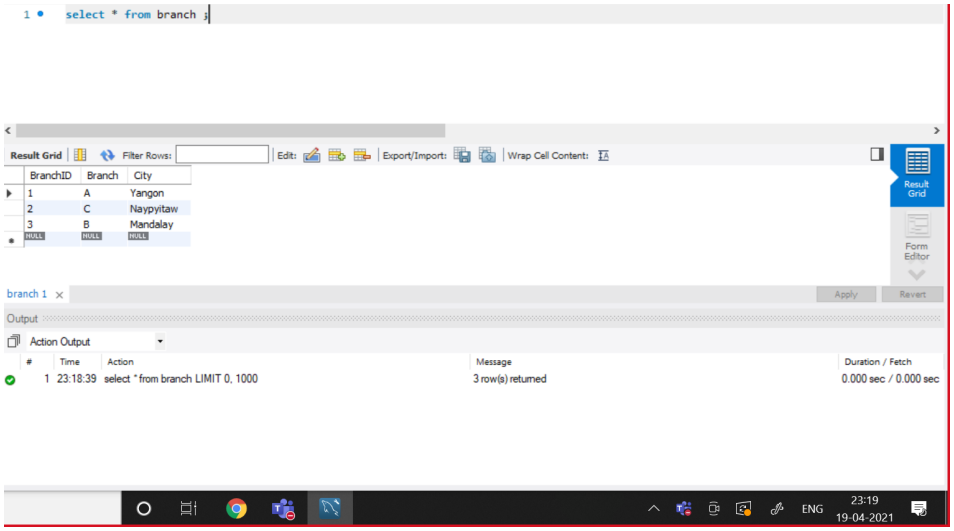
Physical Database Objects

After creating the Entity-Relationship diagram, use forward engineering to create the physical database

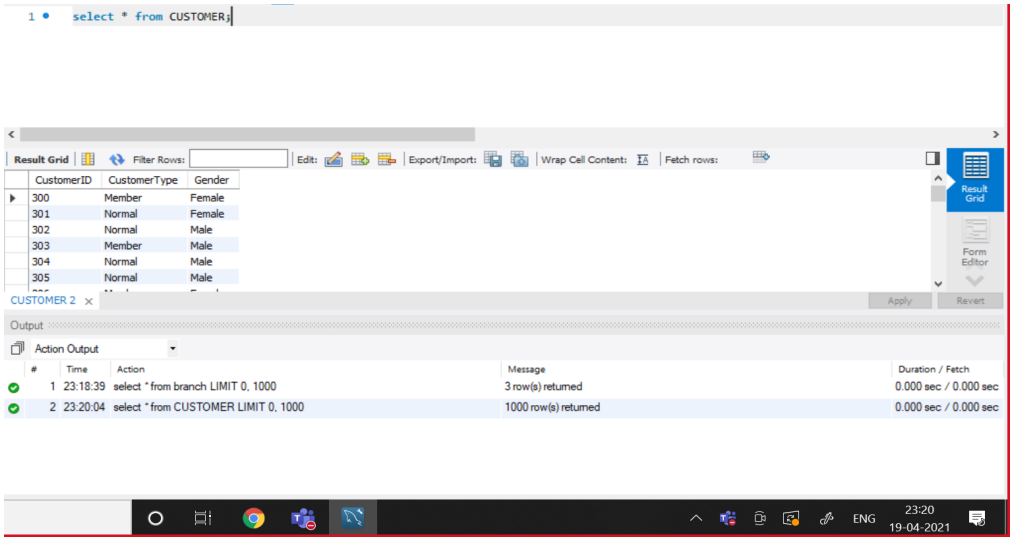


Data in the Database:

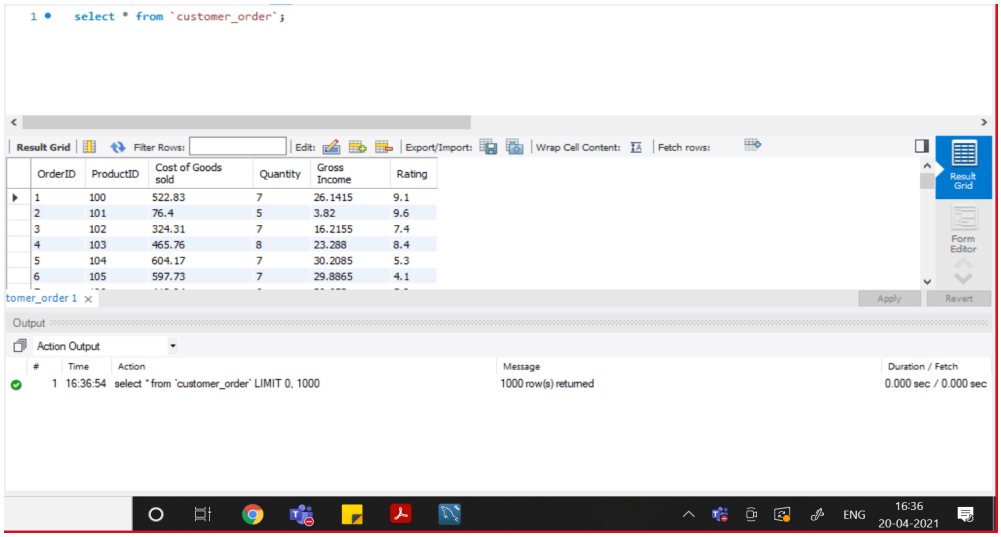
BRANCH Table



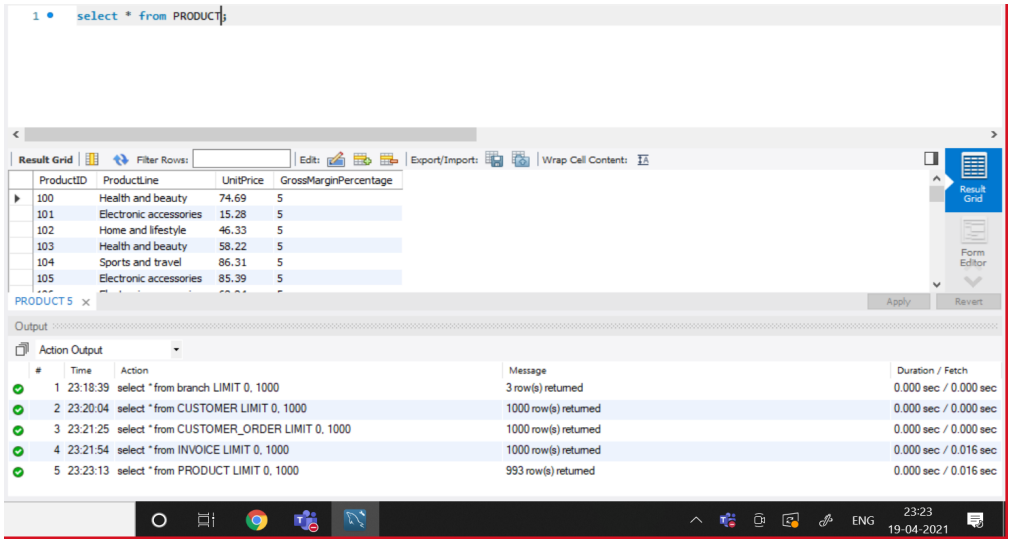
CUSTOMER Table



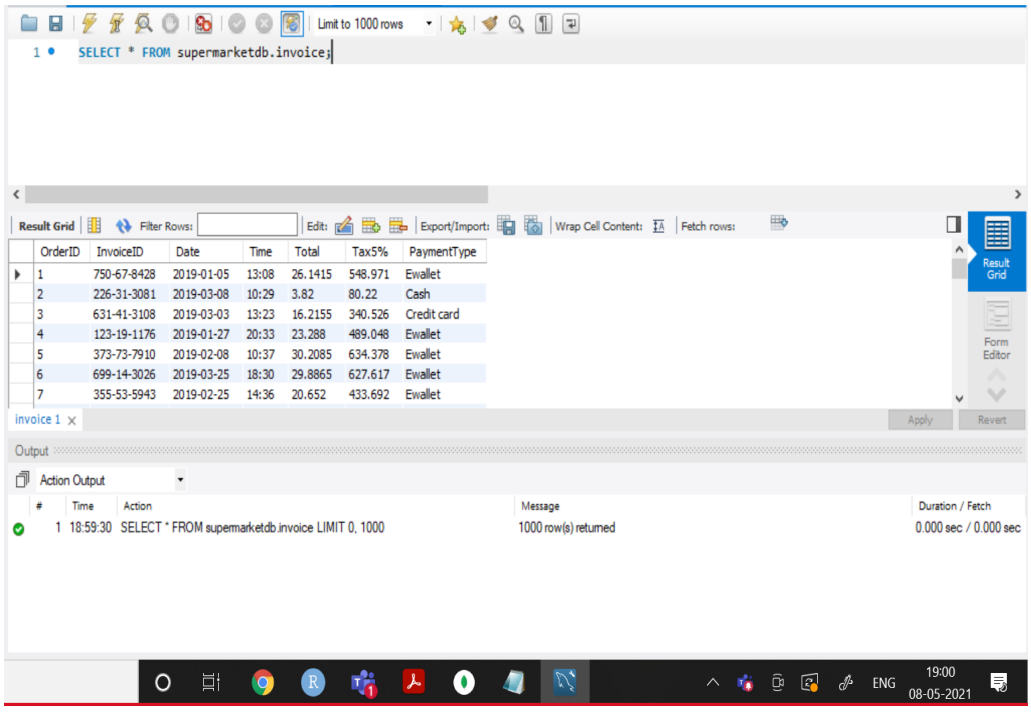
CUSTOMER ORDER Table



PRODUCT Table



INVOICE Table



DATA IN THE DATABASE

Table Name	Primary Key	Foreign Key	# of rows in the table
Branch	Branch ID	Product ID	3
Product	ProductID	Branch ID	993
Customer	CustomerID	Order ID	1000
Invoice	InvoiceID	Order ID	1000
Customer Order	OrderID,ProductID	Invoice ID	1000

Reasons for revising data :

From Phase 1, We have just made a small correction which includes an identifying relationship between order id and invoice table.

SQL QUERIES

Question 1:

Some retailers believe that there is more money to be made in selling fashion accessories to men than sports and travel to women. Is this true?

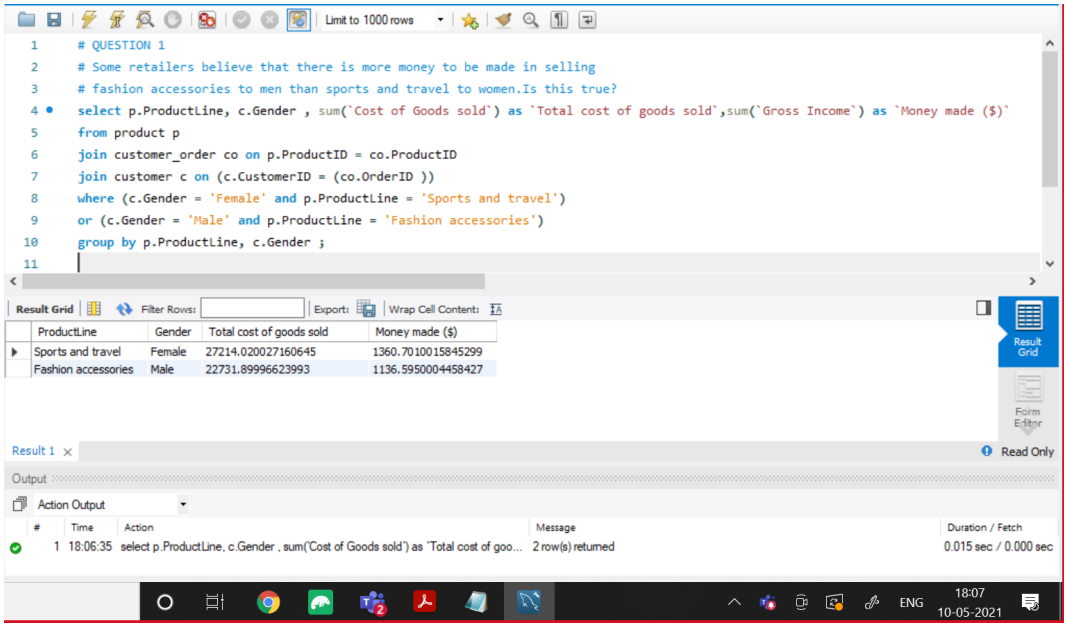
Translate:

Select Product Line, Gender, the sum of Cost of Goods Sold as Total cost of Goods Sold, the sum of Gross Income as Money Made $ from the Product table joined with the Customer Order table on ProductID in the Product table matching ProductID in Customer Order table joined with Customer table on OrderID in the Customer Order table matching CustomerID in the customer table where gender of the customer equal female and Product Line equal 'Sports and Travel' or Customer Gender of customer equal Male and Product Line equal Fashion Accessories grouped by Product Line and Gender.

Cleanup:

Select Product.Product Line, Customer.Gender, sum(Cost of Goods Sold) as Total cost of Goods Sold, sum(Gross Income) as Money Made $ from Product join Customer Order on Product.Product ID = Customer Order.Product ID join Customer on Customer Order.Order ID=Customer.Customer ID where (customer.gender= 'Female' and Product Line='Sports and Travel') or (Customer.gender ='Male' and Product Line='Fashion Accessories') group by Product.Product Line and Customer.Gender

Screenshot of SQL Query and Results



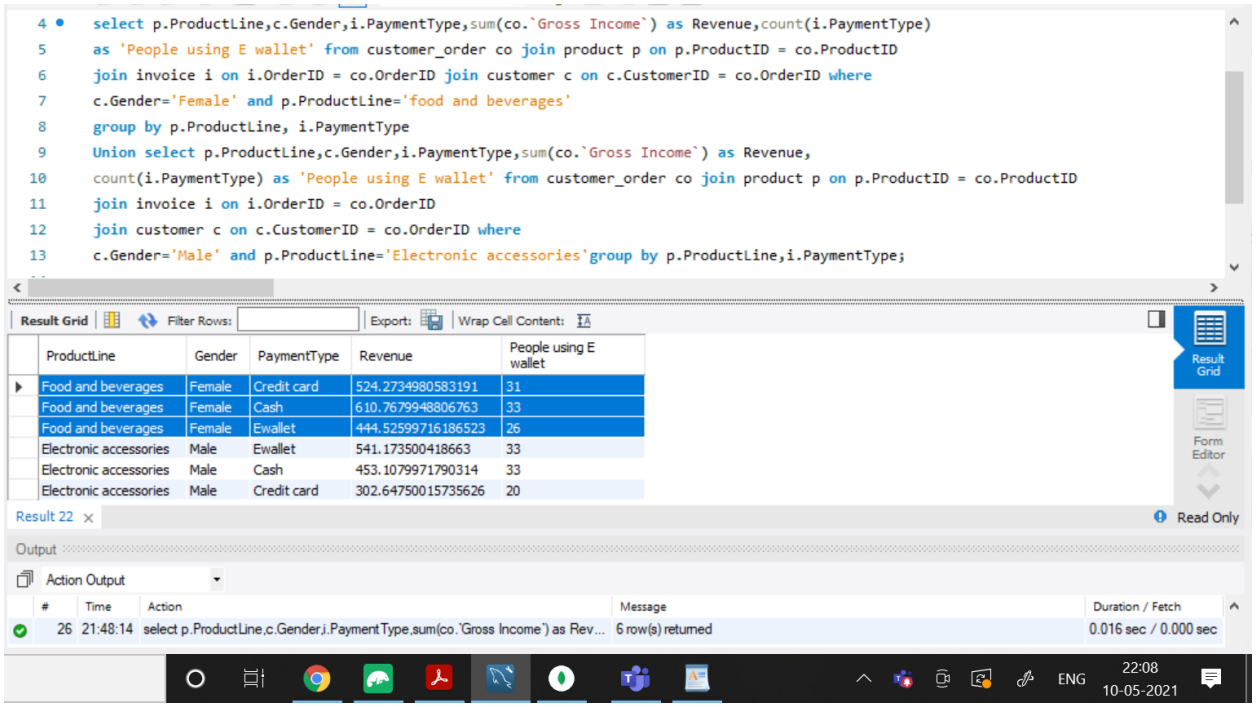
Notes/Comments:
Retailers' belief is **wrong**. There is more gross income generated by selling 'Sports and Travel' related products to Female customers which is $1360.70 than selling Fashion accessories to male customers which is $1136.59. Also the conclusion can be further supported by verifying Total cost of goods sold for both female and male customers who purchased 'Sports and Travel' and 'fashion accessories' respectively.Number of rows returned=2

Question 2: Some retailers believe that revenue in food and beverages can be increased amongst women by focusing on E-wallets, while others believe EWallets are more popular with men buying electronic accessories. Who is right?

Translate:
Select Product Line,Customer Gender,the count of Payment Type as 'People using E wallet', the sum of Gross Income as Revenue, from Customer_order table joined with the Product table on Product ID in the Customer Order table matching Product ID in Product table joined with Invoice table on OrderID in the Customer Order table matching Order ID in the invoice table joined with Customer table on Order ID in Customer Order table matching with Customer ID in Customer table where Customer gender equal Female, Product Line equal 'Food and Beverages' and union the result with the result of Customer gender equal Male, Product Line equal 'Electronic Accessories' and Payment Type equal EWallet grouped by Payment type and Product Line

Cleanup:
Select Product.ProductLine, Customer.Gender , count(Invoice.PaymentType) sum(Customer Order.Gross Income) as Revenue, as 'People using E wallet' from Customer_order join Product on Customer Order.Product ID=Product.Product ID join Invoice on Customer Order.OrderID=Invoice.Order ID join Customer on Customer Order.Order ID=Customer.Customer ID where (Customer.Gender= 'Female' and Product.Product Line='Food and Beverages') union (Customer.Gender='Male' with Product.Product Line='Electronic Accessories') group by Invoice.PaymentType and Product.Product Line.

Screenshot of SQL Query and Results



```
select p.ProductLine,c.Gender,i.PaymentType,sum(co.`Gross Income`) as Revenue,count(i.PaymentType)
as 'People using E wallet' from customer_order co join product p on p.ProductID = co.ProductID
join invoice i on i.OrderID = co.OrderID join customer c on c.CustomerID = co.OrderID where
c.Gender='Female' and p.ProductLine='food and beverages'
group by p.ProductLine, i.PaymentType
Union select p.ProductLine,c.Gender,i.PaymentType,sum(co.`Gross Income`) as Revenue,
count(i.PaymentType) as 'People using E wallet' from customer_order co join product p on p.ProductID = co.ProductID
join invoice i on i.OrderID = co.OrderID
join customer c on c.CustomerID = co.OrderID where
c.Gender='Male' and p.ProductLine='Electronic accessories'group by p.ProductLine,i.PaymentType;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

ProductLine	Gender	PaymentType	Revenue	People using E wallet
Food and beverages	Female	Credit card	524.2734980583191	31
Food and beverages	Female	Cash	610.7679948806763	33
Food and beverages	Female	Ewallet	444.52599716186523	26
Electronic accessories	Male	Ewallet	541.173500418663	33
Electronic accessories	Male	Cash	453.1079971790314	33
Electronic accessories	Male	Credit card	302.64750015735626	20

Result 22 | Read Only

Output

Action Output

#	Time	Action	Message	Duration / Fetch
26	21:48:14	select p.ProductLine,c.Gender,i.PaymentType,sum(co.`Gross Income`) as Rev...	6 row(s) returned	0.016 sec / 0.000 sec

ENG 22:08 10-05-2021

Notes/Comments:

Retailers who believe “e-wallets are more popular with men buying electronic accessories” are **right.** Total number of payments by male customers in purchasing electronic accessories which was a total of 33 payments that were made and generated a revenue of $ 541.17 is higher compared to the total number of payments by male customers using the other mode of payments. The total number of payments by female customers in purchasing products related to Food and beverages which were only 26 with a generated revenue of $ 444.52 which is lesser compared to the other modes of payments. Number of rows returned=2.

Question 3: Some retailers believe payment method is a bigger indicator of health and beauty purchases while other retailers believe gender is a bigger factor. Who is right?

Translate:

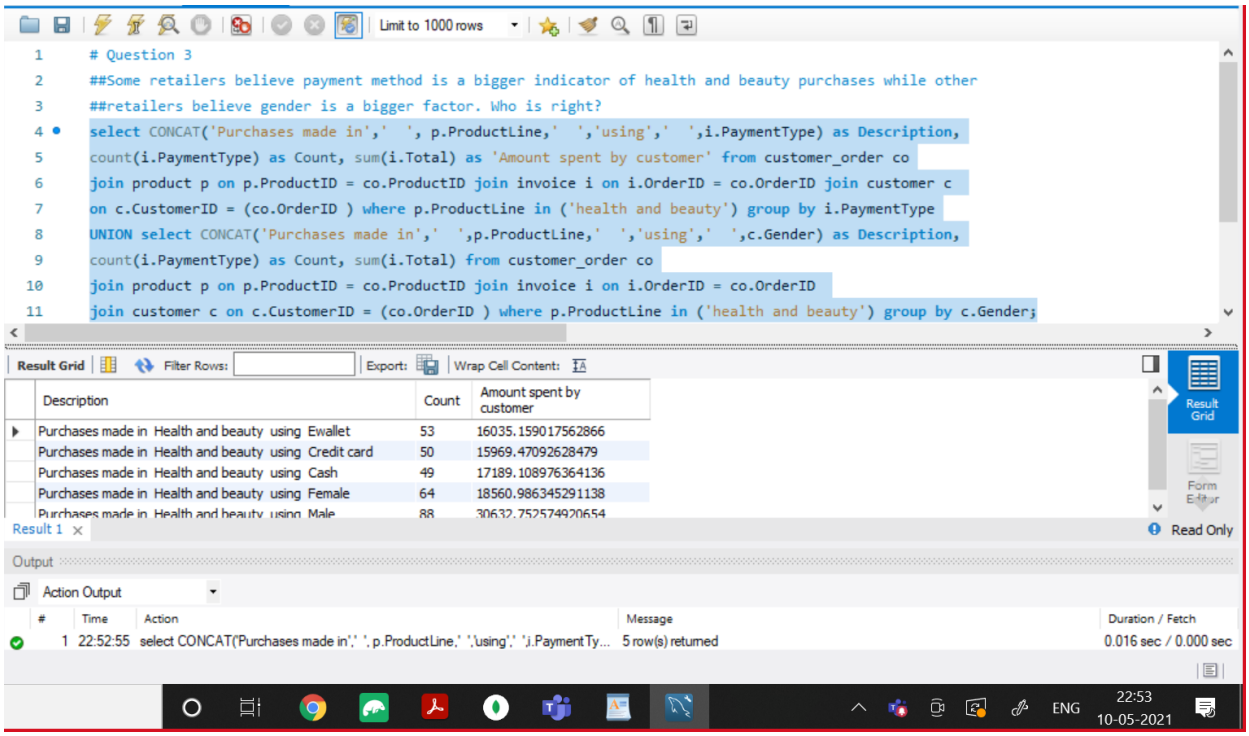
Select concatenation of “Purchase made in”,” “, Product Line,” “,”using”,” “, and Payment Type as Description,the count of payment type as Number of Payments, the sum of total as ‘Amount spent by customer’ from customer_order table joined with Product table on Product ID in Customer Order table matching Product ID in Product table joined with Invoice table on Order ID in Customer Order table matching Order ID in Invoice table joined with Customer table on Order ID in Customer Order table

matching Customer ID in Customer table where Product Line is Health and Beauty grouped by Payment Type Union Select concatenation of “Purchase made in”,” “, Product Line,” “,”using”,” “, and Payment Type as Description,the count of payment type as Number of Payments, the sum of total as ‘Amount spent by customer’ from customer_order table joined with Product table on Product ID in Customer Order table matching Product ID in Product table joined with Invoice table on Order ID in Customer Order table matching Order ID in Invoice table joined with Customer table on Order ID in Customer Order table matching Customer ID in Customer table where Product Line is Health and Beauty grouped by Payment Type

Cleanup:

Select Customer.Customer Type, avg(Cost of Goods Sold) as Average amount made, avg(Total) as Average amount spent from Customer_Order join Invoice on Customer Order.Order ID = Invoice.Order ID join Customer on Customer.Customer ID=Customer Order.Order ID group by Customer.Customer Type

Screenshot of SQL Query and Results



Notes/Comments:

Retailers who believe **Gender is a bigger factor are right.** Total number of payments and Amount spent by customers using different payments methods is not much impacted by the Payment type as Count is 53,50,49 and Amount spent is

$16035.16,$15969.47,$17189.10 for Payment types Ewallet, Credit Card and Cash respectively. However, on the other hand, there is a significant impact of gender on Number of payments made and amount spent by Customers where purchases made by male customers are almost 70% more than purchases made by female customers. Number of rows returned=2.

Question 4:

Some retailers believe that their members are spending more per purchase while members believe they are spending less per purchase. Who is right?

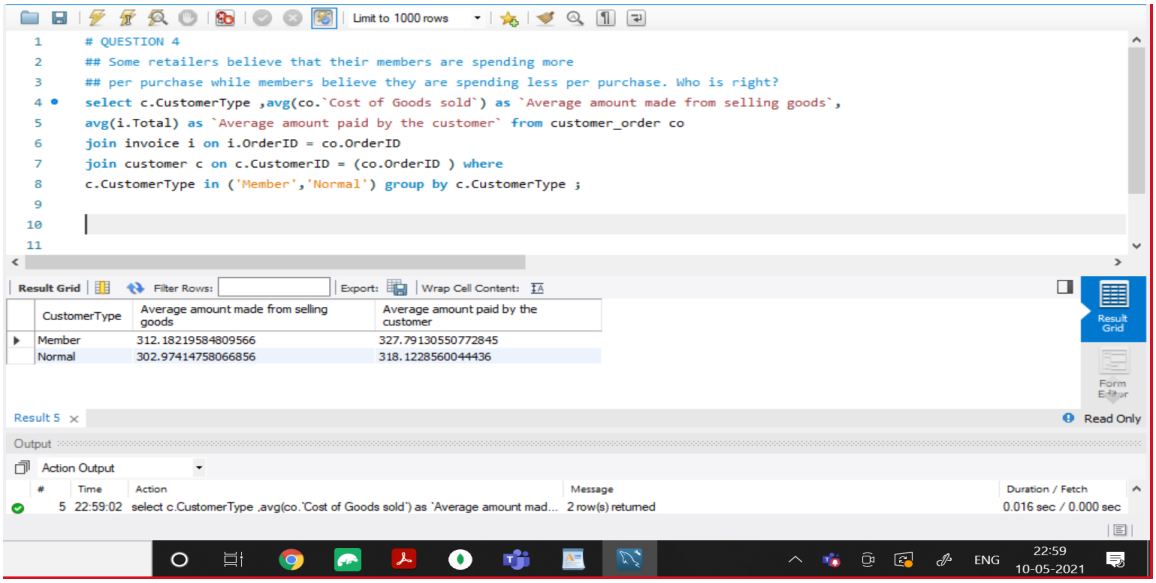
Translate:

Select Customer Type, the average Cost of Goods Sold as Average amount made, the average of Gross Income as average amount made $ from Customer order table joined with the Invoice table on Order ID in the Customer Order table matching Order ID in Invoice table joined with Customer table on OrderID in the Customer Order table matching CustomerID in the customer table grouped by Customer Type

Cleanup:

Select Customer.Customer Type, avg(Cost of Goods Sold) as Average amount made, avg(Total) as Average amount spent from Customer_Order join Invoice on Customer Order.Order ID = Invoice.Order ID join Customer on Customer.Customer ID=Customer Order.Order ID group by Customer.Customer Type

Screenshot of SQL Query and Results



Notes/Comments:
Retailers' belief that members spend more per purchase(which is the average purchase of members for all the orders) is **right.** Members spend on average $327.79 whereas Non-members spend an average $302.97. This might be because members would have some discounts on purchasing in bulk which in turn is more purchase value for order but price per unit decreases on the member side. Number of rows returned=2

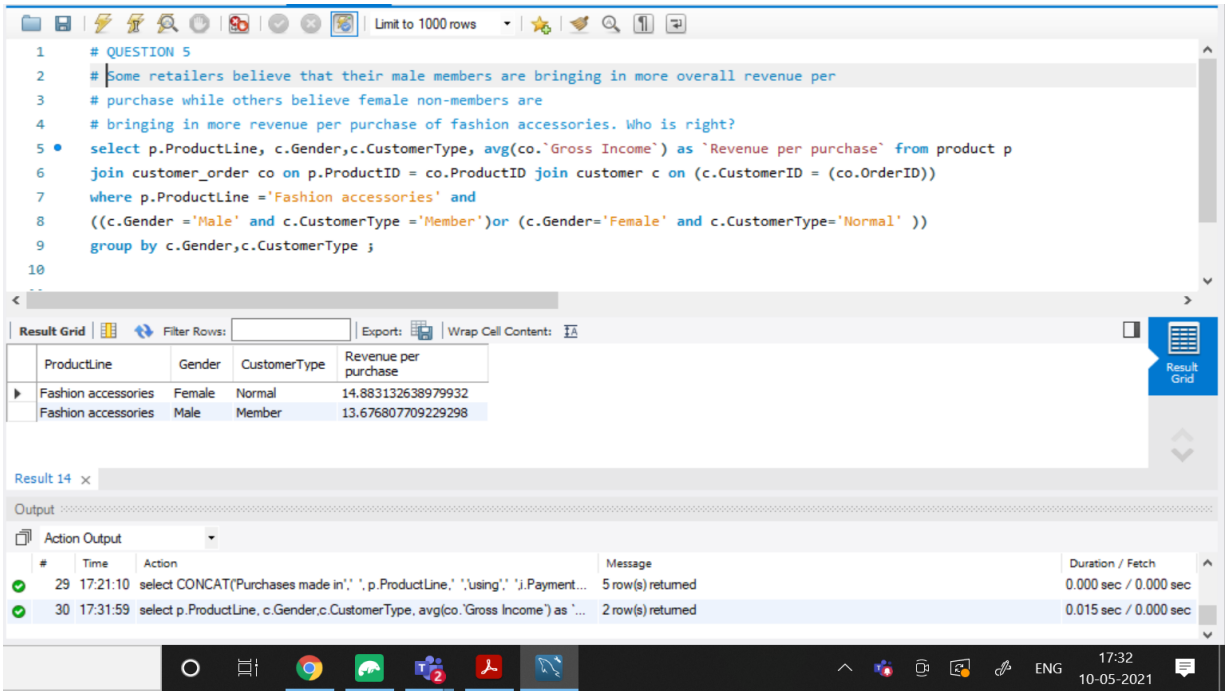
Question 5:

Some retailers believe that their male members are bringing in more overall revenue per purchase while others believe female non-members are bringing in more revenue per purchase of fashion accessories. Who is right?

Translate:
Select Product Line, Customer Gender, Customer Type, the average Gross Income as Revenue per purchase from Product table joined with the Customer_Order table on Product ID in the Product table matching Product ID in Customer Order table joined with Customer table on OrderID in the Customer Order table matching CustomerID in the customer table where Product Line is Fashion Accessories and Gender of Customer is Male and Customer Type is Member or Gender of Customer is Female and Customer Type is Normal grouped by Customer Gender and Customer Type

Cleanup:
Select Product.Product Line, Customer.Gender,Customer.Customer Type, avg(Gross Income) as Revenue per Purchase from Product join Customer Order on Customer Order.Product ID = Product.Product ID join Customer on Customer Order.Order ID=Customer.Customer ID where Product.Product Line='Fashion Accessories' and ((Customer.Gender='Male' and Customer.Customer Type='Member') or (Customer.Gender='Female' and Customer.Customer Type='Normal)) 'group by Customer.Gender,Customer.Customer Type

Screenshot of SQL Query and Results



Notes/Comments:
Retailers who believe that female non-members are bringing in more revenue per purchase of fashion accessories are **right.** Female Non members who purchase fashion accessories bring a revenue per purchase of $14.88 on average whereas Male Members bring in revenue per purchase of $13.67 on average. Number of rows= 2

Data Review for MongoDB

Assumptions/Notes About Data Collections

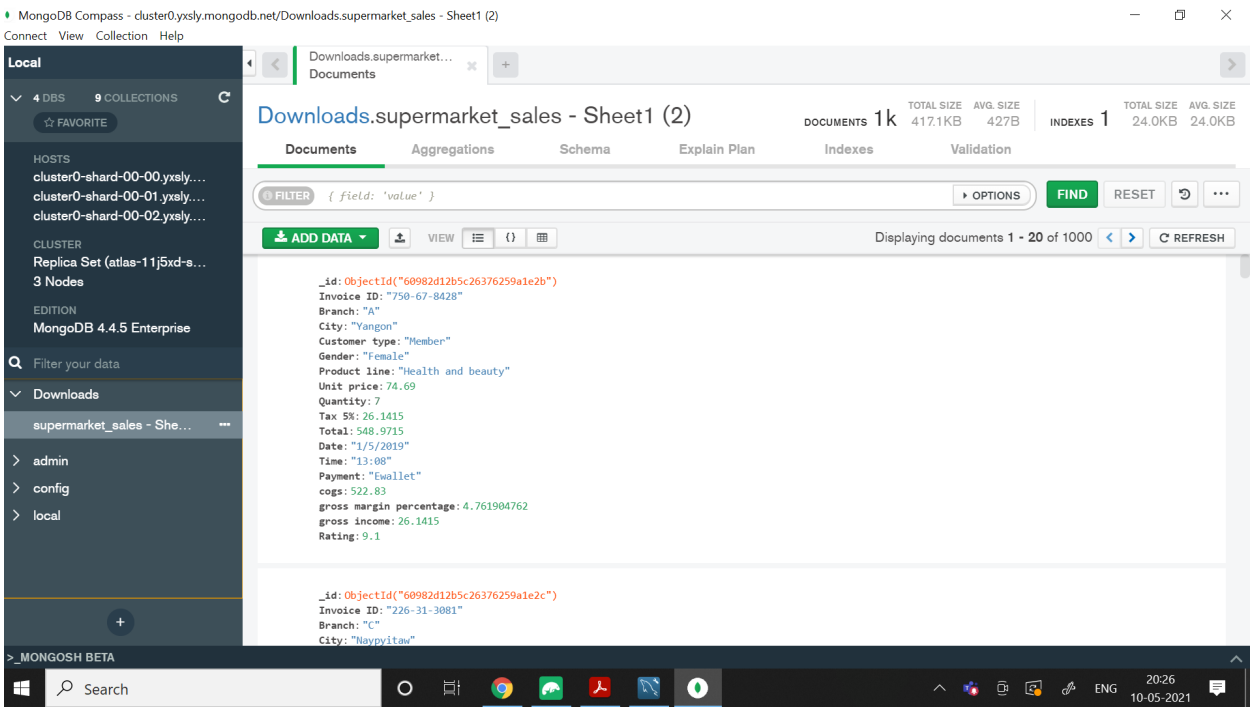
A dataset consisting of a supermarket sales of 1000 customers is given. Since the customer uniqueness cannot be obtained from customer type and gender and the demographic information is not specified, we are assuming that there are unique customers i.e 1000 unique orders and 1000 unique customers.

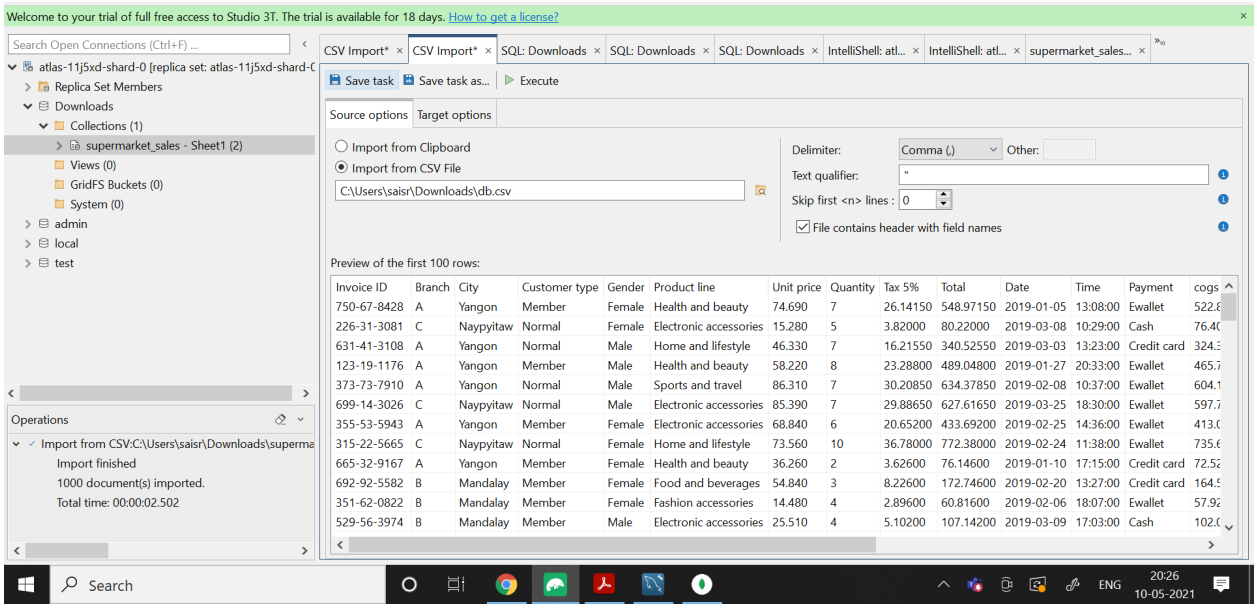
Attributes and Relationships in Collection

- Every customer has one and only one customer order. Each customer order is associated with one and only one customer.
- Every customer order has one and only one product. Each product is present in many customer orders.
- Every product is available in many branches. Each branch is associated with many products.
- Every customer order has one and only one invoice. Each invoice is associated with one and only one customer order.

Physical Data in Mongo Database :

We have downloaded the **raw dataset** into the Mongo Database and generated all our mongo queries using **Studio 3T.**





MONGO QUERIES

Question 1:

Some retailers believe that there is more money to be made in selling fashion accessories to men than sports and travel to women. Is this true?

Translate:

Using Supermarket_Sales collection,utilize aggregate framework to group documents by Gender and Product Line and get the sum of 'cost of goods sold' for each gender and each Product Line from purchases made by female customers in the "Sports and Travel" product line and purchases made by male customers in the "Fashion accessories" product line.

```
db = db.getSiblingDB("Downloads");
db.getCollection("supermarket_sales - Sheet1 (2)").aggregate(
    [{ "$match" : {
        "$or" : [ { "$and" : [ { "Gender" : {  "$in" : ["Female"]}},
                  { "Product line" : {  "$in" : [ "Sports and travel"]   }}]},
                  {"$and" : [{  "Gender" : { "$in" : ["Male"] } },
                  {"Product line" : {  "$in" : ["Fashion accessories"] } }]} ] }  },
                 { "$group" : {  "_id" : { "Gender" : "$Gender", "Product line" : "$Product line" },
                   "SUM(cogs)" : {  "$sum" : "$cogs" }} },
                  {  "$project" : { "Product line" : "$_id.Product line", "Gender" : "$_id.Gender",
                    "SUM(cogs)" : "$SUM(cogs)",
                    "_id" : NumberInt(0)}}], { "allowDiskUse" : true}
                 );
```

Raw Shell Output | Aggregate Query (line 3)

50 | Documents 1 to 2 | Pin Result | Query Code | Explain Query | Table View

Output > Product line

_id	Product line	Gender	SUM(cogs)
	Sports and travel	Female	27214.02
	Fashion accessori...	Male	22731.9

1 document selected | Count Documents | 00:00:00.051

ENG 15:10 09-05-2021

Screenshot of MongoDB Query and Results

Notes/Comments

Retailers' belief is **wrong**. Cost of goods sold for female customers purchasing in by 'Sports and Travel' Product Line $27214.02 is more compared to Cost of goods sold for male customers purchasing in Fashion accessories Product Line $22731.9. Number of documents =2

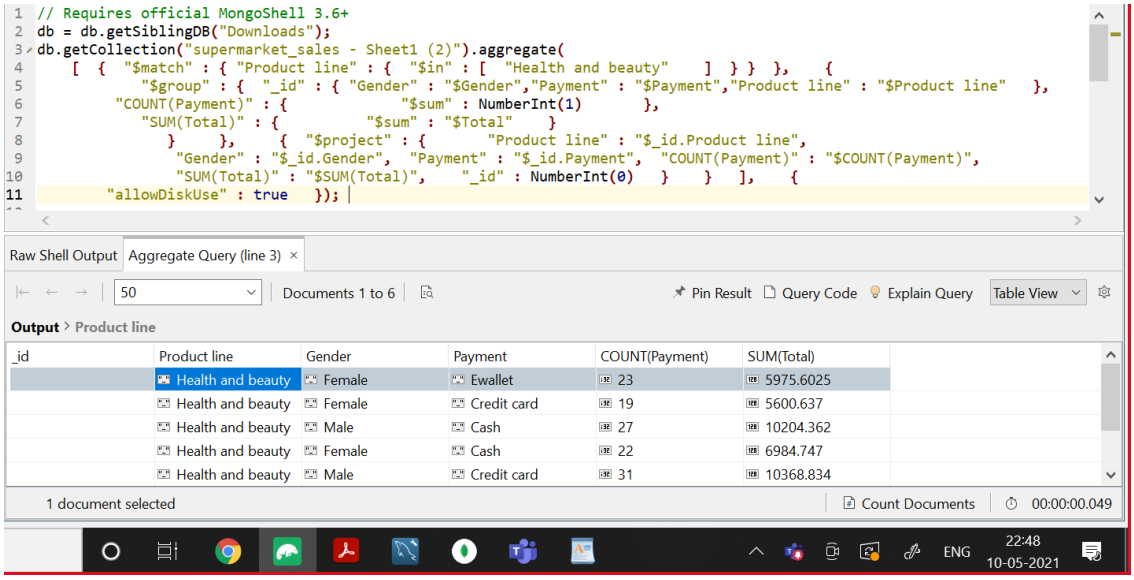
Question 3:

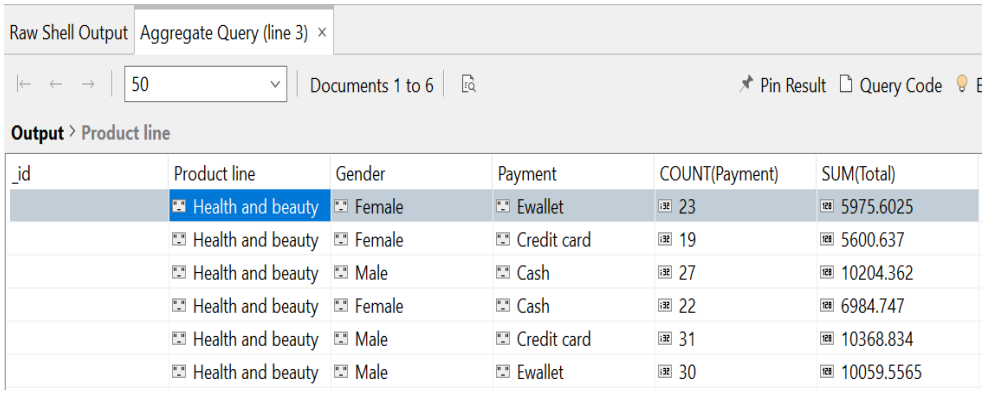
Some retailers believe payment method is a bigger indicator of health and beauty purchases while other retailers believe gender is a bigger factor. Who is right?

Translate:

Using Supermarket_Sales collection,utilize aggregate framework to group documents by Payment Type and Gender and get the sum of Total paid for each gender and each Payment Type in Product Line "Health and Beauty" and count the number of documents for each Payment Type

Screenshot of MongoDB Query and Results





_id	Product line	Gender	Payment	COUNT(Payment)	SUM(Total)
	Health and beauty	Female	Ewallet	23	5975.6025
	Health and beauty	Female	Credit card	19	5600.637
	Health and beauty	Male	Cash	27	10204.362
	Health and beauty	Female	Cash	22	6984.747
	Health and beauty	Male	Credit card	31	10368.834
	Health and beauty	Male	Ewallet	30	10059.5565

Notes/Comments:

Retailers who believe **Gender is a bigger factor are right.** Total number of payments and Amount spent by customers using different payments methods is not much impacted by the Payment type as Count is 53,50,49 and Amount spent is $16035.16,$15969.47,$17189.10 for Payment types Ewallet, Credit Card and Cash respectively. However, on the other hand, there is a significant impact of gender on Number of payments made and amount spent by Customers where purchases made by male customers are almost 70% more than purchases made by female customers. Number of documents=6

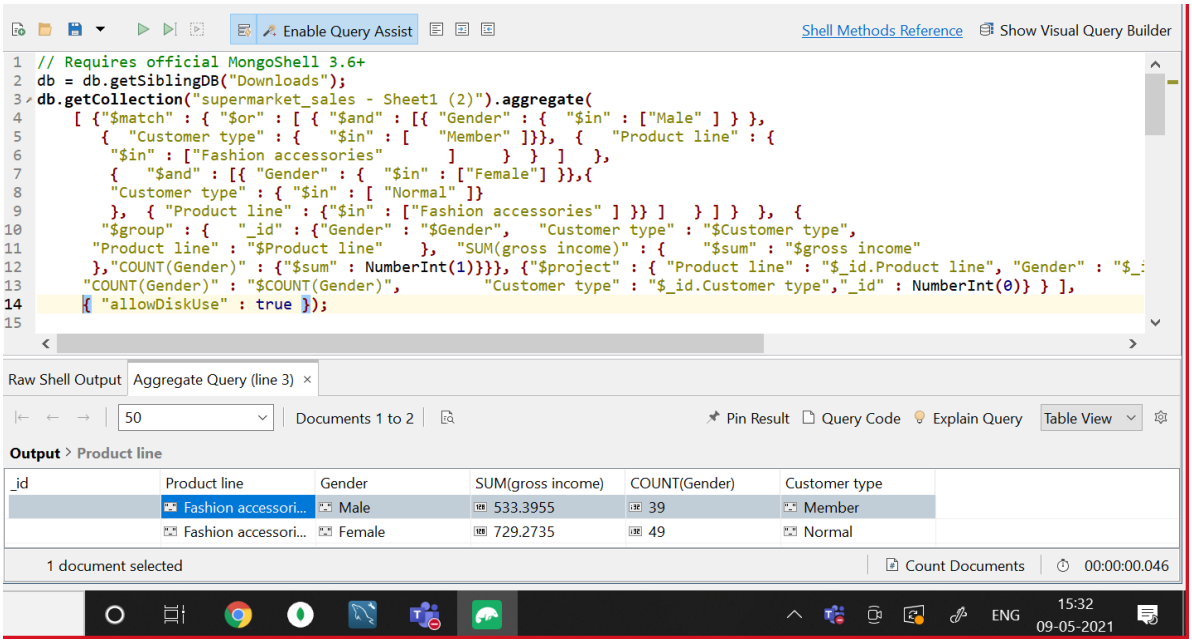
Question 5:

Some retailers believe that their male members are bringing in more overall revenue per purchase while others believe female non-members are bringing in more revenue per purchase of fashion accessories. Who is right?

Translate:

Using Supermarket_Sales collection,utilize aggregate framework to group documents by Gender and Customer Type and get the average of ‘Gross Income’ for each gender and each Customer Type from purchases made in “Fashion accessories” product line and by male members and female non-members(normal customer type)

Screenshot of MongoDB Query and Results



Notes/Comments:

Retailers who believe that female non-members are bringing in more revenue per purchase of fashion accessories are **right.** Female Non members who purchase fashion accessories bring a revenue of $729.27 for 49 purchases which is $14.88 on average whereas Male Members bring in revenue of $533.39 for 39 purchases per purchase of $13.67 on average. Number of documents=2