SQL CAPSTONE PROJECT

Project Title

Superstore Sales Analysis Using SQL

Sri Sathya Sai Organisation

Submitted By: G.Vivekchary

Course: SQL

Table of Contents

What is a Database

A database is an organized collection of structured information or data, typically stored electronically in a computer system. It allows users to easily access, manage, modify, update, and retrieve data efficiently and securely. Databases are fundamental to numerous software applications and systems that require structured data handling.

Key Components of a Database

- **Data**: Data is defined as collection of information such as names, numbers, or transactions.
- **Database Management System (DBMS)**: Software that interacts with the database, users, and applications to manage data effectively. Examples include MySQL, PostgreSQL, Oracle, and MongoDB.
- **Schema**: Defines the structure or blueprint of the database, detailing how data is organized (tables, fields, relationships).

Key Features of a Database

- ✅ Data Integrity and Consistency
- 🔐 Data Security and Access Control
- 🧠 Efficient Storage and Retrieval
- 💾 Backup and Recovery Support
- 🔄 Concurrency and Transaction Management
- 📈 Scalability and Performance

Types of Databases

- **Relational Databases (RDBMS)** – Store data in tables using rows and columns (e.g., MySQL, PostgreSQL).
- **NoSQL Databases** – Designed for unstructured or semi-structured data (e.g., MongoDB, Cassandra).
- **Distributed Databases** – Data is stored across different physical locations.
- **Cloud Databases** – Hosted in the cloud (e.g., Firebase, AWS RDS).
- **Object-oriented Databases** – Store data as objects, similar to OOP programming.

What is SQL?

SQL (Structured Query Language) is a standard programming language used for accessing and managing data in relational databases. It enables users to create, read, update, and delete (CRUD) data efficiently. It is particularly useful for handling structured data, which involves relations among entities and variables.

SQL was standardized by the American National Standards Institute (ANSI) in 1986. It also gained an International Standard designation from the International Organization for Standardization (ISO) and has been adopted as a standard by numerous governments and organizational bodies worldwide.

Features of SQL

- 🔍 Powerful querying using SELECT, WHERE, ORDER BY
- 🛠️ Data manipulation via INSERT, UPDATE, DELETE
- 🧱 Table and schema management using CREATE, ALTER, DROP
- 💳 Transaction control with COMMIT, ROLLBACK, SAVEPOINT
- 🔗 JOINs for combining data from multiple tables
- 🛡️ Access control using GRANT and REVOKE
- ⚡ Functions, subqueries, views, indexing, and stored procedures

Introduction

In the world of retail, understanding product sales, customer behavior, and profit trends is crucial. This project uses a real-world Superstore dataset to perform comprehensive sales analysis using SQL. The primary objective is to uncover patterns in sales, identify profitable products and regions, and generate actionable business insights through structured querying and data modeling.

Dataset Overview

The dataset used for this project contains transactional data of a retail superstore including details on customers, products, orders, sales, discounts, and profit. The original Superstone Sales dataset contains the following features:

No.	Column Name	Data Type	Description
1	Row ID	int64	Unique identifier for the row
2	Order ID	object	Unique order identifier
3	Order Date	object	Date the order was placed
4	Ship Date	object	Date the order was shipped
5	Ship Mode	object	Shipping mode used
6	Customer ID	object	Unique customer identifier
7	Customer	object	Name of the customer

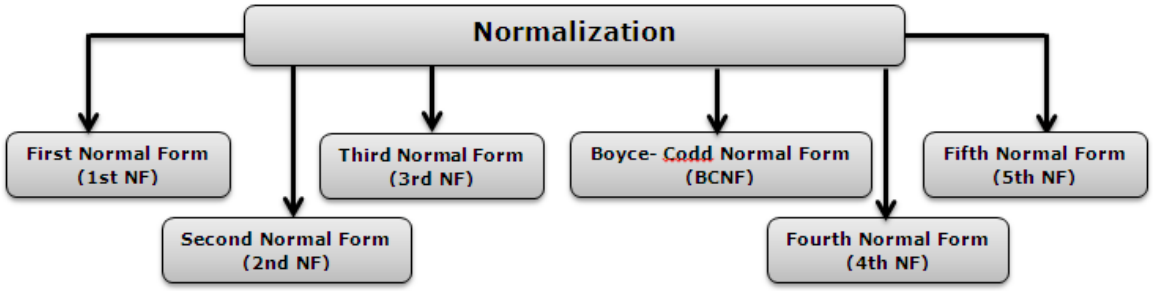
No.	Column Name	Data Type	Description
	Name		
8	Segment	object	Customer segment (e.g., Consumer)
9	Country	object	Country (e.g., United States)
10	City	object	City of the customer
11	State	object	State of the customer
12	Postal Code	int64	Postal code
13	Region	object	Region (e.g., West, East)
14	Product ID	object	Unique product identifier
15	Category	object	Product category
16	Sub-Category	object	Product sub-category
17	Product Name	object	Name of the product
18	Sales	float64	Sales amount in currency
19	Quantity	int64	Number of units sold
20	Discount	float64	Discount applied
21	Profit	float64	Profit earned on the sale

Data Normalization

What is Normalization in Databases?

Normalization is a process in relational database design that organizes data into multiple related tables to minimize redundancy and ensure data integrity.

Levels of Normalizations



1. **First Normal Form (1NF)** ensures that each column contains atomic values (indivisible) and that each record is unique. For example, a table with multiple values in a single cell is restructured so each value occupies its own row.
2. **Second Normal Form (2NF)** builds on 1NF by eliminating partial dependencies. This means non-key attributes must depend on the entire primary key, not just a part of it. For composite keys, data is split into separate tables to remove partial dependencies.
3. **Third Normal Form (3NF)** eliminates transitive dependencies, where non-key attributes depend on other non-key attributes. Each non-key attribute should depend only on the primary key.

4. **Boyce-Codd Normal Form (BCNF**) is a stricter version of 3NF. It ensures that for every functional dependency, the left-hand side is a superkey, addressing cases where 3NF might still allow redundancy.
5. **Fourth Normal Form (4NF)** handles multivalued dependencies, ensuring that no table contains two or more independent multivalued attributes.
6. **Fifth Normal Form (5NF)** eliminates join dependencies, ensuring that data cannot be further decomposed without losing information.

Then this Original Dataset was provided in a flat CSV format and later normalized into four CSV files such as:

- Customers.csv

- Products.csv

- Orders.csv

- OrderDetails.csv

Data Preprocessing

- Converted date formats to YYYY-MM-DD
- Removed duplicate records from Customers, Orders, and Products
- Ensured no missing values in key fields
- Split flat file into normalized CSV files for import into MySQL

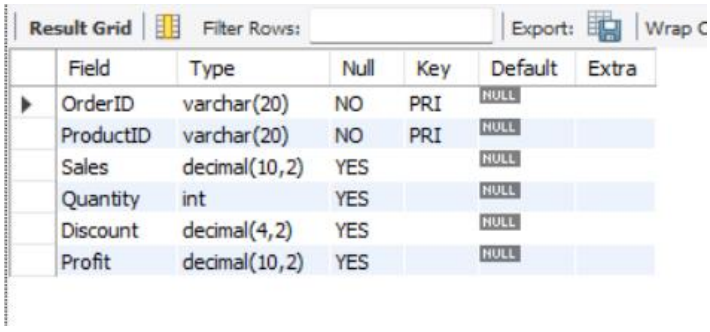
Database Design and Schema

To reduce redundancy and improve data integrity, the dataset was normalized into the following structure:

- Customers (CustomerID, CustomerName, Segment, City, State, Region, PostalCode)
- Products (ProductID, ProductName, Category, SubCategory)
- Orders (OrderID, OrderDate, ShipDate, ShipMode, CustomerID)
- OrderDetails (OrderID, ProductID, Sales, Quantity, Discount, Profit)

Customers Table Schema

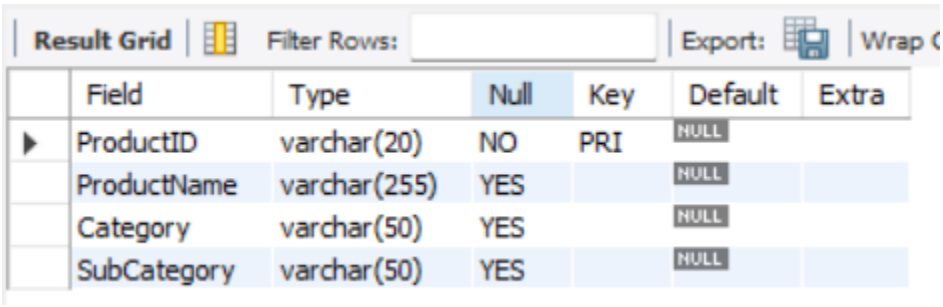
SQL Query: desc customers;


Result Grid | Filter Rows: | Export: | Wrap C

Field	Type	Null	Key	Default	Extra
OrderID	varchar(20)	NO	PRI	NULL	
ProductID	varchar(20)	NO	PRI	NULL	
Sales	decimal(10,2)	YES		NULL	
Quantity	int	YES		NULL	
Discount	decimal(4,2)	YES		NULL	
Profit	decimal(10,2)	YES		NULL	

Products Table Schema

SQL Query: desc products;

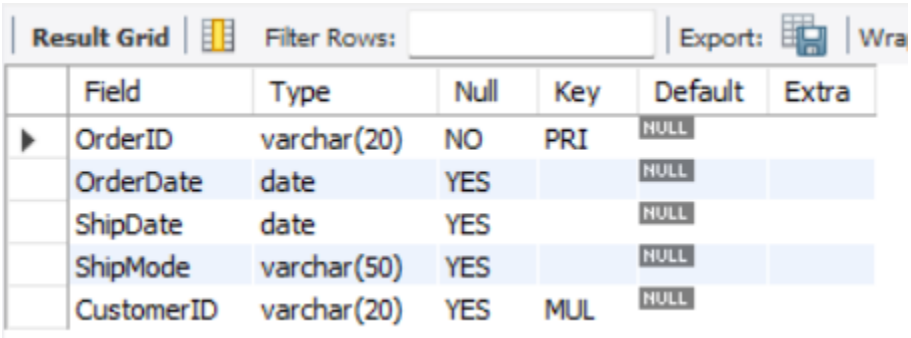


Result Grid | Filter Rows: | Export: | Wrap

Field	Type	Null	Key	Default	Extra
ProductID	varchar(20)	NO	PRI	NULL	
ProductName	varchar(255)	YES		NULL	
Category	varchar(50)	YES		NULL	
SubCategory	varchar(50)	YES		NULL	

Orders Table Schema

SQL Query: desc orders;

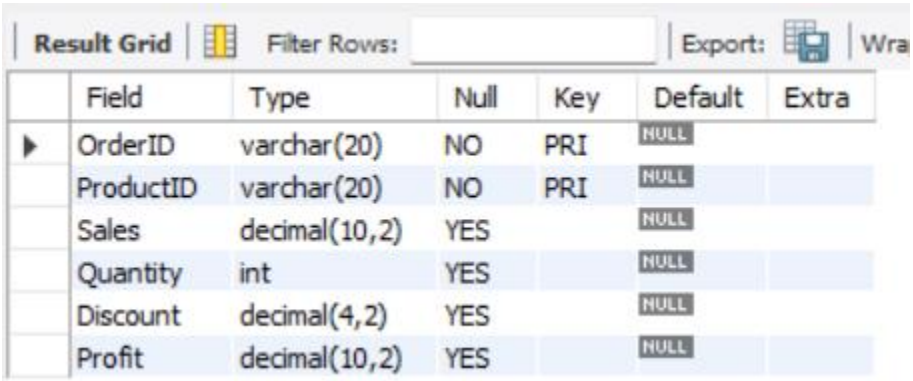


Result Grid | Filter Rows: | Export: | Wra

Field	Type	Null	Key	Default	Extra
OrderID	varchar(20)	NO	PRI	NULL	
OrderDate	date	YES		NULL	
ShipDate	date	YES		NULL	
ShipMode	varchar(50)	YES		NULL	
CustomerID	varchar(20)	YES	MUL	NULL	

Orderdetails Table Schema

SQL Query: desc orderdetails;

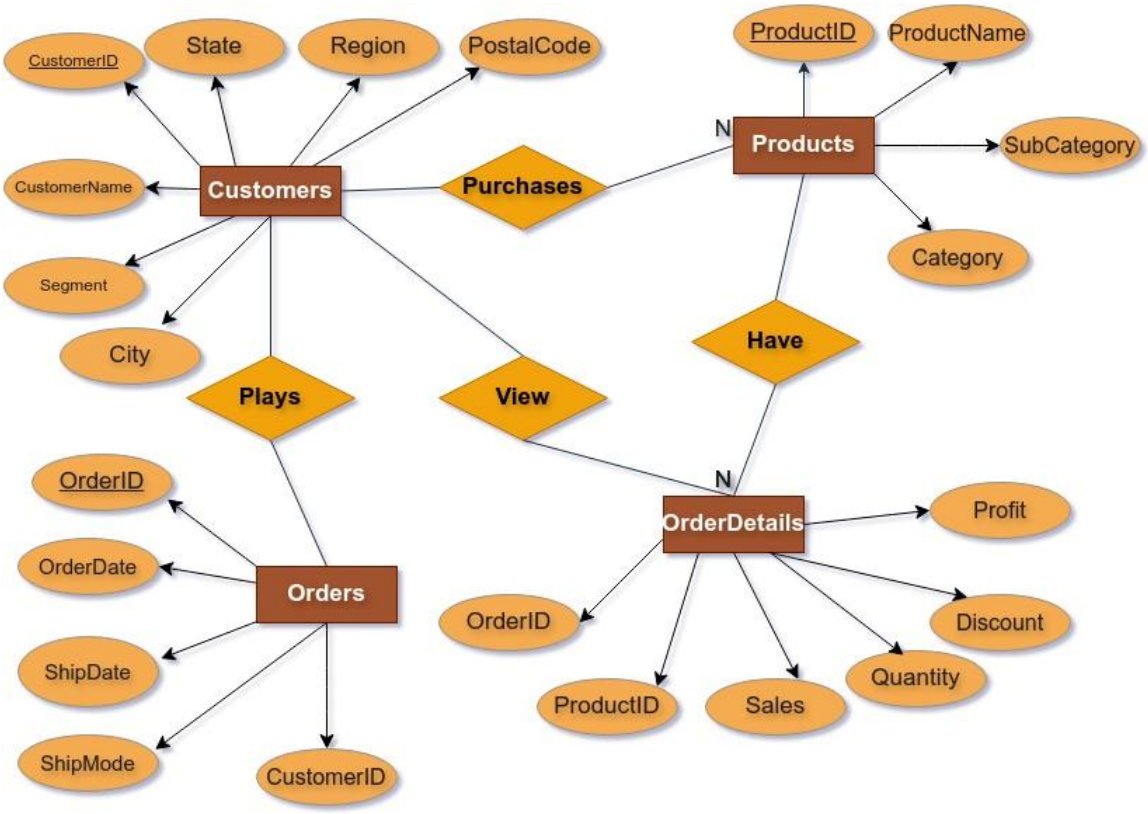


Result Grid | Filter Rows: | Export: | Wra

Field	Type	Null	Key	Default	Extra
OrderID	varchar(20)	NO	PRI	NULL	
ProductID	varchar(20)	NO	PRI	NULL	
Sales	decimal(10,2)	YES		NULL	
Quantity	int	YES		NULL	
Discount	decimal(4,2)	YES		NULL	
Profit	decimal(10,2)	YES		NULL	

ER Diagram

The Entity-Relationship (ER) Diagram for the Superstore Sales Analysis project illustrates a normalized relational database design. It identifies the main entities, their attributes, and the relationships between them to ensure efficient data storage, minimal redundancy, and support for complex queries.



ER DIAGRAM FOR SUPERSTONE SALES ANALYSIS

SQL Implementation

The SQL implementation involved:

- Create the Database:

```
CREATE DATABASE SuperstoreSales;
USE SuperstoreSales;
```

- Creating tables using SQL CREATE statements
  - **Customer table Creation**

```
CREATE TABLE Customers (
    CustomerID VARCHAR(20) PRIMARY KEY,
    CustomerName VARCHAR(100),
    Segment VARCHAR(50),
    City VARCHAR(50),
    State VARCHAR(50),
    Region VARCHAR(50),
    PostalCode INT
);
```

  - **Products table Creation**

```
CREATE TABLE Products (
    ProductID VARCHAR(20) PRIMARY KEY,
    ProductName VARCHAR(255),
    Category VARCHAR(50),
    SubCategory VARCHAR(50)
);
```

  - **Orders table Creation**

```
CREATE TABLE Orders (
    OrderID VARCHAR(20) PRIMARY KEY,
    OrderDate DATE,
    ShipDate DATE,
    ShipMode VARCHAR(50),
    CustomerID VARCHAR(20),
```

```
    FOREIGN KEY (CustomerID) REFERENCES
Customers(CustomerID)
);
```

- **Orderdetails table Creation**

```
CREATE TABLE OrderDetails (
    OrderID VARCHAR(20),
    ProductID VARCHAR(20),
    Sales DECIMAL(10, 2),
    Quantity INT,
    Discount DECIMAL(4, 2),
    Profit DECIMAL(10, 2),
    PRIMARY KEY (OrderID, ProductID),
    FOREIGN KEY (OrderID) REFERENCES
Orders(OrderID),
    FOREIGN KEY (ProductID) REFERENCES
Products(ProductID)
);
```

Importing The Data Into the Tables

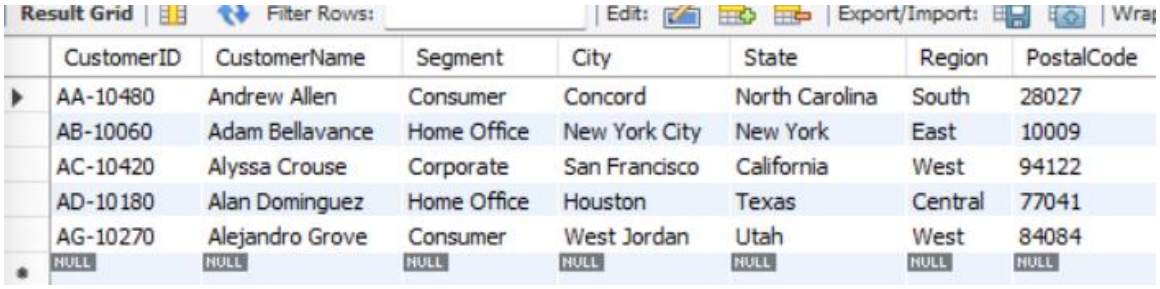
Using MySQL Workbench's Import Wizard:

- Open the MySQL Workbench
- Right-click on the table > Table Data Import Wizard
- Choose the corresponding CSV file (e.g., customers.csv)
- Match column headers to table fields

SQL queries and results

- **Basic Queries: SELECT, WHERE, ORDER BY**

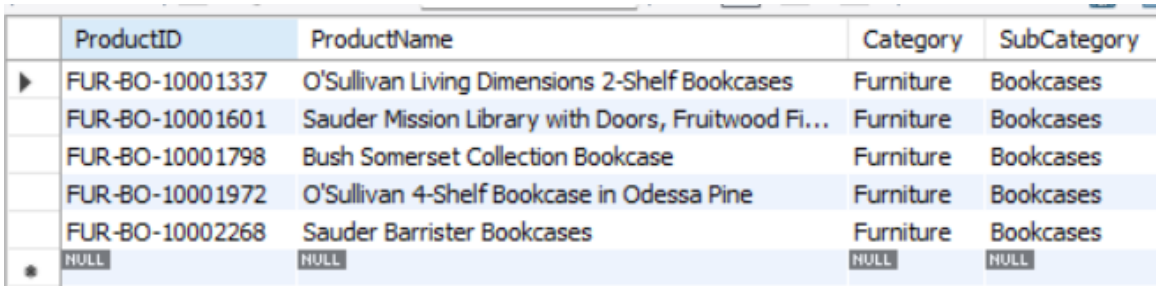
1. **select * from customers limit 5;**



Result Grid | Filter Rows: | Edit: | Export/Import:

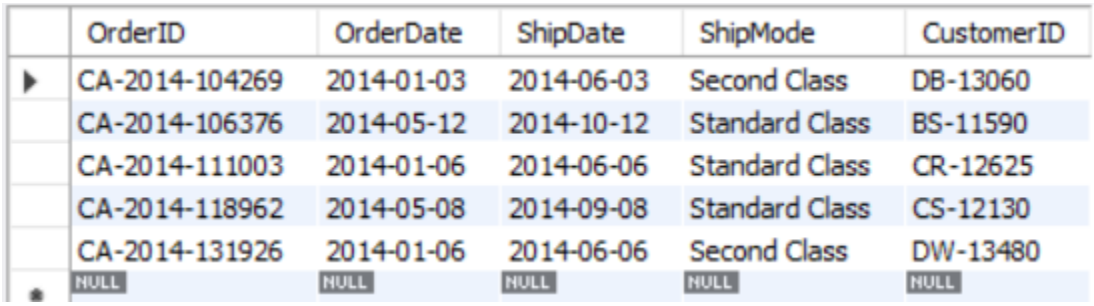
CustomerID	CustomerName	Segment	City	State	Region	PostalCode
AA-10480	Andrew Allen	Consumer	Concord	North Carolina	South	28027
AB-10060	Adam Bellavance	Home Office	New York City	New York	East	10009
AC-10420	Alyssa Crouse	Corporate	San Francisco	California	West	94122
AD-10180	Alan Dominguez	Home Office	Houston	Texas	Central	77041
AG-10270	Alejandro Grove	Consumer	West Jordan	Utah	West	84084
NULL	NULL	NULL	NULL	NULL	NULL	NULL

2. **select * from products limit 5;**



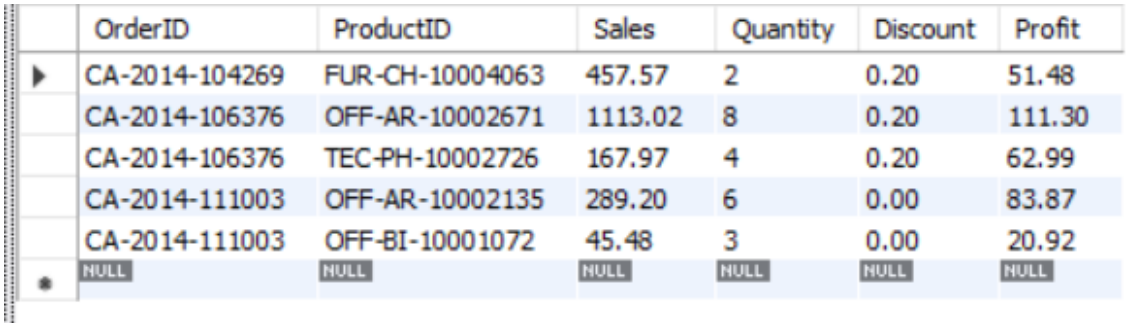
ProductID	ProductName	Category	SubCategory
FUR-BO-10001337	O'Sullivan Living Dimensions 2-Shelf Bookcases	Furniture	Bookcases
FUR-BO-10001601	Sauder Mission Library with Doors, Fruitwood Fi...	Furniture	Bookcases
FUR-BO-10001798	Bush Somerset Collection Bookcase	Furniture	Bookcases
FUR-BO-10001972	O'Sullivan 4-Shelf Bookcase in Odessa Pine	Furniture	Bookcases
FUR-BO-10002268	Sauder Barrister Bookcases	Furniture	Bookcases
NULL	NULL	NULL	NULL

3. **select * from orders limit 5;**



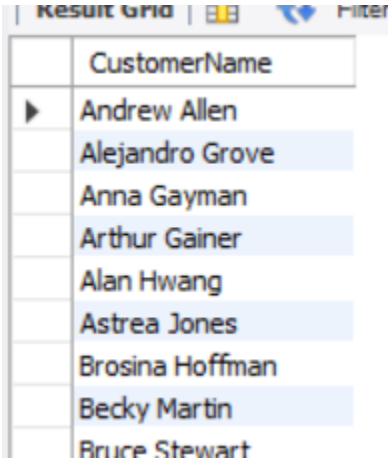
OrderID	OrderDate	ShipDate	ShipMode	CustomerID
CA-2014-104269	2014-01-03	2014-06-03	Second Class	DB-13060
CA-2014-106376	2014-05-12	2014-10-12	Standard Class	BS-11590
CA-2014-111003	2014-01-06	2014-06-06	Standard Class	CR-12625
CA-2014-118962	2014-05-08	2014-09-08	Standard Class	CS-12130
CA-2014-131926	2014-01-06	2014-06-06	Second Class	DW-13480
NULL	NULL	NULL	NULL	NULL

4. **select * from orderdetails limit 5;**



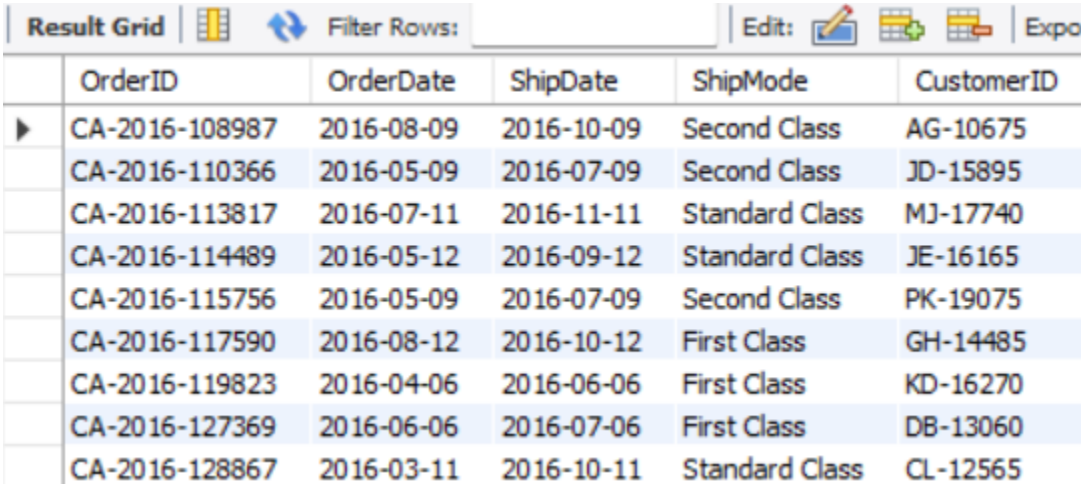
OrderID	ProductID	Sales	Quantity	Discount	Profit
CA-2014-104269	FUR-CH-10004063	457.57	2	0.20	51.48
CA-2014-106376	OFF-AR-10002671	1113.02	8	0.20	111.30
CA-2014-106376	TEC-PH-10002726	167.97	4	0.20	62.99
CA-2014-111003	OFF-AR-10002135	289.20	6	0.00	83.87
CA-2014-111003	OFF-BI-10001072	45.48	3	0.00	20.92
NULL	NULL	NULL	NULL	NULL	NULL

1. **Get customer names from "Consumer" segment**
   select CustomerName from customers where Segment="Consumer";



CustomerName
Andrew Allen
Alejandro Grove
Anna Gayman
Arthur Gainer
Alan Hwang
Astrea Jones
Brosina Hoffman
Becky Martin
Bruce Stewart

2. **Get all orders placed in the year 2016**
   select * from orders where YEAR(OrderDate)=2016;



Result Grid | Filter Rows: | Edit: | Expo

OrderID	OrderDate	ShipDate	ShipMode	CustomerID
CA-2016-108987	2016-08-09	2016-10-09	Second Class	AG-10675
CA-2016-110366	2016-05-09	2016-07-09	Second Class	JD-15895
CA-2016-113817	2016-07-11	2016-11-11	Standard Class	MJ-17740
CA-2016-114489	2016-05-12	2016-09-12	Standard Class	JE-16165
CA-2016-115756	2016-05-09	2016-07-09	Second Class	PK-19075
CA-2016-117590	2016-08-12	2016-10-12	First Class	GH-14485
CA-2016-119823	2016-04-06	2016-06-06	First Class	KD-16270
CA-2016-127369	2016-06-06	2016-07-06	First Class	DB-13060
CA-2016-128867	2016-03-11	2016-10-11	Standard Class	CL-12565

3. **Get all orders shipped using "Standard Class"**
   Select * from orders where ShipMode="Standard Class";

	OrderID	OrderDate	ShipDate	ShipMode	CustomerID
▶	CA-2014-106376	2014-05-12	2014-10-12	Standard Class	BS-11590
	CA-2014-111003	2014-01-06	2014-06-06	Standard Class	CR-12625
	CA-2014-118962	2014-05-08	2014-09-08	Standard Class	CS-12130
	CA-2014-134677	2014-06-10	2014-10-10	Standard Class	XP-21865
	CA-2014-139892	2014-08-09	2014-12-09	Standard Class	BM-11140
	CA-2014-164973	2014-04-11	2014-09-11	Standard Class	NM-18445
	CA-2015-110457	2015-02-03	2015-06-03	Standard Class	DK-13090
	CA-2015-110744	2015-07-09	2015-12-09	Standard Class	HA-14920
	CA-2015-122756	2015-03-12	2015-07-12	Standard Class	DK-13225

4. **Get top 5 highest-profit order lines**

   select * from orderdetails order by profit DESC limit 5;

	OrderID	ProductID	Sales	Quantity	Discount	Profit
▶	CA-2014-164973	TEC-MA-10002927	3991.98	2	0.00	1995.99
	CA-2016-129714	OFF-BI-10004995	4355.17	4	0.20	1415.43
	CA-2016-114489	FUR-CH-10000454	1951.84	8	0.00	585.55
	CA-2014-131926	FUR-CH-10004063	2001.86	7	0.00	580.54
	CA-2014-131926	OFF-AP-10002945	1503.25	5	0.00	496.07
*	NULL	NULL	NULL	NULL	NULL	NULL

5. **Get unique product categories**

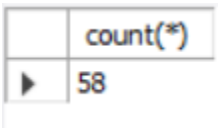
   select DISTINCT productName from products;

	productName
▶	O'Sullivan Living Dimensions 2-Shelf Bookcases
	Sauder Mission Library with Doors, Fruitwood Fi...
	Bush Somerset Collection Bookcase
	O'Sullivan 4-Shelf Bookcase in Odessa Pine
	Sauder Barrister Bookcases
	Atlantic Metals Mobile 3-Shelf Bookcases, Custo...
	Atlantic Metals Mobile 4-Shelf Bookcases, Custo...
	Bush Mission Pointe Library
	O'Sullivan 2-Door Barrister Bookcase in Odessa ...

- **Group By And Aggregations: COUNT, SUM, AVG, MAX, MIN**

1. **Count total number of orders**

select count(*) from orders;



	count(*)
▶	58

2. **Total sales by product category**

select p.Category,sum(od.Sales) as total_sales from products as p join orderdetails as od on p.productID=od.ProductID group by p.Category;

	Category	total_sales
▶	Furniture	18250.81
	Office Supplies	14550.17
	Technology	20978.16
	5 1/2"" X 4""	62.65
	Ream"	91.36
	Black"	9.71

3. **Average discount per sub-category**

select p.SubCategory,avg(Discount) as AVG_Discount from products as p join orderdetails as od on p.productID=od.productID group by p.SubCategory;

SubCategory	AVG_Discount
Chairs	0.127273
Art	0.071429
Phones	0.150000
Binders	0.341176
Paper	0.023529
Appliances	0.300000
Storage	0.073684
Tables	0.328571
Accessories	0.088889

4. **Total number of orders placed by each customer**

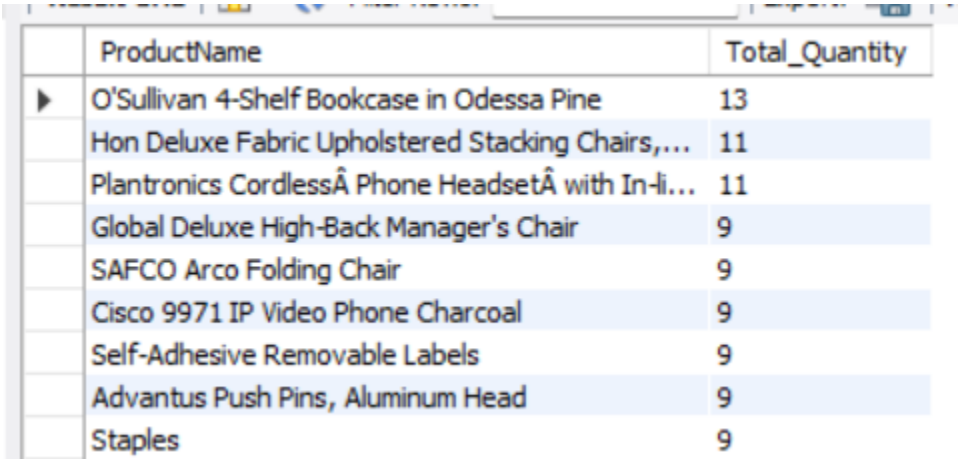
```
select c.CustomerName,Count(o.OrderID) as no_of_orders from
customers as c join orders as o on c.CustomerID=o.CustomerID
group by c.CustomerName;
```

CustomerName	no_of_orders
Raymond Buch	1
Roger Barcio	1
Sung Pak	1
Tamara Dahlen	1
Troy Staebel	1
Ted Trevino	1
Victoria Brennan	2
Victoria Wilson	1
Xylona Preis	1

5. **Total quantity sold per product**

```
select p.ProductName,sum(od.Quantity) as Total_Quantity from
products as p join orderdetails as od on
p.ProductID=od.ProductId
group by p.ProductName
```

ORDER BY Total_Quantity DESC;



ProductName	Total_Quantity
O'Sullivan 4-Shelf Bookcase in Odessa Pine	13
Hon Deluxe Fabric Upholstered Stacking Chairs,...	11
Plantronics CordlessÂ Phone HeadsetÂ with In-li...	11
Global Deluxe High-Back Manager's Chair	9
SAFCO Arco Folding Chair	9
Cisco 9971 IP Video Phone Charcoal	9
Self-Adhesive Removable Labels	9
Advantus Push Pins, Aluminum Head	9
Staples	9

6. Average sales per month

select month(o.OrderDate),avg(od.Sales) as AVG_Sales

from orderdetails as od join orders as o on od.OrderID=o.OrderID

group by month(o.OrderDate)

order by month(o.OrderDate) ASC;

month(o.OrderDate)	AVG_Sales
1	359.143871
2	380.279000
3	137.180909
4	350.785625
5	284.007826
6	161.815714
7	305.990000
8	670.465000
9	165.231000

7. **Highest and lowest profit per sub-category**

select p.SubCategory,max(od.profit) as Maximum_profit,min(od.profit) as Minimum_profit

from products as p join orderdetails as od on p.ProductID=od.ProductID

group by p.SubCategory;

SubCategory	Maximum_profit	Minimum_profit
Chairs	585.55	-24.86
Art	111.30	0.50
Phones	123.47	-14.70
Binders	1415.43	-115.72
Paper	65.73	2.69
Appliances	496.07	-453.85
Storage	207.15	-58.63
Tables	165.38	-248.25
Accessories	129.60	-7.72

- **Grouping: GROUP BY, HAVING**

1. **Get sub-categories with total sales over ₹5,000**

   select p.SubCategory,sum(od.Sales) as Total_sales from

   orderdetails as od join products as p on od.ProductID=p.ProductID

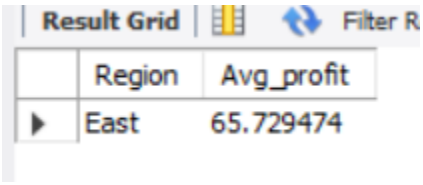
   group by p.SubCategory

   having Total_sales>5000;

SubCategory	Total_sales
Chairs	8104.36
Binders	5275.61
Machines	13822.27

2. **Regions with average profit above ₹50**

```
select c.Region,avg(od.Profit) as Avg_profit
from orderdetails as od join orders as o on
od.OrderID=o.OrderID
join customers as c on o.CustomerID=c.CustomerID
group by c.Region
having Avg_profit>50;
```



Result Grid | Filter R

	Region	Avg_profit
▶	East	65.729474

- **Multi-table analysis: JOIN, LEFT JOIN, RIGHT JOIN, INNER JOIN**
    1. **INNER JOIN**

```
SELECT O.OrderID, C.CustomerName, P.ProductName,
OD.Sales
FROM Orders O
JOIN Customers C ON O.CustomerID = C.CustomerID
JOIN OrderDetails OD ON O.OrderID = OD.OrderID
JOIN Products P ON OD.ProductID = P.ProductID;
```

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

OrderID	CustomerName	ProductName	Sales
CA-2016-129714	Adam Bellavance	Acco Hanging Data Binders	3.05
CA-2016-129714	Adam Bellavance	GBC DocuBind P400 Electric Binding System	4355.17
CA-2016-129714	Adam Bellavance	Xerox 1881	24.56
CA-2016-129714	Adam Bellavance	Sabrent 4-Port USB 2.0 Hub	6.79
CA-2016-108987	Anna Gayman	Riverside Palais Royal Lawyers Bookcase, Royal...	2396.27
CA-2016-108987	Anna Gayman	Contico 72"H Heavy-Duty Storage System"	131.14
CA-2016-108987	Anna Gayman	Super Decoflex Portable Personal File	35.95
CA-2016-108987	Anna Gayman	Sony 64GB Class 10 Micro SDHC R40 Memory Card	57.58
CA-2015-144253	Alan Schoenberger	Electrix 20W Halogen Replacement Bulb for Zoo...	26.80

2. LEFT JOIN

→ Returns all rows from the left table + matching rows from the right.

```
select c.CustomerName,o.OrderID from
customers as c left join orders as o on
c.CustomerID=o.CustomerID;
```

CustomerName	OrderID
Andrew Allen	NULL
Adam Bellavance	CA-2016-129714
Alyssa Crouse	NULL
Alan Dominguez	NULL
Alejandro Grove	NULL
Andrew Gjertsen	NULL
Andy Gerbode	NULL
Anna Gayman	CA-2016-108987
Arthur Gainer	NULL

3. RIGHT JOIN

→ Returns all rows from the right table + matching rows from the left.

```
select o.OrderID,od.OrderID from
orderdetails as od right join orders as o on
o.OrderID=od.OrderID;
```

Result Grid | Filter Rows:

	OrderID	OrderID
▶	CA-2016-129714	CA-2016-129714
	CA-2016-129714	CA-2016-129714
	CA-2016-129714	CA-2016-129714
	CA-2016-129714	CA-2016-129714
	CA-2016-108987	CA-2016-108987
	CA-2016-108987	CA-2016-108987
	CA-2016-108987	CA-2016-108987
	CA-2016-108987	CA-2016-108987
	CA-2015-144253	CA-2015-144253

- **Subqueries and nested logic**

  1. Get products with sales greater than average

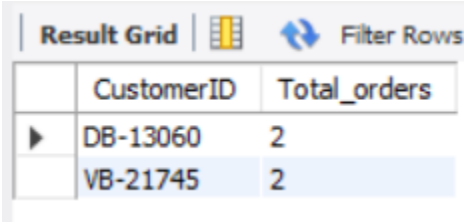
  select p.ProductName,sum(od.Sales) as Total_sales from products as p join orderdetails as od on p.ProductID=od.ProductID
  group by p.ProductID having Total_sales>(select Avg(Sales) from orderdetails);

Result Grid | Filter Rows: | Export:

	ProductName	Total_sales
▶	Global Deluxe High-Back Manager's Chair	2459.43
	Hunt BOSTON Model 1606 High-Volume Electric ...	1113.02
	Honeywell Enviracaire Portable HEPA Air Cleane...	1503.25
	SAFCO Arco Folding Chair	1740.06
	Lexmark MX611dhe Monochrome Laser Printer	8159.95
	Samsung Galaxy Note 3	703.97
	Canon imageCLASS MF7460 Monochrome Digital...	3991.98
	ImationÂ Secure+ Hardware Encrypted USB 2.0...	408.74
	Hon Racetrack Conference Tables	787.53

2. Customers who placed more than 1 orders

```
select CustomerID,count(OrderID) as Total_orders
from orders group by CustomerID
having Total_orders>1;
```



Result Grid | Filter Rows

CustomerID	Total_orders
DB-13060	2
VB-21745	2

Insights and Recommendations

- Technology is the top-selling category — recommend priority marketing and stocking
- West region generates the most profit — recommend regional focus in campaigns
- Certain sub-categories like Phones, Chairs are high-profit — bundle and upsell opportunities

Challenges and Resolutions

- Challenge: Flat, redundant data
  Solution: Normalized to 3NF and created relational model
- Challenge: Date import errors
  Solution: Reformatted to correct MySQL DATE format
- Challenge: Foreign key constraint errors
  Solution: Maintained correct table import order

Conclusion

This capstone project demonstrated how structured query language (SQL) and relational database design can be used to extract meaningful business insights from raw transactional data. With a clean schema, effective querying, and proper preprocessing, the Superstore dataset revealed trends and patterns that can directly inform business strategy.