Group 11 Assignment

Course: Business Intelligence for Data Science (880682-M-6)

Semester: Fall 2024/Block 2

Group 11

Business Topic: Health Trends in Pharmacy: Consumer Behavior and Market Insights Across Dutch Cities

Contributors:

Hui-Chiao Huang, 2136016

Dilruba Turan, 2138180

Boya Chuang, 2137743

Yu-Ling Chen 2135187

CONTENT:

1. Describe motivation and goals for building DW into an enterprise of choice – 2 points.
2. Create and describe the OLTP schema with relationships, entities and attributes with screenshots and submit the file of the model (.mvb file) – 2 points.
3. Transfer the OLTP schema diagram into Tables in MySQL and load with the .csv files – 2 points.
4. Create and transfer DW schema into Tables in MySQL – 2 points.
5. Describe the dimensions and fact for the DW and chosen variant, i.e. star/snowflake/constellation with screenshots and submit the file of the model (.mvb file) – 2 points.
6. Describe and present (both with screenshots and by submitting the .ktr files) three ETL processes used for transformations (more details in Section 6) – 6 points.
7. Create and present one View (both with screenshot and by submitting the file) using some of the keywords (more details in Section 7) – 2 points.
8. Create and present one Report (both with screenshot and by submitting the file) – 2 points.

*Note: Beside submitting this document, it is also required to submit the files for the models OLTP + DW (.mwb files), transformations (.ktr files), query for the view (.sql or any text file), report file (.prpt) and .csv files of the data. Finally, you are submitting only **one** .zip folder where you have one file (this document) + folder with the files mentioned + folder with .csv files of the data.

Only one student from the group is submitting on behalf of all. Name of the .zip folder: **GroupX_Project_BI4DSS. If there are more files submitted for one group, one randomly chosen will be taken for evaluation. -** Delete this note before submitting.

1. Motivation & Goals

The primary motivation for creating this Data Warehouse (DW) is to provide a centralized repository; integrating datasets like customer details, product details, location, and sales transactions to track sales trends and monitor city-wise performance. The aim is to reduce redundancy, avoid inconsistency, and help in maintaining and tracking sales trends at the product level and within the locations in this chain. It enables the pharmacy to uncover customer purchasing behaviors, track medicine costs, analyze spending differences between ill and non-ill customers, and calculate profits accurately. It provides insights into whether supplements or medicines drive higher revenue and consolidates data for audits to ensure regulatory compliance.

With these insights, the pharmacy can identify top-performing products, monitor city-specific sales, address gaps, and optimize supply chain decisions. This empowers the pharmacy to enhance service quality and adapt to healthcare trends while addressing key questions like: What are the top-selling medicines, and how can supply chain processes be optimized?

2. Source OLTP schema

In our OLTP model, there are five tables with 500 rows each. Each table has its own Primary Key.

The Sales table is the fact table, connected with Customer_Details, Location, Product Details, and Date. Which indicates that a customer can have one or more sales records(Foreign Key: Sale.Customer_Details_CustomerID). The same for Location with Sales, which means a specific chain store can associate with multiple sales records. There are two main product types, associated with Sales to see which product sells the best and then know which type is the most purchased. Finally, the specific date can be linked to multiple sales records, hence the foreign Key: Sale.DateID references Time.DateID. This structure forms an OLTP schema where Primary Keys serve as unique identifiers for each table, and Foreign Keys connect the tables to support data consistency and efficient querying.

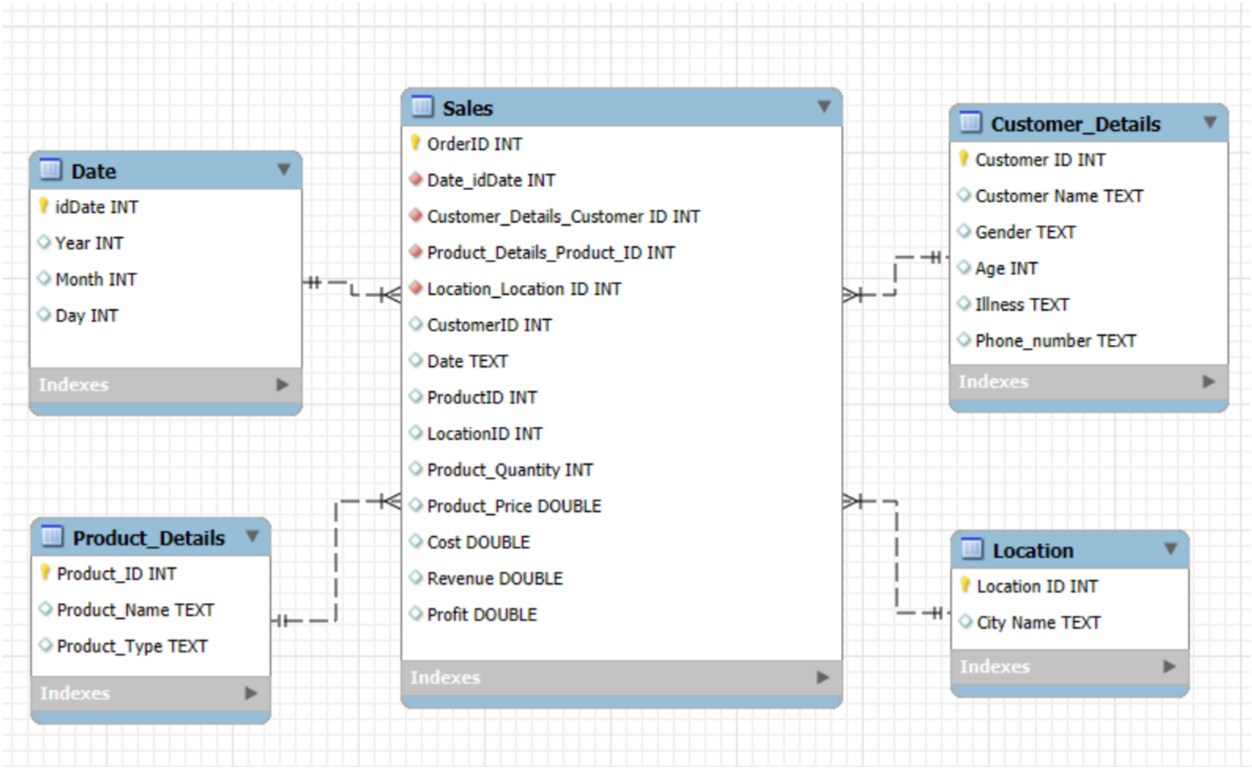
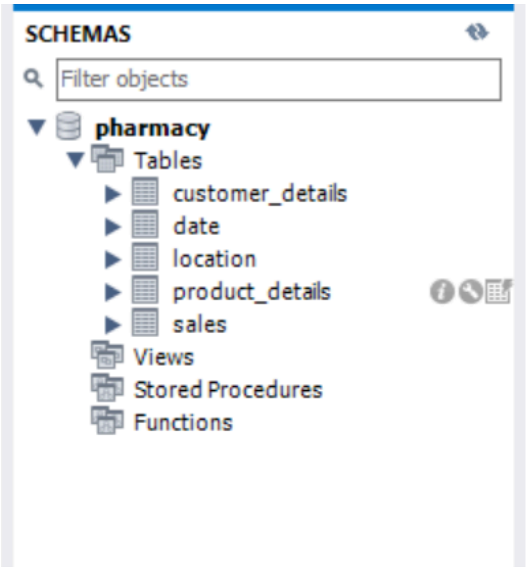


Figure 1. OLTP schema Pharmacy.

3. Transferring OLTP schema to Tables in MySQL

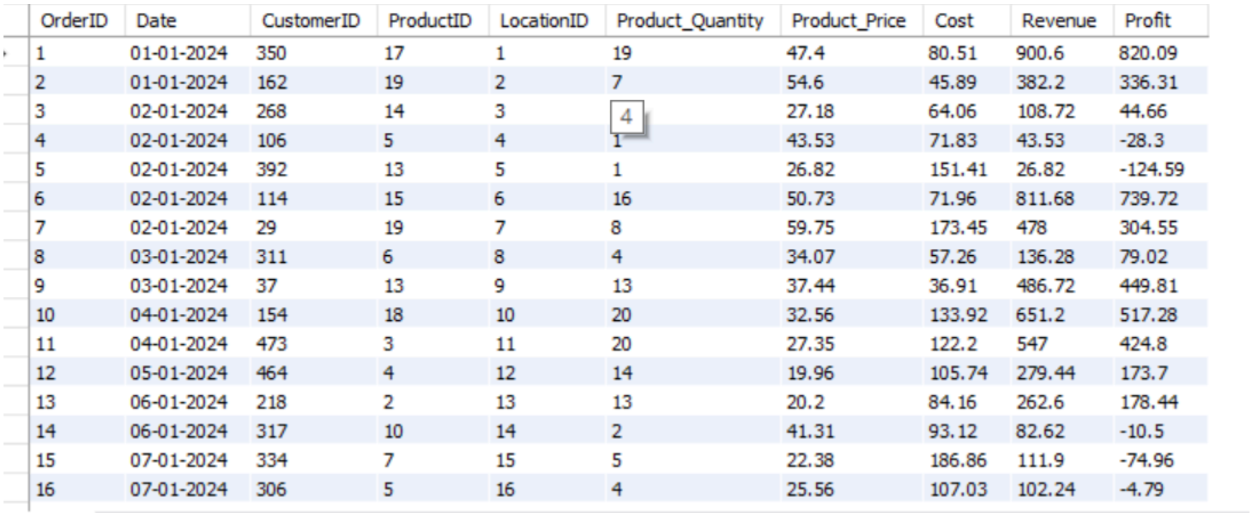

SCHEMAS

Filter objects

- pharmacy
  - Tables
    - customer_details
    - date
    - location
    - product_details
    - sales
  - Views
  - Stored Procedures
  - Functions

:

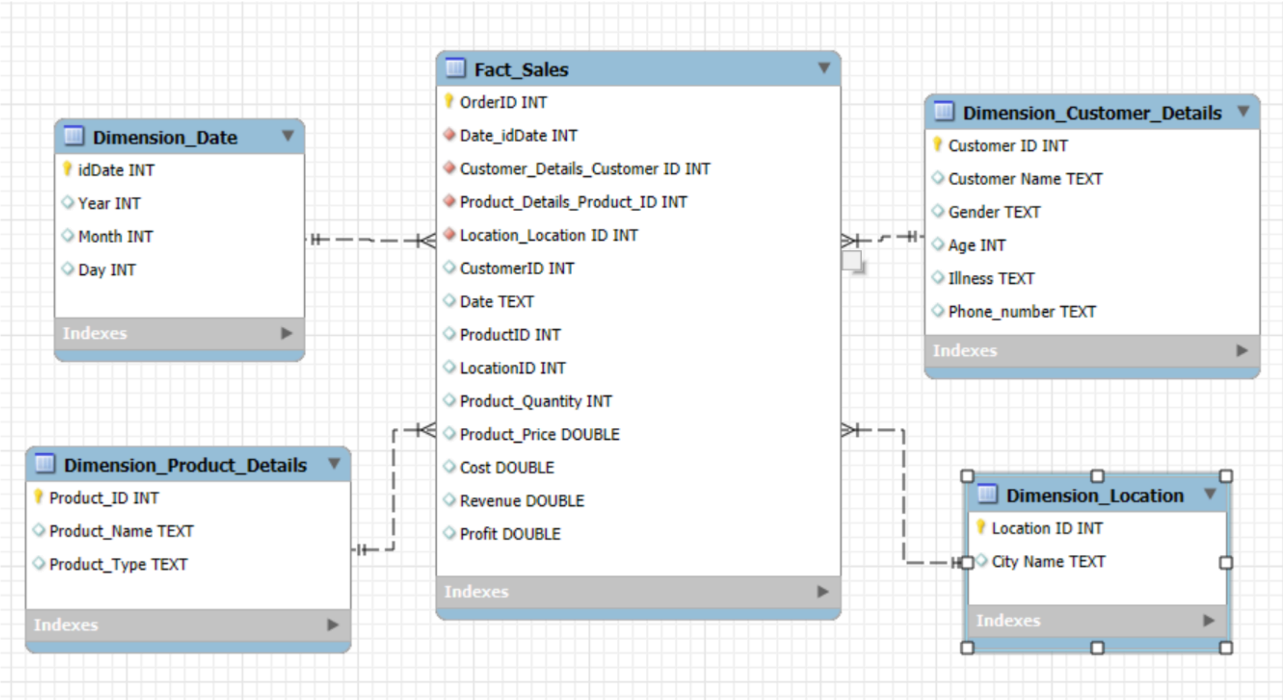
Figure 2. Transferring the OLTP into Tables.



OrderID	Date	CustomerID	ProductID	LocationID	Product_Quantity	Product_Price	Cost	Revenue	Profit
1	01-01-2024	350	17	1	19	47.4	80.51	900.6	820.09
2	01-01-2024	162	19	2	7	54.6	45.89	382.2	336.31
3	02-01-2024	268	14	3	4	27.18	64.06	108.72	44.66
4	02-01-2024	106	5	4	1	43.53	71.83	43.53	-28.3
5	02-01-2024	392	13	5	1	26.82	151.41	26.82	-124.59
6	02-01-2024	114	15	6	16	50.73	71.96	811.68	739.72
7	02-01-2024	29	19	7	8	59.75	173.45	478	304.55
8	03-01-2024	311	6	8	4	34.07	57.26	136.28	79.02
9	03-01-2024	37	13	9	13	37.44	36.91	486.72	449.81
10	04-01-2024	154	18	10	20	32.56	133.92	651.2	517.28
11	04-01-2024	473	3	11	20	27.35	122.2	547	424.8
12	05-01-2024	464	4	12	14	19.96	105.74	279.44	173.7
13	06-01-2024	218	2	13	13	20.2	84.16	262.6	178.44
14	06-01-2024	317	10	14	2	41.31	93.12	82.62	-10.5
15	07-01-2024	334	7	15	5	22.38	186.86	111.9	-74.96
16	07-01-2024	306	5	16	4	25.56	107.03	102.24	-4.79

Figure 3. Imported data for the Sales table.

4. DW schema



Fact_Sales
- OrderID INT
- Date_idDate INT
- Customer_Details_Customer ID INT
- Product_Details_Product_ID INT
- Location_Location ID INT
- CustomerID INT
- Date TEXT
- ProductID INT
- LocationID INT
- Product_Quantity INT
- Product_Price DOUBLE
- Cost DOUBLE
- Revenue DOUBLE
- Profit DOUBLE

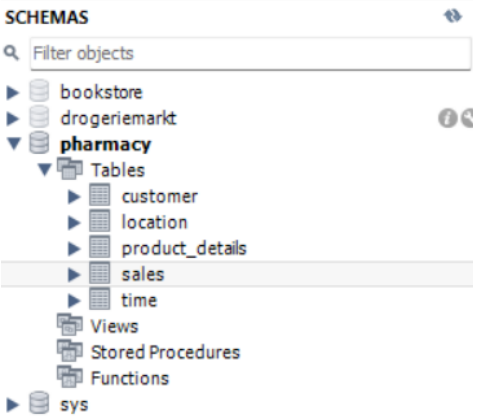
Dimension_Date
- idDate INT
- Year INT
- Month INT
- Day INT

Dimension_Customer_Details
- Customer ID INT
- Customer Name TEXT
- Gender TEXT
- Age INT
- Illness TEXT
- Phone_number TEXT

Dimension_Product_Details
- Product_ID INT
- Product_Name TEXT
- Product_Type TEXT

Dimension_Location
- Location ID INT
- City Name TEXT

Figure 4. DW schema for Pharmacy.



SCHEMAS

Filter objects

- bookstore
- drogeriemarkt
- **pharmacy**
  - Tables
    - customer
    - location
    - product_details
    - sales
    - time
  - Views
  - Stored Procedures
  - Functions
- sys

Figure 5. Transferring the DW into Tables.

5. Dimensions & Facts of DW

There are four key dimensions in our model: Customer Details, Product Details, Location, and Date.

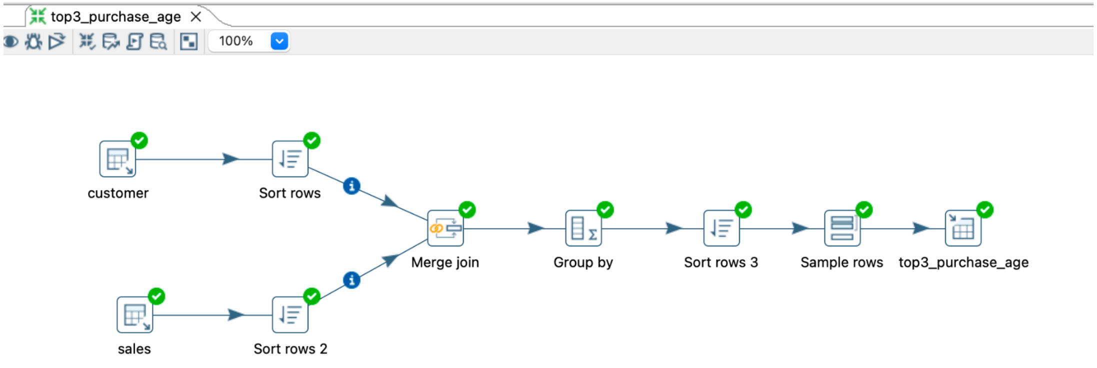
- In the Customer dimension, the most important factor is whether the customer is ill or not. This will help us compare spending patterns between customers who are ill and those who are healthy.
- In the Product dimension, we categorize all products into two types: Health Supplements and Medicines. This allows us to analyze consumer preferences and identify market trends.
- In the Location and Date dimensions, we aim to determine which city and which specific date over the past six months have generated the highest sales, which helps us understand where and when the pharmacy has the greatest profit potential. This aligns with our business objectives.

For the Sales Table, we focus on Profit, which is derived from the Revenue and Cost columns. Our goal is to calculate the daily earnings of a chain store, integrated with the Location, Customer, and Product dimensions to answer the key questions. Each row in the sales table represents a specific transaction, with details such as the date, customer ID, and the purchased product. By multiplying the quantity by the product price and subtracting the cost, we calculate the profit for each transaction. While there may be multiple customers buying the same product on the same day, we calculate the product cost for each transaction, as the cost includes the overall expenses associated with importing the product, such as employee salaries and electricity fees.

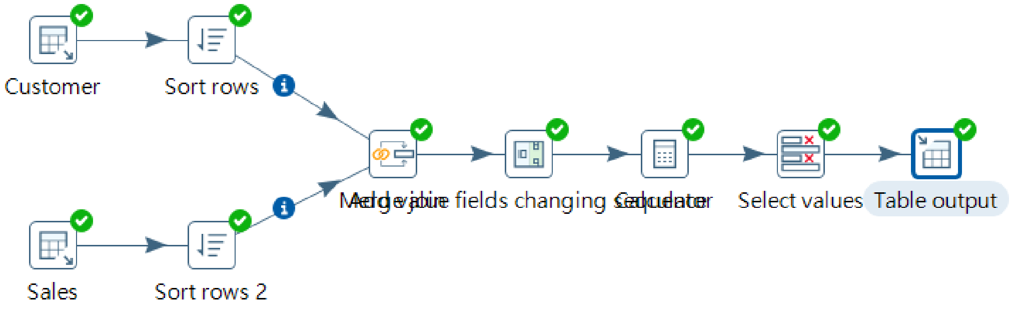
This method allows us to identify which city or date has the best overall sales records and provides insight into which type of product consumers purchase most. We can even analyze whether the consumer is ill or not, contributing to a more effective market trend analysis for the business.

We have chosen the Star Schema model, where the Sales fact table is at the center, surrounded by the Customer, Location, Product, and Date dimension tables. This schema is ideal for simple queries and fast data retrieval, as there are no complex relationships between the dimension tables, making it easy to understand and navigate for analytical purposes.

6. ETL transformations in Pentaho Data Integration

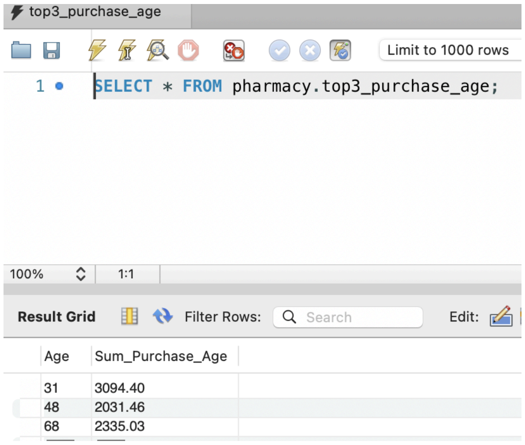


Pentaho: top3_purchase_age



Pentaho: Profit_Margin_Per_Sale

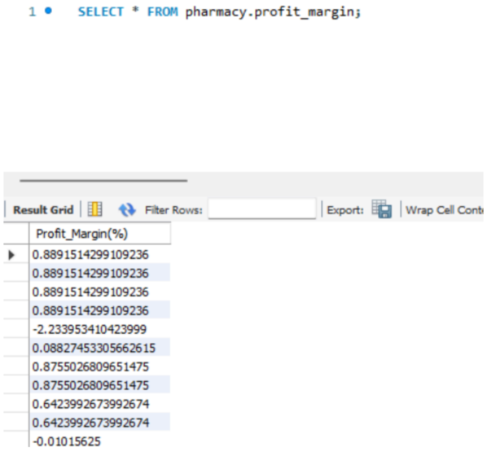
MySQL: top3_purchase_age



```
SELECT * FROM pharmacy.top3_purchase_age;
```

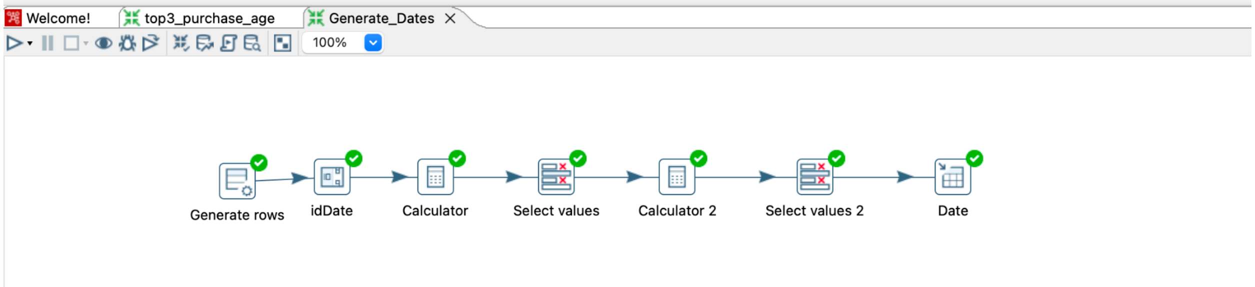
Age	Sum_Purchase_Age
31	3094.40
48	2031.46
68	2335.03

MySQL: Profit_Margin_Per_Sale



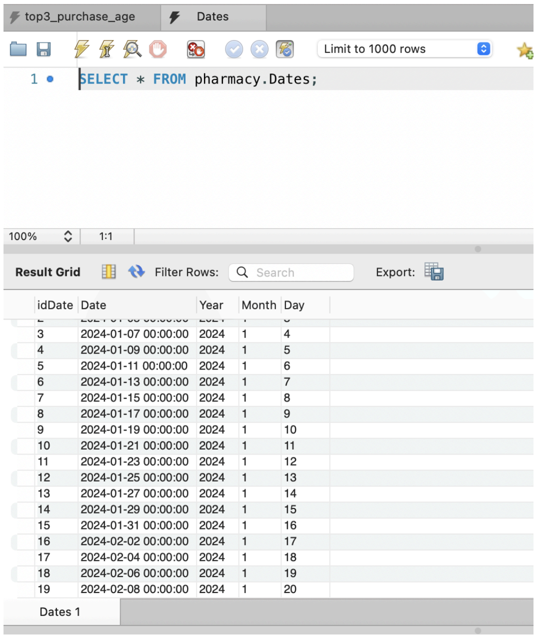
```
SELECT * FROM pharmacy.profit_margin;
```

Profit_Margin(%)
0.8891514299109236
0.8891514299109236
0.8891514299109236
0.8891514299109236
-2.233953410423999
0.08827453305662615
0.8755026809651475
0.8755026809651475
0.6423992673992674
0.6423992673992674
-0.01015625



Pentaho: Dimention_Time

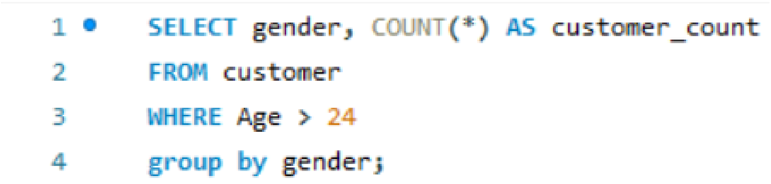
MySQL: Dimension_Time



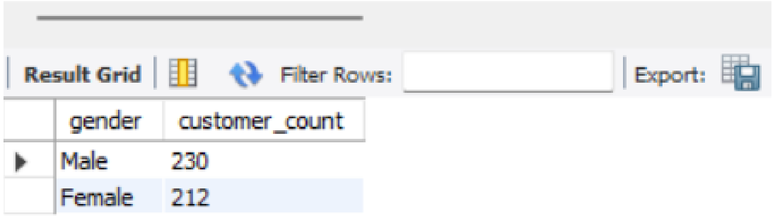
```
SELECT * FROM pharmacy.Dates;
```

idDate	Date	Year	Month	Day
2	[illegible]	[illegible]	1	3
3	2024-01-07 00:00:00	2024	1	4
4	2024-01-09 00:00:00	2024	1	5
5	2024-01-11 00:00:00	2024	1	6
6	2024-01-13 00:00:00	2024	1	7
7	2024-01-15 00:00:00	2024	1	8
8	2024-01-17 00:00:00	2024	1	9
9	2024-01-19 00:00:00	2024	1	10
10	2024-01-21 00:00:00	2024	1	11
11	2024-01-23 00:00:00	2024	1	12
12	2024-01-25 00:00:00	2024	1	13
13	2024-01-27 00:00:00	2024	1	14
14	2024-01-29 00:00:00	2024	1	15
15	2024-01-31 00:00:00	2024	1	16
16	2024-02-02 00:00:00	2024	1	17
17	2024-02-04 00:00:00	2024	1	18
18	2024-02-06 00:00:00	2024	1	19
19	2024-02-08 00:00:00	2024	1	20

7. View in MySQL

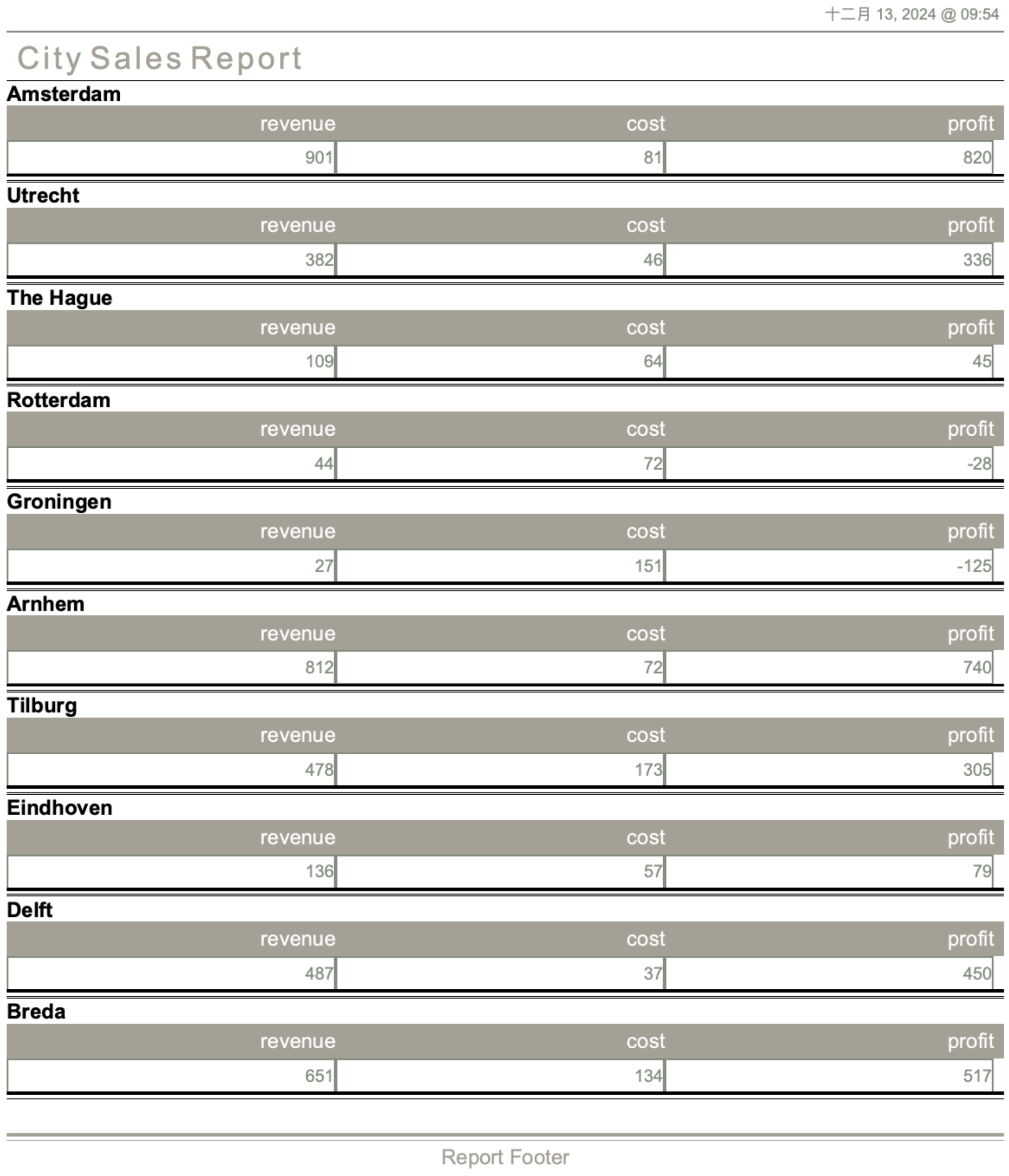


```
SELECT gender, COUNT(*) AS customer_count
FROM customer
WHERE Age > 24
group by gender;
```



gender	customer_count
Male	230
Female	212

8. Report in Pentaho Report Designer



十二月 13, 2024 @ 09:54

City Sales Report

Amsterdam

revenue	cost	profit
901	81	820

Utrecht

revenue	cost	profit
382	46	336

The Hague

revenue	cost	profit
109	64	45

Rotterdam

revenue	cost	profit
44	72	-28

Groningen

revenue	cost	profit
27	151	-125

Arnhem

revenue	cost	profit
812	72	740

Tilburg

revenue	cost	profit
478	173	305

Eindhoven

revenue	cost	profit
136	57	79

Delft

revenue	cost	profit
487	37	450

Breda

revenue	cost	profit
651	134	517

Report Footer

RECAP: Submit the .zip folder containing this document + .mwb files for OLTP and DW + .csv files for imported data + folder with 3 transformations where you have .ktr files + sql query for view + report file .prpt.

DATA: Either use mockaroo.com to generate the data artificially for your model or find EER diagram for a business topic you want to do and populate again with the data from mockaroo in the form you want.