**Final project | business intelligence course   
Semester B | 2023**

Group Q:

Lior Shapira - 315549444

Naomi Shabat - 318298213

Tal Tenenbaum - 316299346

1. **Business question + KPIs**

E-commerce has become a new channel to support businesses development. Through e-commerce, businesses can get access and establish a wider market presence by providing cheaper and more efficient distribution channels for their products or services. E-commerce has also changed the way people shop and consume products and services. Many people are turning to their computers or smart devices to order goods, which can easily be delivered to their homes.

We normalized the dataset into two different tables:

**Table 1 – Transaction\_info**Columns:

* TransactionNo
* Date
* ProductNo
* ProductName
* CustomerNo
* Country

**Table 2 – Transaction\_income**Columns:

* TransactionNo
* Price
* Quantity

**Business Question 1 -** What are the most popular products and categories among our customers, and how do these vary across different countries and time periods?

KPIs:

1. **Total sales revenue by product category**

* Specific - product category.
* Measurable - measures the total revenue generated by each product category.
* Achievable - can be tracked.
* Relevant – relevant to finding the most popular products and categories.
* Time-bound - time periods.

1. **Customer retention rate by country**

* Specific – country
* Measurable - measures the percentage of customers who make a repeat purchase within a specified time period, broken down by country.
* Achievable - can be tracked.
* Relevant – relevant to finding popularity across different countries.
* Time-bound - within a specified time period.

1. **Average number of items per order**

* Specific – transaction.
* Measurable - measures the average number of items purchased per order
* Achievable - can be tracked.
* Relevant - can help identify trends in customer behavior and preferences.
* time-bound - time periods.

**Business Question 2** - Which country is the most profitable, and how can the company increase sales in countries with lower customer engagement?

KPIs for the second business question:

1. **Customer transaction rate by country**

* Specific – country.
* Measurable - customer acquisition rate.
* Achievable - can be tracked.
* Relevant – relevant to finding most profitable countries.
* time-bound - time periods.

1. **Sales revenue generated by each country**

* Specific – country.
* Measurable – measuring sales revenue.
* Achievable - can be tracked.
* Relevant - relevant to finding most profitable countries.
* time-bound - time periods.

1. **Customer retention rate by country**

* Specific – country.
* Measurable – measuring customer retention rate.
* Achievable - can be tracked.
* Relevant - relevant to finding most profitable countries.
* time-bound - time periods.

1. **Data warehouse**
2. The dataset follows a simplified **Star schema**. Here's how the schema could be structured:

Fact Table: The fact table in this dataset would likely be the "Transaction" table, with columns such as TransactionNo, Date, ProductNo, Price, Quantity, CustomerNo, and Country. Each row in the fact table represents a unique sales transaction, with information about the specific product, customer, and transaction details.

Dimension Tables: The dataset doesn't explicitly mention dimension tables, but based on the provided information, possible dimension tables could include:

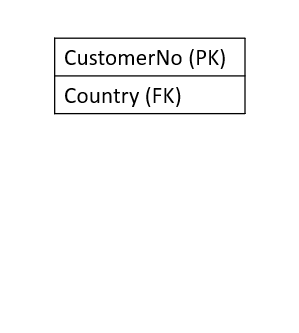
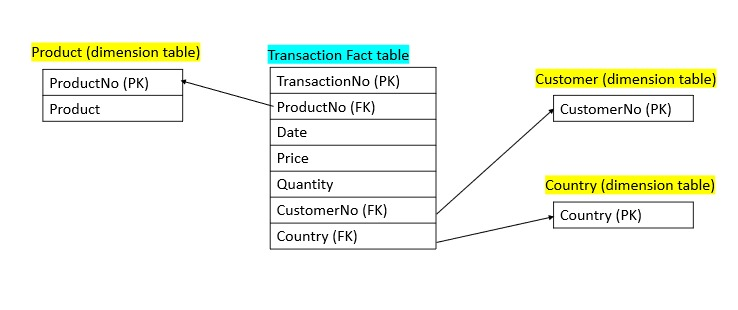
"Product" dimension table: This table could contain information about each product, including the ProductNo and Product name.

"Customer" dimension table: This table could contain information about each customer, including the CustomerNo and possibly additional details such as customer demographics or segmentation such as country.

In this schema, the fact table (Transaction) is at the center, representing the sales transactions, while the dimension tables (Product and Customer) surround it, providing additional descriptive information about the entities involved in the transactions.

*\*\* We’ve fixed/changed the ERD from Matala 2.*

1. ERD:



1. Use cases:

* Use Case 1: Sales Analysis for an E-commerce Business

Scenario: An e-commerce business wants to analyze sales data to gain insights into customer behavior, product performance, and geographical trends.

Efficient Data Analysis: The star schema and ERD design offer efficiency in data analysis for this use case. The fact table "Transaction" contains essential information about each sales transaction, such as the transaction number, date, product details, customer information, and country. This consolidated and denormalized structure enables quick and straightforward queries on sales data without the need for complex joins across multiple tables.

In addition, the simplified structure of the star schema makes it easier for the database query optimizer to generate efficient query plans. With fewer tables and straightforward relationships, the optimizer can better utilize indexing, caching, and other optimization techniques to speed up data retrieval.

* Use case 2: Inventory Analysis for an E-commerce Business

Scenario: An e-commerce business wants to analyze its inventory to optimize stock levels, identify popular products, and improve supply chain management.

Supplier Performance Assessment: By leveraging the "Product" dimension table, the business can evaluate supplier performance. By analyzing sales data associated with specific products, they can identify suppliers whose products are selling well or underperforming. This information can aid in supplier management, contract negotiations, and strategic partnerships to ensure a reliable and efficient supply chain.

Reporting and Visualizations: The star schema's simplicity and the associated ERD provide a user-friendly structure for reporting and visualizations. Users, such as inventory managers or supply chain analysts, can easily navigate the schema, understand the relationships between tables, and generate reports or visualizations to monitor inventory performance, identify trends, and make informed decisions. This accessibility and clarity of the schema empower users to extract insights quickly and take timely actions.

**3. ETL process in our E-commerce business transaction dataset:**

* Fact Table - Transaction:

Extract:

The extraction phase for the Transaction fact table involves retrieving the sales transaction data from the source. This could be a database, CSV files, or any other data storage where the transaction data is stored.

Transform:

In the transformation phase for the Transaction fact table, you may perform various data transformations and manipulations, such as:

Data cleaning: Removing or handling any missing or invalid values in the extracted data.

Aggregations: Aggregating the data at a suitable level, such as summing the quantity or price per transaction, to obtain meaningful metrics for analysis.

|  |  |
| --- | --- |
| **Source** | **Destination** |
| TransactionNo | transaction\_id |
| CustomerNo | customer\_id |
| TotalPrice | total\_price |

Load:

In the loading phase, the transformed data is inserted into the Transaction fact table. The fact table would have columns like TransactionNo, Date, ProductNo, Price, Quantity, CustomerNo, and Country. For each sales transaction, a row is created in the fact table with the relevant data. For example, a row in the Transaction fact table might look like:

TransactionNo: 123456  
ProductNo: P1234  
Price: 10.99  
Quantity: 3  
CustomerNo: C98765

* Dimension Table - Product:

Extract:

In the extraction phase for the Product dimension table, you would obtain the relevant product information from the source. This could include attributes like the ProductNo, Product name, and additional product details. For example, we may extract the following columns from the source: ProductNo, Product.

Transform:

In the transformation phase for the Product dimension table, you may perform various transformations and modifications, such as:

Cleaning and standardizing product names: Ensuring consistent formatting and removing any extraneous characters.

|  |  |
| --- | --- |
| **Source** | **Destination** |
| ProductNo | product\_id |
| ProductName | product\_name |

Load:

In the loading phase, the transformed data is inserted into the Product dimension table. For each unique product, a row is created in the dimension table with the corresponding attributes. For example, a row in the Product dimension table might look like:

ProductNo: P1234  
Product: T-Shirt

* Dimension Table - Customer:

Extract:

In the extraction phase for the Customer dimension table, you would retrieve the relevant customer information from the source. This could include attributes like the CustomerNo, Customer Name, and demographic details. For example, we extract the following column from the source: CustomerNo.

Transform:

In the transformation phase for the Customer dimension table, you may perform various transformations and modifications, such as:

Handling missing or incomplete customer information: Dealing with any missing values or applying data cleansing techniques.

|  |  |
| --- | --- |
| **Source** | **Destination** |
| CustomerNo | customer\_id |
| Country | country |

Load:

In the loading phase, the transformed data is inserted into the Customer dimension table. The dimension table would have columns like CustomerNo, Customer Name..

**In this ETL process, we would like to answer the question: What is the average transaction price for each customer? (not including cancelled transactions).  
In order to answer this question, we will create a new fact table, including the columns below: TransactionNo, TotalPrice, CustomerNo.   
Total price is the outcome of multiplication operation between Price and Quantity columns.**

*\*\* in order to run the code effectively, we minimize the dataset to 20K rows (instead of ~500K).*

1. **Data warehouse analysis**

Query 1:

/\* Selecting the most best-selling product \*/

SELECT TOP 1

ProductName, count (\*) AS total\_transactions

FROM Transaction\_info

GROUP BY ProductName

ORDER BY total\_transactions;

Query 2:

/\* Ranking for each country, product in DESC order \*/

SELECT Country, ProductName,

RANK() OVER(PARTITION BY Country ORDER BY Price DESC) AS RANK

FROM Transaction\_info inf INNER JOIN Transaction\_income inc

ON inf.TransactionNo = inc.TransactionNo

Query 3:

/\* Find the first value of Quantity for each product \*/

SELECT ProductName, Date, Quantity, FIRST\_VALUE(Quantity)

OVER(PARTITION BY ProductName ORDER BY Date) AS FirstQuantity

FROM Transaction\_info inf INNER JOIN Transaction\_income inc

ON inf.TransactionNo = inc.TransactionNo

Query 4:

/\* Find the country with the biggest income from transactions \*/

SELECT Country, total\_income

FROM

(SELECT Country, SUM(Price\*Quantity) AS total\_income,

MAX(COUNT(\*)) OVER () AS max\_ total\_income

FROM Transaction\_info inf INNER JOIN Transaction\_income inc

ON inf.TransactionNo = inc.TransactionNo

GROUP BY Country)

WHERE total\_income = max\_ total\_income

Query 5:

/\* find the customer that spent the most money \*/

SELECT CustomerNo, Country, SUM(Price\*Quantity) AS expenses

FROM Transaction\_info inf INNER JOIN Transaction\_income inc

ON inf.TransactionNo = inc.TransactionNo

GROUP BY CustomerNo, Country

ORDER BY expenses DESC

LIMIT 1

Query 6:

/\* Find the date with the most transactions for each country \*/

SELECT Country, Date, COUNT(TransactionNo) AS total\_transactions,

FIRST\_VALUE(COUNT(TransactionNo))

OVER(PARTITION BY Country ORDER BY Date) AS FirstDate

FROM Transaction\_info inf INNER JOIN Transaction\_income inc

ON inf.TransactionNo = inc.TransactionNo

Query 7:

/\*Find the amount of monthly active users from the last year \*/

WITH ACTIVE\_CTE (CustomerNo, MONTH)

AS

(

SELECT MONTH(Date) AS MONTH, \*

FROM Transaction\_info

WHERE YEAR(Date)=YEAR(DATEADD(Year,-1,GETDATE()))

GROUP BY 1

)

SELECT MONTH,

COUNT(DISTINCT CustomerNo ) AS Customer\_Num

FROM ACTIVE\_CTE

GROUP BY 1

ORDER BY 1

Query 8:

/\*Find the best selling product every month from the last year\*/

SELECT Sales\_month, ProductName, Monthly\_ Quantity

FROM (

SELECT x.\*,

ROW\_NUMBER() OVER(PARTITTION BY Sales\_month ORDER BY Monthly\_ Quantity desc) rn

FROM(

SELECT MONTH(Date) AS Sales\_month,

inf. ProductName,

SUM(Quantity) AS Monthly\_ Quantity

FROM Transaction\_info inf

INNER JOIN Transaction\_income inc

ON inf.TransactionNo = inc.TransactionNo

WHERE YEAR(Date)=YEAR(DATEADD(Year,-1,GETDATE()))

GROUP BY Sales\_month, inf. ProductName

)x

)y

WHERE rn= 1

ORDER BY 1

Query 9:

/\* Find the average number of orders for all years for the customers \*/

WITH Sales\_CTE (CustomerNo, NumberOfOrders)

AS

(

SELECT CustomerNo, COUNT(\*) AS NumberOfOrders

FROM Transaction\_info

WHERE CustomerNo IS NOT NULL

GROUP BY CustomerNo

)

SELECT AVG(NumberOfOrders) AS 'Average orders Per customer'

FROM Sales\_CTE;

Query 10:

/\* Find Monthly revenue per customer \*/

WITH CTE(MONTH,)

AS

(

SELECT MONTH(Date) AS MONTH, \*

FROM Transaction\_info inf

INNER JOIN Transaction\_income inc

ON inf.TransactionNo = inc.TransactionNo

WHERE YEAR(Date)=YEAR(DATEADD(Year,-1,GETDATE()))

GROUP BY MONTH, CustomerNo

)

SELECT MONTH, SUM(Quantity\*Price) as Monthly\_revenue

FROM CTE

GROUP BY 1

1. **Conclusions**

* Best Selling Product: selects the best-selling product by counting the total number of transactions for each product and then ordering them in ascending order. The top result of this query will give you the most popular product based on the number of transactions. This information can be useful for inventory management and marketing strategies.
* Country-wise Ranking of Products: ranks the products within each country based on the price in descending order. This query can provide insights into the popularity of products within specific countries and help identify which products generate higher revenues in different markets.
* First Quantity of Products: retrieves the first recorded quantity for each product, allowing you to analyze the initial sales volume for different products. This information can be helpful in understanding the initial demand and sales patterns for each product.
* Country with the Biggest Income calculates the total income from transactions for each country and identifies the country with the highest income. This information can be useful for understanding the most profitable market and focusing resources on expanding business in that country.
* Customer with the Highest Expenses: identifies the customer who has spent the most money by calculating the total expenses (price multiplied by quantity) for each customer and country combination. This information can be valuable for identifying high-value customers and tailoring marketing efforts or loyalty programs to retain and attract similar customers.

1. **Version management**

GitHub repository link - https://github.com/LiorShapiraa/BIProject