## **Data Analyst Task**

The sales team has the following data from various sources

**Customers.xls - [**[**https://easyupload.io/6zrlyw**](https://easyupload.io/6zrlyw)**]**

**Orders.csv - [**[**https://easyupload.io/9mk8z3**](https://easyupload.io/9mk8z3)**]**

**Shippings.json - [**[**https://easyupload.io/eb3uh6**](https://easyupload.io/eb3uh6)**]**

The team is trying to generate the reports for the below requirements

*the total amount spent and the country for the Pending delivery status for each country.*

*the total number of transactions, total quantity sold, and total amount spent for each customer, along with the product details.*

*the maximum product purchased for each country.*

*the most purchased product based on the age category less than 30 and above 30.*

*the country that had minimum transactions and sales amount.*

**As a Data Analyst, you are required to**

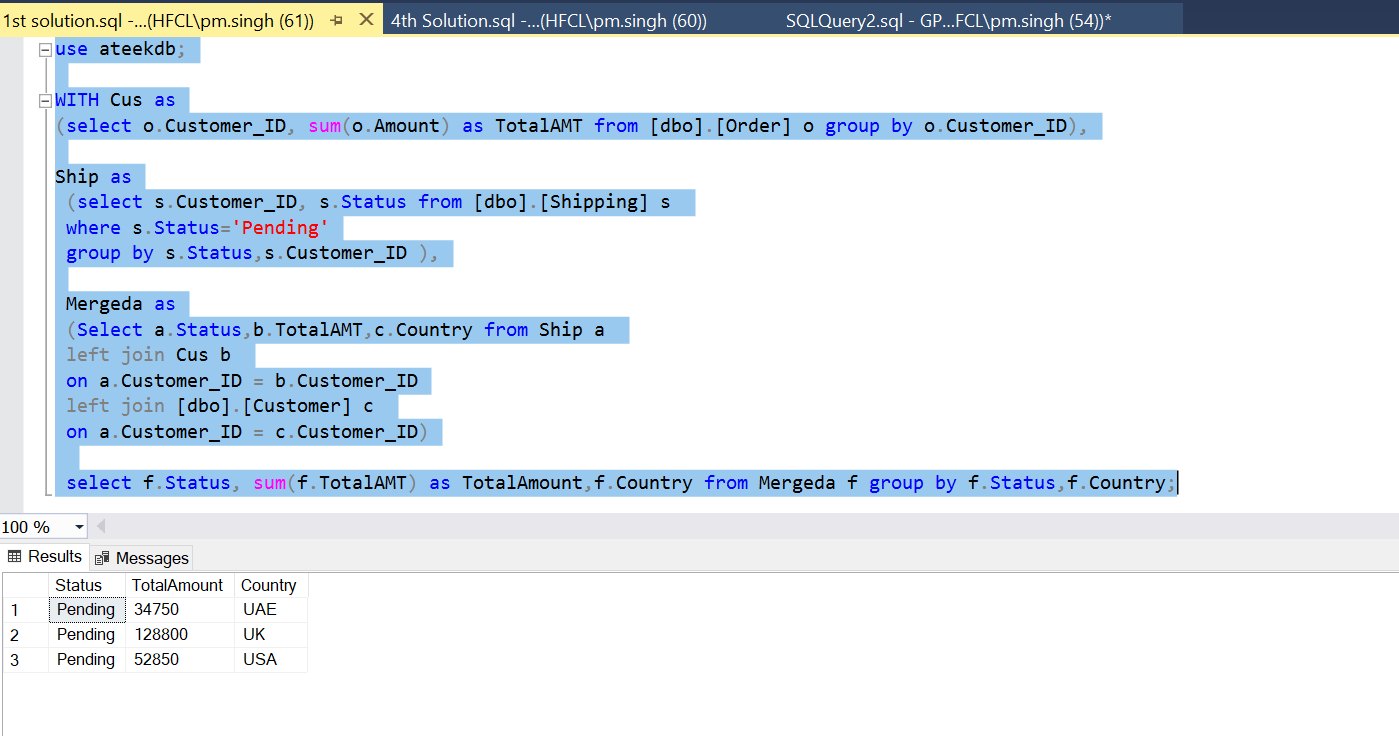
* Verify the accuracy, completeness, and reliability of source data.
* Based on your findings, define and outline the requirements for anticipated datasets, detailing the necessary data components.
* Develop the data models to effectively organise and structure the information and provide a detailed mapping of existing data flows, focussing on the areas of concern.
* Communicate the findings and insights to stakeholders in a visually comprehensive manner.
* What will be your insights to other peer teams of Data Engineers, Data Scientists and other technical and non-technical stakeholders?

**SOLUTION**

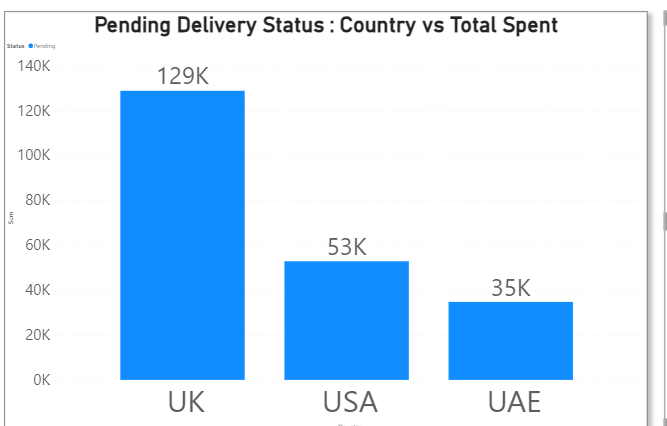
Dear Sir/Ma’am,

Please find the below solutions both in SQL code and Power Bi tool attached with this file :-

*1).The total amount spent and the country for the Pending delivery status for each country.*

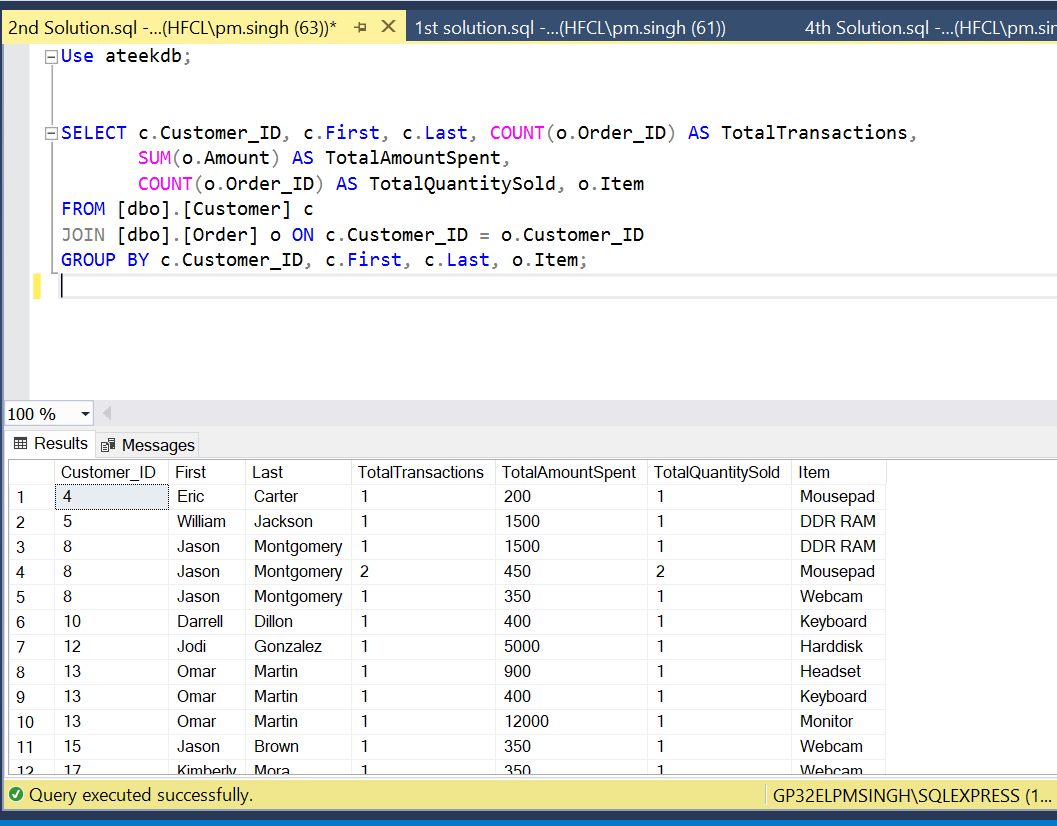
*Ans1)SQLCODE*

*PowerBI*

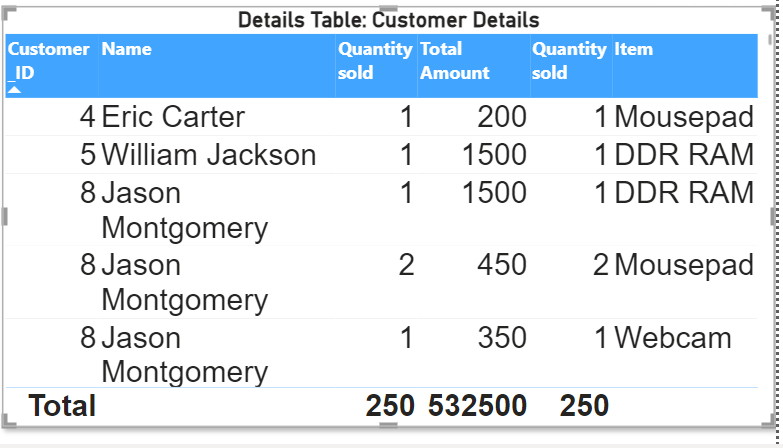
**

*2) The total number of transactions, total quantity sold, and total amount spent for each customer, along with the product details.*

*Ans 2) SQL Code*

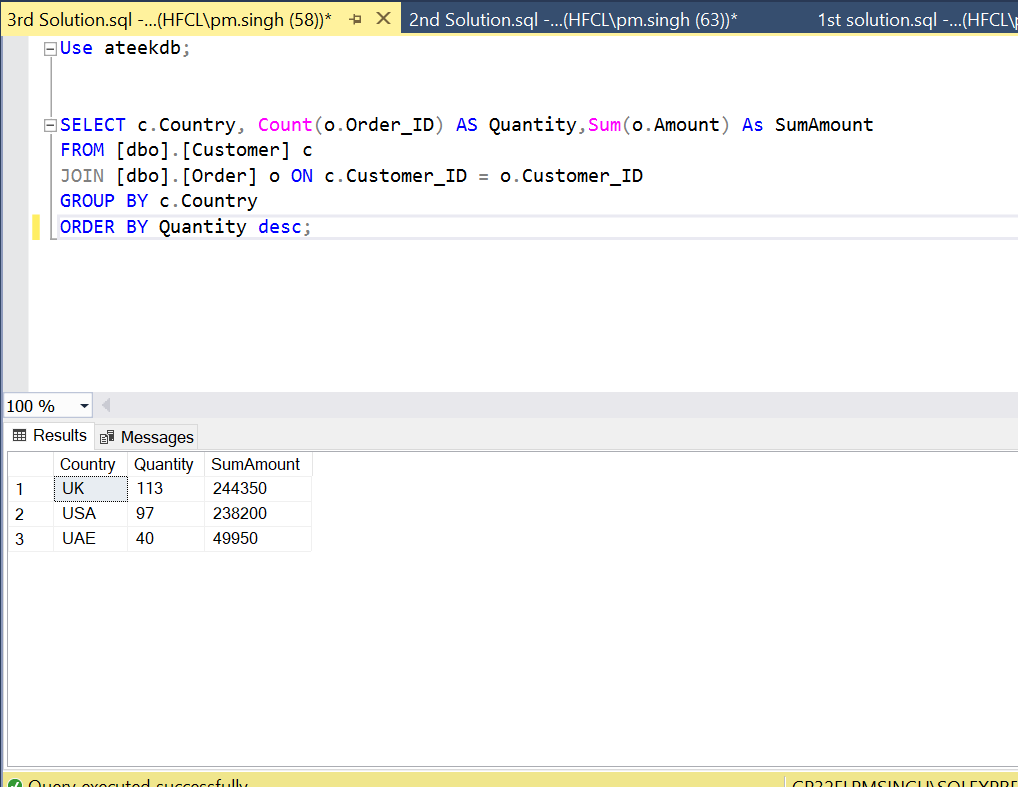
**

*Power Bi*

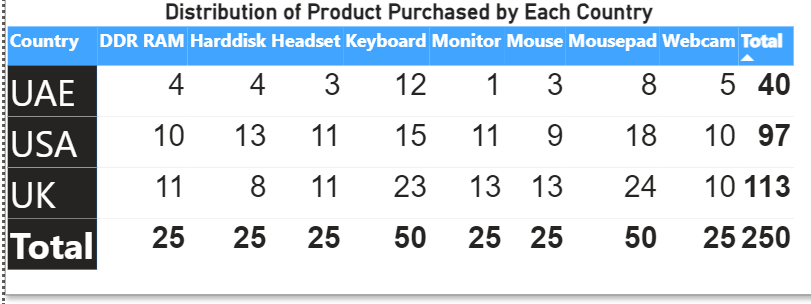
**

*3) The maximum product purchased for each country.*

*Ans 3) SQL Code*

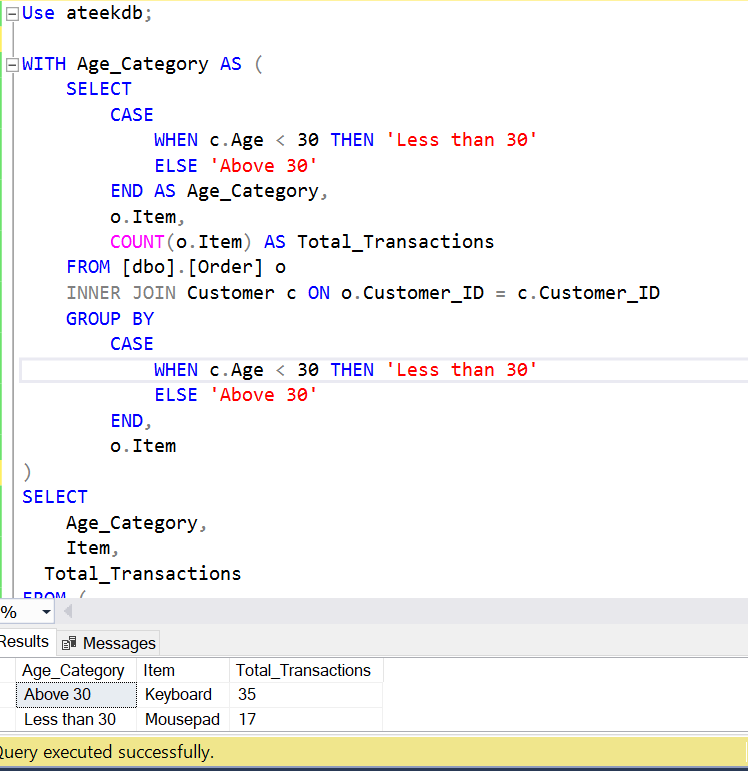
**

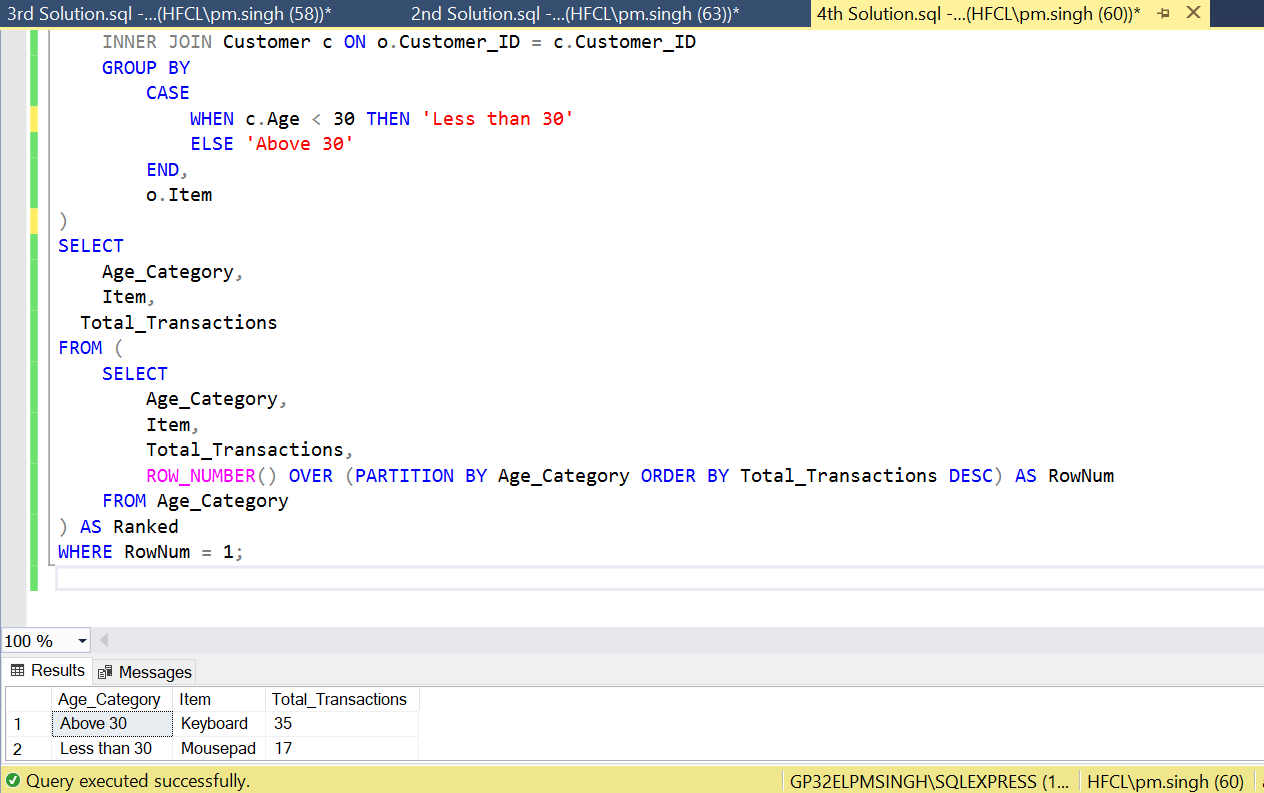
*Power Bi*

**

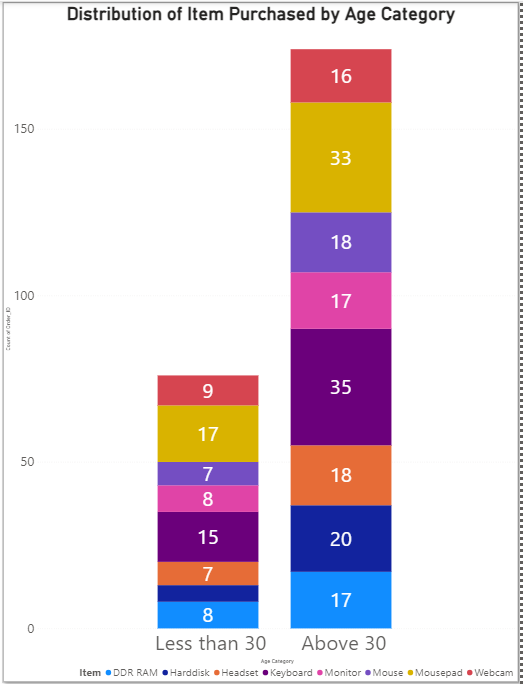
*4) The most purchased product based on the age category less than 30 and above 30.*

*Ans 4) SQL Code*

**

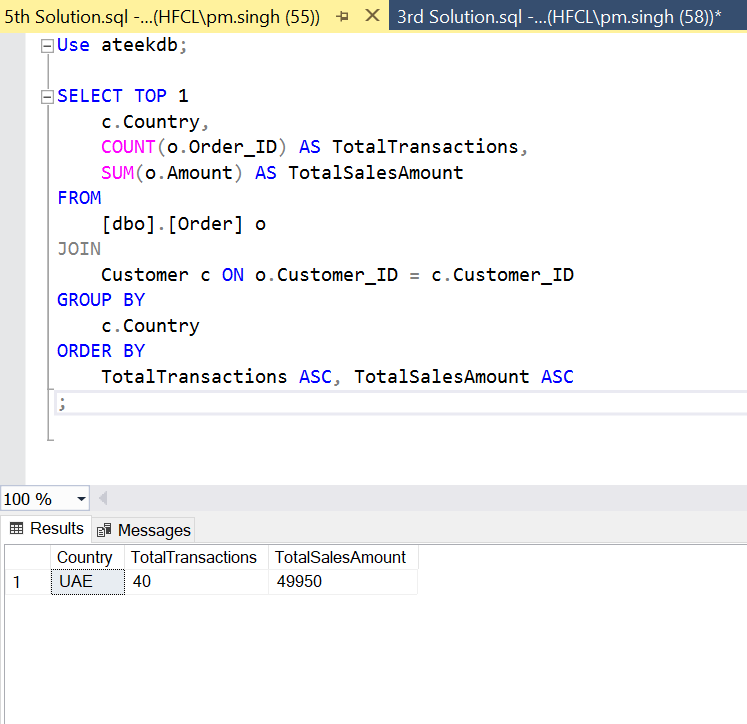
**

*Power BI*

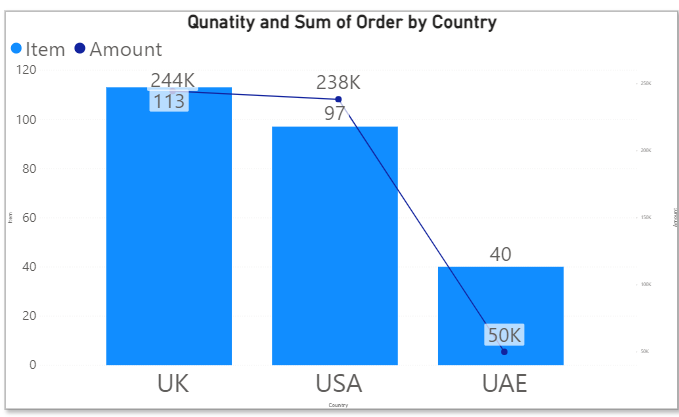
**

*5)The country that had the minimum transactions and sales amount.*

*Ans 5) SQL Code*

**

*PowerBi*

**

**As a Data Analyst, I am required to to begin with, let's assess the accuracy, completeness, and reliability of the provided source data. Here are some initial observations:**

1. Accuracy:
   * Customer data seems accurate in terms of basic information like customer IDs, ages, and countries but few names having spl characters (e.g:- N!cole etc)
   * Order data appears to have accurate information regarding items, amounts, and associated customer IDs.
   * Shipping data seems to reflect the status of orders accurately based on customer IDs.
2. Completeness:
   * The datasets seem complete in terms of basic information about customers, their orders, and shipping status.
   * However, we may need additional information such as timestamps for orders and shipping, as well as more detailed customer profiles like contact information.
3. Reliability:
   * The reliability of the data can't be fully determined without knowing the source and the methods used for data collection as it is just sample data provided for testing.
   * We may need to verify the integrity of the data to ensure it hasn't been tampered with or corrupted.

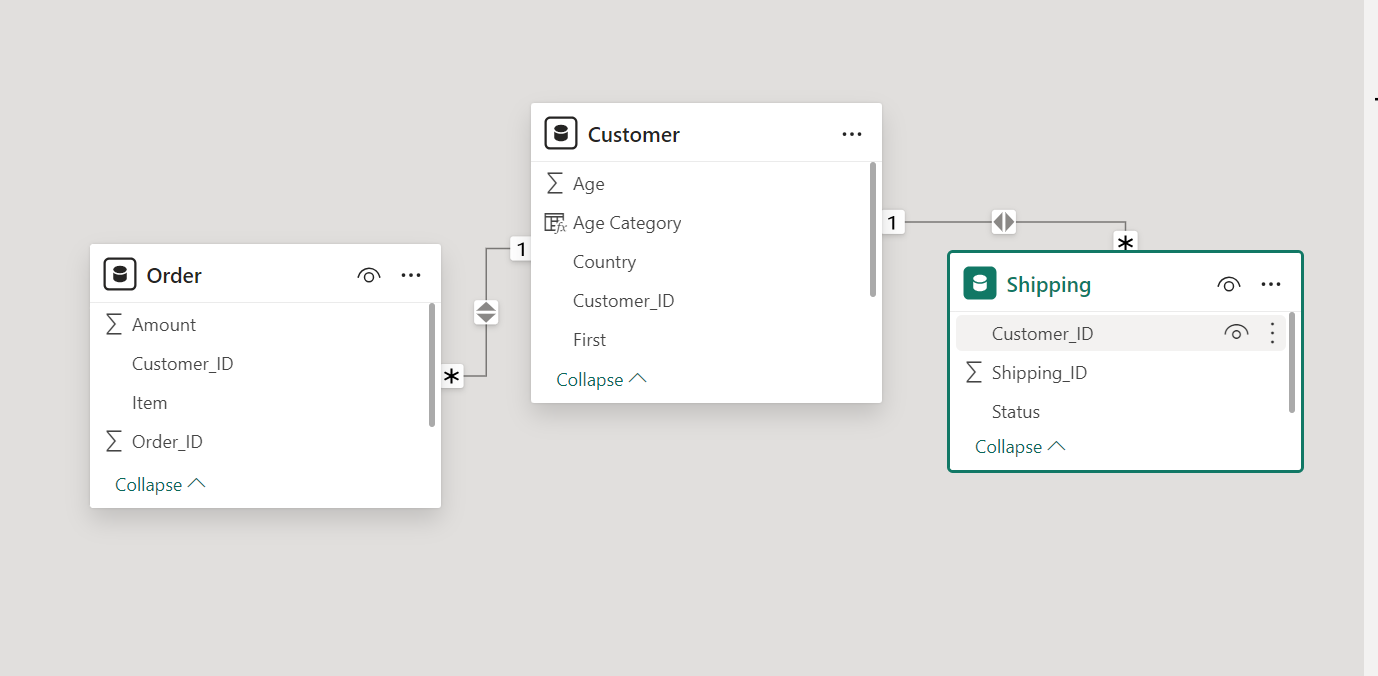
**2) Based on these findings, here are the requirements for anticipated datasets:**

1. Customer Dataset:
   * Basic customer information: Customer ID, First Name, Last Name, Age, Country
   * Additional customer information: Contact details (email, phone), Address
   * Data integrity checks to ensure accuracy and reliability
2. Order Dataset:
   * Order information: Order ID, Item, Amount, Customer ID
   * Timestamps for orders to track order history and patterns
   * Data validation to ensure consistency in item names and amounts
3. Shipping Dataset:
   * Shipping information: Shipping ID, Status, Customer ID
   * Timestamps for shipping to track delivery times and performance
   * Consistency checks to ensure accurate status updates

**3) For data modeling, we organize the information into normalized relational databases:**

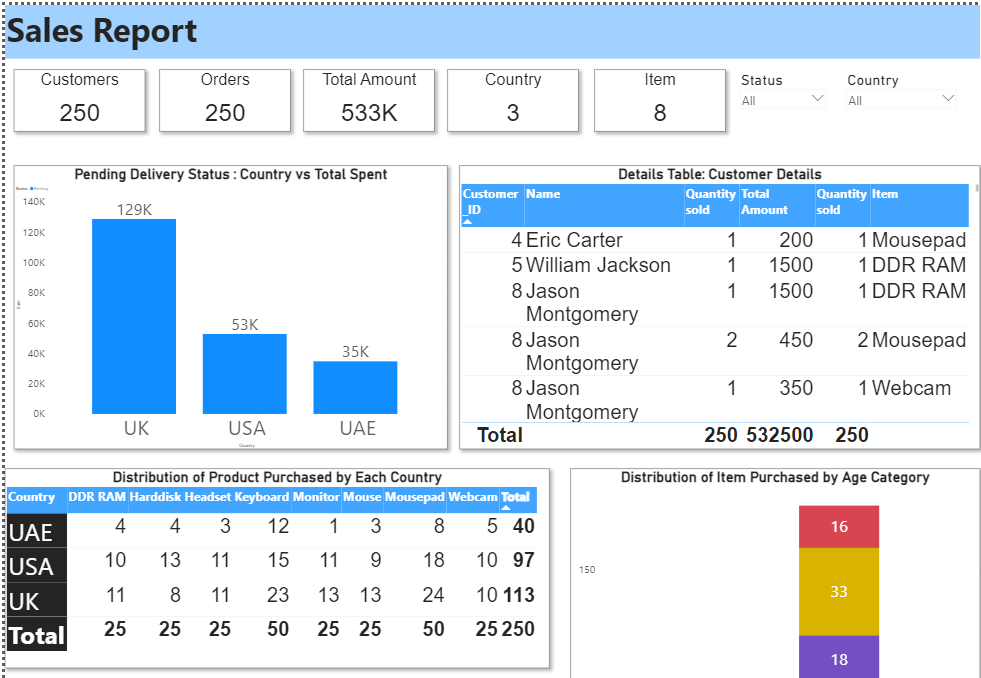
1. Customer Table:
   * Customer\_ID (Primary Key)
   * First
   * Last
   * Age
   * Country
2. Order Table:
   * Order\_ID (Primary Key)
   * Item
   * Amount
   * Customer\_ID (Foreign Key)
3. Shipping Table:
   * Shipping\_ID (Primary Key)
   * Status
   * Customer\_ID (Foreign Key)

**I performed lot of joins to three tables to get the solution as done in SQL codes file and Power BI Relationship from Main Fact Table “ Customer” having connected to “Order” and “Shipping” table having one to many relationship with Customer\_ID as common in all tables and bidirectional flow (both ways) so that all three tables are connected.**

****

**4) For visualization and communication of findings, we can create dashboards and reports highlighting key insights such as:**

Ans 4) Created a “Sales Report” dashboard to represent solutions to different ask by stakeholders through cards, filters, tables, graphs on Power BI report to give insightful information to the stakeholders.



**5) Insights for other peer teams and stakeholders:**

1. Data Engineers:
   * Collaborate to ensure data pipelines are optimized for data ingestion, processing, and storage.
   * Implement data quality checks and monitoring to maintain data integrity.
2. Data Scientists:
   * Provide clean, structured datasets for analysis and modeling.
   * Collaborate on developing predictive models for customer behavior, order forecasting, etc.
3. Technical Stakeholders:
   * Provide regular updates on data quality, integrity, and any anomalies detected.
   * Collaborate on refining data requirements and ensuring data-driven decision-making processes.
4. Non-technical Stakeholders:
   * Present insights and findings in a visually appealing and easy-to-understand format.
   * Provide explanations and context for data-related decisions and recommendations.

By ensuring accurate, complete, and reliable data and effective communication of insights, we can support informed decision-making and drive business success.