



**Aman Mushirul**

Bengaluru

mushirul.aman@gmail.com

+91 9162178254

***[Data Expert / Data Solution Consultant@PWC, Ex-Deloitte, VIT University - B.Tech CSE]***

**Hiring Manager**

PEI GROUP - PEDDA

Dear Hiring Manager,

As a Senior Data Expert, I have been responsible for ensuring that data was accurate, complete, and reliable to build a solid foundation for decision-making processes. Below, I will detail how I executed these responsibilities using PEI assessment example, demonstrating my expertise and approach to managing **data extraction, profiling, cleansing, transformation, modeling, and reporting** it for fact based and data driven decision making for business growth.

### **1. Verified the Accuracy, Completeness, and Reliability of Source Data**

During the assessment with PEI sales data project, I first verified the accuracy, completeness, and reliability of the source data. This involved:

- **Accuracy:** I cross-referenced sales data from "**Customers.xls**," "**Orders.csv**," and "**Shippings.json**" to identify discrepancies or anomalies. For example, I ensured that customer, order details matched the shipping records.
- **Completeness:** I ensured all required fields were populated, checking for missing data points, such as incomplete customer information or orders without shipping records.
- **Reliability:** I assessed the consistency of the data by comparing historical records and business expectations. When I noticed significant deviations in sales data that were not aligned with forecasts, I investigated further to detect potential data entry errors.

Through these efforts, I ensured that the data was prepared for analysis and reported the findings to stakeholders, providing transparency about any issues identified and the solutions

## 2. Outlined Requirements for Anticipated Datasets

Once the data was verified, I outlined the requirements for the anticipated datasets. In this project, I identified the key components necessary for analysis, such as:

- **Required Data Components:** I identified components like customer IDs, product details, sales dates, payment methods, and shipping costs.
- **Data Transformation:** I standardized data from multiple formats by ensuring consistency in fields such as Names and CustomerID. This enabled me to merge the customer, order, and shipping data into a single, coherent dataset.

Close & Apply

New Source

Recent Sources

Enter Data

Data source settings

Manage Parameters

Refresh Preview

Properties

Advanced Editor

Choose Columns

Remove Columns

Keep Rows

Remove Rows

Sort

Split Column

Group By

Use First Row as Headers

Replace Values

Close

New Query

Data Sources

Parameters

Query

Manage Columns

Reduce Rows

Transform

Queries [3]

Customer

Order

Shipping

Table.SelectRows("#Removed Duplicates", each ([First] = "Al!cia" or [First] = "L@rry" or [First] = "N!cole")

123 Customer\_ID

ABC First

ABC Last

123 Age

ABC Country

	Valid	100%	Error	0%	Empty	0%
1	6	N!cole	Jones	33	USA	
2	14	N!cole	Lara	77	UK	
3	109	R0bert	Moore	40	UK	
4	118	R0bert	Shepherd	28	UK	
5	162	N!cole	Bennett	51	USA	
6	171	L@rry	Cole	50	USA	
7	198	R0bert	Bryan	49	UK	
8	211	A!licia	Thompson	38	USA	
9	214	N!cole	Mcintyre	18	UK	
10	236	A!licia	Jensen	19	USA	

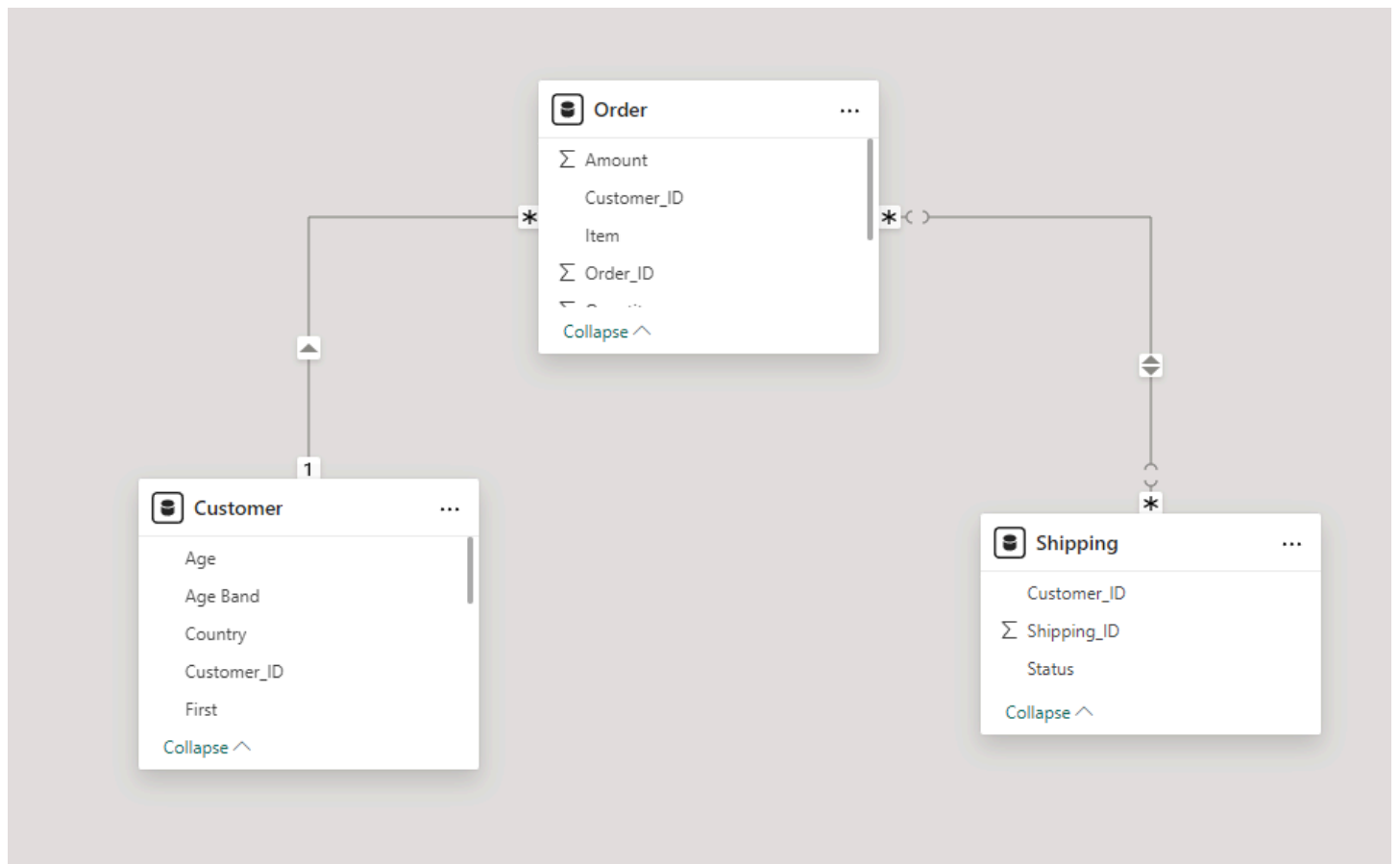
*Expectation of the dataset for improved data driven decision making from data producer / business team :*

<u>Dataset Name</u>	<u>Purpose</u>	<u>Required Data Components</u>	<u>Usage</u>
<b>Customer Dataset</b>	Analyze customer behavior, spending patterns, and demographics	<ul style="list-style-type: none"> <li>- Customer ID</li> <li>- Customer Name</li> <li>- Country</li> <li>- Age</li> <li>- Gender</li> <li>- Total Spending</li> <li>- Transaction Count</li> <li>- Product Preferences</li> </ul>	Provides insights into customer demographics, spending habits, and segmentation analysis.
<b>Orders Dataset</b>	Monitor sales and order trends, focusing on delivery performance	<ul style="list-style-type: none"> <li>- Order ID</li> <li>- Customer ID</li> <li>- Order Date</li> <li>- Product ID</li> <li>- Quantity</li> <li>- Order Status</li> <li>- Country</li> <li>- Total Amount</li> <li>- Delivery Date</li> </ul>	Track sales trends, highlight pending delivery issues, and optimize order processing.
<b>Product Dataset</b>	Analyze product popularity, sales volume, and demand	<ul style="list-style-type: none"> <li>- Product ID</li> <li>- Product Name</li> <li>- Category</li> <li>- Price</li> <li>- Quantity Sold</li> <li>- Stock Availability</li> </ul>	Helps in inventory management, product demand analysis, and marketing strategies.
<b>Shipping Dataset</b>	Manage and analyze shipping details and performance	<ul style="list-style-type: none"> <li>- Order ID</li> <li>- Shipping ID</li> <li>- Shipping Date</li> <li>- Delivery Date</li> <li>- Shipping Status</li> <li>- Shipping Cost</li> <li>- Delivery Country</li> </ul>	Tracks shipping performance, identifies delays, and optimizes logistics operations.
<b>Transaction Dataset</b>	Capture financial transactions for	<ul style="list-style-type: none"> <li>- Transaction ID</li> <li>- Order ID</li> </ul>	Enables tracking of revenue, payments,

	reporting and analysis	<ul style="list-style-type: none"> <li>- Customer ID</li> <li>- Transaction Date</li> <li>- Payment Method</li> <li>- Transaction Amount</li> <li>- Currency</li> </ul>	transaction volumes, and financial forecasting.
<b>Country Dataset</b>	Analyze country-based metrics, such as transaction volume and sales	<ul style="list-style-type: none"> <li>- Country ID</li> <li>- Country Name</li> <li>- Total Transactions</li> <li>- Total Sales</li> <li>- Average Transaction Value</li> <li>- Pending Orders</li> <li>- Most Purchased Product</li> </ul>	Allows comparison of performance across regions, aiding in market expansion and logistics.
<b>Demographics Dataset</b>	Correlate demographic factors with purchasing behavior	<ul style="list-style-type: none"> <li>- Age Group</li> <li>- Gender</li> <li>- Total Purchases</li> <li>- Product Preferences</li> </ul>	Supports analysis of product preferences by demographic, facilitating targeted marketing.
<b>Sales Performance Dataset</b>	Provide a holistic view of sales performance	<ul style="list-style-type: none"> <li>- Sales Region</li> <li>- Product Category</li> <li>- Total Revenue</li> <li>- Sales Growth</li> <li>- Top Selling Products</li> </ul>	Offers a broad perspective on sales performance, aiding in trend identification and growth strategies.

I presented this outline to stakeholders, clearly detailing how the data would be used in further analysis, making sure both technical and non-technical team members understood the requirements.

### 3. Developed Data Models



I then developed the data model to structure the data in an optimized format. This included:

- **Entity-Relationship Mapping:** I built relationships between customers, orders, and shipping data, ensuring proper indexing and minimizing redundancy.
- **Data Transformation:** I created new fields, such as “**Average Spending per Transaction**” or the below measures to enhance the data's analytical capacity.
- **Normalization:** I normalized the data to reduce redundancy and improve consistency, ensuring that future queries would run efficiently.

### Developed DAX and Measures Example:

1. Average Spending per Transaction =

**DIVIDE**(

**SUM**('Consolidated Sales'[Order.Amount]),

**COUNT**('Consolidated Sales'[Order.Order\_ID])

)

2. Distinct Products = **DISTINCTCOUNT**('Consolidated Sales'[Order.Item])

3. Max Product Purchased by Country =

```
CALCULATE(  
  
    MAXX(  
  
        VALUES('Consolidated Sales'[Order.Item]),  
  
        CALCULATE(SUM('Consolidated Sales'[Order.Quantity]))  
  
    ),  
  
    ALLEXCEPT('Consolidated Sales', 'Consolidated Sales'[Country])  
  
)
```

4. Most Purchased Product Age < 30 =

```
CALCULATE(  
  
    MAXX(  
  
        VALUES('Consolidated Sales'[Order.Item]),  
  
        CALCULATE(SUM('Consolidated Sales'[Order.Quantity]),  
  
            'Consolidated Sales'[Age] < 30  
  
        )  
  
    )  
  
)
```

5. Most Purchased Product Age ≥ 30 =

```
CALCULATE(  
  
    MAXX(  
  

```

**VALUES**('Consolidated Sales'[Order.Item]),

**CALCULATE**(**SUM**('Consolidated Sales'[Order.Quantity]),

'Consolidated Sales'[Age] >= 30

)

)

)

6.Total Amount Spent for Pending Deliveries =

**CALCULATE**(

**SUM**('Consolidated Sales'[Order.Amount]),

'Consolidated Sales'[Shipping.Status] = "Pending"

)

7.Total Amount Spent per Customer =

**SUM**('Consolidated Sales'[Order.Amount])

8.Total Orders = **DISTINCTCOUNT**('Consolidated Sales'[Order.Order\_ID])

9.Total Quantity Sold per Customer =

**SUM**('Consolidated Sales'[Order.Quantity])

10.Total Transactions = **DISTINCTCOUNT**('Consolidated Sales'[Order.Order\_ID])

11.Total Transactions per Customer = **COUNTROWS**( 'Consolidated Sales')

## Data



Search

### Customer

- ☐ Age
- ☐ Age Band
- ☒ Country
- ☐ Customer\_ID
- ☐ First
- ☐ Last
- ☐ Max Product Purchased / Country
- ☐ Most Purchased Product / Age < 30
- ☐ Most Purchased Product / Age > 30
- ☐ Total Amount Spent / Customer
- ☐ Total Amount Spent for/ Pending D...
- ☐ Total Pending Orders by Country
- ☐ Total Quantity Sold / Customer
- ☐ Total Transactions / Customer
- ☐ Total Transactions Orders

### Order

- ☐  $\Sigma$  Amount
- ☐ Average Spending / Transaction
- ☐ Count / customer
- ☐ Customer\_ID
- ☐ Item
- ☐  $\Sigma$  Order\_ID
- ☐  $\Sigma$  Quantity
- ☐ Unique Products

### Shipping

- ☐ Customer\_ID
- ☐  $\Sigma$  Shipping\_ID
- ☐ Status



Customer_ID	First	Last	Age	Country	Age Band
1	Joseph	Rice	43	USA	30 and Above
2	Gary	Moore	71	USA	30 and Above
6	Nicole	Jones	33	USA	30 and Above
7	David	Davis	59	USA	30 and Above
12	Jodi	Gonzalez	69	USA	30 and Above
16	David	Benson	61	USA	30 and Above
17	Kimberly	Mora	34	USA	30 and Above
21	Diane	Henson	74	USA	30 and Above
22	Cody	Lyons	47	USA	30 and Above
26	Tina	Moore	34	USA	30 and Above
27	Kylie	White	63	USA	30 and Above
31	Craig	Myers	42	USA	30 and Above
36	Melissa	Smith	75	USA	30 and Above
42	Matthew	Velez	71	USA	30 and Above
46	Anthony	Chavez	36	USA	30 and Above
47	Carolyn	Green	31	USA	30 and Above
51	John	Huber	51	USA	30 and Above
52	Haley	Martinez	75	USA	30 and Above
56	Jose	Poole	60	USA	30 and Above
61	Jade	Wall	51	USA	30 and Above
66	Hannah	Wu	59	USA	30 and Above
67	Guy	Bennett	44	USA	30 and Above
71	Michelle	Edwards	77	USA	30 and Above
72	Alexander	Griffin	76	USA	30 and Above
76	Ricky	Phillips	49	USA	30 and Above
81	Allison	Sweeney	75	USA	30 and Above
82	Joseph	Hernandez	41	USA	30 and Above
86	Emily	Thomas	74	USA	30 and Above
91	Stephanie	Hicks	53	USA	30 and Above
92	Paul	Brown	78	USA	30 and Above
96	Eric	Harvey	52	USA	30 and Above
97	Debbie	Rogers	76	USA	30 and Above
101	Brian	Olson	38	USA	30 and Above
111	Audrey	Richardson	53	USA	30 and Above
112	Maureen	Bryant	64	USA	30 and Above
116	Rhonda	Flores	52	USA	30 and Above
117	Steven	Black	41	USA	30 and Above
121	Pamela	Wall	42	USA	30 and Above

Table: Customer (250 rows) Column: Total Pending Orders by Country (0 distinct values)

Data

Search

- Customer
  - Age
  - Age Band
  - Country
  - Customer\_ID
  - First
  - Last
  - Max Product Purchased / Coun...
  - Most Purchased Product / Age...
  - Most Purchased Product / Age...
  - Total Amount Spent / Customer
  - Total Amount Spent for/ Pendi...
  - Total Pending Orders by Country
  - Total Quantity Sold / Customer
  - Total Transactions / Customer
  - Total Transactions Orders
- Order
  - Amount
  - Average Spending / Transaction
  - Count / customer
  - Customer\_ID
  - Item
  - Order\_ID
  - Quantity
  - Unique Products
- Shipping
  - Customer\_ID
  - Shipping\_ID
  - Status

## 1. Customer Table

Attribute	Data Type	Constraints	Description
Customer_ID	Integer	Primary Key, Not Null	Unique identifier for each customer
Age	Integer	Not Null	Customer's age
Age Band	Text	Not Null	Grouping of customers based on age (e.g., <30, 30+)
Country	Text	Not Null	Customer's country of residence
First	Text	Optional	First name of the customer
Last	Text	Optional	Last name of the customer

**Notes:**

- Customer\_ID uniquely identifies a customer.
- Country can be used for geographic analysis and segmentation.
- Age Band helps in analyzing customer purchases by age group.

**2. Order Table**

<u>Attribute</u>	<u>Data Type</u>	<u>Constraints</u>	<u>Description</u>
Order_ID	Integer	Primary Key, Not Null	Unique identifier for each order
Customer_ID	Integer	Foreign Key (Customer.Customer_ID), Not Null	Links the order to the corresponding customer
Item	Text	Not Null	Product or service ordered
Amount	Decimal	Not Null	Total amount for the order
Order_Date	Date	Not Null	Date when the order was placed

**Notes:**

- **Customer\_ID** as a Foreign Key creates a relationship with the Customer table.
- Order\_ID is a unique identifier for each transaction.
- The Amount represents the total amount spent for each order, which will help in KPIs like "Average Spending per Transaction."
- Order\_Date helps in time-based analysis.

**3. Shipping Table**

<u>Attribute</u>	<u>Data Type</u>	<u>Constraints</u>	<u>Description</u>
Shipping_ID	Integer	Primary Key, Not Null	Unique identifier for each shipping record
Customer_ID	Integer	Foreign Key (Customer.Customer_ID), Not Null	Links the shipping record to the corresponding customer
Status	Text	Not Null	Shipping status (e.g., "Pending", "Completed")
Shipping_Date	Date	Optional	Date when the shipment was made (if applicable)

#### Notes:

- **Customer\_ID** links each shipping record to a customer.
- The Status field is critical for analyzing pending or completed shipments.
- Shipping\_Date can help track shipment timelines.

#### Relationships:

- Customer ↔ Order (1:n)**
  - One customer can place multiple orders.
  - Customer\_ID acts as a foreign key in the Order table, ensuring that each order is linked to a specific customer.
- Customer ↔ Shipping (1:1):**
  - Each customer can have only one corresponding shipping record at a time.
  - Customer\_ID in Shipping is a foreign key that references Customer.Customer\_ID.
- Order ↔ Shipping (1:n)**
  - A single shipping record can link to multiple orders.
  - Shipping\_ID in the Order table establishes this relationship.

#### Data Model Concerns & Solutions:

- **Data Consistency:** Ensure that all orders have corresponding customer and shipping data, and there are no orphan records.
  - **Solution:** Data integrity constraints and Foreign Key relationships must be enforced in the database.
- **Data Completeness:** Ensure no null values in mandatory fields like Order\_Amount, Country, and Status.
  - **Solution:** Implement data validation checks before inserting records into tables.
- **Pending Deliveries:** Analysis of all pending orders and their geographical distribution is important.
  - **Solution:** Use the Status field in the Shipping table to filter pending orders and link them back to customers and countries.

## ISSUES IN THE DATAMODEL :

### STEP 1 : Profiling

There are Customer\_ID from CUSTOMERS which are not present in ORDER and SHIPPING tables resulting in getting NULL records while creating a Golden Consolidation Sheet.

One or more of the datasets could be incomplete, meaning some customers may not have corresponding orders or shipments, resulting in null values during the join.

If a **Customer\_ID** exists in the **Customers** table but not in the **Orders** table, it means that customer hasn't placed any orders. This is expected behavior, as some customers may not have ordered yet.

If a **Customer\_ID** exists in the **Orders** table but not in the **Shipping** table, it could indicate an issue in the shipping process or that the order hasn't been processed yet.

### STEP 2 : Data Cleansing ( Getting the data ready for Reporting )

**Exclude from Analysis:** If these missing customers are minimal or irrelevant, we can filter out these rows in our BI dashboards.

**Investigate Source:** It's important to investigate why these customers are missing in the first place. Ensure the ETL process correctly maps customers to orders.

**Solution for correct reporting** : Do inner join on the 3 tables and only take relevant data with not NULL values ( Creation of Consolidates Sales table which is the source of Truth for Dashboard visualization and reporting.

#### **4. Mapped Data Flows and Highlighted Areas of Concern**

I mapped the data flows, visualizing how the customer, order, and shipping data traveled from their respective sources into the final analysis-ready dataset. This included:

- **Source Integration:** I outlined how the data from multiple sources was combined and transformed to align.
- **Data Transformation Stages:** I visualized how the data was cleaned and standardized, ensuring that stakeholders understood each step.
- **Highlighted Areas of Concern:** I flagged any issues, such as discrepancies between order dates and shipping times, and document these concerns to present to Data Engineering team and Leadership.

This mapping helped ensure that potential bottlenecks were addressed before the final analysis phase.

#### **5. Prepared Technical Specifications for Data Engineers**

##### **Part: Shipping & Pending Orders Flow**

- **Objective:** The goal is to track pending orders efficiently, ensuring that customers whose orders have not yet been shipped are flagged and analyzed geographically.
- **Data Flow:**
  - Use the Shipping.Status column to identify pending deliveries.
  - Join the Shipping table with the Order and Customer tables via Customer\_ID to determine the total number of pending orders for each country.
  - Implement logic to compute pending delivery percentages by region.

##### **SQL for Pending Deliveries:**

```
SELECT c.Country, COUNT(o.Order_ID) AS Total_Pending_Orders  
  
FROM Order o  
  
JOIN Shipping s ON o.Customer_ID = s.Customer_ID
```

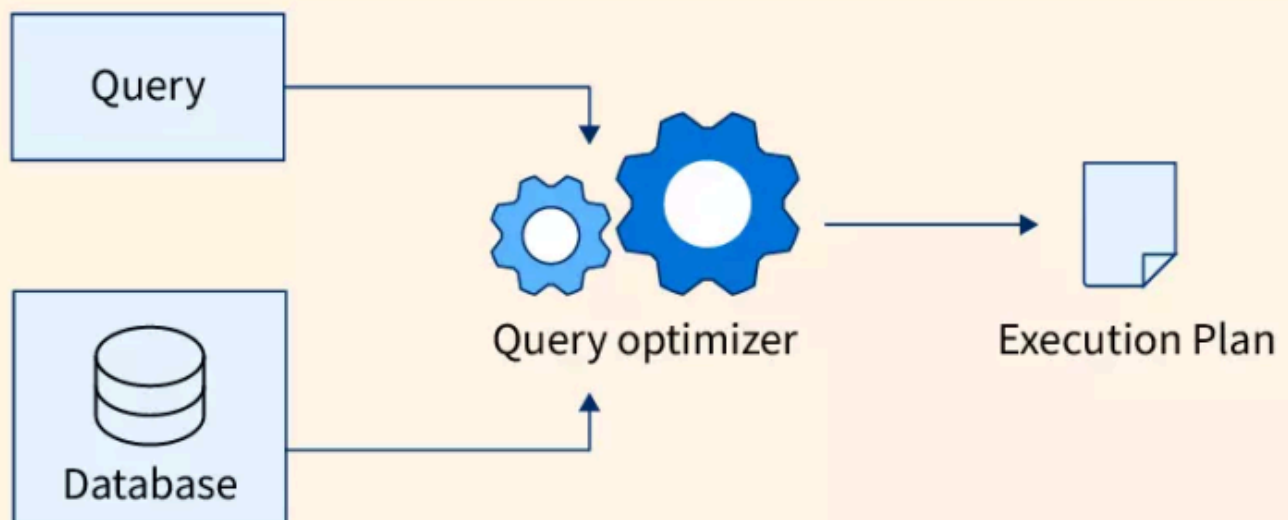
```
JOIN Customer c ON o.Customer_ID = c.Customer_ID
```

```
WHERE s.Status = 'Pending'
```

```
GROUP BY c.Country;
```

For this project, the data engineering team needs to take a look on the below factors. This includes:

- **Indexing for Query Optimization:** I specified which fields needed indexing, such as customer IDs and order dates, to ensure faster querying.
- **Handling Large Datasets:** I outlined strategies for partitioning large datasets, enhancing storage and retrieval speed.
- **Data Validation Rules:** I established validation rules for incoming data, ensuring consistency in future dataset updates.

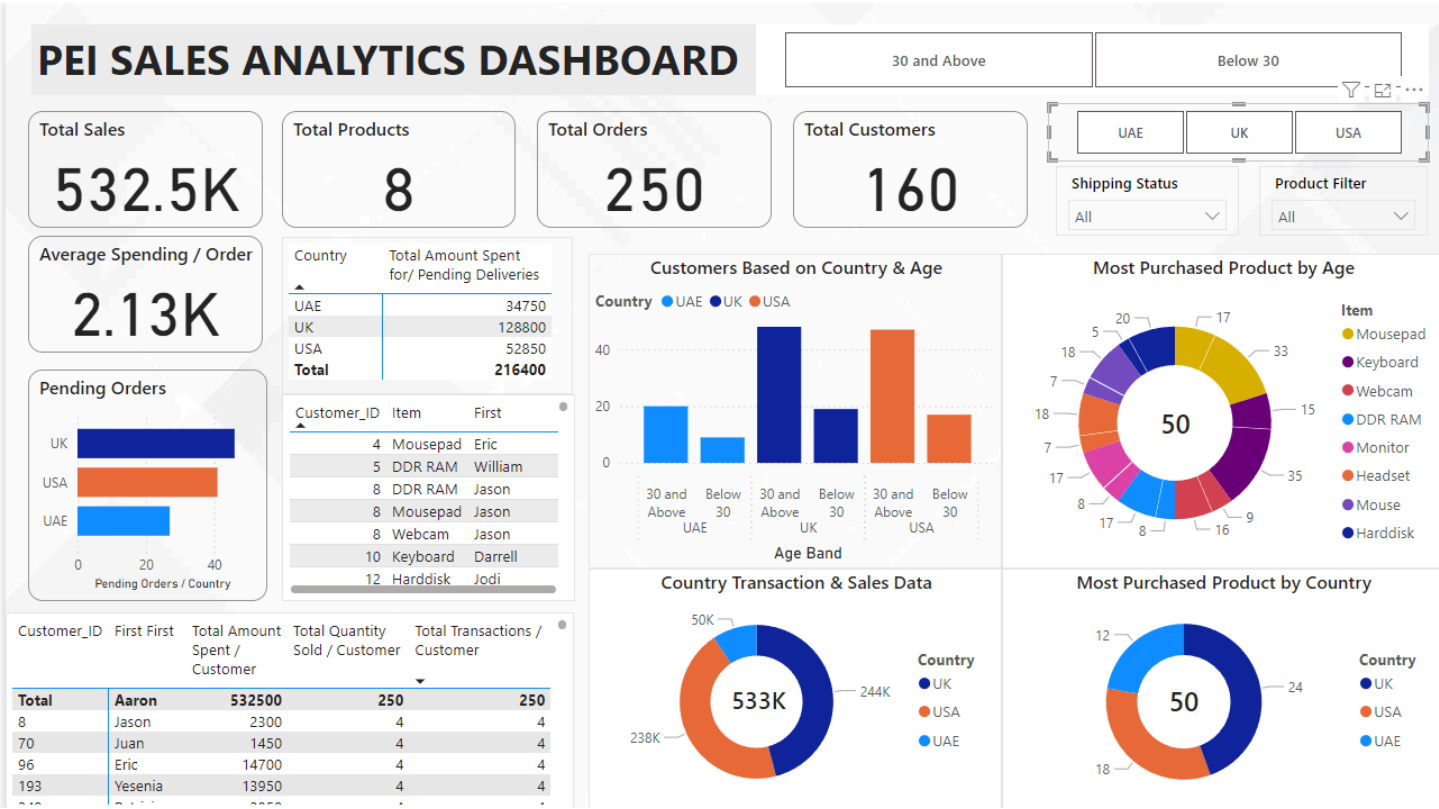


These specifications were documented clearly, enabling the data engineering team to implement the necessary solutions effectively.

## 6. Communicated Findings and Insights to Stakeholders

These reports included:

- KPIs: I displayed key metrics, such as "Total Sales," "Average Order Value," and "Revenue by Region."
- DAX Measures: I used DAX to develop dynamic insights, such as monthly sales trends and customer segmentation.
- Visual Storytelling: I presented insights in a clear and compelling format, ensuring that stakeholders, including executives, could easily grasp the analysis and make informed decisions.



<u>Insight Category.</u>	<u>Description</u>
<b>1. Pending Delivery Status by Country</b>	
<b>Highest Pending Deliveries</b>	Country <b>UK</b> has the highest number of pending deliveries with <b>46 orders</b> , with a total amount spent of <b>\$128800</b> . This indicates potential delays in the supply chain for this region.
<b>Lowest Pending Deliveries</b>	Country <b>UAE</b> has the fewest pending deliveries with <b>27 orders</b> , with a total amount spent of <b>\$34750</b> . This suggests more efficient logistics or lower demand in this area.
<b>2. Customer Spending Patterns</b>	
<b>Top Spending Customer</b>	The top customer, <b>Customer ID 166 Morgan</b> , has spent <b>\$17350</b> across <b>3 transactions, purchasing 3 units</b> of Hard Disk,Monitor,Webcam.This highlights the most valuable customers who could be targeted for loyalty programs.
<b>Most Transactions by a Customer</b>	<b>Below customers</b> completed the <b>most transactions</b> , showing a high frequency of purchases. ( Jason , Juan , Eric , Yesnia)
<b>3. Country-Level Product Demand</b>	
<b>Maximum Product Purchased by Country</b>	<b>UK</b> is the county with most product purchased .Product <b>Keyboard &amp; Mousepad</b> is the most purchased in <b>Country UK</b> , accounting for <b>25% of total sales</b> . This suggests a high demand for this product, which could be further promoted through localized marketing strategies.
<b>4. Age-Based Product Preference</b>	




### Under 30 Age Group

Customers **under the age of 30** primarily purchased Product **Mousepad**, accounting for **22.37 % of their total purchases**. This product could be marketed more heavily towards younger customers.

### Conclusion

In this PEI sales data project, I not only ensured that the data was accurate and well-structured but also developed data models that optimized performance and delivered valuable insights to stakeholders. I was able to deliver a solution that was both technically sound and aligned with business objectives.

As a Data Analyst, my expertise in verifying data quality, structuring data models, and communicating findings allowed me to drive meaningful outcomes for the organization. I ensured that the data solutions were scalable, reliable, and tailored to the company's strategic needs.

**PLEASE CHECK OUT BI  
TECHNICAL SPECIFICATION  
DOCUMENT FOR DETAILS AND  
SCREENSHOTS **