## **Problem Statement:**

## **Data Analyst Task**

The sales team has the following data from various sources

**Customers.xls - [**[**https://easyupload.io/nsu7nm**](https://easyupload.io/nsu7nm)**]**

**Orders.csv - [**[**https://easyupload.io/tqcu2a**](https://easyupload.io/tqcu2a)**]**

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

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

1. *the total amount spent and the country for the Pending delivery status for each country.*
2. *the total number of transactions, total quantity sold, and total amount spent for each customer, along with the product details.*
3. *the maximum product purchased for each country.*
4. *the most purchased product based on the age category less than 30 and above 30.*
5. *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.
* Prepare a story with technical specifications for one part of the data model for a data engineer.
* Communicate the findings and insights to stakeholders in a visually comprehensive manner.

## **Solution:**

**Introduction and details:**

* There are 3 data sources provided whose formats are different
* I’ll be using Power BI as the main data visualization and transformation tool
* Along with that assuming if we can load the data into SQL database using any of the ETL tool, I’ll be creating data model according to it. I’ll be using “Oracle SQL developer data modeler” for the creation of Logical, Relational and Physical models

**Data Modelling:**

To design a data model we start with Conceptual modelling, then Logical modelling later Relational modelling and Physical modelling

**Conceptual Modelling:**

**Objective**: To define the high-level entities without worrying about the actual database structure , we have the data sources and according to our business needs those are our Entities.

**Entities**:

* Customer: Represents the customers.
* Order: Represents the orders placed by customers.
* Shipment: Represents the shipment details of orders and their status.

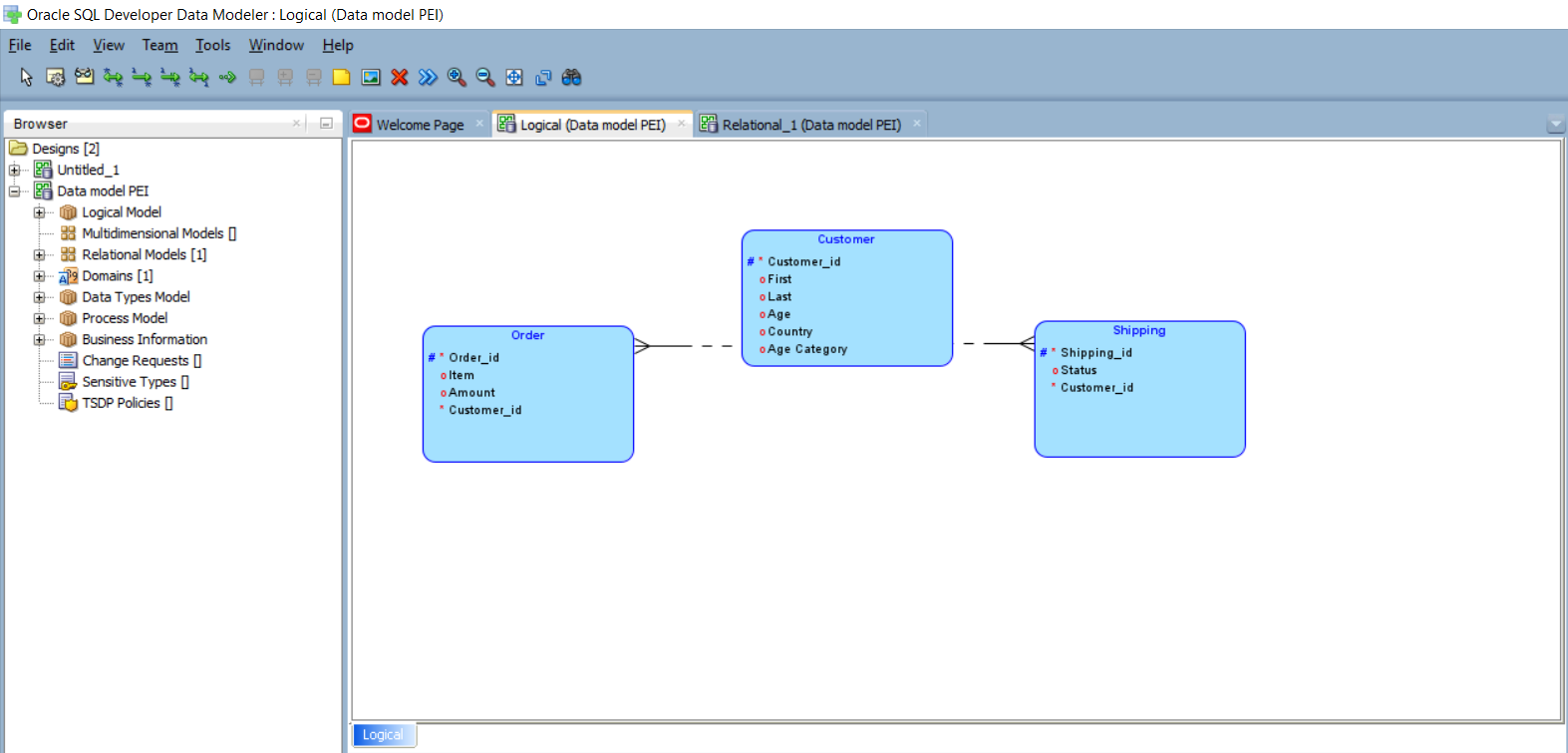
**Logical Modelling:**

**Objective**: To define the structure of the data elements and set the relationships between entities, focusing on the logical structure rather than the physical implementation.

**Attributes**:

* + **Customer**:
    - Customer\_id (Primary Key)
    - Name
    - Age
    - Country
  + **Order**:
    - Order\_id (Primary Key)
    - Item
    - Amount
    - Customer\_id (Foreign Key)
  + **Shipment**:
    - Shipping\_id (Primary Key)
    - Status
    - Order\_id (Foreign Key)

Using Oracle SQL developer data modeler tool, the below Logical model is built along with relationships using Entity Relationship diagram(ERD)



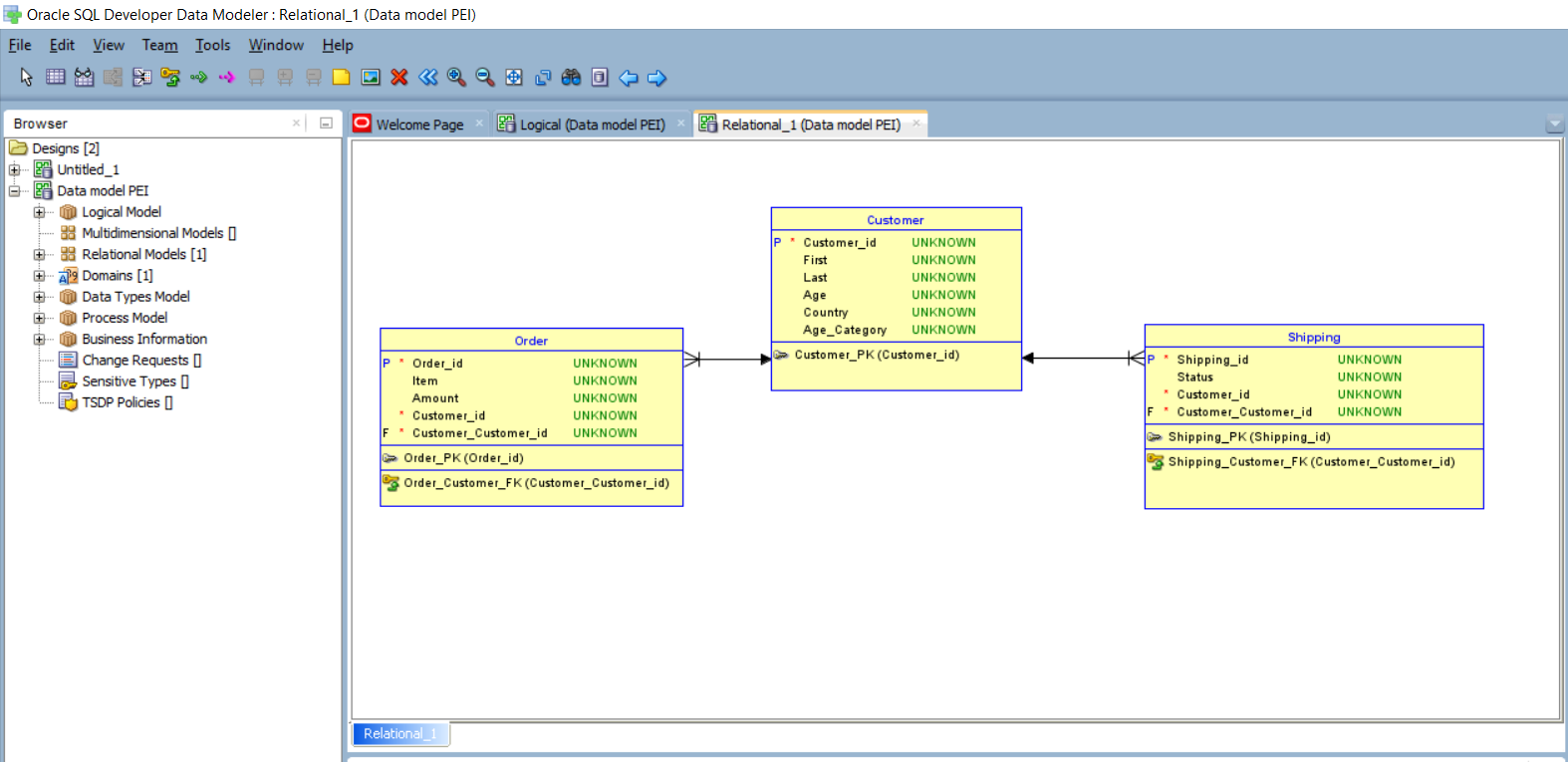
**Relational Modelling:**

**Objective**: To translate the logical model into a relational schema, defining tables, columns, and relationships in a relational database along with data types, contraints with primay key and foreign key

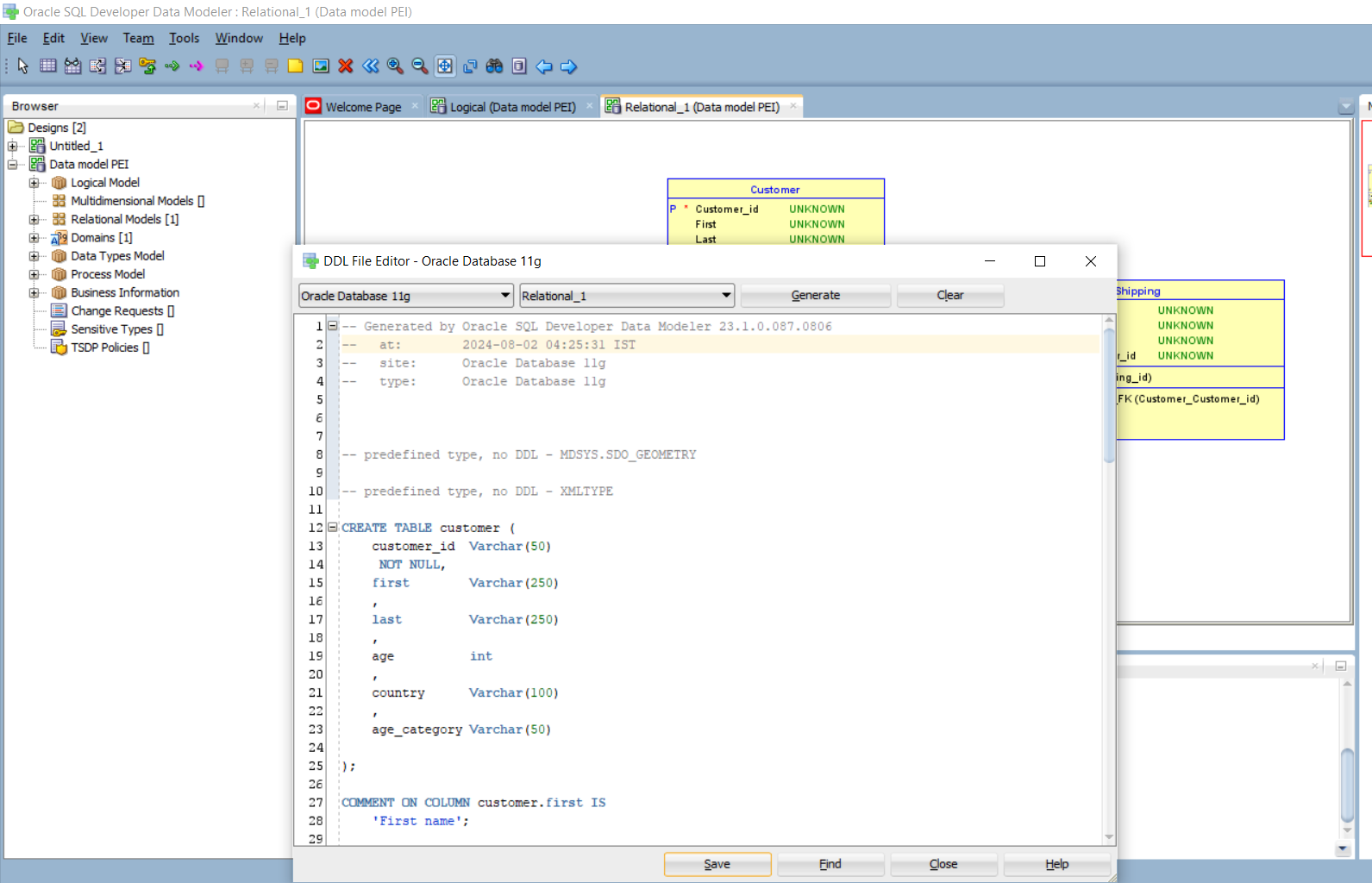
**Entities and Relationships:**

* + **Customer**:
    - Customer\_id (Primary Key)
    - Name
    - Age
    - Country
  + **Order**:
    - Order\_id (Primary Key)
    - Item
    - Amount
    - Customer\_id (Foreign Key)
  + **Shipment**:
    - Shipping\_id (Primary Key)
    - Status
    - Order\_id (Foreign Key)

Using Oracle SQL developer data modeler tool, the below Relational model is built along with relationships using Entity Relationship diagram(ERD), where it will have Table and column details with proper relationship between the Entities as shown.



Generating DDL: The DDL scripts or Physical model is generated using this tool, which provides the scripts to run in relational database to create physical tables.



CREATE TABLE customer (

customer\_id Varchar(50)

NOT NULL,

first Varchar(250)

,

last Varchar(250)

,

age int

,

country Varchar(100)

,

age\_category Varchar(50)

);

COMMENT ON COLUMN customer.first IS

'First name';

ALTER TABLE customer ADD CONSTRAINT customer\_pk PRIMARY KEY ( customer\_id );

CREATE TABLE "Order" (

order\_id Varchar(50)

NOT NULL,

item Varchar(50)

,

amount Varchar(50)

,

customer\_id Varchar(50)

NOT NULL

);

ALTER TABLE "Order" ADD CONSTRAINT order\_pk PRIMARY KEY ( order\_id );

CREATE TABLE shipping (

shipping\_id Varchar(50)

NOT NULL,

status Varchar(50)

,

customer\_id Varchar(50)

NOT NULL

);

ALTER TABLE shipping ADD CONSTRAINT shipping\_pk PRIMARY KEY ( shipping\_id );

ALTER TABLE "Order"

ADD CONSTRAINT order\_customer\_fk FOREIGN KEY ( customer\_customer\_id )

REFERENCES customer ( customer\_id );

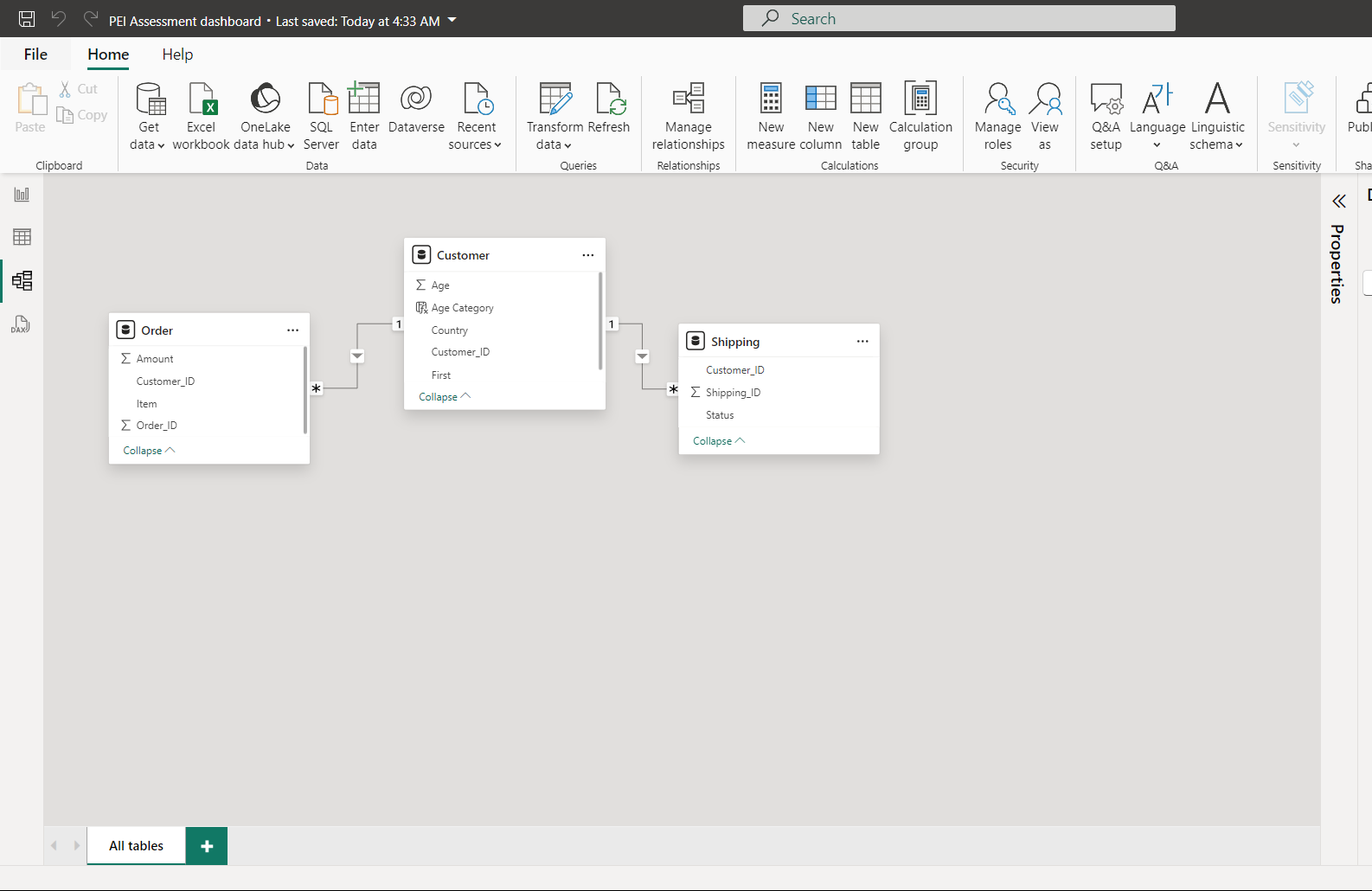
ALTER TABLE shipping

ADD CONSTRAINT shipping\_customer\_fk FOREIGN KEY ( customer\_customer\_id )

REFERENCES customer ( customer\_id );

**Power BI (Data transformation and Visualization):**

As discussed above that creating data model and run the DDL scripts to create tables in database is the procedure, I’m not taking that approach and I’ll be creating the data model and transformation in the Power BI itself using Power query editor tool as it’s more feasible in this scenario.



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

**Customers Table:**

* **Customer\_id:** Check for uniqueness and non-null values.
* **Name:** Ensure no missing values and validate against known patterns (e.g., no numbers).
* **Age:** Verify that values are within a reasonable range (e.g., 0-120).Treat the blank values with mean value
* **Country:** Ensure consistency in country names

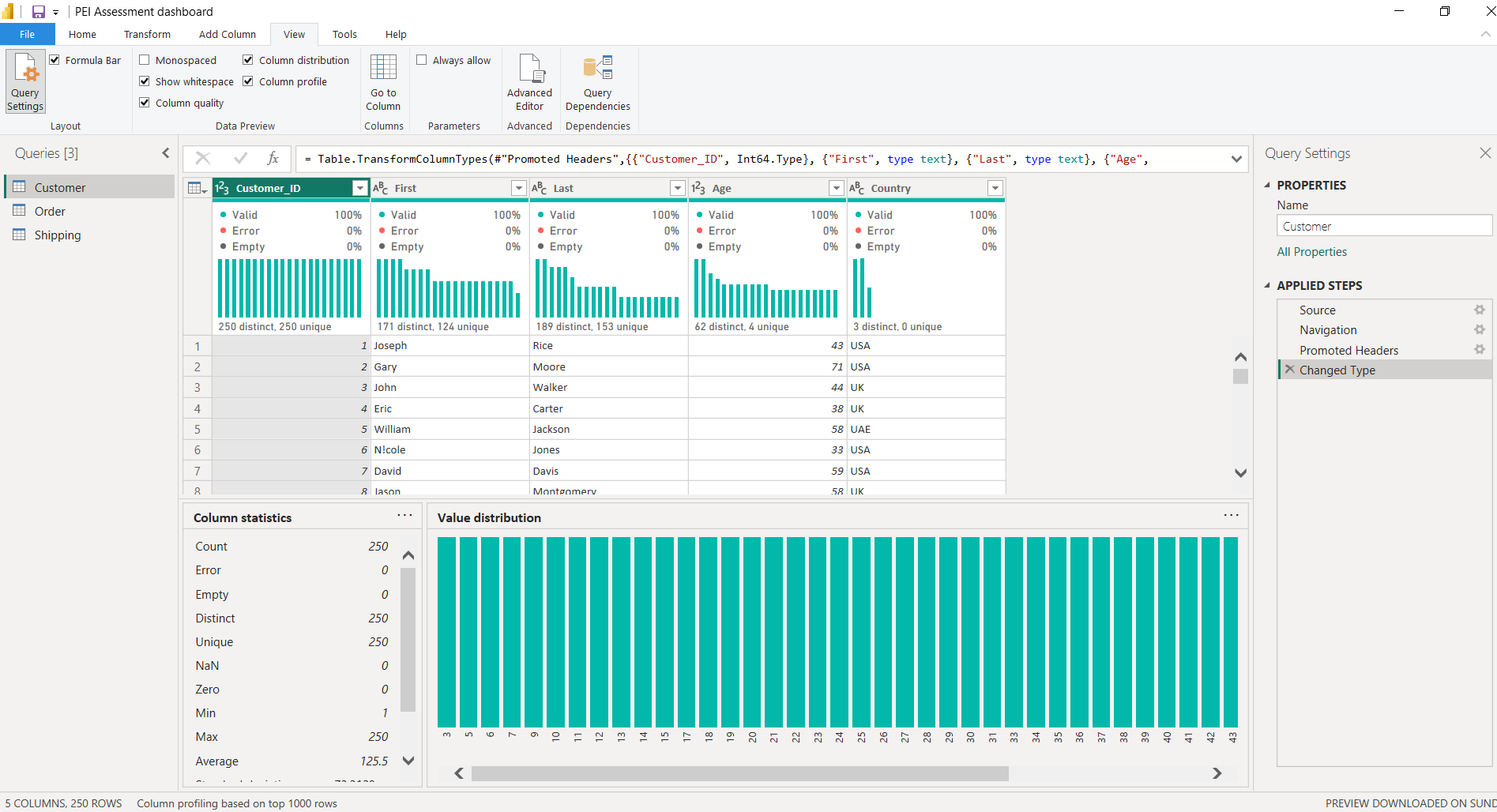
**Orders Table:**

* **Order\_id:** Check for uniqueness and non-null values.
* **Item:** Ensure no missing values.
* **Amount:** Validate that values are positive numbers.
* **Customer\_id:** Ensure it matches a valid Customer\_id in the Customers table.

**Shipment Table:**

* **Shipping\_id:** Check for uniqueness and non-null values.
* **Status:** Ensure it contains valid statuses (e.g., “Delivered”, “Pending”).
* **Customer\_id:** Ensure it matches a valid Customer\_id in the Customers table.

We can do Column quality, profiling check in the Power Query editor itself as shown for one example, using which we can treat the columns



**2. Define and Outline Requirements for Anticipated Datasets**

Based on the verification, the anticipated datasets should include:

* **Customer Demographics:** Customer\_id, Name, Age, Country.
* **Order Details:** Order\_id, Item, Amount, Customer\_id.
* **Shipment Status:** Shipping\_id, Status, Customer\_id.

**3. Develop Data Models and Map Existing Data Flows**

As discussed earlier,

**Conceptual Data Model:**

* **Entities:** Customers, Orders, Shipments.
* **Relationships:**
  + Customers have Orders (one-to-many).
  + Customers have Shipments (one-to-many).

**Logical Data Model:**

* **Customers Table:** Customer\_id (PK), Name, Age, Country.
* **Orders Table:** Order\_id (PK), Item, Amount, Customer\_id (FK).
* **Shipment Table:** Shipping\_id (PK), Status, Customer\_id (FK).

**Physical Data Model:**

* Implement the logical model in a database management system (DBMS) like SQL Server, MySQL, etc.

**4. Technical Specifications for Data Engineer**

**Example: Orders Table**

* **Table Name:** Orders
* **Columns:**
  + **Order\_id:** INT, PRIMARY KEY, AUTO\_INCREMENT
  + **Item:** VARCHAR(100), NOT NULL
  + **Amount:** INT NOT NULL
  + **Customer\_id:** INT, FOREIGN KEY REFERENCES Customers(Customer\_id)

**5. Communicate Findings and Insights**

**Visual Representation:**

* **ER Diagram:** Show entities (Customers, Orders, Shipments) and their relationships.
* **Data Flow Diagram:** Illustrate how data moves between tables and processes.

**Dashboard:**

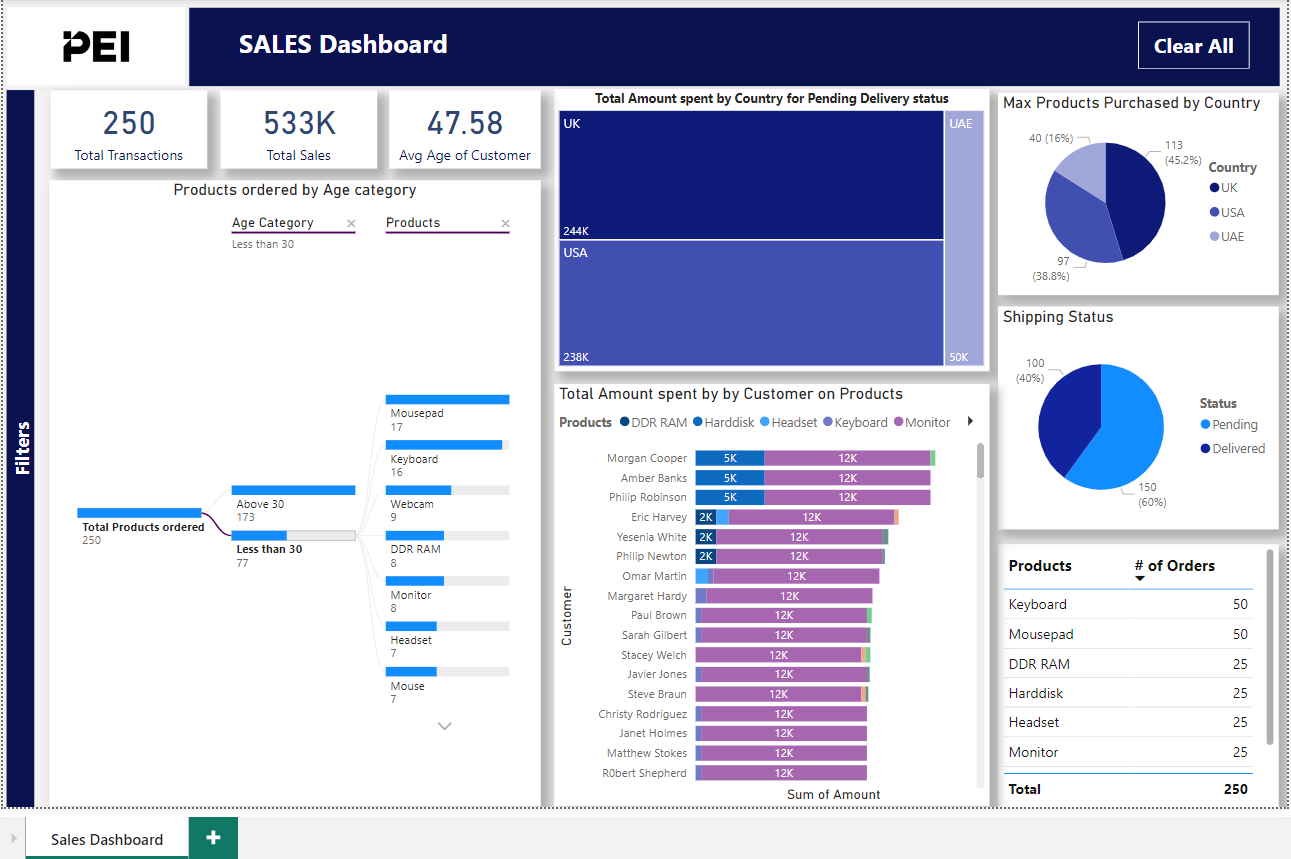
* + Find more details in the attached pbix (Power bi dashboard)

**Dashboard:**

Attached the dashboard(pbix) for analysis.

****

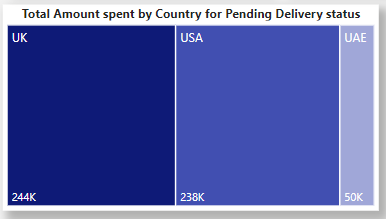
Below is the screen shot from the dashboard. Towards left side there is filter button, Ctrl + Click will open a window to use different slicers. Also at top right “Clear All” button will clear all the filter selections made in the page



**Findings and requirements:**

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

Using Treemap, so when there are multiple countries, its easy to visualize the countries with higher Total Amount spent for Pending Delivery status

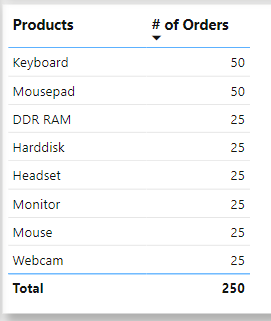


1. *the total number of transactions, total quantity sold, and total amount spent for each customer, along with the product details.*

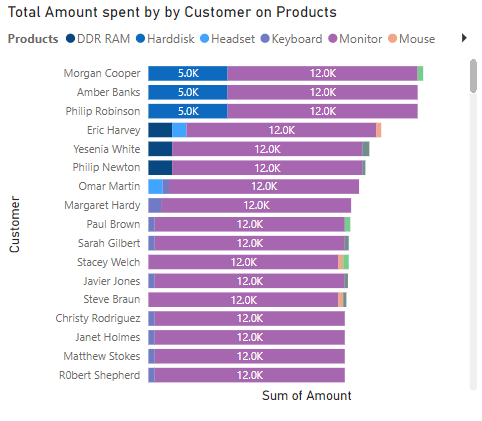
Total transactions or total amount spent



Total quantity sold by products

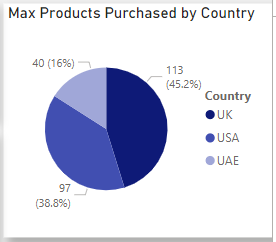


Total Amount spent by Customers on products individually



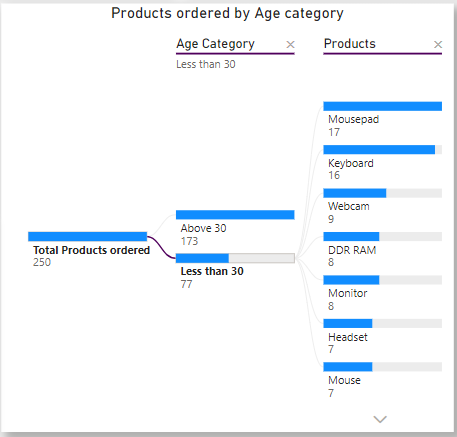
1. *the maximum product purchased for each country.*

Maximum products purchased by Country



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

Products ordered/purchased by Age less than 30. This Decomposition tree is flexible to analyze for both age category as shown below. Just click on + to expand it and X to close it



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

*This shows country with least sales amount*

