**1.Verify the accuracy, completeness, and reliability of source data.**

*Customer Table:* 1. Customer Id- Unique Identifier as there are 250 distinct and 250 unique rows.

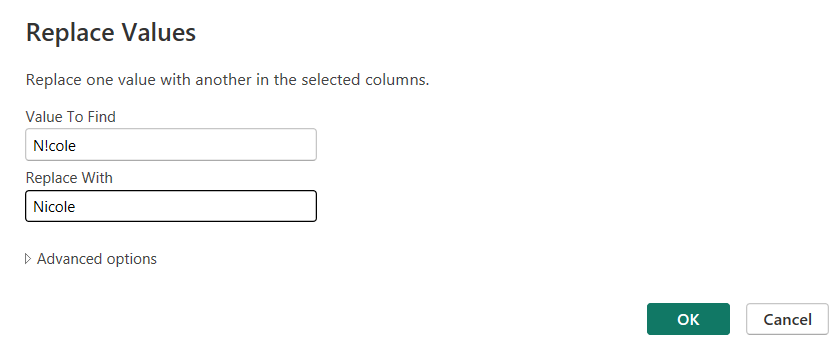
2. No missing values found.

3. Renamed First to FirstName and Last to LastName for better understanding.

4. Checked data type for columns, they are correct:

Customer Id-Numeric, FirstName -Text, LastName-Text, Age- Numeric, Country-Text

5. Few values for FirstName and LastName column are incorrect/inaccurate, replaced N!cole to Nicole, L@rry to Larry, R0bert to Robert, and Al1cia to Alicia.

A screenshot of a computer

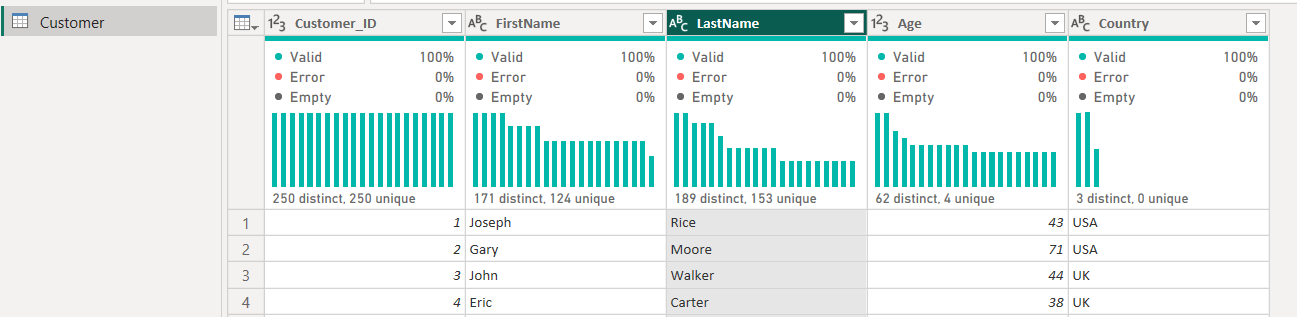
Description automatically generatedA screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

6. Data is now clean and error free, and ready to use for visualization.



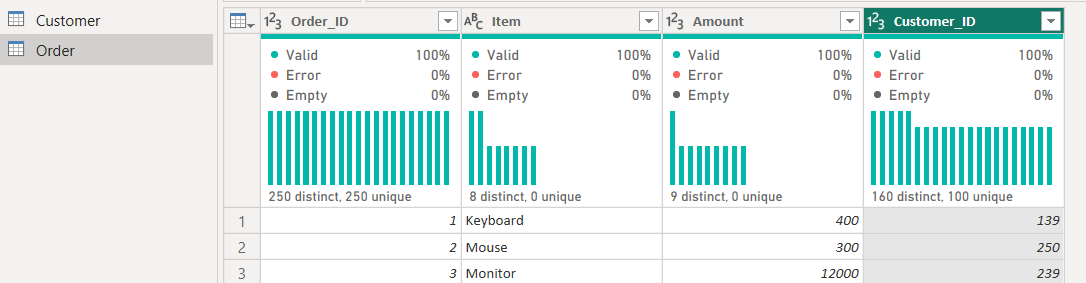
*Orders Table:* 1. Order Id- Unique Identifier as there are 250 distinct and 250 unique rows and Customer ID is Foreign Key.

2. No missing values found.

3. Checked data type for columns, they are correct:

Order Id - Numeric, Item - Text, Amount - Numeric, Customer Id – Numeric

4. Data is now clean and error free, and ready to use for visualization.



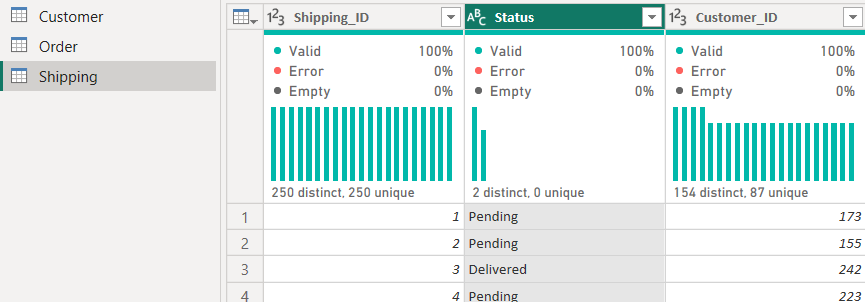
*Shipping Table:* 1. Shipping ID - Unique Identifier as there 250 distinct and unique rows and Customer ID is Foreign Key.

2. No missing values found.

3. Checked data type for columns, they are correct:

Shipping ID - Numeric, Status - Text, Customer ID-Numeric

4. Data is now clean and error free, and ready to use for visualization.



*Observation*: In shipping table, we have customer ID but not Order ID. It would have been more reliable and complete if we have Order ID in shipping table. If customer has more than one order than based on current mapping (Customer\_ID 🡪 Shipping\_ID) we will not be able to get shipping information for a specific order.

**2.Based on your findings, define and outline the requirements for anticipated datasets, detailing the necessary data components.**

**3.Develop the data models to effectively organize and structure the information and provide a detailed mapping of existing data flows, focusing on the areas of concern.**

Below is the answer for both the questions.

*Data Model/Data flow:*

1.Identify the data structure and fields – Customer (XLS), Order (CSV), Shipping (JSON).

2.Design the unified schema.

3.Creating the data model- Create three SQL tables using CREATE and define the relation among them.

Data Components:

Customer Entity:

Attributes: Customer\_ID, Name, Email, Address

Order Entity:

Attributes: Order\_ID, Item, Amount, Customer\_ID

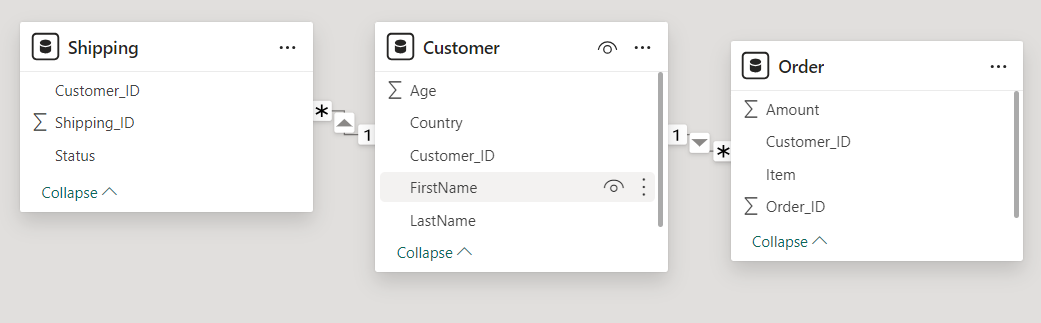
Relationship: Many Orders belong to one Customer.

Shipping Entity:

Attributes: Shipping\_ID, Status, Customer\_ID

Relationship: Many Shipping’s belong to one Customer

Entity Relation diagram:



4. Import the data directly into the tables that we created or populate the tables with values using INSERT SQL function, former one seems ideal in case of huge data.

5. Use the schema under which you have created the tables as a single data source for visualization.

*Area of concerns:*

1.Data quality (Accuracy, Completeness, Consistency)

2.Data Integration (Schema matching, data merging)

3.Data Redundancy and Normalization

4.Error handling

5.Performance and Scalability

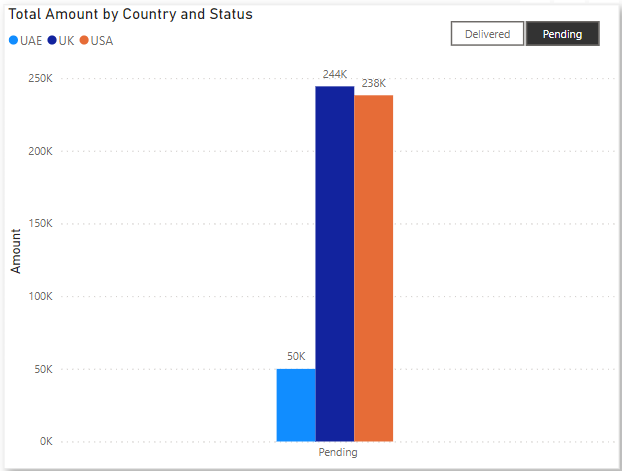
5.Documentation and Maintenance

**4.Communicate the findings and insights to stakeholders in a visually comprehensive manner.**

UsedPower BI as it is easy to connect to three different data sources and uncover insights from it.

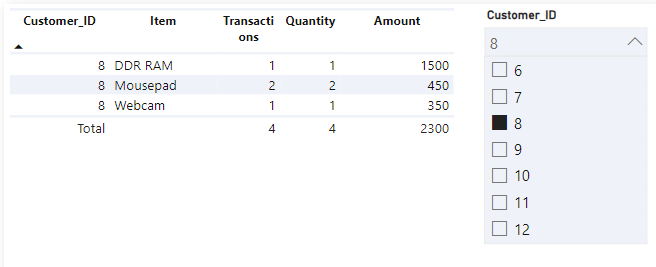
1.The total amount spent and the country for the pending delivery status for each country.

* Used Clustered column chart to represent status and country, Slicer to exchange the data between delivered and pending status.



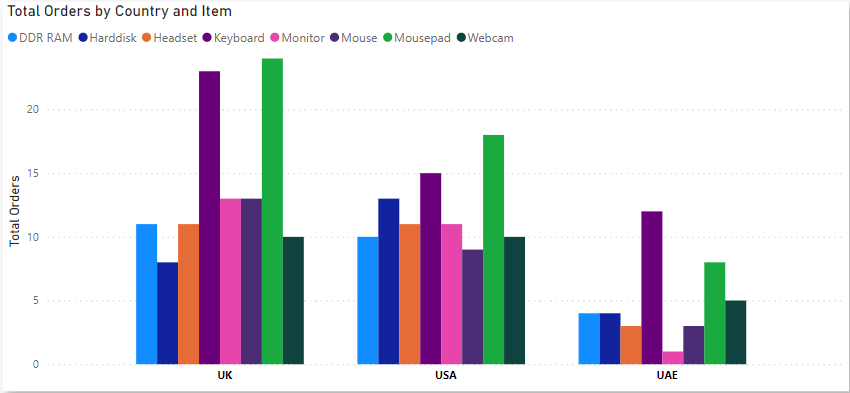
2.The total number of transactions, total quantity sold, and total amount spent for each customer, along with product details.

* Used table visualization for Customer, Item, and quantity details. User can use slicer to filter on customer details.

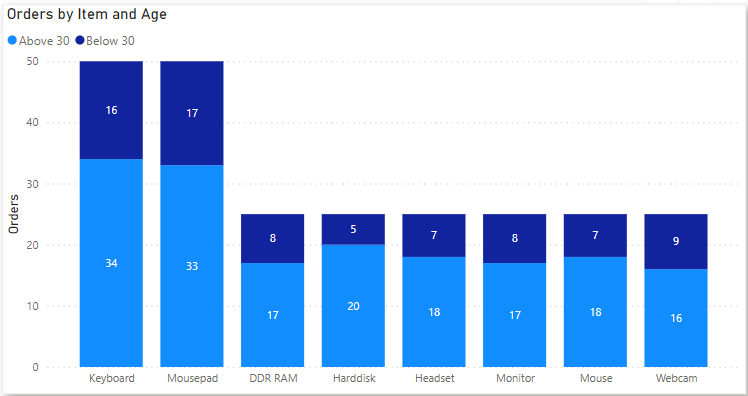


3. The maximum product purchased for each country.

* Used clustered column chart to depict maximum product purchased by country, by looking at below visualization we can say Mousepad is max product in UK, USA, and Mouse in UAE.

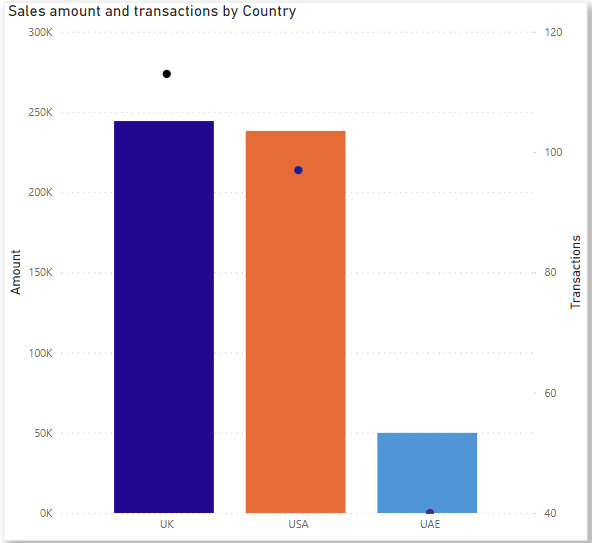


4. The most purchased product based on the age category less than 30 and above 30.

* Used Stacked Column Chart to show product and age details, by looking at below visualization we can say Mousepad is most purchased for age below 30 and Keyboard for above 30.
* Used DAX function to create new measure for age (If age<=30, below 30, above 30). 

5. The country that had minimum transactions and sales amount.

* Used Line and Stacked column chart to show both amount and order values, we can say UAE is the country with minimum transactions and sales amount.



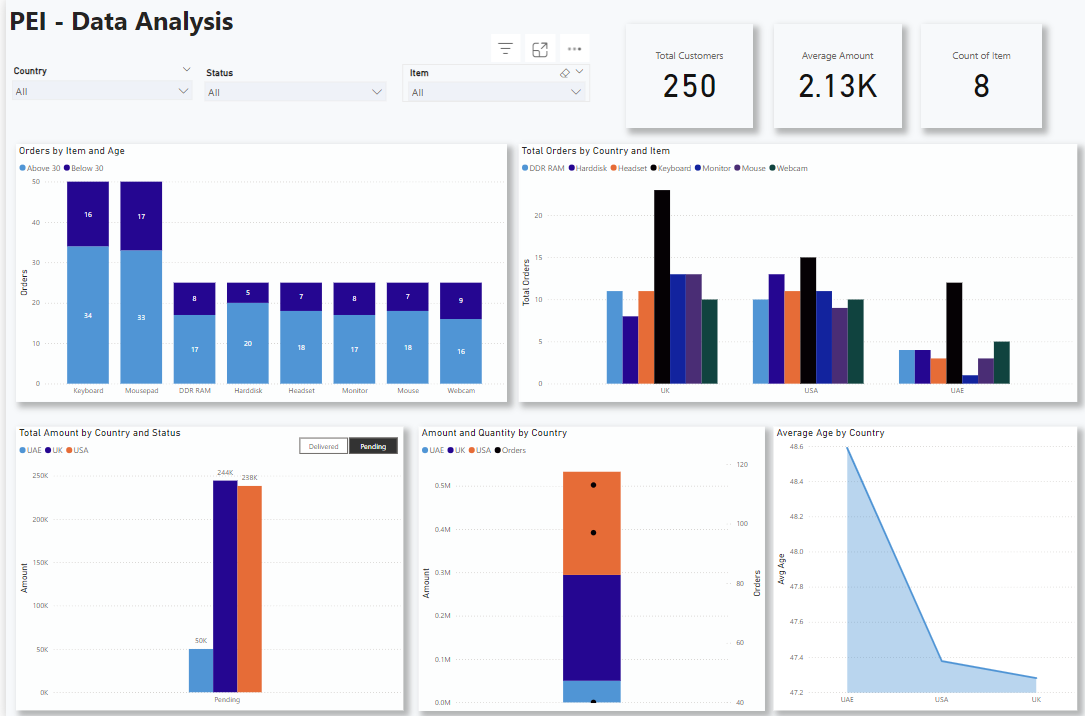
Final sample Dashboard to send to users:

1.

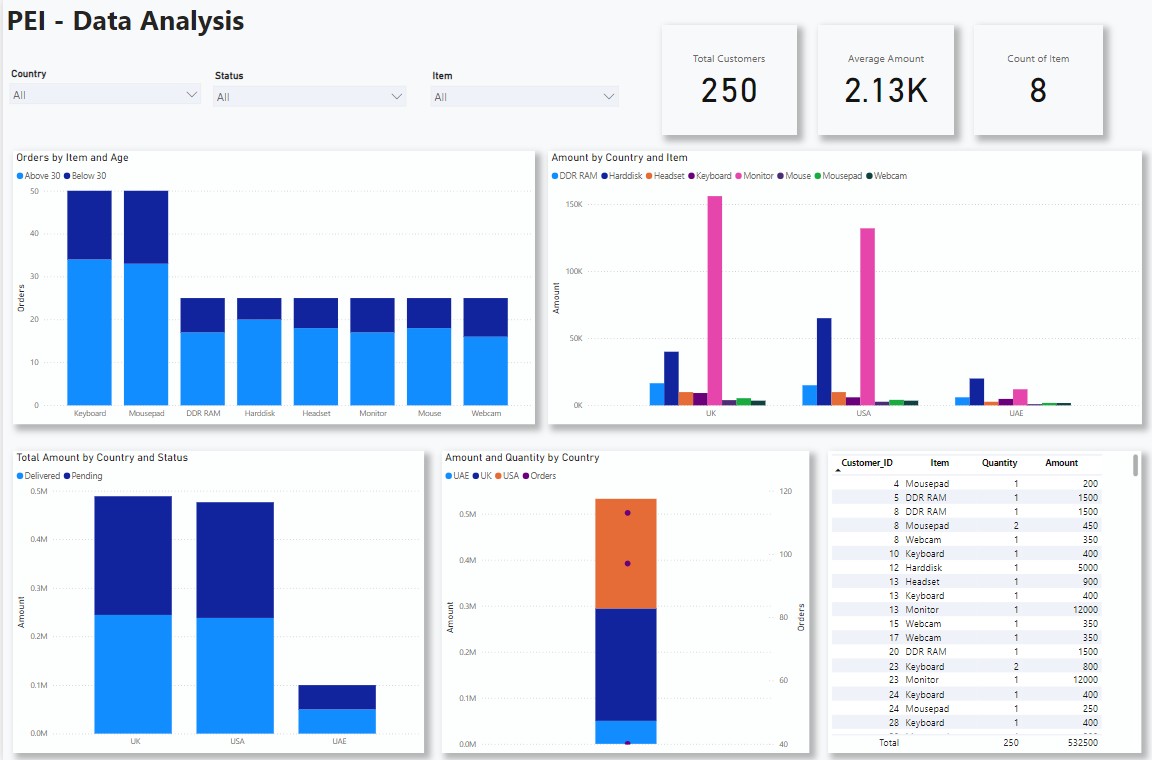
A screenshot of a computer

Description automatically generated

2.



3.



Based on the user requirement we can make changes to the visualizations, also we can build the same in tableau as well.

**5.What will be your insights to other peer teams of Data Engineers, Data Scientists, and other technical and non-technical stakeholders.**

**Data Engineers:** My suggestion for data engineers would be to combine all three data sources into single source to ensure ETL processes are robust to handle data integration.

1.Using DBeaver tool.

Explanation: Using DBeaver tool import data from each source (excel, csv, JSON) and load into tables and then connect to snowflake or SQL server management studio through data connection. The tables will automatically reflect in PROD environment of snowflake/SQL server.

2.Using Azure Databricks

Explanation: Import data from three different sources and write to snowflake/SQL server by creating databricks notebook.

**Data Scientists:**

Dopredictive analytics on customer orders for each country and by product. Use “Where we are” visualization insights to predict “where we can in future/where we should be”.

**Applicable insights that we can communicate with technical and non-technical stakeholders:**

1.Build Delivery Performance metrics (Average delivery time) and show them by status, Country, and Item.

2.Breakdown of shipping status (pending and delivery) by country and product.

3.If we have date column like order date, communicate to users about orders over time(monthly/quarterly/yearly) trend.

4.Create a line chart to show them previous, actual and target sales amount difference by adding performance year as a filter.

5.Customer distribution by location.

6.Performance of product in each country using sales amount as value.

7.Average age depiction by country and item.

8.Bar charts for average order value by product segment.

9.Pie/Bar charts for shipping status distribution.