**Sales Data-Analysis**

Data Verification

* Three distinct data sources—Order (CSV), Shipping (JSON), and Customers (Excel)—have been provided for analysis.
* Data normalization and verification were conducted using Python to ensure consistency across formats thus enabling streamlined and accurate analytical processing.
* Below Python commands were utilized to structure and prepare it for further business intelligence activities.  
  A group of green text

  AI-generated content may be incorrect.
* Next, we check for Missing Values in all 3 data sources

A screen shot of a computer code

AI-generated content may be incorrect.  
A white background with black text

AI-generated content may be incorrect.

A screen shot of a computer

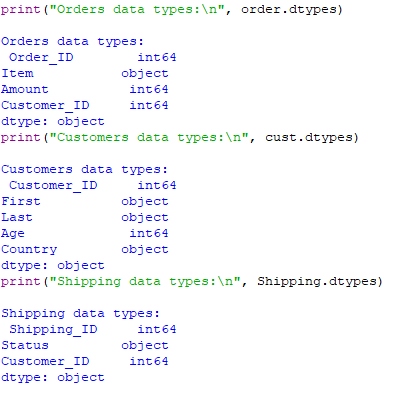
AI-generated content may be incorrect.

* Duplicate records were proactively identified and removed, where present, to eliminate data redundancy and enhance the overall quality

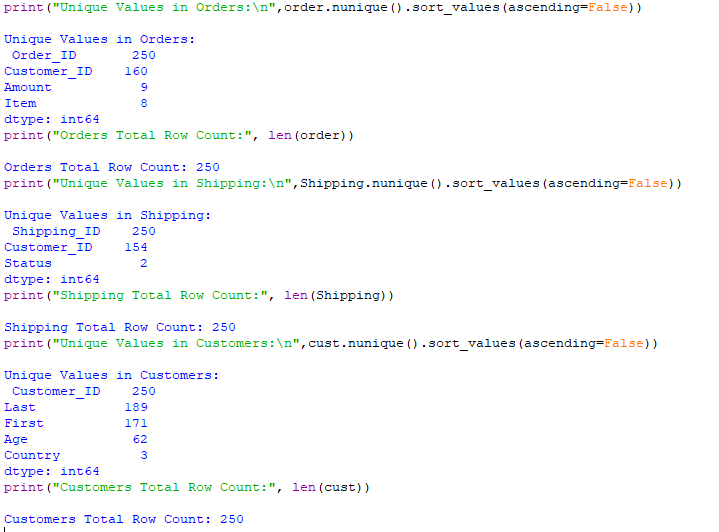
A white background with black text

AI-generated content may be incorrect.

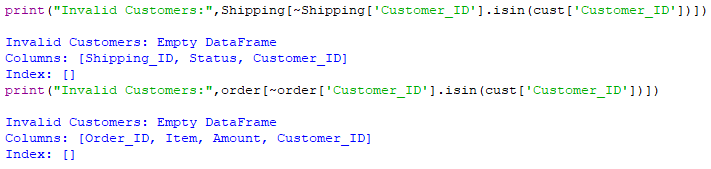
* Data types for each column are verified



* Each data source should have a Primary Key column which is Non-Null and Unique, next step is to ensure this constraint is met in all 3 data sources. We compare the number of Unique and non-null values with the complete row count in each.



* Hence, by observation **Order\_ID** from Orders, **Shipping\_ID** from Shipping and **Customer\_ID** from Customers data can be used as **Primary Keys**.
* Next, we check the referential Integrity Constraint, to ensure there is no Customer ID in Orders whose details are not present in Customers data and similarly in Shipping data



* Above output indicates that Customers table has information about all customers who have placed an order or whose order has been shipped
* Upon validation, it was observed that certain customers listed in the Orders dataset do not have corresponding Shipping details, and vice versa, indicating incomplete data linkage across sources.A screenshot of a computer

  AI-generated content may be incorrect.

Conclusions: The 3 datasets have non-null values with no duplicates and **Order\_ID** from Orders, **Shipping\_ID** from Shipping and **Customer\_ID** from Customers data can be used as **Primary Keys**. With **Customer\_ID** being the foreign key in Orders and Shipping data. The datatype of each column is known and also we noticed that Orders and Shipping data cannot be a complete join.

ER Diagram

The below represents an Entity-Relation Diagram which serves as a reference for the Data Engineer to build the tables.  
  
A screenshot of a computer

AI-generated content may be incorrect.

Customers:

* The Customer\_ID is the primary key.
* A customer can have multiple orders and shipping records, but each record in Orders and Shipping refers to exactly one customer (1:N relationship).

Orders:

* The Order\_ID is the primary key.
* Each order is associated with a Customer\_ID (foreign key from Customers).
* A customer can place multiple orders, but an order is associated with exactly one customer.

Shipping:

* The Shipping\_ID is the primary key.
* Each shipping record is associated with a Customer\_ID (foreign key from Customers).
* A customer can have multiple shipping records, but a shipping record is associated with exactly one customer.
* As depicted in the diagram, the Customer\_ID fields in both the Orders and Shipping dataset’s function acts as foreign key columns referencing the Customers dataset. These fields are expected to be non-null, to maintain referential integrity.
* Additionally, Quality Assurance Engineers should ensure that these columns adhere to the defined data types during testing to uphold data consistency and schema compliance.

Part-2

Business Reporting Queries

To execute the Queries 3 tables : CUSTOMERS\_TEST, ORDERS\_TEST, SHIPPING\_TEST is created as below using SQL

A screen shot of a computer program

AI-generated content may be incorrect.

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

A screenshot of a computer

AI-generated content may be incorrect.

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

A screenshot of a computer

AI-generated content may be incorrect.

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

A screenshot of a computer program

AI-generated content may be incorrect.

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

A screenshot of a computer

AI-generated content may be incorrect.

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

A screenshot of a computer

AI-generated content may be incorrect.