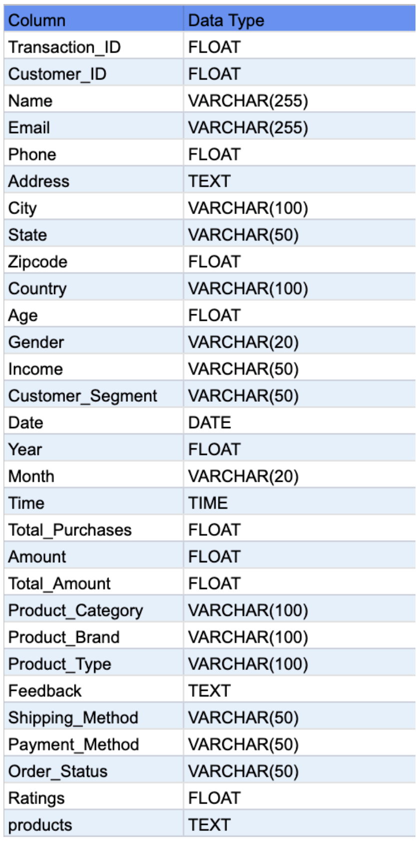
Data Structure Overview

The dataset represents retail transactional data taken from Kaggle for a hypothetical clothing startup, Lokii. It includes comprehensive information about customers, their purchases, and transaction details. The key attributes in the dataset are:

- **Customer Information:**
  - Customer ID, Name, Email, Phone, Address, City, State, Zipcode, Country
  - Age, Gender, Income, Customer Segment
- **Transaction Details:**
  - Last Purchase Date, Total Purchases, Amount Spent
- **Product Information:**
  - Product Category, Product Brand, Product Type
- **Additional Attributes:**
  - Feedback, Shipping Method, Payment Method, Order Status



Column	Data Type
Transaction_ID	FLOAT
Customer_ID	FLOAT
Name	VARCHAR(255)
Email	VARCHAR(255)
Phone	FLOAT
Address	TEXT
City	VARCHAR(100)
State	VARCHAR(50)
Zipcode	FLOAT
Country	VARCHAR(100)
Age	FLOAT
Gender	VARCHAR(20)
Income	VARCHAR(50)
Customer_Segment	VARCHAR(50)
Date	DATE
Year	FLOAT
Month	VARCHAR(20)
Time	TIME
Total_Purchases	FLOAT
Amount	FLOAT
Total_Amount	FLOAT
Product_Category	VARCHAR(100)
Product_Brand	VARCHAR(100)
Product_Type	VARCHAR(100)
Feedback	TEXT
Shipping_Method	VARCHAR(50)
Payment_Method	VARCHAR(50)
Order_Status	VARCHAR(50)
Ratings	FLOAT
products	TEXT

Data Issues and Inconsistencies

Upon initial inspection, several issues and inconsistencies were identified in the dataset:

1. **Erroneous Values:**
   - Date columns (`Date`, `Month`, `Year`) contained mismatched values, including future dates (up to December 2024).
   - Product categories and brands were incorrectly placed.
   - Age values were unrealistic or inconsistent.
2. **Duplicate Entries:**
   - Duplicate Transaction IDs and Customer IDs were found, which should be unique.
3. **Data Discrepancies:**
   - Incorrect city, state, and country combinations.
   - Incorrect data types, particularly for decimal values.
4. **Missing Values:**
   - Several columns contained null values that required imputation or removal.

Data Cleaning and Preparation Steps

To ensure the integrity and reliability of the dataset, the following cleaning steps were performed using Python:

1. **Handling Missing Values:**
   - Used **mode imputation** for categorical variables and **mean imputation** for numerical variables where relevant.
   - Irrelevant columns with a high proportion of missing data were removed.
2. **Data Type Conversion:**
   - Converted data types to ensure consistency, especially for numeric and date fields.
3. **Date Mismatch Resolution:**
   - Carefully extracted the correct date values from the "Date" column and synchronized the "Month" and "Year" columns accordingly.
4. **City-State-Country Corrections:**
   - Monitored and studied these columns together to detect and correct inconsistencies early on.
5. **Duplicate Removal:**
   - Identified and removed duplicate Transaction IDs and Customer IDs to ensure data uniqueness.

By addressing these issues, the dataset was transformed into a clean, reliable resource, ready for further analysis and visualization.

Libraries used: Pandas, Numpy, matplotlib (for EDA)