**Project Background**

This is a retail sales data cleaning and Exploratory Data Analysis (EDA) project based upon the an abridged version of this dataset: <https://www.kaggle.com/datasets/sahilprajapati143/retail-analysis-large-dataset/data>

I have taken a sample of 9,880 rows for example purposes and worked on the following:

* **Data cleaning** (including:
* creating a staging copy of the database table,
* splitting fields into more logical/understandable names using regular expressions
* reformatted some numbers with rounding for readability
* appended phone numbers and date formats with SUBSTRING functionalities
* appended blank values to NULLs, sought to update these fields/rows where possible and found and removed duplicates using window functions.
* **Exploratory Data Analysis (EDA)** where analysis was conducted with the following insights to be gleaned, in focus:

**Sales Performance Insights**

* Which products or services are generating the most revenue?
* How many products in total were sold in 2023 vs 2024? Year on year (YOY)

comparison

* What is the average order value (AOV)?

**Customer Insights**

* Do top spending customers follow a particular: gender, age and/or income

pattern?

* Percentages of customers by: gender, age and income
* How much revenue is driven by repeat vs new customers?

**Product (Market) Trends**

* Which products are underperforming?
* Which product categories are performing? and which are not?
* Are there particular brands performing better/worse than others?

And any other, not previously noted, notable trends in general prevailing throughout the data.

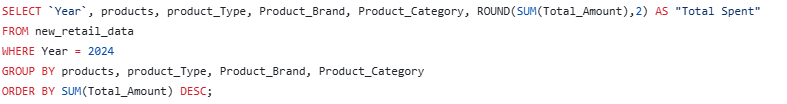
**Data Structure**

|  |  |
| --- | --- |
| **Table name**: **new\_retail\_data** |  |
|  |  |
| **Columns:** | **Data Type** |
| Transaction\_ID | int |
| Customer\_ID | int |
| Email | text |
| Phone | varchar(12) |
| Full\_Address | text |
| City | text |
| State | text |
| Zipcode | int |
| Country | text |
| Age | int |
| Gender | text |
| Income | text |
| Customer\_Segment | text |
| Date | text |
| Year | int |
| Month | text |
| Time | text |
| Total\_Purchases | int |
| Amount | double |
| Total\_Amount | double |
| Product\_Category | text |
| Product\_Brand | text |
| Product\_Type | text |
| Feedback | text |
| Shipping\_Method | text |
| Payment\_Method | text |
| Order\_Status | text |
| Ratings | int |
| products | text |
| first\_name | varchar(60) |
| last\_name | varchar(60) |
| street\_number | varchar(10) |
| street\_name | varchar(60) |
| Apt\_Number | varchar(10) |
| Suite | varchar(60) |
| Transaction\_Date | date |
| New\_Customer\_ID | varchar(50) |

**Executive Summary**

**Key Insights Obtained**

This section will detail the different SQL queries used and the corresponding resulting information and insights that were uncovered from the queries:

**Products Generating the Most Revenue in 2024** 

****

* (Being the tough year that it was for some... perhaps unsurprisingly 😊) Coffee appears to be the top selling product for 2024 followed by thriller novels, ginger ale, t-shirts and mineral water
* Interestingly Nestle and Pepsi brands featured predominantly in the top 10 highest grossing products

**Products That Generated the Most Revenue in 2023**

**A close-up of a text

Description automatically generated**

**A screenshot of a product list

Description automatically generated**

**Products That Generated the Most Revenue in 2023**

**Recommendations**