Advanced Analytics & Data Visualisation Project

RETAIL DATA ANALYSIS

By: Ishita Shukla

Under the mentorship of Mr.Manoj K, EY India

Summer Internship May-July, 2025

TABLE OF CONTENT

DATASET OVERVIEW

U	V	W	X	Y	Z	AA	AB
l_Am	Product_C	Product_B	Product_T	Feedback	Shipping_I	Payment_	Order
.0863	Clothing	Nike	Shorts	Excellent	Same-Day	Debit Carc	Shippe
.7078	Electronic	Samsung	Tablet	Excellent	Standard	Credit Car	Proces
3.433	Books	Penguin B	Children's	Average	Same-Day	Credit Car	Proces
6.854	Home Dec	Home Dep	Tools	Excellent	Standard	PayPal	Proces
8.553	Grocery	Nestle	Chocolate	Bad	Standard	Cash	Shippe
5.167	Electronic	Apple	Tablet	Good	Express	PayPal	Pendi
.1153	Electronic	Samsung	Television	Bad	Standard	Cash	Proces
58807	Clothing	Zara	Shirt	Bad	Same-Day	Cash	Proces
0.714	Grocery	Nestle	Chocolate	Bad	Same-Day	Cash	Delive
6.112	Home Dec	Home Dep	Decoratio	Excellent	Standard	Cash	Delive
.9275	Home Dec	Home Dep	Tools	Average	Standard	Credit Car	Shippe
.8306	Books	Random H	Non-Fictio	Average	Standard	Credit Car	Pendi
8.794	Grocery	Coca-Cola	Water	Bad	Standard	PayPal	Delive
7.353	Grocery	Nestle	Snacks	Excellent	Express	PayPal	Delive
6.356	Clothing	Adidas	T-shirt	Bad	Same-Day	Cash	Shippe
.1717	Books	Random H	Literature	Bad	Express	Credit Car	Pendi
.2716	Grocery	Pepsi	Water	Average	Same-Day	Debit Carc	Shippe
.3606	Electronic	Apple	Tablet	Good	Express	Cash	Shippe
.1379	Grocery	Coca-Cola	Juice	Bad	Express	Credit Car	Shippe
.3193	Home Dec	IKEA	Furniture	Average	Standard	Cash	Shippe
25.25	Grocery	Nestle	Coffee	Average	Same-Day	PayPal	Proces
.1485	Books	HarperCol	Non-Fictio	Average	Standard	PayPal	Proces
.7456	Books	Penguin B	Non-Fictio	Average	Same-Day	PayPal	Proces
2.863	Home Dec	Bed Bath &	Bathroom	Good	Standard	PayPal	Pendi
.5967	Grocery	Coca-Cola	Juice	Bad	Same-Day	PayPal	Proces
.2596	Home Dec	Bed Bath &	Kitchen	Bad	Standard	Debit Carc	Pendi
.5707	Electronic	Samsung	Smartpho	Good	Express	PayPal	Shippe
7.411	Electronic	Apple	Tablet	Bad	Same-Day	Credit Car	Shippe

Source

retail_data.csv contained 20+ columns with 10k+ records. It's key fields being Transaction ID, Product Categories, Types and Brands, Total Revenue generated, Customer details like income, age segments etc.

Initial Observed Issues

-Income column contains mixed formats ("HIGH" , "$40,000", etc.)
-Date column has inconsistent formats and missing values.
-Feedback column had null values.
-Ratings column had some outlier values.
-Duplicate Customer_IDs and Transaction_IDs.

Purpose of Dataset

The dataset simulates sales, operational and customer satisfaction data across multiple countries, intended to be analysed for:
-Sales trends
-Customer Segmentation
-Product Performance
-Delivery & Payment Efficiency

DATA CLEANING & PREPARATION

STEP 1: PYTHON

– Explored the dataset shape and structure using df.info(), df.describe(), and df.isnull().sum().

– Standardized the Date column using pd.to_datetime() for uniform format.

– Cleaned the Income column by removing currency symbols, converting strings to numeric, mapping values like high, medium, low to numeric values.

– Handled missing values in Feedback, Ratings, and Income columns.

```
Shape: (302010, 30)

Column Names:
['Transaction_ID', 'Customer_ID',

Data Types:
Transaction_ID      float64
Customer_ID         float64
```

```
Missing Values Report:
                Missing Values  Percentage (%)
Transaction_ID             333            0.11
Customer_ID                308            0.10
Name                       382            0.13
Email                      347            0.11
```

```
df['Date'] = pd.to_datetime(df['Date'], errors='coerce')
df['Year'] = df['Date'].dt.year
df['Month'] = df['Date'].dt.month
df['Weekday'] = df['Date'].dt.day_name()
```

```
df['Ratings'].fillna(df['Ratings'].mean(), inplace=True)
df['Gender'].fillna(df['Gender'].mode()[0], inplace=True)
df['Feedback'].fillna('Unknown', inplace=True)
```

```
    if isinstance(val, str):
        income_map = {'Low': 1, 'Medium': 2, 'High': 3}
        val = val.strip()
        if val in income_map:
            return income_map[val]
        else:
            try:
                return float(val.replace('$', '').replace(',', ''))
            except ValueError:
                return np.nan
```

DATA CLEANING & PREPARATION

STEP 2: SQL

– Imported the cleaned CSV into SQL table, verified data types, ran SELECT TOP 100* FROM RetailSales to inspect raw imported data.

– Used ALTER TABLE to convert Customer_ID and AGE to INT and formatting Ratings as FLOAT.

– Calculated Revenue by Product Category:

```
SELECT Product_Category, SUM(Total_Amount) AS Revenue FROM Fact_Retail
GROUP BY Product_Category;
```

– Calculated Average Ratings by Country:

```
SELECT Country, AVG(Ratings) AS AvgRating FROM Dim_Location L
JOIN Fact_Retail R ON L.Location_ID=R.Location_ID
GROUP BY Country;
```

– Designed a STAR SCHEMA for scalable reporting in Power BI.

Fact Table: Fact_Retail containing transactional data (amount, ratings, etc.)

Dim Table:

1. Dim_Customer containing customer details name, gender, age, income label.
2. Dim_Product containing product info like category, brand, type.
3. Dim_Date having calendar info like year, quarter, month, weekday.
4. Dim_Location having geographical details country, state, city, zip.

DATA VISUALISATION

The Data Visualisation part was done by using Power BI Dashboards. They were created using the cleaned retail dataset to derive business insights for decision making. The report focuses on key areas such as revenue trends, customer behvaious, product performance, and operational metrics.
The dashboard is divided into two well-structured pages:

1. Revenue and Orders Overview

It answers questions like:

How are we performing?

Which categories or brands are driving revenue?

What regions and methods are most effective?

2. Customer Behavior & Product Insights

It answers questions like:

Who are our customers?

How do they rate us?

Which products perform best?

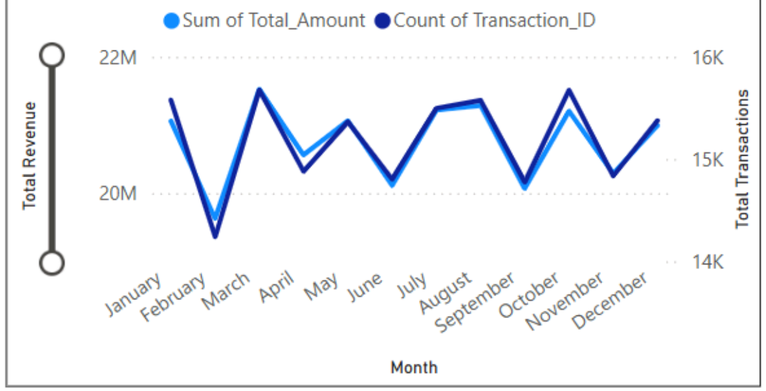
What kind of feedback are we receiving?

PAGE 1: REVENUE OVERVIEW

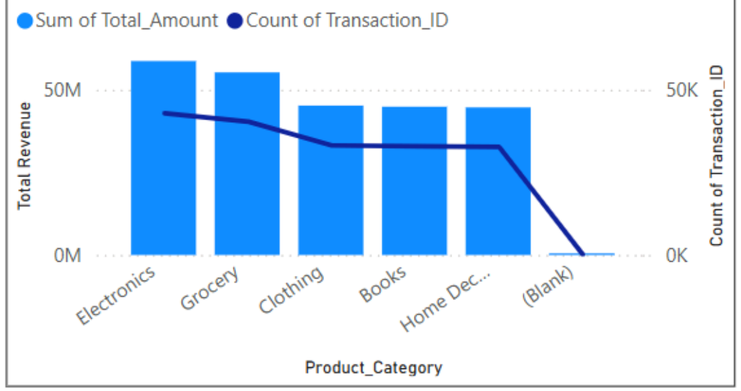
REVENUE AND ORDERS OVERVIEW

Total Revenue	Total Orders	Average Order Value	Unique Customers	Average Ratings
249.00M	182.27K	1.37K	77.80K	3.17

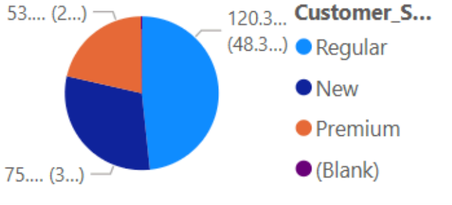
Revenue Trend



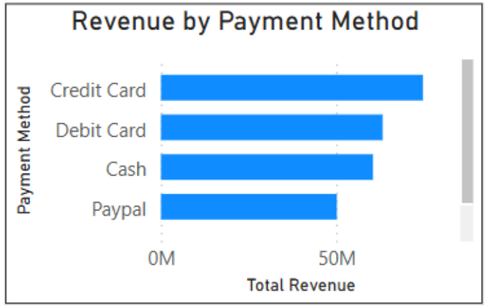
Revenue by Product Category



Revenue by Customer Segment



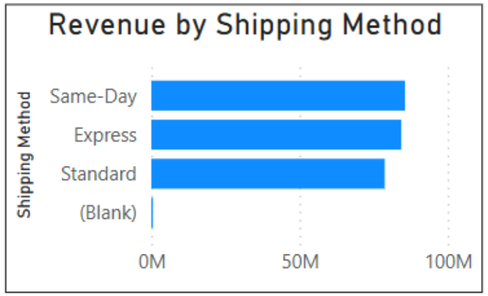
Revenue by Payment Method



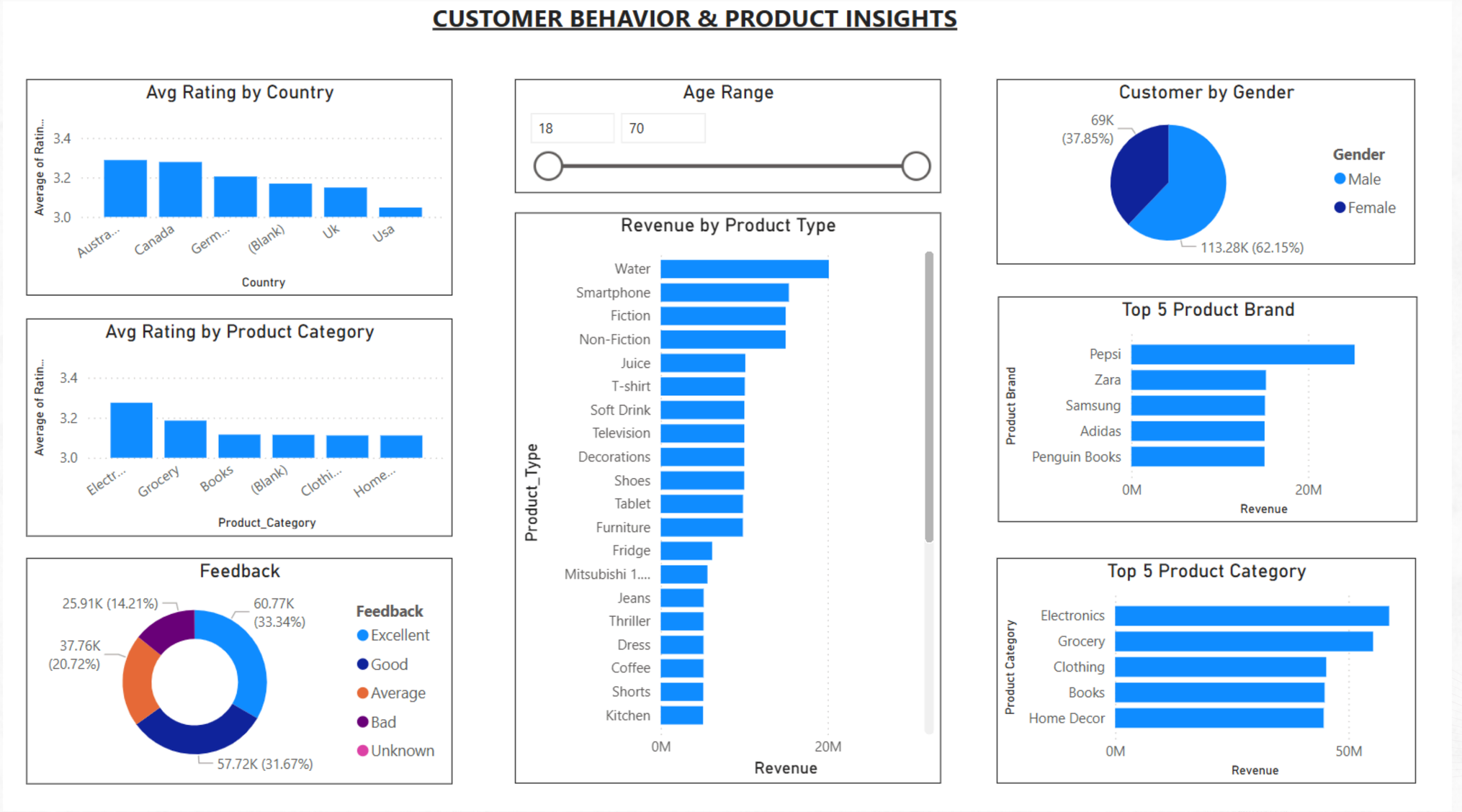
Geographical Location

Product Brand

Revenue by Shipping Method



PAGE 2: CUSTOMER & PRODUCT INSIGHTS



KEY INSIGHTS

1. Revenue

– The total revenue generated is $249 million from 182,270+ orders with major sales activity in North America, Europe and Australia.
– USA contributes significantly to the revenue but has the lowest average customer rating.
– Same-Day and Express Shipping methods generate higher revenues compared to Standard indicating customer's preference.

2. Customer Segments

– Female customers represent approx 62% of the customer base and contribute to more revenue than male.
– Age 25-40 group is the most engaged and profitable, both in terms of frequency and order value.
– Feedback distribution is 33% Excellent, so customer experience can be improved.

3. Product Performance

– Electronics is the leading brand in revenue and order volume. Average rating also highest for electronics.
– Top 5 brands are Pepsi, Zara, Samsung, Adidas and top products are Water, Smartphones and Fiction books.

4. Time-wise Analysis

– Quarter 3 records the highest revenue likely due to seasonal demand, festive and vacation season.
– Quarter 2 records the lowest sales.

THANK YOU